Sampling Designs for Exploratory Multivariate Analysis

by

Julie Anne Hopkins

Thesis submitted to the University of Sheffield for the degree
of
Doctor of Philosophy

Department of Probability and Statistics

Submitted: November 1999
Accepted: 2000

Sampling Designs for Exploratory Multivariate Analysis

Julie Anne Hopkins

Summary

This thesis is concerned with problems of variable selection, influence of sample size and related issues in the applications of various techniques of exploratory multivariate analysis (in particular, correspondence analysis, biplots and canonical correspondence analysis) to archaeology and ecology. Data sets (both published and new) are used to illustrate these methods and to highlight the problems that arise — these practical examples are returned to throughout as the various issues are discussed. Much of the motivation for the development of the methodology has been driven by the needs of the archaeologists providing the data, who were consulted extensively during the study.

The first (introductory) chapter includes a detailed description of the data sets examined and the archaeological background to their collection. Chapters Two, Three and Four explain in detail the mathematical theory behind the three techniques. Their uses are illustrated on the various examples of interest, raising data-driven questions which become the focus of the later chapters. The main objectives are to investigate the influence of various design quantities on the inferences made from such multivariate techniques. Quantities such as the sample size (e.g. number of artefacts collected), the number of categories of classification (e.g. of sites, wares, contexts) and the number of variables measured compete for fixed resources in archaeological and ecological applications. Methods of variable selection and the assessment of the stability of the results are further issues of interest and are investigated using bootstrapping and procrustes analysis. Jack-knife methods are used to detect influential sites, wares, contexts, species and artefacts.

Some existing methods of investigating issues such as those raised above are applied and extended to correspondence analysis in Chapters Five and Six. Adaptions of them are proposed for biplots in Chapters Seven and Eight and for canonical correspondence analysis in Chapter Nine. Chapter Ten concludes the thesis.

Acknowledgements

I would like to thank my supervisor, Nick Fieller, for both his support throughout this Ph.D. and for offering me somewhere to stay during the last few months. Thanks also go to Scott for persuading me not to give up after two weeks and to Mark, Matt and Jon for making my third year so much more fun. Thanks to Sofia (Dias) for being a fantastic office mate and to Sofia (Lopes) for being a good friend. Jeremy and Marc deserve a mention for coming from Nottingham to liven up the department as do Andy (Fox) and Alan for just being around. Thanks to Nikhil for starting at the same time as me and for lunch in Loxleys. Thanks to Maralyn for providing continual gossip and to Pauline for being so efficient.

Contents

Chapter Six: Category Selection Methods and Correspondence Analysis 204

Chapter Eight: Stability, Sample Size and Biplots 288

Chapter Nine: Stability, Selection Methods, Sample Size and Canonical Correspondence Analysis 330

List of Figures

List of Tables

Chapter One

Introduction

1.1 Background and Motivation

This thesis has been motivated by both an interest in multivariate techniques of analysis and an interest in archaeology (and other related 'field studies' such as some areas of ecology), although the methods described are by no means exclusively tied to 'field studies' — they are applicable in other areas. The aim, however, is to develop practical guidelines regarding data collection for archaeologists in particular, in order to enable sensible statistical analysis to be carried out post-excavation. Because time and money for excavations are severely limited, a balance has to be struck between numbers of variables recorded on each artefact (or other item), numbers of artefacts collected at each site, numbers of categories into which artefacts are classified and numbers of sites examined. Ways of approaching this multivariate design problem are illustrated in the chapters that follow, using the data sets described in Section 1.2 and listed in the Appendix.

1.1.1 Aims and Objectives

The aim of many techniques of exploratory multivariate analysis is to give an informal assessment of the structure of a data set and to give initial answers to questions such as: are particular observations similar or distinct; are variables correlated or independent; do the data subdivide into groups; which observations are particularly associated with which variables? Available techniques include principal component analysis, correspondence analysis, canonical correspondence analysis and the various forms of biplot. Which one is appropriate in any particular study depends on the form of the data (e.g. continuous measurements or counts) and the archaeological or ecological problems posed. For example, in archaeology, pottery fragments might be collected at various sites and classified according to fabric type, function and decoration; in ecology, various species of beetle might be recorded at various sites. The objective might be to determine which past human activities were associated with which sites, or which species (and hence environmental regimes) characterise which sites.

Sometimes the data analysed are all the data that were potentially available. However, in other cases, in particular ecological, archaeological and other 'field studies', there are more data that could be collected if necessary. In such cases it is desirable to be economical in data collection, yet still be able to obtain conclusive results from statistical analyses. Thus, there is a need to design the data collection stage in terms of numbers of observations made and which variables are recorded, bearing in mind which multivariate technique is to be used for analysis.

If the data are of the form of a sites-by-types matrix then analysis might start with a graphical display obtained by using correspondence analysis. The questions of statistical design raised are, for example: how many pieces of pottery are needed at each site (this depends on the number of sites examined) and how fine a classification should be recorded? Given that there is always a limited budget available (in terms of both time and money), a choice is forced between many samples at few sites or few samples at many sites, as well as between detailed classification on few objects or less detail on more objects. Distinguishing between very similar fabric types might be time consuming and returning to a site to supplement an inadequate sample might be additionally expensive. The requirement is that there should be sufficient data for the inferences drawn from the analysis to be adequately 'reliable'.

If, alternatively, the data are in the form of an observations-by-variables matrix then analysis might start with a graphical display of the matrix, obtained by using a biplot.

Statistical questions raised include: how many variables should be recorded; is it worth distinguishing between strongly correlated ones; do we have enough observations to reveal any (known or unknown) group structure; what are the effects of measuring fewer observations? This time the choice is primarily between measuring many variables on few observations or few variables on many observations.

This thesis is directed towards answering such questions as those outlined above. This requires investigating these various exploratory multivariate techniques, formalising the way 'informal' assessments are made, particularly from graphical displays (largely by bootstrapping) and investigating how sample size (in terms of numbers of sites visited, numbers of classifications made and numbers of artefacts measured) influences these methods. All of the graphical display techniques in this thesis are based on the singular value decomposition of a matrix but note that we do not directly consider principal component analysis (PCA) here because, as we will see in Chapter Three, PCA is encompassed within the biplot framework. Section 1.2 introduces the data sets (both new and published) which are returned to throughout and Section 1.3 explains the notation and methodology for the chapters which follow. Whichever technique is used for analysis, the steps involved in the graphical display of the data remain similar and these are described in 1.4. Section 1.5 explains the historical background of the techniques and 1.6 describes the structure of the thesis, chapter by chapter.

1.2 Data Sets

The three multivariate techniques — correspondence analysis, biplots and canonical correspondence analysis (which are described fully in Chapters Two, Three and Four) — are illustrated on various recently collected data sets as well as on some published data. The data sets used are described in the following sections and listed in the Appendix. They all arise from either archaeological or ecological studies.

1.2.1 Memphis Pottery Sherds

The first extensive data set arises from excavations carried out by the Egypt Exploration Society in Memphis, Egypt (approximately 30 km south of Cairo), over the last 20 years. These data have been provided by Janine Bourriau of the Macdonald Institute, Cambridge and are listed in Table A.1 of the Appendix. The data consist of excavated pottery sherds that form a stratigraphic sequence. The sequence consists of weights (in grams) of pottery sherds classified into 13 contexts (where a context can be thought of as the situation or circumstances in which an artefact is found e.g. soil conditions and is the unit of excavation) and 48 pottery 'wares' (where a ware is considered to be a combination of vessel form, fabric and decoration). The total weight of all sherds is 261 kg. The contexts form a chronological sequence with that nearest to the current ground surface being the most recently 'used' and that deepest below ground the least recent. Archaeological interest lies partly in investigating how pottery typology has altered (where typology means chronological evolution of an artefact), partly in examining how pottery function has altered with stratigraphy and partly in providing a reference collection of pottery to be used on smaller sites (because Memphis is a large site). The complete stratigraphy at Memphis covers a period of perhaps 1500 years but the subject of the data used here is restricted to only a few hundred years. The sherds were weighed rather than counted in order to save time, but it is known that in this assemblage 1 sherd ≅ 10 grams (Bourriau, *pers. comm.*) and so we treat the data as counts of sherds. Statistical interest is concerned with using correspondence analysis to identify the relationships between the various wares and contexts and seeing whether the chronological nature of the contexts is reflected in the analysis. Also, with such a large number of wares, the effects of 'merging' wares on the interpretation of contexts is of interest — i.e. how does the relationship between contexts alter (as revealed by statistical analyses) if we don't distinguish between certain types of wares. In addition to the above we can assess how 'reliable' the contexts are — i.e. if we were able to repeat the data collection procedure then how would this alter the observed relationships between the contexts. Figure 1.1 illustrates the stratigraphic sequence with context 377 being closest to the current ground surface.

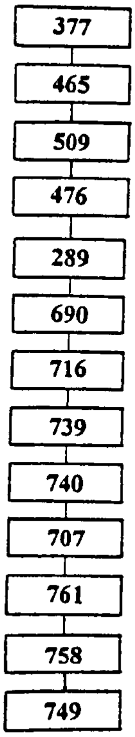

Figure 1.1 Stratigraphic Sequence of Memphis Sherds

1.2.2 Amarna Pottery Sherds

These data were obtained from Paul Nicholson at Cardiff University and are listed in Table A.2 of the Appendix. The data consist of the surface collection of 12693 pottery sherds from (to date) 12 'sites' over the city of Amarna, Egypt. Archaeological interest lies in establishing which areas of the city were used for which type of activities, such as domestic, ceremonial, craft etc. The sherds were collected by selecting a target point in the centre of an area of visibly high sherd density on the ground surface (conventionally taken to indicate the previous occurrence of activity below the surface), scribing a circle of radius 10 feet, collecting all sherds within the scribed circle and classifying them into pottery wares (10 in total). Each circle is taken to be a separate site and there are a total of 12 across the city with the number of sherds at the various sites ranging from 243 to 2589. Statistical interest is mainly concerned with investigating sample size issues using correspondence analysis. If, for example, a circle of smaller radius had been scribed, or if only a fixed number of sherds had been examined, then it is important to understand how this would alter the results and conclusions of any analysis (i.e. would it lead to changes in which pottery wares are associated with which sites and which sites are most similar with regard to pottery wares). These questions are of interest because the study is ongoing and so far the 12 sites examined cover only a small proportion of the total area of the city. Decreasing the sample size per site would allow more sites to be examined in the available time. Conversely, if sample sizes are inadequate at some or all of the sites then there is time to remedy this. Further questions of interest include how the various relationships between wares would be altered if fewer sites had been visited and whether any site is particularly unusual in terms of the pottery it contains.

1.2.3 Melanesian Starch Grains

Carol Lentfer at Southern Cross University, Australia has provided data consisting of the abundances of each of 96 types of starch grain, retrieved from soil samples at each of 15 sites of known environment (e.g. plantation, garden, village, forest), in Gauru, New Britain, together with the sizes of each grain (length and width). There are 3336 grains in total. These data are being used in a new area of archaeological research because, whilst there is a belief that any single plant species gives rise to only one 'type' of starch grain, there is a suspicion that different species could give rise to the same grain 'type'. However, it is suspected that grains of the same type from different species might be differentiated on the basis of size. Thus, interest lies partly in establishing the effect on the site and type relationship when grain types that appear multimodal, from histograms of grain length, are divided up into several categories on the basis of this measurement, but also in detecting information on past vegetation, crops and climate from this fossil plant material. We use correspondence analysis to investigate these issues and we also examine the associations between sites of different environments. These data are listed in Table A.3 of the Appendix and the environmental descriptions of the sites are given in Table A.4.

1.2.4 Early Stone Age Tools

These published data (Bølviken *et al.*, 1982) consist of 899 worked stone artefacts from 28 Early Stone Age sites (8000-4000 BC) published originally by Odner (1966), together with similar data from 15 sites published originally by Simonsen (1961). The artefacts have been grouped by Bølviken *et al.* (1982) into 16 functional types and then into 7 functional classes. The reason for collecting such data was to test Odner's hypothesis that the largest sites in the inner part of the fjords of the Varangerfjord area of Scandinavia reflect larger aggregates of people during longer periods of time than the smaller sites which are located in the outer fjord-coast area. If Odner is right then it is expected that the subsistence dichotomy should include two groups of geographically different sites, with an emphasis on different artefact types. Statistical analysis is concerned with using correspondence analysis to compare the effects of artefact groupings based on archaeological arguments with those obtained from groupings based purely on statistical methods, by looking specifically at the locations of the sites in the ordination map. These data are listed in Tables A.5 and A.6 of the Appendix.

1.2.5 Ceramic Pots

Finds of many sherds of one particular shape and size from the 17th century porcelain kiln-site of Hyakken, near Arita, in Japan, prompted Impey (1979) to speculate on the number of potters working at the kiln. He thought that if the sherds were complete enough for several measurements to be taken on each piece, then these measurements could be analysed to see if there were natural groupings. If there were such groups then the number of groups might correspond to the number of potters working on that shape at the kiln, which could have implications for output, trade distribution, craft specialisation and population size of the site. Furthermore, Impey & Pollard (1985) thought that even within a given pottery shape, the individual characteristics of both the thrower and the turner would be detectable by taking measurements and thus they hypothesised that the thickness of the rim of a vessel would be determined by the thrower, the width of the foot by the turner and the overall height by both. In order to investigate this hypothesis (and to try to shed light on the issues raised above) Impey & Pollard (1985) commissioned an experiment whereby three potters were shown the kiln-site material from Japan and asked to make 10 replicate pots each. Thirteen measurements (in cm) were then taken on each of the 30 pots and these are listed in Table 1.1 below, along with their associated codes. The aim was to investigate whether the pots divide into three groups on the basis of these measurements. The data can be found in Table A.8 of the Appendix.

Table 1.1 Ceramic Pot Measurements (cm)

Measurement Code	Description
1	Internal height at centre
2	External diameter at lip
3	Internal diameter 2cm from base
4	External diameter 2cm from base
5	Internal diameter at lip
6	Overall height
7	Height from point of angle
8	Diameter at point of angle
9	External diameter of footring at base
10	Internal diameter of footring at base
11	Internal depth of footring at centre
12	Thickness of wall at 2cm from base
13	Thickness of lip

The pot measurements are illustrated in Figure 1.2 where the inside of the pot is represented by the left-hand side of the diagram and the exterior, with any decoration, by the right hand side. The dark shaded area on the far left represents the thickness of the pot.

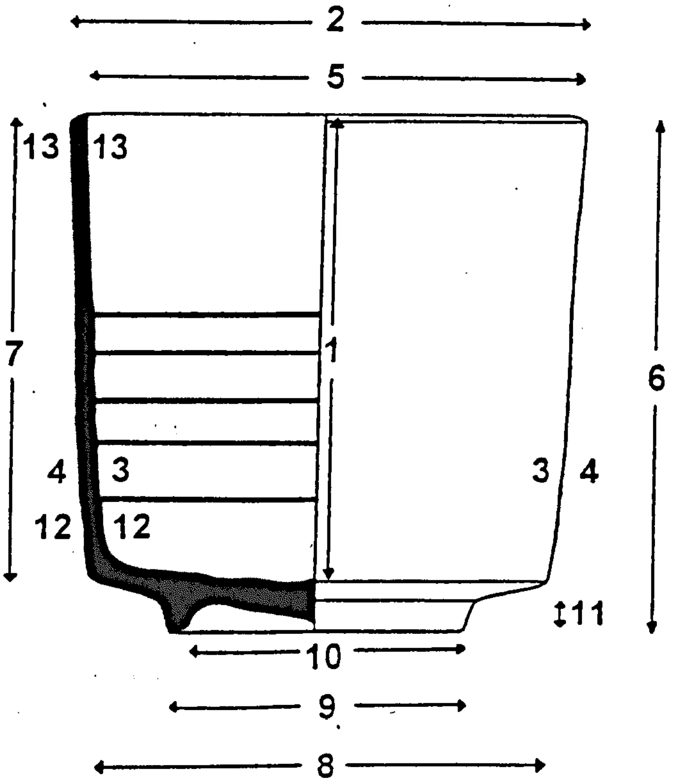

Figure 1.2 Ceramic Pot Measurements

Statistical interest lies in whether biplots enable us to distinguish between the pots made by each of the three potters using the available measurements, how the separation of groups is altered when variable selection methods are implemented and the effects on the analysis when fewer pots are considered.

1.2.6 Simpson Desert Flakes

These data were obtained from Huw Barton at the University of Sydney, Australia and consist of dimensional measurements (made using callipers) on flakes (flint tools and flake debitage) from the Simpson Desert, Australia. Additionally, the weight of each flake (in grams) was recorded using an electronic balance. Flint tools were recorded at two sites, coded 08 and 09. Archaeologically, the landform at site 08 is described as 'escarpment', whereas site 09 is described as 'plain with drainage'. The measurements (in mm) taken on the flakes are described in Table 1.2, along with their associated codes and are illustrated in Figure 1.3.

Table 1.2 Simpson Desert Flake Measurements

Measurement	Code	Description
Length (mm)	**1**	Length from the point of force of application to the most distal point on the flake.
Width (mm)	**2**	A measurement perpendicular to the length axis taken at the midpoint of that axis.
Thickness (mm)	**3**	A measurement taken at the intersection of the length and width axes from the ventral to the dorsal flake surface.
Platform width (mm)	**4**	Along the plane of the striking platform from one lateral flake margin to the next.
Platform thickness (mm)	**5**	Measurement from the point of force of application, perpendicular to the bulb of percussion, from the ventral to the dorsal flake surface.
Weight (grams)	**6**	Weight of flake to the nearest tenth of a gram. The lower limit of sensitivity is 0.5g and the upper limit is 1000g.

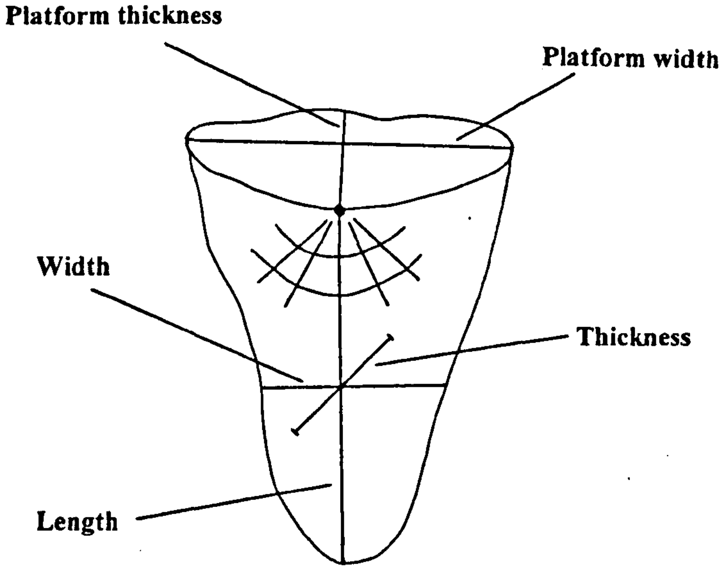

Figure 1.3 Simpson Desert Flake Measurements

These data were collected as follows. A 5 metre (m) grid square was sub-divided into 1m squares, where placement of each 5m x 5m grid was determined largely on the basis of visible flake density: low density patches were avoided in order to increase the amount of data recovered from each recording unit. At each major sample location, a total of five, 5m square grids were laid out and within each metre square every flake bigger than 5mm was recorded (flakes smaller than 5mm were ignored because collection and analysis of such material would have been almost impossible, Barton, *pers. comm.*). At minor sample locations the area examined was smaller and so the quantities of flakes measured at different locations are not comparable, because the total surface area examined differs. The aim is to discover which measurements are important in discriminating between the material from each type of terrain for both flint tools and, separately, flake debitage (Barton, *pers. comm.*). The tool data, after incomplete tools have been deleted, consist of six measurements on 53 tools from site 08 and on 26 tools from site 09. The debitage data consist of six measurements on 2767 flakes from 28 sites. Statistical interest lies in assessing how 'reliable' the measurements are when sampling methods are used to reduce the number of tools or flakes analysed and also when the number of variables measured is reduced. Methods of identifying outlying and influential flakes are also of interest. In addition to the above, the ability of biplots to identify differences between the sites based on their landform and access to water sources is investigated, as well as how this alters when variable selection methods are implemented. The tool measurements and site descriptions are listed in Tables A.9 and A.10 of the Appendix respectively.

1.2.7 Bone Engravings

Data consisting of the abundances of 44 designs, engraved on bones from five sites in Spain, were obtained from Kaufman (1998), but were originally investigated by Conkey (1980). Conkey (1980) used these data to try to distinguish between aggregation and dispersion sites for the Early Magdalenian occupation of Cantabrian Spain, arguing that aggregation sites should exhibit a greater diversity of designs than dispersion sites, because bands of hunter-gatherers would congregate at these sites. We use these data to introduce the diversity biplot into archaeology. Examples of the designs are illustrated in Figure 1.4 and the data are given in Table A.11 of the Appendix.

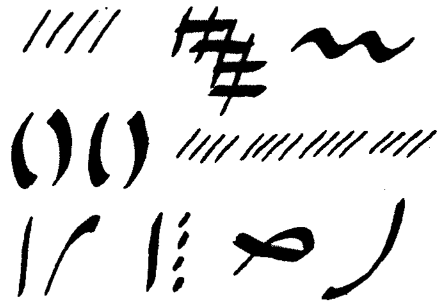

Figure 1.4 Bone Engravings (after Conkey, 1980)

1.2.8 Hunting Spiders

Data on the distributions of 12 species of hunting spider across sites in a Dutch dune area, together with measurements of environmental characteristics at the various sites, have been taken from van der Aart & Smeenk-Enserink (1975) and were previously analysed by ter Braak (1986). The species data consist of the numbers of individuals of each species caught in pitfall traps over a period of 60 weeks, with 26 environmental variables measured at each of the 28 traps. The reason for collecting these data was to trace the main environmental factors that have influenced the distributions of the species studied. In ter Braak (1986) the number of variables was considered too large to sort out their independent effects on community composition and 18 were removed on a priori grounds; two more were removed because they were strongly correlated with one of the remaining six variables. These data and their descriptions are listed in Tables A.12-A.14 of the Appendix and are used both to illustrate canonical correspondence analysis and to examine how 'reliable' the sites are when there are small changes in the data collected — i.e. if we were to repeat the data collection procedure then how would this alter the observed relationships between the species, sites and environmental variables. We are also interested in the effects on the analysis of transforming both the species and environmental data and the data provide scope for examining both existing and new methods of selecting environmental variables. In addition, we can also examine the effects on the analysis of visiting fewer sites i.e. how do the relationships between the species and environmental variables alter.

1.2.9 Dune Meadow Vegetation

Data relating to dune meadow vegetation originate from a research project by Batterink & Wijffels (1983, unpublished) on the Dutch island of Terschelling, but were taken from ter Braak (1988). The objective of the original project was to investigate the differences in vegetation among dune meadows that have been subjected to different management regimes. Thirty species have been recorded across 20 sites according to the ordinal scale of van der Maarel (1979) and, additionally, five environmental variables have been measured at each site. These are (a) thickness of the A1 soil horizon (measured in centimetres), (b) moisture content of the soil (on a 5 point scale), (c) grassland management type — (standard farming (SF), biological farming (BF), hobby-farming (HF) and nature conservation management (NC)), (d) agricultural grassland use — (hayfields (H), pasture (P) or a combination (C) of these) and (e) quantity of manure applied. These data are used to again illustrate canonical correspondence analysis and to look at the 'reliability' of the sites — i.e. how representative are they of the true population of dune meadow vegetation data. They also raise important questions regarding how to deal with ordinal and nominal variables and how the scales on which vegetation abundances are usually measured affect the reliability assessment. In addition, interest lies in using these data to help develop methods for detecting influential sites, species and variables. These data can be found in Tables A.15 and A.16 of the Appendix.

1.3 Notation

As explained in Sections 1.1 and 1.2, the graphical display techniques in which we are interested are correspondence analysis, the various forms of biplot and canonical correspondence analysis. All three methods make use of the singular value decomposition and it is useful, therefore, to define some notation that is used throughout the thesis and to describe the methodology that is common to all three techniques.

1.3.1 The Norm

The norm of a vector is the distance of the vector from the origin. Therefore, the norm of a vector v with r entries is called the Euclidean norm and is denoted by:

$$\|v\| = \sqrt{\sum_{i=1}^{r} v_i^2}\ .$$

The norm of a matrix A (n x m) is defined to be the square root of the sum of its squared entries:

$$\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}\ .$$

1.3.2 The Singular Value Decomposition

The singular value decomposition (SVD) is used in graphical display techniques to find a lower rank matrix that approximates the data matrix (Eckart & Young, 1936). The SVD of any real matrix A (n x m) of rank r, can be expressed as:

$$A = UD_\mu V^T$$

where $D_\mu = \text{diag}\ (\mu_1,\ldots, \mu_r)$ contains the matrix of singular values of A in decreasing order of magnitude;

$D_\mu^2 = \text{diag}\ (\mu_1^2,\ldots, \mu_r^2)$ contains the matrix of eigenvalues of A in decreasing order of magnitude;

U (n x r) and V (m x r) are orthonormal i.e. $U^TU = V^TV = I_r$;

U and V are the eigenvectors of A^TA and AA^T respectively.

If the singular values are all distinct then the singular value decomposition of a matrix is unique up to a simultaneous reflection of the corresponding columns of U and V.

Having expressed A in terms of its SVD we can find a least-squares rank p approximation of A, denoted by $A_{[p]}$ (p < r):

$$A_{[p]} = U_{[p]}D_{\mu[p]}V_{[p]}^T$$

where $D_{\mu[p]}$ = diag ($\mu_1,\ldots, \mu_p$) contains the matrix of singular values of $A_{[p]}$ in decreasing order of magnitude;

$U_{[p]}$ (n x p) and $V_{[p]}$ (m x p) are orthonormal i.e. $U_{[p]}^TU_{[p]} = V_{[p]}^TV_{[p]} = I_p$.

$A_{[p]}$ is the closest of all possible rank p approximations to A in the sense that it minimises the sum of the squared differences between corresponding entries of A and $A_{[p]}$, i.e. it minimises:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij} - a_{ij[p]})^2.$$

1.3.3 The Generalised Singular Value Decomposition

We can generalise the above so that any matrix Q (n × m) of rank r, can be decomposed as:

$$Q = ND_{\mu}M^T$$

where $N^T\Omega N = M^T\Phi M = I_r$;

Ω and Φ are positive definite symmetric matrices.

This is called the generalised singular value decomposition (GSVD) in the metrics Ω and Φ. The rank p approximation to Q in the metrics Ω and Φ is given by:

$$Q_{[p]} = N_{[p]} D_{\mu[p]} M_{[p]}^T.$$

1.4 Graphical Displays in Low-Dimensional Space

From a practical viewpoint we are interested in displaying the results from applying graphical multivariate techniques in a low number of dimensions (typically two), because these are the easiest to interpret. There are three steps to displaying a data matrix in low-dimensional space:

Step 1: The data matrix is scaled. Forms of scaling include column-centring (subtracting the mean of each variable from the appropriate column) and row-centring. Call this scaled matrix H.

Step 2: Compute the GSVD of H:

$$H = ND_\mu M^T$$

and obtain its two-dimensional approximation.

Step 3: Obtain the co-ordinates of the row and column points of H (see below).

The row and column co-ordinates in a p-dimensional display are given by $F_{[p]}$ (n x p) and $G_{[p]}$ (m x p) respectively, which are both of rank p:

$$F_{[p]} \equiv N_{[p]} D^a_{\mu[p]}$$

$$G_{[p]} \equiv M_{[p]} D^b_{\mu[p]}$$

where $N_{[p]}$ and $M_{[p]}$ are the first p columns of N and M respectively;

$D_{\mu[p]}$ is a diagonal matrix consisting of the first p singular values of H;

a and b are real numbers, where a + b = 1.

The individual row and column points are given by the rows f_i^T of $F_{[p]}$ and g_j^T of $G_{[p]}$, respectively.

1.4.1 Axis Scaling

It is known that the scales of the displayed axes must be equal for the various forms of biplot (see Chapter Three) and also for canonical correspondence analysis (see Chapter Four). This is because the interpretations of the variables utilise angles between the vectors representing them and the relative lengths of these vectors. However, in the biplot, if the observation points vary greatly in magnitude from the variable points then one set of points can be multiplied by a suitable constant before displaying the data, without altering the interpretation of the display. This also applies to both the category and variable points in canonical correspondence analysis. Because correspondence analysis (see Chapter Two) can only compare distances between row points and, separately, distances between column points, the scales of the displayed axes do not have to be equal.

1.5 Historical Background

In this section we give a brief guide to the early development of the main three multivariate methods discussed in this thesis. Correspondence analysis is thought to originate with a paper by Hirschfield in 1935, although at around the same time Horst was independently suggesting similar ideas and labelling them the 'method of reciprocal averages'. Fisher was also deriving the same theory in an ecological context and calling it 'dual scaling'. Correspondence analysis, or rather, 'Analyse des Correspondances', was developed by the French linguist and data analyst Benzécri in the late 1960's and was subsequently described and popularised in English by Greenacre (who studied with Benzécri in the early 1970's). Biplots were originally developed by Gabriel (1971, 1972) and have since been summarised by Greenacre & Underhill (1982), Greenacre (1984), Gower (1984) and Gower & Hand (1996).

Jolliffe (1972, 1973) and then Krzanowski (1987) have published methods in the area of variable selection in principal component analysis and Krzanowski (1993) has also published work on attribute selection in correspondence analysis. Ter Braak (1986, 1988) was the first to develop canonical correspondence analysis, although Lebreton was independently exploring similar ideas at about the same time. Canonical correspondence analysis is widely used in ecology and to some extent in archaeology. Its popularity is growing.

1.6 Thesis Structure

This section explains the structure of the thesis chapter by chapter. Chapters Two, Three and Four set the scene for the later chapters and consist mainly of explanations of the theory behind the three multivariate techniques of correspondence analysis, biplots and canonical correspondence analysis respectively, with illustrations of their application to the various data sets described in Section 1.2. These chapters also highlight what we believe to be problems with applying these methods to archaeological and ecological data, but a full consideration of these issues and possible solutions is deferred to the later chapters (Five, Six, Seven, Eight and Nine).

Chapter Two reviews the technique of correspondence analysis and explains how to interpret the results using new data sets from Memphis (1.2.1, pottery sherds), Amarna (1.2.2, pottery sherds) and Melanesia (1.2.3, starch grains). Correspondence analysis is suitable for data in the form of a contingency table and looks for informal patterns between row categories and between column categories. We identify various problems with the application of this method to archaeological data, including the problem of displaying large numbers of categories simultaneously, the effects on the analysis of the number of categories into which artefacts are classified and the influence of overall sample size. These problems and others, are discussed in detail in Chapters Five and Six.

Chapter Three explains the theory behind the various forms of biplot, collating the information on the various types from the fragmentary literature and describing their application to ceramic pots (1.2.5, published data), bone engravings (1.2.7, published data) and to new data on flint tools and flake debitage from the Simpson Desert (1.2.6). Biplots are suitable for data consisting of variables measured on a number of observations and display both the observations and the variables simultaneously. Interpretation rests on examining the correlations between variables and identifying group structure among the observations. The relative merits of the various forms of biplot are discussed, including which is most appropriate for the particular question in hand and we introduce the diversity biplot into archaeology. Because of the large numbers of variables which are often measured by archaeologists, it is important to examine how the observations and remaining variables are affected if fewer variables are measured. On a similar theme, interest also lies in how the relationship between variables and group structure of observations is altered if fewer observations are measured. These questions and others, are addressed in Chapters Seven and Eight.

In Chapter Four we introduce the less well known technique of canonical correspondence analysis and apply it to published data sets consisting of hunting spiders (1.2.8) and dune meadow vegetation (1.2.9). The structure of a data set suitable for canonical correspondence analysis is abundances of a multitude of species across a number of sites, together with a set of environmental variables measured at each site. This method is concerned with identifying which environmental variables are most important in explaining the distributions of the species across the sites. We investigate the effects of various transformations of the data (raw abundances, log and square root transformations, conversion to presence/absence) on the results of the analysis and explain how the method could be used much more widely in archaeology. We also consider how the number of environmental variables can influence the analysis, before implementing variable selection methods in Chapter Nine.

The methodology of Chapter Two forms the basis of Chapters Five and Six. In Chapter Five we introduce methods of investigating the effects of varying sample sizes on the results of the correspondence analysis. These methods aid us in developing general guidelines to help archaeologists when sampling and classifying artefacts, in order to ensure that enough data are collected for statistical methods to be used effectively. In addition, we examine how reliable our particular data sample is by looking at the stability of the two-dimensional maps; we do this by using bootstrapping to resample from the multinomial and the hypergeometric distributions and obtain confidence regions using convex hulls and concentration ellipses. We also assess stability by applying a jack-knife approach and emphasise that any resampling must be implemented in a way appropriate to the method by which the data were originally collected.

Chapter Six investigates the effect of the number of categories into which artefacts are classified on the results of the correspondence analysis. We discuss existing statistical methods of selecting categories and suggest improvements, before implementing a new method. We also consider the implications of these statistical methods on archaeology and use archaeological expertise to suggest alternative category groupings, before discussing why the two approaches may not agree. A method of detecting influential categories is also introduced; this is based on a jack-knife approach.

Chapters Seven and Eight extend the methodology of Chapter Three. Because of the large numbers of variables that are often present in archaeological data, Chapter Seven adapts the existing methods of variable selection used in principal component analysis to the various forms of biplot and comments on their validity. This chapter also develops and implements other methods of variable selection and discusses their relative merits (by analogy with linear regression).

Chapter Eight discusses replicating the data matrix by using the multivariate normal distribution, in order to investigate the stability of the biplot variables and also to examine the effects of varying sample size on the biplot interpretation. Confidence intervals for the true directions of the variables (i.e. for the whole population of data) are also developed, using both traditional bootstrap and directional data methods. In addition, jack-knifing as a means of both assessing stability and identifying influential observations is introduced.

Canonical correspondence analysis (CCA) is investigated in Chapter Nine, which expands on the methodology of Chapter Four. By resampling from the multinomial distribution and considering confidence regions based on convex hulls and concentration ellipses, we can assess the reliability of our particular data sample; an alternative method of investigating stability is to use a jack-knife approach. We also consider the effects of the number of sites visited on the CCA map and compare an existing method of variable selection with a method that we introduce. We propose using jack-knifing to identify influential species, sites and environmental variables (i.e. those which have a large influence on the ordination diagram) and we assess the impact of changes in species abundance (i.e. sample size) on the interpretation of the map.

This thesis is concluded in Chapter Ten.

Chapter Two

Correspondence Analysis

2.1 Introduction

This chapter presents a review of the technique of correspondence analysis. The purpose is to bring together the algebraic details of the method and the various steps involved in interpreting the results and to illustrate these on several new data sets which were introduced in Chapter One and which are listed in the Appendix. In addition, questions generated by the particular problems underlying the data sets are raised, such as the influence of overall sample size, the influence of the number of categories and the effects of amalgamating and dividing categories on the analysis. These and other issues, are addressed in Chapters Five and Six.

Correspondence analysis is a graphical exploratory multivariate technique that displays the rows and columns of a matrix of non-negative data as points in low-dimensional vector spaces. These spaces can be superimposed to obtain a joint display of rows and columns. The most basic form of correspondence analysis, known as simple correspondence analysis, is its application to a two-way contingency table. All other forms of correspondence analysis are the application of the same algorithm to other types of data matrices.

Section 2.2 describes the algebraic details of the technique, whereas 2.3 explains the interpretation of the results with illustrations on pottery sherds from Memphis (1.2.1) and Amarna (1.2.2) and on Melanesian starch grains (1.2.3). Questions arising as a result of applying correspondence analysis to these data sets are also raised in this section. Some faults of correspondence analysis are illustrated in Section 2.4, using data on frequency seriation and the role of seriation in archaeology is also discussed. A brief comparison of correspondence analysis with principal component analysis is given in 2.5 and Section 2.6 concludes the chapter, identifying several areas of concern which may arise when applying the method, particularly to archaeological data. These concerns are addressed in Chapters Five and Six.

2.2 The Theory of Correspondence Analysis

Correspondence analysis aims to obtain a graphical representation of both the rows and columns of a matrix in as few dimensions as are deemed 'adequate'. The method can be defined in three steps:

Step 1: Define two clouds of points (one for rows and one for columns) in corresponding multidimensional space.

Step 2: Impose a metric structure on each cloud of points i.e. define distances between rows and between columns.

Step 3: Define the fit of each cloud of points to a low-dimensional subspace onto which the points are projected for subsequent display and interpretation. Typically, we use two-dimensional space.

These steps are described algebraically in Section 2.2.1.

2.2.1 Algebraic Definition

A variety of approaches lead to the equations of correspondence analysis (Tenenhaus & Young, 1985), but here we use the idea of the singular value decomposition (SVD) of a matrix (Eckart & Young, 1936) to provide the theoretical background.

Step 1

We introduce the following definitions.

Let X (n × m) = data matrix of rank r, with elements x_{ij};

P = data matrix X divided by the sum of all its elements, with elements p_{ij} (profiles);

r = vector of row sums of P (row masses or average column profile);

c = vector of column sums of P (column masses or average row profile);

$D_r = \text{diag}(r)$;

$D_c = \text{diag}(c)$;

$R = D_r^{-1}P$;

$C = D_c^{-1}P^T$.

We want to represent graphically the distance between row (or column) profiles and so we orientate the configuration of points at the 'centroid' of both sets. The centroid of the set of row points in its space is c, the vector of column masses or 'average row profile'. The centroid of the set of column points in its space is r, the vector of row masses or 'average column profile'. To perform the analysis relative to the centre of gravity, P is centred 'symmetrically' by rows and columns i.e. we compute P - rc^T (Hoffman & Franke, 1986).

Step 2

In step 1 we defined the set of row points and their masses in r-dimensional space and calculated their centroid. This space needs to be structured so that we can compute distances between profiles. However, the usual Euclidean distance function is not suitable and so a weighted Euclidean metric is used, called the chi-squared metric, where each squared difference between row profiles z_i and $z_{i'}$ is divided by the respective element of the average row profile:

$$d_c^2(z_i, z_{i'}) = (z_i - z_{i'})^T D_c^{-1}(z_i - z_{i'}).$$

The usual chi-squared statistic, χ^2, that tests the null hypothesis of row-column independence can be expressed as:

$$\chi^2 = x_{..}\sum_{i=1}^{n} r_i (z_i - c)^T D_c^{-1}(z_i - c)$$

where $x_{..}$ is the sum of all the elements of X. In other words, $\frac{\chi^2}{x_{..}}$ can be defined geometrically as the weighted average of the squared distances of the row profiles to their centroid. This is termed the total inertia of the data matrix.

Step 3

In steps 1 and 2 we have defined the cloud of row-profile points with masses in a space structured by the chi-squared metric. The problem is finding the k-dimensional subspace through the centroid of the cloud that is closest to all the points. The measure of closeness is defined as the weighted sum of squared distances from the points to the subspace, where the weights are the row masses and the distances are computed using the chi-squared metric.

It can be shown that the first k right and the first k left singular vectors respectively of $D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$, corresponding to the k largest singular values, represent the k-dimensional subspace of the row and column clouds which are closest to the points in terms of the weighted sum of squared distances (see e.g. Greenacre & Hastie, 1987). Let the SVD of $D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$ be:

$$D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}} = ND_\mu M^T \tag{2.1}$$

where N and M are orthonormal i.e. $N^TN = M^TM = I_r$ and $D_\mu = \text{diag}\ (\mu_1, \ldots, \mu_r)$ is a diagonal matrix of singular values. The columns of N and M define the principal axes of the row and column clouds respectively.

The trace of the matrix in (2.1) is equal to $\frac{\chi^2}{x_{..}}$ and so its eigenvalues, or principal inertias, are a decomposition of the total inertia and give an idea of the quality of the representation with respect to the individual principal axes. The co-ordinates of the row profiles with respect to their principal axes (i.e. the row principal co-ordinates) are given by:

$$F = D_r^{-\frac{1}{2}}ND_\mu. \tag{2.2}$$

2.2.2 The Dual Problem

The above methodology can be applied in an equivalent fashion to the columns of the data matrix P — i.e. by repeating the above steps on the transposed matrix P^T. We now look for the principal axes of the column profiles, weighted by masses that are the elements of c, in a space with a chi-squared metric defined by the diagonal matrix D_r^{-1}. Thus, the elements of r and c play dual roles, weighting the profiles on the one hand and rescaling the dimensions on the other. If we divide each row of this transposed matrix by its total, we obtain a matrix C of column profiles. There is no need to recompute the dual solution because it can be obtained from the first problem via the transition formulae (see 2.2.3 and Greenacre, 1984). The total inertia and its decomposition into principal inertias (i.e. along principal axes) is exactly the same in the two problems. Because of the transition formulae, in their respective subspaces a row point is attracted to the region of the column points for which the row profile is large and vice versa. These reasons justify the merging of the respective plots of the row and column profiles into one and the representation of the row and column points on the same principal axes. If the first two principal axes are plotted then the inter-row and inter-column distances may be interpreted as approximate chi-squared distances, but row to column distances are meaningless.

The co-ordinates of the column profiles with respect to their principal axes (i.e. the column principal co-ordinates) are given by:

$$G = D_c^{-\frac{1}{2}} M D_\mu . \tag{2.3}$$

2.2.3 The Transition Formulae

The transition formulae for these principal co-ordinates express the row co-ordinates in terms of the column co-ordinates and vice versa. They are used to describe the relationship between the row and column points in the display and to plot the supplementary points (see 2.2.4). The two sets of co-ordinates, F and G, are related to each other by the following equations, known as the transition formulae:

$$F = RGD_{\mu}^{-1}$$

$$G = CFD_{\mu}^{-1}.$$

The usual display of row and column points as defined by the transition formulae is in a symmetric map. In this case both sets of points are called principal co-ordinates. For a symmetric map in the first two dimensions, co-ordinates in the first row of F are plotted against those in the second row (these are the first and second principal axes) and similarly for the first and second rows of G. Occasionally, the third principal axis will explain a 'substantial' amount of variation in the data and so the first and third axes, or second and third axes, will be plotted.

2.2.4 Supplementary Points

Sometimes there are additional rows and columns of data which are not the primary data of interest, but which are useful in interpreting features discovered in the primary data. Any additional row (or column) of a data matrix can be superimposed onto an existing map, as long as the profile of this row (or column) is meaningful. Such a row (or column) is known as a supplementary point and takes no part in the determination of the axes. However, the contribution of the axes to the supplementary point is meaningful and allows us to judge whether the point lies to a greater or lesser extent in the space of the map, rather than out of it (Greenacre, 1993b).

2.2.5 The Principle of Distributional Equivalence

If two row points occupy identical positions in multidimensional space, then they may be merged into one point, whose mass is the sum of the two masses, without affecting the masses and interpoint distances of the column points. Similarly, a row of data may be subdivided into two rows of data, each of which is proportional to the original row, leaving the geometry of the column points invariant. This principle also applies to column points.

2.2.6 The Standard Co-ordinates

The standardisation of the principal co-ordinates is the ‘natural’ standardisation imposed by our definition of the two dual and symmetric geometries. Another standardisation differs from the co-ordinates in equations (2.2) and (2.3) by the absence of the scaling factors D_μ; these are the standard co-ordinates. The rows are denoted in (2.4) and the columns in (2.5):

$$\Phi = D_r^{-\frac{1}{2}}N \tag{2.4}$$

$$\Gamma = D_c^{-\frac{1}{2}}M. \tag{2.5}$$

Thus, the weighted average of the squared principal co-ordinates of the rows, or the columns, on a principal axis, is equal to the squared singular value (or ‘principal inertia’) associated with that axis, whereas the weighted average of the squared standard co-ordinates is equal to 1 (Greenacre, 1993a).

2.3 The Ordination Diagram and its Interpretation

This section describes how to interpret the correspondence analysis display, with applications to new sets of archaeological data.

2.3.1 Symmetric and Asymmetric Displays

In a symmetric display (or map), the separate configurations of row profiles and column profiles are overlaid in a joint display, even though they emanate from different spaces and both row and column points are displayed in principal co-ordinates (Greenacre, 1993b). The convenience of such a display is that we always have both clouds of points equally spread out across the plotting area.

An asymmetric display means that the standardisations imposed on the two sets of points are different. Usually, one of the sets is represented in principal co-ordinates and the other is represented in standard co-ordinates, known as vertices (Greenacre, 1984). We refer to an asymmetric row plot when the rows are in principal co-ordinates and the columns are in standard co-ordinates (and vice versa for an asymmetric column plot). In asymmetric maps the principal co-ordinates are often bunched up in the middle of the display, far from the outer standard co-ordinates, especially if the principal inertias are low.

When row points are in principal co-ordinates, the row-to-row distances approximate the inter-row chi-squared distances. This is the case in both the symmetric map and in the asymmetric row map. The danger of symmetric maps is in interpreting row-to-column distances directly because no such distance is defined or intended in this map. However, the degree of association between a row point and a column point is determined by a comparison of their distances from the origin. Whether the joint map is produced using asymmetric or symmetric scaling, there is a style of interpretation that remains universally valid — the dimensional interpretation. This involves interpreting one axis at a time and using the relative positions of one set of points to give a descriptive name to the axis.

The row and column points are displayed in the subspace of the first few principal axes and the quality is gauged by the moments of inertia (or eigenvalues) expressed as percentages of the total inertia. The quality of representation in the subspace of k dimensions, rather than in the full r-dimensional space, is given by the ratio of the sum of the eigenvalues:

$$t_l = 100 \times \frac{\sum_{t=1}^{k} \mu_t^2}{\sum_{t=1}^{r} \mu_t^2} .$$

It is hoped that the first two dimensions explain 'most' of the variability in the data.

2.3.2 Inertia

The overall spatial variation in the set of row points and set of column points assists in the interpretation of the maps. This variation, the total inertia, is defined as the weighted sum of squared distances from the points to their respective centroids and is equivalent for both sets of points. It is given by:

$$\frac{\chi^2}{X_{..}} = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j} .$$

The total inertia can be decomposed along the principal axes. Each eigenvalue, μ_t^2, indicates the weighted variance (inertia) explained by the t-th principal axis of the display; summed over all principal axes, these eigenvalues represent the total inertia of the spatial representation. Additionally, the inertia of an axis can be decomposed among the different points so that each point's contribution to the position of that axis can be found. The total inertia of a point can also be decomposed along the different axes in order to show how well each point is represented by each axis. However, a two-dimensional display does not indicate which points have had the most impact in determining the orientation of the axes and so we need additional information.

2.3.2.1 Inertia of the Points

The inertia of the i-th row point is equal to:

$$v_i = r_i \left[\sum_{j=1}^{m} \frac{\left(\frac{p_{ij}}{r_i} - c_j\right)^2}{c_j} \right] = r_i \sum_{t=1}^{r} f_{it}^2 .$$

This is the contribution of the i-th row point to the total inertia. A similar definition holds for each column point.

2.3.2.2 Absolute Contributions to Inertia

The inertia along the t-th axis, μ_t^2, consists of the weighted sum of squared distances of the displayed row (or column) profiles to the origin, where the weights are the masses for each row (or column) point. For the row profiles, this inertia can be expressed as:

$$\mu_t^2 = \sum_{i=1}^{n} r_i f_{it}^2 .$$

Thus, each eigenvalue also represents the inertia of the projections of the set of row (or column) points onto each axis. Each term in the summation is expressed relative to the inertia 'explained' by each axis (i.e. as a percentage) and so the absolute contribution of the i-th row to the t-th principal axis is obtained. The absolute contributions quantify the importance of each point in determining the direction of the principal axes.

2.3.2.3 Relative Contributions to Inertia

The 'Quality' of the representation of each point in the display can also be determined. The relative contribution of the t-th principal axis to the inertia of the i-th row is given by the quantity:

$$q_{it} = \frac{f_{it}^2}{\sum_{t=1}^{r} f_{it}^2}$$

which indicates how well each point is ‘fit’ by the representation. For a particular point, the sum of the relative contributions across all r axes is equal to 1, because an angle cosine can be given a geometric interpretation as a correlation coefficient. We can find out how close each point lies to the k-dimensional subspace by looking at the angle θ between the true profile point vector and the principal axis. The inertia of the profile point vector is decomposed along the principal axes and the value of $\cos^2\theta$ is called the contribution of the axis to the inertia of the point. If $\cos^2\theta$ is high then the axis explains the point’s inertia very well; equivalently, θ is low and the profile vector is said to lie in the direction of the axis or ‘correlate’ with the axis. The values of $\cos^2\theta$ are called the relative contributions because they are independent of the mass of the point. Generally, a high contribution of the point to the inertia of the axis implies a high relative contribution of the axis to the inertia of the point, but not conversely (Greenacre, 1984).

2.3.3 Application to Memphis Pottery Sherds

In this section we illustrate correspondence analysis by using data consisting of weights of pottery sherds obtained from excavations at Memphis, Egypt (1.2.1). The weights of the sherds were recorded according to their contexts within the stratigraphic sequence and their fabric type (ware). The contexts are listed in reverse stratigraphical order (see Figure 1.1), from that closest to the current ground surface (context 377) to that furthest below ground (context 749) and the data are given in Table A.1 of the Appendix. Some of the results from applying symmetric correspondence analysis to these data are shown in Tables 2.1 and 2.2. Because there are so many wares, the points representing the contexts and wares are displayed in two separate diagrams (Figures 2.1 and 2.2 respectively), but it is still difficult to identify the precise relationships between the wares — the effect of large numbers of categories on the analysis is discussed in detail in Chapter Six. Since there are 48 wares (rows) and 13 contexts (columns) in the data matrix, there are 12 dimensions to the solution. Thus, 12 principal inertias are obtained which are shown in Table 2.1

below.

Table 2.1 Principal Inertias for the Memphis Pottery Sherds

Inertia	Percentage of Inertia	Cumulative Percentage of Inertia
0.743	44.22	44.22
0.254	15.09	59.31
0.224	13.30	72.61
0.131	7.81	80.42
0.099	5.88	86.29
0.077	4.61	90.90
0.055	3.29	94.19
0.042	2.50	96.69
0.029	1.74	98.43
0.019	1.16	99.58
0.004	0.23	99.81
0.003	0.19	100.00
1.681		

We can see from this table that the first principal axis explains 44.2% of the inertia of the data and that the second axis explains 15.1%. These percentages are high enough for us to be confident in our interpretations of the two-dimensional display (there is no rule for deciding whether a percentage is 'high enough', but a figure of at least 50% has proven to be a reasonable rule of thumb for archaeological data). Considering Figure 2.1 and using the dimensional interpretation of Section 2.3.1, it is clear that the first axis is a contrast between context 377 on the far left, contexts {465, 476, 509} in the middle and the remaining contexts on the right. Thus, going from left to right across the display corresponds to going from closest to the current ground surface to furthest below ground in the stratigraphic sequence (see Figure 1.1). The second axis separates context 289 from the remaining contexts, which is particularly interesting because context 289 is common to another stratigraphic sequence at Memphis (Bourriau, *pers. comm.*).

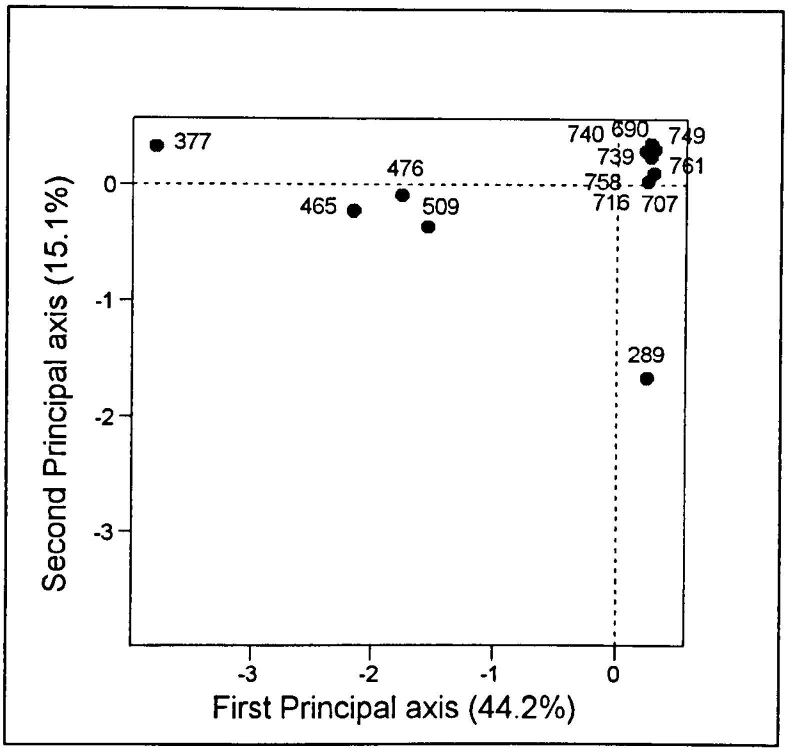

Figure 2.1 Correspondence Analysis Map of Memphis Contexts

We now consider Figure 2.2, the codes for which are listed in Table A.1 of the Appendix. The first axis separates wares {5, 8, 9, 31, 33} — which we refer to as 'group 1' — from {4, 6, 7, 10, 22, 23, 24, 36, 37, 38, 39} — 'group 2' — and both these from the wares in the top right-hand corner which we call 'group 3'. The second axis splits wares 25 and 29 (which are undatable Nile fabrics) from all the remaining wares.

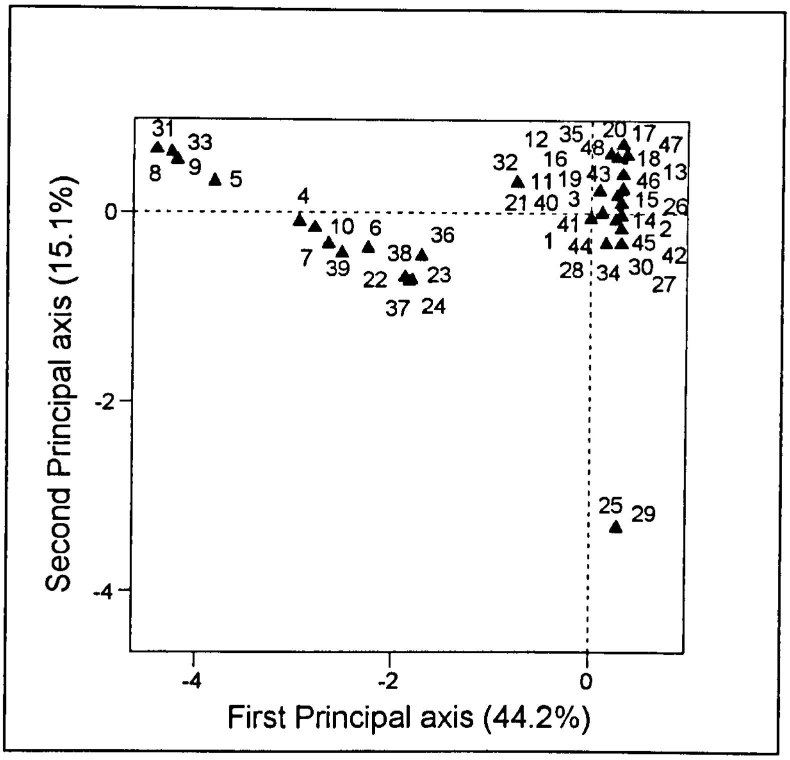

Figure 2.2 Correspondence Analysis Map of Memphis Wares

Overlaying Figures 2.1 and 2.2 suggests that the wares in 'group 1' have a strong association with context 377 (because they are located at similar distances from the origin), the wares in 'group 2' have a strong association with contexts {465, 476, 509} and the wares in 'group 3' have a strong association with the remaining contexts. Wares 25 and 29 are highly associated with context 289. These inferences are confirmed by a close look at Table A.1, where the weights of the wares are relatively large in their associated contexts just described.

The Memphis sherds comprise a large amount of data that have been examined in detail by archaeologists. Many of the sherds could be sorted much more quickly into broader categories of wares which would save time (which could be spent collecting other archaeological information) and which would also mean that there are fewer categories to display on the correspondence analysis map — relationships between wares could then be more clearly identified. It is, therefore, important to investigate whether sorting the sherds into broader categories would alter the interpretation of the correspondence analysis map: this can be assessed by amalgamating categories based on either archaeological expertise or statistical methods. Combining categories in this way reduces the number of categories without losing all the information collected. This issue is addressed in Chapter Six. Table 2.2 displays the numerical output from the correspondence analysis for the contexts only.

Table 2.2 Correspondence Analysis Output for the Memphis Contexts

				First Principal axis			Second Principal axis		
Context	**Quality**	**Mass**	**Inertia**	**Co-ord**	**Correl**	**Contr**	**Co-ord**	**Correl**	**Contr**
377	0.858	0.031	0.312	-3.785	0.851	0.600	0.334	0.007	0.014
465	0.579	0.024	0.116	-2.166	0.573	0.150	-0.213	0.006	0.004
509	0.430	0.038	0.132	-1.552	0.408	0.122	-0.356	0.021	0.019
476	0.442	0.009	0.038	-1.770	0.441	0.038	-0.081	0.001	0.000
289	0.939	0.075	0.136	0.250	0.021	0.006	-1.670	0.918	0.827
690	0.174	0.077	0.038	0.234	0.066	0.006	0.299	0.108	0.027
716	0.255	0.063	0.015	0.295	0.211	0.007	0.136	0.044	0.005
739	0.020	0.011	0.032	0.298	0.018	0.001	0.113	0.003	0.001
740	0.409	0.095	0.030	0.284	0.154	0.010	0.366	0.256	0.050
707	0.256	0.316	0.049	0.256	0.249	0.028	0.041	0.006	0.002
761	0.235	0.011	0.004	0.275	0.130	0.001	0.248	0.105	0.003
758	0.354	0.149	0.025	0.286	0.294	0.016	0.129	0.060	0.010
749	0.161	0.101	0.073	0.314	0.081	0.013	0.313	0.080	0.039

We see from the above table that, for example, the mass of context 377 is 0.031, its inertia in full 12-dimensional space is 0.312, its principal co-ordinate ('Co-ord') on the first axis is -3.785 and its principal co-ordinate on the second axis is 0.334. We obtain similar information for the wares. The following section describes the interpretation of the other entries in the table in more detail.

2.3.3.1 Absolute Contributions to Inertia

For each principal axis we look down the column headed 'Contr', in order to interpret the dimension. The inertia along the first axis is 0.74 (see Table 2.1), which is equal to the weighted sum of squared distances of the displayed column or row profiles to the origin. Each term in this sum can be expressed as a percentage of this first principal inertia and we call these 'contributions by the points to the principal axis'

For example, the point 377 has a mass of 0.031 and a distance from the centroid of -3.785. Its absolute contribution to the first principal inertia is thus 0.031 x $(-3.785)^2$ = 0.44 which is 60.0% of 0.74. It is the points such as this one, with high contributions, that have played the major role in determining the final orientation of the first principal axis. An alternative method of calculating the absolute contribution of context 377 to the inertia of the first axis is to multiply its contribution ('Contr') of 0.60 by the inertia of this first axis, 0.74, giving $0.60 \times 0.74 = 0.44$. We can also carry out similar calculations for the second axis, for the other contexts and for the wares. In Chapter Six we investigate an alternative method of identifying influential points.

2.3.3.2 Relative Contributions to Inertia

For each point we scan across the values in the 'Correl' columns in order to identify the axes which represent the point well. Considering context 377, its squared correlation with the first axis is 0.851 and with the second is 0.007. This is not surprising given that context 377 is well separated from most of the other contexts on the first axis. Looking at context 289, this has a very high squared correlation with the second axis (0.918), which confirms our interpretation of it being distinguished from the remaining contexts on this axis. The quality ('Quality') of representation of context 377 in the two-dimensional display is 0.858, the squared correlation (cosine) with the plane, which is the sum of the individual squared correlations. A similar interpretation can be made for the other contexts and for the wares.

2.3.4 Application to Amarna Pottery Sherds

A second application of correspondence analysis concerns pottery sherds from Amarna, Egypt, which were described in Section 1.2.2 of Chapter One. The sherds are classified according to pottery ware and site and the data are given in Table A.2 of the Appendix. The first two dimensions explain 56.3% of the inertia of the data and this is a high enough percentage for us to be confident in our interpretations of the two-dimensional display. Figure 2.3 displays the symmetric correspondence analysis map of the Amarna sherds. Interpreting the first principal axis, there is a contrast between site c and pottery ware 10 on the right, which appear to be associated with each other, site k and ware 9 in the middle and the remaining sites and pottery wares on the left.

As for the Memphis sherds, absolute and relative contributions to inertia can also be obtained.

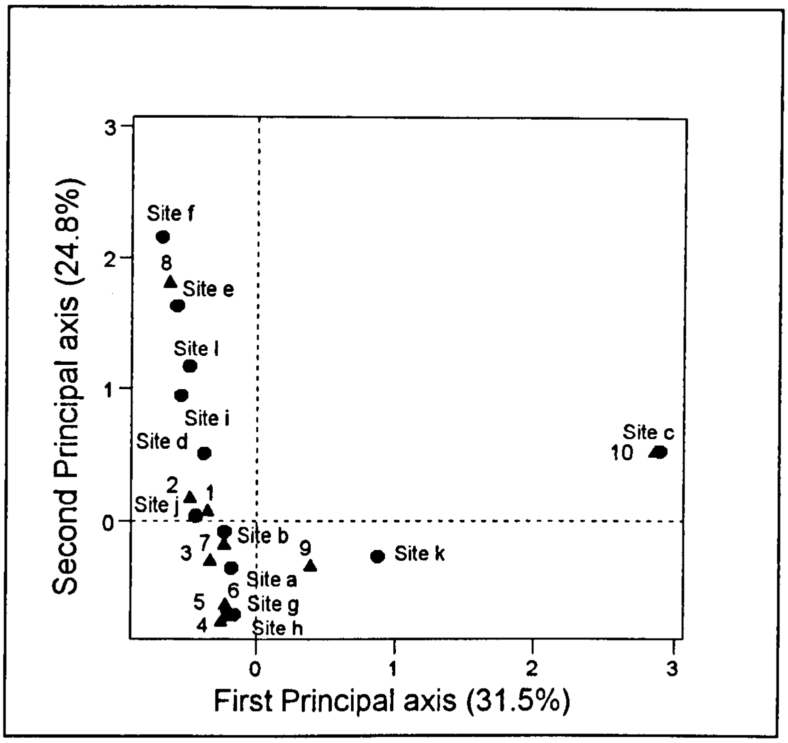

Figure 2.3 Correspondence Analysis Map of Amarna Sherds

Key questions which arise with these data concern sample sizes. The existing data consist of sherds collected from 12 sites, but it is important to understand how the relationships between the pottery wares would be affected if fewer sites had been visited. In addition, we need to consider how the relationship between wares and sites might be altered if fewer sherds had been collected at one or more sites, with a view to making inferences about the minimum sherd numbers required to make analysis worthwhile. Questions such as these are considered in detail in Chapters Five and Six.

2.3.5 Application to Melanesian Starch Grains

In a third example, correspondence analysis is applied to counts of starch grains from Melanesia, cross-classified into type and site, described in Section 1.2.3 and listed in the Appendix. Only 40.0% of the total inertia is explained in the first two dimensions, which is fairly low and so it may be of interest to consider the third dimension (which accounts for 11.4% of the total inertia) in addition to the first two. Figure 2.4 shows the sites displayed in the first two dimensions. Along the first axis site S16 (a garden site) appears to be quite unusual in comparison with the remaining sites, whereas sites S1, S2 and S5 (the rock island sites) are slightly removed from the bulk of the sites in the second dimension. Figure 2.5 displays the types of starch grain but there are so many types that it is almost impossible to obtain a clear picture of the relationships between them. Despite this, it seems reasonable to suggest that types 13, 19, 20, 45, 62 and 65 are slightly separated both from each other and also from the remaining types on the first axis.

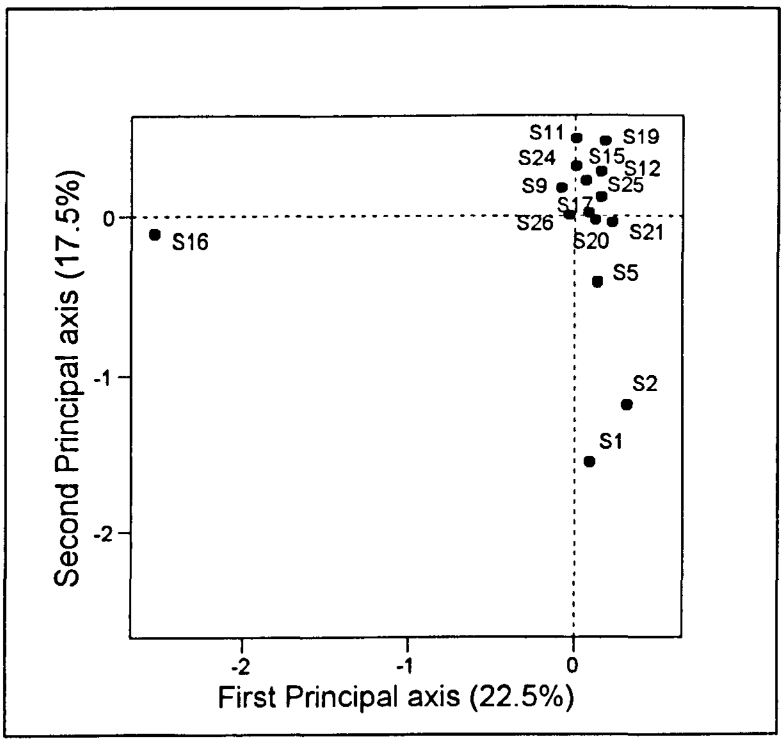

Figure 2.4 Correspondence Analysis Map of Melanesian Sites (First and Second Principal Axes)

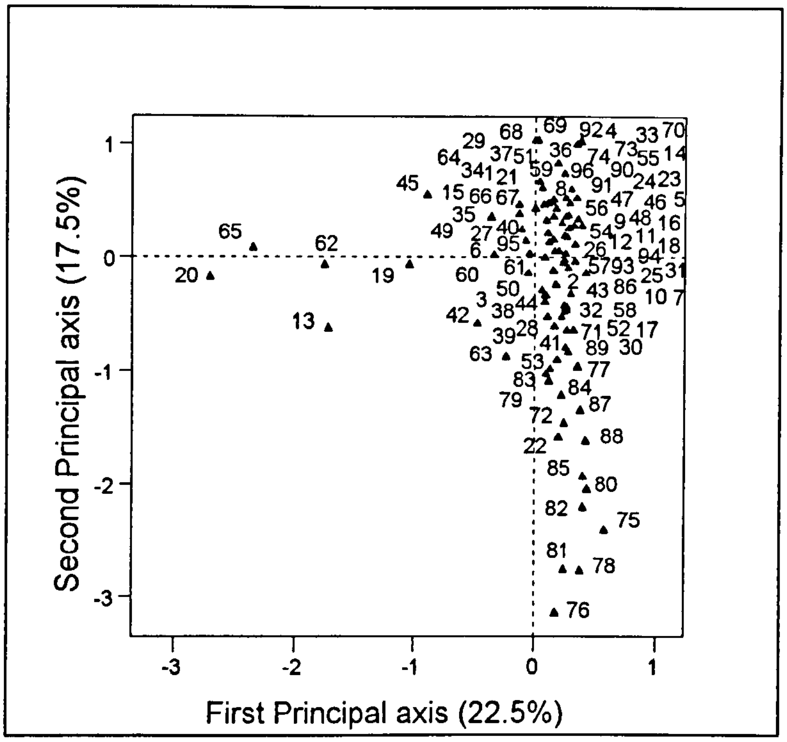

Figure 2.5 Correspondence Analysis Map of Melanesian Starch Grain Types

The sites are displayed in the second and third dimensions in Figure 2.6. We see from this figure that the second axis contrasts the three rock island sites (S1, S2 and S5) from the other sites and there appears to be a disturbance gradient on this second axis going from the least disturbed rock island sites on the left to the most disturbed sites on the right.

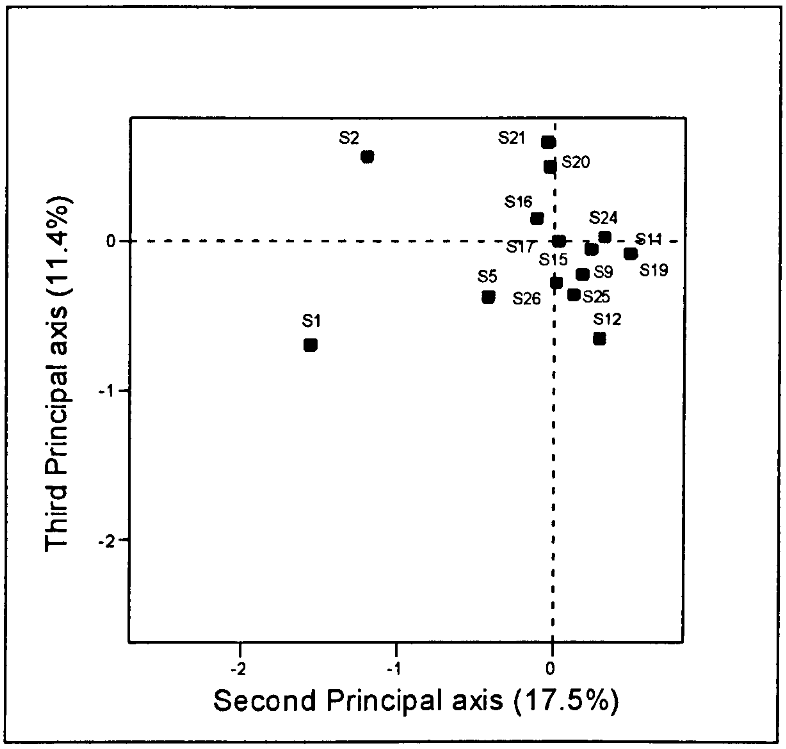

Figure 2.6 Correspondence Analysis Map of Melanesian Sites (Second and Third Principal Axes)

Because of the large number of types which need to be displayed and the relatively sparse nature of most of the types (many with fewer than 10 grains in total), it is important to consider how the relationships between the sites are altered when fewer types, each consisting of a 'reasonable' number of grains, are used in the analysis. Reasons why we may want to delete categories of starch grain can be summarised as follows:

- The data collected for some types of grain are so sparse that they may be hiding relationships between other categories and with other categories. This can cause points to be all bunched up together in the correspondence analysis map. Deleting one or more of these sparse categories can lead to 'true patterns' emerging (or at least more interpretable ones).

- It is almost impossible to identify patterns in the data because we cannot easily visualise all the 96 types of grain that have been collected. Deleting some types may help us to identify patterns.

In addition to the above points and as explained in Section 1.2.3, it is of great importance to archaeologists to investigate the possibility that different plant species produce different sized grains of the same type. It is therefore necessary to allow for the division of types on the basis of grain size and this is considered in Chapter Six.

2.4 Frequency Seriation

The concept of seriation can be attributed to Flinders Petrie (1899), although the essential theory behind the seriation method was 'formalised' by Brainerd and Robinson (Robinson, 1951). Kendall (1971) reviewed and evaluated Petrie's work, explaining the similarity between the approaches of Petrie and Robinson, despite their apparent differences. It is from here that the term 'Petrie matrix' (see below) appears to originate.

The application of correspondence analysis to seriation problems (usually in archaeology) has revealed some unwelcome features of the technique which we explain below. Frequency seriation problems concern determining a plausible ordering of sites or assemblages of artefacts on the basis, typically, of only the presence or absence of each of various categories of artefact. This is taken to indicate their ordering in time. It is, perhaps, the fact that the data matrix is binary — an 'extreme' form of the data — that reveals the deficiencies in correspondence analysis clearly.

Table 2.3 illustrates an incidence matrix, which we have invented, but which is based on an idea from Lock & Wilcock (1987), where the rows represent six site assemblages (collections of artefacts) and the columns five typical groups of artefacts from different archaeological Periods. The occurrence of a '1' indicates that the particular type of artefact is found at the given site. This matrix is unordered and the 1's are widely scattered.

Table 2.3 Incidence Matrix of Assemblages and Artefacts

	Artefacts				
Site Assemblage	**Iron Tools**	**Beaker Pottery**	**Stone Tools**	**Samian Ware**	**Bronze Tools**
A	1	0	0	1	0
B	0	0	1	0	0
C	0	1	0	0	1
D	0	0	0	1	0
E	1	0	0	0	1
F	0	1	1	0	0

To obtain a two-way Petrie matrix from this data (a Petrie matrix is an incidence matrix that has a block of consecutive 1's in every row; the matrix is two-way Petrie if the matrix also has a block of consecutive 1's in every column, the block in the first column starting in the first row and the block of the last column ending in the last row), the aim is to get the 1's as close as possible to the central diagonal by reordering the rows and columns. Both the artefact types and the sites will hopefully then be seriated.

Table 2.4 shows the data rearranged into a two-way Petrie matrix. For any table that permits such a rearrangement we can discover the correct ordering of sites and artefacts from the scores of the first axis of a correspondence analysis (Figure 2.7). However, correspondence analysis does not reveal the structure if the 1's and 0's are interchanged (their role is asymmetrical) — the 1's are important but the 0's are disregarded (Jongman *et al.*, 1995).

Table 2.4 Two-way Petrie Matrix of Assemblages and Artefacts

	Artefacts				
Site Assemblage	**Samian Ware**	**Iron Tools**	**Bronze Tools**	**Beaker Pottery**	**Stone Tools**
D	1	0	0	0	0
A	1	1	0	0	0
E	0	1	1	0	0
C	0	0	1	1	0
F	0	0	0	1	1
B	0	0	0	0	1

Using archaeological knowledge in conjunction with this seriation, we would infer that B is the earliest site (stone tools only, from the Palaeolithic Period) and that D is the most recently occupied site (Samian ware, typical of the Roman Period) and also that the age order of the artefact types is stone tools, beaker pottery, bronze tools, iron tools and Samian ware. Archaeological knowledge is important for the interpretation of a Petrie matrix because otherwise it would be unclear whether site B or site D contained the earliest material.

2.4.1 Seriation in Archaeology

In archaeology we distinguish between two types of seriation — frequency seriation and contextual seriation. In contextual seriation it is the duration of different artefact styles which governs the seriation and frequency seriation is not just applicable to presence or absence of artefacts, but it also measures changes in the frequency of a (ceramic) style. There are three basic assumptions behind frequency seriation:

- (Pottery) styles gradually become more popular, reach a peak popularity and then fade away.
- At a given time period, a (pot) style popular at one site would similarly be popular at another site.
- Sites must cover a single Period in the archaeological record.

Seriation by itself does not tell us which end of a given sequence is first and which is last — the true chronology has to be determined by other means e.g. by links with excavated stratigraphic sequences.

2.4.2 Faults of Correspondence Analysis

The ordination of Table 2.4 illustrates two 'faults' of correspondence analysis which are outlined below.

2.4.2.1 Changes in Artefact Composition

A correspondence analysis of the data in Tables 2.3 and 2.4 is illustrated in Figure 2.7. The artefacts are coded as it = iron tools, bp = beaker pottery, st = stone tools, sw = Samian ware and bt = bronze tools; the sites are labelled as a-f. From Table 2.4 we observe that the change in artefact composition between two consecutive assemblages is constant and we would therefore like this constant change to be reflected in equal distances between the correspondence analysis scores of neighbouring assemblages along the first axis of the map in Figure 2.7. However, this is not the case — the assemblage scores at the ends of the first axis are closer together than those in the middle of the axis.

2.4.2.2 The 'Arch Effect'

The artefact composition is explained perfectly by the ordering of the assemblages and artefacts along the first axis and it has therefore been argued that the importance of the second axis should be zero. However, the first axis of Figure 2.7 explains 45.2% of the variation in the data and the assemblage and site scores on the second axis show a quadratic relation to the first axis. This is termed the 'arch effect' (Gauch *et al.*, 1977) and denotes the phenomenon which sometimes occurs in ordination methods, where all or most of the plotted points appear in a curve. This is because although the axes are orthogonal, non-linear relationships may still exist between them — the axes are not independent. The 'arch effect' is a mathematical phenomenon, which does not correspond to any real structure in the data.

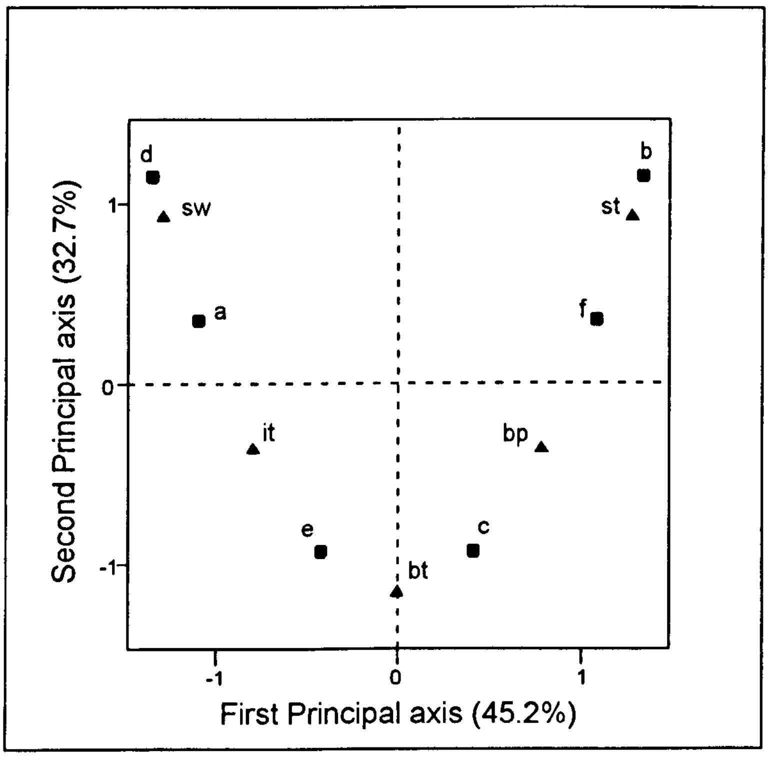

Figure 2.7 The 'Arch Effect'

Hill & Gauch (1980) believe that the 'arch effect' occurs fairly often in ecological data sets and developed the technique of 'detrended' correspondence analysis to solve the problem. This technique has since come under criticism but is not discussed here.

2.5 Correspondence Analysis and Principal Component Analysis

In the following sections we give a brief comparison of the techniques of correspondence analysis and principal component analysis. However, an explanation of principal component analysis is not given here but is deferred to Chapter Three — 'Biplots' — because is has closer similarities with these methods of analysis.

2.5.1 Joint Plots

In correspondence analysis, both the rows and columns of the data matrix are plotted together on the same picture, whereas in principal component analysis it is just the scores that are plotted. These scores correspond to the rows.

2.5.2 Size and Shape

Correspondence analysis pays more attention to 'shape' and less to 'size' than principal component analysis (Ringrose, 1990). In particular, if two rows have the same relative abundances but different absolute abundances (e.g. one is exactly twice the other) then in correspondence analysis these will have exactly the same co-ordinates on the axes, differing only in their contributions to the overall positioning of the axes. However, in principal component analysis they will have different positions and the row with the larger abundance will be further away from the origin than the other row on each axis.

2.5.3 Decomposition of Variance

The inertia in correspondence analysis is decomposed along the principal axes, just as variance is in principal component analysis. Thus, a decision can be made on how many dimensions adequately describe the data in both techniques.

2.6 Summary and Conclusions

This chapter has reviewed the technique of correspondence analysis, which allows data in the form of contingency tables or incidence matrices to be displayed in two-dimensional space and also allows us to visually identify relationships between row categories and between column categories, which is of great use in archaeology. The method is known, however, to have several faults which we illustrated using a frequency seriation problem. Seriation, as defined in archaeology, was also discussed and some of the similarities between correspondence analysis and principal component analysis were explained.

The Memphis sherd data (1.2.1) and the Melanesian starch grains (1.2.3) suggested that correspondence analysis maps can become difficult to interpret if there are many categories to display (Figures 2.2 and 2.5) and that it may therefore be advisable either to classify artefacts into broader categories initially, which will also save the archaeologist time, or to amalgamate or delete categories at a later stage. There are various methods available, both archaeological and statistical, for choosing which categories to combine and methods also exist for deleting categories. We explained that with certain types of data (e.g. starch grains, phytoliths, microfossils) it is necessary to allow for the division of categories after data collection, on the basis of an external variable. Existing methods of category selection and our extensions of these are examined in Chapter Six.

In this chapter we also discussed (using the Amarna sherds of 1.2.2) the importance of considering the effect of the number of artefacts collected, both in total and within each site, on the relationships between categories. For example, it may be that, depending on the number of artefacts collected, the relationships between wares and sites in the correspondence map vary. These issues need to be considered in detail and this forms part of Chapter Five. In addition, by looking at the output from correspondence analysis, we explained how to identify which points have played the biggest part in determining the orientations of the first two principal axes. In Chapter Six we investigate an alternative method of detecting influential points, using a jack-knife approach.

Chapter Three

Biplots

3.1 Introduction

A second, relatively recently developed collection of techniques for displaying data matrices is the set of different forms of biplot. The aim of this chapter is to bring together the algebraic details of these various forms, from a large number of sources and to combine them with guides to the important aspects of biplot interpretation, giving illustrations using both published and new data sets. Potential problems with applying biplots to archaeological data are also highlighted, for example the effects of large numbers of artefacts on the interpretation, the influence of outliers and the need for variable selection methods when vast numbers of variables have been measured. We also use concentration ellipses for summarising large numbers of observations and extend their use to compare groups of artefacts. In addition, the influence of the number of variables measured and their ability to discriminate between groups of observations is briefly discussed and we introduce the diversity biplot into archaeology for the first time. The discussion in this chapter is a prelude to the developments described in Chapters Seven and Eight, where problems such as those just described are addressed and solutions are suggested.

A biplot is a method of visualising the elements of a rectangular data matrix by representing the rows and the columns of the matrix as points or vectors in a joint display in low-dimensional space. Often, the data matrix (X) is in an observations-by-variables format; the observations are usually represented by points and the variables by vectors extending from the origin of the display. There are various forms of biplot, all of which rely on a generalised singular value decomposition (GSVD) of different scalings of the data matrix. These include the covariance, correlation, coefficient of variation, Spearman rank correlation, principal component and diversity biplots. Biplots are related to techniques such as principal component analysis, correspondence analysis and multivariate analysis of variance and just as with these their importance lies in revealing structure within the data which may, or may not, be suspected.

This chapter describes the theory behind the various forms of biplot and applies the most common ones to archaeological data, collating the fragmentary literature into a coherent account. Section 3.2 presents the standard technique of principal component analysis in preparation for Section 3.3, which introduces biplots. In Section 3.4 we define and compare the biplots of the Correlation Biplot Family, describe their properties and develop the Spearman rank correlation biplot to consider the two cases of tied ranks and absence of tied ranks separately. We describe the main biplot of the Principal Component Biplot Family in Section 3.5 and in 3.6 we describe how to assess the quality of representation of a biplot in two dimensions. The interpretation of a biplot, with applications to ceramic pots and Simpson Desert flakes, is illustrated in Section 3.7, where concentration ellipses are used to summarise large numbers of observations and to assess similarities between groups of observations. The concept of diversity is discussed in 3.8, where the diversity biplot is introduced into archaeology and comparisons are made with correspondence analysis. In 3.9 we describe another form of biplot, the symmetric biplot and the relationships between biplots and a selection of other multivariate techniques are described in 3.10. The chapter is concluded in Section 3.11 where the particular problems driven by archaeological data are summarised. These are discussed in Chapters Seven and Eight.

3.2 Principal Component Analysis

Principal component analysis (PCA) is one of the most widely used techniques of multivariate analysis and yet it has many similarities with biplots, which are much less well known. The main idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components, which are uncorrelated with each other, each containing a proportion of the variance explained in decreasing order of magnitude, so that the first component retains most of the variation present in all of the original variables (Jolliffe, 1986).

Consider a data matrix X (n × m) which consists of m variables (columns) measured on n observations (rows). For PCA, the matrix X is scaled in one of two ways, to form matrix H. If X is mean-centred by columns then we refer to PCA on the covariance matrix, but if X is standardised so that each variable has zero mean and unit variance, useful when the variables have been measured in different units, then we refer to PCA on the correlation matrix. Consider the singular value decomposition of H:

$$H = UD_{\mu}V^{T}$$

where $D_{\mu} = \text{diag}(\mu_1, \ldots, \mu_m)$ is a diagonal matrix of singular values;

$D_{\mu}^{2} = \text{diag}(\mu_1^{2}, \ldots, \mu_m^{2})$ is a diagonal matrix of eigenvalues;

U (n x m) and V (m x m) are orthonormal i.e. $U^{T}U = V^{T}V = I_m$;

U is the matrix of left eigenvectors of HH^{T};

V is the matrix of right eigenvectors of $H^{T}H$.

The matrix V defines a rotation of the original axes to a new set of axes (the principal axes). The rotation is applied to the data by postmultiplying the matrix H by V to obtain the co-ordinates of the points for the observations relative to their principal axes. These co-ordinates are often termed 'scores'. Thus, if a_{ij} is the score for the i-th observation along the j-th principal axis, it is given by:

$$a_{ij} = h_{i1}v_{1j} + \ldots + h_{im}v_{mj} = \sum_{l=1}^{m} h_{il}v_{lj}$$

and the matrix of scores, A (n x m), is given by:

$$A = HV = UD_{\mu}.$$

If h is the vector of m variables then hv_1 denotes the scores on the first principal axis where v_1 is a column vector containing $v_{11},\ldots, v_{1m}$. It is a standard result that the first principal axis is the linear function of the elements of h that has maximum variance; the second principal axis is the linear function hv_2 that has maximum variance, subject to being uncorrelated with hv_1.

For any specified k < m, PCA finds the subspace of k dimensions for which the sum of squared perpendicular distances of $h_1,\ldots, h_m$ to the subspace is minimised. Each point is then represented by the projection of its original position onto that subspace and to obtain the best fitting subspace of k dimensions we take the first k columns of scores of A. Up to m principal components could be found but it is hoped that most of the variation in y will be accounted for by the first k principal components, where k << m. Using k principal components instead of m variables considerably reduces the dimensionality of the problem when k << m, but usually the values of all m variables are still needed in order to calculate the principal components, because each principal component is generally a function of all m variables.

Because we hope to reduce the dimensionality of the data from the original m dimensions to a much smaller number, k, we are interested in measuring the percentage of variation in the data accounted for by the first k principal components. This is given by:

$$t_k = 100 \times \frac{\sum_{j=1}^{k} \mu_i^2}{\sum_{j=1}^{m} \mu_i^2}.$$

We mentioned at the beginning of this section that principal component analysis has close similarities with certain forms of biplot (particularly the principal component biplot) and these are described in Section 3.10.1. However, PCA only displays the observations, whereas biplots display both observations and variables and this is their major advantage. We describe biplots in Section 3.3 below.

3.3 Introduction to Biplots

Any matrix X (n x m), of rank r, can be factorised as:

$$X = FG^T$$

where F (n x r) and G (m x r) are both of rank r. Each x_{ij}, the (i,j)-th element of X, can be expressed as the inner product of f_i and g_j, where f_i^T is the i-th row of F and g_j is the j-th column of G^T. In a biplot the co-ordinates of the rows (observations) of matrix X are usually represented by points and the columns (variables) are usually represented by vectors, where the j-th variable is represented by a vector from the origin to the point g_j. Biplots represent each element x_{ij} geometrically, as in Figure 3.1, where a perpendicular is dropped from point f_i onto vector g_j and the distance from the origin to the foot P of this perpendicular is multiplied by the length of vector g_j. The product corresponds to the inner product $f_i^Tg_j$. The geometrical interpretation of these points is in terms of the distances of each point from the origin and the cosines of the angles which pairs of the vectors subtend at the origin. There is no particular reason why the variables are represented by vectors rather than points, but it allows us to more easily visualise the correlations between the variables, which are represented by the size of the angle subtended between their vectors at the origin.

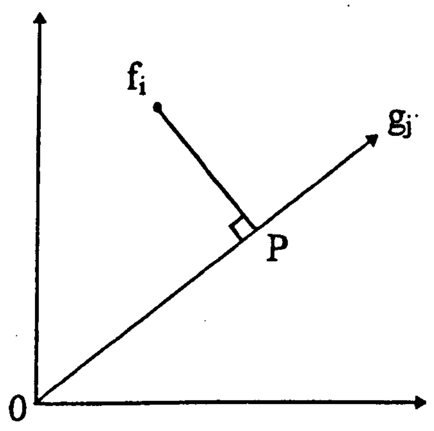

Figure 3.1 Graphical Representation of the Elements of a Matrix

If X is of rank two then the rows and columns are displayed exactly on a two-dimensional plot, but otherwise they are a least squares approximation to the full rank matrix. The quality of representation of the two-dimensional display is evaluated by expressing as a percentage the ratio of the sum of eigenvalues in two dimensions to the sum of eigenvalues in full r-dimensional space. The interpretation of a biplot focuses on the relationship between the variables — those variables represented by vectors which subtend a small angle at the origin are highly correlated, whereas those with an angle of approximately 90° between them are considered to be uncorrelated. More specific details regarding the interpretation of biplots are given in Sections 3.6 and 3.7. These vary according to the form of biplot.

Each time a biplot is implemented a GSVD is required. However, when the GSVD is carried out, the left and right eigenvectors are determined independently and this can lead to arbitrary sign changes in the eigenvectors and hence in the resulting co-ordinates (see 5.2.4). There are two main families of biplots — these are described in Sections 3.4 and 3.5 — they differ in the scalings of the matrix of singular values obtained from the GSVD of matrix X. The diversity biplot is discussed in 3.8, where we propose applying it to archaeological data and the symmetric biplot which has another alternative scaling is described in 3.9. In the subsequent sections of this chapter we use the notation of Chapter One.

3.4 The Correlation Biplot Family

Considering the GSVD of a matrix X for the simplest case of $\Omega = I_n$ and $\Phi = I_m$, as described in Chapter One, we have a = 0 and b = 1 for all biplots in this family. Thus, the co-ordinates of the observations and the variables in two dimensions are given by the first two columns of F and G respectively, where:

$$\begin{aligned} F &= U \\ G &= VD_{\mu}. \end{aligned} \tag{3.1}$$

In Sections 3.4.1-3.4.4 we describe the four most commonly used biplots of the Correlation Biplot Family. These include the covariance biplot, which is most suitable for variables measured in the same units, the correlation biplot, which is useful when the variables are measured on different scales and the coefficient of variation biplot which is suited to data matrices in which the relative variability of the variables, rather than the absolute variability, is of main interest. We also describe the Spearman rank correlation biplot, which is useful when there are large discrepancies between the magnitudes of the observations because it ranks the observations within each variable. All these biplots involve scaling a data matrix X (n x m) of rank r. In addition, the covariance, correlation and coefficient of variation biplots are all scaled by $\left(\frac{1}{n-1}\right)^{\frac{1}{2}}$ to ensure that properties 3.4.6.1 and 3.4.6.2 (see below) hold automatically. These various types of biplot are illustrated on the ceramic pots (1.2.5), flint tools and flake debitage (1.2.6) in Section 3.7.

3.4.1 The Covariance Biplot

The covariance biplot is a common form of biplot, which involves column-centring matrix X to form matrix Y:

$$Y = \left(\frac{1}{n-1}\right)^{\frac{1}{2}} \left(X - \overline{X}\right)$$

where $\overline{X}$ is a matrix of column means. By calculating a GSVD of matrix Y, we obtain the co-ordinates of the observations and the variables, as defined in the previous section. The squared norm of matrix Y is of interest because it allows us to identify (loosely) which variables are of most importance in the biplot analysis, by calculating their relative contributions to the squared norm (Underhill, 1990) and this is discussed further in 3.4.5.7. The squared norm of Y is given by:

$$\|Y\|^2 = \sum_{j=1}^{m} s_j^2$$

where s_j is the standard deviation of variable j. The relative contribution of variable j to the squared norm is denoted by:

$$r_j = \frac{s_j^2}{\|Y\|^2}.$$

This depends on the magnitude and scale of measurement of the variables.

3.4.2 The Correlation Biplot

In the correlation biplot the matrix X is column-standardised as follows to form matrix C:

$$C = \left(\frac{1}{n-1}\right)^{\frac{1}{2}} \left(X - \overline{X}\right)\left(\mathrm{diag}(s_1, \ldots, s_m)\right)^{-1}$$

where $s_1, \ldots, s_m$ are the standard deviations of the variables. Because of the standardisation imposed on matrix X, the standard deviations of the variables of C are all equal to one and the length of the vector representing variable j, given by $\|g_j\|$, is equal to one for each variable, in full r-dimensional space. The squared norm of matrix C is given by:

$$\|C\|^2 = m$$

and thus each variable j makes an equal contribution to the squared norm of $r_j = \frac{1}{m}$.

3.4.3 The Coefficient of Variation Biplot

The coefficient of variation biplot was developed by Underhill (1990). The matrix X is scaled as follows:

$$E = \left(\frac{1}{n-1}\right)^{\frac{1}{2}}\left(X - \overline{X}\right)\left(\mathrm{diag}\left(\overline{x}_1, \ldots, \overline{x}_m\right)\right)^{-1}$$

where $\overline{x}_1, \ldots, \overline{x}_m$ are the means of the variables. Because of this scaling of matrix X, the standard deviations of the columns of E are the coefficients of variation of the variables. Thus the length of the vector representing variable j in the display, given by $\left\|g_j\right\|$, gives the coefficient of variation of the variable, because:

$$\left\|g_j\right\| = \frac{s_j}{\overline{x}_j}.$$

This display is useful for data matrices in which the relative variability of the columns, rather than the absolute variability, is of prime interest. However, the variables must be such that the coefficient of variation is meaningful — they need to be measured on a ratio scale. Underhill (1990) says that the coefficient of variation biplot is a compromise between leaving the variables in their original scales and units (so that the variables with the largest standard deviations dominate) and transforming by the standard deviations so that each variable has equal importance. The squared norm of matrix E is given by:

$$\left\|E\right\|^2 = \sum_{j=1}^{m}\left(\frac{s_j}{\overline{x}_j}\right)^2$$

and the relative contribution of variable j to the squared norm is denoted by:

$$r_j = \frac{\left(\frac{s_j}{\bar{x}_j}\right)^2}{\|E\|^2},$$

which is proportional to the coefficient of variation of the variable. Variables with small coefficients of variation make a small contribution to the squared norm and vice versa, regardless of their original scales of measurement.

3.4.4 The Spearman Rank Correlation Biplot

The Spearman rank correlation biplot was introduced by Iloni (1991). In this form of biplot the observations within each column of matrix X are ranked and the matrix of ranks, multiplied by $\left(\frac{1}{n-1}\right)^{\frac{1}{2}}$, is given by Z, with elements z_{ij}. In the next two sections we extend the comments by Iloni (1991) to consider the cases of tied ranks and absence of tied ranks, separately.

3.4.4.1 Absence of Tied Ranks

Firstly, we consider the case of no ties. If there are no tied observations then the variables of Z have the same norm and the squared norm of matrix Z with no ties is given by:

$$\|Z\|^2 = \frac{n(n+1)(2n+1)m}{6}.$$

Each individual variable has squared norm:

$$\|Z_j\|^2 = \frac{n(n+1)(2n+1)}{6}.$$

If none of the variables have tied ranks then they each have an equal relative contribution to the squared norm, given by:

$$r_j = \frac{1}{m}.$$

3.4.4.2 Tied Ranks

In this section we develop expressions for the squared norm and relative contributions to the squared norm for the case of tied ranks. The squared norm of a variable j which has w (≥2) tied observations out of n, is given by:

$$\left\|Z_j\right\|^2 = \frac{n(n+1)(2n+1)}{6} - \frac{(w-1)w(w+1)}{12}. \tag{3.2}$$

Therefore, for p variables with w ties and m variables with no ties, the squared norm of matrix Z is:

$$\left\|Z\right\|^2 = \frac{2mn(n+1)(2n+1) - wp(w-1)(w+1)}{12}$$

and the relative contribution of a variable j with w ties to the squared norm of a matrix of which p variables have w ties is:

$$r_j = \frac{2n(n+1)(2n+1) - w(w-1)(w+1)}{2mn(n+1)(2n+1) - pw(w-1)(w+1)}. \tag{3.3}$$

Similarly, the relative contribution of a variable with no ties to the squared norm of a matrix of which p variables have w ties is:

$$r_j = \frac{2n(n+1)(2n+1)}{2mn(n+1)(2n+1) - pw(w-1)(w+1)}. \tag{3.4}$$

We see that the denominators of (3.3) and (3.4) are equal and so if a variable has tied observations then in the Spearman rank correlation biplot this variable contributes less to the squared norm and can be loosely considered to be less important. In fact, the more observations within a variable that are tied, the less contribution that that variable makes to the squared norm. The contribution of a variable to the squared norm also varies with the number of pairs of ties, triples of ties and so on and we can see from (3.2) that more information is lost (i.e. there is less contribution to the squared norm) when three values tie together, compared with when two pairs of values tie. Because the other forms of biplot all use a transformation of the actual measurement data and not ranks, no information is lost when tied values are obtained for these biplots.

3.4.5 Comparisons between Biplots

Having described the most common types of biplot, the following sections comment on the various scalings of the data matrix X which result in the covariance, correlation, coefficient of variation and Spearman rank correlation biplots. It is important to be aware of the various features of each before undertaking any data analysis so that the most suitable biplot can be chosen.

3.4.5.1 Units of Measurement

The elements in the scaled data matrices of C, E and Z in the cases of the correlation, coefficient of variation and Spearman rank correlation biplots respectively, have the potential advantage of being dimensionless, while those of Y, in the covariance biplot, are in the units of measurement of the original variables. Thus, for the flint tool data introduced in 1.2.6 and discussed in 3.7.2, where some measurements are in millimetres and one is in grams, the covariance biplot is not really suitable.

3.4.5.2 Robustness to Outliers

The Spearman rank correlation biplot is useful when there are large discrepancies between the magnitudes of the observations for a particular variable, because it is robust with respect to outliers (we will see this with the flint tool data in Figure 3.9). However, if we want to preserve differences in the magnitude of observations, information is clearly lost in this form of biplot.

3.4.5.3 Correlations between Variables

In all the biplots of the Correlation Biplot Family, the cosine of the angle subtended at the origin between the vectors representing two variables approximates the correlation between the two variables, but in the case of the Spearman rank correlation biplot this is the rank correlation.

3.4.5.4 Variability

The correlation biplot and Spearman rank correlation biplot do not display variability because the variables are scaled to have a standard deviation of one if perfectly represented in two dimensions. However, the coefficient of variation biplot does display variability. If, in the covariance biplot, the scales of measurement of the variables are different, then only relative variabilities can be compared.

3.4.5.5 Standardisation of Variables

In the correlation biplot all the variables are standardised to have a mean of zero and variance of one. This prevents the plot from being dominated by a few variables, but has the disadvantage that the relative variabilities are not displayed. Also, by standardising the scales of measurement in the correlation biplot, the relative weights of variables having small standard deviations are effectively inflated (and vice versa). This may be a desirable feature in some applications but might divert attention away from particular observations occurring as extremes in some variables.

3.4.5.6 Scales of Measurement

If the scales of measurement of the variables differ greatly, the variables on larger scales dominate the plot in the covariance biplot, at the expense of the other variables whereas in the correlation biplot and Spearman rank correlation biplot the relative importance of each variable is the same. In the coefficient of variation biplot, variables with large coefficients of variation tend to be associated with the large singular values and therefore have a high quality of display and are relatively dominant (and vice versa), but this does not depend on the original scale of measurement. Variables that are highly correlated with those with large coefficients of variation will also be well displayed.

3.4.5.7 Relative Contributions to the Squared Norm

The relative contributions of each variable to the squared norm may loosely be thought of as the 'weight' or importance of each variable in the analysis, although these quantities do not appear as weights in any equation (Underhill, 1990). The relative contributions of variable j to the squared norm of the matrix were defined in

Sections 3.4.1-3.4.4, but we summarise the main points below.

- The contributions of each variable to the squared norm in a covariance biplot depend on the magnitudes and the scales of measurement of the variables. Variables making larger contributions tend to dominate the biplot and so the covariance biplot may fail to find any subtle multivariate structure in a data matrix if there are large relative differences between the smallest and largest standard deviations (Underhill, 1990).

- In the correlation biplot, each variable makes an equal contribution to the squared norm. This inflates the relative contributions to the squared norm of variables having small standard deviations (and vice versa).

- In the coefficient of variation biplot, the contribution of each variable to the squared norm is proportional to the coefficient of variation. Variables with small coefficients of variation make a small contribution to the squared norm and vice versa, regardless of their original scales of measurement.

- If a variable has tied observations then in the Spearman rank correlation biplot this variable contributes less to the squared norm than other variables and variables with greater numbers of ties give a smaller relative contribution. The contribution to the squared norm also varies with the number of pairs of ties, triples of ties and so, for example, there is less contribution from a variable when three values tie together compared with when two pairs of values tie.

3.4.6 Geometrical Properties of the Correlation Biplot Family

The geometrical interpretation of the biplots in this family is in terms of the distances of each observation from the origin and the cosines of the angles which pairs of the vectors representing variables subtend at the origin. Four important properties are listed below and it is useful to bear these in mind when interpreting a biplot.

3.4.6.1 Standard Deviations of Variables

The distance of g_j, the j-th column of G, from the origin is given by $\|g_j\|$ and for the covariance, correlation and Spearman rank correlation biplots this approximates the standard deviation, s_j, of variable j. We show this below for the covariance biplot. As reported in Barr, Underhill & Kahn (1990), it follows from equation (3.1) that:

$$\begin{aligned} GG^T &= VD_\mu D_\mu V^T \\ &= Y^T Y \\ &= \frac{1}{n-1}(X-\overline{X})^T(X-\overline{X}) \\ &= S. \end{aligned}$$

Thus:

$$\begin{aligned} \|g_j\|^2 &= \frac{1}{n-1}\sum_{t=1}^{n}(x_{tj}-\overline{x}_j)^2 \\ &= s_j^2. \end{aligned} \tag{3.5}$$

Similar arguments follow for the correlation and Spearman rank correlation biplots, but the standard deviation can take a maximum value of one in these cases. However, for the coefficient of variation biplot, $\|g_j\|$ approximates the coefficient of variation of the variable, denoted by $\frac{s_j}{\overline{x}_j}$, because:

$$
\begin{aligned}
GG^T &= VD_\mu D_\mu V^T \\
&= E^T E \\
&= \frac{1}{n-1}\left(X-\overline{X}\right)^T\left(X-\overline{X}\right)\mathrm{diag}\left(\overline{x}_1,\ldots,\overline{x}_m\right)^{-2} \\
&= \frac{S}{\overline{X}^2}.
\end{aligned}
$$

Thus:

$$
\begin{aligned}
\left\|g_j\right\|^2 &= \frac{1}{n-1}\sum_{t=1}^{n}\frac{\left(x_{tj}-\overline{x}_j\right)^2}{\overline{x}_j^2} \\
&= \frac{s_j}{\overline{x}_j^2}.
\end{aligned}
$$

3.4.6.2 Covariances between Variables

We now consider the covariances between variables. For the covariance, correlation and Spearman rank correlation biplots, the inner product of variables j and j', given by $g_j{}^T g_{j'}$, approximates the covariance $s_{jj'}$ between columns j and j' of X, because:

$$
g_j \cdot g_{j'} = \frac{1}{n-1}\left(x_{tj}-\overline{x}_j\right)\left(x_{tj'}-\overline{x}_{j'}\right). \tag{3.6}
$$

However, this does not hold for the coefficient of variation biplot because:

$$
g_j \cdot g_{j'} = \frac{1}{n-1}\sum_{t=1}^{n}\frac{\left(x_{tj}-\overline{x}_j\right)}{\overline{x}_j}\frac{\left(x_{tj'}-\overline{x}_{j'}\right)}{\overline{x}_{j'}}.
$$

3.4.6.3 Correlations between Variables

The cosine of the angle between g_j and $g_{j'}$ (or, equivalently, their inner product) approximates the correlation between the variables j and j' of X. Thus, if two variables j and j' are highly positively correlated then g_j and $g_{j'}$ will lie in the same direction from the origin and if they are negatively correlated then they will lie on opposite sides of the origin. If the correlation is close to zero, then they will tend to lie at right angles to each other.

For all biplots in the Correlation Biplot Family (covariance, correlation, coefficient of variation and Spearman rank correlation), the cosine of the angle between g_j and $g_{j'}$ is given by:

$$\cos(\theta_{jj'}) = \frac{g_j \cdot g_{j'}}{\|g_j\| \|g_{j'}\|}.$$

Using equations (3.5) and (3.6), for the covariance, correlation and coefficient of variation biplots we obtain:

$$\cos(\theta_{jj'}) = \frac{\sum_{t=1}^{n}(x_{tj} - \bar{x}_j)(x_{tj'} - \bar{x}_{j'})}{\sqrt{\sum_{t=1}^{n}(x_{tj} - \bar{x}_j)^2 \sum_{t=1}^{n}(x_{tj'} - \bar{x}_{j'})^2}}.$$

For the Spearman rank correlation biplot x is replaced with z in the above.

3.4.6.4 Distances between Variables

Distances between variable points (the tips of the vectors) in the display represent the Euclidean distances between the columns of the matrices Y, C, E or Z of the covariance, correlation, coefficient of variation and Spearman rank correlation biplots respectively. For the covariance biplot, denoting the j-th column of Y by $Y_{(j)}$, we have:

$$\begin{aligned}\|g_j - g_{j'}\|^2 &= \left(Y_{(j)} - Y_{(j')}\right)^T \left(Y_{(j)} - Y_{(j')}\right) \\ &= \sum_{t=1}^{n}\left[\left(x_{tj} - \bar{x}_j\right) - \left(x_{tj'} - \bar{x}_{j'}\right)\right]^2\end{aligned}$$

and so the square of the distance between g_j and $g_{j'}$ is proportional to the Euclidean distance between the centred columns j and j' of X.

3.4.6.5 Distances between Observations

Distances between observations in the display represent the Mahalanobis distances between the rows of the matrices. From (3.1) it follows that:

$$F = U = YVD_{\mu}^{-1}.$$

Therefore:

$$\begin{aligned} FF^{T} &= YVD_{\mu}^{-2}V^{T}Y^{T} \\ &= YS^{-1}Y^{T} \\ &= \frac{1}{n-1}\left(X - \overline{X}\right)S^{-1}\left(X - \overline{X}\right)^{T}. \end{aligned}$$

Thus:

$$\left\|f_{k}\right\|^{2} = \frac{1}{n-1}\sum_{t=1}^{m}(x_{kt} - \overline{x}_{t})^{2}s_{kk}^{-1}$$

where s_{kk} is the k-th standard deviation. Therefore, the distance between f_k and $f_{k'}$ is, from Barr, Underhill & Kahn (1990):

$$\begin{aligned} \left\|f_{k} - f_{k'}\right\|^{2} &= \left(Y^{(k)} - Y^{(k')}\right)S^{-1}\left(Y^{(k)} - Y^{(k')}\right)^{T} \\ &= \sum_{t=1}^{m}\left(y_{kt} - y_{k't}\right)S^{-1}\left(y_{kt} - y_{k't}\right) \end{aligned}$$

which is the Mahalanobis distance between $Y^{(k)}$ and $Y^{(k')}$, where $Y^{(k)}$ is the k-th row of Y.

Because the displays are often a two-dimensional approximation to a higher dimensional data matrix, there are always distortions and Underhill (1990) says that these distortions are unevenly distributed over the displayed observations, so that while some (or most) of the observations may be well represented, others are poorly displayed.

3.4.7 Significance Testing

Whilst biplots are essentially exploratory techniques, they can be adapted to testing the significance of the association between any two variables as explained by Gabriel (1995). If we recall that the chi-squared statistic equals the square of the correlation coefficient multiplied by the sample size n and that the chi-squared statistic with one degree of freedom is the square of a Standard Normal random variable, then significance can be established by checking whether the absolute value of the inner product that approximates the correlation coefficient in a biplot exceeds the appropriate percentage point of the Standard Normal distribution, divided by $\sqrt{n}$. However, the vectors are obtained by projection onto the two-dimensional plane which may distort the angles between them and so it is safer to use the test only for variables which are well represented on a biplot (see 3.6). If the associations of all pairs of variables are to be tested simultaneously then the test must be adjusted by using a Bonferroni correction. In Section 3.7.1.2 we will see an application of this test to the ceramic pots (1.2.5).

3.5 The Principal Component Biplot Family

In this section we describe the main biplot of the Principal Component Biplot Family. Considering the GSVD of a matrix X, for the simplest case of $\Omega=I_n$ and $\Phi=I_m$, described in Chapter One, we have a = 1 and b = 0. The co-ordinates of the observations and the variables in two dimensions are given by the first two columns of F and G respectively, where:

$$\begin{aligned} F &= UD_\mu \\ G &= V. \end{aligned} \tag{3.7}$$

One disadvantage of this choice of F and G, as compared with the Correlation Biplot Family of Section 3.4, is that properties 3.4.6.1 and 3.4.6.2 are no longer valid, because:

$$GG^T=VV^T\neq S.$$

The main biplot in this family is the principal component biplot, in which the data matrix X is usually either column-centred:

$$Y=\left(\frac{1}{n-1}\right)^{\frac{1}{2}}\left(X-\overline{X}\right)$$

or column-standardised:

$$C=\left(\frac{1}{n-1}\right)^{\frac{1}{2}}\left(X-\overline{X}\right)\left(\mathrm{diag}\left(s_1,\ldots,s_m\right)\right)^{-1}.$$

Having introduced the principal component biplot, we now describe the geometrical properties of biplots in the Principal Component Biplot Family.

3.5.1 Geometrical Properties of the Principal Component Biplot Family

Biplots in the Principal Component Biplot Family have three important properties. These should be considered when interpreting a biplot and are explained below.

3.5.1.1 Correlations between Variables

The cosine of the angle between g_j and $g_{j'}$ approximates the correlation between the variables j and j' of X. Using scaling (3.7), we see that:

$$G = V = Y^T U D_\mu^{-1}.$$

Therefore:

$$\begin{aligned} GG^T &= VV^T \\ &= Y^T U D_\mu^{-2} U^T Y \\ &= Y^T S^{-1} Y \\ &= \frac{1}{n-1}\left(X - \overline{X}\right)^T S^{-1}\left(X - \overline{X}\right). \end{aligned}$$

Thus: $$\left\|g_j\right\|^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_{tj} - \overline{x}_j)^2 s_{jj}^{-1}$$

and $$g_j \cdot g_{j'} = \frac{1}{n-1}\sum_{t=1}^{n}(x_{tj} - \overline{x}_j) s_{jj}^{-\frac{1}{2}} s_{j'j'}^{-\frac{1}{2}} (x_{tj'} - \overline{x}_{j'}).$$

For the principal component biplot the cosine of the angle between g_j and $g_{j'}$ is given by:

$$\cos(\theta_{jj'}) = \frac{g_j \cdot g_{j'}}{\left\|g_j\right\|\left\|g_{j'}\right\|}.$$

Therefore:

$$\cos(\theta_{jj'}) = \frac{\sum_{t=1}^{n}(x_{tj} - \bar{x}_j)(x_{tj'} - \bar{x}_{j'})}{\sqrt{\sum_{t=1}^{n}(x_{tj} - \bar{x}_j)^2 \sum_{t=1}^{n}(x_{tj'} - \bar{x}_{j'})^2}}$$

which is the Pearson product-moment correlation coefficient between variables j and j'.

3.5.1.2 Distances between Variables

The distances between variable points (the tips of the vectors) in the display represent the Mahalanobis distances between the columns of the matrix Y (Gabriel, 1995). The 'proof' of this could not be found stated explicitly in the literature, but is clearly as follows. From 3.5.1.1, denoting the j-th column of Y by $Y_{(j)}$, it follows that:

$$\begin{aligned} \|g_j - g_{j'}\|^2 &= \left(Y_{(j)} - Y_{(j')}\right)^T S^{-1}\left(Y_{(j)} - Y_{(j')}\right) \\ &= \sum_{t=1}^{n}\left(y_{tj} - y_{tj'}\right) S^{-1}\left(y_{tj} - y_{tj'}\right) \end{aligned}$$

This is the Mahalanobis distance between columns j and j' of Y.

3.5.1.3 Distances between Observations

Distances between observations in the display represent the Euclidean distances between the rows of the matrix. The 'proof' of this is as follows. From (3.7) we see that:

$$\begin{aligned} FF^T &= UD_\mu D_\mu U^T \\ &= YY^T \\ &= \frac{1}{n-1}\left(X - \bar{X}\right)\left(X - \bar{X}\right)^T. \end{aligned}$$

Thus:

$$\|f_k\|^2 = \frac{1}{n-1}\sum_{t=1}^{m}(x_{kt} - \bar{x}_t)^2.$$

Therefore, the distance between f_k and $f_{k'}$ is given by:

$$\|f_k - f_{k'}\|^2 = \left(Y^{(k)} - Y^{(k')}\right)\left(Y^{(k)} - Y^{(k')}\right)^T$$

$$= \sum_{t=1}^{m}\left[(x_{kt} - \bar{x}_t) - (x_{k't} - \bar{x}_t)\right]^2.$$

Thus, the square of the distance between f_k and $f_{k'}$ is proportional to the Euclidean distance between the centred columns k and k' of X.

3.6 Quality of Representation in Two Dimensions

When applying biplots we usually consider two-dimensional displays, because these give the best visual representation of the data. However, this means that only the first two columns of F and the first two columns of G, which correspond to the two largest singular values μ_k (k = 1,2), are used. When implementing a biplot, there are four main goodness of fit measures to help us decide whether the two-dimensional representation of our data is adequate. These are explained below and applied to the ceramic pots (1.2.5) and Simpson Desert flint tools (1.2.6) in Section 3.7.

3.6.1 The Data Matrix

For both families of biplots previously described, the elements of the scaled data matrix X are represented in a biplot where the goodness of fit is given by the ratio of the sum of the eigenvalues in two dimensions to the sum of the eigenvalues in full r-dimensional space. This is usually expressed as a percentage:

$$d_p = 100 \times \frac{\sum_{k=1}^{2} \mu_k^2}{\sum_{k=1}^{r} \mu_k^2}. \tag{3.8}$$

This is the most important measure for assessing goodness of fit and we have found that as a rough rule of thumb, at least 50% of the variation in archaeological data should be explained in the first two dimensions in order to provide a useful two-dimensional representation.

3.6.2 The Principal Co-ordinates

In the Correlation Biplot Family the co-ordinates of the variables are known as principal co-ordinates because these involve the singular values, whereas in the Principal Component Biplot Family it is the co-ordinates of the observations which are the principal co-ordinates. In the Correlation Biplot Family, the elements of S, the variance-covariance matrix, have goodness of fit given by the ratio of the sum of the squared eigenvalues in two dimensions to the sum of the squared eigenvalues in full r-dimensional space. Again, this is usually expressed as a percentage:

$$q_p = 100 \times \frac{\sum_{k=1}^{2} \mu_k^4}{\sum_{k=1}^{r} \mu_k^4}.$$

A high value of q_p indicates that both variances and covariances are closely approximated by the squared lengths of g-vectors and their inner products respectively. For the Principal Component Biplot Family, as for the Correlation Biplot Family, q_p is a measure of goodness of fit of the principal factors. Note that q_p is higher than the goodness of fit of X itself (see (3.8)), which is approximated jointly by both principal and standard factors (Gabriel, 1995).

3.6.3 Inter-row and Inter-column Distances

The inter-row (Mahalanobis) distances for the Correlation Biplot Family and the inter-column (Mahalanobis) distances for the Principal Component Biplot Family are approximated, in two dimensions, with goodness of fit:

$$t_s = \frac{2}{r}.$$

This is because it is the inner products that are being directly approximated — the approximation of the inter-row or inter-column distances is indirect (Barr, Underhill & Kahn, 1990). The value of t_s appears low, but this is because it evaluates the goodness of fit of Mahalanobis distances in standard form (Gabriel, 1995) — it is the values in principal form which are well represented (as we saw in 3.6.2) and thus the type of biplot must be chosen appropriately, depending on whether the observations or variables are of main interest.

3.6.4 Variables

The quality of the representation of variable j in both biplot families may, as in correspondence analysis (Greenacre, 1984), be defined as:

$$q_j = 100 \times \frac{\left\| g_{j[2]} \right\|^2}{\left\| g_j \right\|^2}.$$

This is the squared cosine of the angle between the vector g_j in r-space and $g_{j[2]}$, the vector in the displayed two-dimensional subspace, expressed as a percentage. When very large numbers of variables are measured, as is common in archaeology, we are often interested in reducing this number whilst still retaining any group structure amongst the artefacts, because this will save the archaeologist time (and sometimes money). In Chapter Seven we discuss existing variable selection methods for use with principal component analysis and extend and develop them to biplots. However, other factors may become relevant such as ease of measurement of variables and the goodness of fit measure above could also help us with the selection process.

3.7 Interpreting Biplots

Having described, in considerable detail, the theory of the most common biplots and also explained how to assess whether a two-dimensional representation is adequate, we now illustrate these biplots using two of the most common types of artefacts recovered in archaeology — namely pottery and flint. We return to these data throughout Chapters Seven and Eight.

3.7.1 Application to Ceramic Pots

The various types of biplot are illustrated on the ceramic pot data (1.2.5), which consist of 13 measurements (all in cm) on each of 30 ceramic pots. Archaeological interest lies mainly in assessing whether three groups of pots can be distinguished, corresponding to the three potters who made them, on the basis of these measurements. Also of interest is whether any groupings can be identified when using fewer measurements (hence saving time and money) and this is addressed in Chapter Seven.

The four types of biplot from the Correlation Biplot Family — the covariance, correlation, coefficient of variation and Spearman rank correlation biplots — and the principal component biplot, are illustrated in the sections that follow. Not all of these are necessarily appropriate to answer these particular questions on these data (usually the correlation or principal component biplot is the most appropriate), but it is a useful illustration to see each of them applied to the same data set. In 3.7.1.6 we discuss their relative merits, relevance to the problem and the interpretation of the results in relation to the underlying archaeological objectives. In the plots that follow, each of the 30 pots is represented by a circle and each of the 13 variables by a line (vector) emanating from the origin. For biplots in the Correlation Biplot Family, each biplot has quality of representation of inter-row distances of 15.38% in two dimensions which is clearly low, but not unexpected (see 3.6.3). For the principal component biplot it is the inter-vector distances which have quality of representation 15.38%. The qualities of representation of individual variables as discussed in 3.6.4 are listed in Table 3.1 for all five biplots.

Table 3.1 Quality of Representation of Individual Variables (%) for the Ceramic Pots

	Biplot				
Variable	**Covariance**	**Correlation**	**Coefficient of Variation**	**Spearman Rank Correlation**	**Principal Component**
1	88.3	84.6	19.8	85.2	19.9
2	45.6	42.3	9.8	55.5	8.2
3	95.4	86.6	67.2	85.4	17.5
4	75.5	77.0	26.1	76.3	14.9
5	68.5	64.1	7.2	58.0	12.5
6	85.7	91.2	36.0	86.3	23.3
7	69.7	72.8	3.4	78.8	17.0
8	70.5	76.3	11.4	81.8	15.0
9	68.8	76.4	57.7	79.8	18.3
10	80.0	85.1	64.1	78.8	18.6
11	14.8	19.9	99.4	15.9	5.1
12	79.2	81.3	99.7	76.0	18.9
13	38.9	50.7	6.7	45.0	10.7

From Table 3.1 we see that the quality of representation for all variables in the principal component biplot is low, because of the scaling involved in this type of biplot (see 3.5). We also note that the quality of representation of variables in the coefficient of variation biplot varies considerably from as low as 3.4% for variable 7, to as high as 99.7% for variable 12 and this is probably because variable 7 has a low and variable 12 a high, coefficient of variation. Variable 11 is poorly represented in all but the coefficient of variation biplot. Investigating further, we see that in the first three dimensions variable 11 has a quality of representation of 15.6% in the covariance biplot, 59.0% in the correlation biplot and 87.3% in the Spearman rank correlation biplot and thus a considerable proportion of the variation is hidden in the third dimension. These differences in quality of representation are interesting because if we consider variable 11 to be particularly important, or variable 7 is very difficult to measure, we might consider a coefficient of variation biplot to be most appropriate. In addition, given that the majority of variables in the biplots of the Correlation Biplot Family are well represented in two dimensions, it may not be worth looking at three dimensions. This is aside from problems of visualisation which clearly favour two dimensions, although the first and third, or second and third, principal axes could be plotted against each other if a substantial percentage of variation is explained in the third dimension. The next few sections consider each biplot in turn.

3.7.1.1 The Covariance Biplot

Constructing a covariance biplot for these data, as explained in 3.4.1 and representing the results in the first two dimensions produces Figure 3.2.

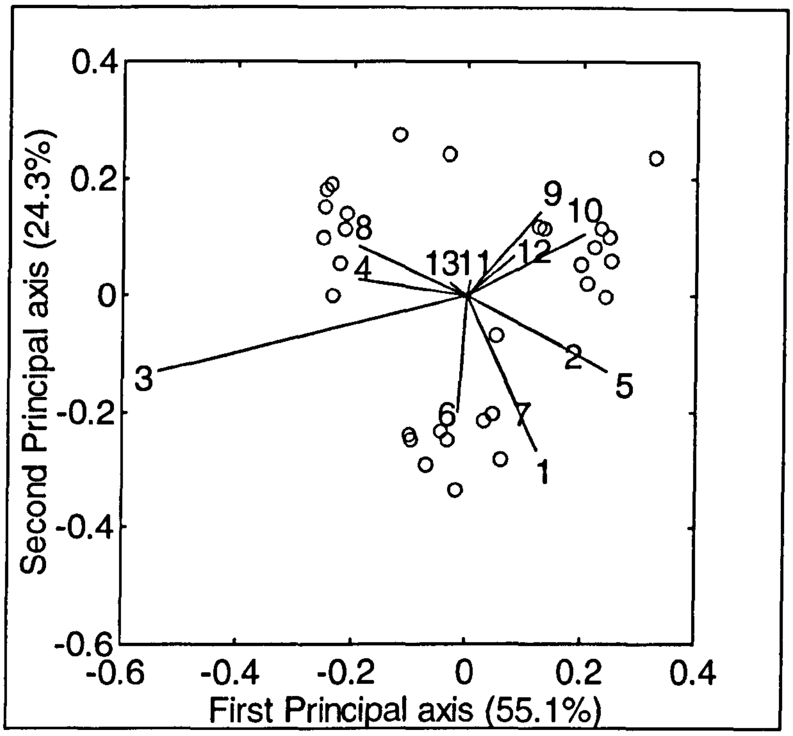

Figure 3.2 Covariance Biplot of Ceramic Pots

The data matrix Y is represented with goodness of fit 79.4%, whereas the goodness of fit of the variance-covariance matrix is 96.7% and so both these values are more than adequate for us to be confident in our interpretations of the display. We interpret the biplot as follows: pairs of variables with small angles between them, such as 1 & 7 and 2 & 5, are highly positively correlated; pairs of variables with an angle of approximately 90° between them, such as 1 & 10 and 1 & 3, are uncorrelated; and pairs of variables with an angle of approximately 180° between them, such as 2 & 8, are highly negatively correlated. Pots which are similar as regards measurements are located close together on the plot and thus there appear to be three groups of pots represented: one group towards the top left, one group towards the top right and one group towards the bottom of the picture. We see from the plot that variables 11, 12 and 13 have short vectors and from 3.4.6.1 we know that the length of the vector approximates the standard deviation of the variable. This is not really surprising because variables 11, 12 and 13 are internal depth of footring, wall thickness and lip thickness respectively and we would expect these to have low variability compared with variables such as pot height.

3.7.1.2 The Correlation Biplot

Figure 3.3 displays the data in the form of a correlation biplot in two dimensions, as described in 3.4.2.

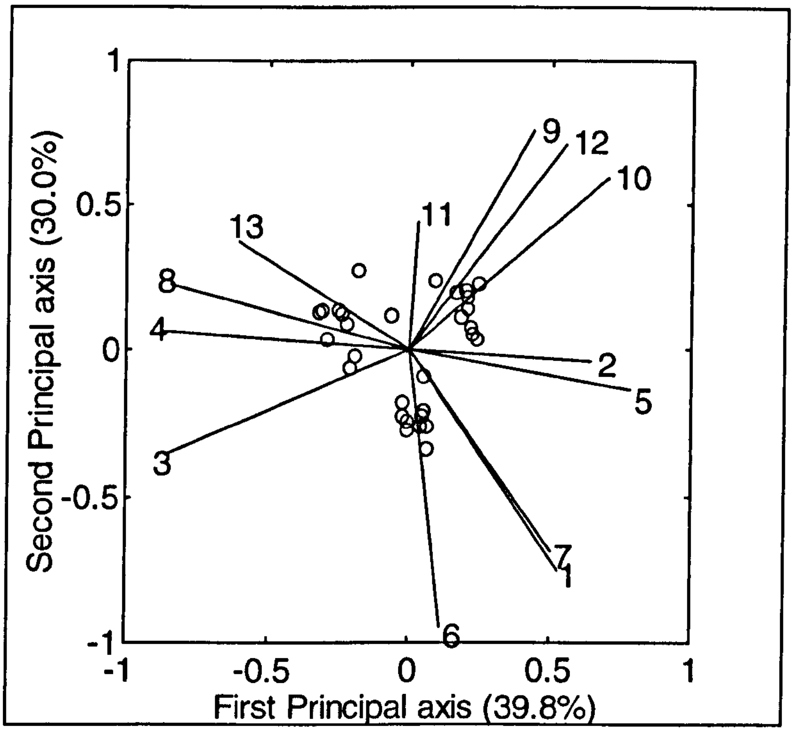

Figure 3.3 Correlation Biplot of Ceramic Pots

Data matrix C is represented with goodness of fit 69.8% and the goodness of fit of the variance-covariance matrix is 92.8%. Both these values are lower than for the covariance biplot but are still very high. As in the covariance biplot, there appear to be three groups of pots and similar pairs of variables to those in the covariance biplot appear to be highly correlated. We can see that all variables except 11 are fairly well represented in the plot (this is indicated by the closeness of the lengths of the lines to one, which indicates perfect representation) and that variable 6 is the best represented, having the longest vector. If we draw a unit circle on the correlation biplot then it becomes even more obvious which variables are best represented.

Applying the idea described by Gabriel (1995) and explained in 3.4.7, to variables 1 & 7, because these are both well represented on the plot (see Table 3.1), we can test for any significance of association between them. The value of their inner product is 0.785 and comparing this with the 5% critical point of the standard normal distribution divided by the square root of 30 ($= \frac{1.96}{5.477} = 0.358$), we conclude that the variables are significantly associated with each other. Repeating the test for variables 1 & 3, which we would interpret as being uncorrelated by 3.4.6.3, we obtain an inner product with absolute value of 0.178. Comparing this with the same percentage point of the normal distribution leads to the conclusion that these variables are not associated with each other and are measuring different aspects of the data.

3.7.1.3 The Coefficient of Variation Biplot

Using a coefficient of variation biplot to display the data (see 3.4.3) and representing the results in two dimensions produces Figure 3.4. The matrix E is represented with goodness of fit 81.8%, whereas the goodness of fit of the variance-covariance matrix is 94.7% in the first two dimensions and so once again, these values are extremely high.

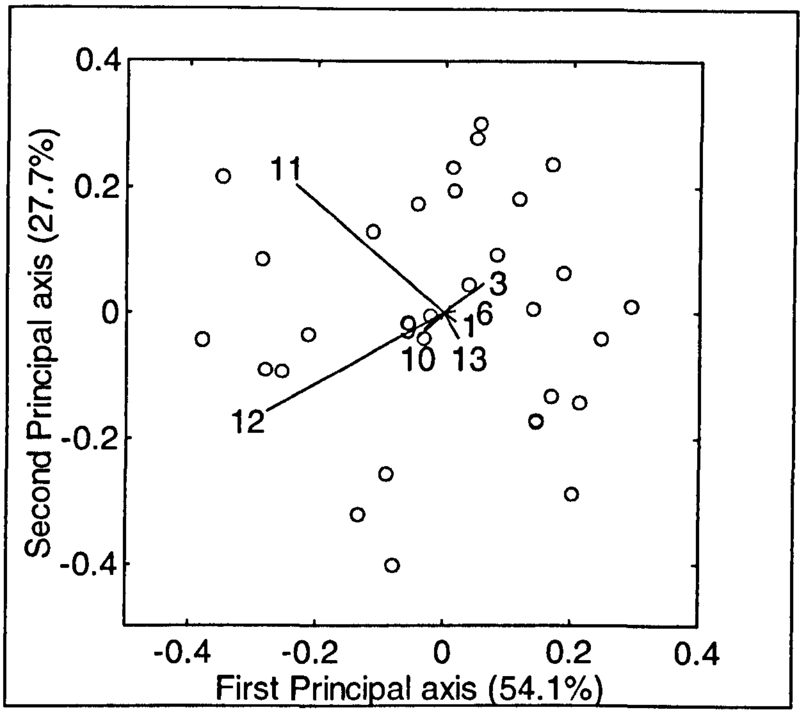

Figure 3.4 Coefficient of Variation Biplot of Ceramic Pots

Variables 11 (internal depth of footring) and 12 (thickness of wall) dominate the plot and so we know, from 3.4.3, that these have larger coefficients of variation than the remaining variables. Some variables have such small coefficients of variation that they are difficult to see on the plot — namely 2, 4, 5, 7 and 8 and because of this it is difficult to assess which pairs of variables are highly correlated and which are not. However, in this biplot there no longer appear to be three groups of pots as in the previous two biplots (even after plotting the pots without the variables to reduce the scale). We suggest that one reason for this could be that because variables 11 and 12 are so dominant, it may be that they do not contain much grouping information — i.e. it is not these variables which differentiate between pot groups and they may even hinder group separation. Also, given that some variables are very poorly represented (see Table 3.1), it may be worth excluding these from the analysis. Alternatively, it may be that this form of biplot is not suitable for identifying groups of observations, perhaps because of the data themselves. The idea of variable selection and assigning importance to variables in their ability to discriminate between groups of observations is discussed in Chapter Seven.

3.7.1.4 The Spearman Rank Correlation Biplot

Figure 3.5 illustrates the Spearman rank correlation biplot applied to the same data (see 3.4.4).

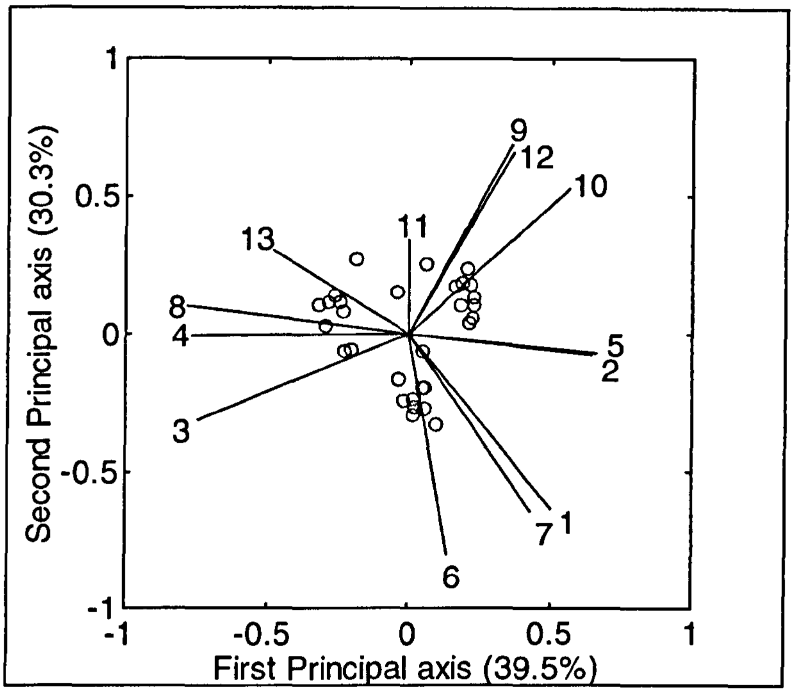

Figure 3.5 Spearman Rank Correlation Biplot of Ceramic Pots

The matrix Z is represented with goodness of fit 69.5%, which is equal to that of the correlation biplot and the goodness of fit of the variance-covariance matrix is 93.1% in the first two dimensions. As in the covariance and correlation biplots, there appear to be three groups of pots and we see that variable 11 is poorly displayed because the vector representing it is short. As for the correlation biplot, vectors with lengths close to one indicate near perfect representation of the corresponding variables.

3.7.1.5 The Principal Component Biplot

The principal component biplot of Section 3.5 is illustrated in Figure 3.6, where the matrix Y of the ceramic pot data is represented with goodness of fit 69.8%.

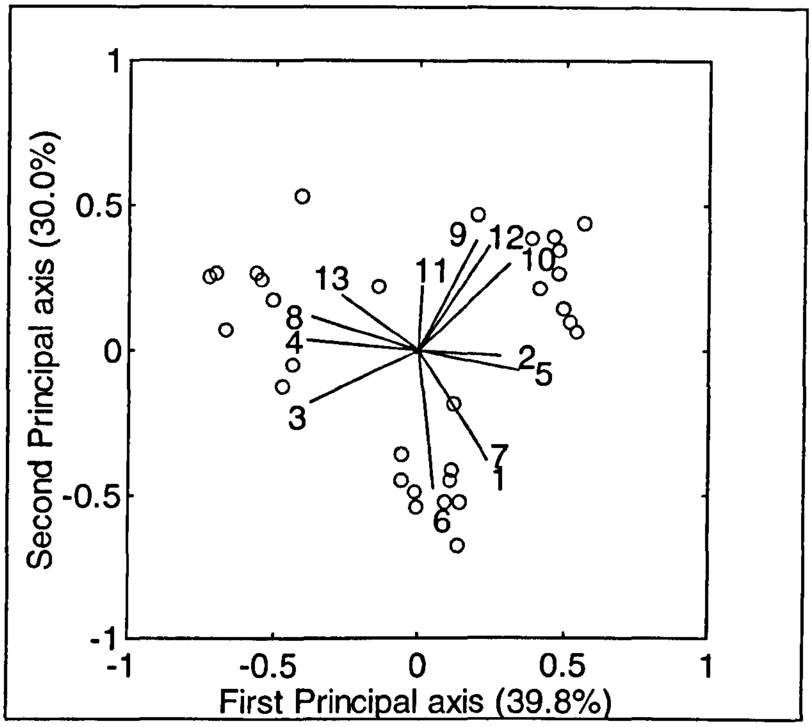

Figure 3.6 Principal Component Biplot of Ceramic Pots

Again, the plot suggests three groups of pots and is very similar to the correlation and Spearman rank correlation biplots in terms of which pairs of variables appear to be highly correlated.

3.7.1.6 Summary and Comparisons

The covariance, correlation, Spearman rank correlation and principal component biplots separate the pots into three groups and these groupings are discussed further in 3.7.3.1. The coefficient of variation biplot does not reveal these pot groupings. We suggested in 3.7.1.3 that one explanation for this lack of grouping is that variables 11 and 12 dominate the plot and are perhaps obscuring group structure which might be revealed if they were not present (these variables dominate the plot because they have large coefficients of variation compared with the other variables). In Chapter Seven we examine, adapt and improve existing methods of variable selection, but removing these two variables makes the pot groups become only slightly more distinct. We therefore believe that the lack of grouping must either be due to the scaling involved in obtaining a coefficient of variation biplot, or the data set itself. Looking in more detail at this form of biplot, it appears that when variable means are subtracted and the variables are then divided by their means, the difference in each measurement between observations tends to be reduced, which in turn could result in group structure being less likely to be revealed. Thus, if we use a biplot because we want to display groups of observations, then the coefficient of variation biplot may not be suitable unless, perhaps, there are very large differences between observations from different groups.

All five biplots explain approximately 70% or more of the variation in the data and the positions of the variables in the various biplots are almost identical, except for the coefficient of variation biplot. From what we have seen in Section 3.7.1 and using our a priori knowledge of the data, we believe that the correlation and principal component biplots are always likely to be the most useful because they do not require variables to be in the same units (unlike the covariance biplot). In addition, they make use of all the data (unlike the Spearman rank correlation biplot) and they are able to separate out groups of observations (unlike the coefficient of variation biplot). The choice between these two may therefore come down to the percentage of variation explained and the quality of representation of the variables. However, it is of course hazardous to make generalisations on the basis of one data set and we now consider a further example.

3.7.2 Application to Simpson Desert Flint Tools

Several types of biplot are illustrated on the Simpson Desert flint tool data (1.2.8) which consist of six measurements on 52 flint tools from site 08 and 26 from site 09. In the analysis the data are treated as 78 tools i.e. sites are not distinguished, but tools are labelled according to site in the resulting plots. Site 08 is considered to be of landform 'escarpment', whereas site 09 is described as 'plain with drainage'. The archaeological aim is to identify whether there is any distinction between tools measured at the two landforms on the basis of the available measurements and therefore whether the sites were used for different activities in the past (Barton, *pers.*

comm.).

Five of the variables are measured in millimetres and one, weight, in grams, so we did not consider the covariance biplot to be appropriate (see 3.4.5.6). However, the remaining four biplots which were described in Sections 3.4 and 3.5 are illustrated and the results compared and summarised in 3.7.2.5. For biplots in the Correlation Biplot Family each biplot has a quality of representation of inter-row distances of 33.3%, which seems reasonable (see 3.6.3). The qualities of representation of individual variables for the various biplots are given in Table 3.2. There is also interest in seeing whether weight (the most expensive variable to obtain) is really necessary — if it were dropped then the covariance biplot would be available.

Table 3.2 Quality of Representation of Individual Variables (%) for the Simpson Desert Flint Tools

	Biplot			
Variable	**Correlation**	**Coefficient of Variation**	**Spearman Rank Correlation**	**Principal Component**
Length	87.9	69.7	93.5	60.3
Width	85.0	75.5	80.7	28.8
Thickness	79.3	77.3	81.6	23.0
Platform Width	83.4	94.4	85.1	35.1
Platform Thickness	69.4	75.7	58.3	20.1
Weight	89.0	76.4	96.9	32.7

The qualities of representation of individual variables in the principal component biplot are lower than for the other biplots and all variables are adequately represented in the correlation, coefficient of variation and Spearman rank correlation biplots. The various biplots are illustrated below where tools from site 08 are represented by circles and tools from site 09 by crosses.

3.7.2.1 The Correlation Biplot

The correlation biplot was described in 3.4.2 and is illustrated in Figure 3.7.

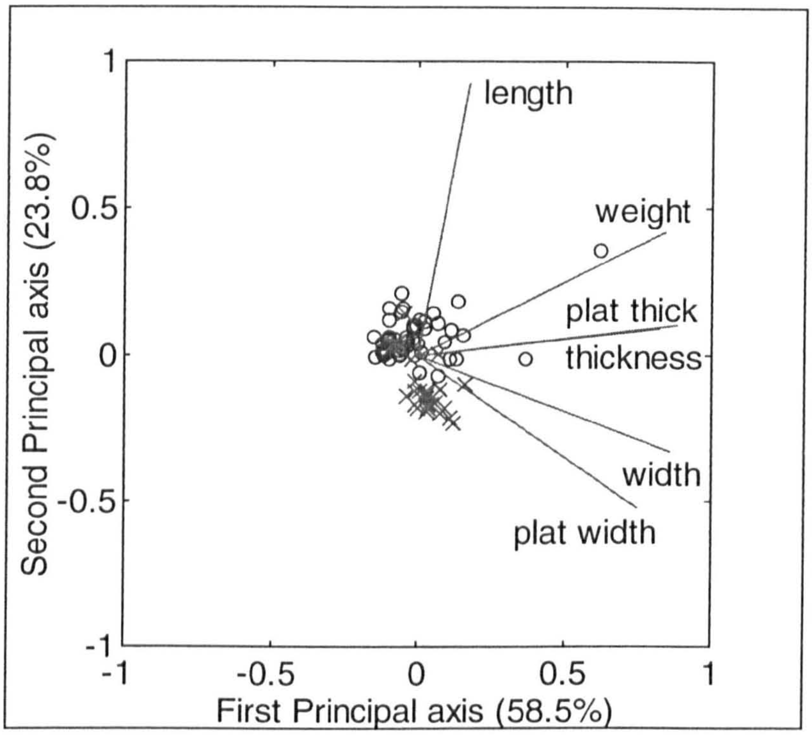

Figure 3.7 Correlation Biplot of Simpson Desert Flint Tools

Data matrix C is represented with goodness of fit 82.3% which is high enough for us to be confident in our interpretations of the display and we see from the lengths of the variables, as well as from Table 3.2, that the variables are all well represented. Clearly, platform thickness and thickness are very highly correlated, judging by the small angle between them on the plot and the tools appear to divide into groups which correspond to the different landforms, although there is some overlap. There appear to be two outlying tools to the right of the picture which are associated mainly with variable weight and either or both of thickness and platform thickness. These tools are from site 08.

3.7.2.2 The Coefficient of Variation Biplot

Figure 3.8 illustrates the coefficient of variation biplot, which was described in 3.4.3.

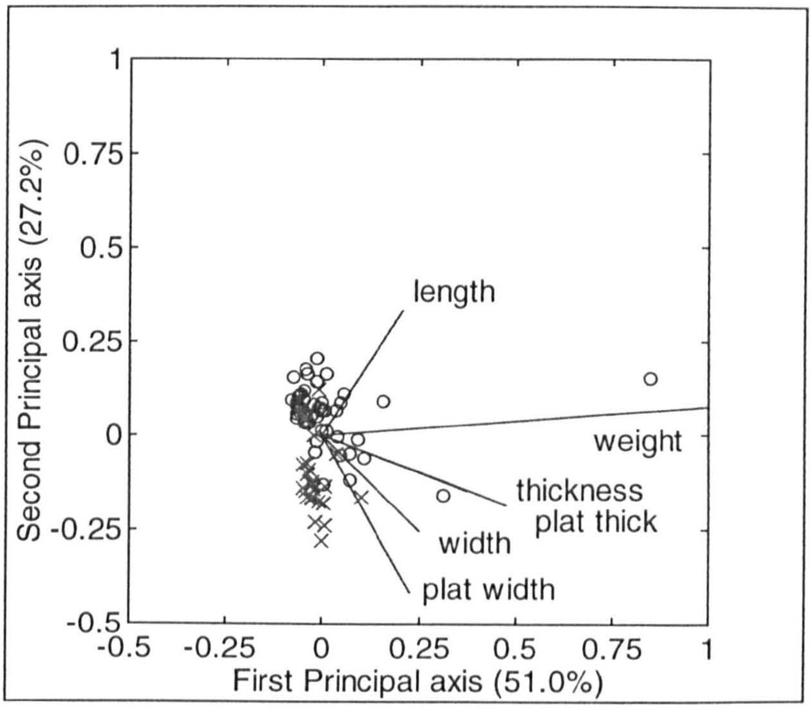

Figure 3.8 Coefficient of Variation Biplot of Simpson Desert Flint Tools

The matrix E is represented with goodness of fit 78.2% and the coefficient of variation biplot produces a more distinct division of tools into two groups than was seen in the correlation biplot, with some tools from site 09 (crosses) appearing with those from site 08. There are again two outlying tools, which are located towards the right of the plot. In contrast to the coefficient of variation biplot on the ceramic pot data, it appears that the variables are able to distinguish between groups of observations in this type of biplot for the flint tool data. This could be because tools from the two sites differ by a relatively large amount on the basis of these measurements, or because particularly 'appropriate' variables for distinguishing between tools at different landforms have been measured.

3.7.2.3 The Spearman Rank Correlation Biplot

The Spearman rank correlation biplot of 3.4.4 is illustrated in Figure 3.9, where the data matrix Z is represented with goodness of fit 82.7% in two dimensions.

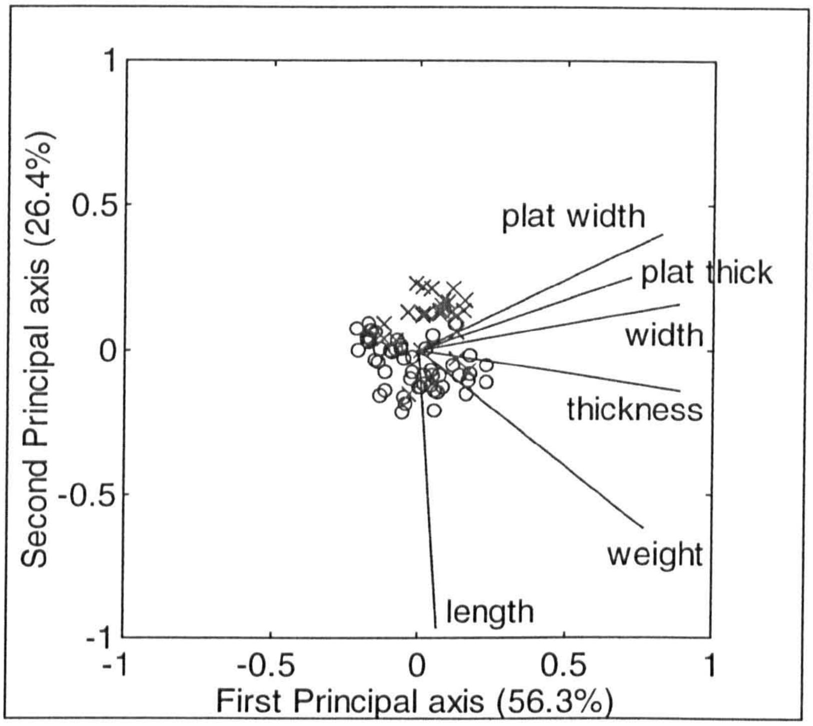

Figure 3.9 Spearman Rank Correlation Biplot of Simpson Desert Flint Tools

Figure 3.9 is more similar (apart from arbitrary reflection, see 5.2.4) to the correlation biplot than to the coefficient of variation biplot, both in terms of the division of tools into groups and in terms of highly correlated variables. We also notice that there are no outlying tools on this plot — tools with extreme measurements have been removed by ranking the observations within each variable. All six variables are well represented in the plot because their vectors have lengths close to one, although this is already evident from Table 3.2. Calculating the correlation coefficients on the raw data confirms the correlations observed in Figure 3.9.

3.7.2.4 The Principal Component Biplot

In Section 3.5 we introduced the principal component biplot and this is illustrated in Figure 3.10.

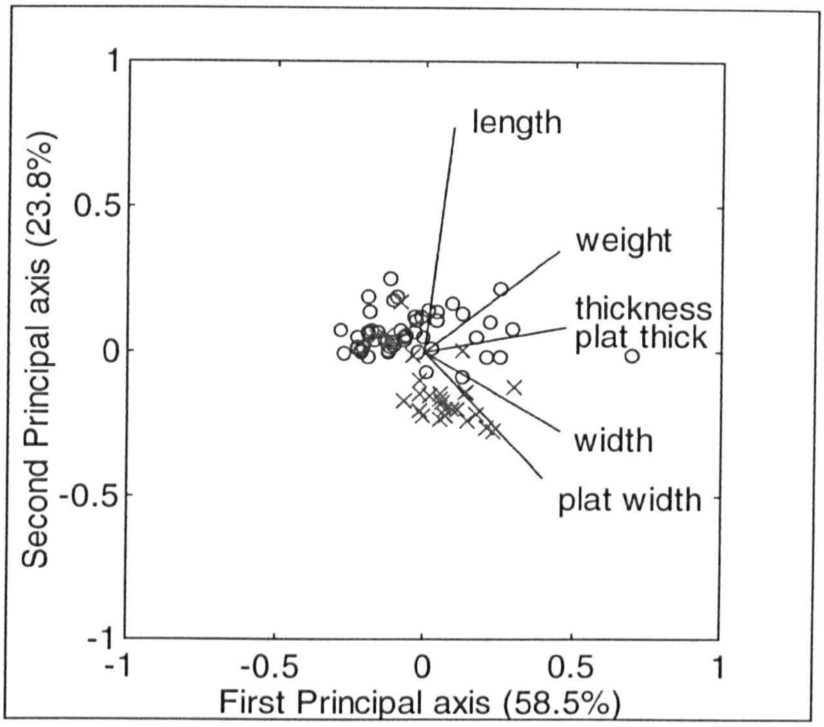

Figure 3.10 Principal Component Biplot of Simpson Desert Flint Tools

The principal component biplot is most similar to the coefficient of variation biplot, both in terms of grouping of similar tools and in terms of pairs of highly correlated variables. The majority of tools from the different sites are well separated, although a few from site 08 overlap with those from site 09.

3.7.2.5 Summary and Comparisons

Each biplot produces a separation of tools from landform ‘escarpment’ (site 08) from those from ‘plain with drainage’ (site 09). In both the correlation and coefficient of variation biplots there appear to be two ‘outlier’ tools from site 08, whereas in the principal component biplot there is only one. If we consider it necessary to identify and remove outliers then the Spearman rank correlation biplot is not appropriate. It may also be the case that these outliers have a large influence on the orientation of the vectors representing the variables and on the relationships between other observations.

This is something we consider in Chapter Eight, where we develop a method to detect influential observations. All biplots explain approximately 80% of the variation in the data, which is more than adequate for us to be confident in our interpretations of the displays.

3.7.3 Displaying Groups of Observations

One method of displaying particular groups of observations on a biplot is by using a concentration ellipse for the points of each group of interest. Gabriel (1981) comments that use of concentration ellipses is of particular importance when large sets of data need to be displayed (i.e. when there are more row markers than can be displayed effectively). We found this to be the case with the flake debitage data (1.2.6) that is described in 3.7.3.2, but we also propose using ellipses to identify similarities between groups of observations. Concentration ellipses are based on the multivariate normal distribution (Mardia *et al.*, 1979) and if $X \sim MN_m(\mu, \Sigma)$ then the equation given by:

$$(X-\mu)^T \Sigma^{-1} (X-\mu) = c^2$$

provides ellipses centred on the mean μ of constant density, where c is a constant. Mardia *et al.* (1979) explain that:

$$(X-\mu)^T \Sigma^{-1} (X-\mu) \sim \chi_m^{\ 2}. \qquad (3.9)$$

By analogy with confidence intervals for normal distributions, we suggest taking percentage points of the chi-squared distribution at 68% (~ one standard deviation from the centroid) and 95% (~ two standard deviations from the centroid), although applied to non-normal data, use of such ellipses effectively imposes some non-parametric smoothing on the data. Concentration ellipses are illustrated below for the co-ordinates obtained from the ceramic pot and Simpson Desert flake debitage data.

3.7.3.1 Application to Ceramic Pots

The ceramic pot data (1.2.5) used in Section 3.7.1 are actually a control group of ceramics. Impey approached a potter and showed him some kiln-site material. The potter was asked to look at, but not measure, these sherds and show them to three other potters. The three potters were then asked to make ten similar pots each and pots from each potter are labelled 1, 2 or 3 in the plots. With these pot data, we suggest using concentration ellipses in order to identify whether pots from the different groups are similar (rather than to summarise large amounts of data) and we show that overlapping ellipses are indicative of similar groups, whereas distinct ellipses are likely to result from dissimilar pot groups.

We propose displaying six ellipses, two for each group of pots (one at the 68% point of the chi-squared distribution and one at the 95% point), with different symbols representing each group of pots. This is done in Figure 3.11 for the covariance biplot, where the ellipses are meant to be used as an informal assessment of similarities between groups and not as a formal significance test.

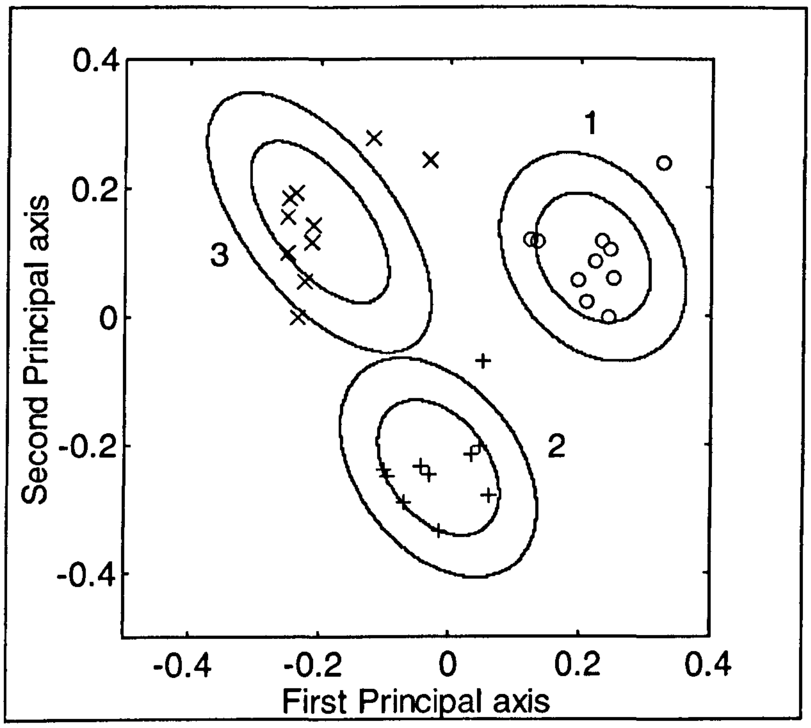

Figure 3.11 Concentration Ellipses for the Ceramic Pots (Covariance Biplot)

Figure 3.11 shows that the ellipses of the different groups are non overlapping and appears to confirm that there are three distinct groups of pots, corresponding to the three different potters. The inner ellipse of each group should contain 68% of the points, if equation (3.9) holds and the outer ellipse should contain 95% of the points. However, for each group there is at least one pot which does not fall within the inner ellipse. If we apply the same methodology to the coefficient of variation biplot where, in Figure 3.4, there was no evidence of three groups of pots, then we obtain Figure 3.12.

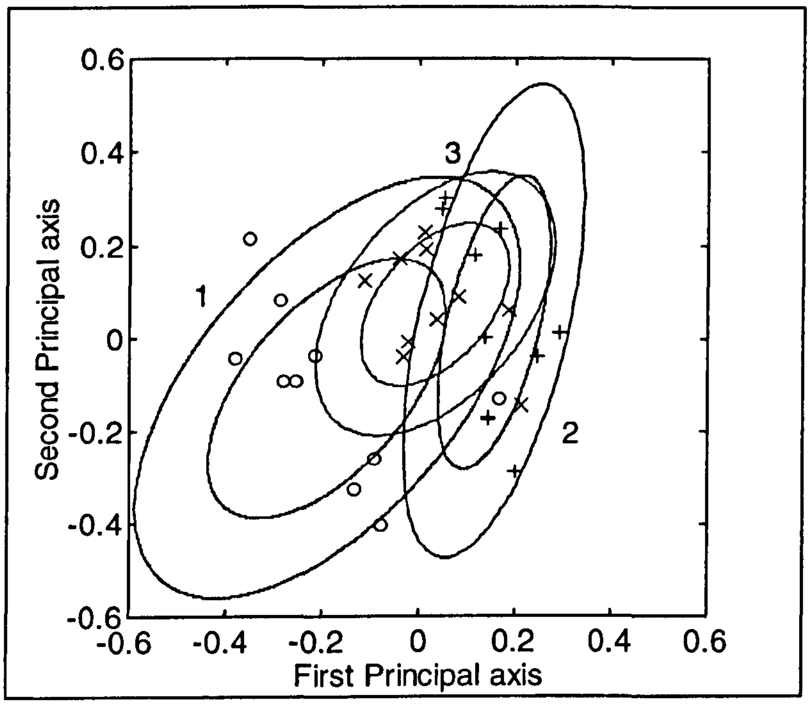

Figure 3.12 Concentration Ellipses for the Ceramic Pots (Coefficient of Variation Biplot)

We see from the figure that the inner ellipses of groups 1 and 2 touch but do not intersect, although both intersect with the inner ellipse of group 3 and there is considerable overlap between the outer ellipses of all three groups. Thus, if we display only ellipses and not pots, we still conclude that the pots cannot be separated into groups. We propose that as a rule of thumb, if the centroid of one ellipse lies within another ellipse, then the corresponding groups cannot be distinguished.

In conclusion, we can use ellipses as an informal method of identifying similarity between groups of observations (in this case pots). The covariance biplot enables us to distinguish between the pots produced by three potters (as do the correlation, Spearman rank correlation and principal component biplots which are not shown), whereas the coefficient of variation biplot does not. We favour either the correlation biplot or the principal component biplot because these account for variables in different units and display outlying observations. In archaeology, establishing groups of pots (or indeed other artefacts, as we will see with the flake debitage) has far reaching implications for reconstructing the past, for example the number of people which existed at a site, the extent of craft activities and division of labour.

3.7.3.2 Application to Simpson Desert Flake Debitage

In this section we consider the flake debitage (flint waste material, not flint tools, 1.2.6) which consist of six measurements taken on 2767 flakes from 28 sites (including 08 and 09 — see 3.7.2) across the Simpson Desert. Archaeological interest lies in establishing whether biplots are able to distinguish between groups of flakes, according to water permanency and land terrain at the sites where they were found (Barton, *pers. comm.*). However, it is extremely difficult to display all the flakes on one plot and to identify relationships between them and so we need to consider other methods of representing the data. We also develop methods for looking at the influence of large samples on observation groupings and on relationships between variables in Chapter Eight. Concentration ellipses, as advocated by Gabriel (1981), are used to summarise this large data set and their use is extended to assess group structure.

A correlation biplot was carried out on all the data but it was impossible to distinguish between sites of different terrain and water permanency by displaying individual points. A random sample of 10% of the data was taken and it was again impossible to distinguish between sites, because of the volume of data. We therefore summarise the data by displaying concentration ellipses using the 95% point of the chi-squared distribution, one for each type of water permanency, in Figure 3.13.

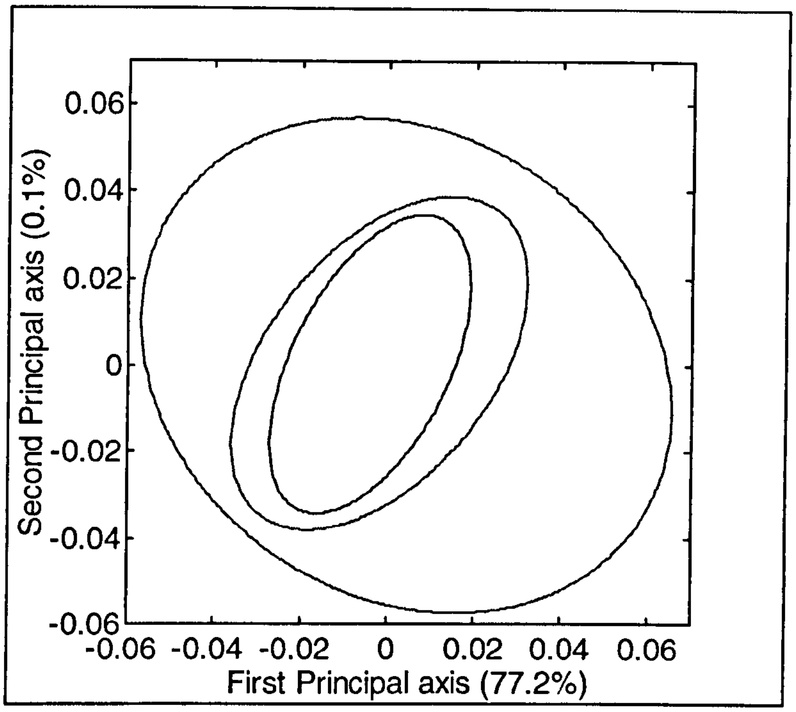

Figure 3.13 95% Concentration Ellipses for the Flake Debitage according to Water Permanency (Correlation Biplot)

The largest ellipse is that for ephemeral water sources and there are 1200 flakes from sites with this type of water permanency; the smallest ellipse is for flakes from sites with permanent water sources (878 flakes) and the middle one is for sites which had semi-permanent water sources (689 flakes). It is interesting to see that the ephemeral ellipse completely encompasses the other two ellipses and is approximately three times their sizes. However, the overlapping nature of the ellipses suggests that differences between flakes from different water sources either do not exist, or cannot be detected using the available data. This could be because the measured variables were not the most appropriate ones to identify differences, or because certain variables mask the effects of others. This raises the question of whether variable selection methods would be useful to identify the most important variables in distinguishing between groups of observations and this is discussed in Chapter Seven. Because the orientations of the ellipses in Figure 3.13 are different, it is evident that the flakes from the different water sources have a different correlation structure.

In Figure 3.14 we display five ellipses using the 95% point, one for each type of terrain. It is the dissected residual terrain that has the largest ellipse and this encompasses the ellipses of all the other terrains, which are all of similar size, but different correlation structure, to the dissected residual terrain. It therefore seems that we cannot separate out the flakes according to their terrain.

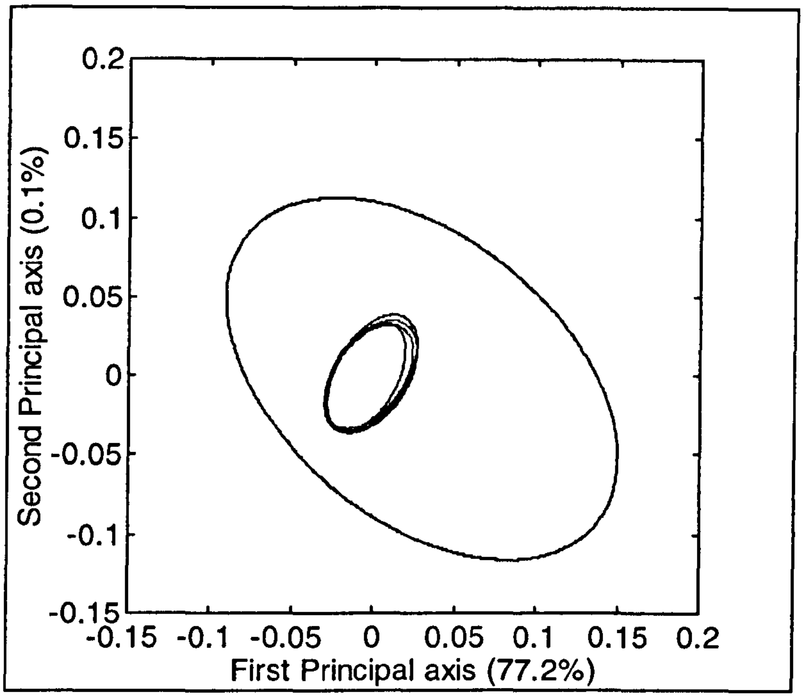

Figure 3.14 95% Concentration Ellipses for the Flake Debitage according to Land Terrain (Correlation Biplot)

The variable co-ordinates obtained from the correlation biplot are illustrated in Figure 3.15. We see from the figure that the relationships between the variables are similar to those in the correlation biplot of flint tools in Figure 3.7, but in Figure 3.15 the variables thickness and platform thickness are not so highly correlated. Platform thickness and weight also appear to be uncorrelated in this figure, in contrast to Figure 3.7. However, these differences are not surprising because flint tools are likely to be a different shape to waste flakes by definition of their purpose (Barton, *pers. comm.*).

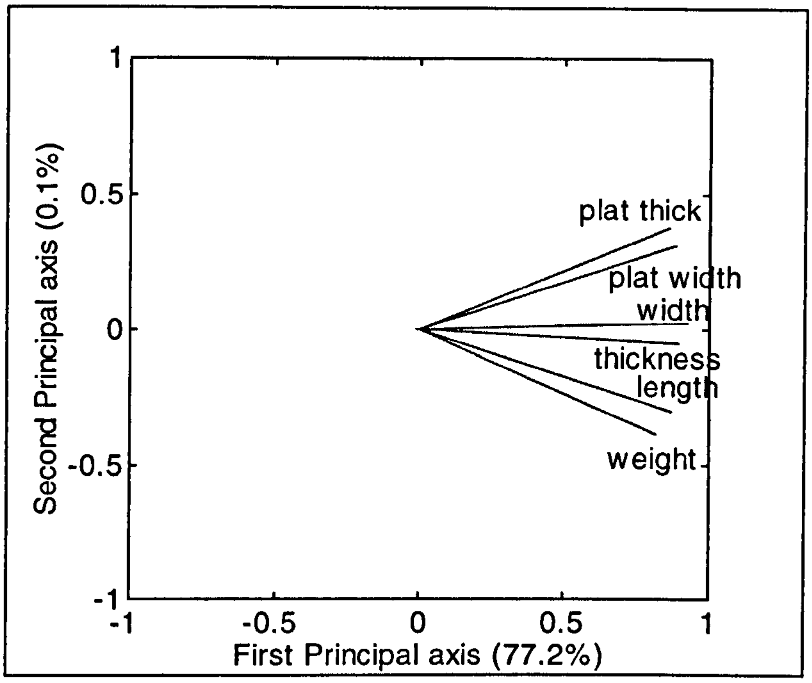

Figure 3.15 Correlation Biplot Variables for the Flake Debitage

Other forms of biplot were also obtained for these data (although for reasons already discussed we prefer the correlation or principal component biplot), but none were able to distinguish between groups of flakes. Concentration ellipses using the 95% point for the coefficient of variation biplot, for water permanency, are illustrated in Figure 3.16. The figure does show some areas where there is no ellipse overlap and is an improvement on the correlation biplot of Figure 3.13 in terms of separating out groups of flakes, but there are no clear groupings and all three ellipses are of a similar size.

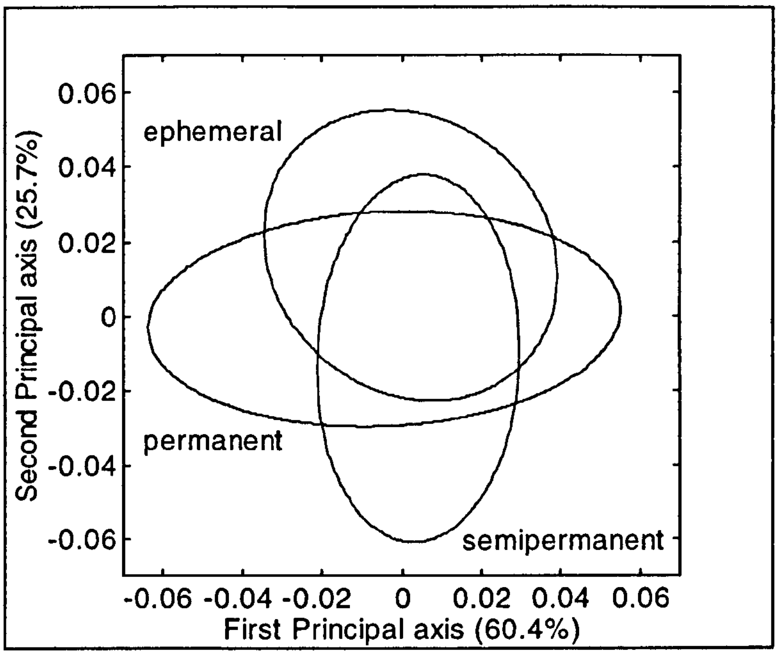

Figure 3.16 95% Concentration Ellipses for the Flake Debitage according to Water Permanency (Coefficient of Variation Biplot)

3.8 Diversity

The concepts and methods of diversity, as developed primarily in the field of ecology, have been increasingly represented in the archaeological literature over the last twenty years (Conkey, 1980; Grayson, 1984; Leonard & Jones, 1984; Rhode, 1988; McCartney & Glass, 1990; Ringrose, 1990; Kaufman, 1998), although the diversity biplot does not appear to have been introduced. In the sections that follow we discuss the role of diversity in archaeology and we also extend the role of the diversity biplot to cover this area.

Numerous measures of diversity have been developed in ecology, some of which have been adapted to archaeology. The most basic measure is the number of artefact types recovered from a site, but this does not allow for evenness considerations and is really an artefact type richness measure. It also ignores varying sample sizes, which can affect the number of artefact types sampled at a site and can therefore make comparisons between sites of dubious relevance. The next few sections briefly describe indices of richness, evenness and diversity and provide references where further details can be found.

3.8.1 Richness Indices

Bobrowsky & Ball (1989) describe many richness indices, which include those due to Margalef (1958), Odum *et al.* (1960) and Menhinick (1964). These three indices attempt to correct for sample size, because we would expect that as more individual artefacts are counted, the variety of types encountered increases. In practice we have found that there is very little difference between these indices.

3.8.2 Evenness Indices

Evenness describes the distribution of artefact abundances, or the relative frequencies of individual artefacts within each of the artefact types and is most often expressed as the ratio between an observed index of diversity and an expected maximum diversity, where all types are equally abundant. Pielou (1975, 1977) described two evenness indices:

- Ratio of Brillouin's Indices (originally introduced by Zar, 1974).
- Ratio of Shannon Indices.

3.8.3 Diversity Indices

Diversity indices attempt to combine richness and evenness measures into a single index and the main diversity indices are listed below. If X (n x m) is a data matrix of rank r, with columns corresponding to sites and rows corresponding to artefact types observed at those sites, then let P (n x m) be a matrix with entries p_{ij}, where $p_{ij} = \frac{x_{ij}}{x_{.j}}$ is the proportion of artefacts of type i at site j (where $x_{.j}$ is the sum of entries in column j) and let n_i be the number of individual artefacts counted of type i. For a particular site, N is the total number of individual artefacts and S is the number of artefact types found at the site.

3.8.3.1 The Simpson Index

Simpson (1949) defined the probability of any two individual artefacts drawn at random from an infinitely large site belonging to the same type, as:

$$E = \sum_{i=1}^{s} p_i^2$$

where $p_i = \frac{n_i}{N}$.

However, in order to calculate the index, the form appropriate to a finite site is used:

$$E = \sum_{i=1}^{s} \frac{n_i(n_i - 1)}{N(N-1)}.$$

It is evident that as E increases, diversity decreases and so the index is usually expressed as 1-E (Greenberg, 1956; Pielou, 1969) or 1/E (Williams, 1964; Whittaker, 1972; Hill, 1973). The index is also heavily weighted towards the most abundant type in the sample.

3.8.3.2 Squared Euclidean Distance

The Squared Euclidean distance between two sites (k and l), given by:

$$d_{kl}^2 = \sum_{i=1}^{s} (p_{ik} - p_{il})^2 \tag{3.10}$$

is a measure of dissimilarity between sites and may be regarded as a measure of 'beta' diversity.

3.8.3.3 Other Diversity Indices

Other diversity indices, which are not explained here, include:

- The Shannon Index.
- The Berger-Parker Index.
- Brillouin's Index (due to Margalef, 1958 and then Brillouin, 1962).
- The McIntosh Index (McIntosh, 1967).
- Hill's Family of Diversity Indices.

3.8.4 Diversity in Archaeology

In archaeology, an assemblage is defined to be a collection of artefacts and differences in assemblage diversity have been thought to represent, amongst other things, important differences in settlement function (e.g. a wider range of artefacts could indicate a more permanent settlement), social relations and subsistence patterns. So far, the emphasis in the literature has been on investigating the relationships between richness and sample size with a view to the fact that as more artefacts are collected, the number of artefact types within the collection increases. Rhode (1988) comments that this issue is critical for comparative studies between sites because, if the diversity measures vary as a function of assemblage size, then they may be telling us more about collection strategy, or rate of deposition, than about differences in past human behaviour.

An assemblage's diversity must not be described purely in terms of its diversity index because a site with a few, evenly represented artefact types can have the same diversity index as one with many, unevenly represented types. These two components of diversity — type richness (the variety of types) and type evenness (the relative abundance of types) — must be studied separately. To help us understand diversity, we can think of 'alpha' diversity as within-assemblage diversity and 'beta' diversity as between-assemblage diversity. Beta diversity is a measure of how different (or similar) a range of sites are in terms of the variety (and sometimes the abundances) of types found at them; the fewer types that are shared by two sites, the higher the beta-diversity will be. Both of these measures can be represented on a type of principal component biplot known as the diversity biplot, introduced into ecology by ter Braak (1983). We extend its use to archaeology in Section 3.8.9.

3.8.5 Application of Richness, Evenness and Diversity Indices to Bone Engravings

In this section we use the data described in Section 1.2.7, which consist of counts of 44 designs on bones from five sites in Spain. Conkey (1980) attempted to distinguish between aggregation and dispersion sites for the Early Magdalenian occupation of Cantabrian Spain and working mainly with the design elements on engraved bone artefacts, argued that aggregation sites should exhibit a greater diversity of elements than would be found in dispersion sites, because bands of hunter-gatherers would congregate at these sites.

3.8.5.1 Richness Indices

We calculated the three richness indices mentioned in 3.8.1 for the bone engraving data. Two of them place Altamira as being the richest in design classes, followed by Cueto de la Mina, El Cierro, El Juyo and then La Paloma; the third index has the same ordering except that Altamira and Cueto de la Mina change places.

3.8.5.2 Evenness Indices

The two evenness indices referred to in 3.8.2 were also calculated for the bone engraving data. Clearly, these indices are measuring different aspects of the data because Brillouin's evenness index is lowest for La Paloma, yet the Shannon index is highest for this site. This leaves us with the obvious problem of which, if either, to choose.

3.8.5.3 Diversity Indices

The diversity indices of 3.8.3 attempt to combine richness and evenness into one single index and we have evaluated all those mentioned for the bone designs. For all the indices, Altamira has the highest diversity, whereas Cueto de la Mina has the second highest diversity for all except the McIntosh Index, although this index does produce similar values for all sites. Beta diversity (equation (3.10)) is measured by squared Euclidean distances between sites and is evaluated for the five sites in Table 3.3.

Table 3.3 Squared Euclidean Distances between Bone Engraving Sites

	Altamira	El Cierro	El Juyo	Cueto de la Mina	La Paloma
Altamira	0.000	0.050	0.069	0.049	0.074
El Cierro	0.050	0.000	0.034	0.064	0.067
El Juyo	0.069	0.034	0.000	0.074	0.096
Cueto de la Mina	0.049	0.064	0.074	0.000	0.051
La Paloma	0.074	0.067	0.096	0.051	0.000

We see that the lowest value in the table is between El Juyo and El Cierro, which suggests that these sites are the most similar in terms of artefact types; we also see that La Paloma and El Juyo are the most different. Conkey (1980) used the Shannon Index of diversity and concluded that Altamira was high in diversity while El Cierro, El Juyo and La Paloma were low in diversity. Conkey also thought that Cueto de la Mina was intermediary between Altamira and the other sites and was therefore also thought to represent an aggregation site. The richness, evenness and diversity indices suggest that Altamira and Cueto de la Mina are the most diverse sites.

3.8.6 The Jack-knife Technique and Diversity

Kaufman (1998) has recently developed a method of testing for differences between diversity measures. When we obtain diversity, richness or evenness indices we usually comment on which sites have larger or smaller values than which other sites, but we don't know if these differences are 'significant'. Two approaches currently used to assess differences are simulation (Kintigh, 1984, 1989) and regression (Grayson, 1984) methods, but these can give contradictory results when applied to the same data (Kaufman, 1998). Kaufman has therefore developed an approach based on the jack-knife technique, which we believe is a valuable contribution to the literature. Its main advantages are that it does not assume a theoretical sampling model and no a priori assumptions regarding an underlying distribution are required. It is not illustrated here but we believe that it has a sound statistical basis.

3.8.7 The Diversity Biplot and its Interpretation

Using the notation of 3.8.3, each of the m sites can be represented by a vector p_j in r-dimensional Euclidean space. For each site, the proportion of artefacts of a particular type is equal to the length of the orthogonal projection of the site onto the axis that represents that type. The length of each site vector is p_j, which is the distance between the site vector and the origin and is denoted $\|p_j\|$. The squared length of each site vector p_j is given by:

$$\|p_j\|^2 = \sum_{i=1}^{n} p_{ij}^2,$$

which is equivalent to the formula for the Simpson Index discussed in 3.8.3.1.

Co-ordinates of the row points (types) and column points (sites) are obtained by applying a singular value decomposition to the matrix of proportions P. A non-centred principal component biplot of proportion data gives site ordinations that display approximate alpha diversities of sites and beta diversities of groups of sites, as measured by the Simpson Index (see 3.8.3.1) and squared Euclidean distance (see 3.8.3.2) respectively (ter Braak, 1983). However, type-centring of the matrix of proportions allows a better approximation to beta diversities and alpha diversities can still be visualised if the true origin is projected onto the plane of ordination.

If both artefact types and sites are displayed as vectors from the origin to the points representing them, then type vectors pointing in roughly the same direction as a site vector are present in high proportions at that site. A site with low diversity will have only a few long type vectors pointing in its direction, whereas a site with high diversity will have several, shorter type vectors that point in its direction. The biplot thus displays which types make a site as diverse as it is. The type-centred diversity biplot is explained in the next section.

3.8.8 The Type-Centred Diversity Biplot

In type-centred diversity biplots the data matrix of proportions, P, is row-centred, so that the entries are now p_{ij} - $p_{i.}$, where $p_{i.} = \frac{1}{m}\sum_{j=1}^{m} p_{ij}$. This means that the origin of the co-ordinate system is translated to the centroid of the sites and the proportions of a type at each site are approximated as deviations from the mean proportion of the types at the site. Distances between sites are not affected by this translation of origin, but lengths of the vectors representing them are, hence Euclidean distances are displayed more accurately than before, but the Simpson Index values are not. For the Simpson Index values we need to know the position of the true origin and this is determined by projecting the true origin (in full-dimensional space) onto the plane of the biplot and calculating the distance from the true origin to its projection, which has co-ordinate z_r on the r-th principal component. The squared distance from the origin to z in a two-dimensional biplot is equal to:

$$t^2 = \left\{\sum_{i=1}^{n} p_{i.}^2\right\} - \left(z_1^2 + z_2^2\right).$$

The Simpson Index of a site in the biplot is approximated by t^2 plus the squared distance between the site and the projection of the origin. The order of the Simpson Indices of the sites can be seen by looking at distances from z, where high-diversity sites will be near to z. We apply the type-centred diversity biplot to the bone engravings in section 3.8.9 below.

3.8.9 Application of the Diversity Biplot to Bone Engravings

Diversity indices for the bone engravings have been discussed in several papers (Kaufman, 1998; Rhode, 1988; Kintigh, 1984, 1989), but a diversity biplot has never been introduced and the present author is unable to find any diversity biplots in the literature which relate to archaeology. Figure 3.17 illustrates a design-centred diversity biplot, but because there are so many designs to display we represent these by plusses and the sites by lines. The five sites are labelled (ALTAMIRA=Altamira; CUETO=Cueto de la Mina; EL JUYO=El Juyo; CIER=El Cierro; PALOMA=La Paloma) and the position of the true origin is indicated by an asterisk (*), which happens to be located at the present origin.

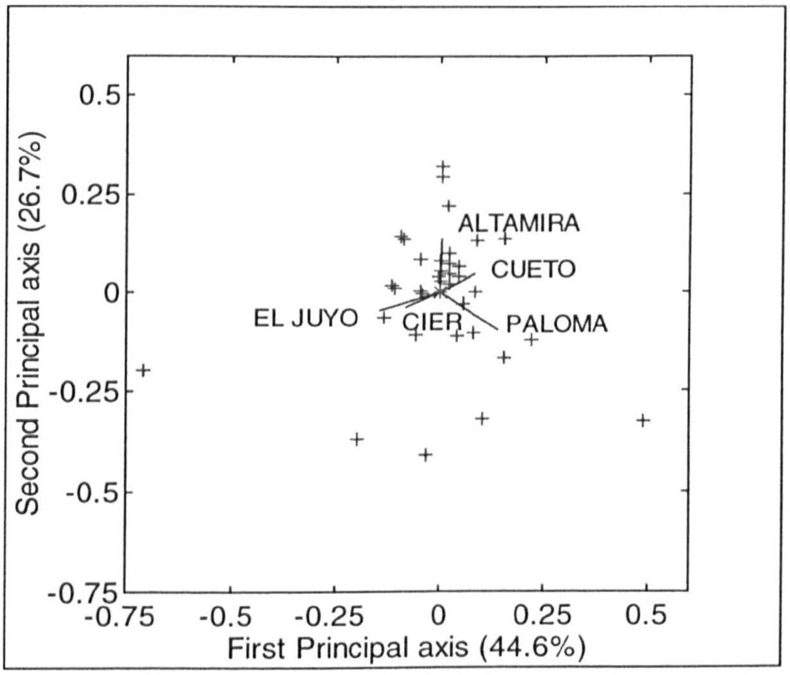

Figure 3.17 Design-Centred Diversity Biplot of Bone Engravings

The percentage of variation accounted for in these first two dimensions is over 70%, which is more than adequate for us to be confident in our interpretations of the display. Design 8 is the point located on the extreme left of the diagram and projection of the sites onto a vector in this direction indicate that it is present in high proportions at El Juyo and El Cierro, but in very low proportions at Altamira and Cueto de la

Mina. Beta diversity is lowest between El Juyo and El Cierro because these have the smallest Euclidean distance of any pair of sites and thus they are most similar (the angle between them is also small suggesting that they are highly correlated in some sense). The biplot in Figure 3.17 agrees to a certain extent with Conkey's interpretation of diversity based on the Shannon Index, because we see that there are more designs in the direction of Altamira compared with Cueto de la Mina and even less for the remaining sites. However, sites nearest to the asterisk have high diversity and this would suggest that El Cierro is most diverse, rather than Altamira.

To a certain degree we suggest that the diversity biplot 'corrects' for sample size problems, because it is applied to the data as proportions. We believe that just as there is probably a sample size above which richness does not increase for a given assemblage, there is also a sample size below which it is not sensible to compare richness across sites. Clearly, the sample size should be at least as big as the total number of categories identified after considering all relevant sites. In the case of the bone engravings, the sites of La Paloma and El Cierro have sample sizes of 23 and 35 respectively, which are both less than the number of design elements (44). We therefore have no hope of achieving the same richness at these sites as at the other three, which all have sample sizes greater than 44. Conkey (1980) briefly discusses this and concedes that:

> '... One must carve at least 44 design elements in order to achieve the maximum diversity of the Lower Magdalenian design element repertoire. Only two sites in this study yielded fewer instances of the use of design elements than the total (44) of different design elements ...'

Whilst there are only two sites with less than 44 designs, we must remember that this is 40% of the sites and so there is an argument for either obtaining more bone engravings at these sites or limiting our interest to the remaining three sites. We are not convinced that it is worthwhile to estimate the minimum sample size above which there will be little increase in assemblage richness, because the type of finds we are dealing with are generally limited in number and some information is better than none at all.

Looking at the raw data (Table A.11 in the Appendix) we see that there are only six design elements which are found at all five sites and there are 11 designs found only at Altamira. Only one other site — Cueto de la Mina — yielded design elements unique to it. There are six designs not present at Altamira, five of which are unique to Cueto de la Mina and so there is evidence to suggest that this site may also have been an aggregation locale (Conkey, 1980).

3.8.10 The Diversity Biplot and Correspondence Analysis: Bone Engravings

This section compares the interpretation of the diversity biplot with that of the correspondence analysis map for the bone engraving data. The diversity biplot is applied to data in the form of design-centred proportions for each site and it is therefore interesting to compare the interpretation of this biplot with that of a correspondence analysis map (described in Chapter Two), which is based on relative frequencies. Figure 3.18 illustrates the correspondence analysis map with designs represented as circles and sites as triangles. The site labelling is as follows: ALT=Altamira; CUE=Cueto de la Mina; JUY=El Juyo; CIE=El Cierro; PAL=La Paloma.

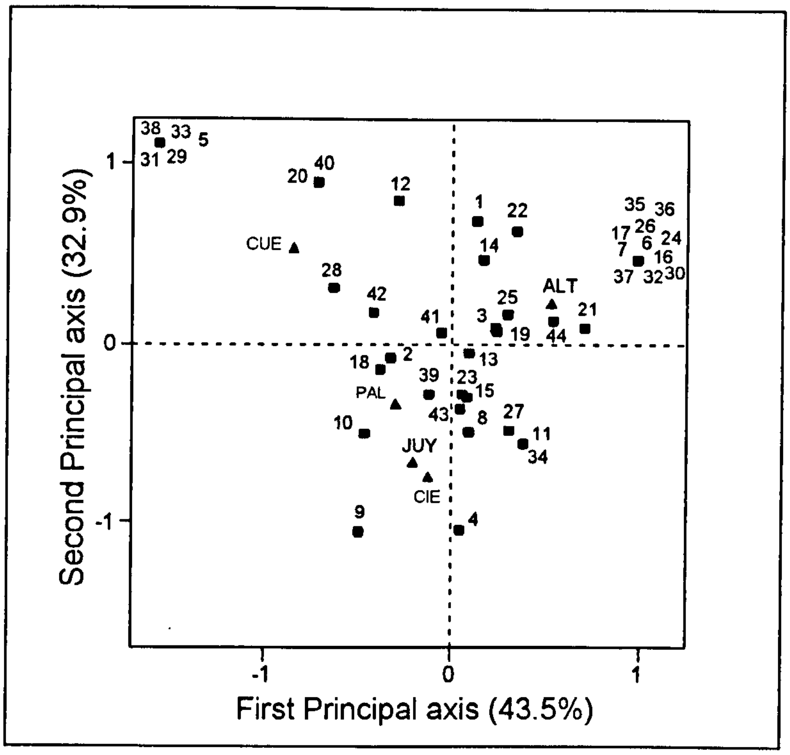

Figure 3.18 Correspondence Analysis Map of Bone Engravings

Interpreting this map we infer that Altamira is highly associated with many types of design, mainly because of the large cluster of designs in the top right-hand corner of the display; these are in fact the 11 designs unique to this site. It is also evident that in the top left-hand corner are five designs highly associated with Cueto de la Mina and these are the five designs mentioned in 3.8.9 that are unique to this site. In the correspondence analysis map abundances of designs at sites cannot be approximately recovered by projecting a design onto the line from the origin to the site point, as is the case for biplots. Also, in a symmetric correspondence analysis map the designs are placed at the centroid of the sites in which the design occurs and vice versa. Additionally, distances between sites are in terms of the chi-squared distance in the correspondence analysis plot, rather than in terms of Euclidean distance in biplots. However, relationships between designs and sites do seem to be more clearly displayed in the correspondence analysis map than in the diversity biplot.

3.9 The Symmetric Biplot

Having described and applied biplots of the Correlation and Principal Component Biplot Families, we briefly introduce another form of biplot that has fewer properties than these other biplots. The symmetric biplot is a combination of the row and column scalings of both the Correlation Biplot Family and the Principal Component Biplot Family. The mean of all elements of the data matrix, $\overline{x}$, is subtracted from each element of the matrix:

$$Y = X - \overline{x}$$

and because the overall mean of the matrix is subtracted, the variables must be measured in the same units. Considering the GSVD of a matrix X for the simplest case of $\Omega = I_n$ and $\Phi = I_m$, described in Chapter One, we have $a = \frac{1}{2}$ and $b = \frac{1}{2}$. The row and column co-ordinates in two dimensions are given by the first two columns of the matrices F and G respectively, where:

$$F = UD_{\mu}^{\frac{1}{2}}$$

$$G = VD_{\mu}^{\frac{1}{2}}.$$

3.10 Relationships with other Techniques

Biplots are known to have close similarities with both principal component analysis and correspondence analysis and these similarities are described below.

3.10.1 Biplots and Principal Component Analysis

Both biplots and principal component analysis usually operate on data where rows represent observations and columns represent variables. There is a simple algebraic transformation from one technique to the other, which was explained in Baxter (1994), although the interpretations of the two representations are different.

To obtain either a correlation biplot or a principal component biplot from a principal component analysis, it is necessary to carry out the following steps:

Step 1: Start with a data matrix X (n x m), where rows are observations and columns are variables.

Step 2: Let L be matrix X, column-standardised so that each column has zero mean and unit variance.

Step 3: Carry out a PCA on matrix L. Obtain principal component scores and coefficients.

Step 4: Carry out a singular value decomposition on matrix L, so that:

$$L = U\Lambda V^T$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ is a diagonal matrix of singular values;

U is the eigenvector of L^TL;

V is the eigenvector of LL^T.

The principal component scores and coefficients obtained in step 3 are exactly the same as the row co-ordinates and the column co-ordinates respectively in the principal component biplot. The correlation biplot can be obtained from the principal component biplot (with column standardisation) by dividing the column co-ordinates by λ_i and multiplying the row points by λ_i, where λ_i is the singular value from the singular value decomposition of L.

Whereas in PCA only the total variance (the sum of the column variances) is decomposed along the principal axes, in the covariance biplot the individual variable variances are also displayed (represented by vectors). The relative variances are also displayed in the correlation and Spearman rank correlation biplots and these can take a maximum value of one.

3.10.2 Biplots and Correspondence Analysis

Correspondence analysis was originally developed for data in the form of a contingency table, although it is often extended to frequency and categorical data. In contrast, biplots are more appropriate for data matrices of continuous data, where rows represent observations and columns represent variables. However, both biplots and correspondence analysis are ways of interpreting a joint map of row and column points and the main differences between the techniques are listed in the following sections. Table A.1 of Appendix A in Greenacre (1984) explains the relationship between various multivariate techniques in terms of the singular value decomposition.

3.10.2.1 Interpreting the Displays

The interpretation of a biplot is in terms of row-to-column scalar products with respect to the origin — biplots are designed to recover, approximately, the individual elements of the data matrix in these scalar products. In contrast, the correspondence analysis map is interpreted in terms of interpoint distances. Thus, row-to-column scalar products can be interpreted in a biplot but row-to-column distances cannot be interpreted in symmetric correspondence analysis because the rows and columns are in different low-dimensional spaces (Greenacre, 1993a).

3.10.2.2 Types of Data Matrices

Two-dimensional correspondence analysis plots cannot, in general, be compared in any direct way with biplots because the data matrices are of different types. However, both biplots and correspondence analysis determine the plotting positions for rows and columns of a data matrix X (n × m) from the singular value decomposition (SVD) of a matrix. For biplots the SVD is calculated for X (after scaling), whereas in correspondence analysis the SVD is found for a matrix of residuals after subtracting expected values, assuming independence of rows and columns (i.e. from $\frac{X}{x_{..}}$, where $x_{..}$ is the sum of all the entries of X; Jolliffe, 1986).

3.10.2.3 Approximation of Data Matrices

The steps of approximation and factorisation of biplots are reversible and the possibility of reproducing the data, at least approximately, from the display is a unique feature of biplots. However, in correspondence analysis we start with a matrix X, calculate a function of the matrix and then produce a map of these distances of correlations by the chi-squared metric. We cannot even approximately retrace the step from the map of the distances or correlations to the original data because the functions that have been used to summarise the data are generally not one-to-one functions.

3.11 Summary and Conclusions

Both biplots and principal component analysis are used to display data that consist of a series of variables measured on each of a number of observations. However, whereas PCA displays only observations, biplots display both observations and variables simultaneously and this is their strength. There are many types of biplot, each of which is useful in a different situation, depending on the aims of the analysis and the form that the data take (e.g. whether the variables are in the same units; whether there are any outlying observations etc.). This chapter has collated the most common forms of biplot together (from the fragmentary literature) and illustrated each of them on both new and published data. We have also explained in detail the goodness of fit measures that help us to assess whether the display in our chosen dimensionality, typically two, is adequate and whether individual variables are well represented. (The quality of representation of individual variables is discussed in depth in Chapter Eight, in relation to assessing the stability of these variables.) In addition, we have expanded the Spearman rank correlation biplot in order to enable us to assess the influence of tied, as compared with untied, observations. From our analyses we suggested that as a rule of thumb at least 50% of the variation in archaeological data should be explained in the first two dimensions of the ordination diagram.

Biplots are particularly useful in identifying groups of observations and in revealing which pairs of variables are highly correlated. However, because of the typically large numbers of variables that are measured in archaeology and because of the limited time and money available, we propose the introduction of variable selection methods to reduce the number of variables needed to reveal group structure. By using variable selection methods we can assign importance to variables in their ability to discriminate between groups of observations, although other factors such as ease of measurement may also come into consideration. In Chapter Seven we apply the methods of variable selection which exist for PCA to the various biplots and we also develop alternative methods.

It is known that concentration ellipses are useful if there are large numbers of observations to display (as is often the case in archaeology), because they can be used to summarise those from different 'groups' and avoid over-cluttering of the diagram. In this chapter we proposed extending their use to allow informal assessments of the similarity between groups of observations to be made, by looking for ellipse overlaps. By making an analogy with confidence intervals based on the normal distribution, we suggested taking percentage points of the chi-squared distribution at 68% (~ one standard deviation from the centroid) and 95% (~ two standard deviations from the centroid) when plotting the ellipses. We also suggested that if the centroid of an ellipse representing one group of observations lies within the ellipse representing another group, then the groups can be considered to be indistinguishable on the basis of the available measurements. We discuss the overlap of concentration ellipses in connection with correspondence analysis and canonical correspondence analysis in Chapters Five and Nine respectively.

Although the diversity biplot has been used in ecology, no references could be found to its use in archaeology and so we have introduced it into archaeology in this chapter. The diversity biplot is applicable to data which consist of the proportions of artefacts of different types observed at a number of sites — it allows us to assess visually which sites are particularly rich in which artefact types, rather than using one or more of the numerous diversity indices which we also discussed. In addition, we used the bone engraving data (1.2.7) to compare the diversity biplot with the symmetric correspondence analysis map — the interpretations of both diagrams proved to be very similar. We also explained the similarities between biplots and both PCA and correspondence analysis.

Finally, it is clear that outlying observations are revealed by their aberrant locations in the ordination diagram, but observations that have been influential in determining the display are not obvious and so in Chapter Eight we propose using the jack-knife technique to identify such observations. We also use a jack-knife approach to help us to detect which categories have been influential in determining the correspondence analysis (Chapter Six) and canonical correspondence analysis (Chapter Nine) displays.

Chapter Four

Canonical Correspondence Analysis

4.1 Introduction

A rather different development of statistical methodology (as compared with the previous two chapters) for examining the structure of certain types of data matrix has focused on exploring the relationship between species abundances and environmental variables that have been observed at various sites. The resulting technique is known as canonical correspondence analysis and is usually implemented by using the commercially available package CANOCO, although it is straightforward to implement in any standard programming language because it relies on the singular value decomposition and what is essentially a least squares method of regression.

The aim of this chapter is to provide a coherent account of the technique with algebraic details and a guide to use and interpretation, partly because it is much less widely used than other exploratory multivariate methods and partly to expand the framework for further work in Chapter Nine. Much of the development of the methodology has been driven by problems arising in ecology (specifically, community ecology) and it is helpful to describe the technique with close reference to this specific area of application. There is also considerable potential for the technique to be more widely used in archaeology and this is something we discuss later in the chapter.

Community ecology is the study of assemblages of plants and animals that live together and their interaction with environmental variables. Typically, data are collected on abundances of a multitude of species at a number of sites (a site is the basic sampling unit, separated in space or time from other sites e.g. a quadrat, or a trap) and sometimes environmental variables are also measured at these sites.

Before canonical correspondence analysis (CCA) was developed, the available statistical methods for analysing such data either assumed linear relationships between species abundances and environmental variables, or were restricted to regression analysis of the response of each species separately. To analyse the generally non-linear, non-monotone response of a community of species it was necessary to use ordination and cluster analysis — indirect methods that are generally less powerful than the direct method of regression analysis. Recently, regression and ordination have been integrated into techniques of (multivariate) direct gradient analysis, which are collectively called canonical ordination. One of these techniques, canonical correspondence analysis, developed by ter Braak (1986), escapes the assumption of linearity and is able to detect unimodal relationships between species and external (environmental) variables.

Ordination techniques such as correspondence analysis are commonly used to explain the variation in community composition by displaying points representing sites and points representing species in an ordination diagram. Subsequently, the diagram is interpreted with the help of external data, for example by calculating correlation coefficients between environmental variables and ordination axes, or by multiple regression of the ordination axes on the environmental variables. It is known that one difficulty with these methods is that the ordination axes are just particular orthogonal directions in the ordination diagram; other directions may well be better related to the environmental variables. Canonical ordination is a solution to this problem and with this method the regression model is inserted into the ordination model. As a result the ordination axes appear in order of the variance explained by linear combinations of the environmental variables.

This chapter describes the technique of CCA, beginning in Section 4.2 with a discussion of the various scales of measurement used when collecting vegetation data, followed by an explanation of the theoretical background and two approaches to implementing the method in 4.3. Section 4.4 describes the ordination diagram and associated quantities of interest, which aid its interpretation. Applications to published data on hunting spiders and dune meadow vegetation are discussed in 4.5 and 4.6 respectively and connections with other multivariate techniques are explained in 4.7. This chapter is concluded in Section 4.8.

4.2 Data Collection and Transformation

Data suitable for CCA consist of abundances of species (animal or, separately, vegetation) at a number of sites and of environmental variables measured at each site. The next section describes the most commonly used scales when collecting vegetation abundance data, one of which applies to the dune meadow vegetation of Section 4.6.

4.2.1 Scales of Measurement for Vegetation Data

Typically, the taxonomic unit employed in sampling within community ecology is the species. Species abundance measures include presence/absence, percentage cover, density (of number of individuals), frequency (percentage of quadrats having a species present), biomass (dry weight) or some weighted average of two or more such quantities. Abundance relates to the density of the individuals of a given species in a plot, whereas percentage cover is measured as the vertical projection of all aerial parts of plants of a given species as a percentage of the total plot area. Estimation of coverage is made by quick visual inspection and most scales have between five and ten ordinal values. Abundance and percentage cover are usually estimated together in a single 'combined estimation' or 'cover-abundance' scale and the Braun-Blanquet and Domin scales have been most commonly used (Gauch, 1982). The Braun-Blanquet scale is an ordinal scale that was extended by Barkman *et al.* (1964) to include subdivisions 2m, 2a and 2b and recoded to numeric values by van der Maarel (1979). Table 4.1 illustrates these scales, although for extensive surveys with very diverse communities it has been argued that the bulk of the information lies in qualitative differences i.e. in species presences and absences. We should note that, for example with Domin's scale, a value of 10 is not equal to twice a value of 5 in terms of cover-abundance; this is also the case for the other scales in the table and this is crucial when investigating the stability of the CCA map for the dune meadow vegetation in Chapter Nine.

Table 4.1 Cover-Abundance Scales for Vegetation

Braun - Blanquet		Barkman's refinement of Braun - Blanquet		van der Maarel	Domin	
Symbol	**Cover (%)**	**Symbol**	**Cover - abundance**	**Symbol**	**Symbol**	**Cover - abundance**
1	< 5	r	one or few individuals	1	+	one individual,
2	5 - 25	+	occasional and less than 5% of total	2		reduced vigour
			area		1	rare
3	25 - 50	1	abundant and with very low cover, or	3	2	sparse
			abundant but with higher cover; less		3	< 4%, frequent
			than 5% cover of total plot area		4	5 - 10%
4	50 -75	2m	very abundant	4	5	11 - 25%
5	> 75	2a	5 - 12.5% cover, irrespective of	5	6	26 - 33%
			number of individuals		7	34 - 50 %
		2b	12.5 - 25% cover, irrespective of	6	8	51 - 75%
			number of individuals		9	76 - 90%
		3	25 - 50% cover of total plot area,	7	10	91 - 100%
			irrespective of number of individuals			
		4	50 - 75% cover of total plot area,	8		
			irrespective of number of individuals			
		5	75 - 100% cover of total plot area,	9		
			irrespective of number of individuals			

4.2.2 Data Transformation

Ter Braak (1987a) is one of the prime references for CCA and he comments that species abundances are highly variable and nearly always show a skew distribution with respect to a quantitative environmental variable. He goes on to say that if the abundance of each species has a highly skew distribution, with many small values and a few large values, then the data can be square rooted or logged to down-weight high abundances. This is necessary because it is commonly believed that the abundance of a species tends to have a single-peaked response function to an environmental variable. This is because not only does it require a certain minimum amount of a resource, but also it cannot tolerate more than a certain maximum amount of a resource (this is essentially Shelford's Law of Tolerance (Shelford, 1911; Odum, 1971)). Each species' occurrence is thus confined to a limited range, known as its niche. Relationships between the species and quantitative environmental variables are therefore generally non-linear, but a unimodal curve may appear monotonic if only a limited range of the environmental variable is sampled. Ter Braak (1987a) believes that a good choice of environmental variable should minimise the number of species with more complex distributions than unimodal. However, if any of the environmental variables do follow a skewed distribution then they can be transformed to a symmetric distribution by taking logarithms, although any transformation of the species abundance data may influence the results of an analysis. In Section 4.5 we consider, for the hunting spiders, the effect of transformations of both species abundances and environmental variables on the interpretation of the ordination diagram because the consequences are potentially quite far-reaching. Chapter Nine discusses the effect of data transformations on the stability of the CCA map.

4.3 The Theory of Canonical Correspondence Analysis

Ter Braak (1987a) and ter Braak & Verdonschot (1995) are the main references for explaining the theory behind and the application of, the technique of CCA and the next few sections are heavily based on these. There are two methods of implementing CCA — an iterative algorithm and a singular value decomposition, both of which are explained below.

4.3.1 Background and Notation

In canonical correspondence analysis abundances of species are assumed to have bell-shaped (i.e. unimodal) response curves with respect to linear combinations of the environmental variables (which are known as environmental gradients). CCA also assumes a response model that is common to all species and the existence of a single set of underlying environmental gradients to which all species respond. When the response curves are not unimodal but approximately linear, the results can be expected to be adequate, but it is conventionally recommended in this situation to utilise instead the linear counterpart of CCA, known as redundancy analysis (van den Wollenberg, 1977). Redundancy analysis is a constrained form of multiple regression of the species' responses on the explanatory (environmental) variables (constrained so that the site scores are linear combinations of environmental variables), but there is no weighted averaging as in CCA. Instead, there is a two-way weighted summation.

CCA forms a linear combination of environmental variables that maximally separates the niches of the species. The first synthetic gradient is the first ordination axis, where the achieved maximum amount of niche separation is described by the eigenvalue of the ordination axis i.e. the eigenvalue is a measure of separation of the species' distributions along the ordination axis. Subsequent ordination axes are also linear combinations of the environmental variables that maximally separate the niches, but subject to the constraint that they are uncorrelated with the axis or axes extracted previously. In principle, as many ordination axes can be extracted as there are environmental variables.

We can define CCA in terms of a singular value decomposition (SVD), or equivalently, by means of an iterative algorithm of reciprocal averaging and multiple regression. In this chapter we give details of both approaches, although the latter almost exclusively dominates the relevant literature. There are two types of explanatory variables — environmental variables and covariables: by environmental variables we mean variables of prime interest; covariables are those variables whose effect is to be removed (ter Braak, 1987b). When covariables are not present, the steps involving these variables are omitted. We first introduce some notation:

Let $Y = [y_{ik}]$ $(i = 1,..., n; k = 1,..., m)$ be a species-by-sites matrix containing the observations of n species at m sites. The observation y_{ik} must be greater than, or equal to, 0;

$Z_2 = [z_{2kj}]$ $(k = 1,..., m; j = 1,..., q)$ be a sites-by-environmental variables matrix containing the observations of q environmental variables at the m sites;

$R = \text{diag}\,(y_{i.})$ be a diagonal matrix of species totals;

$W = \text{diag}\,(y_{.k})$ be a diagonal matrix of site totals;

$W^* = \text{diag}\left(\frac{y_{.k}}{y_{..}}\right)$

If covariables exist then let $Z_1 = [z_{1kl}]$ $(k = 1,..., m; l = 0,..., p)$ be a sites-by-covariables matrix containing the observations of p covariables at the m sites. The first column is a column of 1's to account for the intercept in the regression analysis and the observations z_{1kl} and z_{2kj} may take any real value. We can also define products of variables in order to examine whether the effect of one variable depends on the value of another variable (in the same way as in multiple regression).

4.3.2 Preliminary Calculations

Some preliminary calculations are necessary before CCA is implemented, regardless of whether we use the iterative algorithm or the singular value decomposition approach. These calculations are to standardise all the environmental variables (both environmental variables of interest and covariables which are not of direct interest).

The effects of the covariables are then removed by calculating the residuals of the linear regression of each of the environmental variables on the set of covariables. However, we suggest that it may be preferred not to make a distinction between covariables and environmental variables but to leave both in the complete analysis, unless it is absolutely certain which variables are the most appropriate for explaining species abundance (i.e. variables such as time and temperature would be classed as covariables but other, less obvious variables, would not). We describe the preliminary calculations below.

P1. It is generally recommended to standardise the environmental variables to zero mean and unit variance. Although ter Braak (1987b) does not say so explicitly, this is presumably important only when the variables are in different units or have widely differing variances. If there is only one variable and no covariables then it is clearly not necessary (although only one canonical ordination axis can be extracted). We also believe that there may be some situations, when the variables are in the same units, where they should not be standardised so as to allow those with greater values to have relatively higher weight in the calculations.

To standardise the variables, calculate the mean and variance for environmental variable j:

$$\bar{z}_2 = \sum_k wz_{2kj}, \qquad v_2 = \sum_k w\left(z_{2kj} - \bar{z}_2\right)^2$$

and set

$$z_{2kj} = \frac{\left(z_{2kj} - \bar{z}_2\right)}{\sqrt{v_2}}.$$

P2. If covariables exist, standardise these to zero mean and unit variance.

Calculate the mean and variance for covariable l:

$$\bar{z}_l = \sum_k w z_{lkl}, \qquad v_l = \sum_k w\left(z_{lkl} - \bar{z}_l\right)^2$$

and set $$z_{lkl} = \frac{\left(z_{lkl} - \bar{z}_l\right)}{\sqrt{v_l}}.$$

P3. Calculate for each environmental variable j the residuals of the multiple regression of the environmental variables on the covariables:

$$c_j^* = (Z_l^T W Z_l)^{-1} Z_l^T W z_{2j}$$

$$\tilde{z}_{2j} = z_{2j} - Z_l c_j^*$$

where $z_{2j} = \left(z_{21j}, \ldots, z_{2nj}\right)^T$ and c_j^* is the vector of coefficients of the regression of z_{2j} on Z_1. Define $\tilde{Z}_2 = \left[\tilde{z}_{2ij}\right]$ (i = 1,..., n; j = 1,..., q).

Having performed the above calculations, canonical correspondence analysis can now be implemented. The iterative algorithm approach is described in the following section.

4.3.3 The Iterative Algorithm Approach

We denote the species and site scores on the s-th ordination axis by u = [u_i] (i = 1,..., n) and x = [x_k] (k = 1,..., m) respectively. The canonical coefficients of the environmental variables are denoted by c = [c_j] (j = 1,..., q) and the site scores of the previous (s-1) ordination axes are denoted as columns of the matrix A. We carry out the following steps.

Step 1 : Start with arbitrary, but unequal site scores $x = [x_k]$. Set $x_k^0 = x_k$.

Step 2: Derive new species scores from the site scores by weighted averaging:

$$u_i = \frac{\sum_k y_{ik} x_k}{y_{i.}}.$$

Step 3: Derive new site scores $x^* = [x_k^*]$ from the species scores by weighted averaging:

$$x_k^* = \frac{\sum_i y_{ik} u_i}{y_{.k}}.$$

Step 4: Make $x^* = [x_k^*]$ uncorrelated with the covariables by calculating the residuals of the multiple regression of x^* on Z_1:

$$x^* = x^* - Z_1\left(Z_1^T W Z_1\right)^{-1} Z_1^T W x^*.$$

Step 5: If $q > s_A$ where s_A is number of axes already extracted, then calculate a multiple regression of the site scores x^* on the environmental variables Z_2:

$$c = \left(Z_2{}^T W Z_2\right)^{-1} Z_2{}^T W x^*$$

and take as new site scores the fitted values:

$$x = \tilde{Z}_2 c.$$

Step 6: If $s_A > 0$, make $x = [x_k]$ uncorrelated with previous axes by calculating the residuals of the multiple regression of x on A:

$$x = x - A\left(A^T W A\right)^{-1} A^T W x.$$

Step 7: Standardise $x = [x_k]$ to zero mean and unit variance:

$$\bar{x} = \sum_k w_k^* x_k$$

$$s^2 = \sum_k w_k^* (x_k - \bar{x})^2 \qquad (4.1)$$

$$x_k = \frac{(x_k - \bar{x})}{s}.$$

Step 8: Stop on convergence i.e. when the new site scores are sufficiently close to the site scores of the previous iteration. If:

$$\sum_k w_k^* (x_k^0 - x_k)^2 < 10^{-10}$$

then go to step 9. Otherwise, set $x_k^0 = x_k$ and go to step 2.

Step 9: Set the singular value λ equal to s in (4.1) and add $x = [x_k]$ as a new column to the matrix A.

Step 10: Set $s_A = s_A + 1$ and go to step 1 if further ordination axes are required; otherwise stop.

4.3.3.1 The Transition Formulae

Each time step 10 is reached, site scores which are a linear combination of the environmental variables are obtained for a particular axis (x). From these we can obtain, for each axis, by using the following transition formulae with the appropriate λ, species scores (u) and site scores (x^*) which are weighted mean species scores:

$$u = \lambda^{-\alpha} R^{-1} Y^T x \qquad (4.2)$$

$$x^* = \lambda^{\alpha-1} (I - Z_1 (Z_1^T W Z_1)^{-1} Z_1^T W) W^{-1} Y u \qquad (4.3)$$

$$c = \left(\tilde{Z}_2{}^T W \tilde{Z}_2\right)^{-1} Z_2{}^T W x^*$$

$$x = (I - A(A^T W A)^{-1} A^T W)\,\tilde{Z}_2 c$$

where α $(0 \leq \alpha \leq 1)$ is specified by the analyst.

There is no natural way of selecting α and in most fields outside community ecology only α = 0.5 has been used (Oksanen, 1987), but in community ecology either α = 0 or α = 1 is usually used. When α = 0, species scores are weighted means of the site scores; when α = 1, site scores are weighted means of the species scores; and when α = 0.5, sites and species are treated in a symmetric way, so that neither is the weighted mean of the other. Ter Braak (1985) believes that in species-by-sites matrices the choice of α = 1 is more appealing because there are nearly always species whose optimum is outside the sampled range of sites — therefore the species should have a greater range of scores and the site scores should be the direct weighted averages of species scores. In contrast, when α = 0 the species' optima all lie inside the range of sample scores, although this is the default in CANOCO 3.1 and ter Braak (1987b) does comment that the choice of scaling is less critical the higher the eigenvalues of the ordination axes. We use both α = 0 and α = 1 in this chapter and in Chapter Nine.

Having explained the iteration algorithm approach to CCA we now describe the singular value decomposition approach.

4.3.4 The Singular Value Decomposition Approach

The species and sites co-ordinates produced by the algorithm of 4.3.3 can also be obtained by the following singular value decomposition. The details are taken from Jongman *et al.* (1995) and use the notation defined in 4.3.1.

Define:

$$S_{12} = YZ_2$$

$$S_{22} = Z_2^T RZ_2.$$

Calculate the SVD of:

$$W^{-0.5}S_{12}S_{22}^{-0.5} = P\Lambda^{0.5}Q^T$$

where $\Lambda^{0.5} = \text{diag}\,(\lambda_1^{0.5},\dots,\lambda_m^{0.5})$ are singular values;

P (m × m) and Q (m × m) are orthonormal i.e. $P^TP = Q^TQ = I_m$.

For the case of $\alpha = 1$ the species co-ordinates are given by:

$$U = W^{-0.5}P\Lambda^{0.5},$$

the site scores are given by:

$$X = Z_2S_{22}^{-0.5}Q$$

and the canonical coefficients are obtained from:

$$C = S_{22}^{-0.5}Q.$$

Having obtained species and site co-ordinates, there are some quantities that aid the interpretation of the ordination diagram and these are described in 4.3.5, using the notation of 4.3.3. For the questions addressed in Chapter Nine, such as investigating the influence of sample size on the analysis, the detection of influential categories and variables and the development of methods for assessing the stability of the ordination diagram, we believe that the SVD is the more useful approach.

4.3.5 Quantities of Interest

The following sections explain the main quantities of interest which aid the interpretation of the ordination diagram. They are discussed in Sections 4.5 and 4.6 for the hunting spiders and dune meadow vegetation data respectively.

4.3.5.1 Intraset Correlations

The species scores (u) and site scores (x^*) are the co-ordinates which are plotted in the ordination diagram, whereas the intraset correlations are the correlation coefficients between the environmental variables ($\tilde{Z}_2$) and the ordination axes (x). They relate to the rate of change in community composition per unit change in the corresponding environmental variable, where the other environmental variables covary with that one environmental variable. Any arbitrariness in the units of measurement of the variables was removed when the variables were standardised prior to the analysis (see 4.3.2).

4.3.5.2 Environmental Variable Co-ordinates

The co-ordinate of an environmental variable on axis s is $\lambda_s^{0.5}$ multiplied by the intraset correlation (see 4.3.5.1) of the environmental variable with that axis, where λ_s is the singular value of axis s. Thus, environmental variables can be displayed along with species and sites in the ordination diagram.

4.3.5.3 Canonical Coefficients

The final regression coefficients (c) as defined in 4.3.3.1 are called canonical coefficients and these define the ordination axes as linear combinations of the environmental variables, along which the distributions of the species are maximally separated. Canonical coefficients relate to the rate of change in community composition per unit change in the corresponding environmental variable, where the other environmental variables are held constant. When the environmental variables are strongly correlated with each other e.g. when the number of environmental variables approaches the number of sites, the effects of different environmental variables on community composition cannot be separated out and so the canonical coefficients are unstable. This is known as the multicollinearity problem and we see an example of this in Section 4.5 for the hunting spider data.

4.3.5.4 Species-Environment Correlations

The multiple correlation coefficient of the final regression is called the species-environment correlation and is a measure of how well the extracted variation in community composition can be explained by the environmental variables. It is equal to the weighted correlation between the site scores (x^*) which are weighted mean species scores and the site scores (x) which are a linear combination of the environmental variables. However, McCune (1997) comments that as the number of environmental variables increases, the species-environment correlation always converges to 1, so that it is a poor measure of the success of an ordination.

4.3.5.5 Variance Inflation Factors

Variance inflation factors (VIFs) relate to the (partial) multiple correlation between environmental variable j and the other environmental variables in the analysis. If the VIF is large, say > 20, then the variable is almost perfectly correlated with the other variables and therefore has no unique contribution to the regression equation and its canonical coefficient is unstable. When implementing environmental variable selection methods in Chapter Nine, it is helpful if we can ensure that the VIFs of the selected variables are low.

4.3.5.6 Inter-set Correlations

The inter-set correlations of the environmental variables with the axes are the weighted correlation coefficients between the environmental variables and the site scores (x^*) which are weighted mean species scores. In contrast to the canonical coefficients, the inter-set correlations do not become unstable when the environmental variables are strongly correlated with each other i.e. when the VIFs are large. The mean squared inter-set correlation is the fraction of the total variance in the standardised environmental data that is extracted by each species axis.

4.3.5.7 Ordinal Variables

CCA cannot directly cope with ordinal variables — these must be treated either as if they are quantitative, or as nominal variables; nominal variables must be transformed to dummy variables. This causes some problems when assessing the stability of the variables in Chapter Nine.

4.4 The Ordination Diagram and its Interpretation

The CCA map is usually examined in two dimensions, because this gives the most convenient visual representation of the data. The following sections describe the interpretation of the species points, site points and environmental variables in the ordination diagram and Chapter Nine investigates how this interpretation changes when the number of sites or variables included in the analysis alters. Each pair of {sites, species, environmental variables} forms a biplot and when the CCA map contains all three quantities it is known as a triplot.

4.4.1 Displaying Species and Sites

In the ordination diagram, sites and species are each represented by points that form a biplot. These points jointly represent the dominant patterns in community composition insofar as these can be explained by the environmental variables. By taking $\alpha = 0$ (equations (4.2) and (4.3)), species scores are weighted mean site scores and each species point then lies at the centroid of the sites points at which it occurs, with the origin of the plot lying at the centroid of the species points. We can then infer which species are likely to be present at a particular site (i.e. those located close to the site). We consider methods of assessing the stability of the site points (i.e. how representative they are of the true population of data) in Chapter Nine.

4.4.2 Displaying Qualitative Environmental Variables

Each class of a nominal environmental variable is represented separately by a point located at the centroid of the sites belonging to that class. Classes consisting of sites with high values for a species will then tend to lie close to the point representing that species. If the environmental data consist of a single qualitative variable, the points for classes and species in the CCA diagram are identical to those obtained from a correspondence analysis on the species-by-classes table, the entries of which are the total abundance of each species in each class (see Chapter Two). The stability of nominal environmental variables is discussed in Chapter Nine.

4.4.3 Displaying Quantitative Environmental Variables

Quantitative environmental variables are represented by vectors (lines), each of which determines a direction or axis in the diagram. It is only the directions and relative lengths of the vectors that convey information, so the lengths can be reduced or extended to fit into the ordination diagram. Each vector points in the direction of maximum change in the value of the associated variable and the vector length is proportional to this maximum rate of change. An environmental variable with a long vector is one for which abundances vary rapidly as the variable changes. The length of a vector also indicates the importance of the variable: the length is equal to the multiple correlation of the variable with the displayed ordination axes and thus indicates how well the values of the variable are displayed in the biplot of sites and environmental variables.

If we can extend a vector in both directions then from each species point we can drop a perpendicular to this axis, where the end points of these perpendiculars indicate the relative positions of the centres of the species distributions along the environmental axis (see Figure 4.2). In general, the approximate ranking of the weighted averages for a particular environmental variable can be seen from the order of the endpoints of the perpendiculars of the species along the axis for that variable: the inferred weighted average is higher than average if the endpoint of a species lies on the same side of the origin as the head of a vector and vice versa. The grand mean of each environmental variable is represented by the origin of the plot. We discuss methods for investigating the stability of these quantitative variables in Chapter Nine.

4.4.4 Species and Environmental Variables

The species points and the vectors of the environmental variables jointly represent the species' distributions along each of the environmental variables and this joint plot is a biplot. This biplot provides a weighted least squares approximation of the weighted averages of the species with respect to the environmental variables. There is a measure of goodness of fit which expresses the percentage of variation of the weighted averages accounted for by the two-dimensional diagram of vectors and species and this is given by:

$$t_1 = 100 \times \frac{\sum_{j=1}^{2} \lambda_j}{\sum_{j=1}^{q} \lambda_j}.$$

Ter Braak (1987b) comments that the percentage of variation accounted for is dependent on the number of variables in the analysis: with only two environmental variables, two canonical axes always explain 100% of the variation, regardless of whether or not the result is ecologically meaningful.

4.4.5 Ordination Axes

The first eigenvalue is equal to the maximised dispersion of species scores along the first ordination axis. The second and further axes also select linear combinations of environmental variables that maximise the dispersion of the species scores, but these are subject to being uncorrelated with previous axes; in principle as many axes can be extracted as there are environmental variables. CCA is in fact restricted correspondence analysis (CA), but the restrictions become less strict with the more environmental variables that are included in the analysis: if $q \geq m-1$, then there are no restrictions and CCA is then CA. In Chapter Nine we compare the species and site points obtained from CA and CCA for the hunting spider data.

Eigenvalues indicate how long the extracted gradients are (where higher values mean longer gradients). If the gradients are long then the scores (optima) of most species lie close to the centre region where the sites lie and there is some evidence that the probability of occurrence of species along the gradients is unimodal as required. By looking at the signs and relative magnitudes of the intraset correlations (4.3.5.1) and the canonical coefficients (4.3.5.3), we can infer the relative importance of each environmental variable for predicting the community composition.

4.4.6 Tolerances

The tolerance or weighted standard deviation of a species is a measure of its niche-breadth and is calculated as:

$$t_i = \sqrt{\sum_{k=1}^{m} \frac{y_{ik}}{y_{i.}} \left(x_k - u_i\right)^2}$$

where x_k is the site score at site k (ter Braak, 1985; ter Braak & Verdonschot, 1995). The tolerance can be calculated for each species on each extracted ordination axis and plotted as a cross with the species points as the centres and the tolerances on each axis as lines through this point. We illustrate the tolerances on the first two axes for each of the hunting spider species in Figure 4.3.

4.4.7 Supplementary Points

As was the case for correspondence analysis in Chapter Two, samples and species in CCA can be made passive so that they do not influence the determination of the ordination. Their scores on the ordination axes are calculated after CCA has been implemented.

4.4.8 Ranking Environmental Variables

It is known that the environmental variables can be ranked in order of their importance for determining the species composition. A related aim is to reduce a large set of variables to a smaller set that suffices to explain the variation in species composition and this forms part of the focus of Chapter Nine. Ter Braak & Verdonschot (1995) comment that environmental variables can be ranked and selected in CCA in a similar way to how predictors can be ranked and selected in regression, with the species and the environmental variables taking the roles of the response and explanatory variables respectively. However, CCA aims to explain the variation in the species composition i.e. in relative abundance values, whereas linear regression aims to explain the variation in absolute abundances. Ter Braak & Verdonschot (1995) describe a forward selection method and apply it to macro-fauna data from the

Netherlands. We explain this method in detail and introduce an alternative in Chapter Nine.

4.4.9 The 'Arch Effect'

The 'arch effect' was described in the context of correspondence analysis in 2.4.2.2 and is an approximately quadratic dependence between the scores of the first two axes, which occurs whenever a short gradient is dominated by a long gradient (Gauch, 1982). In CCA it is known that the 'arch effect' can be removed by dropping superfluous environmental variables; variables that are highly correlated with the 'arched' axis (often the second axis) are most likely to be superfluous. Detrended CCA (Hill & Gauch, 1980) also removes the 'arch effect', but as with detrended correspondence analysis it has been heavily criticised.

4.4.10 Canonical Correspondence Analysis and Archaeology

The main application area of CCA is the field of ecology, but there is scope for its use in other areas. In archaeology or palaeoecology for example, it may be that the environmental conditions at the sites are unknown and that this is what we are hoping to discover, but we may have information on the environmental preferences of species. In this situation we can just reverse the roles of sites and species, taking Z_2 in Section 4.3.1 to be species-by-environmental variables and obtain species scores which are constrained to be linear combinations of the environmental preferences (i.e. obtaining species scores in equation (4.1) instead of site scores). This gives us information on the site-environment relationship rather than on the species-environment relationship. We therefore obtain values for site-environment correlations, which measure how well the environmental variables explain the variation between sites. Additionally, we can project the sites onto the environmental variables and see the relative positions of the centres of the site distributions along the axes. In summary, rather than measuring environmental conditions at each site and inferring species preferences from this, we use our knowledge of the environmental preferences of species to infer past environmental conditions at sites. Of course, environmental preferences can alter over time and we need to bear this in mind. One of the most recent references of CCA applied to archaeology is Bogaard *et al.* (1999).

4.5 Application to Hunting Spiders

In this section CCA is applied to data consisting of the distributions of 12 species of hunting spider in a Dutch dune area, taken from van der Aart & Smeenk-Enserink (1975) and analysed by ter Braak (1986) in relation to environmental data. These data were described in Chapter One and comprise the numbers of individuals of each species caught in 28 pitfall traps (sites) with 26 environmental variables measured at each site. This is a considerable number of sites and variables and Chapter Nine introduces one method of adjusting the overall number of sites used in the analysis, as well as a technique for assessing which sites are particularly influential. It also discusses variable selection methods which reduce the number variables used in the analysis. In ter Braak (1986) the number of environmental variables was considered too large to sort out their independent effects on community composition and 18 were removed on a priori grounds; two more were removed because they were strongly correlated with one of the remaining six variables. We implement CCA on these remaining variables which are labelled 1, 4, 5, 6, 7 and 26 in the analysis below. Ter Braak (1986) took $\alpha = 1$ (see 4.3.3.1), transformed the species-by-sites data by taking square roots to down-weight high abundances and transformed the environmental variables by taking logarithms. However, we take $\alpha = 0$ and implement CCA on both the original data and the transformed data. We also treat the species data as presence/absence in order to investigate how the inferences made alter depending on the form of the data.

4.5.1 The Original Data

Implementing CCA on the untransformed data (Tables A.12-A.13 of the Appendix), we obtain the ordination diagram of Figure 4.1, where the environmental variables are represented by lines from the origin (labelled with codes from Table 4.2), the species are represented by circles (three of which are labelled) and the sites are indicated by pluses.

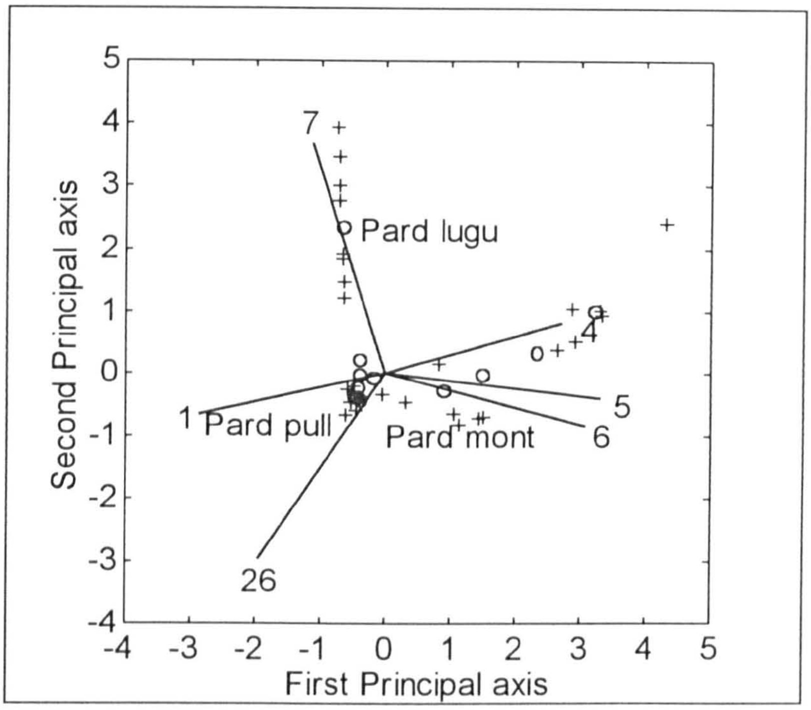

Figure 4.1 Canonical Correspondence Analysis Map of Hunting Spiders

The species-environment correlations (see 4.3.5.4) of the first two axes are 0.97 and 0.90 and because these values are high we can be confident that the environmental variables are sufficient to explain the major variation among the different species of spider. The environmental variable vectors in conjunction with the species points account for 77.2% of the variance in the weighted averages of the 12 spiders with respect to the six environmental variables in two dimensions (see 4.4.4), which is also fairly high. The species and site points show some evidence of an 'arch effect' as discussed in 4.4.9.

The canonical coefficients (final regression coefficients), intraset correlations (correlations between the environmental variables and the ordination axes) and variance inflation factors which were described in sections 4.3.5.3, 4.3.5.1 and 4.3.5.5 respectively are displayed in Table 4.2. We see that variables moss, twigs and herbs have VIFs of 30.32, 38.40 and 58.72 respectively, which implies strong multicollinearity among the environmental variables and unstable canonical coefficients, so we must be careful when interpreting these variables. Interpreting the axes, the first axis appears to be a moisture gradient on which the drier sites have a high percentage of moss or bare sand, whereas the correlations of the second axis show a contrast between sites with a high cover of herbs and sites without.

Table 4.2 Canonical Coefficients, Intraset Correlations and Variance Inflation Factors for the Hunting Spiders

		Canonical Coefficient		Intraset Correlation		Variance Inflation
Code	**Environmental Variable**	**Axis 1**	**Axis 2**	**Axis 1**	**Axis 2**	**Factor**
1	**Water Content**	-0.0017	-0.0031	-0.905	-0.302	2.59
4	**Bare Sand**	0.0026	0.0041	0.849	0.362	12.33
5	**Fallen Twigs**	-0.0040	0.1380	-0.358	1.663	38.40
6	**Cover Moss**	0.0021	0.0017	1.036	-0.180	30.32
7	**Cover Herbs**	-0.0049	0.0051	-0.607	-1.352	58.72
26	**Light Reflection**	0.0024	0.0000	0.965	-0.386	3.01

In Figure 4.1, the small angle between the vectors representing twigs and moss demonstrates the high correlation between these two variables and the angle of approximately 180° between variables water content and sand indicates that these are highly negatively correlated. It is also evident that all the vectors are of roughly the same length, which means that the gradients in abundances are similar for each variable. Considering the species and site points in the ordination diagram, we infer that *Pardosa pullata* reaches its maximum abundance in the pitfall traps on the left of the diagram and that *Pardosa monticola* is most abundant in the pitfall traps in the centre-right of the diagram. *Pardosa lugubris* occupies an aberrant position in the diagram, being the single spider species that occurs mainly in habitats with a high cover of herbs.

The fraction of the total variance in the standardised environmental data that is extracted by each species axis (the mean squared interset correlation) is 40.0% for the first axis and 20.0% for the second axis: this is fairly high. The first two eigenvalues are 0.64 and 0.30, which show that the extracted gradients are reasonably long and there is therefore some evidence that the probability of occurrence of a species along these environmental gradients is unimodal as required.

Extending the bare sand axis and dropping a perpendicular from each species point onto this axis (see Section 4.4.3) means that we can see the relative positions of the centres of the species distributions along the axis. This is illustrated in Figure 4.2, although the species in the bottom left-hand corner are not labelled as this would confuse the diagram. Thus, *Arctosa perita* has the highest weighted average of all the species and is higher than the average of all species (the origin), because it lies on the same side of the origin as the axis. The species with the next highest weighted average is *Alopecosa fabrilis* followed by *Alopecosa accentuata*; we can make similar inferences about the species on the other environmental variables.

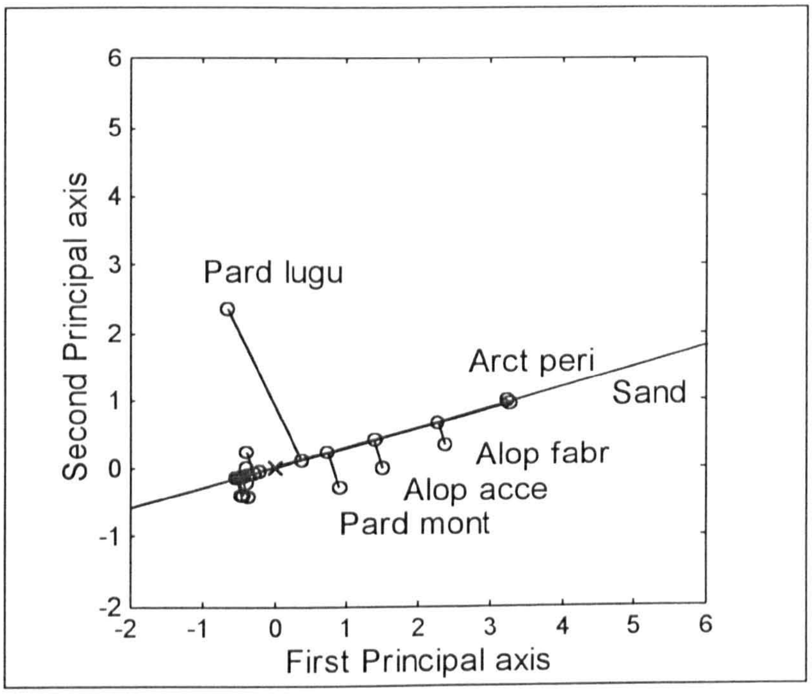

Figure 4.2 Projection of Hunting Spider Species Points onto the Bare Sand Axis

We now consider the tolerances of each species (see 4.4.6) and display them in Figure 4.3.

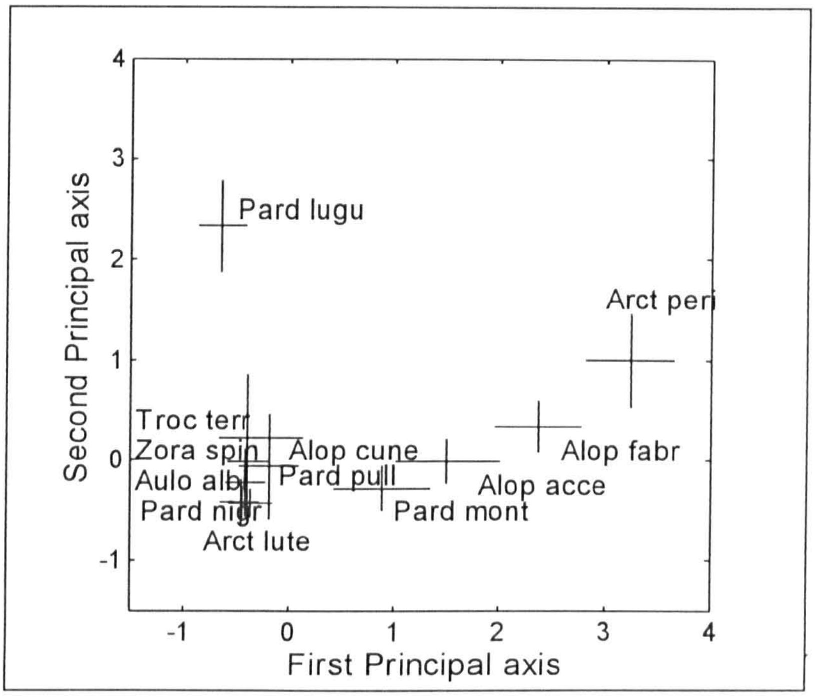

Figure 4.3 Tolerances of Hunting Spider Species

The species in the left-hand corner (*Zora spinimana*, *Aulonia albimana*, *Arctosa lutetiana*, *Pardosa nigriceps*, *Pardosa pullata* and *Alopecosa cuneata*) appear to have similar tolerances on both axes to the linear combination of environmental variables obtained from CCA, whereas *Pardosa lugubris* and *Trochosa terricola* have relatively high tolerances to the environmental variables which are strongly represented on the second axis.

4.5.2 Transformed Data

Transforming both the species data (by taking square roots to down-weight high abundances) and the environmental variables (by taking logs) as in ter Braak (1986), leads to the ordination diagram of Figure 4.4. We are interested in comparing this figure with that of the untransformed data (Figure 4.1).

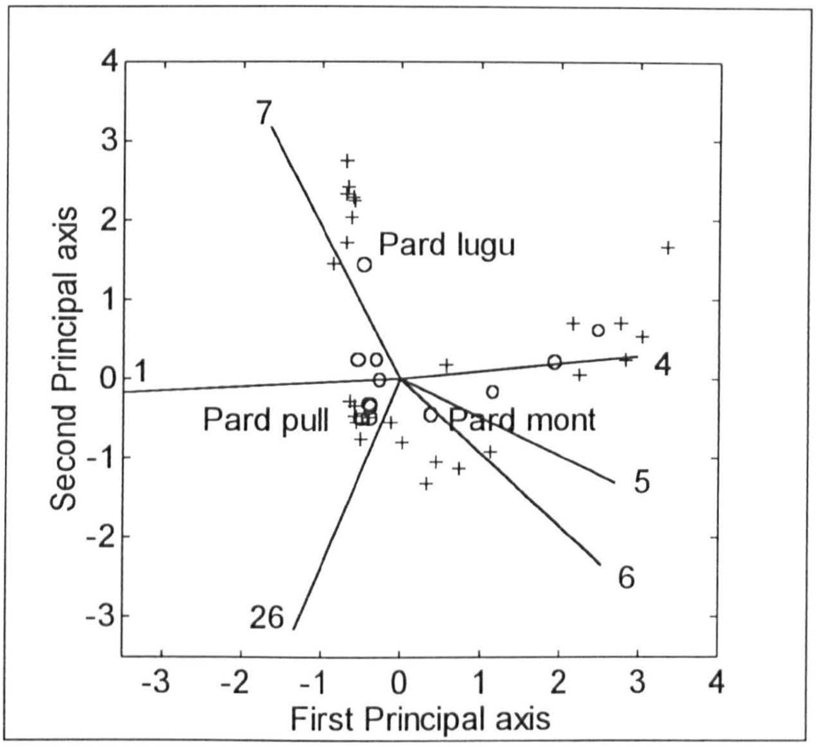

Figure 4.4 Canonical Correspondence Analysis Map of Hunting Spiders (Transformed)

The species-environment correlations on the first two axes are 0.96 and 0.95 which are similar to those for the original untransformed data. However, the vectors for the environmental variables account for, in conjunction with the species points, 88.2% of the variance in the weighted averages of the spiders in the first two dimensions, which is 11% higher than for the untransformed data. The interpretation of the canonical coefficients and intraset correlation coefficients is similar for both sets of data. Considering the transformed data, there are no variables with high variance inflation factors and so we can be confident when interpreting the variables, although we note that their relative positions in the ordination map are similar regardless of the form of the data. The distributions of sites and species across the ordination diagrams are again similar and so for these data the transformations have had little effect on the interpretation of the CCA map.

4.5.3 Presence/Absence Data

In this section we use the untransformed environmental variables, but we convert the species data into a presence/absence format (i.e. replacing any value greater than zero with a one). The resulting ordination diagram is illustrated in Figure 4.5.

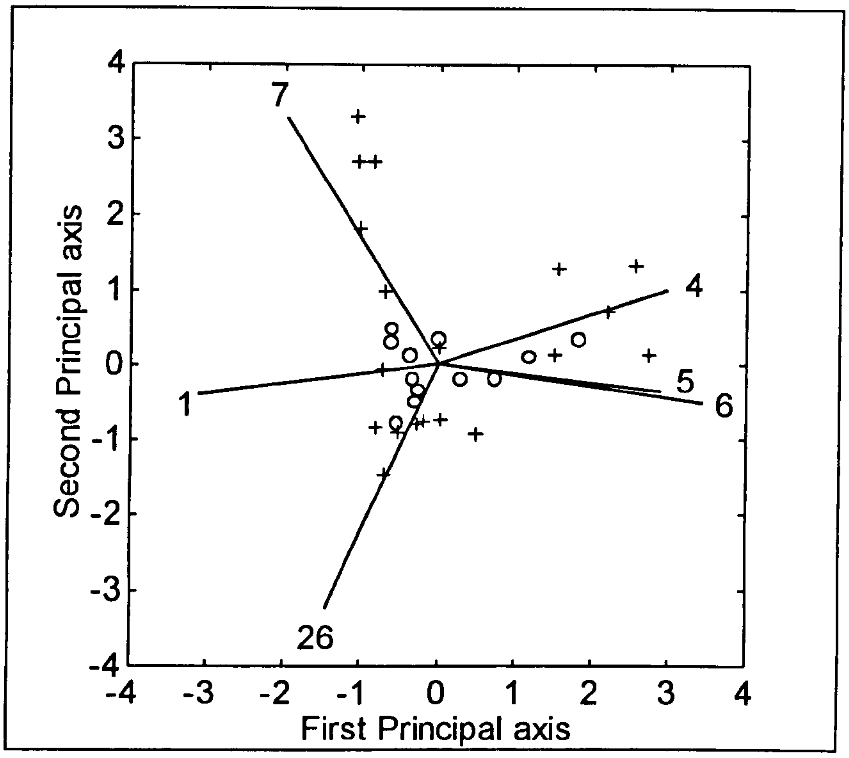

Figure 4.5 Canonical Correspondence Analysis Map of Hunting Spiders (Presence/Absence)

We see from the map that the species *Pardosa lugubris* is no longer situated apart from the other species, although this is as we would expect given that the data no longer consist of absolute abundances. It is also clear that variables 5 and 6 (twigs and moss) are more highly correlated than they were in Figures 4.1 and 4.4, but the positions of the other variables on the map have not really altered. The species-environment correlations of the first two axes are 0.92 and 0.84, which are lower than for both the transformed and untransformed data and so the variation in community composition is slightly less well explained by the environmental variables. It is also evident that the species points no longer form an ‘arch effect’.

The environmental vectors in conjunction with the species points account for 91.8% of the variance in the weighted averages of the 12 spiders in the first two dimensions, which is extremely high. However, four of the six variance inflation factors are also very high (sand=21.75, moss=36.35, twig=55.04, herb=63.45) and so there is greater multicollinearity than for the untransformed data. The first two eigenvalues are 0.37 and 0.11 and so the first extracted gradient is reasonably long but the second is not and there is therefore little evidence of unimodality in the occurrence of species along the environmental gradients. On the basis of the above interpretation it therefore appears that the presence/absence form of the data is the least useful form of data for canonical correspondence analysis.

4.5.4 The Original Data and all 26 Environmental Variables

As we explained at the beginning of Section 4.5, 26 environmental variables were originally measured at the 28 sites. In this section we apply CCA to all these variables in combination with the untransformed species data in order to assess the influence of the number of environmental variables measured on the interpretation of the results. The CCA map is illustrated in Figure 4.6, where most of the environmental variables are labelled; asterisks indicate the reduced set of six variables which were used in the previous CCA maps. It is clear from the figure that as was the case in Figures 4.1 and 4.4, both species and site points form an 'arch effect'.

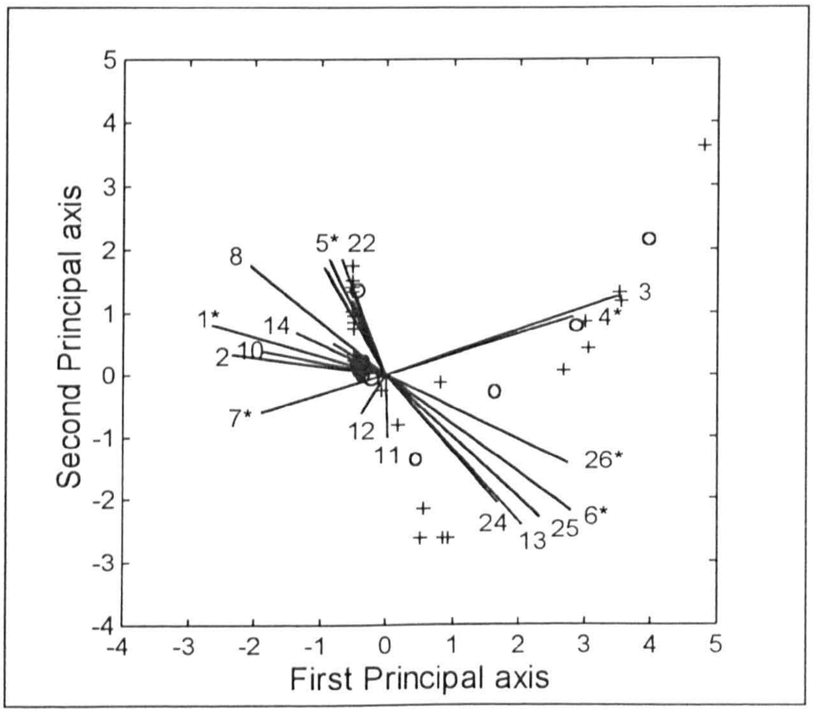

Figure 4.6 Canonical Correspondence Analysis Map of Hunting Spiders (all 26 Variables)

Comparing Figure 4.6 with Figure 4.1 we can see that the relationships between the environmental variables have altered considerably, for example variables 6 & 26 are now highly correlated when previously they were uncorrelated. We also see that there is a group of variables on the bottom right that are highly correlated — {6, 13, 24, 25, 26} — variables 3 & 4 are also highly positively correlated. Examining the variance inflation factors reveals that these are very large (all are greater than 24) and so multicollinearity is severe. We cannot, therefore, be confident in any of our interpretations based on this figure. There is, however, considerable scope for implementing variable selection methods and we discuss this further in Chapter Nine.

4.6 Application to Dune Meadow Vegetation

In this section we apply CCA to data consisting of the abundances of 30 plant species (measured on van der Maarel's scale of 1-9) in 20 sample plots on the Dutch island of Terschelling. These data were described in Chapter One and comprise five environmental variables, two of which are considered to be nominal; the data were taken from ter Braak (1987b) but originate in Batterink & Wijffels (1983, unpublished).

Implementing CCA (this time with $\alpha = 1$) as in ter Braak (1987b), leads to the ordination diagram of Figure 4.7 where species are represented by circles, samples by crosses, quantitative environmental variables by lines and nominal environmental variables by asterisks. All the variables are labelled.

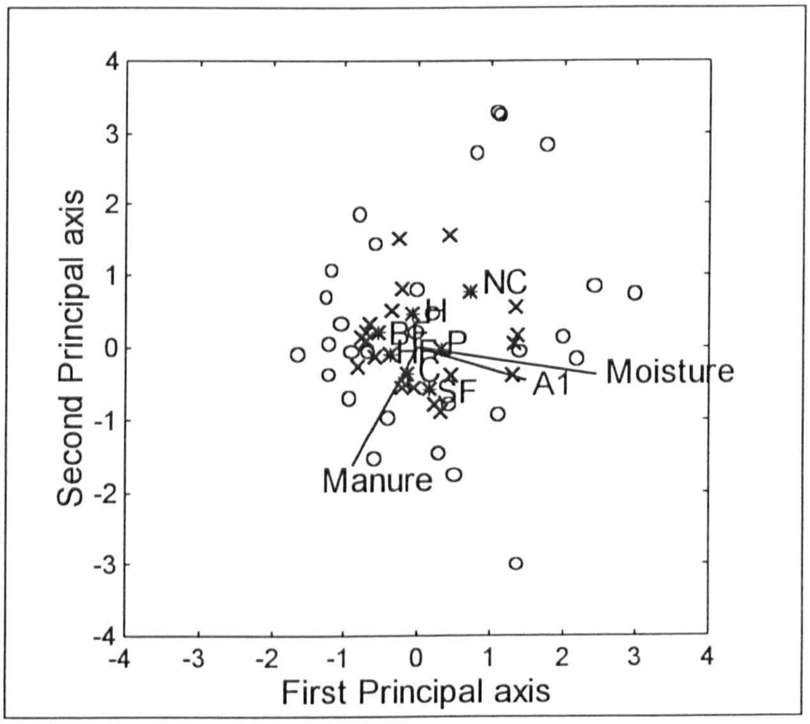

Figure 4.7 Canonical Correspondence Analysis Map of Dune Meadow Vegetation

In contrast with the ordination maps for hunting spiders (Figures 4.1, 4.4, 4.5 and 4.6), there is no 'arch effect' in Figure 4.7. It is also clear from the figure that variables moisture and A1 are highly correlated (because there is a small angle between the vectors representing them). The first two eigenvalues are 0.49 and 0.27, which suggest that the environmental gradients are reasonably long. The variance inflation factors are fairly low (less than 12) so we can be confident in our interpretation of this figure. In addition, the species-environment correlations are 0.97 and 0.92 and so the variation in community composition is well explained by the environmental variables. The vectors representing the environmental variables account for, in conjunction with the species points, 62.3% of the variance in the weighted averages of the plants in the first two dimensions, which is reasonably high.

4.7 Connections with other Techniques

It is known that canonical correspondence analysis has close connections with other multivariate techniques and we describe some of these below.

4.7.1 Canonical Correspondence Analysis and Correspondence Analysis

In the context of community ecology, correspondence analysis is applied to data consisting of a species-by-sites matrix: it constructs from these data principal axes that maximise niche separation. Canonical correspondence analysis, however, also requires environmental variables to be measured at each site and constructs principal axes by linearly combining the measured environmental variables; this has the advantage that the environmental basis of the ordination is guaranteed. However, if there are nearly as many environmental variables as there are sites then CA and CCA are reported to produce the same site and species ordination. It is also known that correspondence analysis is very susceptible to species-poor sites containing rare species in that it places such aberrant sites (and the rare species occurring there) at extreme ends of the first ordination axis, relegating the major vegetation trends in the data to later axes. In contrast, canonical correspondence analysis does not show this 'fault', provided that the sites that are aberrant in species composition are not so aberrant in terms of the environmental variables. A practical problem with both techniques, however, is that species that are unrelated to the ordination axes tend to be placed in the centre of the ordination diagram and are not distinguished from species that have true optima there (ter Braak, 1985, 1986). This can be circumvented by looking at a species-by-sites matrix in which species and sites are arranged in order of their scores on one of the ordination axes (see the discussion in Section 2.4).

4.7.2 Canonical Correspondence Analysis and Discriminant Analysis

Chessel *et al.* (1987) and Lebreton *et al.* (1988) were among the first to recognise the formal equivalence between canonical correspondence analysis and discriminant analysis. Multiple discriminant analysis works on measurements of variables on individuals belonging to different groups, where the usual aim is to assign new individuals with unknown group membership to groups on the basis of the measured variables. To investigate whether it is possible to discriminate between groups using fewer dimensions, canonical variates are obtained (linear combinations of the variables that maximally separate the groups). Replacement of 'groups' by 'niches of species' yields similar definitions for discriminant analysis and CCA, but with discriminant analysis the variables are measured on each individual, whereas with CCA the (environmental) variables are measured at each site. Furthermore, discriminant analysis is only appropriate if the number of sites is much greater than the number of species and the number of classes (Schaafsma & van Vark, 1979). Consequently, many ecological data sets cannot be analysed by discriminant analysis without dropping many species, but CCA can be used regardless of the number of species.

4.7.3 Canonical Correspondence Analysis and Canonical Correlation Analysis

In canonical correlation analysis the species scores are parameters estimated by a multiple regression of the site scores on the species variables and this regression means that the number of species plus the number of environmental variables must be smaller than the number of sites. In contrast, canonical correspondence analysis has no upper limit to the number of species that can be analysed.

4.8 Summary and Conclusions

Canonical correspondence analysis is a multivariate ordination method most commonly used in community ecology, although it does have applications in other areas, for example archaeology, but these are relatively underdeveloped. Within ecology, it is an appropriate method to use when interest lies in the response of a community of species to environmental variables measured at certain sites. With CCA, environmental variables are directly included in the ordination so that the resulting axes are linear combinations of these variables, measuring particular environmental gradients. It is usual for all variables to be standardised prior to analysis, but we suggested that there may be some situations, for example when the variables are in the same units, where this should be avoided so as to allow those variables with greater values to have relatively higher weight in the calculations. It is also known that CCA cannot directly cope with ordinal variables and we discuss this in Chapter Nine in the context of assessing the stability of the environmental variables.

This chapter has described the algebraic details of CCA, two methods of implementing the technique (by an iterative algorithm and by a singular value decomposition) and given a guide to the interpretation of the results, focusing on displaying the data in two dimensions. We have investigated the effect of the form of the data (raw, transformed, or presence/absence) on the results of the analysis and concluded that it is not advisable to implement CCA on presence/absence data, although this was the only form of data for which there was no evidence of an 'arch effect'. In Chapter Nine we discuss how the form of the data affects the stability of the CCA map.

We also raised the issue of how the number of environmental variables measured at each site influences the results of the analysis and discovered that there are severe problems with multicollinearity when large numbers of variables are measured, although even with smaller numbers of variables there can be high variance inflation factors, depending on the form of the data. Chapter Nine describes an existing method of selecting environmental variables and proposes a different approach, based on procrustes analysis. It was evident that ordination diagrams can become difficult to interpret when there are large numbers of categories (species or sites) to display and so in Chapter Nine we discuss methods of selecting categories. These are again based on the procrustes statistic. The question of how to detect the influence of individual categories on the determination of the CCA map (i.e. how does the interpretation of the map alter if these are removed) was also highlighted in this chapter and we address this in Chapter Nine.

Finally, the similarities between CCA and each of correspondence analysis, discriminant analysis and canonical correlation analysis were explained and in Chapter Nine the similarities between the interpretation of CCA maps and both CA and biplot ordination diagrams are discussed, using the hunting spider data (1.2.8).

Chapter Five

Stability, Sample Size and Correspondence Analysis

5.1 Introduction

Chapter Two explained the theory behind correspondence analysis and illustrated its application to pottery sherds and starch grains. It also raised the questions of how to deal with sparse data and how the number of artefacts collected influences the results of the analysis. For example, it may be that there are a minimum number of categories or artefacts below which it is not worthwhile carrying out a correspondence analysis because there is not enough information to distinguish between the categories in the resulting display. Considering our data sets, this could apply to the number of Memphis (1.2.1) and Amarna (1.2.2) wares that were identified, to the number of Amarna sites visited, to the number of Memphis contexts identified or to the total number of sherds collected. It could also apply to the number of starch grain types (1.2.3) or to the total number of grains obtained.

Similarly, there may be a maximum number of categories (pottery wares, sites or types of grain), artefacts or grains above which the resulting correspondence analysis map becomes too cluttered for patterns to be revealed. Considering the starch grains, if many different types of grain have been identified then groups of similar types, or the identification of which sites are similar in terms of vegetation, can be almost impossible. Investigation of such issues using specific examples aids us in developing general guidelines to help archaeologists (and others) when sampling and classifying artefacts, not least because the same type of data is frequently collected in archaeology, but also because pottery in particular is one of the most common artefacts found. In addition, we examine the stability of the two-dimensional maps obtained as a result of the analysis, by considering confidence regions based on convex hulls and concentration ellipses (i.e. how confident are we that the data collected are a representative sample of all possible data). This chapter, therefore, aims to combine the theory of correspondence analysis with other techniques such as bootstrapping and jack-knifing, in order to investigate problems such as those listed above.

The remainder of Section 5.1 explains the concepts of bootstrapping and 'stability' of a display and briefly discusses problems in assessing stability. Two bootstrap sampling methods involving the multinomial distribution are explained and applied in 5.2 (which one is applicable depends on how the data were collected) and methods of dealing with sparse contingency tables (i.e. trace and absolute zeroes), which are common in archaeology, are developed in 5.3. Convex hulls and concentration ellipses as methods of summarising stability and investigating similarities between categories are explained and extended in Section 5.4 and in 5.5 we discuss the jack-knife as a method for assessing stability. Section 5.6 introduces an alternative method of resampling which does not involve the multinomial distribution and investigates the influence of sample size on correspondence analysis maps. This section also discusses problems with estimating minimum required sample sizes in archaeology. We conclude the chapter in 5.7.

5.1.1 The 'Bootstrap'

The bootstrap (Efron, 1979) is used to assign measures of accuracy to statistical estimates. The idea is to resample from the original data — either directly or via a fitted model — to create replicate data sets, from which the variability of the quantities of interest can be assessed. Bootstrapping is applied throughout most of this chapter and again in Chapters Eight and Nine, where it is used in conjunction with biplots and canonical correspondence analysis respectively.

biplots and canonical correspondence analysis respectively.

5.1.2 Stability of the Displays

The purely algebraic technique of correspondence analysis is an exploratory method that can be regarded as a generalisation of a scatterplot for investigating the structure and relationships between two sets of categories. It is not a method of estimation (nor of hypothesis testing) but nevertheless bootstrapping methods can be used to assess the 'stability' of the displays, answering informal questions of how 'distinct' are different categories and how sensitive is observed structure to sampling variability in the data. Strictly, we should differentiate between internal stability and external stability: internal stability is at the level of the data matrix itself and external stability is at the level of the wider population (see Greenacre, 1984). Our main interest lies in investigating whether small differences in the original data matrix can produce relatively large differences in the correspondence analysis map, because this could indicate that either our data sample is not representative of the true population of data, or that we have a particularly influential category or cell and hence our interpretations of the display could be misleading.

Ringrose (1990) comments that if it were possible to obtain more data in exactly the same way as the data already collected (i.e. by using the same sampling scheme), then this process could be repeated many times to obtain a set of replicate data matrices, each of which could be subjected to correspondence analysis to produce a new set of points. Thus, each point (i.e. category) in the original analysis would lead to a cloud of points, one from each replicate matrix. This represents the uncertainty of a point's true position and the overlapping nature of the clouds of points could be used informally to assess the similarities between categories. However, if this repeated sampling is not possible, as is usually the case, then the observed sample can be treated as a proxy for the underlying distribution and new samples can be drawn from it. This is called 'resampling' or 'bootstrapping' and in Section 5.2.1 we briefly describe two methods of resampling using the multinomial distribution.

Having obtained a series of replicate matrices, the next two sections explain two different schools of thought regarding how to apply correspondence analysis to these replicates. They differ in whether the co-ordinates of each bootstrap should be relative to different axes or whether they should be related to the original co-ordinate system: both of these are illustrated in Section 5.2.

5.1.2.1 Greenacre: Partial Resampling

Greenacre (1984) holds the view that a new correspondence analysis should not be carried out on each replicate matrix, because this leads to the points' co-ordinates being relative to different axes. Instead, he advocates converting the bootstrapped sets of row and column profiles into points on the co-ordinate system calculated from the original data, using the transition formulae (see 5.2.2) i.e. the original plane is fixed as the viewing plane for the replications. The replicated row or column points, depending on which are of main interest, are then projected onto this plane in order to explore the stability of the points themselves as well as, indirectly, the stability of the original plane.

5.1.2.2 Milan & Whittaker: Filtering

Milan & Whittaker (1995) argue that because the partial resampling of Greenacre does not repeat the singular value decomposition (SVD) on each facsimile matrix, it does not give a full simulation of the sampling variation and the size of the region produced may be quite different from that obtained when a new correspondence analysis is applied to each of the bootstrapped samples. They argue that Greenacre's form of partial resampling generates less nuisance variation and that the bootstrap regions can be wrongly centred and too small. Thus, they advocate carrying out a new correspondence analysis on each replicate matrix. They also comment that the possible effects of carrying out a new SVD on each matrix are arbitrary changes in the sign of singular vectors, inversion of the order of singular values and rotation of the plotted co-ordinates. Because it is not possible to avoid the SVD constraints, Milan & Whittaker (1995) propose filtering techniques to minimise their effects and these are explained in Section 5.2.4.

5.2 Assessing Stability by using the Multinomial Distribution

In order to generate replicate matrices to assess stability, as described in 5.1, we need to assume a distribution for the data. For a contingency table we can resample, with replacement, individuals in the sample, noting their original row and column classifications. The data are usually treated either as a series of multinomial distributions, one for each column (or row, depending on which we are primarily interested in), or as a single multinomial distribution for the whole matrix, although a binomial distribution for each cell can also be considered. There are two algorithms appropriate for the two types of multinomial sampling described above and these are briefly explained in 5.2.1 below.

5.2.1 Bootstrap Methods

In this section we describe two bootstrap methods. The first method views the data matrix as a series of separate multinomial samples, one for each column (or row) and the second views the data as a single multinomial sample for the whole matrix. Before implementing a resampling method we need to decide which form of multinomial sampling to use and we can use our knowledge of how the data were collected to help us decide. For example, in archaeology, if the data were originally obtained by collecting samples of a predetermined size from a number of sites, then clearly modelling each site as a separate multinomial sample is most suitable as a resampling method. This is because the sample size is fixed and the counts from one site are independent of the counts from another site. If, however, pottery sherds were collected and subsequently cross-classified into say, fabric and glaze, then modelling the data as a single multinomial sample is the most suitable resampling method, because before collection it was unknown how many sherds would be found and recorded i.e. the total number of sherds in each fabric category or glaze category was not pre-determined. The next two sections briefly explain the two proposed methods.

5.2.1.1 Method One: Separate Multinomial Samples

When we apply bootstrapping to the data matrix, we may decide to compare all columns (or rows) together, treating each column as a simple multinomial sample and allocating each cell an appropriate probability (which depends on its proportion of the column total). Thus, column totals are fixed but row totals can vary; this is equivalent to sampling each column with replacement.

5.2.1.2 Method Two: A Single Multinomial Sample

If, instead, we decide to treat the whole data matrix as a single multinomial sample, then neither row or column totals are fixed, but just the overall matrix sum and each cell is allocated a probability which depends on its proportion of this matrix sum. This is equivalent to sampling the whole matrix with replacement.

5.2.2 Obtaining Bootstrap Co-ordinates by using Partial Resampling

In order to apply Greenacre's method of obtaining new co-ordinates for the replicate matrices (using one of the algorithms just described), we use the following procedure, where it is supposed that the main interest lies in the column co-ordinates. The notation is the same as that in Chapter Two.

Step 1: Carry out a correspondence analysis on the original data matrix. Store the matrix FD_{μ}^{-1}.

Step 2: Carry out bootstrapping on the data matrix using the most appropriate method from 5.2.1 and obtain B replicate matrices.

Step 3: Calculate replicated column profiles, $D_c^{-1*} P^*$, for each of the generated matrices.

Step 4: Apply the relevant transition formula to relate the bootstrapped matrices to the original co-ordinate system. The matrix of projected column co-ordinates is now given by:

$$G^* = D_c^{-1*}P^{*T}FD_\mu^{-1}.$$

G^* contains the principal co-ordinates of the replicate column profiles (the column co-ordinates) and F contains the principal co-ordinates of the row profiles from the original matrix (the row co-ordinates). This is equivalent to each replicate column being projected onto the display as a supplementary column (see 2.2.4).

5.2.2.1 Application to Memphis Pottery Sherds

Using the Memphis pottery sherds, classified according to context and ware and described in 1.2.1, we generate 200 bootstraps from each context using method one and apply Greenacre's partial resampling to obtain Figure 5.1. This is not the most appropriate method to use for these data because, before excavation, it had not been decided that a specific number of sherds from each context would be collected (and indeed this would not have been possible), but for illustrative purposes we compare the two methods in order to examine the effects of making an inappropriate choice. However, because there are so many zero cells in the data matrix, some of the replicate matrices have columns with all zero entries which means that correspondence analysis cannot be applied. We therefore arbitrarily alter one of the cells in these columns from zero to one (see Section 5.3 for a full discussion of our proposals for dealing with zeroes in data matrices).

The aim of applying either bootstrap method is to assess whether the pottery samples obtained from the contexts are really representative of the whole population of pottery (the bigger the cloud, the less representative they are) and whether the contexts are similar in terms of the pottery wares excavated from them (the greater the overlap of clouds, the more similar the contexts). Similar contexts as suggested by overlapping clouds could have implications for the popularity or availability of a particular type of ware (i.e. restricted availability may mean that similar wares were in use for a long period of time and are therefore common to several neighbouring contexts). We should emphasise, however, that our inferences are informal and we have no method of calibrating differences in cloud size: interpretations are based purely on visual display.

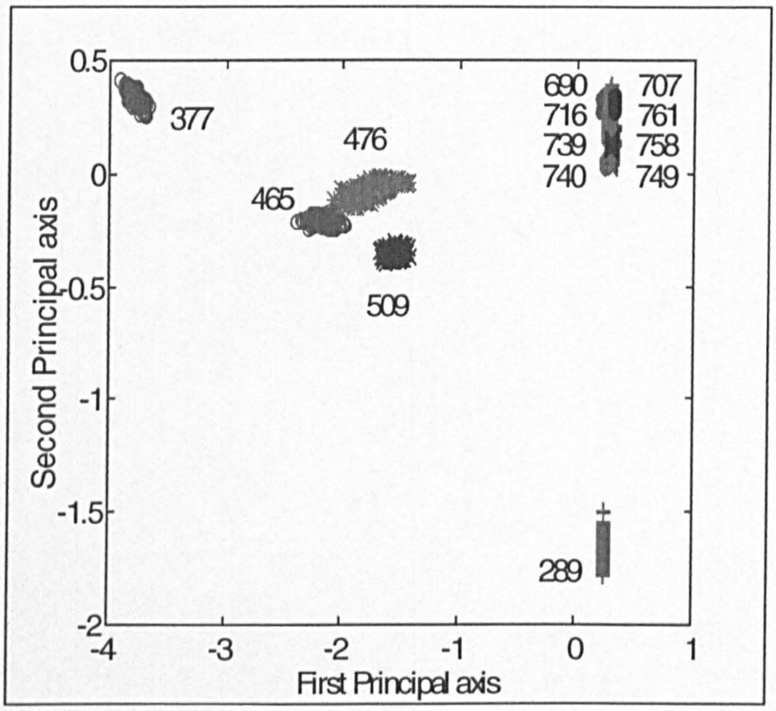

Figure 5.1 Two Hundred Bootstrap Points of Memphis Contexts (Method One)

From Figure 5.1 it is clear that the contexts nearest to the ground surface (see Figure 1.1) — {377, 465, 509, 476, 289} — are well spaced from the remaining contexts which are all bunched together in the top right of the diagram. It is also evident that context 476 contains more variability than the other contexts (because it has a larger bootstrap cloud) and so we are less certain that the pottery from this context is representative of the true population of pottery. Additionally, there is considerable overlap between the contexts in the top right of the picture, suggesting that it is difficult to distinguish between them using the available data, although all their bootstrap clouds are relatively small and they must therefore consist of very similar proportions of wares. Given the practical difficulty sometimes involved in excavations in identifying where one context ends and another begins (and hence the element of arbitrariness in defining contexts), it is interesting to examine the effects of combining those contexts in the top right which are also next to each other in the stratigraphic sequence illustrated in Figure 1.1 and we give this full consideration in Chapter Six.

Generating 200 bootstraps from each context using method two and Greenacre's partial resampling leads to Figure 5.2.

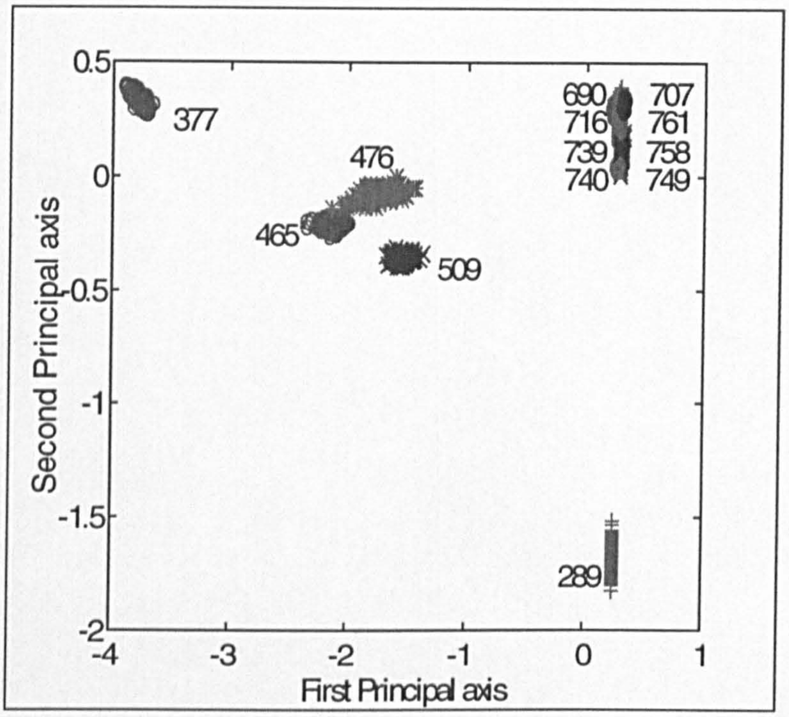

Figure 5.2 Two Hundred Bootstrap Points of Memphis Contexts (Method Two)

There are no obvious visual differences between Figures 5.1 and 5.2, which suggests that the choice of resampling method is not crucial with these data and number of replications. Repeating the process with 1000 bootstraps using method two and Greenacre's resampling produces Figure 5.3. From Figure 5.3 it is clear that the bootstrap clouds are larger than those in Figure 5.2, suggesting that the number of replications is relevant when assessing stability, but this is not really surprising. We believe that the reason for this is because as more and more replicate matrices are generated, the chances of generating more and more unusual matrices increases, probably up to a limiting size of cloud and this is because there are a finite number of replicate matrices which can be obtained by applying the multinomial distribution to the original data.

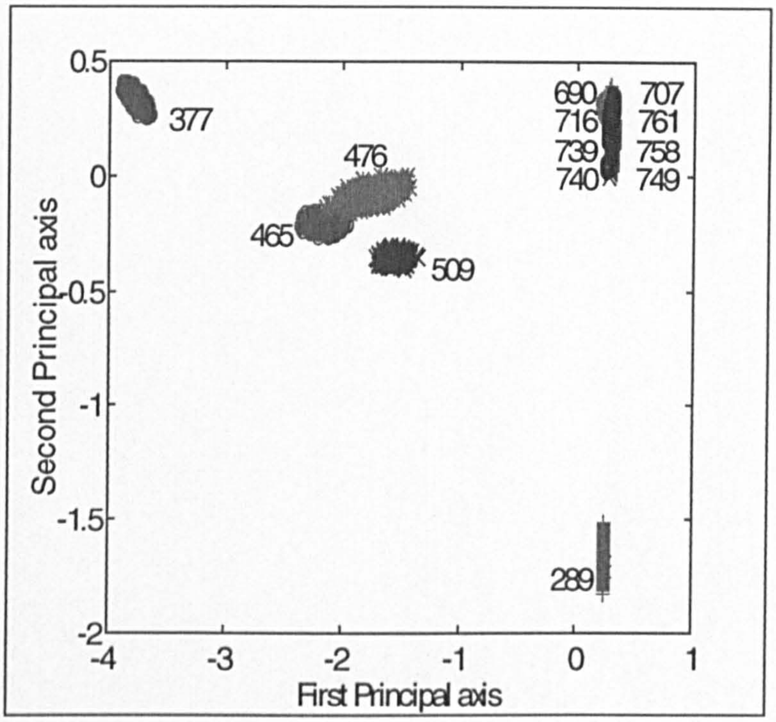

Figure 5.3 One Thousand Bootstrap Points of Memphis Contexts (Method Two)

5.2.2.2 Application to Amarna Pottery Sherds

Using the Amarna pottery sherds, which are classified according to site and ware and which were described in 1.2.2, we generate 100 bootstraps from each site using method one and apply Greenacre's partial resampling to obtain Figure 5.4. We use method one because the sites were originally sampled independently and this is therefore the most appropriate method. Interest lies in assessing how representative the sample at each site is (in terms of pottery) of the true population of pottery at that site and in determining which sites are similar in terms of their distributions of wares. Figure 5.4 reveals that all the sites except 7 and 8 are fairly distinct, which suggests that they contain different proportions of wares. Sites 4, 5 and 12 are the most variable because they have the largest clouds and we are therefore less certain that the samples from these sites are representative of the true population of data. There is also a possible 'arch effect' (see 2.4.2.2).

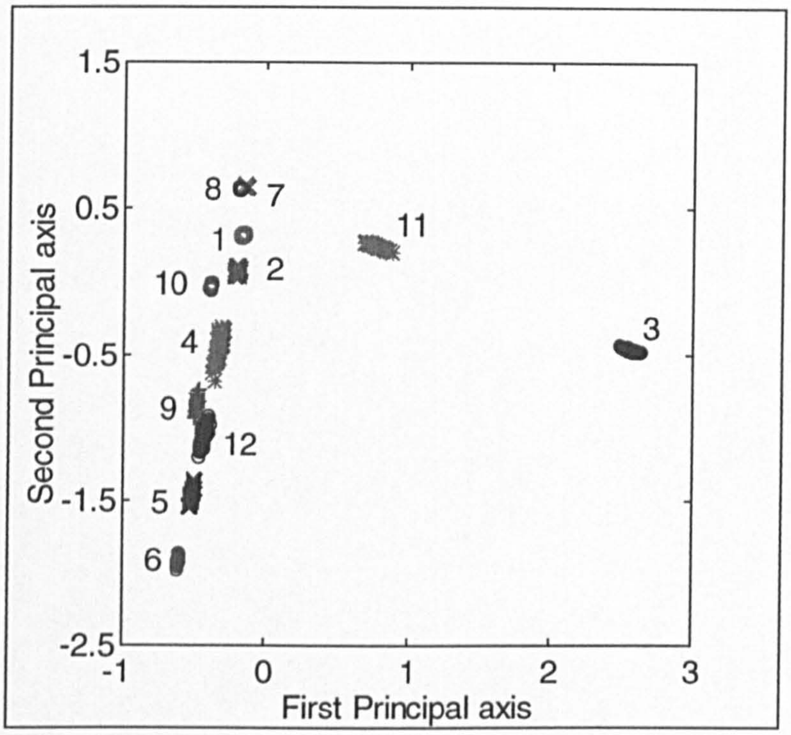

Figure 5.4 One Hundred Bootstrap Points of Amarna Sites (Method One: Greenacre)

If we generate 100 bootstraps from each context using method two (which is not strictly appropriate, but it is interesting to compare it with method one) then we obtain an almost identical picture, suggesting that there is little difference between the two methods for these data and number of replications. Using greater numbers of bootstraps the clouds become larger and with 1000 bootstraps sites 9 and 12 overlap using both methods. Generating 1000 bootstraps using method two leads to the clouds for sites 5 and 12 touching each other. This appears to suggest that with 1000 bootstraps choosing the wrong method has some effect on which sites overlap, although our inferences based on any overlapping clouds are still informal. Because the overlapping nature of the clouds appears to depend on the number of bootstraps, we believe that these cannot be used directly to ascertain the stability of the display, nor to assess the similarity between row or column categories. A trimmed measure is therefore needed and we introduce possible measures in Sections 5.4.4 and 5.4.5.

5.2.3 Obtaining Bootstrap Co-ordinates by using Milan & Whittaker's Method

We can also apply Milan & Whittaker's method of obtaining row and column co-ordinates, which involves carrying out a new correspondence analysis on each replicate matrix (see 5.1.2.2) and this is illustrated in Section 5.2.4.4 below. One problem with implementing correspondence analysis on each replicate matrix, which is not present in Greenacre's method, is the arbitrary sign of eigenvectors obtained from the singular value decomposition (and hence the arbitrary sign of the row and column co-ordinates which are based on these eigenvectors) which is part of the correspondence analysis (see Chapter Two). Another difference between the methods of Greenacre and Milan & Whittaker is that because a new correspondence analysis is carried out on each replicate matrix under the latter method, the co-ordinates of each matrix are relative to different axes. We must therefore decide which method we believe is best and so we implement that due to Milan & Whittaker in order to compare them.

5.2.4 Filtering Techniques

Correspondence analysis involves the singular value decomposition of a matrix (see Chapter Two), although associated with this are standard orthogonality conditions. Possible effects of the singular value decomposition are:

- Arbitrary changes in the sign of the singular vectors.
- Inversion of the order of the singular values.
- Rotation of the plotted co-ordinates.

However, such effects only become apparent when more than one set of co-ordinates is to be displayed. Milan & Whittaker (1995) propose 'filtering' techniques to avoid these problems which they say are a result of the resampling and they define the Frobenius norm for the difference between two matrices, U (n × m) and V (n × m), by:

$$\|U - V\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(u_{ij} - v_{ij})^2.$$

Using the notation of Milan & Whittaker (1995), if w_o is the matrix of original co-ordinates (known as the reference) and $w^{[r]}$ is the matrix of co-ordinates from the r-th bootstrap, then for each bootstrap a set:

$$W = \{w_k^{[r]};\ K\in k\}$$

is identified and compared with the reference. The three filtering techniques are described below.

5.2.4.1 Reflection

To minimise the effect of arbitrary reflection we apply all possible reflections to the new set of co-ordinates, compare these with the reference set and take the closest to be the new co-ordinates. For each bootstrap, $w^{[r]}$, each of the possible reflections of the co-ordinates from the set:

$$W^R = \{w_k^{[r]};\ w_k^{[r]} = w^{[r]}R_k,\ K\in k\}$$

are compared with w_o. The identified set of co-ordinates are those which minimise:

$$\left\|w_o - w_k^{[r]}\right\|_F^2 \text{ over } w_k^{[r]} \in W^R. \qquad (5.1)$$

In the two-dimensional case, the R_k are given by:

$$R_1 = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, R_2 = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, R_3 = \begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix}, R_4 = \begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix} \qquad (5.2)$$

5.2.4.2 Reordering

During the simulation process two singular vectors may change order. The effect of inversion of the order of singular values and vectors is minimised by minimising expression (5.1) over W^o, instead of over W^R, where W^o is the set of combinations of all possible reflections and all possible inversions of order. The points whose co-ordinates have the least distance to the reference among all possible combinations of reflections and order inversions are considered to be the new co-ordinates and in two dimensions there are only two possible permutations of point co-ordinates. Each element from the set of all combinations of reflections and inversion of order:

$$W^{\circ} = \{w^{[r]}O_k; K \in k\}$$

is obtained by multiplying $w^{[r]}$ by one matrix R_i ($i=1,2,3,4$) from the matrices displayed in (5.2) and one matrix (j=1,2) from (5.3). We therefore obtain , where:

(5.3)

5.2.4.3 Rotation

A further filtering is obtained by rotating the co-ordinates. The identified points are , where the orthogonal matrix rotates the points to the closest position to the reference set. The technique used to select the best rotation is called orthogonal procrustes and can be described by:

(5.4)

subject to $Q^TQ = I_m$, where Q rotates the points to the position closest position to w_o in a least squares sense. The matrix , the solution to (5.4), is given by where U and V are obtained through the SVD:

$$(w^{[r]})^T w_o = UD_mV^T$$

and where D_m= diag ($\lambda_1,\ldots, \lambda_m$) is a diagonal matrix of singular values.

5.2.4.4 Filtering Applied to Amarna Pottery Sherds

In this section we apply the reflection and reordering of Sections 5.2.4.1 and 5.2.4.2 respectively to the Amarna sherds (1.2.2), having generated 100 bootstraps using method one. It turns out, however, that the resulting plots are the same for both these types of filtering. We do not believe that it is necessary to consider rotation because there is no translation of the co-ordinates as a result of the SVD and procrustes rotation may therefore 'overcorrect' for nuisance variation that does not really exist.

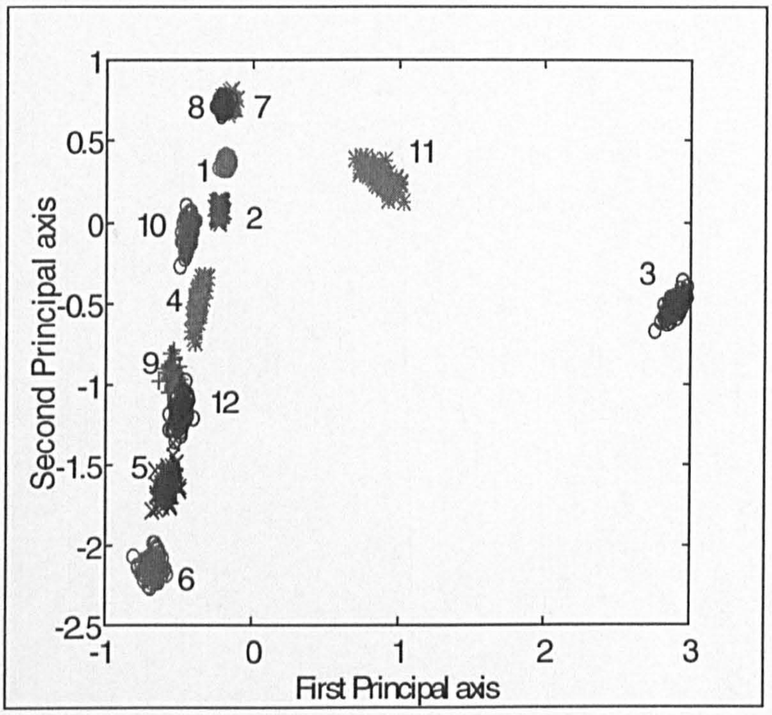

Figure 5.5 One Hundred Bootstrap Points of Amarna Sites (Method One: Filtering)

Comparing Figure 5.5 (filtering) with Figure 5.4 (partial resampling), it is evident that the bootstrap clouds are much larger under filtering (as suggested by Milan & Whittaker, 1995). Thus, if we use filtering rather than the transition formulae, then we conclude that the data are less representative of the true population of data. The relative stability of each site is also different compared with Greenacre's resampling: sites 4 and 12 no longer have larger clouds as compared with the other sites. Generating 1000 replicates of the Amarna sherds, the bootstrap clouds are much larger than those from 100 bootstraps and so clearly we cannot use size of bootstrap cloud as a measure of stability, but we discuss this problem further in 5.4.5. We appreciate the arguments for both resampling and filtering, but we mainly focus on Greenacre's partial resampling method in this chapter, although filtering is used in Section 5.5.

5.3 Sparse Contingency Tables

Sometimes zeroes occur in the data matrix, for example with the Memphis sherd data listed in Table A.1 of the Appendix. This can be a problem when generating replicate matrices based on the multinomial distribution, because each zero cell will be allocated zero probability. However, if it is clear that zero counts occurred because of an absence in the population from which the sample was taken of that particular artefact (these are called essential zeroes or absolute zeroes) then the problem is not serious. If, on the other hand, the zero counts occurred because the sampling technique was not adequate to detect rare artefacts (these are called trace zeroes) then it may be advisable to adjust the bootstrapping procedure to account for this.

It is not always possible, however, when faced with a data set, to determine which type of zero is present. If, for example, all non-zero cells contain counts of several thousand, then zero cells may well indicate essential zeroes. But, if cells contain small numbers of say, less than ten, then the nature of the zeroes may not be clear. In principal, increasing the sample size can eliminate trace zeroes, but in practice this is too costly. We therefore propose adjusting the probabilities assigned by the bootstrapping algorithm to each cell in order to account for trace zeroes (i.e. use a 'smoothed bootstrap') and thus the generated matrices may then contain non-zero counts in those cells which previously contained zeroes.

We propose two methods of adjusting the probabilities assigned to cells which contain trace zeroes, both of which use the binomial distribution and this is because we are considering the cell with trace zero or not. The methodology that we have developed is introduced in Section 5.3.1.

5.3.1 Methodology for Adjusting for Trace Zeroes

In this section we introduce two methods of adjusting for trace zeroes, which we refer to as A1 and A2.

Method A1: For cells containing trace zeroes, we take the upper one-sided (1-α)% confidence limit for the cell probability (based on the observed zero) and we use this as the multinomial probability when generating replicate matrices. It is now possible to obtain a non-zero count in a cell that previously contained a zero. It is easily seen that this upper (1-α)% confidence limit for the cell probability, given the observation zero in n trials, is $1-\alpha^{\frac{1}{n}}$. We can interpret this as the highest value for the cell probability that is consistent with the observed zero. We could take α to be the conventional value of 0.05 but perhaps a higher value is preferable — taking $\alpha = 0.5$ gives the smallest value for the cell probability where we are 'more certain than not' that it is consistent with the trace zero.

Method A2: For the cell containing a trace zero, take the expected count for the cell to be no lower than a certain value z (for example z = 0.5). Because, for the binomial distribution, the expected value of a cell is given by np, we have:

$$np > z$$

$$\Rightarrow \qquad p > \frac{z}{n}.$$

Here, we take $p = \frac{z}{n}$ to be the cell probability used in generating replicate matrices. Taking z = 0.5 in particular has the informal interpretation that this is the smallest value for the cell probability where the expected cell count would 'just avoid being rounded down to a [trace] zero'. Whatever value we impute for the cell probability corresponding to a trace zero, we need to ensure that it is at least large enough in relation to the number of bootstraps performed to ensure that a reasonable number of bootstrap matrices do occur with non-zero frequencies in those cells.

Having introduced these methods of accounting for trace zeroes, in the next few sections we illustrate and adapt A1 to account for:

- the number of trace zeroes.
- whether method one or two of 5.2.1 is appropriate for resampling from the data.

5.3.1.1 One Trace Zero for Column Comparisons

When resampling from each column separately, we propose the following methods of accounting for trace zeroes (B1 and B2). When one zero occurs in a column, the probability assigned to that zero cell is calculated as in A1 above and the initial probability assigned to each non-zero cell is obtained by dividing the cell value by the column total. However, how these probabilities are adjusted to account for the non-zero probability assigned to the zero cell is open to discussion and we propose two ways:

Method B1: Divide the calculated probability for the zero cell by the number of non-zero cells in the column. Subtract this value from the initial probabilities obtained for each of these non-zero cells. The sum of the probabilities allocated to each cell in the column should then equal one.

Method B2: For each non-zero cell, multiply the initial probability assigned to the cell by the calculated probability for the zero cell and subtract this value from the initial probability.

These two methods are illustrated in a simple example.

Example

Suppose the data matrix is as follows:

$$\begin{bmatrix} 0 & 2 & 6 \\ 4 & 3 & 9 \\ 6 & 8 & 12 \end{bmatrix}.$$

Taking $\alpha = 0.2$ we have, for the first column:

$$p_1 = 0.15$$

$$p_2 = 0.40$$

$$p_3 = 0.60$$

where p_i $(i=1,2,3)$ is the probability assigned to each cell in a particular column. Calculating the adjusted probabilities produces:

Method B1	**Method B2**
$p_2 = 0.40 - \dfrac{0.15}{2} = 0.33$	$p_2 = 0.40 - (0.40 * 0.15) = 0.34$
$p_3 = 0.60 - \dfrac{0.15}{2} = 0.53$	$p_3 = 0.60 - (0.60 * 0.15) = 0.51.$

We prefer method B2 because it accounts for the relative magnitude of the cells with non-zero entries. In the next section we consider the whole matrix.

5.3.1.2 One Trace Zero for the Whole Matrix

When considering resampling from the matrix as a whole, we need to consider the total number of zero cells in the matrix. If there is only one zero cell then the probability assigned to that cell is calculated as in the beginning of Section 5.3.1 and the initial probability assigned to each non-zero cell is obtained by dividing the cell value by the matrix total. However, how these probabilities are adjusted to account for the zero cell is again open to discussion, although we can apply methods B1 and B2 introduced in 5.3.1.1 to the whole matrix rather than to each column. Of course, the adjusted probabilities of the non-zero cells will be very little different to the unadjusted probabilities for the case of only one trace zero.

5.3.1.3 Two or More Trace Zeroes for Column Comparisons

Sometimes, two or more trace zeroes occur in a column. We first consider the case of two zeroes. The probability allocated to a zero cell is calculated as before and we propose that either one of the following methods is chosen to adjust this value:

Method C1: The probability is divided equally amongst the number of zero cells in the column.

Method C2: The probability is divided amongst the number of zero cells in the column proportionally, by conditioning on row totals.

The probabilities assigned to the non-zero cells can then be adjusted and we can also apply similar methodology to the case of more than two zeroes in a column. However, even with these 'sparse algorithms' there is no guarantee that a generated column will contain an entry other than zero, which causes problems for correspondence analysis. A value of one can then be placed in a cell chosen at random from within the column consisting only of zeroes. For the case of two or more trace zeroes in the whole matrix, similar methodology can be applied.

5.3.2 Application to Amarna Pottery Sherds

In order to illustrate how bootstrapping can be adapted to account for trace zeroes, method A1 and either B2 or C2, whichever is appropriate, is applied to the Amarna sherds, taking α=0.25, generating 100 bootstraps and using method one. Greenacre's partial resampling method is also applied and it is necessary to assume that all zeroes are trace zeroes (which may or may not be realistic). Comparing the resulting figure with Figure 5.4 there is no visual difference and this is because the probabilities assigned to each zero cell are still extremely small which, in turn, is because the total sample sizes obtained from each site are reasonably large. Our experience reveals that even with 'small' sample sizes, the probabilities assigned to the zero cells are too small to make any real difference to the bootstrap clouds and so it is probably advisable to apply the multinomial distribution directly to the original data, without adjustment.

5.4 Convex Hulls, Peeling and Ellipses

In this section we describe the convex hull, which we use as a method of summarising the clouds of points resulting from bootstrapping. Each set of replicate points, regardless of the method used to obtain them, can be enclosed by a convex hull of the points which connects the outermost points of each set.

Convex hull peeling (Green, 1981) involves constructing the convex hull of the data, deleting it and then constructing the convex hull of the remaining points. This procedure may be repeated until no points are left and in the bivariate case, the successive shells so formed are called the convex hull peels of the data. The Green-Silverman peeling routine, due to Green & Silverman (1979), is one of a number of peeling algorithms. This routine, however, makes an attempt to deal with degeneracies caused by rounding errors, which other methods tend to ignore and is the algorithm used throughout the work below.

Rather than displaying all the bootstrap points, just the convex hulls of the clouds are usually shown and non-overlapping hulls are taken to indicate that differences exist between row or between column categories. (Alternatively, concentration ellipses at a given probability level can be drawn and again, non-overlapping ellipses indicate that differences exist between categories.) Ringrose (1992) comments that the points which have the greatest spread in the bootstrap display are those with similar numbers in all their cells and low frequencies, while the reverse will give a smaller group. A large spread of points can also be due to the category being poorly represented in the given dimensions. If categories overlap then we conclude that they are hard to distinguish on statistical grounds and categories that remain separate after many bootstraps have been generated are unlikely to have the same profiles across rows. It should be remembered that it is not appropriate to compare clouds of row points with clouds of column points because there is no definition of distance between columns and rows in correspondence analysis (see 2.3.1).

Ringrose (1992) carried out a simulation study in order to investigate the reliability of bootstrap confidence regions, where the idea was to investigate the significance level they represent. Ringrose commented that the more bootstrap replications that are used in the construction of the hulls, the larger they are likely to be and so the greater the probability of overlaps. He noted that the 100% hulls are likely to be the most affected by differences in the number of bootstrap replications and so it might be more useful to use the 90% hulls instead. Ringrose also noted that the hull overlap rate for 1000 replications is 1.5 times greater than that for 100 replications and this is something we address in the next few sections.

5.4.1 Application to Memphis Pottery Sherds

Having generated 200 bootstraps using method two and applied Greenacre's partial resampling, the outer convex hulls for the Memphis contexts (1.2.1) are shown in Figure 5.6.

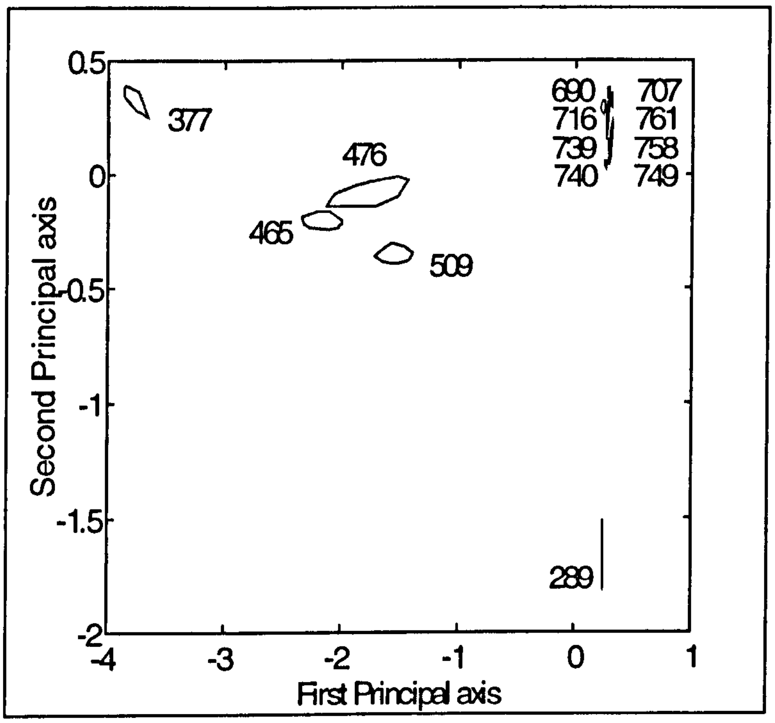

Figure 5.6 Outer Convex Hulls of Memphis Contexts

As explained in Ringrose (1992), the context with the greatest spread (476) has similar numbers across the wares and low numbers in each cell. In fact, context 476 has a larger hull than all of the top right contexts put together and one reason for this could be because the period of occupation that it corresponds to is greater than for the other contexts (and it therefore has a greater variety of wares), or because greater changes in pottery typology were occurring during the phase represented by this context.

5.4.2 Application to Amarna Pottery Sherds

Having generated 100 bootstraps using method one and applied Greenacre's resampling, Figure 5.7 illustrates convex hull peeling for the Amarna sherds (1.2.2) for site 12.

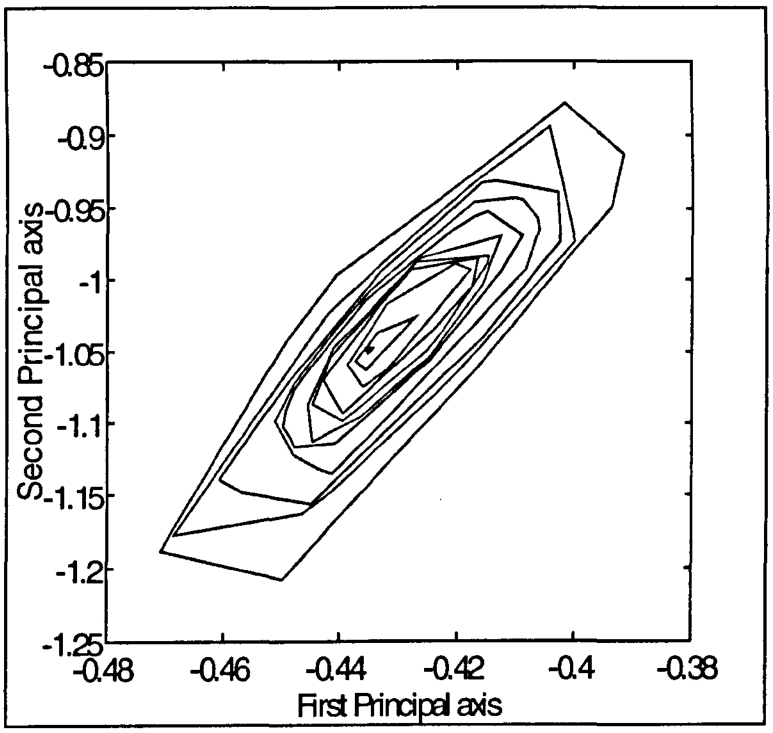

Figure 5.7 Convex Hull Peels of Amarna Site 12

We see that there are 11 hull peels in total and that the two outer hulls are some distance away from the remaining hulls. It is also clear that taking the outer hull as compared with the hull containing approximately 50% of the points, leads to a very different estimate of site stability. We address this problem in the next section.

5.4.3 Multivariate Summaries

Seheult, Diggle and Evans (1976) suggested that convex hull peels could be used to define a median and interquartile set, with the median point being taken as the centroid of the innermost convex hull. Green (1981) used these summaries, with the outermost convex hull, as a bivariate analogue of the box-and-whisker plot. Alternative ideas have included the vector of marginal medians, which ignores the bivariate structure of the data and the mediancentre, which is that point from which the aggregate distance to the data points is minimised (Gower, 1974).

Because one of our objectives is to assess the stability of the displays, we have, in Section 5.2, suggested various resampling methods. However, as we saw in 5.2.2, the size of the resulting clouds depends on the number of bootstraps generated and we must therefore develop confidence regions to account for this. Alternative suggestions for summarising two-dimensional data are discussed below, although they are easily extended to higher dimensions.

5.4.3.1 Alternative Methods

Problems arise with comparing summaries of data based on different numbers of bootstraps because, as we saw in 5.2.2, the size of a bootstrap cloud increases as the number of bootstraps increases. When obtaining confidence intervals in one dimension, it is often a 95% interval that is obtained and ideally we would like to investigate the equivalent in two dimensions. However, in the case of bootstrap clouds, the size of the hull which contains closest to 95% of the points is still extremely dependent on the number of bootstraps, as is the 75% hull and we therefore need to consider either a hull based on a much smaller number of points (say 50%), or develop a method for adjusting the hull according to the number of bootstraps. Depending on the number of bootstraps, we observe that it is often the case that the hull containing closest to e.g. 75% of the points actually contains anything between approximately 60%-90% of the points (because there is no hull that contains close to 75% of the points). At first we considered the idea of ranking the hulls and taking, say, the third peel as a summary measure, but in reality the number of hull peels vary considerably with the value of ε in the Green-Silverman routine (Green & Silverman, 1979) and also with the particular data set being analysed and so this is not sensible.

We believe that one relatively stable choice of obtaining similar sized hulls, regardless of the number of bootstraps, is to plot those hulls containing closest to 25% and closest to 75% of the points for 100 bootstraps. This can be thought of as being analogous to the interquartile range in one dimension. However, because the cloud size increases with the number of bootstraps, we need to adjust the region to account for this. Going from 100 bootstraps to 1000 increases the 'size' of the outer hull by approximately 50% (see Table 5.1) and going from 100 to 5000 bootstraps increases the 'size' of the outer hull by roughly 100%. Sections 5.4.4 and 5.4.5 explain the summary measures in more detail.

5.4.4 Measuring Stability by Area

In this section we propose investigating informally the differences in category stability as assessed by the methods of Greenacre and Milan & Whittaker. We do this by calculating the areas of the convex hulls and concentration ellipses of the bootstrap points. Stability is visually assessed by hull size and so by calculating the area of the hulls we believe that the distortion which sometimes occurs when plots are displayed in less than full dimensionality and when computer packages are used, can be avoided. We also investigate the effect of the number of bootstraps on convex hull areas.

We believe that two methods discussed by Jennrich & Turner (1969) in the context of animal home range have direct applicability to measuring category stability. We describe these methods below.

5.4.4.1 Area of a Convex Polygon

Given a set of points, we can draw the smallest convex polygon which contains all the points (or, say, 75% of them) and take:

$$A_1 = (\text{area of the convex polygon})$$
$$= \frac{1}{2}\sum_{i=1}^{n}\left(x_i y_{i+1} - x_{i+1} y_i\right)$$

as an index of the variability of points, where (x_i,y_i) is the i-th ordered point out of n moving in an anticlockwise direction on the convex hull and where $(x_{n+1},y_{n+1}) = (x_1,y_1)$.

5.4.4.2 Area of a Concentration Ellipse

Jennrich & Turner (1969) developed an index that measures non-circular as well as circular clusters of points. Let:

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix}$$

be the mean vector and variance-covariance matrix respectively of the bivariate normal distribution. The regions of the most intense numbers of points are shown to be bounded by concentric, constant density ellipses of the form:

$$(z-\mu)^T\Sigma^{-1}(z-\mu) = \text{constant}$$

where z denotes an arbitrary point on the ellipse. An ellipse of this form that accounts for a proportion p of the total number of points has an area given by:

$$a = \pi \ln(1-p)^{-2}|\Sigma|^{\frac{1}{2}}.$$

By setting $p = 1-e^{-3} = 0.95$, this simplifies to:

$$a = 6\pi|\Sigma|^{\frac{1}{2}}. \qquad (5.5)$$

Equation (5.5) is the definition of variability. It is the area of the smallest region that accounts for 95% of the total number of points and is estimated by the statistic:

$$A_4 = 6\pi|S|^{\frac{1}{2}}.$$

Here, |S| is the determinant of the sample variance-covariance matrix.

To use the above method we must assume that our bootstrapped points can be described by a bivariate normal distribution. The variability of the bootstrap points can then be thought of as the area of the smallest sub-region that accounts for a specified proportion, p, of its total area.

5.4.4.3 Application to Amarna Pottery Sherds

Considering the Amarna sherd data (1.2.2), varying numbers of bootstraps are generated (100, 1000 and 5000), using method one and Greenacre's partial resampling. The measures A_1 and A_4 are calculated and are shown in Table 5.1, although the measures themselves cannot be compared.

Table 5.1 Stability Measures according to the Number of Bootstraps

	A_1 (× 1000)			A_4 (× 10)		
Site	**100**	**1000**	**5000**	**100**	**1000**	**5000**
1	0.921	1.570	1.987	0.011	0.011	0.010
2	4.204	6.460	8.446	0.039	0.037	0.041
3	1.684	3.128	3.968	0.018	0.019	0.019
4	15.373	24.010	33.663	0.149	0.147	0.150
5	3.108	4.376	6.227	0.031	0.031	0.031
6	0.305	0.408	0.558	0.003	0.003	0.003
7	0.554	0.884	1.304	0.007	0.007	0.006
8	0.286	0.406	0.545	0.003	0.003	0.003
9	3.031	5.193	7.447	0.036	0.034	0.035
10	0.335	0.508	0.711	0.003	0.003	0.003
11	1.792	2.764	3.211	0.017	0.018	0.018
12	9.530	16.591	24.799	0.115	0.116	0.115

It is clear from reading down the columns of the above table that sites 4 and 12 have the largest bootstrap clouds. It is also evident that in contrast to measure A_1, measure A_4 does not appear to depend on the number of bootstraps. Figure 5.8 displays 95% ellipses for the 12 Amarna sites based on 100 replicate matrices.

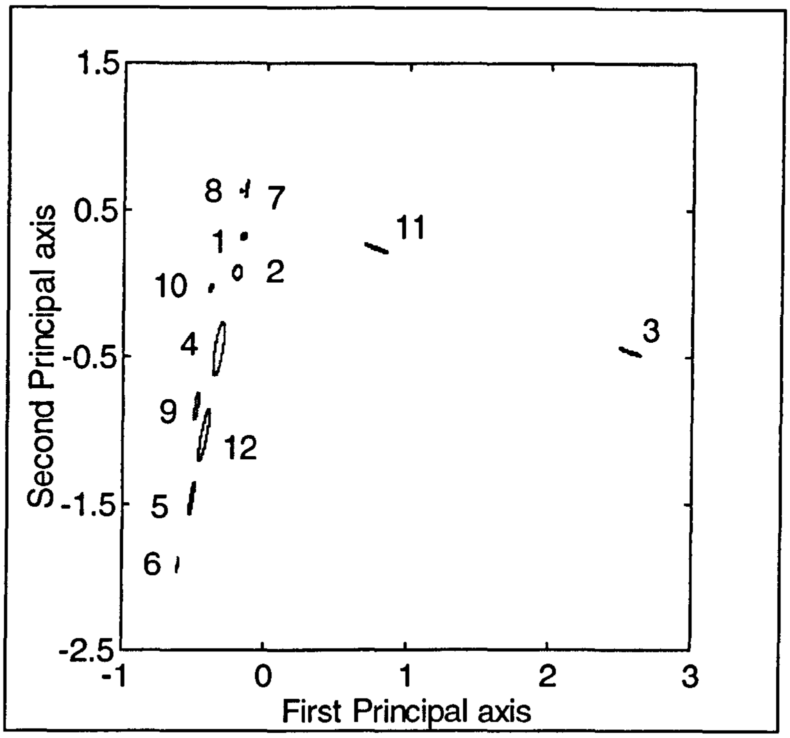

Figure 5.8 95% Concentration Ellipses for the Amarna Sites

It is evident from the above figure that none of the ellipses overlap and so we conclude that all the sites are distinct with regard to the wares that they contain (again, this is an informal inference). However, we should consider at what degree of overlap we would no longer view sites as being distinct. We suggest that if the centroid of a 95% ellipse representing one site is included in the 95% ellipse of another site, then these sites can be considered to be virtually indistinct in terms of their profiles of wares.

5.4.5 Application of Multivariate Summaries to Amarna Pottery Sherds

In this section we use method one to generate 100 bootstraps from each Amarna site, before applying Greenacre's partial resampling method and implementing convex hull peeling on the resulting site co-ordinates. Calculating the area of each peel and, if necessary, interpolating between peels, we find the areas of the peels containing 25%, 75%, 95% and 100% of the bootstrap points. We then find the percentages of points needed from 1000 and 5000 bootstraps in order to obtain these same areas. Because the percentages will vary for each site, we obtain ranges of percentages, based on the lowest and highest values across all sites and these are shown in Table 5.2.

Table 5.2 Approximate Percentages of Points in the Hulls

	Number of Points in the Bootstrap		
	100	**1000**	**5000**
% Hull	25%	13.8-26.5%	12.0-24.4%
	75%	58.6-75.5%	56.3-72.1%
	95%	85.6-95.8%	85.1-93.6%
	100%	92.9-98.2%	92.4-96.9%

From Table 5.2 we see that as the number of bootstraps increases, the percentages of points in the hulls corresponding to those of 100 bootstraps decreases i.e. the hulls containing say, 75% of the points for 100, 1000 and 5000 bootstraps vary in size considerably.

Figure 5.9 shows, for site 1, the hulls closest to those containing 25% and 75% of the points for 100 bootstraps and also those hulls for 1000 and 5000 bootstraps with areas closest to these. Sometimes, there is no hull that contains close to 25% or 75% of the points.

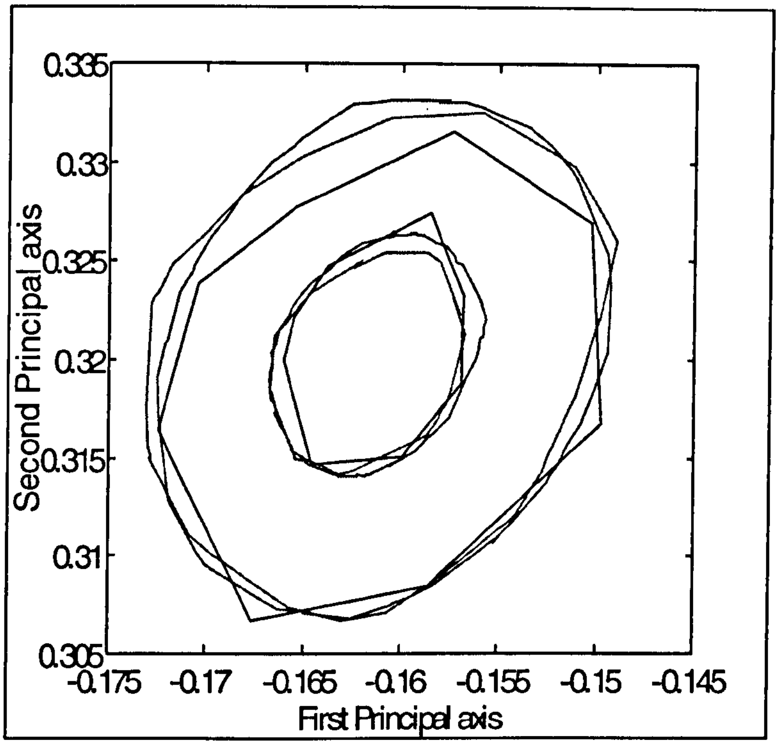

Figure 5.9 25% and 75% Hulls of Amarna Site 1

It is clear from Figure 5.9 that the hulls are fairly similar and that our proposed method is a good means of adjusting for the number of bootstraps.

5.5 Assessing Stability by using a Jack-knife Approach

So far in this chapter we have used the multinomial distribution to help us to assess the stability of the categories in a correspondence analysis map. We now introduce an alternative method, based on the jack-knife technique. We propose that each column (or row) category of a contingency table is deleted in turn, correspondence analysis is implemented on the reduced matrices and the size of the resulting cloud of row (or column) points is then examined. However, because the dimensions of the data matrix are reduced each time a column is deleted, Greenacre's method of partial resampling cannot be applied. It is, therefore, necessary to implement filtering and so the 'jack-knife clouds' should be compared with those clouds obtained from using multinomial sampling with filtering. We believe that this method can also be used as a means of detecting influential categories and this is explained in Chapter Six.

5.5.1 Application to Amarna Pottery Sherds

In this section we apply the jack-knife method introduced above to the Amarna sherd data. We omit each ware in turn, implement correspondence analysis and display the site points in Figure 5.10 (there are arbitrary reflections as compared with Figure 5.4 — see 5.2.4), where there are 10 points for each site, each corresponding to a deleted ware. We see that the clouds are generally slightly larger under jack-knifing as compared with bootstrapping (although this will depend on the number of bootstraps generated and the method of obtaining replicate co-ordinates), which is probably due to particularly influential wares (see 6.8). It may not, therefore, be sensible to produce hulls, ellipses or other summary measures for each site, when such influential wares exist (although this depends on the archaeological importance of the wares). One advantage of a jack-knife approach is that it gives us a benchmark with which to compare the results of other methods of assessing stability.

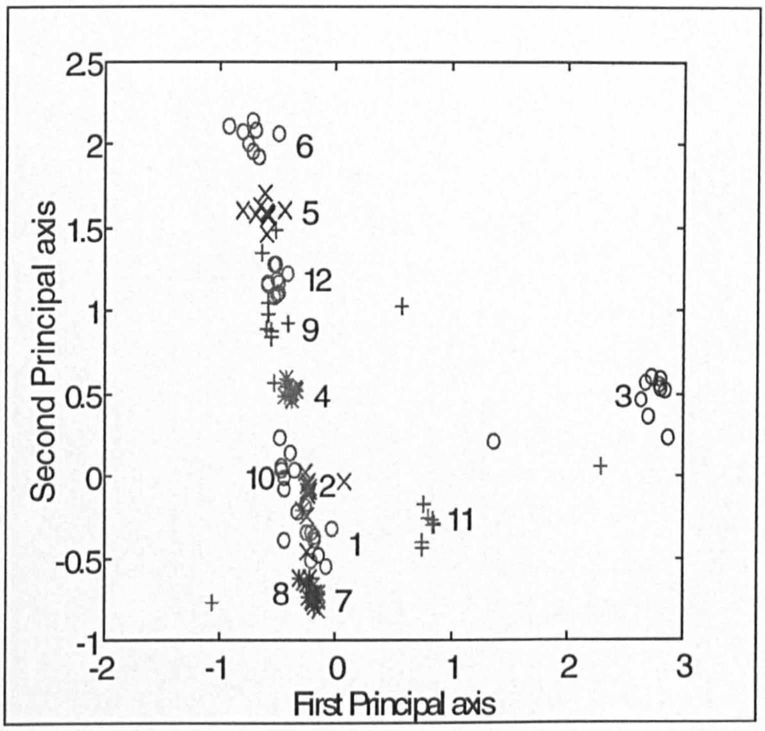

Figure 5.10 Amarna Site Clouds (Jack-knifing)

5.6 The Influence of Sample Size

So far we have proposed assessing stability of the correspondence analysis map by resampling using the multinomial distribution or by jack-knifing. We now compare the results from the first method with those from 'sampling without replacement'. This is really sampling using the hypergeometric distribution (multinomial sampling is the same as sampling with replacement).

Using sampling without replacement, observations can only be retained at most once and so this form of resampling is only suitable for answering questions concerning smaller sample sizes than those actually obtained. We suggest using sampling with and without replacement to assess some of the questions posed in Chapter Two, namely, the influence of sample size on both the relationship between row and column categories and on the stability of the categories in the correspondence analysis map.

If we take many samples of size $h < n$, without replacement, where n is either the original matrix sum, or the sum of a particular category (depending on how the data were originally collected), then this enables us to evaluate the relationship between row and column categories as if we had originally taken a sample of size h. We can also investigate whether a particular smaller sample is representative of the true population of data (we do this by obtaining one sample without replacement and then resampling this with replacement). We apply these suggestions to the Amarna sherds in the following section.

5.6.1 Application to Amarna Pottery Sherds

The two types of resampling so far discussed — using the multinomial distribution and sampling without replacement — are illustrated below for the Amarna sherds (1.2.2). We generate 100 bootstraps for a series of sample sizes that consist of varying proportions of the original numbers of sherds obtained from each site. Because the measure A_4 does not appear to vary with the number of bootstraps (see Table 5.1), we use this to assess site stability.

5.6.1.1 Sampling by using the Multinomial Distribution

Using method one from 5.2.1.1 and Greenacre's partial resampling, we calculate the measure A_4 for each of the Amarna sites: the results are shown in Table 5.3. The numbers of sherds generated from each site vary according to the column headings of the table (where 2 = double the original number etc.).

Table 5.3 The Measure A_4 (×10) for Varying Sample Sizes (Multinomial Distribution)

	Sample Size (proportion of original)					
Site	**2**	**1**	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
1	0.005	0.011	0.015	0.019	0.035	0.078
2	0.020	0.039	0.049	0.090	0.151	0.322
3	0.010	0.018	0.026	0.036	0.064	0.135
4	0.071	0.149	0.216	0.333	0.601	1.413
5	0.014	0.031	0.039	0.059	0.093	0.172
6	0.001	0.003	0.003	0.005	0.010	0.000
7	0.003	0.007	0.010	0.011	0.025	0.044
8	0.001	0.003	0.004	0.005	0.012	0.023
9	0.018	0.036	0.041	0.068	0.139	0.322
10	0.002	0.003	0.004	0.008	0.012	0.031
11	0.008	0.017	0.026	0.035	0.077	0.082
12	0.059	0.115	0.149	0.207	0.407	0.917

Reading from left to right across the table, we see that the smaller the sample size the larger the value of A_4 and so the less stable the site i.e. the less confident we are that the sample collected at the site is representative of the true population of wares. The corresponding 95% ellipses for each sample size are illustrated in Figure 5.11 for site 12 and we see that the ellipses are not quite concentric, although we believe that this is due to the inherent variation when using bootstrapping. The smallest ellipse corresponds to a sample of size $\frac{1}{8}$th of the original and the largest ellipse to double the original sample size. Clearly, sample size is very influential in our interpretation of site similarity (i.e. with smaller samples we obtain larger ellipses and greater numbers of site overlaps).

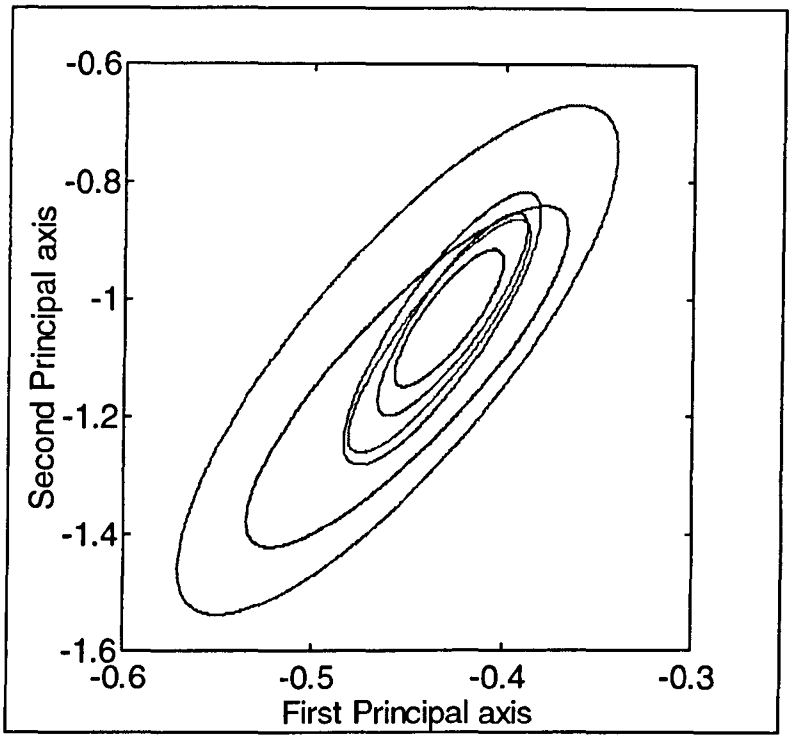

Figure 5.11 95% Concentration Ellipses for Amarna Site 12 (Varying Sample Sizes)

5.6.1.2 Sampling Without Replacement

In this section we use sampling without replacement in order to assess the effects on the correspondence analysis display of smaller samples than that actually obtained. We sample different proportions of the original sherds collected at each Amarna site, without replacement, 100 times and calculate A_4 for each site. The values are shown in Table 5.4.

Table 5.4 The Measure A_4 (×10) for Varying Sample Sizes (Without Replacement)

	Sample Size (proportion of original)			
Site	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
1	0.003	0.013	0.033	0.086
2	0.014	0.036	0.139	0.290
3	0.006	0.022	0.048	0.148
4	0.051	0.139	0.528	0.890
5	0.010	0.032	0.100	0.208
6	0.001	0.003	0.010	0.016
7	0.002	0.006	0.017	0.045
8	0.001	0.003	0.008	0.018
9	0.011	0.037	0.098	0.255
10	0.001	0.003	0.008	0.024
11	0.006	0.020	0.060	0.121
12	0.033	0.123	0.337	0.808

Reading from left to right across the table, it is clear that as was the case for multinomial sampling, the smaller the sample size, the more unstable the site, although sampling from the multinomial distribution produces larger values of A_4 than sampling without replacement.

5.6.1.3 Stability of a Particular (Smaller) Sample

Here, we use sampling without replacement to generate a smaller sample than that actually obtained, before generating 100 bootstraps using multinomial sampling to assess the stability of this particular smaller sample. Table 5.5 displays the values of A_4 and it is evident that smaller sample sizes lead to greater instability of the sites. As expected, the table contains similar values to those in Table 5.3.

Table 5.5 The Measure A_4 (×10) for a Particular Sample for Varying Sample Sizes

	Sample Size (proportion of original)			
Site	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
1	0.013	0.016	0.041	0.099
2	0.042	0.077	0.197	0.407
3	0.024	0.035	0.022	0.129
4	0.204	0.314	0.498	1.599
5	0.035	0.063	0.140	0.175
6	0.004	0.006	0.011	0.000
7	0.009	0.013	0.023	0.064
8	0.005	0.007	0.010	0.018
9	0.047	0.062	0.177	0.284
10	0.005	0.008	0.013	0.027
11	0.023	0.027	0.070	0.101
12	0.141	0.221	0.354	1.386

5.6.2 Minimum Sample Sizes

Given that, for each data set, we have only a sample of all possible data, we would like to know how large a sample is required to estimate a proportion of artefacts with a particular attribute to a certain level of accuracy, with a required probability. We use the notation of Barnett (1991) and define:

n = total number of artefacts in the sample;

N = total number of artefacts in the population;

$f = \frac{n}{N}$ = finite population correction;

p = sample proportion;

P = population proportion;

r = number of artefacts with the attribute in the sample;

R = number of artefacts with the attribute in the population;

Q = 1 - P.

5.6.2.1 A Single Proportion

Using the above notation, we will effectively assume that:

$$p \sim N(P, \frac{(1-f)P(1-P)}{n}).$$

However, this is not the immediate extension of the argument supporting the binomial distribution for p because, by incorporating the finite population correction in var(p), some account is taken of the 'lack of replacement'. Formulae are available to calculate the sample size required in order to estimate a given proportion to a certain degree of accuracy. These involve choosing n to ensure that:

$$\Pr(\,|p\text{-}P| > d\,) \le \alpha$$

where d is known as the tolerance. Ignoring the finite population correction and using the normal approximation to the binomial, leads to:

$$\Pr\left[\frac{|p-P|}{\sqrt{\frac{PQ}{n}}} > \frac{d}{\sqrt{\frac{PQ}{n}}}\right] \le \alpha$$

$$\Rightarrow \qquad 1-\Phi\left(\frac{d}{\sqrt{\frac{PQ}{n}}}\right) \le \alpha$$

$$\Rightarrow \qquad n \ge \frac{PQ\left(\Phi^{-1}(1-\alpha)\right)^2}{d^2} \qquad (5.6)$$

We estimate P and Q by their sample equivalents p and q.

5.6.2.2 Several Proportions

We now turn our attention to data matrices which consist of one or more row categories and two or more column categories (or vice versa) and where the entries in the matrix correspond to counts or abundances. Rather than calculating the sample size required to estimate one proportion to a given level of accuracy, we consider the case of estimating several proportions simultaneously. For example, with the Amarna sherds we want to determine, for each site, the minimum sample size (n) necessary in order to estimate the proportions of sherds of several wares simultaneously, to a certain level of accuracy. After some algebra and assuming the column categories are independent, we obtain a similar formula to (5.6).

For a particular row, we take P_i to be the proportion of 'artefacts' in column category i, where i=1,..., A and Q_i=1-P_i. If we also assume that P_i is the same for all column categories, then we obtain the following inequality:

$$n \geq \frac{P_i Q_i (\Phi^{-1}((1-\alpha)^{\frac{1}{A}}))^2}{d^2} \qquad (5.7)$$

where P_i ,Q_i>0.

If, instead, we allow P_i to vary across the A column categories, then we need to solve the following inequality numerically:

$$\prod_{i=1}^{A} (\Phi(\frac{d}{\sqrt{\frac{P_i Q_i}{n}}})) \geq 1-\alpha \qquad (5.8)$$

5.6.2.3 Application of Several Proportions to Amarna Pottery Sherds

In this section we are interested in estimating minimum sample sizes for each of the 12 Amarna sites (1.2.2) separately (because, in the original sampling scheme, each site was sampled independently). In order to estimate sample size when we have several proportions to consider simultaneously, we need to make an analogy with the well known case of estimating one proportion. It is known that the largest estimate of sample size for one proportion is obtained from (5.6) by taking P=0.5. Therefore, the most natural analogy for several proportions is to apply inequality (5.7) with $P_i = \frac{1}{A}$ and take A=10 Amarna wares, but this does not produce the largest sample size estimate. If, instead we take P_1=P_2=0.49 and $P_i = \frac{0.02}{A-2}$ for i =3,..., A and apply (5.8), we appear to obtain the highest estimates (assuming we take the proportions to two decimal places, which seems reasonable). These estimates are listed in Table 5.6 and clearly become higher as the significance level α decreases. For a tolerance of d=0.1, all the estimated sample sizes are smaller than those actually collected in the field, but for d=0.05 and all three values of α, the actual number of sherds collected from site 4 is lower than the estimated numbers.

Table 5.6 Estimated Minimum Required Sample Sizes for the Amarna Sites

	Tolerance (d)	
α	**0.1**	**0.05**
0.2	67	267
0.1	96	383
0.05	126	502

We now apply the above methodology to the actual Amarna sherd data. Considering various values of d and α we take, for each site, P_i and P_j to be the actual proportions which are closest to 0.49 for two wares i and j, say $P_i = P_j = p_t$ and we also take $P_k = \left(\frac{1-2p_t}{A-2}\right)$ for the remaining eight wares. The estimated minimum sample sizes for estimating all 10 ware proportions simultaneously for each site are given in Table 5.7 below. These are not necessarily the largest minimum sample sizes but they are likely to be very close — if a site has one ware that accounts for the vast majority of sherds at that site then it may be possible to obtain slightly higher estimates of sample size by using different proportions (e.g. site 3 contains one ware which forms 90% of the total sherds at the site and we can obtain higher estimates for this site). This method assumes, however, that sherds of each ware are found at each site and if we use the number of different wares that are actually found at each site then we obtain smaller estimates. The actual numbers of sherds collected at each site are given in the last column of Table 5.7.

Table 5.7 Estimated Minimum Required Sample Sizes for the Amarna Sites (Actual Data)

	Tolerance (d)						
	0.1			0.05			Number
Site	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	Collected
1	47	70	95	186	277	379	881
2	50	68	90	197	270	358	1447
3	19	27	37	74	108	146	960
4	54	76	100	215	303	399	243
5	37	54	72	146	215	288	590
6	11	15	20	41	60	80	555
7	43	66	92	172	263	365	1788
8	52	76	102	207	302	406	2589
9	64	91	120	254	364	477	576
10	53	77	102	210	306	408	1951
11	43	61	81	170	244	321	779
12	49	73	98	196	289	389	334

If we consider a tolerance of 0.1 to be adequate, then we can recommend collecting considerably less artefacts in future (at all sites) than the numbers that were actually obtained, for all three values of α. However, if we take d=0.05 and α=0.05 then the actual numbers of sherds collected from sites 4 and 12 are less than the numbers which we recommend, based on inequality (5.8). For d=0.05 and α=0.2, the numbers of sherds collected at all sites exceed the recommendations.

5.6.2.4 Application of Several Proportions to Memphis Pottery Sherds

In this section we estimate minimum sample sizes for the Memphis sherds (1.2.1). Because the original numbers of Memphis sherds obtained from each context were not fixed in advance, it is more appropriate to estimate a minimum sample size for the total number of sherds obtained (although, given the fact that we are looking at contexts, we would not actually be able to choose our sample size on excavation). To account for all AB cells of the data matrix simultaneously we can use a similar argument to that in 5.6.2.2, taking P_i to be the same for all i =1,..., AB (where P_i is the proportion of the total number of sherds in cell i, $Q_i=1-P_i$ and $P_i,Q_i>0$). We then obtain the inequality:

$$n \geq \frac{P_i Q_i (\Phi^{-1}((1-\alpha)^{\frac{1}{AB}}))^2}{d^2}$$

where n is the required sample size. If, instead, we allow P_i to vary across each cell, then we obtain the following:

$$\prod_{i=1}^{AB} (\Phi(\frac{d}{\sqrt{\frac{P_i Q_i}{n}}})) \geq 1-\alpha \qquad (5.9)$$

Considering various values of d and α and applying (5.9) we can estimate an overall minimum sample size (although, because the Memphis data consist of sherd weights rather than sherd counts, we are actually estimating a minimum weight). Table 5.8 shows the estimated weights when P_1 = P_2= 0.49 and $P_i = \frac{0.02}{AB-2}$ for i=3,..., 624. These proportions appear to produce the highest estimates of weight.

Table 5.8 Estimated Minimum Required Weights for the Memphis Contexts (grams)

	Tolerance (d)	
α	**0.1**	**0.05**
0.2	67	267
0.1	96	383
0.05	126	502

Comparing the values in the table with the actual weight of sherds collected (261280 grams), suggests that regardless of our choice of tolerance and significance level, a much smaller weight of sherds is required than that actually obtained. Taking P_i and P_j to be the actual cell proportions from the data which are closest to 0.49, say

$P_i = P_j = p_s$, taking $P_k = \left(\frac{1-2p_s}{AB-2}\right)$ for the remaining 622 cells and using (5.9) we obtain the values in Table 5.9. There are 168 non-zero cells in the data matrix and so in brackets are the estimated minimum weights using this number of cells. Again, much smaller weights are required than that that was actually obtained, although it may be possible to obtain higher estimates of sample size, depending on how the abundance of sherds is distributed across the sites and wares.

Table 5.9 Estimated Minimum Required Weights for the Memphis Contexts (Actual Data)

	Tolerance (d)	
α	**0.1**	**0.05**
0.2	388 (314)	1551 (1255)
0.1	429 (354)	1714 (1416)
0.05	469 (394)	1875 (1575)

5.6.2.5 Problems with Estimating Minimum Sample Sizes for Pottery Sherds

Having used sampling fractions to calculate minimum required sample sizes in order to estimate several proportions simultaneously, we now consider the arguments against this approach. Orton *et al.* (1995) criticise attempts to answer the question of 'what is a minimum viable sample size below which it is not worth quantifying any assemblage' for two reasons:

[1] They argue that we expect to merge assemblages (an assemblage is a collection of artefacts) into different groupings for different purposes, for example chronological groupings or functional groupings and so even an assemblage that is 'too small' by itself may form a useful part of some larger grouping.

[2] They argue that a lower limit would be in terms of pies (pottery information equivalents: numbers which are obtained from eves — estimated vessel equivalents: estimates of the number of pots represented from sherds — and which have the same statistical properties as counts of objects — one pie contains as much statistical information as one whole pot), because we seek a lower limit on the information contained in an assemblage. However, we cannot measure pies directly, only from eves and so to know whether we are above or below a threshold, we must quantify the pottery first, by which time it is too late to save time by not doing so.

We believe that the first criticism can be discounted, because we propose looking at sample size recommendations for either a specific collection of data, collected to answer a particular question (as with the Memphis and Amarna sherds), or, if several groupings have been envisaged prior to data collection, then recommendations can be made based on all these groupings. We also propose using the methods for data other than pottery sherds, for example starch grains, which are not subject to these criticisms.

We accept the second criticism, but propose that because excavations commonly take place over several seasons, an idea of the likely type and quantity of finds can be gained in the first or second season (similar to trial trenching as a means of gaining an idea of the type of subsoil and likely finds) and this information can then be used to estimate sample sizes for future seasons. It is also extremely probable that similar studies as far as the statistical aspects are concerned, will be carried out in the future, for which we can make recommendations.

Orton *et al.* (1995) criticise the traditional statistical approach of using sampling fractions because they say that we have no idea of the original size of the population. They argue that such an approach is not an adequate description of the sampling process because it does not take into account the fact that the pots are nearly always found broken and incomplete and that in general, brokenness varies between wares and according to size within the same ware, so that sherd counts are biased as measures of the proportions of wares. They also believe that correspondence analysis cannot be applied to sherd counts because the requirements of independence and/or lack of bias are not met. By this they presumably mean that because sometimes many sherds are from the same pot, each sherd is not independent and so some pots are overrepresented in the sample. At Memphis, the sherds were weighed because there were too many to count individually and if they had been counted then the amount of data collected would have been reduced. We believe that the bias argument is overcome by using sherd weights.

5.7 Summary and Conclusions

This chapter has investigated various methods of examining the stability of the categories displayed in a correspondence analysis map (i.e. how representative are the samples that they contain of the true population of data) and examined the effect of varying the sample size on the results of the analysis. Two methods of assessing stability were discussed, both of which involved fitting the multinomial distribution to the original data and generating replicate matrices (bootstraps) before implementing correspondence analysis, remembering that resampling should be carried out in the same way as the original data were collected. Typically, this means either fitting a single multinomial distribution to the whole matrix or fitting a series of multinomial distributions, one to each column (or row). Regardless of which method is used, each column category in the original analysis leads to a cloud of points, one from each replicate matrix, from which stability is assessed. We revealed that for a given data set, the size of the bootstrap clouds does not really alter according to which resampling method is used, although it is known that cloud size is affected by the number of bootstraps generated (more bootstraps lead to larger clouds). We also developed a third method for assessing stability, which is based on a jack-knife approach and involves deleting each column (or row) category in turn, before implementing correspondence analysis on the reduced data. We revealed that the resulting clouds of points for each category are much smaller under this method than those obtained from fitting one or more multinomial distributions (although this depends on the number of replicate matrices generated) and so the jack-knife method provides a useful standard against which other methods of assessing stability can be measured.

Having generated a series of replicate matrices, there are two known methods of obtaining the category co-ordinates to display in the correspondence analysis map, both of which were discussed in detail and compared. The first method involves relating the replicate matrices to the original co-ordinate system via the transition formulae; the second approach is to carry out a new correspondence analysis on each matrix. The latter method leads to larger bootstrap clouds and filtering is required to overcome the arbitrary sign changes resulting from the singular value decomposition (which forms part of correspondence analysis). This chapter has focused on the former method, but the equivalent of the latter method for biplots forms the basis of Chapter Eight. The stability of the categories was summarised using the known method of (non-parametric) convex hulls, but we also introduced (parametric) concentration ellipses for this purpose. We proposed using the areas of hull peels to assess stability (larger areas indicate greater instability) and to obtain comparable areas between clouds resulting from differing numbers of bootstraps. We also introduced the idea of using the area of an ellipse to measure stability and this method has the advantage of being unaffected by the number of bootstraps generated. In addition, we suggested using ellipse overlaps to assess similarities between categories. In particular, we suggested that if the centroid of a 95% ellipse representing one category is included in the 95% ellipse of another category, then the categories can be considered to be virtually indistinct. Sampling from the multinomial distribution was compared with sampling without replacement and inferences were drawn regarding the effect of sample size (e.g. the number of artefacts collected) on the correspondence analysis map and on stability. It is clear that the smaller the sample size the less stable the category, but also that sampling using the multinomial distribution leads to greater instability than sampling without replacement.

Sometimes, large numbers of trace zero cells occur in archaeological data (i.e. the sampling technique is not adequate to detect rare artefacts). This can be a problem when generating replicate matrices based on the multinomial distribution, because each zero cell is allocated zero probability. We therefore developed two methods based on the binomial distribution to adjust the probabilities assigned to these cells. However, the sizes of the bootstrap clouds appear unchanged by these methods unless the sample size is very small and this is because the probabilities assigned to the zero cells are also very small. We have therefore concluded that it is not worth accounting for trace zeroes in the data when assessing for stability.

Finally, we investigated how the actual numbers of artefacts collected by archaeologists compare with recommendations based on statistical calculations, obtained by using traditional sampling theory i.e. using sampling fractions. Because our data consist of several categories, we made an analogy with the well known case of estimating sample size for one proportion and assumed that the categories are independent. Criticisms of applying this traditional approach to archaeological data were also considered and largely refuted. It is clear that the actual sample sizes collected by archaeologists tend to exceed those required based on statistical criteria, sometimes by as much as 600% for any particular site.

Chapter Six

Category Selection Methods and Correspondence Analysis

6.1 Introduction

The theory behind correspondence analysis (CA) was explained in Chapter Two, where its application to pottery sherds and starch grains was illustrated. Chapter Two also highlighted problems that arise when applying this technique to archaeological data, in particular the difficulty in interpreting the ordination map when large numbers of categories are displayed and the effect of the number of row categories on the relationships between column categories. We also discussed the fact that it is sometimes necessary to divide categories after data collection, on the basis of an external variable. This chapter, therefore, aims to combine the theory of correspondence analysis with other techniques such as bootstrapping, procrustes analysis and jack-knifing in order to investigate issues such as those raised above.

Section 6.2 discusses the rationale behind category selection methods and describes the various strategies available for selecting the number of categories into which artefacts are classified. Section 6.3 explains and applies an existing method of selecting categories for deletion, proposed by Krzanowski (1993) and introduces the use of a scree-plot to aid category selection. This section also suggests ways in which Krzanowski's method could be adapted. A method of clustering categories discussed by Greenacre (1988, 1993b) is explained in 6.4, where we also introduce terminology for distinguishing between statistical 'clustering' of categories and 'merging' categories based on archaeological grounds. Correspondence analyses on the resulting categories from both methods are then compared. In Section 6.5 we propose using correspondence analysis to assess the effects of dividing categories and in 6.6 we discuss reasons for leaving categories unchanged. In Section 6.7 we develop a method which accounts for both combining and deleting categories simultaneously, which is based on work by Krzanowski (1993). We also compare two methods of combining categories in this section and other possible methods are suggested, but not implemented. In addition, 6.7 investigates the stability of and the influence of sample size on, the correspondence analysis map resulting from category selection. In Section 6.8 we extend the method of jack-knifing first introduced in Chapter Five to the detection of influential categories and we conclude the chapter in 6.9. Throughout this chapter we illustrate the various methods on the Amarna and Memphis pottery sherds (1.2.1, 1.2.2), the Melanesian starch grains (1.2.3) and also on Early Stone Age tools (1.2.4).

6.2 Selecting Categories

Sometimes, artefacts are classified into such a large number of categories that it is difficult to distinguish between them on the correspondence analysis map — we saw this in Chapter Two with both the Melanesian starch grains and the Memphis pottery sherds. Also, the data collected are often sparse with many zero counts, making it difficult to distinguish between categories based on these (insufficient) data. Additionally, it may be that some categories are expensive to obtain and that both time and money can be saved if fewer are 'needed' in order to reveal the same relationships between the row and between the column classifications. However, whether numbers of categories can be reduced depends on the objectives of the study. If, for example, the main aims are to answer archaeological objectives, with statistical analyses playing a small part in this, then it makes sense for an archaeologist to differentiate between all pottery wares, because this is extremely important for answering archaeological questions (i.e. category reduction is redundant). It may also be the case that either the row or the column classifications in a correspondence analysis are beyond the control of the archaeologist. There may, for example, be predefined categories (e.g. Bronze Age, Iron Age, Roman Period etc.), but it may be possible to sample more or fewer 'sites' to compensate for this fixed number of categories. We believe that there are five possible options available for deciding on the number of categories to include in a correspondence analysis, namely: deleting, clustering, merging or dividing the categories, or leaving them unchanged. We give a critical approach to each of these possibilities in the sections that follow.

6.3 Deleting Categories

Krzanowski (1993) comments that the focus of an analysis of a contingency table is on determining whether the grouping categories (e.g. the rows) can be sufficiently distinguished from each other on the basis of the observed characteristics (i.e. the columns). However, we disagree with this for two reasons. Firstly, it is not always clear that there are grouping categories and, separately, observed characteristics: it may be that there are two sets of observed characteristics (for example, pottery sherds are often cross-classified into fabric and ware). Secondly, we often have such large amounts of data (e.g. the Melanesian starch grains and Memphis pottery sherds described in Chapter One and listed in the Appendix) that our aim is only to look for relationships between rows and between columns, which cannot be seen from looking at the raw data alone, i.e. we merely want to display our data.

Considering category deletion methods, one means of assessing the effect that the deletion of a complete row or column of the contingency table has on the correspondence analysis is to use the influence function (Pack & Jolliffe, 1992). This considers the change in eigenvalues or eigenvectors, thereby ranking the importance of rows or columns. However, Krzanowski (1993) believes that the problem with this is that it ranks the importance of the overall goodness of fit of the r-dimensional configuration and does not pay attention to the individual row points.

Krzanowski (1993) therefore introduced a method to select those columns of a contingency table that are the most important in describing the differences between rows. Krzanowski first considered this aspect in the context of principal component analysis and proposed a procrustean measure of importance for each variable, which he subsequently adapted to correspondence analysis; we extend this method to the various forms of biplot and to canonical correspondence analysis in Chapters Seven and Nine respectively. The method works as follows:

Stage 1: Carry out a correspondence analysis on the data and retain the co-ordinates of the rows in the reference configuration X.

Stage 2: Omit each column of the data matrix in turn, implement correspondence analysis and retain the row co-ordinates of the reduced data set in matrix Y.

Stage 3: Apply procrustes analysis to minimise trace{(X-Y)(X-Y)^T} under translation, rotation and reflection of Y. This results in a residual sum of squares M^2, where the smallest M^2 corresponds to the least important variable because deleting it results in a configuration that is the least different from the reference.

Krzanowski explains that there are extra considerations that arise with the application of this technique to the co-ordinates obtained from correspondence analysis. In general, we can expect a substantial change in the pattern of entries of the contingency table when columns are deleted from it. The rotated configurations after column deletions are thus likely to undergo considerable scale changes and if this is felt to be problematic then the configurations should be rescaled to a common size before each calculation of M^2. Krzanowski suggests that a simple way of doing this is to rescale X and Y so that the sum of squares of elements in each matrix is equal to a constant value, say one. The second consideration is that in correspondence analysis masses are attached to each point in the configuration (see 2.2.1) and so it can be argued that in calculating M^2 it is more appropriate to minimise a weighted sum of squares (i.e. we are more willing to tolerate an error in the position of a point with low mass than in the position of a point with high mass). For this reason the preliminary translation of the configurations should be so that their weighted centroids coincide. However, it is not clear which are the appropriate weights to use since the row masses will change each time that a column is deleted from the table. Krzanowski says that a relatively stable choice is to use the masses obtained from the original table, because this provides the reference configuration at each step of the analysis. Thus, using the notation of Chapter Two, we weight each row of X and Y by the mass r_i and we obtain M^2 as the minimum of trace{D(X-Y)(X-Y)^T}, where D = diag (r_1,..., r_r). We discuss later why we believe that the original co-ordinates may not be the best choice of reference configuration for each step of the method.

Krzanowski goes on to say that the ideal solution for the selection of the best q columns is to compute M^2 between the new and reference configurations for each possible choice of q columns and to select the q columns that correspond to the smallest M^2. However, he also believes that a backward elimination algorithm can be considered to be an acceptable alternative. He chooses the dimensionality of the reference configuration (m) to be the smallest dimensionality greater than two which accounts for at least 80% of the total inertia and takes q to be as close as possible to m as seems reasonable in each data set. We apply Krzanowski's method to the Memphis pottery sherds in Section 6.3.2, where we implement both the rescaling and weighting of rows. We suggest that a minimum of 80% is often too stringent and that the number of columns selected should be chosen independently of the dimensionality used in the calculations (see the discussion in 6.3.5).

6.3.1 Reasons for Deleting Categories

Before implementing any category deletion methods we believe that it is important to list the three main reasons why deletion may be appropriate.

[1] The data collected on certain categories may be so sparse so as to hide relationships both between and with, other categories. This can cause the points to be all bunched up together in the correspondence analysis map and deleting one or more of these sparse categories can lead to 'true patterns' emerging (or at least more recognisable and interpretable ones).

[2] Data on a very large number of categories may have been collected and too many categories make it almost impossible to identify patterns in the data because they cannot all be visualised. Deleting some of these categories may, therefore, considerably aid interpretation.

[3] Time and effort in future studies can be saved if a suitable number of categories can be recommended before data collection begins. This is of particular importance in archaeology where there can be a tendency to 'overcollect' because of the difficulty (both in time and expense) of returning to a site (which may no longer exist).

6.3.2 Application to Memphis Pottery Sherds — Deleting Wares

The aim of this section is to use the Memphis sherd data (1.2.1) to try to establish which ware categories are most important in explaining the differences between contexts and which contexts are most important in explaining the differences between wares. This is an example of data that do not consist of grouping categories and observed characteristics (see the discussion at the beginning of 6.3). We know that, by definition, there are differences between wares. It is, however, more difficult to distinguish between contexts because, using the stratigraphic method of excavation, this relies on identifying changes in colour, texture and smell of the sediment or soil. Both contexts and wares therefore need to be identified by experienced archaeologists. By examining the wares we can identify which of them are most important in allowing us to differentiate between different levels of activity and time periods in the past: we can also investigate how the total quantity and number of different wares alter over time (this has relations with frequency seriation, see 2.4). By considering the contexts we can investigate whether there are some which are dominated by particular wares, or whether the wares are spread evenly across the contexts. However, contexts by definition form a sequence and it is not sensible to consider the effects of deleting any one context. We therefore introduce the idea of combining neighbouring contexts and the justification for this later in the chapter (see 6.7.5).

It is evident that with as many as 48 wares, computing all subsets of wares in order to establish the 'most important' ones would be extremely time consuming and so we follow the backward elimination procedure proposed by Krzanowski. However, we initially choose the dimensionality of the reference configuration to be two (rather than basing it on the percentage of variation explained, as Krzanowski suggested), because we will always display the data in two dimensions. Table 6.1 lists the order in which the wares are deleted and the corresponding M^2 values for the first 12 steps of the procedure.

Table 6.1 Order of Deletion of Memphis Wares

Step	M^2 (× 10^6)	Ware Deleted
1	0.139	43
2	0.123	46
3	0.132	45
4	0.144	42
5	0.149	44
6	0.178	17
7	0.219	13
8	0.319	21
9	0.452	18
10	1.077	14
11	2.293	15
12	3.899	39

We introduce a 'scree-plot' in Figure 6.1, which plots the ware deleted at each step against the corresponding M^2 value and we see that we stop deleting wares after ware 18 because this is the point at which there is a large change in slope. We can think of the vertical axis as a goodness of fit measure, with the bottom representing the best fit (i.e. all categories included) and the top the worst fit. For clarity, we only plot the first 12 steps of the scree-plot, which includes the sudden rise after ware 18 is deleted. However, we should always implement all steps when applying this method of category selection, because the magnitude of M^2 can alter substantially between steps. If, instead, we produce a cumulative scree-plot i.e. we sum the values of M^2 across the steps, then again we stop after deleting ware 18, but this time the change in slope of the plot is more pronounced (and the plot will always be monotonically increasing).

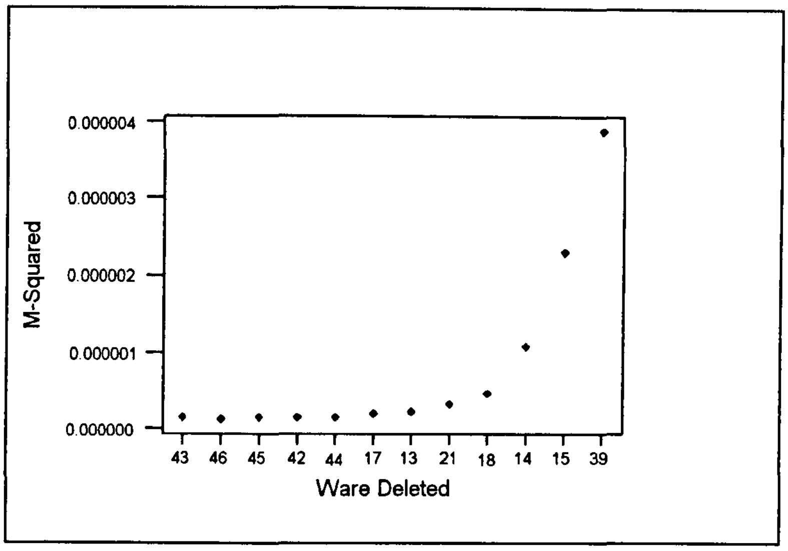

Figure 6.1 Scree-Plot for the Memphis Wares (Backward Elimination)

Having deleted 9 wares, we need to carry out a correspondence analysis on the remaining wares in order to examine how the relationships between contexts have altered, compared with when all the data were retained. We find that the correspondence map is little changed (see Figures 2.1 and 2.2), which suggests that the wares we deleted are the 'redundant' wares in some sense and that little information has been lost by deleting these wares. Thus, for these data two dimensions (which explained 59.3% of the inertia of the original data) are sufficient and many more than two wares are retained (see the discussion in 6.3 on choosing dimensionality and number of categories). We believe that there is, therefore, scope for adapting Krzanowski's method.

Besides using a reference configuration in two dimensions and introducing scree-plots and cumulative scree-plots to detect 'surplus' categories, we believe that there are other possibilities for selecting which wares to delete:

- Choose the dimensionality of the reference configuration based on the percentage of variation explained (as in Krzanowski, 1993). This is discussed in the following section for the Amarna sherds.

- Use a forward selection, all subsets or stepwise procedure for selecting wares, although the first and last of these may not produce any 'better' identification of redundant categories than backward elimination (the all subsets approach is also too time consuming to be implemented for the Memphis sherds). It may be that all these methods will select different categories for deletion but that there are several possibilities (i.e. there is no unique solution), none of which substantially affect the correspondence analysis map. We believe that the ideal method of category deletion should enable several subsets of categories to be obtained and as we will see in the next section, the selected categories vary depending on the number of dimensions used in the procrustes calculations.

- Compare the co-ordinates obtained at each step with those from the previous step rather than with those obtained from the whole data set (and change the weights appropriately). Our justification for this is partly that the selection process is more closely monitored, but also that this provides closer similarities with variable selection methods in regression.

6.3.3 Application to Amarna Pottery Sherds — Deleting Wares

In this section we apply Krzanowski's backward elimination method to the Amarna sherds (1.2.2). More specifically, we attempt to establish which wares are most important in explaining the differences between sites i.e. if a particular ware is not identified (either ignored or classified with another ware) then how is the relationship between sites affected. We can also establish which sites are most important in determining the differences between wares i.e. how do the relationships between wares alter if a particular site is not visited; this is the focus of 6.3.4. We apply Krzanowski's backward elimination method (again using a reference configuration in two dimensions) and the resulting scree-plot is displayed in Figure 6.2. Inspection of the scree-plot (and cumulative scree-plot) indicates that there is little change in M^2 from deleting wares 1, 4 and 5, a small change after deleting 3 and more substantial changes from deleting further wares. This suggests that the elimination of wares should cease after ware 3, although other factors might be used to decide whether or not to include ware 3, because the scree-plot has no clear large change of slope. One such factor might be the quantity of pottery available to record, because if ware 3 is rare then there may be little point in including it in the analysis.

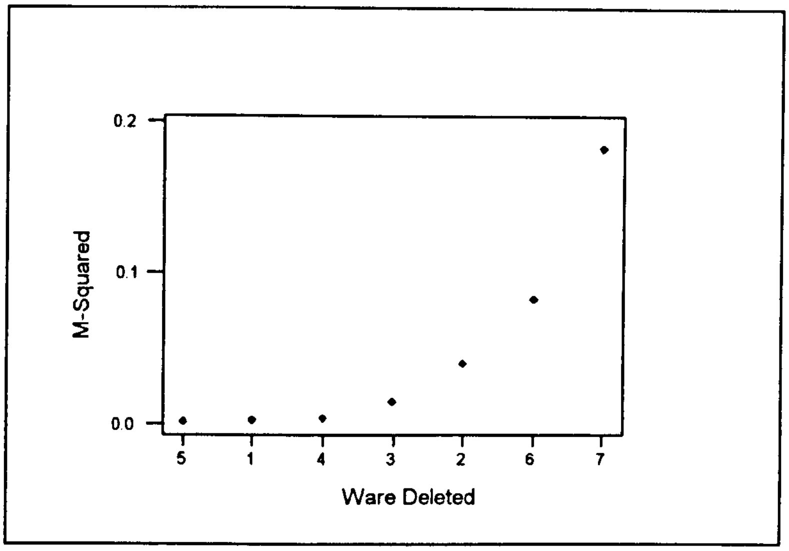

Figure 6.2 Scree-plot for the Amarna Wares (Backward Elimination: Two Dimensions)

Implementing correspondence analysis after deleting wares 1, 4 and 5 produces a very similar site configuration to that of Figure 2.3, which included all wares and so this again indicates that the scree-plot is useful in deleting wares which do not add information to the relationships between sites. If ware 3 is also deleted then the resulting correspondence analysis map is still similar to Figure 2.3. We should recall from Section 1.2.2 that the Amarna sherds were collected by scribing a circle of given radius and classifying all pottery wares within that area. However, with other sampling schemes there may be scope for deliberately including wares that are known to be common.

If, instead of using two dimensions to calculate M^2 when selecting wares, we choose the dimensionality of the reference configuration according to the percentage of variation explained, as Krzanowski suggested, then we use four dimensions and this produces the following scree-plot (Figure 6.3). However, using four dimensions in the calculations means that we are only able to take five steps in the backward elimination process.

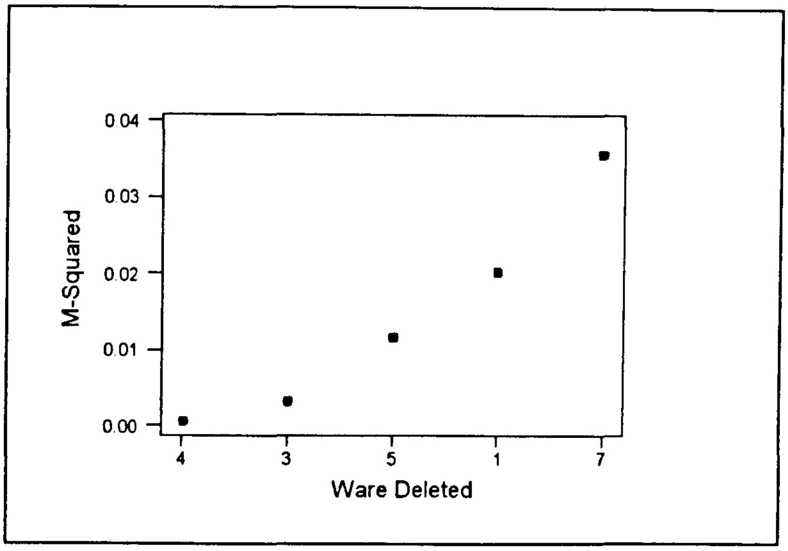

Figure 6.3 Scree-plot for the Amarna Wares (Backward Elimination: Four Dimensions)

Figure 6.3 suggests that we should delete wares 3 and 4; the results of a correspondence analysis with these wares deleted are little altered from those with all 10 wares. However, when using two dimensions in our calculations we are able to delete more wares based on the scree-plot, without altering the pattern in the correspondence analysis map. It is interesting to note that the inertia explained in the original data in two dimensions was 56.3% and so it seems that it is not always necessary to choose dimensions that account for 80%, as suggested by Krzanowski.

6.3.4 Application to Amarna Pottery Sherds — Deleting Sites

Besides looking at which wares are most important for describing differences between sites, as we did in the previous section, we can also establish which sites are most important in determining the differences between wares i.e. how do the relationships between wares alter if a particular site is not visited. We apply Krzanowski's backward elimination method (using a reference configuration in two dimensions), but the resulting scree-plot is not monotonically increasing. The cumulative scree-plot is illustrated in Figure 6.4. With this method and these data we are only able to carry out six steps because on the 7-th step there is a ware that occurs at none of the remaining sites and so correspondence analysis cannot be applied (because the technique requires that the row and column sums are non-zero).

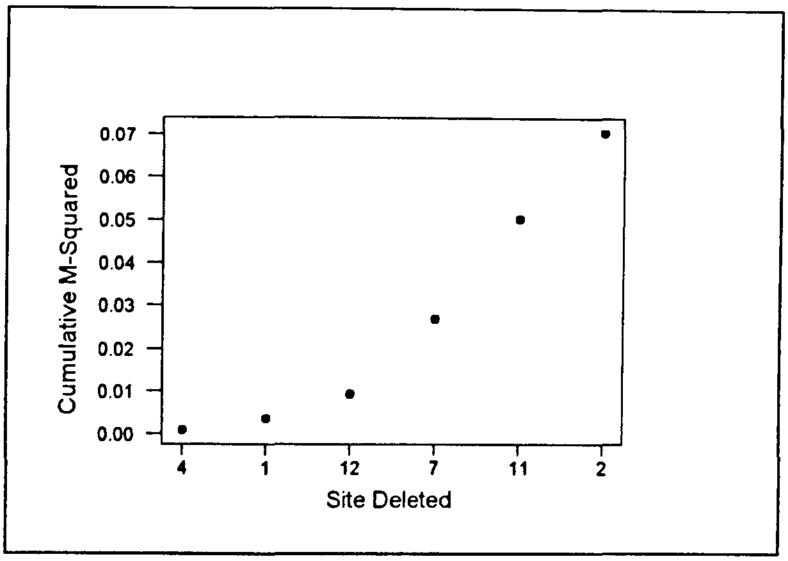

Figure 6.4 Cumulative Scree-plot for the Amarna Sites (Backward Elimination)

Inspection of the above scree-plot suggests that, because of the change in slope at step 3, we should stop the selection process after site 12 is deleted. If, instead of using two dimensions to calculate M^2 when selecting wares, we choose the dimensionality of the reference configuration according to the percentage of variation explained being greater than 80%, as Krzanowski suggested, then we use four dimensions. We would probably delete sites 4 and 6 on the basis of the resulting scree-plot (not shown), although possibly 7 and 12 as well. Deleting sites 4, 6, 7 and 12 produces a correspondence analysis map that is very similar to that of Figure 2.3, which suggests that the plot has picked out the 'least important' sites (in terms of their ability to highlight the distinction between pottery types) for deletion. If we look at the stability of the sites in Chapter Five, then we recall that sites 4 and 12 are the most unstable (they have the largest bootstrap clouds) under Greenacre's partial resampling method and so there does not appear to be any relationship between site stability and site deletion. We see from the above analysis that site 6 is deleted in four dimensions but not in two, although both resulting correspondence analysis maps are little changed from the original. This is evidence, we believe, that there are many subsets of categories which can be retained and not one 'unique' set. Formalising Krzanowski's category deletion method is therefore inappropriate.

6.3.5 Summary of Category Deletion Methods

Having implemented Krzanowski's backward elimination method and variations of it, in the previous sections, we make the following comments. Krzanowski recommended that the number of dimensions used in the calculations should together explain more than 80% of the variation in the original data. While this is only a guide, we saw with the Memphis and Amarna sherds that sensible subsets of categories are obtained from using just two dimensions, which account for a much smaller percentage of variation in the data and in some cases retain fewer categories.

We suggest that the methodology can be improved as follows. Rather than choosing the number of categories to correspond to the dimensionality, we propose choosing them independently. Firstly, we suggest that the dimensionality for the procrustes calculations should be chosen. This can be done informally (e.g. because the results will be displayed in two dimensions) or based on the percentage of variation explained. We believe that the aim of category deletion is to reduce the number of categories to a more 'manageable' number, without losing information i.e. if the patterns displayed in the correspondence analysis map vary according to the number of column (or row) categories included in the analysis, then the data are not sufficiently stable for us to be confident of our inferences. We therefore believe that formalising the selection process is not necessary, because there are likely to be several sets of categories which can be deleted without altering the inferences of the map, rather than one unique set. Both the scree-plot and cumulative scree-plot proved to be useful aids, because they allow us to visually monitor the selection process as it progresses. However, sometimes the scree-plot is non-monotonic, which can lead to inconclusive evidence concerning which categories to delete (it is nearly always more monotonic in higher dimensions and so for this reason it may be worthwhile choosing, for example, three dimensions), although the cumulative scree-plot overcomes this problem. Because the procrustes statistic is based on comparisons of sets of points, the dimensionality used in the calculations affects the number of categories that can be deleted. For example, if we have category co-ordinates in four dimensions then we cannot have fewer than five categories remaining after category selection.

6.4 Clustering and Merging Categories

Depending on the aims of the analysis, an alternative to deleting categories can be to combine them. Again, this is likely to be most appropriate when there are large numbers of categories to display, for example with the Memphis sherds and Melanesian starch grains discussed in Chapter Two (see also Section 6.2). Greenacre (1988, 1993b) describes how to investigate clusters of rows or columns, based on reductions in the chi-squared statistic and builds on work by Hirotsu (1983) in which only row-wise and/or column-wise multiple comparisons are considered.

Using the notation of Chapter Two (where the data consist of n rows and m columns), we consider the row problem. Suppose we have performed a hierarchical clustering of the rows using a method such as that due to Lance & Williams (1967); the result of the clustering can then be depicted in the form of a binary tree, with $H \equiv n-1$ nodes. It is possible to decompose the total inertia (and chi-squared statistic) with respect to this set of nodes. Every hierarchical clustering method will imply different decompositions of inertia, but Greenacre says that one method is of special interest. This method minimises the 'pseudo-distance' between clusters at each node, where the pseudo-distance between two clusters is the squared chi-squared distance between the two cluster centroids multiplied by a weighting factor, which depends on the masses of the profiles in the two clusters. This is a weighted version of the clustering criterion described by Ward (1963).

The pseudo-distance between two rows is defined by:

$$v_h = \left\{ r_{[1]} r_{[2]} / \left(r_{[1]} + r_{[2]} \right) \right\} \left\| \bar{a}_{[1]} - \bar{a}_{[2]} \right\|^2 \qquad (6.1)$$

where $r_{[1]}$ and $r_{[2]}$ are the masses of the profiles in the two respective clusters merged at node h and $\bar{a}_{[1]}$ and $\bar{a}_{[2]}$ are the centroids of the two respective clusters. It is known that if the pseudo-distance given by (6.1) is multiplied by the sample size n to obtain the equivalent chi-squared component, then we obtain the statistic given by Hirotsu (1983) to perform multiple comparisons on the rows of the contingency table. Hirotsu shows that nv_h for any two subsets of rows, or the equivalent statistic for any two subsets of columns, is bounded above by the largest eigenvalue of a matrix which has an asymptotic Wishart distribution. The relevant Wishart matrix variate, $W_r(s)$, has order r ≡ min{n-1, m-1} and degrees of freedom s ≡ max {n-1, m-1}.

In archaeology, it is not necessarily sensible to cluster categories on a statistical basis; it may be advisable to merge them based on the expertise of an archaeologist. If, using Greenacre's method, we find archaeologically non-sensible categories being clustered, then this can be argued to indicate that not enough data have been collected to show archaeological differences (i.e. the data do not contain information on the differences) and we should therefore still combine categories on mathematical reasoning. Furthermore, if the categories remain unclustered then they may contribute noise to the analysis and by clustering them we may obtain a better picture of the remaining classifications. What we believe to be the advantages and disadvantages of Greenacre's method of clustering are discussed in the next two sections and then applications of the method are illustrated.

6.4.1 Advantages of Clustering

There are three main advantages of clustering:

[1] Some information from the clustered categories is retained, whereas this is completely lost if any categories are deleted.

[2] Categories with sparse data can dominate the analysis, or contribute noise and by clustering them we can stop these effects from occurring.

[3] Clustering categories can help suggest how fine a classification is needed. This is particularly useful for future studies where similar data are to be collected.

6.4.2 Disadvantages of Clustering

There are two main disadvantages of clustering:

[1] Because the clustering is based on the chi-squared statistic, it involves only minimising the difference between the observed and expected counts. Thus, it can be argued that there is no real archaeological basis for the clustering.

[2] Clustering is only appropriate for some types of data. It is not necessarily sensible to cluster e.g. pottery wares if they are archaeologically very different.

Similarly, it would not usually be sensible to cluster sites.

We now introduce a distinction between merging and clustering, because we feel that as far as practical applications are concerned there are two approaches that need to be compared. We define merging to be combining categories as a result of archaeological information, but clustering to be combining categories as a result of statistical criteria. We discuss these two methods in the following sections.

6.4.3 Application to Memphis Pottery Sherds — Merging Wares

In consultation with Janine Bourriau, from whom the Memphis sherd data (1.2.1) were obtained, categories of wares that can be merged on archaeological grounds (i.e. similar wares) were identified. These are listed in Table A.7 of the Appendix and reduce the ware categories from 48 to 30. We implement correspondence analysis on these grouped wares in order to examine the effect of the mergings (Figure 6.5) on our interpretation of the map.

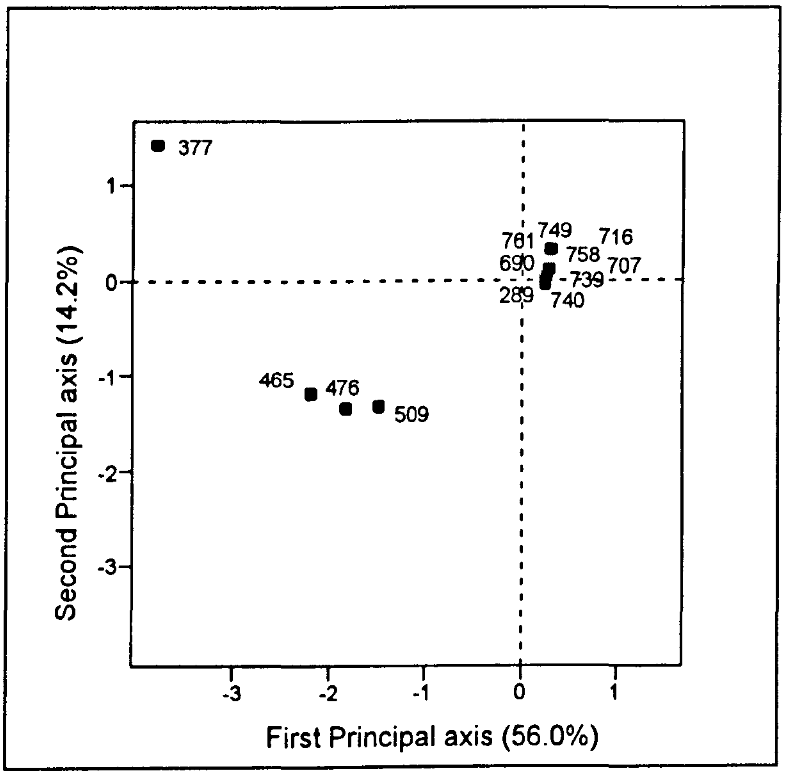

Figure 6.5 Correspondence Analysis Map of Memphis Contexts (Merged Wares)

This figure differs slightly from Figure 2.1 because context 289 is now located close to the contexts in the top right of the figure, when previously it was located some distance away from these. We can therefore conclude that it is at least one of the merged wares which is able to separate out this context from the remainder. Thus, depending on how important it is to distinguish between this context and others, compared with the advantages of reducing the number of categories in the analysis, the broader categorisation of wares illustrated above may, or may not, be acceptable. In Section 6.8 we introduce a method of detecting influential categories and if we apply this to the Memphis wares then we may be able to ascertain which wares are responsible for altering the position of context 289 on the map.

6.4.4 Application to Early Stone Age Tools — Merging and Clustering Tools

The effect of merging archaeologically similar categories can also be investigated by using the Early Stone Age tool data described in Section 1.2.4. If we compare the mergings into seven categories as defined by Bølviken *et al.* (1982), which are listed in Table A.5 of the Appendix, with the clusterings obtained from Greenacre's method applied until seven categories remain, we discover that the statistical mergings do not agree with the archaeological ones. A chi-squared test on these data produces a statistic of 1238.88 which indicates that the tools and sites are not independent and so it was sensible to proceed with Greenacre's method of clustering. We display the clusterings and mergings in the table below.

Table 6.2 Clustering and Merging Early Stone Age Tool Categories

Method	
Greenacre's Clusterings	**Bølviken's Mergings**
1, 5	1, 2, 3, 4
2, 9, 10, 14	5
3, 13	6, 7, 8
4	9, 10, 11
6, 7, 11, 12	12
8, 16	13, 14, 15
15	16

To fully implement Greenacre's method we need to compare the chi-squared values with the upper percentage point of the W_{15} (42) distribution, in order to find the stopping point. However, because we are comparing an archaeological method with a statistical one, we do not believe that this is really necessary.

We see from Table 6.2 that the groupings obtained from the two methods are very different. We believe that the reason for this is because Greenacre's criterion is based solely on the relative frequencies of tools across sites and there is no reason why the relative frequencies of tools should imply similar archaeological use and only arguably imply a similar distribution across sites. For these data, post-depositional destruction (i.e. what happened to an artefact between its deposition and its discovery) is probably not relevant, because the tools are all made of stone (and are therefore likely to have survived equally well in the archaeological record). In addition, many of the cells of the data matrix contain zero frequencies, which affects the clustering method but not the mergings of the archaeologist.

The above example illustrates how incompatible statistical and archaeological criteria for combining categories can be and that care, thought and preferably the expertise of an archaeologist should, where possible, be sought before combining categories. Figures 6.6a-6.6c below illustrate the Early Stone Age tool sites obtained from a correspondence analysis using:

[a] The original data

[b] Greenacre's clusterings

[c] Bølviken's mergings.

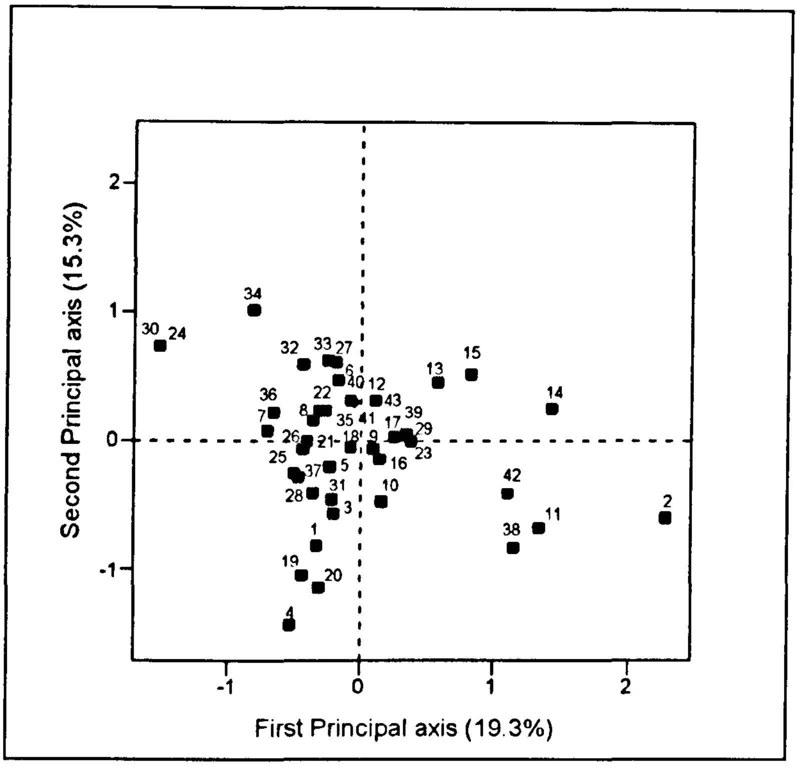

Figure 6.6a Correspondence Analysis Map of Early Stone Age Tool Sites

Figure 6.6a allows us to distinguish sites {24, 30, 34} on the left, from the group in the middle, from sites {2, 11, 14, 38, 42} on the right. Sites {1, 4, 19, 20} towards the bottom of the plot are also separated from the bulk of the points. However, the map has not revealed any clear patterns and only 34.6% of the variation in the data has been explained. Figure 6.6b reveals even less differentiation between sites, although sites {24, 30} are still separated out from the remainder and sites {6, 7, 34, 36} are slightly separated.

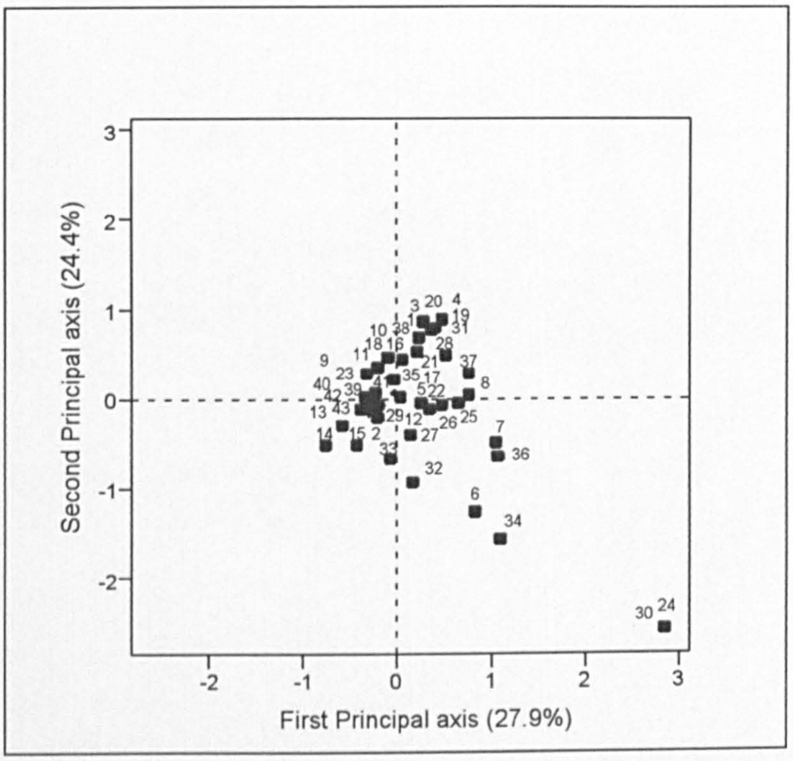

Figure 6.6b Correspondence Analysis Map of Early Stone Age Tool Sites (Greenacre's Clusterings)

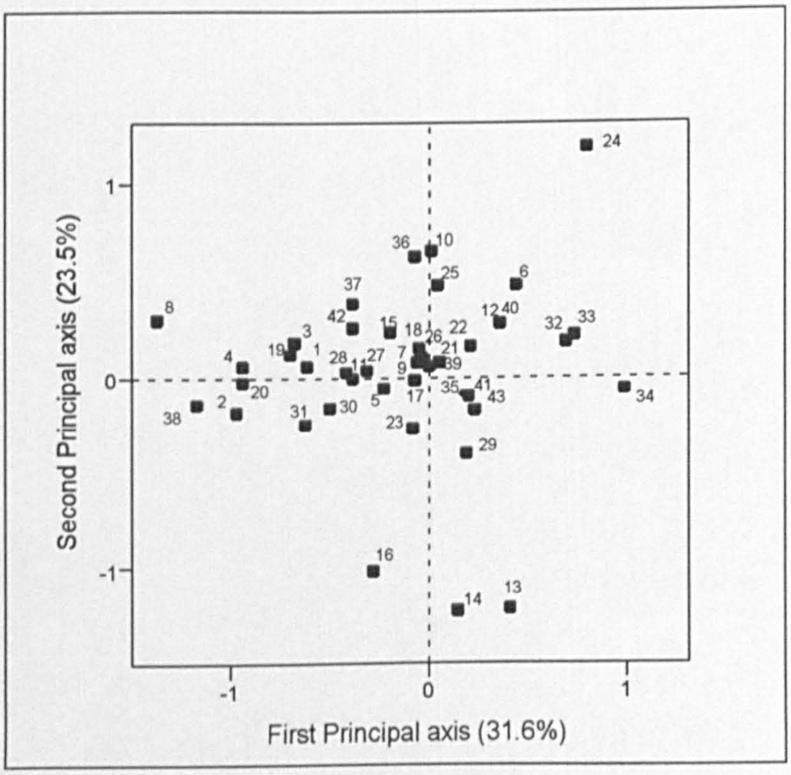

Figure 6.6c Correspondence Analysis Map of Early Stone Age Tool Sites (Bølviken's Mergings)

Figure 6.6c represents the sites resulting from the tool mergings based on the opinion of the archaeologist (Bølviken) and shows a well-spaced out plot, although some of the sites which are located close together in 6.6a are no longer located close together e.g. sites 24 and 30. However, 55.1% of the variation in the data is explained in the plot. We recall from Chapter One that the original aim of the project was to test the hypothesis that the largest sites in the inner part of the fjords of the Varangerfjord area of Scandinavia reflect larger aggregates of people during longer periods of time than the smaller sites which are located in the outer fjord-coast area. However, the map does not reveal this (sites from the inner fjords are not located together and away from the sites of the outer fjords) and despite the fact that all three figures consist of the same number of sites, Figure 6.6c is clearly the most easy to interpret. Based on the above three figures, we conclude that it may be worthwhile considering other methods of clustering categories, because if the relationship between sites varies with the tools included in the analysis, it is difficult to draw sensible archaeological conclusions. If, for example, another study is carried out, but only a selection of tool categories are obtained, we need to be confident in our interpretation of the correspondence analysis map. We therefore consider other methods of clustering categories in Section 6.7.

6.5 Dividing Categories

For some types of archaeological data the categories may contain counts that need to be divided based on some external variable and this was briefly mentioned in Chapter Two. However, it may not be clear at the time of data collection that a finer division of categories is needed, or the external information needed to subdivide them may not be available. We propose using correspondence analysis to assess the effect of category division and we illustrate this in the following section. Category division is particularly important when considering organic plant materials such as starch grains, which we focus on below, but also for phytoliths and microfossils.

6.5.1 Application to Melanesian Starch Grains

Whilst there is a belief among palynologists that any single plant species gives rise to only one 'type' of starch grain, there is a suspicion that different species could give rise to the same grain 'type'. However, grains of the same type from different species might be differentiated on the basis of their size and by looking at histograms of the lengths of starch grains for each type, it is clear that some types do consist of grains of several distinct sizes. Dividing the types into groups based on the median size or the antimode of grains within a type means, however, that a proportion of the grains are misclassified, if their sizes form a mixture of two or more distributions.

Figures 2.4 and 2.5 in Chapter Two illustrated correspondence analysis on all the Melanesian starch grain data (1.2.3), but we now consider only types that consist of more than 10 grains, because otherwise the plot becomes too crowded. We could, of course, implement the category deletion methods of 6.3 (it is not sensible to combine grains of different types) in order to reduce the number of points displayed on the correspondence analysis map, although types with fewer than 10 grains make it difficult to assess whether distinct groups of different sized grains exist. Examining histograms of the starch grain lengths (not shown) reveals that types 6, 28, 32, 40, 92 and 142 might reasonably be subdivided into two groups based on size.

By dividing type x into xa and xb at the antimode and implementing correspondence analysis, we propose that if xa and xb are located some distance apart on the correspondence analysis map then there may be some evidence that they are from different species. However, if they are located close together then they are likely to be different sized grains of the same type. If there is a grain type that has a unimodal distribution when considering a histogram of grain lengths, then there is no reason to suppose that it originates from more than one species. We can, however, divide it into two groups at the median and examine whether both groups occur together in the resulting correspondence analysis map.

A series of correspondence analyses were carried out on these data, with each type that may feasibly originate from two distinct species separately subdivided and then with all these types divided simultaneously (Figure 6.7).

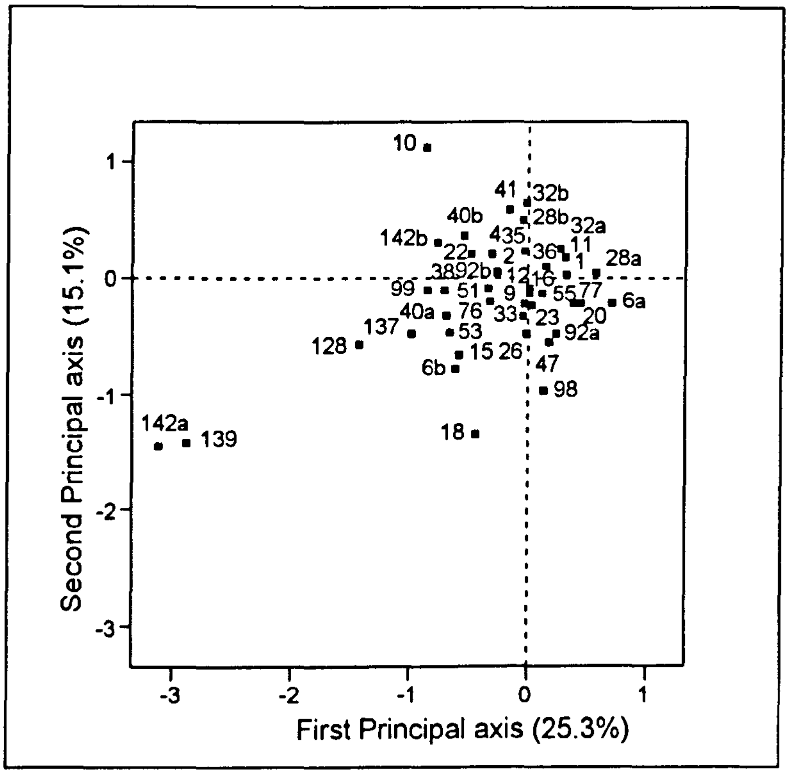

Figure 6.7 Correspondence Analysis Map of Melanesian Starch Grain Types (With Subdivisions)

We see from the above figure that when type 6 is split, 6a and 6b are located some distance apart and similarly for 142a and 142b. Subdivisions {28a, 28b}, {40a, 40b} and {92a, 92b} are reasonably close and {32a, 32b} closer still. Dividing all types separately, whilst the other types remain undivided, does not really alter these patterns. This is an advantage because it means that types can be considered separately, without confusion and that new data can easily be incorporated into the analysis. We should also bear in mind that types that are located close together after category division support the hypothesis that they are from the same species, but they do not prove it. Also, the decision of how close the points have to be, to be considered to be from the same species is in part subjective, but could be aided by convex hulls and concentration ellipses. As in Chapter Five, we could generate replicate data matrices with these types divided, by fitting a series of multinomial distributions, one for each site. Then, for example, if the centroids of the 95% concentration ellipses of the two divisions of a type overlap, we may infer that the types are from the same species.

We believe that this method of assessing the effects of category division could come under criticism for the following reason. Correspondence analysis is based on relative frequencies (of grains). If dividing a type into two groups based on size leads to both groups having similar frequencies across sites, then they will be located together in the CA map, but it is not clear why similar relative frequencies of different size grains should imply that they originate from the same species.

6.6 Leaving Categories Unchanged

In the previous sections we have discussed various methods for altering the number of categories into which the data are classified. However, before any category selection method is applied, we need to appreciate the reasoning behind why these categories were originally chosen. For example, the data may have been collected and classified into particular categories because these were testing a specific hypothesis of the archaeologist. Deleting, clustering, merging or dividing them therefore alters the question(s) originally posed. Often, the archaeologist has only one chance at collecting artefacts and unless s/he has retained them, the corresponding categories cannot be subdivided at a later stage, but they can be merged, clustered or removed from analysis.

One method of overcoming the problem of a heavily cluttered correspondence analysis map is to implement correspondence analysis on all the data, but to display only some of the resulting row and column points at any one time. In this way, all the data are used in the analysis (and thus no information is lost), but the plot is not too confusing and patterns can be revealed. It may also be advisable to exclude categories consisting of sparse data from the analysis and project them onto the resulting display as supplementary points (see 2.2.4). Leaving categories unchanged clearly retains more information than the other methods of altering category numbers.

6.7 Combining Methods of Category Selection

In the following sections we introduce and apply a method that allows us to simultaneously consider deleting and clustering categories. This is useful when it is clear that either there are too many categories to display, or when some categories consist of sparse data.

6.7.1 Backward Elimination Procrustes Analysis

In this section we describe a method which we have developed to allow, simultaneously, for the possibility of deleting and clustering categories. In order to decide which column categories best distinguish between row categories we propose the following method:

Stage 1: Each column category is deleted in turn and each pair of column categories are combined in turn.

Stage 2: Correspondence analysis is applied to each reduced matrix and row co-ordinates are obtained.

Stage 3: Each set of row co-ordinates is compared with the reference configuration (the co-ordinates from the original data) using procrustes analysis, scaling each configuration and weighting by the original row masses as suggested by Krzanowski (1993) and as described in Section 6.3. The residual sum of squares, M^2, is obtained in each case.

Stage 4: The column deletion or column clustering that results in the smallest M^2 is implemented.

Stage 5: Stages 1-4 are repeated for the reduced matrix. The values of M^2 at each step are then plotted in a scree-plot to assess the stopping point i.e. the number and combination of categories to retain.

6.7.1.1 Application to Amarna Pottery Sherds

In this section we apply the method just proposed to the pottery sherds from Amarna (1.2.2). The resulting M^2 values and corresponding retained wares are displayed in Table 6.3; the resulting scree-plot is illustrated in Figure 6.8.

Table 6.3 Category Groupings of Amarna Wares

Step	M^2 (× 10^3)	Wares
0	0.000	{1} {2} {3} {4} {5} {6} {7} {8} {9}
1	0.026	{1} {2} {3} {4,5} {6} {7} {8} {9} {10}
2	0.040	{1} {2} {3} {4,5,6} {7} {8} {9} {10}
3	0.068	{1,2} {3} {4,5,6} {7} {8} {9} {10}
4	0.156	{1,2} {3,7} {4,5,6} {8} {9} {10}
5	8.652	{1,2,3,7} {4,5,6} {8} {9} {10}
6	20.430	{1,2,3,7} {4,5,6,9} {8} {10}
7	61.790	{1,2,3,4,5,6,7,9} {8} {10}

We include $M^2 = 0$ in the scree-plot to allow for the possibility that no categories are combined or deleted, but for these particular data the slope of the plot is not altered if it is omitted (although this is not always the case — for example, see Figure 6.10).

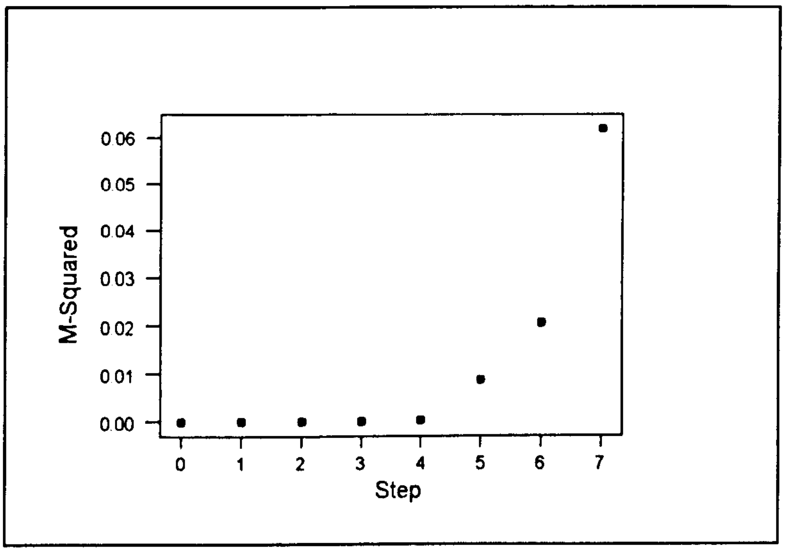

Figure 6.8 Scree-plot for the Amarna Wares (Procrustes Analysis)

Considering Figure 6.8 we stop the elimination process after step 4, taking the associated groupings from Table 6.3. A correspondence analysis map using these groupings is illustrated in Figure 6.9.

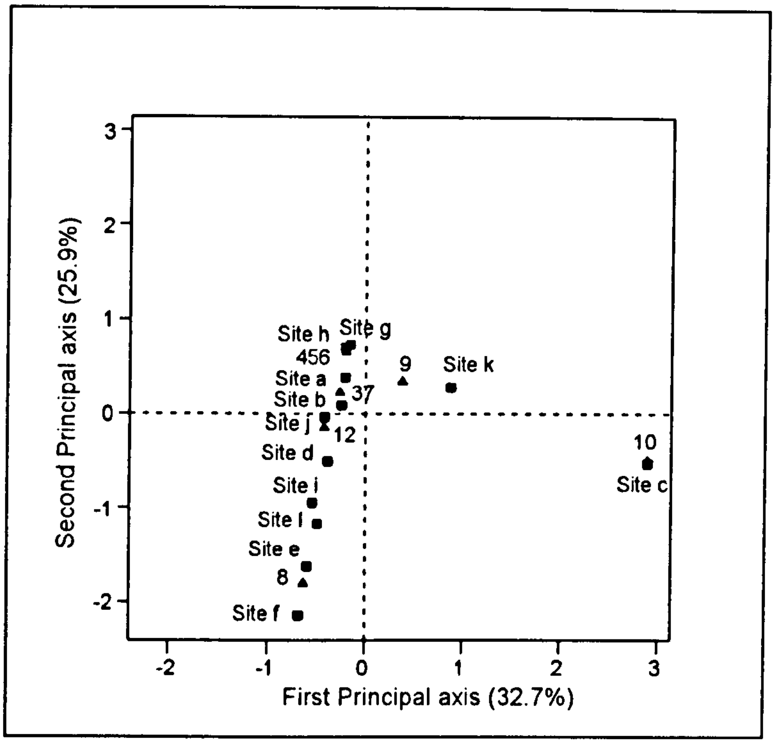

Figure 6.9 Correspondence Analysis Map of Amarna Pottery Sherds (Category Groupings)

Figure 6.9 produces a very similar picture to Figure 2.3 (the original data), after allowing for arbitrary reflections (see 5.2.4). The method therefore appears to work well because the wares that have been combined have not altered the original map and, therefore, our interpretations of the relationships between contexts are unchanged. This gives us confidence in our inferences made from the correspondence analysis map.

6.7.1.2 Application to Early Stone Age Tools

As a second example we apply the above method to the Early Stone Age tools (1.2.4) and the resulting categories are listed in Table 6.4. With these data it is necessary to stop at step 10 because when we try to delete columns {1,3,5,6,7,12,13} in the first stage of step 11, we obtain one row total of zero which means that correspondence analysis cannot be applied. Because the data consist of generally low row counts anyway, it is probably best not to assign a value of 1 to one of the cells in that row at random (as we did in 5.2.2.1). However, if we carry out all other possible combinings and deletions at step 11 and choose that with the smallest M^2, then we obtain the results in the table.

Table 6.4 Category Groupings of Early Stone Age Tools

Step	M^2 (× 10^3)	Categories
1	24.594	{1} {2,9} {3} {4} {5} {6} {7} {8} {10} {11} {12} {13} {14} {15} {16}
2	24.521	{1} {2,9} {3} {4} {5} {6} {7,13} {8} {10} {11} {12} {14} {15} {16}
3	24.607	{1} {2,9} {3} {4} {5, 6} {7,13} {8} {10} {11} {12} {14} {15} {16}
4	25.300	{1} {2,9} {3} {4,8} {5,6} {7,13} {10} {11} {12} {14} {15} {16}
5	26.582	{1} {2,9} {3,7,13} {4,8} {5,6} {10} {11} {12} {14} {15} {16}
6	29.291	{1} {2,9,10} {3,7,13} {4,8} {5,6} {11} {12} {14} {15} {16}
7	31.051	{1,12} {2,9,10} {3,7,13} {4,8} {5,6} {11} {14} {15} {16}
8	35.400	{1,12} {2,9,10} {3,7,13} {4,8} {5,6} {11,15} {14} {16}
9	40.850	{1,5,6,12} {2,9,10} {3,7,13} {4,8} {11,15} {14} {15}
10	53.323	{1,3,5,6,7,12,13} {2,9,10} {4,8} {11,15} {14} {16}
11	63.793	{1,3,5,6,7,12,13} {2,9,10} {4,8} {11,15,16} {14}
12	86.833	{1,3,5,6,7,12,13} {2,9,10,14} {4,8} {11,15,16}

A scree-plot of the results in Table 6.4 is produced in Figure 6.10 and we include M^2 = 0 to allow for the possibility that no categories are combined or deleted.

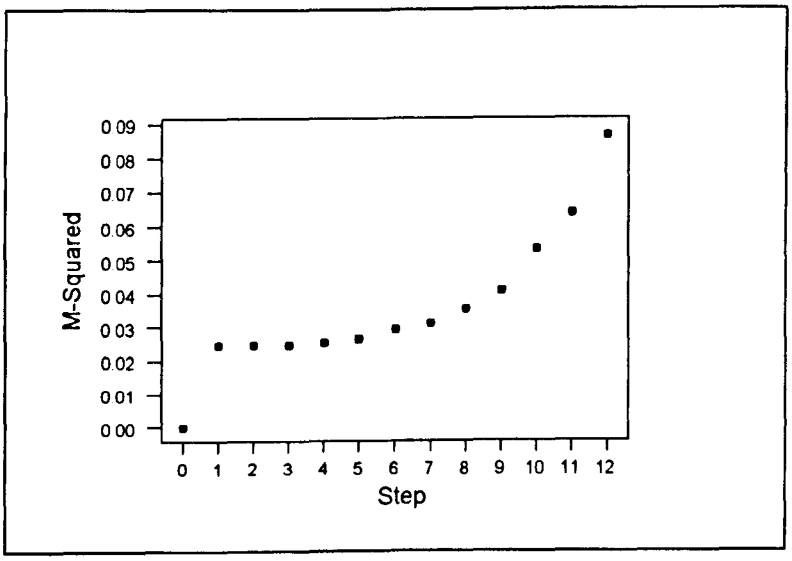

Figure 6.10 Scree-plot for the Early Stone Age Tools (Procrustes Analysis)

From the above plot we conclude that no categories should be combined or deleted and this is because of the large difference in scale between the M^2 at step 0 and the remaining steps. The interpretation of the plot is that combining any categories loses considerable information, but once combining has begun little subsequent information is lost until approximately step 9. A correspondence analysis of the categories obtained as a result of stopping the category selection process at step 9 produces a figure that is very similar to that obtained when all tool groups are considered separately (Figure 6.6a). This contrasts with Figure 6.6c where mergings based on archaeological expertise produce a different pattern of sites and so it now seems that there may be some justification for combining categories on purely statistical grounds. It is also interesting to note that in both our examples categories have always been combined but never deleted.

So far, implementing correspondence analysis on the categories identified by the scree-plot has lead to very similar maps to those of the original data. However, it may be that we can stop at any point on the scree-plot and still obtain a similar ordination map. In order to investigate this issue further, we also implement correspondence analysis using the categories of step 11 in Table 6.4. The results suggest that the patterns between the sites will remain similar regardless of where in the scree-plot we stop the selection process. The advantage of this is that if, for some reason, we are only able to collect information on a subset of tool categories, then we can be confident that our data are not too sparse to mask the relationships between the sites which would be revealed with a larger number of tool categories.

6.7.2 Other Methods of Combining Categories

Having discussed a number of category selection methods in the previous sections, we believe that there are other methods that should be considered.

[1] Firstly, backward elimination procrustes analysis could be implemented as in Section 6.7.1, but rather than comparing each set of row co-ordinates with the original co-ordinates, the co-ordinates could be compared with those of the previous step. Based on the description at the beginning of Section 6.3, we propose using weights equal to the masses of the row categories at the previous step, because these rows are now our reference co-ordinates. This allows us to more carefully monitor the selection process and by using a scree-plot we can obtain a visual assessment of the process.

[2] Secondly, backward elimination procrustes analysis could be used to allow not only for the possibilities of combining and deleting categories, but also for the possibility of leaving them unchanged. For this to be implemented, the co-ordinates would have to be compared with the original set and leaving categories unchanged could not be included in the first step.

[3] Thirdly, we could implement a forward selection method, allowing for the possibilities of deleting and combining categories. At the first step we would choose two individual categories, or a pair of combined categories and an individual category, with the smallest M^2 when compared with the original co-ordinates. For subsequent steps we would choose the option that produced the largest M^2 when co-ordinates are compared with those of the previous step.

[4] A stepwise method of selecting category combinations could also be implemented, using a combination of the backward elimination and forward selection methods and comparing co-ordinates with those of the previous step.

[5] Finally, an all subsets approach could be applied. All possible combinations of categories could be computed; the co-ordinates from each combination can be compared with the original co-ordinates and that with the smallest M^2 chosen.

We believe that the ideal choice in one sense would be method [5], because all category combinations are considered. However, this is the most time consuming method. A stepwise method could therefore be recommended because this compares the co-ordinates at each step with those of the previous step and if we also introduce critical values then the option of leaving categories unchanged is automatically included, although this may suggest that we are seeking one unique set of categories rather than any of several subsets (see 6.3.5).

6.7.3 Measuring Stability by Area

Methods of assessing the stability of the categories in the correspondence analysis map were explained in Chapter Five and included calculating the areas of convex hull peels and concentration ellipses. By implementing these methods on the combined categories we can investigate the effects of altering the original category groupings on stability.

6.7.3.1 Application to Amarna Pottery Sherds

Using the Amarna ware groupings obtained at step 4 of Table 6.3, we generate replicate matrices using the multinomial distribution, as we did in 5.2.1.1 and apply convex hull peeling to the resulting site co-ordinates. The methodology of Section 5.4.5 is followed.

Table 6.5 Approximate Percentages of Points in the Hulls

	Number of Points in the Bootstrap		
	100	**1000**	**5000**
% Hull	25%	13.0-25.3%	12.7-22.8%
	75%	59.5-72.4%	50.8-70.5%
	95%	83.0-95.5%	74.8-94.1%
	100%	87.0-99.2%	83.3-98.5%

Considering Table 6.5, the percentages of points in the hulls from 1000 and 5000 bootstraps are generally slightly lower than those in Table 5.5, but there are no major differences. Thus, any given site exhibits similar stability regardless of the number of categories used in the analysis.

6.7.4 The Influence of Sample Size

Section 5.6 of Chapter Five discussed the influence of sample size on both the stability of categories in the correspondence analysis map and on the relationships between categories. By using our combined categories we can also assess the effects of altering sample size on the results of the analysis. We do this by implementing sampling without replacement and comparing the results with those obtained from the original categories.

6.7.4.1 Application to Amarna Pottery Sherds

Using the category groupings obtained at step 4 of Table 6.3, we sample the Amarna sherds (1.2.2) without replacement 100 times for varying sample sizes. The sample sizes consist of differing proportions of the original sherds obtained from each site and the area of the 95% concentration ellipse, A_4, is calculated for each of them (see 5.4.4.2). The values are given in Table 6.6 below.

Table 6.6 The Measure $A_4(\times 10^3)$ for Varying Sample Sizes

	Sample Size (proportion of original)			
Site	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
1	0.431	1.141	3.108	8.659
2	1.660	4.801	15.188	41.370
3	0.997	2.644	7.840	18.766
4	6.367	20.687	55.789	137.565
5	1.490	4.404	11.292	27.088
6	0.058	0.138	0.475	1.306
7	0.272	0.727	2.850	5.161
8	0.087	0.307	1.039	2.228
9	0.129	4.305	10.741	27.300
10	0.113	0.353	1.086	2.415
11	0.680	2.809	8.042	19.892
12	3.988	15.114	44.275	92.514

Comparing with Table 5.3, we see that the values in the above table are generally slightly higher than those for the original categories, suggesting that there is greater instability in the sites when fewer categories are present. This appears to be true across all sample sizes and the smaller the sample size the greater the instability. This seems reasonable, despite the fact that we have the same total number of sherds, because some information has been lost by combining categories.

For a given proportion of the original sample size obtained at each site, say half, we can calculate A_4 at each step of Table 6.3 in order to compare the stability of different category groupings. The results are illustrated in Table 6.7. From the scree-plot of Figure 6.9 we see that the slope is relatively flat between steps 1-4 and we might therefore expect the value of A_4 to be little changed in this range.

Table 6.7 The Measure $A_4(\times 10^3)$ for Varying Category Groupings ($\frac{1}{2}$ the original Sample)

	Step						
Site	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1.454	1.431	1.434	1.141	1.064	0.000	0.000
2	5.286	5.278	5.304	4.801	4.752	5.267	5.471
3	2.787	2.833	3.106	2.644	1.870	1.888	0.009
4	22.014	21.945	22.171	20.687	1.928	8.359	0.009
5	3.666	3.740	3.708	4.404	3.539	1.646	0.000
6	0.310	0.230	0.231	0.138	0.049	0.308	0.001
7	0.752	0.750	0.717	0.727	0.591	0.460	0.000
8	0.306	0.292	0.293	0.307	0.266	0.000	0.000
9	3.788	3.783	3.557	4.305	3.372	1.231	0.015
10	0.479	0.480	0.436	0.353	0.259	0.148	0.000
11	2.036	2.056	2.173	2.809	1.904	2.772	0.000
12	12.607	12.682	12.696	15.114	12.844	4.057	0.000

Reading across the rows of the table, it is clear that the stability of each site (as measured by A_4) is fairly constant for the first four steps, but that after this point, with fewer categories, stability is increased (i.e. the values in the table fall). This greater stability must, however, be contrasted against the information lost when dealing with fewer categories.

6.7.5 Comparing Methods of Clustering Categories

In this section we propose comparing the results of combining categories using procrustes analysis as described in 6.7.1 (without allowing for category deletion), with the results of clustering using Greenacre's method, described at the beginning of 6.4.

6.7.5.1 Application to Memphis Pottery Sherds

In Chapter Five we explained that for the Memphis sherds (1.2.1), contexts were identified by examining changes in the subsoil (i.e. by using the stratigraphic process) and not by using the arbitrary process of excavation (the arbitrary process is the controlled excavation of the subsoil in measured levels of a predetermined thickness). With this method of excavation it can be very difficult to identify distinct contexts (i.e. where one context ends and another begins) and we therefore propose examining the effect of allowing neighbouring contexts to be combined (i.e. treating them as if they cannot be distinguished). We propose allowing for a maximum of two contexts to be combined, with the justification that it is unlikely that more than two neighbouring contexts would be undifferentiated by the archaeologist. Of course, once excavation has taken place the information is lost and we will never know whether the contexts were reliably identified or not. Our aim is to assess the effect of potential misidentification on the interpretation of the correspondence analysis map. There are 13 contexts for the Memphis sherds and so for the first stage of clustering we compare 12 possible category combinations. The following table reveals how the contexts combine for the two clustering methods.

Table 6.8 Clusterings of Contexts for the Memphis Pottery Sherds

	Method	
Step	Procrustes	Greenacre
1	377 465 509 476 289 690 {716, 739} 740 707 761 758 749	377 465 509 476 289 690 716 739 740 707 {761, 758} 749
2	377 465 509 476 289 690 {716, 739} 740 707 {761, 758} 749	377 465 {509, 476} 289 690 716 739 740 707 {761, 758} 749
3	377 465 {509, 476} 289 690 {716, 739} 740 707 {761, 758} 749	377 465 {509, 476} 289 {690, 716} 739 740 707 {761, 758} 749
4	377 465 {509, 476} 289 690 {716, 739} {740, 707} {761, 758} 749	377 465 {509, 476} 289 {690, 716} 739 {740, 707} {761, 758} 749
5	{377, 465} {509, 476} 289 690 {716, 739} {740, 707} {761, 758} 749	{377, 465} {509, 476} 289 {690, 716} 739 {740, 707} {761, 758} 749
6	{377, 465} {509, 476} {289, 690} {716, 739} {740, 707} {761, 758} 749	

The most obvious difference between the two methods is that a minimum number of seven categories are revealed by procrustes analysis as compared with eight for Greenacre's method. Producing correspondence analysis maps for the most extreme clusterings i.e. when as many contexts as possible are misidentified, reveals that both Greenacre's clustering and the procrustes method produce wares with a similar pattern to Figure 2.2, which suggests that consistently misidentifying a maximum of two contexts does not have serious implications for the relationships between wares and contexts. Greenacre's clustering method has preserved the most important differences between wares (Figure 6.11) and the map is little changed from Figure 2.2.

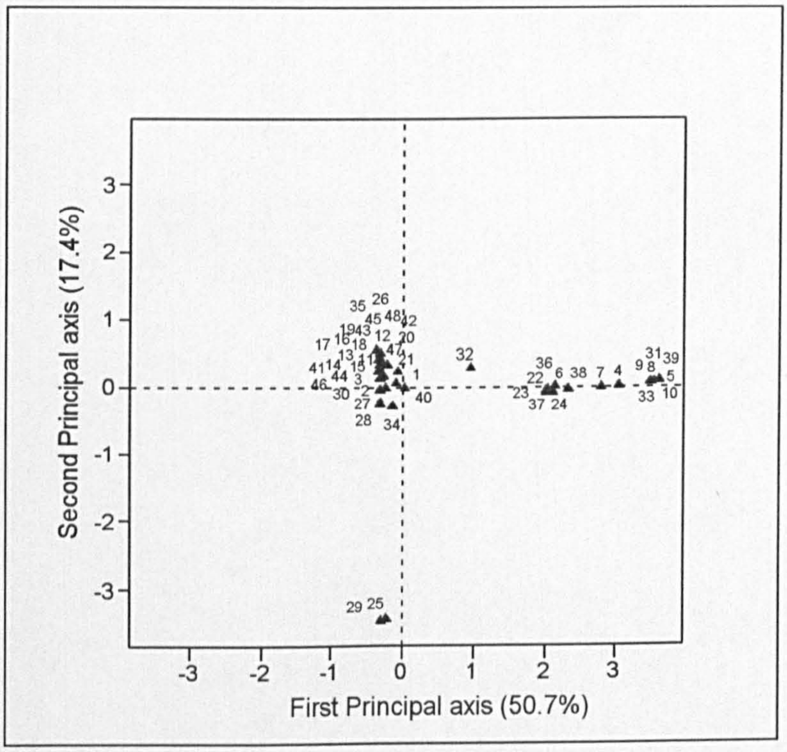

Figure 6.11 Correspondence Analysis Map of Memphis Wares (Greenacre's Clusterings)

The ware groupings obtained from procrustes analysis (not shown) give a fairly similar picture to Figure 6.11, the main differences being that wares 25 and 29 are no longer separated from the remainder and that context 289 is now clustered with other contexts.

6.8 Detecting Influential Categories by using a Jack-knife Approach

In this section we propose using the technique of 'jack-knifing' to identify columns (or rows) that are potentially influential in correspondence analysis and which therefore affect the relationship between rows (or columns). The method is identical to that explained in Section 5.5, but the inferences of the resulting plot are different. We also use a jack-knife approach to detect influential observations in biplots (Chapter Eight) and to identify influential categories in canonical correspondence analysis maps (Chapter Nine).

If we delete each column (or row) category of a contingency table in turn and implement correspondence analysis then we can examine the resulting cloud of row points for unusual points that may indicate unusual column categories (we must use filtering — Greenacre's resampling cannot be used, see 5.5). If removing these column categories affects the relationships between row categories, then the columns are considered to be influential. We discuss how to ascertain which category is 'most influential' in 6.8.2 and we apply this idea of jack-knifing to the Amarna pottery sherds in the next section.

6.8.1 Application to Amarna Pottery Sherds — Influential Wares

Applying the method suggested above, we delete each Amarna ware (1.2.2) in turn, implement correspondence analysis and display the sites. Given that there are 10 wares we should have 10 points for each site, each one representing the deletion of a different ware. A point located some distance away from the bulk of the points within a site indicates a potentially influential ware. The resulting correspondence analysis map is illustrated in Figure 6.12 (note that this is identical to Figure 5.10).

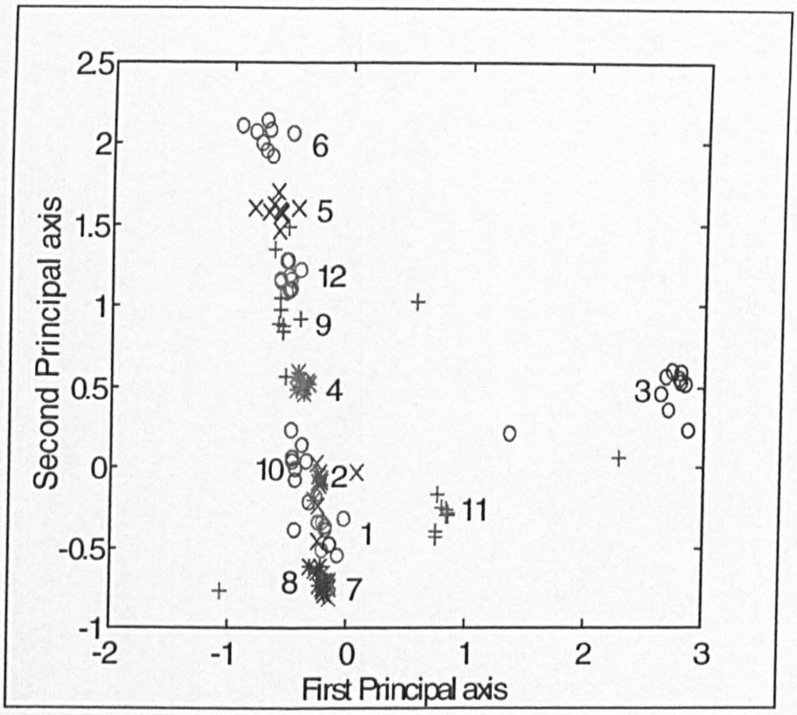

Figure 6.12 Amarna Pottery Site Clouds (Jack-knifing) — Influential Wares

We can see from the above figure that not all the 10 points representing each site are located together. For example, consider site 11: only eight of the points are located together, with an additional one being located at the bottom left of the figure and another one on the right of the figure. The wares represented by these unusual points are wares 9 and 10. However, for other sites, namely 1, 7 and 8, all the points are located close together and none of the wares appear influential. For sites 2, 3, 9 and 10 just ware 10 is located a long distance from the remainder and for sites 4, 5, 6 and 12, ware 8 is located some distance apart.

When we remove the potentially influential wares and implement correspondence analysis, interesting results emerge. Removing ware 8 and, separately, removing ware 10, both cause the resulting correspondence analysis map to change fairly substantially indicating that these are indeed influential wares. Removing wares 9 and 10 together as suggested by site 11 again causes the map to change. However, deleting only ware 9 does not result in any change and this is as we expect because ware 9 alone was not highlighted by any of the sites as being potentially influential. Figure 6.13 shows the correspondence analysis map with ware 10 removed, in order to illustrate the influence of this ware.

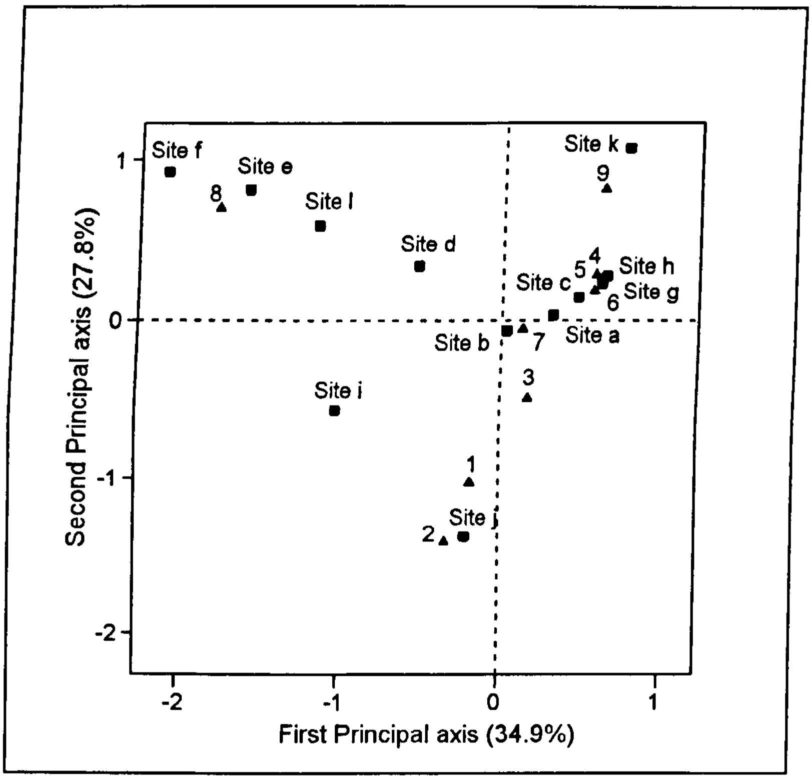

Figure 6.13 Correspondence Analysis Map of Amarna Pottery Sherds (Ware 10 Deleted)

If we compare Figure 6.13 with Figure 2.3 of Chapter Two, we see that site c is no longer separated from the other sites and wares. However, site k and ware 9 are still located a similar distance from the origin and slightly away from the other points. While the relationships between sites and between wares remain similar, the plot is now more spaced out and this time it is sites i & j and wares 1 & 2, which are located some distance apart from the other points.

6.8.2 Application to Amarna Pottery Sherds — Influential Sites

In the previous section we applied the newly introduced jack-knife approach in order to detect influential Amarna wares. In this section we use the jack-knife method to detect influential sites i.e. which sites are most influential in explaining the relationships between pottery wares. We know that, by definition, pottery wares are different, but we would like to know how the correspondence analysis map alters if a particular site is not visited. To investigate this question we delete each site in turn, carry out a correspondence analysis and then display the wares. Given that there are 12 sites we should have 12 points for each ware, each one representing the deletion of a different site. A point some distance away from the bulk of the points within a ware indicates a potentially influential site. The resulting correspondence analysis map is illustrated in Figure 6.14.

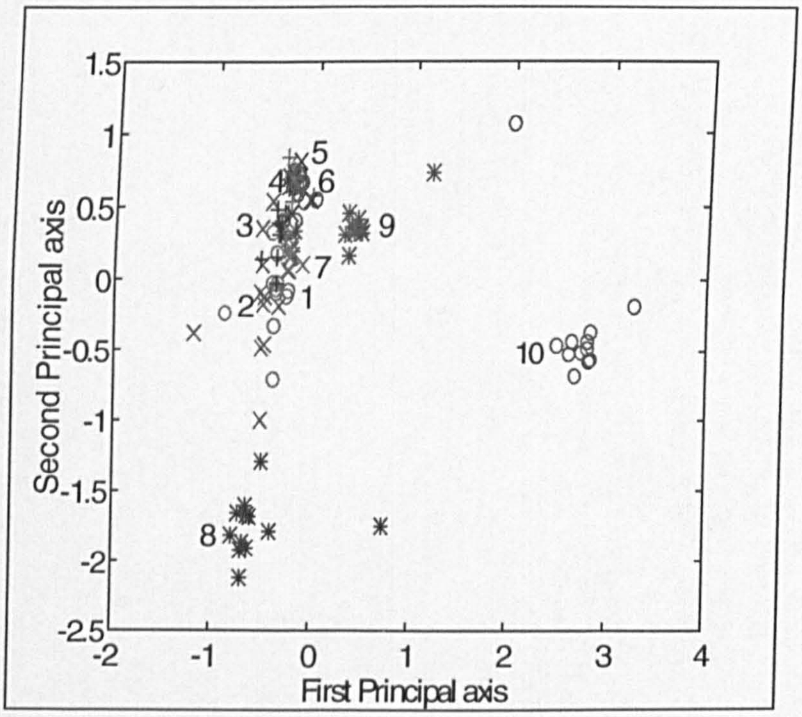

Figure 6.14 Amarna Pottery Ware Clouds (Jack-knifing) — Influential Sites

Considering ware 10, there is one potentially very influential site (the circle at the top of the figure) and one less influential site (to the top right of the majority). These correspond to sites 3 and 11 respectively. Site 3 is also very influential for ware 8, sites 3 and 11 are influential for site 9 and the points representing wares 1 and 2 are very spread out so that there are no sites that are clearly influential. Looking at sites 3 and 11 in the original correspondence analysis map of Figure 2.3, we see that these two influential sites are those which are separated out on the first axis and so perhaps influential column categories tend to be those which are located apart from the majority of column categories on the original CA map. Looking at the bootstrap clouds of Figures 5.4 and 5.5, sites 3 and 11 are no more unstable than any other sites and so using multinomial resampling to obtain bootstrap clouds does not appear to help us in the detection of influential categories. If we delete sites 3 and 11 and implement correspondence analysis we discover that the ordering of the sites from left to right on the first axis of the map is the same as on the second axis in Figure 2.3 and that the same wares are located close to the same sites in both figures. The advantage of this is that our inferences based on the correspondence analysis map are not altered if we are only able to collect data from fewer sites — we can be confident in our interpretations of the ordination diagram.

We now consider how to measure which site (or, more generally, which category) is the 'most influential'. Potentially influential categories can be identified by eye, as we have done in this and in the previous sections, but we propose that the most influential column category could be determined from the jack-knife correspondence analysis map by counting, for each row category, which points (and hence which columns) appear most removed from the bulk of the points. The column with the greatest count is then considered to be the most influential. An alternative and more formal method is to make use of the category deletion methodology of 6.3. Recall that in the first step of the selection process, each column category is deleted in turn and the resulting configurations of row points are compared with the reference configuration, resulting in a procrustes M^2 for each column. With category deletion we look for the smallest M^2, but here we identify that column category with the highest M^2 and this is the 'most influential' category (because the corresponding row co-ordinates differ most from the reference configuration). Implementing both these suggestions for the Amarna sherds leads to site 3 being deemed the 'most influential' site.

We believe that there are other links between the jack-knife method of detecting influential categories and category selection: we might expect that a very influential site would be unlikely to be removed under Krzanowski's backward elimination method, because this latter method is based on removing column categories which result in the least difference between the row co-ordinates and the original configuration. However, Krzanowski's method of category deletion weights the co-ordinates at each stage by the masses from the original correspondence analysis and this could reduce the relationship between the two methods.

6.9 Summary and Conclusions

This chapter has looked at various methods of selecting categories when correspondence analysis is to be used for analysing the data. Firstly, we considered category deletion methods; in particular, Krzanowski's backward elimination method was applied and we made suggestions for adapting it. Specifically, we proposed comparing the co-ordinates at each step of the elimination process with those of the previous step rather than with those of the original data, partly because we believe that this more closely monitors the selection process and partly because this has closer similarities with variable selection methods in regression. We also introduced the use of a scree-plot and a cumulative scree-plot in order to help identify the number of categories to delete (by looking for a change in slope each time a category is removed). These proved to be very successful tools and gave reason to believe that choosing the dimensionality that explains closest to 80% of the variation in the data (as suggested by Krzanowski) is often too stringent: two dimensions are often sufficient. One drawback of the scree-plot is that it may be non-monotonic when low numbers of dimensions are used in the procrustes calculations. However, the cumulative scree-plot overcomes this. We also proposed that the number of categories selected should be chosen independently of the dimensionality used in the calculations (which disagrees with Krzanowski), although it is clear that the higher the dimensionality used in the procrustes calculations, the fewer categories that can be deleted. We believe that the aim of category reduction methods is to identify several subsets of categories rather than one unique set — the all subsets approach is closest to this ideal. Making comparisons with Chapter Five we concluded that there is no relationship between the stability of a site and the first sites deleted in the backward elimination method and so we cannot use the results of one method to make inferences about the results of the other. We also explained that contingency table data do not always consist of grouping categories and observed characteristics as Krzanowski suggested: sometimes there are two sets of observed characteristics.

Secondly, we introduced terminology for distinguishing between combining categories based on archaeological grounds as compared with on statistical grounds. Greenacre's method of clustering categories was applied and the results of this compared with merged categories as defined by an archaeologist. It was clear that for some data sets the expertise of an archaeologist is required before any amalgamation is undertaken (and that some categories should never be combined). We also revealed that large numbers of zeroes in the data affect the deletion and clustering methods, partly because correspondence analysis requires non-zero row and column totals and partly because of the influence of zeroes on clustering methods (but not on the opinion of the archaeologist). In addition, we proposed using correspondence analysis to assess the effect of category division (which is based on external variables) and this was moderately successful. The relative merits of the various methods of category selection were also discussed and we reiterated that sometimes no selection method is appropriate because the given categories are essential in testing a particular hypothesis of the investigator.

In this chapter, we developed a method to account for combining and deleting categories simultaneously and we compared clustering using Greenacre's method with a method based on procrustes analysis, which we introduced — both produced similar results. We also proposed using clustering methods in archaeology in order to assess the effects of misidentifying contexts when the stratigraphic method of excavation is used — the results showed that there are no serious consequences in terms of inferences based on the correspondence analysis map, if two neighbouring contexts are misidentified. After various methods of combining categories were implemented we used the methodology of Chapter Five and calculated the stability of and the influence of sample size on, these categories and compared the results with those obtained from the original categories. It appears that when the data consist of smaller numbers of categories, these categories are less stable. By making comparisons with the backward elimination scree-plot, it is clear that the stability of the categories increases as the slope of the plot rises; where the plot is flat, the stability of the categories remains fairly constant.

Finally, jack-knifing was introduced as a means of detecting influential categories and proved to be a good technique for identifying which categories have a potentially large influence on the ordination diagram. We suggested that the 'most influential' column category could be ascertained by looking for that column with the largest procrustes M^2 at the first step of the backward elimination method (i.e. when each column is removed in turn and the corresponding row co-ordinates are compared with those of the original data). This proved to be very successful.

Chapter Seven

Variable Selection Methods and Biplots

7.1 Introduction

The theory of biplots was explained in Chapter Three and their application to flake debitage, flint tools and ceramic pots was illustrated. Chapter Three also collated information regarding the various forms of biplot, raised the issue of variable selection and explained the relationship between biplots and the more well known technique of principal component analysis. This chapter combines biplots with procrustes analysis, in order to investigate the importance and influence of variable selection when collecting and analysing data. Various methods of selecting variables are introduced and discussed and the influence of the dimensionality used in the calculations is considered. It is clear that there are different issues involved in variable selection for biplots as compared with principal component analysis and this is because biplots are nearly always displayed in two dimensions, whereas the number of principal components tends to be chosen objectively.

The remainder of this section gives a general introduction to the idea of variable selection. Section 7.2 explains the variable selection methods in existence for principal component analysis and presents a critical review of a method introduced by Krzanowski (1987, 1996). A variation of Krzanowski's backward elimination method for principal component analysis is extended to biplots in 7.3 and reasons for its failing for some types of biplot are discussed. The scree-plot and cumulative scree-plot as aids to variable selection are also proposed in this section and the backward elimination method is compared with forward selection, stepwise and all subsets methods, in varying numbers of dimensions. Throughout this chapter, which is concluded in 7.4, the techniques are illustrated on two of the data sets presented in Chapter One and initially investigated in Chapter Three, i.e. the ceramic pots (1.2.5) and Simpson Desert flint tools (1.2.6).

7.1.1 Selecting a Subset of Variables

We saw in Chapter Three with the ceramic pots that it is common in archaeology for many variables to be measured on any given artefact and it is also known that these measurements can be time consuming to obtain. Additionally, some variables can dominate statistical analyses (as we saw with the coefficient of variation biplot in 3.7.1.3) and mask the effects of other variables. There is, therefore, scope for developing variable selection methods for use with biplots in archaeology. When considering variable selection, ease of measurement of variables should also be borne in mind, partly because it is often not possible to take some measurements due to broken or chipped artefacts, but also because some measurements are time consuming to obtain, for example weighing artefacts. The focus of this chapter is on both developing and implementing existing and new methods of variable selection and on investigating the effects of these methods on the relationships between the remaining variables and on the structure of the observations. Jolliffe (1986) explains that when p, the number of variables observed is large, it is often the case that a subset of m variables, with m << p, will contain virtually all the information available in all p variables and thus time can be saved by measuring only m variables.

7.2 Variable Selection and Principal Component Analysis

There are several variable selection methods in existence for use with principal component analysis. The main ones are due to Jolliffe (1972, 1973) and Krzanowski (1987) and these are explained below.

7.2.1 The Work of Jolliffe (1972, 1973)

Two of the first papers to consider variable selection in principal component analysis were those of Jolliffe (1972, 1973). Jolliffe commented that in multivariate analysis when a large number of variables, say 10 or more, are available, then the results are often little changed if a subset of the variables is used, with the remaining variables being considered to be redundant. Jolliffe also observed that variables are often present which complicate the data but which do not contribute any extra information and that time and money are also saved if some of the variables are discarded, computing time is reduced and in future analyses fewer variables need be measured. Jolliffe considered eight rejection methods of variable selection. Two of the principal component methods that he found to be most satisfactory are as follows:

Method One: Carry out a principal component analysis on data matrix X (n × m). If q variables are to be retained, a variable is associated with each of the last (m-q) components. The last (m-q) components are considered consecutively. Starting with the last component, the variable that has the largest coefficient in the component is associated with it, as long as it has not already been associated with a previous component. These (m-q) variables are then rejected.

Method Two: Carry out a principal component analysis on data matrix X (n × m). The first q components are considered successively, starting with the first and the variable with the largest coefficient on a component is associated with it as long as it not already associated with another component. These q variables are retained and the remaining (m-q) rejected.

We should be aware that both the above methods require q to be chosen subjectively.

7.2.2 The Work of Krzanowski (1987, 1996)

Krzanowski (1987) observed that all the variable selection criteria in existence were concerned with overall features, either of the subset data (McCabe, 1984), or of the complete data (Jolliffe, 1972, 1973). Thus, he considered the criteria to be based exclusively on variance-covariance or correlation matrices and their eigenvalues or eigenvectors. Krzanowski therefore suggested that a more appropriate criterion for preserving structure among observations would be one that involved some direct comparisons between the individual points of the subset configuration and the corresponding points of the complete data configuration and he suggested making use of procrustes analysis for this (see 6.3). We describe and discuss his method below, using the following notation, which is taken from Krzanowski (1987):

Let X (n x m) = data matrix of m variables measured on n units, column standardised to zero mean and unit variance;

Y (n x k) = matrix of principal component scores (where $k < m$), yielding the best k-dimensional approximation to the original data configuration;

$\tilde{Z}$ (n x k) = matrix of principal component scores of the reduced data, which contains only q selected variables.

Having defined these matrices, we view Y as the 'true' configuration and $\tilde{Z}$ as the corresponding approximate configuration based on a subset of q variables (we must ensure that sufficient data variability has been explained in the k dimensions). These configurations are then compared using procrustes analysis, which involves finding the sum of squared differences between corresponding points of the two configurations after they have been matched as well as possible under translation, rotation and reflection. The residual sum of squares, M^2, then measures the loss of information about the data structure when only q variables are used, instead of all p variables. The 'best' subset of q variables is the subset that yields the smallest M^2 among all q-variable subsets. However, as with Jolliffe's methods, q needs to be chosen subjectively. Krzanowski suggests choosing the dimensionality of the data, k, by either using the cross-validatory techniques of Wold (1978) or Eastment & Krzanowski (1982), or by convenience (e.g. because the data will be displayed in two dimensions). However, in the latter case, care must be taken to ensure that sufficient data variability has been accounted for in the chosen dimensions.

We question whether the 'true' co-ordinates, which act as the reference set in this method, are the most appropriate co-ordinates to use. In Krzanowski (1987), each time a variable is deleted the resulting co-ordinates are always compared with the original set. However, we are not convinced that this is the most sensible approach, because it differs to the analogous procedures in backward elimination and stepwise regression. With these methods, each time a variable is deleted an F-statistic is computed based on a comparison with the previous step, rather than with the original data. We believe that there are arguments in favour of both methods and we illustrate both of these on different data sets in Section 7.3.

7.2.2.1 A Stopping Rule for Structure-Preserving Variable Selection

Previously, when using variable selection methods, a subjective decision had to be made on how many variables to retain. Krzanowski (1996) introduced some objectivity into the process by providing a stopping rule for the backward elimination method, based on the procrustes residual sum of squares, M_i^2, when i variables have been removed. He considered a stepwise method to be too time consuming in the case of large numbers of variables (which are often present in archaeology), although we consider this possibility for the flint tools in 7.3.2.7.

If we measure m variables on each of n observations and we are interested in k dimensions, then Krzanowski (1996) claims that the procrustes residual sum of squares in the backward elimination process, M_i^2, when i variables have been removed will, if the omitted variables are not structure-carrying, approximately follow a $(1 + c^2)\sigma^2\chi_r^2$ distribution, where $r = nk - \frac{1}{2}k(k+1)$ and $c = \sqrt{(m-i-k)/(m-k)}$. If some of the omitted variables are structure-carrying, then the residual sum of squares will be inflated and we continue deleting variables until the calculated M_i^2 exceeds some critical value. However, σ^2 is unknown and so in practice we will need to replace it by an estimate from the data. Krzanowski suggests that a suitable estimate is given by:

$$\hat{\sigma}^2 = \frac{1}{d}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \sum_{t=1}^{k} u_{it}s_t v_{tj}\right)^2$$

where d = (n-k-1)(m-k) is the number of degrees of freedom left after fitting k principal components and $u_{it}s_tv_{tj}$ are the elements from the singular value decomposition of matrix X. The critical value chosen is a suitable percentage point of the relevant χ^2 distribution, although interpretation of the 'significance level' is complicated by the sequence of repeated 'tests'. Krzanowski (1996) briefly discusses this issue and suggests relaxing the probability levels (i.e. reducing the percentage point) when setting the critical values.

We question whether a formal test is necessary because we believe that there may well be several combinations of variables that are able to distinguish between groups of observations, rather than one 'unique' combination. If this is true then we can use other information (e.g. ease of measurement) to select which combination to use. Additionally, it may be the case that one variable cannot be obtained in a particular study and so to know of other alternative subsets of variables that retain group structure amongst the observations could prove to be important.

7.2.2.2 Choice of Dimensionality

The choice of dimensionality influences both M_i^2 and the percentage point obtained from the chi-squared distribution and so it is important to investigate how this influences the variables selected. However, this must be balanced against the visually appealing nature of two-dimensional displays. Krzanowski (1996) says that underestimation of the dimensionality k means that $\hat{\sigma}^2$ will be too large, which will produce critical values that are too large and hence too few variables may be retained as 'important'. However, he goes on to say that the additional deletions will be the more marginal ones and that the most important variables should not be missed. Krzanowski also believes that overestimation of k will produce a slightly smaller $\hat{\sigma}^2$

and that the final results should not be greatly affected. In our examples in this chapter we consider two-dimensional and three-dimensional space.

7.3 Variable Selection and Biplots

Given the close relationship between principal component analysis and the principal component biplot (see Section 3.10.1), we extend Krzanowski's method of variable selection to this form of biplot and also to the covariance, correlation, coefficient of variation and Spearman rank correlation biplots. We showed in Chapter Three that biplots are more useful than principal component analysis for exploratory analysis because they enable us to display both observations and variables simultaneously. If we wish to use a biplot for our analysis, then we believe that it is more sensible to use it in the variable selection process as well.

In archaeology, variables are often measured on artefacts with the aim of detecting and discriminating between two or more groups (Barton, *pers. comm.*) and thus we consider Krzanowski's method to be more suitable for reducing the number of variables measured than previous methods, because it is based on comparisons between individual observations. However, if we wish to analyse our data using a biplot rather than principal component analysis (for which the method was developed), then it is important to consider whether this method is still reliable. We test this method of variable selection on the observation (row) co-ordinates from the covariance, correlation, coefficient of variation, Spearman rank correlation and principal component biplots, because these correspond to the scores in a principal component analysis. We do this for both the ceramic pots and the Simpson Desert flint tools because, as we saw in Chapter Three, we expect groups to exist in both cases. We also consider these data to be typical of archaeological studies (pottery and flint are the most common artefacts found) and so there is reason to believe that any inferences can be applied more generally.

7.3.1 Application to Ceramic Pots

In this section we introduce a variation of Krzanowski's backward elimination method and apply it to the ceramic pots, which were described in 1.2.5. We recall that the data consist of 13 measurements taken on each of 30 ceramic pots, made by three potters.

7.3.1.1 Backward Elimination using Critical Values (Two Dimensions)

We assume a two-dimensional representation, a 5% significance level and apply an adaption of Krzanowski's backward elimination method of variable selection. Rather than comparing the co-ordinates at each step of the elimination process with those obtained from the original data, we propose comparing them with those from the previous step, although we still remove the variable which results in the lowest M_i^2 at each step. This means, however, that using the critical values from Krzanowski (1996) may not be strictly appropriate, but we use them as an initial guide. We choose to vary the reference set of co-ordinates at each step of the elimination process because we believe that this is the best way of monitoring the process. Table 7.1 lists the order in which the variables were deleted for each form of biplot (explanations of the measurements corresponding to each number are given in Chapter One). We recall, from Chapter Three, that over 69% of the variation in the data is explained in the first two dimensions, for all forms of biplot.

Table 7.1 Krzanowski's Method of Variable Selection for the Ceramic Pots (Two Dimensions)

Biplot	Order in which Variables Deleted	Variables Retained
Covariance	13 11 12 4 8 10 9 2 6	1 3 5 7
Correlation	3 6 7 4 5 12 2 11 10 13 1	8 9
Coefficient of Variation	7 6 2 3 9 5 8 1 4 10 11	12 13
Spearman Rank	8 11 10 1 3 13 9 7 4 2 12	5 6
Principal Component	11 2 3 13 6 9 5 4 8 1 12	7 10

We see from the table that the coefficient of variation biplot removes 11 variables, although this is the form of biplot that did not show any obvious pot groupings when all 13 variables were used (see Figure 3.4). For the correlation, principal component and Spearman rank correlation biplots 11 variables are also deleted (algebraically we cannot remove more than 11), although these are different in each case. The covariance biplot deletes nine variables. The reason why different variables were deleted in each case is because of the different pre-scaling used in each type of biplot (see 3.4 and 3.5) — each form of biplot is important in different situations and measures different aspects of the data. However, because this is a comparative study, we have applied variable selection to five types of biplot, but this would not usually be acceptable.

We now compare the biplots which use only the variables retained after the backward elimination procedure, with those biplots obtained from using all the variables (illustrated in Chapter Three), in order to assess whether it is possible to employ this method of variable selection with this type of data and still retain pot groupings where these exist (e.g. no groupings occurred in the coefficient of variation biplot, so we should not really be surprised if none occur using fewer numbers of variables). Figure 7.1 illustrates the covariance biplot and we see that there is a very good separation between one group of pots, on the bottom left and the remaining pots, with the other two groups being reasonably separated from each other (those below the vector representing variable 5 form one group corresponding to one potter and those above form another group).

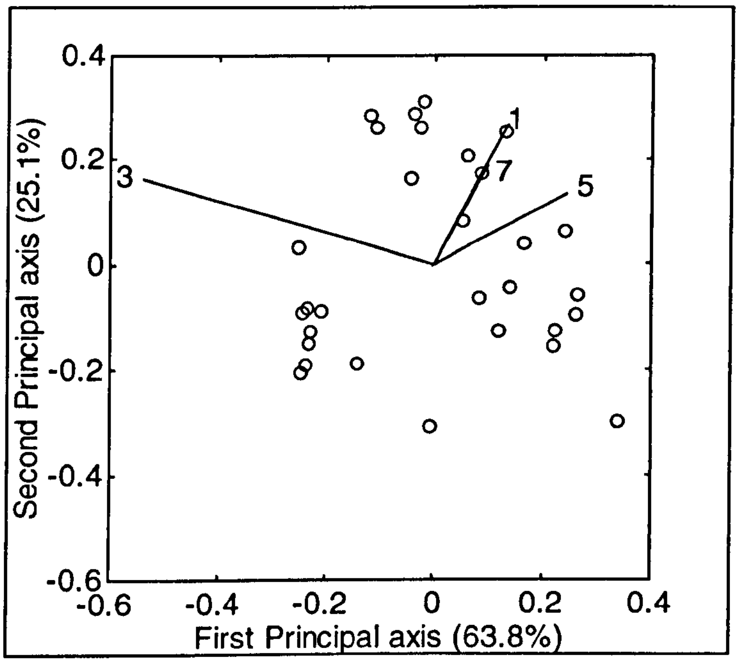

Figure 7.1 Covariance Biplot of Ceramic Pots (Backward Elimination)

Comparing the above figure with the covariance biplot that used all 13 variables (Figure 3.2), the separation of pots into three groups in Figure 7.1 is poor and so the backward elimination method has not selected appropriate variables. This could be either because the critical values are wrongly set, because comparing the co-ordinates with those of the previous step is inappropriate, or because the method itself is not very good at selecting variables for biplots (recall that it was originally introduced for principal component analysis).

We now turn to the correlation biplot in Figure 7.2, which shows that it is not possible to distinguish three groups of pots using only variables 8 and 9 (diameter at point of angle and external diameter of footring at base, respectively). Figure 3.3 illustrates a much better division of pots into groups, using all 13 variables.

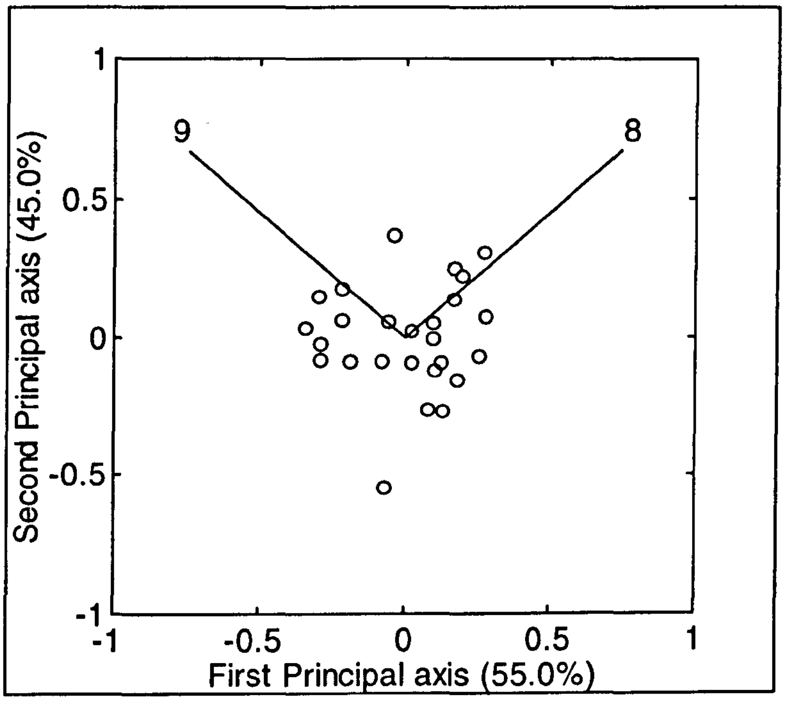

Figure 7.2 Correlation Biplot of Ceramic Pots (Backward Elimination)

The coefficient of variation biplot was also obtained but it is still not possible to distinguish between three groups of pots in this figure, using variables 12 and 13 (thickness of wall at 2cm from base and thickness of lip respectively), although it is no worse than Figure 3.4 which used all 13 variables.

The Spearman rank correlation biplot (not shown) indicates that it is no longer possible to separate out the three groups of pots when using just variables 5 and 6 (internal diameter at lip and overall height respectively), although one group of pots is fairly well separated from the remainder. Figure 3.5, which used all the variables, provides a better separation of pots into groups.

The principal component biplot is illustrated in Figure 7.3 and this produces a separation of the three groups of pots which is nearly as good as that obtained using all 13 variables (see Figure 3.6), although there is an unusual observation at the top of the picture which belongs to the group on the middle left.

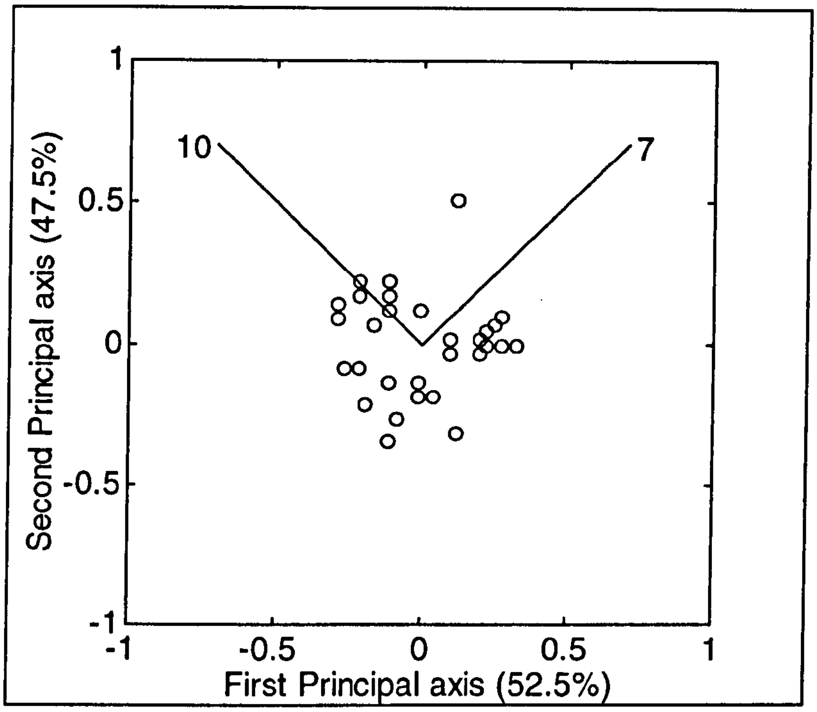

Figure 7.3 Principal Component Biplot of Ceramic Pots (Backward Elimination)

Because the row co-ordinates (pots) are the same as the scores in principal component analysis, which is what the technique was originally developed for and because the principal component biplot is the only biplot to produce as good a group separation of pots after variable selection, we believe that Krzanowski's technique may need some adaption for the other biplots.

7.3.1.2 Reasons for the Breakdown of the Backward Elimination Technique

In this section we suggest four reasons for the breakdown of the backward elimination technique.

[1] We propose that one reason why the technique does not work well on biplots from the Correlation Biplot Family could be due to the nature of the row and column factorisation. In this family the inter-row distances are poorly represented (see 3.6.3) and thus it is possible that these biplots are unable to separate out the distances between observation groups appropriately (recall that the procrustes criterion is based purely on observation differences). In the principal component biplot, however, the inter-row differences are well represented. When all 13 variables were used in Chapter Three, all biplots except the coefficient of variation biplot were able to identify three groups of pots, which could suggest that the procrustes method itself is flawed as a means of selecting variables.

[2] It is possible that the critical values are inappropriate because we are comparing each set of co-ordinates with those obtained from the previous step in the backward elimination procedure, whereas Krzanowski (1987) uses the original co-ordinates as the reference set.

[3] The critical values may be wrongly set and so too few variables are retained in biplots of the Correlation Biplot Family. This problem can be overcome by altering the critical values, or by using a scree-plot or cumulative scree-plot that we introduce in 7.3.1.3.

[4] It is possible that two dimensions are insufficient in this variable selection problem, although they were sufficient when all 13 variables were used in Chapter Three.

We discuss points [3] and [4] in more detail below. However, we believe that the principle of comparing row points (observations) using procrustes analysis is just as applicable to biplots as it is to principal component analysis. Bearing in mind

Krzanowski's comments on variable selection in 7.2.2.1, we re-evaluate the critical points for all the biplots using the 10% significance level. However, it turns out that the variable selection remains the same in all cases, although this may be because we compared the co-ordinates with those of the previous step.

7.3.1.3 The 'Scree-plot' and 'Cumulative Scree-plot' (Two Dimensions)

We believe that the critical values against which the procrustes M_i^2s are compared are an initial guide to selecting variables, but that they need fine-tuning and so we introduce a scree-plot and cumulative scree-plot as alternatives, which may be more informative (although less formal). These operate on a similar basis to the scree-plots used in principal component analysis to choose dimensionality and so we stop deleting variables at the point where there is a 'kink' in the graph. On the vertical axis we plot the values of M_i^2 at each elimination step (or the cumulative sum of the M_i^2 across the steps) and the corresponding variables deleted at each step are shown on the horizontal axis. We can allow for the possibility that no variables are deleted by including $M_i^2 = 0$. Making an analogy with Jolliffe (1986), we also investigate using log (M_i^2), but this does not give good results. Figure 7.4 illustrates the scree-plot for the covariance biplot after using our adaption of the backward elimination method.

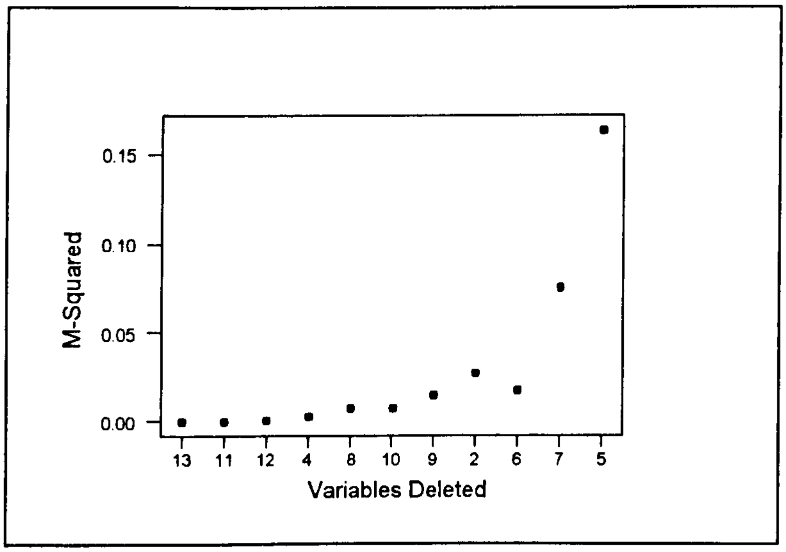

Figure 7.4 Scree-plot for the Ceramic Pots (Covariance Biplot: Two Dimensions)

It is evident from the above figure that there is clearly an anomaly at the point where variable 6 is deleted (step 9) and this means that the plot is not monotonically increasing. If we think of the vertical axis as a goodness of fit measure with the bottom being the best fit (i.e. all variables included, $M_i^2 = 0$) and the top the worst fit, then this plot is effectively saying that deleting variable 6 as well as the previous eight variables improves the fit. This is clearly inappropriate unless we believe that variable 6 is masking the effects of other variables. Examining the scree-plot for the correlation biplot also indicates an anomaly at step 9, i.e. the plot is non-monotonic, but this time it is when variable 10 is deleted.

We believe that one possible explanation for these anomalies could be that the scaling on M_i^2 is inappropriate; another explanation could be that the dimensionality is not sufficiently high: these are effects that would be missed if we just used Krzanowski's critical values with a chosen dimensionality. The scree-plots for the coefficient of variation and Spearman rank correlation biplots (not illustrated) also show that something anomalous occurs at step 9 and for the former biplot also at step 4. If, instead, we use a cumulative scree-plot (i.e. we sum the values of M_i^2 cumulatively across the steps) then we obtain, for the covariance biplot, Figure 7.5.

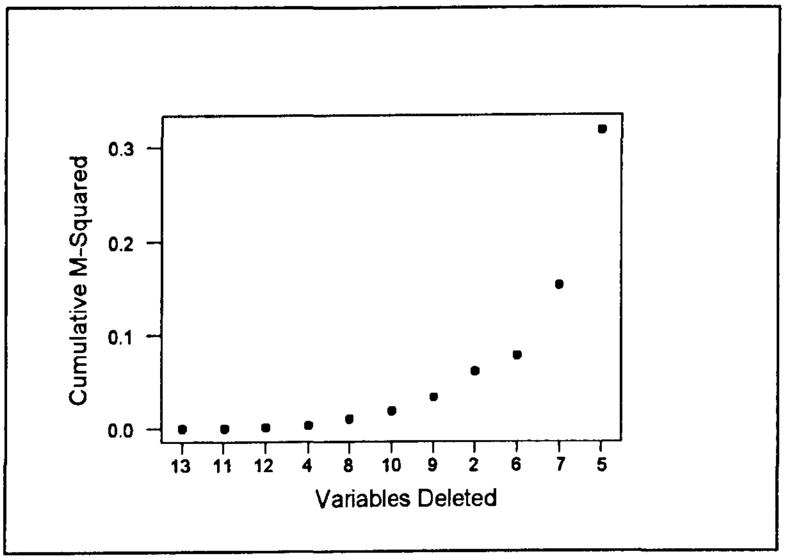

Figure 7.5 Cumulative Scree-plot for the Ceramic Pots (Covariance Biplot: Two Dimensions)

From the above figure we stop deleting variables either after variable 9, or after variable 6 and so the cumulative plot has allowed us to make a decision on the number of variables to retain, where the scree-plot could not. Considering the principal component biplot, we see that the scree-plot is monotonically increasing and we delete eight variables because it is at this point that the M_i^2 increases considerably (but, if we include $M_i^2 = 0$ in the plot, then no variables are deleted). The principal component biplot on the remaining five variables is shown in Figure 7.6.

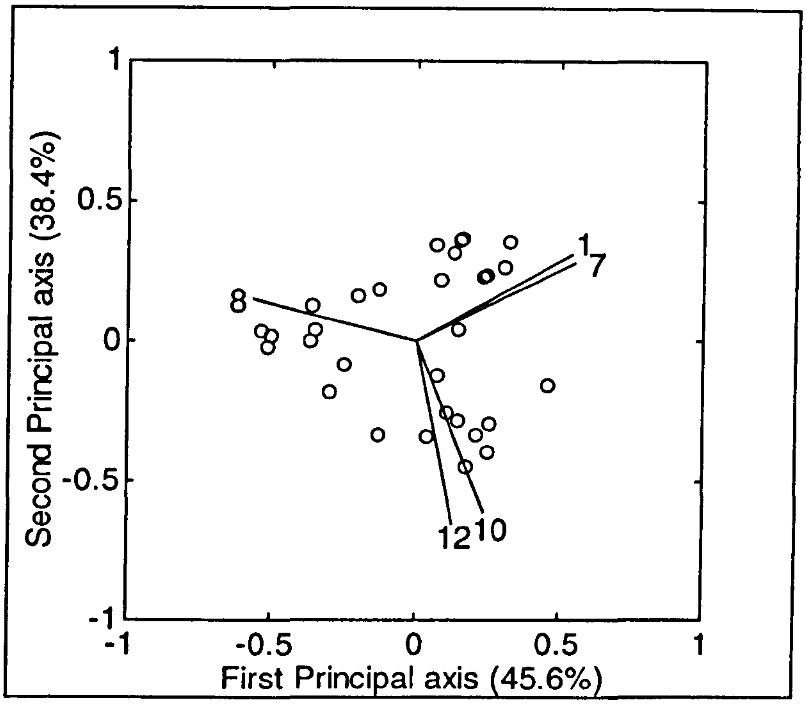

Figure 7.6 Principal Component Biplot of Ceramic Pots (Backward Elimination)

We see from the above figure that retaining five variables has not improved the separation of pot groups over that obtained by retaining just two variables (Figure 7.3) and the group separation is considerably worse than when all 13 variables are included (Figure 3.6). It therefore appears that in the two-dimensional case Krzanowski's critical values perform better than the scree-plot and cumulative scree-plot in terms of choosing variables for the principal component biplot. This is not the case for the other forms of biplot.

7.3.1.4 Backward Elimination using Critical Values (Three Dimensions)

In this section we apply, for comparative purposes, the backward elimination method using three dimensions (rather than two). We use a 5% significance level, compare the co-ordinates with those of the previous step and display the plots in two dimensions. Table 7.2 indicates the variables retained for each form of biplot and these can be compared with those in Table 7.1.

Table 7.2 Krzanowski's Method of Variable Selection for the Ceramic Pots (Three Dimensions)

Biplot	Order in which Variables Deleted	Variables Retained
Covariance	13 11 12 4 6 10 8 9	1 2 3 5 7
Correlation	6 5 3 10 8 1 4 12 7 2	9 11 13
Coefficient of Variation	6 4 2 9 8 7 1 10 5 3	11 12 13
Spearman Rank Correlation	8 3 1 9 6 4 5 12 2 7	10 11 13
Principal Component	8 6 10 3 1 7 4 5 12 2	9 11 13

From Table 7.2 it is evident that one more variable is included in all forms of biplot when three dimensions are used in the selection process, compared with when two dimensions are used. We believe that this is most likely to be a direct result of the dimensionality because, by definition, the least number of variables that we can retain in a two-dimensional plot is two and in a three-dimensional plot is three. We also believe that the most likely explanation for this 'dimensionality effect' is that the critical values are not appropriate. Critical values were introduced by Krzanowski (1996) to give some formality to the selection process, but they do not appear to be satisfactory for the various forms of biplot and we dispute whether they are useful in problems such as these.

Using the variables retained in Table 7.2, none of the biplots show any improvement in separation of pots into groups compared with those obtained from two dimensions and all plots are worse than when all 13 variables were used. In summary, Krzanowski's backward elimination method is no more successful when three dimensions are used in the selection process, compared to when two dimensions are used.

7.3.1.5 The 'Scree-plot' and 'Cumulative Scree-plot' (Three Dimensions)

In this section, as in Section 7.3.1.3, we examine the scree-plot and cumulative scree-plot as alternatives to using the critical points of Krzanowski (1996), but this time we use three dimensions in the procrustes calculations. The scree-plot for the covariance biplot is illustrated in Figure 7.7 and we see that it is monotonically increasing. However, it is not convex and there are two possible reasons for this. Firstly, whilst deleting variables worsens the fit in terms of the M_i^2, there is no reason why the deletion of each subsequent variable should reduce the fit more than the previously deleted variable. Secondly, the non-convexity could be because the dimensionality used in the calculations is too low. It appears from the plot that eight variables should be deleted, because it is at this point that the kink in the graph occurs and in fact this agrees with the number of variables deleted using Krzanowski's critical values. However, the resulting pot groupings are not really satisfactory and we know from Figure 3.2 that when all 13 variables are used it is possible to distinguish between three pot groups.

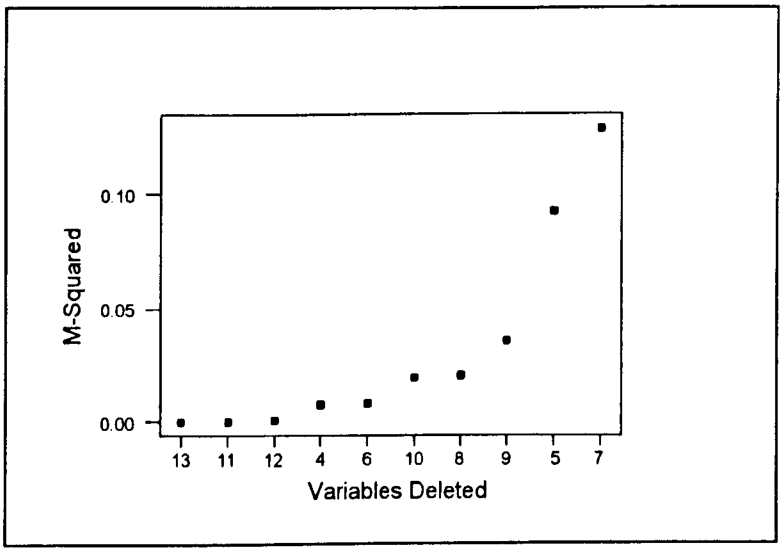

Figure 7.7 Scree-plot for the Ceramic Pots (Covariance Biplot: Three Dimensions)

Considering the correlation biplot, the resulting scree-plot is also monotonically increasing and we would again delete eight variables, whereas using Krzanowski's critical values we deleted ten variables. Deleting these eight variables and producing a biplot provides a slightly better distinction between pot groups than that which we obtained from using critical values, although it is not as good as Figure 3.3 where all 13 variables are used. The scree-plot for the coefficient of variation biplot is still not monotonically increasing (— we could use the cumulative scree-plot), but the scree-plot for the Spearman rank correlation biplot (not shown) indicates that we should delete seven variables (which corresponds to the 10 deleted using the critical values of Krzanowski). A biplot carried out on the remaining six variables is illustrated in Figure 7.8 and similar pot groupings are shown to those obtained using two variables in two dimensions, although again information on pot groupings has been lost from when all 13 variables are used.

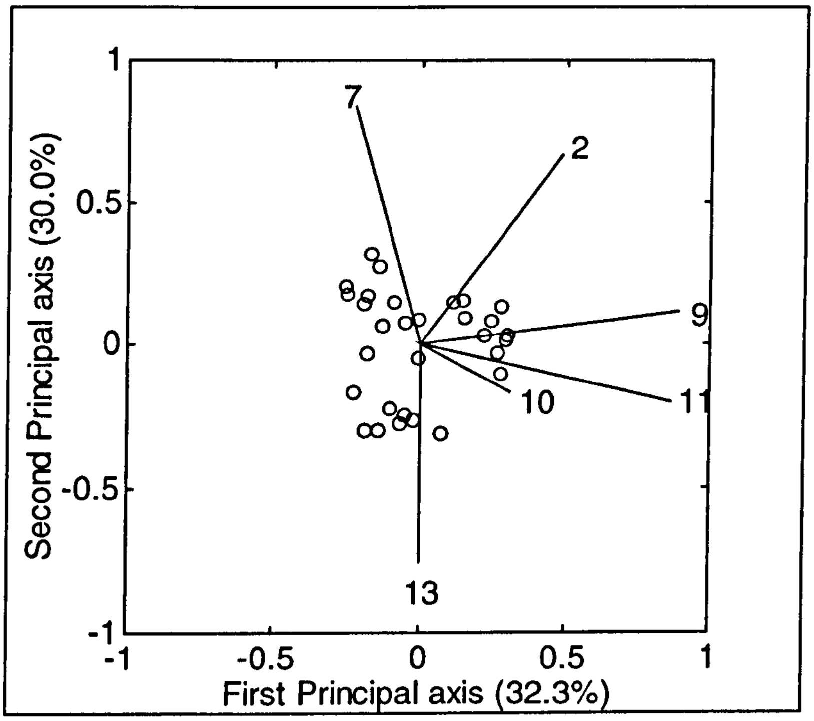

Figure 7.8 Spearman Rank Correlation Biplot of Ceramic Pots (Backward Elimination)

Even though the scree-plot for the principal component biplot is monotonically increasing in two dimensions, it is not monotonically increasing in three dimensions (not shown). This is an important discovery because if we had believed that a monotonically increasing plot implied adequate dimensionality for variable selection then we would not have considered three dimensions (although it could be argued that we should stop considering higher dimensionality when we obtain a monotonically increasing plot and bear in mind that we are going to display the resulting plots in two dimensions regardless. However, we could plot first and third components, say, rather than just first and second). The non-monotonicity of the principal component scree-plot could be due to comparing the co-ordinates at each step of the backward elimination process with those of the previous step, rather than with the original co-ordinates. It therefore seems that either an alternative method of choosing dimensionality is required, or an alternative method of variable selection is needed (i.e. not backward elimination, or not the procrustes statistic). Figure 7.9 shows the cumulative scree-plot obtained when three dimensions are used in the calculations. By looking at the change in slope of the plot, we stop deleting variables after variable 12.

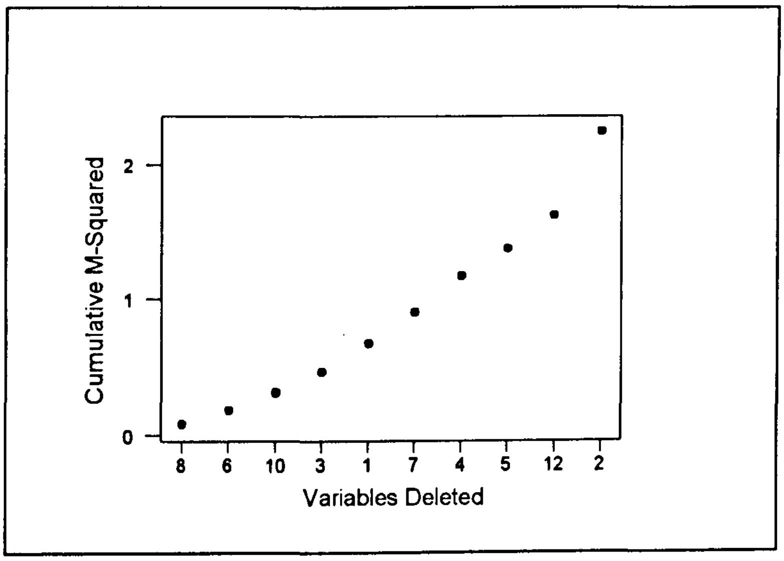

Figure 7.9 Cumulative Scree-plot for the Ceramic Pots (Principal Component Biplot: Three Dimensions)

When all the above work is repeated using the original co-ordinates as a reference and using three dimensions, monotonically increasing plots are obtained for all forms of biplot except the coefficient of variation biplot. However, the resulting biplots show no improvement in separating the pots into three groups. Eastment & Krzanowski (1982) suggest a cross-validation approach for choosing dimensionality, but this is too time consuming in practice and we believe it is too technical for the archaeologist to use unaided.

7.3.1.6 Summary of Ceramic Pots

Variable selection procedures can take the form of backward elimination, forward selection, all subsets and stepwise regression, but for the ceramic pots we only considered the first of these (the others are discussed in the next section for the flint tool data). Within the backward elimination framework, procrustes residual sums of squares (M_i^2) were used as discussed in Krzanowski (1987, 1996), in order to decide on which variable to delete at each step of the process. Krzanowski (1987) used the original co-ordinates as reference co-ordinates, but we suggested, by making an analogy with backward elimination and stepwise regression, that the co-ordinates at each step of the process should be compared with those obtained in the previous elimination step. Both methods have advantages and disadvantages. Krzanowski (1996) introduced critical values as a stopping criterion, with which the procrustes statistic at each step should be compared and we applied these, although we are not convinced of the need for formal testing. This is because we believe that we are looking for any subset of the variables that preserves the data structure, of which there could be many, rather than one unique set of variables.

We recall from Chapter One that the main objective of analysing the ceramic pot data was to see whether it is possible to identify three groups of pots, each corresponding to a different potter, on the basis of the available measurements. A second objective was to investigate whether any groupings were altered by the elimination of some variables. Using two dimensions and applying Krzanowski's critical values to select variables, none of the biplots produced as good a separation of pots into groups as were obtained in Chapter Three using all 13 variables. We proposed several reasons as to the causes of this and introduced a scree-plot and cumulative scree-plot as methods of helping us to select variables. In two dimensions only the principal component biplot produced a monotonically increasing scree-plot, but the group separation of pots was still worse than that obtained using 13 variables. Three dimensions were then used in our variable selection procedure and more variables are retained using the critical values of Krzanowski than are retained in two dimensions for every form of biplot. However, the resulting biplots show no improvement on those obtained using two dimensions. Scree-plots were again used, making use of three dimensions, but the resulting biplots based on the variables selected from these plots are still considerably worse than when all 13 variables are used and the scree-plot for the principal component biplot is no longer monotonically increasing. Cumulative scree-plots are monotonically increasing and so we believe that these are more helpful in the selection process.

For these ceramic pot data backward elimination methods were not able to select appropriate variables to distinguish between the pot groups which we know exist from Chapter Three. Comparing co-ordinates at each step of the process with the original set, rather than with those obtained at the previous step, still does not improve the separation of pot groups.

7.3.2 Application to Simpson Desert Flint Tools

In this section we use the flint tool data described in 1.2.6 in order to investigate the backward elimination method of variable selection, using both critical values from Krzanowski (1996) and scree-plots. We also introduce forward selection, all subsets and stepwise methods.

7.3.2.1 Backward Elimination using Critical Values (Two Dimensions)

Using the methodology of 7.2.2, we apply backward elimination procrustes analysis to the row co-ordinates obtained from the correlation biplot, coefficient of variation biplot, Spearman rank correlation biplot and principal component biplot. Because of differences in the units of measurement between the variables, the covariance biplot is not considered suitable. Assuming a two-dimensional representation and a 5% significance level, Table 7.3 indicates the order in which the variables are deleted for each biplot. For these data we compare the co-ordinates obtained at each step with those obtained from the full (original) data set. The measurements that correspond to the codes 1-6 are listed in Section 1.2.6.

Table 7.3 Krzanowski's Method of Variable Selection for the Simpson Desert Flint Tools (Two Dimensions)

Biplot	Order in which Variables Deleted	Variables Retained
Correlation	None Deleted	1 2 3 4 5 6
Coefficient of Variation	None Deleted	1 2 3 4 5 6
Spearman Rank Correlation	6 4 5	1 2 3
Principal Component	3 5 4	1 2 6

We see from the above table that all the variables are retained for the correlation and coefficient of variation biplots, but that three variables are deleted for the Spearman rank correlation and principal component biplots, although these are different in each case. By implementing the Spearman rank correlation and principal component biplots on the variables retained we can investigate whether any grouping of flint tools occurs and we can compare these groupings with those in the biplots obtained from using all the original six variables. The grouping obtained from the Spearman rank correlation biplot (not shown) is very similar to that obtained when the original six variables are used (Figure 3.9). Figure 7.10 illustrates the principal component biplot, where tools from site 08 are represented by circles (o) and tools from site 09 are represented by crosses (×); it appears that removing three variables (thickness, platform width, platform thickness) does not alter the tool groupings very much (compared with Figure 3.10). Thus, the backward elimination method has selected adequate variables for these two forms of biplot.

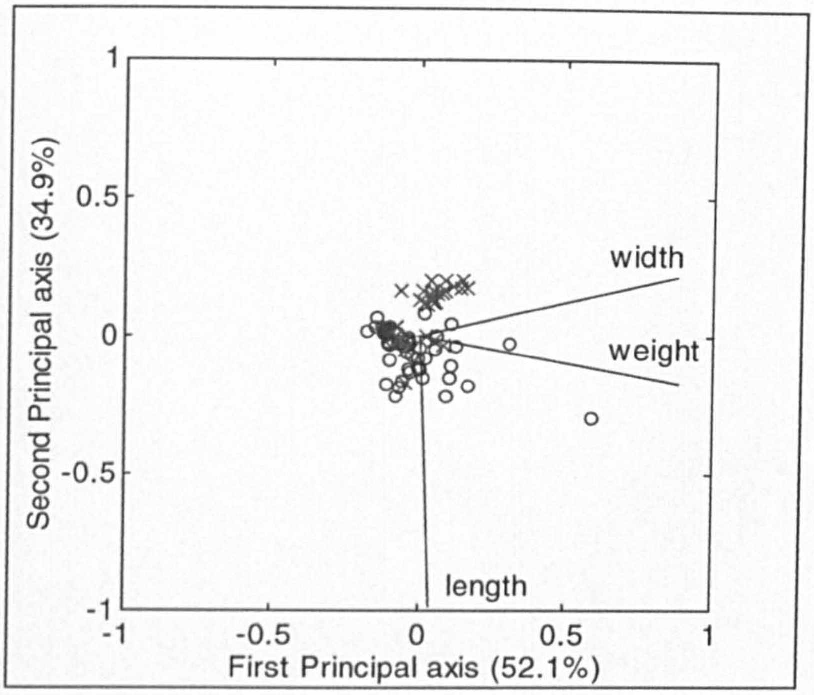

Figure 7.10 Principal Component Biplot of Simpson Desert Flint Tools (Backward Elimination)

7.3.2.2 Backward Elimination using the Scree-plot (Two Dimensions)

In Section 7.3.1.3 we introduced the scree-plot and cumulative scree-plot as possible alternatives to using Krzanowski's critical values in variable selection problems. This section applies the scree-plots to the various forms of biplot for the Simpson Desert flint tools. Figure 7.11 illustrates the scree-plot for the correlation biplot and this suggests that we should delete three variables (thickness, platform thickness and width), which contrasts with none deleted when using Krzanowski's critical values. A biplot on this reduced number of variables is illustrated in Figure 7.12.

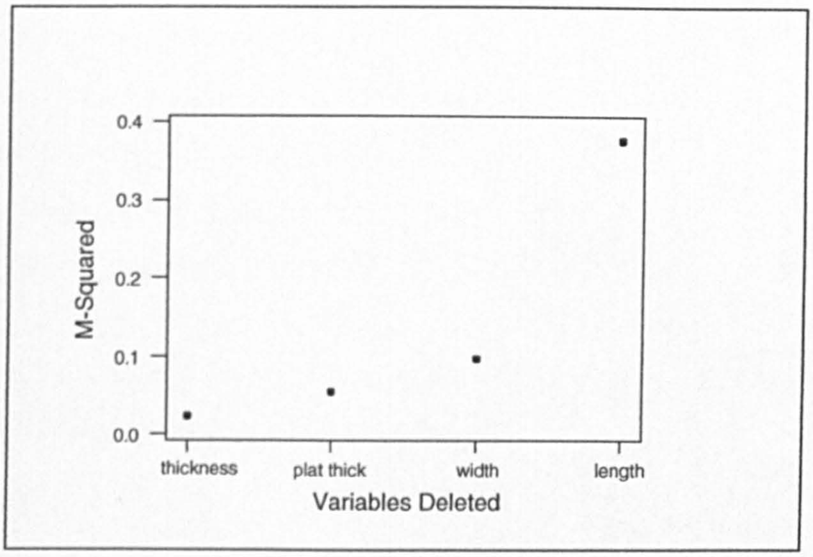

Figure 7.11 Scree-plot for the Simpson Desert Flint Tools (Correlation Biplot: Two Dimensions)

We see that the separation of tools obtained in Figure 7.12 is as good as that in Figure 3.7 where all six variables are used and thus the scree-plot has succeeded where Krzanowski's method has failed.

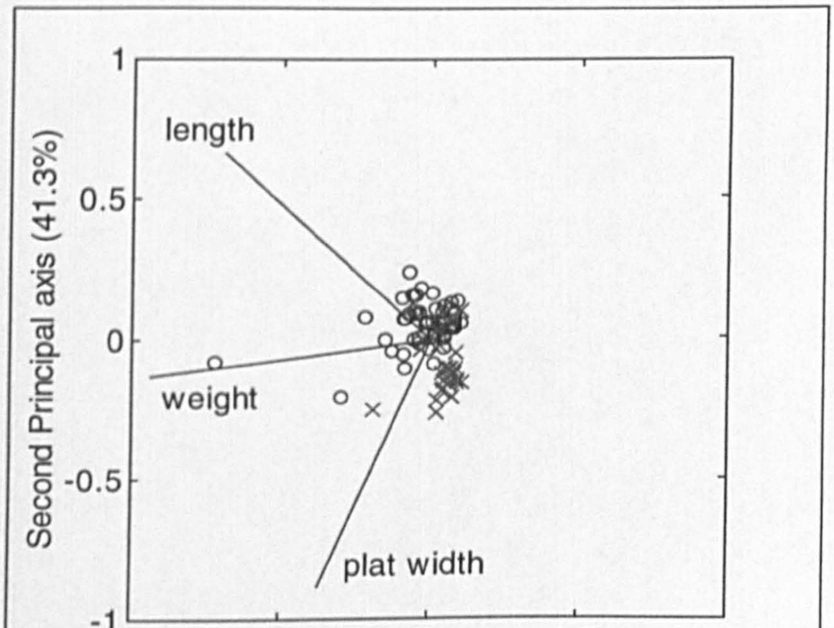

Figure 7.12 Correlation Biplot of Simpson Desert Flint Tools (Backward Elimination)

A scree-plot for the coefficient of variation biplot shows that we should delete two variables in contrast to none deleted using Krzanowski's critical values. The resulting biplot is as good as Figure 3.8 in terms of separation of tools and so again a scree-plot has proven useful. We also delete two variables for the Spearman rank correlation biplot based on the scree-plots, rather than three using critical values. The resulting biplot is as good at separating tool sites as Figure 3.9. Using the scree-plot for the principal component biplot we would probably delete three variables, which agrees with those deleted using the critical values of Krzanowski (1996).

In summary, all the scree-plots are monotonically increasing in two dimensions and the separation of tools into groups is as good when using the reduced numbers of variables as it is with all the original six variables and so we do not believe that it is necessary to consider higher dimensionality at this stage. The scree-plot was useful here because it enabled us to reduce the number of variables measured for the correlation and coefficient of variation biplots whilst still retaining tool groups, whereas Krzanowski's criteria did not select any variables. We now introduce the methodology of other methods of variable selection.

7.3.2.3 The 'All Subsets' Approach (Two Dimensions)

In this section we introduce and discuss the 'all subsets' approach to variable selection, by which we mean that the observation co-ordinates obtained from every combination of two or more variables are compared with the original co-ordinates and we choose that combination with the smallest M_j^2, where j variables are included. Because, in 7.3.2.1, the stopping criterion of Krzanowski (1996) suggested that three variables should be retained for the Spearman rank correlation and principal component biplots, we examine M_j^2 for each combination of three variables (20 combinations in total), to see whether this produces the same three variables as the backward elimination algorithm. We call this the 'all subsets' approach. However, because we only have six variables we could use the all subsets method on each possible combination of variables (56 combinations in total). Table 7.4 lists the variables which are retained under both the backward elimination and the corresponding all subsets approaches (where ---- indicates not appropriate), using both critical values and scree-plots.

Table 7.4 Variables Retained in the Backward Elimination and All Subsets Methods of Variable Selection for the Simpson Desert Flint Tools

Biplot	Backward Elimination (Krzanowski)	All Subsets	Backward Elimination (scree-plot)	All Subsets
Correlation	1 2 3 4 5 6	----	1 4 6	1 4 6
Coefficient of Variation	1 2 3 4 5 6	----	1 3 4 5	1 3 4 5
Spearman Rank Correlation	1 2 3	1 3 4	1 2 3 5	1 3 4 5
Principal Component	1 2 6	1 4 6	1 2 6	1 4 6

It is clear from the first two rows of the table that the variables retained under backward elimination using the scree-plot and under the corresponding all subsets method, agree for the correlation and coefficient of variation biplots. However, from rows three and four we see that different variables are retained under the all subsets and backward elimination approaches using critical values and that different variables are retained using a backward elimination scree-plot as compared with an all subsets approach, for both the Spearman rank correlation and the principal component biplots. We are now interested in obtaining biplots based on the variables retained in the all subsets approach and comparing these with the biplots resulting from the backward elimination method (using both Krzanowski's critical values and scree-plots) and, more importantly, with the original biplots of Chapter Three.

Figure 7.13 illustrates the Spearman rank correlation biplot for the all subsets approach and it is evident that it has very similar tool groupings to that of Figure 3.9 (six variables). Because the scree-plot retains more variables than Krzanowski's critical values, but the biplots on three variables are adequate, it is not worth considering four-variable biplots. However, a comparison of the M_j^2s obtained from applying all subsets with three variables, with the M_j^2s obtained from all subsets with four variables is useful, as we will see below.

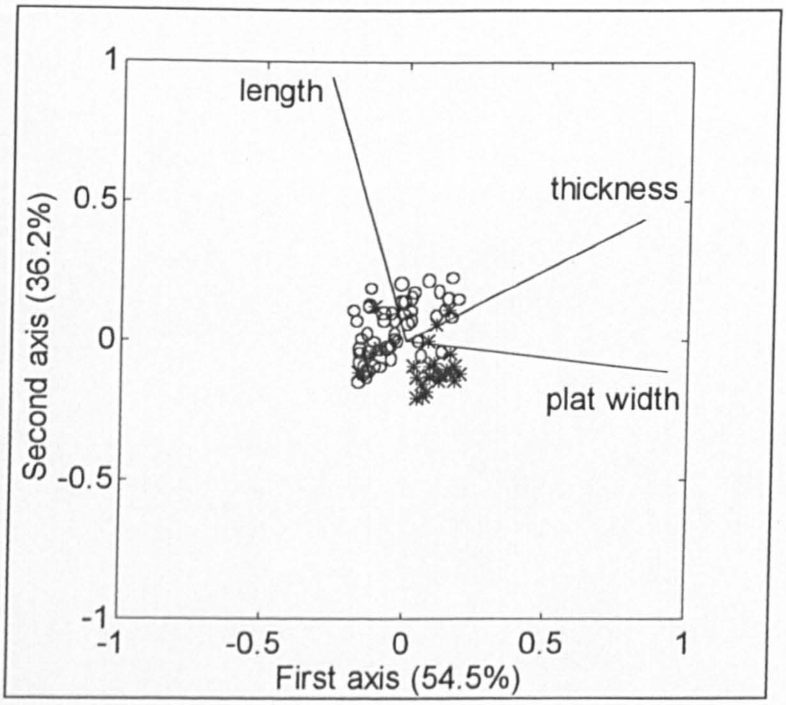

Figure 7.13 Spearman Rank Correlation Biplot of Simpson Desert Flint Tools (All Subsets)

For the principal component biplot (not shown), there is little difference between the tool groupings and Figure 3.10 (the original biplot). We believe that it is probably because the data contain so much noise that there is little visual difference between tool separation for any subset of three variables or more. The M_j^2 statistics are calculated for the different forms of biplot, for three and four variable subsets, so that we can identify how close the choice is between different subsets. For both the three variable subsets (correlation, Spearman rank correlation and principal component biplots) and, separately, the four variable subsets (coefficient of variation and Spearman rank correlation biplots), the choice is clear-cut. Because the separation of tool groups is as good with three variables as it is with four or six, we consider pairs of variables and the pair with the smallest M_i^2 is listed in Table 7.5.

Table 7.5 M_j^2 Statistics for the Simpson Desert Flint Tools: All Subsets Approach

Biplot	Variables Retained	M_j^2
Correlation	1 2	0.304
Coefficient of Variation	1 2	0.301
Spearman Rank	4 6	0.217
Principal Component	1 2	1.298

For the correlation and Spearman rank correlation biplots the M_j^2 values from the best subset of two variables are considerably larger than those for the corresponding best three variable and four variable subsets, indicating that pairs of variables do not need to be considered. However, for the coefficient of variation biplot the M_j^2 for the best two variable subset is smaller than that of the best four variable subset, albeit only slightly. The resulting biplot of these two variables provides a good distinction of tool groups, which is as good as that obtained from using all six variables.

7.3.2.4 Forward Selection (Two Dimensions)

Having implemented backward elimination and all subsets approaches to variable selection, we consider how these compare with forward selection. We propose two possible methods and these are described below.

Method One: Comparing Co-ordinates with those of the Previous Step

In this method the co-ordinates obtained at each step of the selection process are compared with those of the previous step. The first step is to consider all combinations of two variables. The combination with the smallest M_j^2 is chosen (where M_j^2 is the procrustes statistic for j variables included), where this gives the smallest difference between the subset and the full set of variables. Next, each remaining variable is added in turn to this pair of variables and the combination of three variables with the largest M_j^2 when compared with the pair of variables is chosen. Each remaining variable is added in turn and the process continues until all six variables are included. In order to implement forward selection we need to start with pairs of variables. This is in contrast to linear regression where we can obtain a measure of fit for each variable separately.

Method Two: Comparing Co-ordinates with the Original Co-ordinates

In this method the co-ordinates obtained at each step of the selection process are compared with the original co-ordinates. The first step is to consider all combinations of two variables. The combination with the smallest M_j^2 is chosen (where M_j^2 is the procrustes statistic for j variables included), where this gives the smallest difference between the subset and the full set of variables. Next, each variable is added to this pair in turn and the set with the smallest M_j^2 as compared with the original six variables is chosen. Variables are added one by one and the variable with the smallest M_j^2 is chosen each time.

7.3.2.5 Forward Selection using the Scree-plot (Two Dimensions)

In this section we use the scree-plot to assess which variables should be retained for the flint tool data. We can include $M_j^2 = 0$ in the plot, which occurs when all variables are selected, in order to allow for the possibility that five out of the six variables are needed. However, the point at which we stop including variables differs from that of the backward elimination procedure — we are looking for a substantial fall in M_j^2 in return for the inclusion of just one more variable. The scree-plot for the Spearman rank correlation biplot suggests that three variables should be selected and a biplot on these three variables (not illustrated) shows a similar division of tools into groups to that in Figure 3.9.

Figures 7.14 and 7.15 illustrate scree-plots for the principal component biplot, using methods one and two respectively. Based on these plots we would select three and four variables respectively. The principal component biplots of these selected variables are produced in Figures 7.16 and 7.17; both give as good a separation of tools into groups as was obtained using all six variables in Figure 3.10.

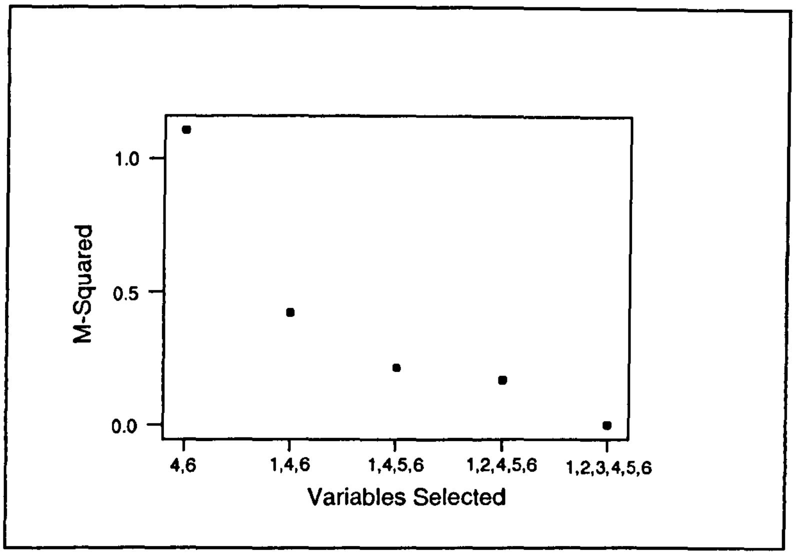

Figure 7.14 Scree-plot for the Simpson Desert Flint Tools (Principal Component Biplot: Method One)

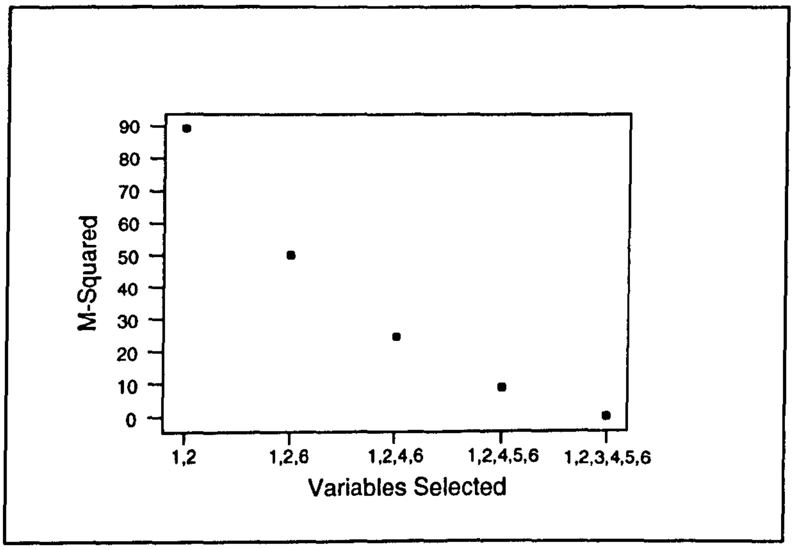

Figure 7.15 Scree-plot for the Simpson Desert Flint Tools (Principal Component Biplot: Method Two)

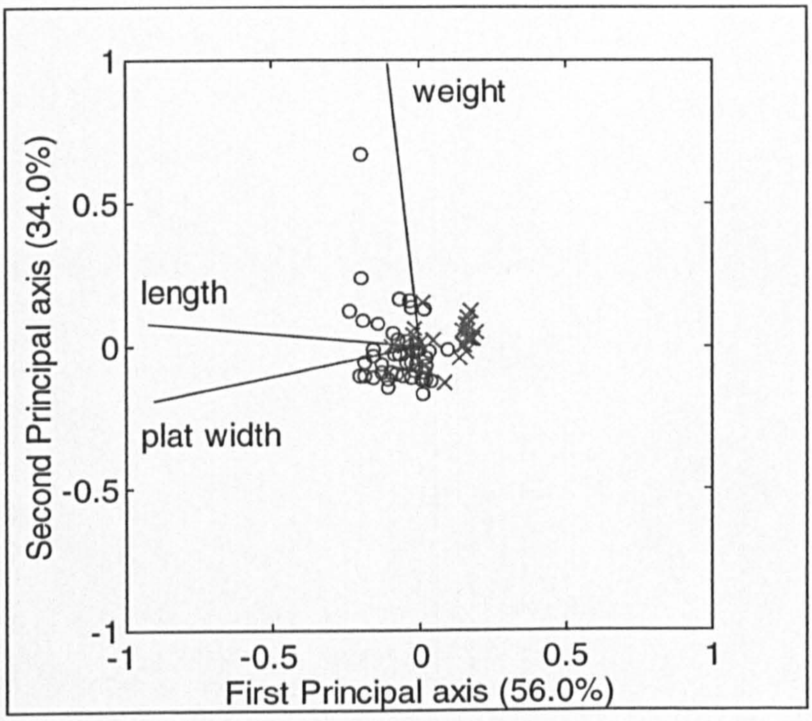

Figure 7.16 Principal Component Biplot of Simpson Desert Flint Tools (Method One)

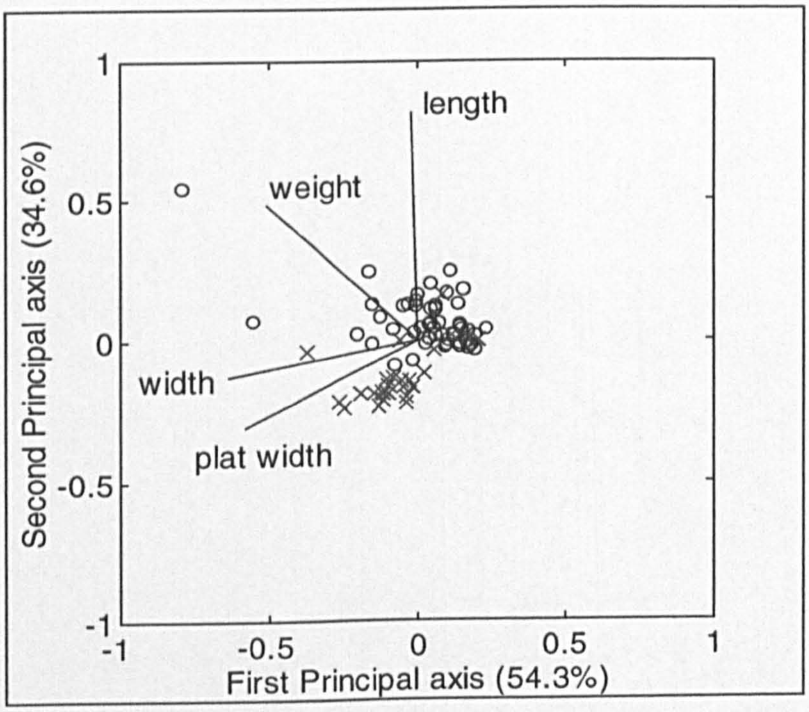

Figure 7.17 Principal Component Biplot of Simpson Desert Flint Tools (Method Two)

The variables retained based on the scree-plots for method one, for all types of biplot are listed in the following table. The spaces between the variables indicate which pair of variables is selected first (---- indicates a non-monotonic plot and so it is not possible to choose any variables).

Table 7.6 Forward Selection for the Simpson Desert Flint Tools (Method One)

Biplot	Order in which Variables Selected	Variables Retained
Correlation	1 2 5 4 3 6	1 2 5 4
Coefficient of Variation	1 2 5 6 4 3	1 2 5 6
Spearman Rank Correlation	1 6 5 2 4 3	----
Principal Component	4 6 1 5 2 3	1 4 6

7.3.2.6 Forward Selection using the Scree-plot (Three Dimensions)

If we begin the forward selection process by selecting two variables then it is not possible to use three dimensions in the M_j^2 calculations at this first step. We can, therefore, either use two dimensions for considering all pairs of variables at the first step and three dimensions for subsequent steps, or we can begin the forward selection process by choosing three variables (and not allow for the possibility of selecting only two). We implement the first option in this section, but for the correlation biplot the scree-plot is non-monotonic and this could be directly related to this decision. We can, however, look at the cumulative scree-plot.

A scree-plot for the coefficient of variation biplot indicates that we should select variables length, platform width and platform thickness. The resulting biplot is shown in Figure 7.18 and produces a similar tool group division to that in Figure 3.8.

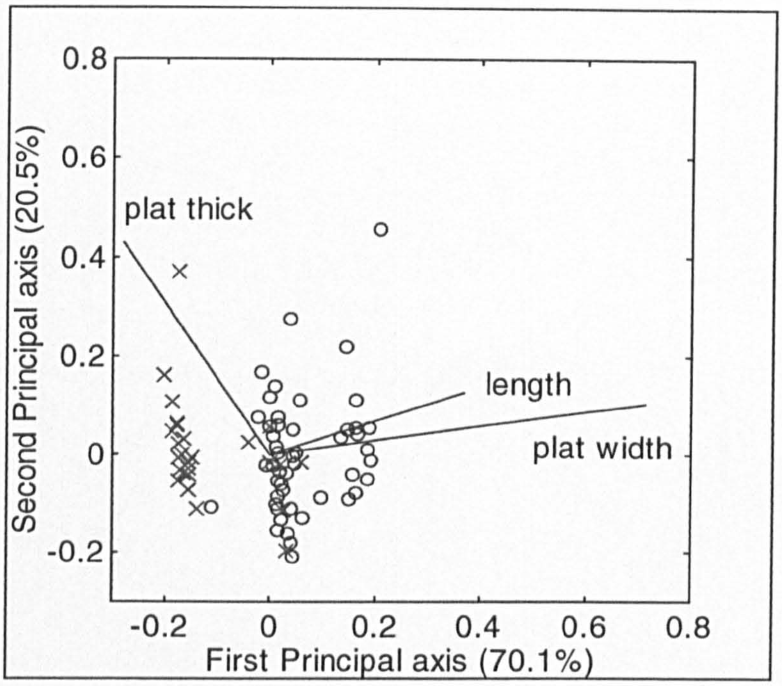

Figure 7.18 Coefficient of Variation Biplot of Simpson Desert Flint Tools (Forward Selection)

7.3.2.7 The Stepwise Approach (Two Dimensions)

In this section we suggest combining the forward selection and backward elimination methods to form a stepwise approach to variable selection which works as follows. The first step is to consider all combinations of two variables and calculate M_j^2 (where j is the number of variables included) by comparing the corresponding row co-ordinates with those of the original data, for each pair. The combination of variables with the smallest M_j^2 is chosen, providing this is greater than a threshold value. Next, each remaining variable is added in turn to this pair of variables and the combination of three variables with the largest M_j^2 is chosen, when compared with this pair, provided that this is greater than the threshold value. If no combination of three variables is greater than the threshold value then we stop the procedure with the pair of variables. The third step is to delete each of the variables in turn (except the one most recently added) and choose the smallest M_i^2 (where i variables are deleted). If M_i^2 is smaller than the threshold value then we remove the variable that was deleted, but otherwise we retain it. The fourth step is to add in each of the remaining variables (provided that they have not just been deleted in the third step) and compare M_j^2 with a threshold value. This process continues until no further variables have an M_j^2 large enough to be added, or an M_i^2 small enough to be removed.

The main problem with this stepwise method is how to determine the threshold values. In this section we use the same threshold value (F_{thr}) for determining whether to add or to delete a variable. We also compare the co-ordinates with those of the previous step, rather than with the original set, although this differs from the method used for the Simpson Desert flint tool applications earlier in the chapter.

We now implement the stepwise procedure described above and the variables retained under ranges of the threshold value are listed in Tables 7.7 and 7.8, for the various biplots.

Table 7.7 Variables Retained for the Simpson Desert Flint Tools (Stepwise Method: Correlation and Coefficient of Variation Biplots)

Correlation Biplot		Coefficient of Variation Biplot	
F_{thr}	Variables Retained	F_{thr}	Variables Retained
$F_{thr} > 2.083$	None	$F_{thr} > 1.844$	None
$1.612 < F_{thr} \leq 2.083$	3 5	$1.456 < F_{thr} \leq 1.844$	3 5
$0.249 < F_{thr} \leq 1.612$	1 5	$0.307 < F_{thr} \leq 1.456$	1 5
$0.239 < F_{thr} \leq 0.249$	1 2*	$0.269 < F_{thr} \leq 0.307$	4 6
$0.179 < F_{thr} \leq 0.239$	1 2 5*	$0.238 < F_{thr} \leq 0.269$	1 5 6
$0.153 < F_{thr} \leq 0.179$	1 3 5*	$0.093 < F_{thr} \leq 0.238$	1 4 6
$0.095 < F_{thr} \leq 0.153$	1 3 4*	$0.052 < F_{thr} \leq 0.093$	1 3 4 6
$0.037 < F_{thr} \leq 0.095$	1 3 4 5*	$0 < F_{thr} \leq 0.052$	1 2 3 4 6*
$0 < F_{thr} \leq 0.037$	1 2 3 4 5* or 1 3 4 5 6*		

Implementing a biplot on each set of retained variables indicates that sensible biplots (in terms of separation of flint tools) are obtained by choosing F_{thr} corresponding to the asterisk (*).

Table 7.8 Variables Retained for the Simpson Desert Flint Tools (Stepwise Method: Spearman Rank Correlation and Principal Component Biplots)

Spearman Rank Correlation Biplot		Principal Component Biplot	
F_{thr}	Variables	F_{thr}	Variables Retained
$F_{thr} > 1.878$	None	$F_{thr} > 1.106$	None
$0.231 < F_{thr} \leq 1.878$	5 6	$0.423 < F_{thr} \leq 1.106$	4 6
$0.210 < F_{thr} \leq 0.231$	1 2*	$0.413 < F_{thr} \leq 0.423$	1 4*
$0.155 < F_{thr} \leq 0.210$	4 5 6*	$0.368 < F_{thr} \leq 0.413$	1 5*
$0.114 < F_{thr} \leq 0.155$	1 3 5*	$0.362 < F_{thr} \leq 0.368$	1 4 5*
$0.053 < F_{thr} \leq 0.114$	1 4 5 6*	$0.217 < F_{thr} \leq 0.362$	1 4 6*
$0 < F_{thr} \leq 0.053$	1 2 4 5 6*	$0.203 < F_{thr} \leq 0.217$	1 3 4 5*
		$0 < F_{thr} \leq 0.203$	1 3 4 5 6*

The above tables suggest that for these data there are always threshold values which we can choose in order to retain anything from 2 to p-1 variables (where p is the number of original variables). If we are unsure where to set our threshold values within the stepwise method then we can select a specific number of variables to retain instead. We can also alter the threshold value for entering a variable so that it is different from that used to delete a variable and we can change the values at each step in the selection process so that they reflect the appropriate degrees of freedom.

7.3.2.8 Summary of Simpson Desert Flint Tools

All the variables selected under Krzanowski's backward elimination method using critical values and from backward elimination scree-plots provide as good a separation of tools into groups as the original six variables. All four scree-plots associated with backward elimination are monotonically increasing in two dimensions and they enable us to reduce the number of variables required to distinguish between tool groups in the correlation and coefficient of variation biplots, where Krzanowski's critical values do not. It is interesting to note that variable length was retained in all selections. Forward selection scree-plots selected sensible variables, in terms of tool separation, for all but the Spearman rank correlation biplot where the plot is non-monotonic. However, the cumulative scree-plot enables a selection to be made (and is, by definition, monotonic). The biplots produced from the all subsets approach in Table 7.4 provide a good discrimination between tool groupings. The stepwise method indicated that a very precise choice of threshold is required in order for the coefficient of variation biplot to select appropriate variables, but for the other biplots there are a range of values that produce a selection which separates the tool groups.

7.4 Summary and Conclusions

The various forms of biplot are useful tools for exploratory data analysis, particularly in field studies. Archaeologists in particular are interested in variable selection methods (Barton, *pers. comm.*) because these can save them valuable time and hence limited resources when collecting data. In this chapter we investigated the backward elimination method of variable selection introduced into principal component analysis by Krzanowski (1987, 1996) and extended this to the various forms of biplot with varying degrees of success. In particular, we proposed an adaption of the method by suggesting that it may be more sensible to compare the co-ordinates at each step of the selection process with those of the previous step, rather than with the original configuration using all the measured variables. This proposal was based on an analogy with linear regression methods.

We also looked at how the forward selection, all subsets and stepwise approaches compare with backward elimination and introduced the idea of a scree-plot and cumulative scree-plot to aid the selection process for the backward elimination and forward selection methods. The scree-plot reveals any non-monotonicity in successive procrustes values and both show a 'kink' in the graph when adding or deleting a variable leads to a particularly large or small difference in the procrustes residual sum of squares. In addition, we discussed the effect of the dimensionality used in the calculations and how this forces a minimum number of variables into the selection methods. There may, therefore, be a trade off between 'adequate' dimensionality in order for a 'good' subset of variables to be selected and the number of variables selected. It was also revealed that an increase in dimensionality when calculating the procrustes residual sum of squares does not always lead to monotonic scree-plots and can sometimes lead to previously monotonic plots becoming non-monotonic. This is overcome by using the cumulative scree-plot, which is, by definition, monotonic. We emphasised that there may be several subsets of variables that are able to distinguish between groups of artefacts, rather than one unique set: this is a major disadvantage of Krzanowski's method, which only obtains a single subset of variables. Because of this we believe that less formal methods of selection, for example scree-plots, have advantages over critical values, not least because they provide a graphical means of following the selection process at each step. With Krzanowski's variable selection method for principal component analysis it is important that k is chosen 'correctly', so that a good representation of the data in k dimensions is achieved, but with biplots we expect to see them in two dimensions only.

With the two data sets we analysed, it was not the dimensionality used in the calculations that was the important factor in the success of the methods, but the data itself. When the initial separation of observations into groups, using all the measured variables (as for the ceramic pots) was good, then regardless of whether two or three dimensions were used in the calculations and regardless of whether critical values or scree-plots were used, the variable selection was poor. However, when there was more 'noise' in the data and the group separation based on the original variables was not so clear cut (as for the flint tools), then two dimensions were adequate for variable selection to be successful, for all forms of biplot and all methods that were implemented, regardless of whether critical values or scree-plots were used.

Chapter Eight

Stability, Sample Size and Biplots

8.1 Introduction

This chapter uses the theory of biplots as described in Chapter Three, in combination with techniques such as bootstrapping and directional data methods, in order to assess how representative our data are of the true population of data within the biplot framework — i.e. to assess the stability of biplots. Chapter Three raised the issues of the effect of the number of artefacts measured on the interpretation of a biplot and the identification of outlying or influential observations. Investigating these, using various resampling methods and jack-knifing, forms part of this chapter. The Simpson Desert flint tools and the ceramic pots first described in Chapter One and discussed extensively in Chapters Three and Seven are used throughout to illustrate the methods that we develop.

Section 8.2 describes how the multivariate normal distribution can be used to replicate the data matrix in order to assess the stability of biplot variables, before explaining why, in contrast to correspondence analysis, there is only one method of obtaining observation and variable co-ordinates from these replicates. We also develop methods of projecting supplementary observations and variables onto the original biplot axes. Traditional bootstrap confidence intervals are extended to biplots to assess the true directions of the variables in Section 8.3 and in 8.4 intervals are obtained by applying directional data methods. In this section we also propose, for some types of biplots, an adaption of the usual method of calculating a mean direction. In Section 8.5 we introduce an alternative method of assessing the stability of biplot variables, which uses the jack-knife. Section 8.6 investigates the influence of sample size (e.g. the number of artefacts measured) on biplots and 8.7 discusses the overlap between variable selection methods and sample size issues. A method of identifying influential observations by using a jack-knife approach is introduced in Section 8.8. Finally, conclusions are drawn in 8.9.

8.2 Assessing Stability by using the Multivariate Normal Distribution

Because our data consist of a number of variables measured on a series of observations for only a sample of all possible data, we need to consider how representative these data are of the true population of data (i.e. how stable are the variables). Whereas it is possible to investigate the stability of categories within a correspondence analysis map by bootstrapping the original data matrix and treating each column, row, or the whole data matrix as a sample from the multinomial distribution (see Chapter Five), for biplots we are dealing with a different type of data. One possibility is to fit the multivariate normal distribution to the data matrix and then bootstrap from this distribution. Another possibility (only appropriate for examining stability when we are interested in smaller sample sizes than that actually obtained) is to sample the observations without replacement (see 8.6).

In the rest of Section 8.2 we describe fitting a multivariate normal distribution to the data (after any necessary transformations), because this seems to fit well; it is also one of the most mathematically tractable distributions. However, in principle the multivariate normal distribution could generate negative values (negative values are inappropriate because the data suitable for biplots consist of 'measurement' variables and as such should be greater than zero); if these occur then an alternative sampling method should be used. Various methods for assessing multivariate normality are given in Gnanadesikan (1977), although in practice marginal normality is usually considered sufficient because large numbers of observations are required to test for multivariate normality. The computational details for fitting a multivariate normal distribution are described below.

8.2.1 Computational Details

Consider a data matrix X (n × m) with n rows (observations) and m columns (variables). We assume that in the population X (or some transformation of X) has a multivariate normal distribution, i.e. $X \sim MN_m(\mu, \Sigma)$. In our sample, $\bar{x}$ approximates μ and is a vector of means of the variables; S approximates Σ and is the variance-covariance matrix. We use the following steps to obtain one bootstrap of X:

Step 1: Calculate the sample mean vector:

$$\overline{x}^T = (\overline{x}_1, \ldots, \overline{x}_m) = \frac{1}{n} 1_n{}^T X$$

where 1_n is the column vector of n 1s.

Calculate the variance-covariance matrix:

$$\text{var (X)} = S = \frac{1}{n-1}(X - \overline{X})^T (X - \overline{X})$$

where $\overline{X}(n \times m)$ is a matrix with all rows equal to $\overline{x}^T$.

Step 2: Assume that the data come from the multivariate normal distribution with mean μ and variance-covariance matrix Σ. To resample from this distribution we must find a matrix G such that $\Sigma = G^T G$ (Morgan, 1995, says that usually a Choleski factorisation is used for Σ, in which G^T is a lower triangular matrix).

Step 3: Generate p independent standard univariate normal random variables $z_1, \ldots, z_m$ and let $z = (z_1, \ldots, z_m)^T$.

Step 4: Let $y = \mu + G^T z$. The vector y is then an observation from a $MN_m(\mu, \Sigma)$ population.

Step 5: Repeat steps 3 and 4 n times.

To obtain B bootstrap matrices carry out steps 3-5 B times.

8.2.1.1 Transformations

It is well known that a minimum requirement for data to follow a multivariate normal distribution is that each separate variable should follow a univariate normal distribution and we therefore need to ensure that this is true for at least the majority of variables. If the variables are of the same type, it is better for the sake of conformity that either the data are left as they are or that all the variables are subjected to the same transformation (usually a logarithm or square root will suffice). We use the Kolmorgorov-Smirnov Test to test each variable for normality before deciding whether a transformation is appropriate or not, but any test of univariate normality can be used, it does not have to be the Kolmorgorov-Smirnov Test.

8.2.1.2 Grouped Data

If we know, a priori, that the data are grouped (as in the cases of the flint tools of 1.2.6, where there are two sites and the ceramic pots of 1.2.5, where there are three potters), then we should allow for this in the generation of replicate data. For each variable we should first subtract group means from each group, assess for normality and if the data are non-normal then we can try various transformations (e.g. Box-Cox, logarithm, square root). We then transform each variable before we subtract group means, assess for normality and if the majority of variables are normally distributed then the transformation is applied to the raw data. Regardless of whether or not a transformation has been used, we generate data from separate multivariate normal distributions for each group. However, the generated data are combined before implementing a biplot.

Even if we know that the data consist of groups of observations, if the original biplot is not able to show group differences then there is an argument for treating the data as a homogeneous set. For example, with the ceramic pot data, a priori we know that there are three potters, but we don't know whether the pots that they produce can be distinguished on the basis of the available measurements. In fact, this is one of the aims of the analysis.

8.2.2 Bootstrap Co-ordinates

Having obtained replicate matrices by fitting the multivariate normal distribution to the original data, we discover that there is only one way of obtaining observation and variable co-ordinates for biplots — this is to implement a biplot on each replicate matrix. This is in contrast to correspondence analysis (see 5.1.2.1 and 5.1.2.2) where we can either project replicate matrices onto the original co-ordinate system or we can carry out a new correspondence analysis on each generated matrix. In the following sections we describe why it is inappropriate to project replicate matrices onto the original co-ordinate system for biplots. We also develop the methodology that enables supplementary observations and variables to be displayed.

8.2.2.1 Separate Biplots for each Replicate Matrix

One method of obtaining biplot co-ordinates from the generated matrices is to carry out a biplot analysis on each replicate matrix and overlay the co-ordinates on the same plot. This method could, however, be criticised on the grounds that the co-ordinates are all relative to different axes and are thus not directly comparable (see the discussion in 5.1.2.1 for correspondence analysis). In the next section we describe why relating co-ordinates to the original axes is inappropriate.

8.2.2.2 Relating Biplot Co-ordinates to the Original Axes

As in correspondence analysis, it can be argued that biplots should not be applied to the replicate matrices directly, because this leads to the co-ordinates being relative to different axes; instead, the replicates should be related to the original co-ordinate system. However, this is inappropriate. Suppose we have a data matrix X (n × m), with singular value decomposition given by:

$$X = UD_{\mu}V^{T}.$$

Recall, from Chapter Three, that for the Principal Component Biplot Family, the original observations have co-ordinates given by:

$$F = UD_{\mu} = XV.$$

Therefore, for each replicated matrix X^r, with singular value decomposition:

$$X^r = U^rD_{\mu}{}^rV^{rT},$$

we can define the observation co-ordinates as given by:

$$F^r = U^rD_{\mu}{}^r = X^rV^r$$

and for each replicate matrix the co-ordinates will be relative to different axes.

In order to obtain the observation co-ordinates we project the matrices (standardised as in Chapter Three) onto the original space spanned by V. The observation co-ordinates are then given by:

$$F^r = X^r V. \tag{8.1}$$

In effect, we are using the original variable co-ordinates (V) as the reference co-ordinates. However, because we have fitted a multivariate normal distribution to our data, which is based purely on the means and variance-covariance matrix of the original data, there is no reason why the first row (observation) of a replicate matrix should correspond to the first observation of the original data i.e. the relative row positions are lost when the replicate matrices are obtained. We do not, therefore, obtain sensible co-ordinates for the observations and variables from the replicate matrices.

8.2.2.3 Supplementary Data for the Principal Component Biplot Family

In this section we propose a method for projecting supplementary observations and variables onto the original co-ordinate system. This is useful if the data contain any unusual observations that we do not want to influence the ordination diagram; these can be projected onto the biplot after the axes have been determined. It is also useful when 'extra' observations are measured, or when another variable is measured on existing observations. We recall that V is the matrix of the original variable co-ordinates (see 3.5) and we take x^s to be a new observation with dimensions 1 × m. Because we rescaled the original data before carrying out a biplot analysis, we should standardise this new observation by subtracting the original means for each variable and dividing by the standard deviations. We should also divide by $\sqrt{n-1}$, where n is the number of observations in the original data matrix. We can now project a new observation onto the display by using (8.1); the co-ordinates of this observation, f^s, are given by:

$$f^s = x^s V.$$

If we obtain a supplementary variable i.e. an extra variable is measured on each observation, then we can also project this onto the original plot. If we take x_v to be a new variable with dimensions n × 1 and standardise it by subtracting its mean, dividing by its standard deviation and dividing by $\sqrt{n-1}$, then the co-ordinates of this variable are given by:

$$g_v = x_v^T U D_\mu^{-1}.$$

8.2.2.4 Supplementary Data for the Correlation Biplot Family

In the previous section we explained how to project supplementary data onto the biplot axes for biplots in the Principal Component Biplot Family. In this section we consider the Correlation Biplot Family. Recall from Chapter Three that the observation co-ordinates are given by:

$$F = U = XVD_\mu^{-1}. \qquad (8.2)$$

For biplots of the Correlation Biplot Family, VD_μ^{-1} is retained from the original data and we take x^s to be a new observation with dimensions 1 × m. Because we rescaled the original data before obtaining a biplot, we should again standardise this new observation. We can then project a new observation onto the display using (8.2), so that its co-ordinates are given by:

$$f^s = x^s V D_\mu^{-1}.$$

If we obtain a supplementary variable i.e. an extra variable is measured on each observation, then again we can project this variable onto the original plot. If we take x_v to be a new (standardised) variable with dimensions n × 1, then the co-ordinates of this variable are given by:

$$g_v = x_v^T U.$$

8.2.2.5 Bootstrap Fans

The aim of Section 8.2.1 was to illustrate one method of generating replicate matrices in order to assess the stability of the biplot variables (i.e. to investigate how representative our data sample is of the true population of data). If any of the variables are unstable then it may be unwise to include them in the variable selection methods discussed in Chapter Seven. We assess stability by obtaining variable co-ordinates from each replicate matrix and thus we have as many vectors for a particular variable as we have replicate matrices. We develop the notion of a 'bootstrap fan' to describe a set of these bootstrap vectors for a particular variable (so-called because they typically resemble a fan) and we propose using these fans to obtain confidence intervals for the true directions of the variables i.e. for the whole population of data.

Having obtained bootstrap fans for each variable, we intuitively hope that the original variable co-ordinates lie roughly in the centre of the fans. However, it turns out that this is not always the case and the reasons why are complex. The root cause appears to be that a biplot in two dimensions is not always appropriate for a particular data set; this is because two dimensions can be insufficient to both explain a high proportion of the variation in the data and also for each individual variable to have a high quality of representation. The problem is confounded by low correlations between variables. We discuss these issues in the context of analysing the flint tool data in 8.2.4.1.

8.2.2.6 Application to Ceramic Pots

In this section we illustrate, for the ceramic pots described in 1.2.5, the bootstrap fans obtained when a new biplot is implemented on each replicate matrix. However, we first need to account for any arbitrary sign changes resulting from the singular value decomposition, as discussed in 5.2.4 and again in 8.2.4. We generate 100 replicate matrices and illustrate the fan for variable 13 (thickness of lip) of the correlation biplot in Figure 8.1. In this figure and throughout this chapter, an asterisk indicates the direction of the variable obtained from the original data.

Figure 8.1 suggests that separate biplot analyses on each matrix produce sensible bootstrap fans, because the width of the fan illustrated only covers approximately 90° (as do those of the other variables which are not shown) and the replicate vectors are of similar lengths to the vector representing the original variable.

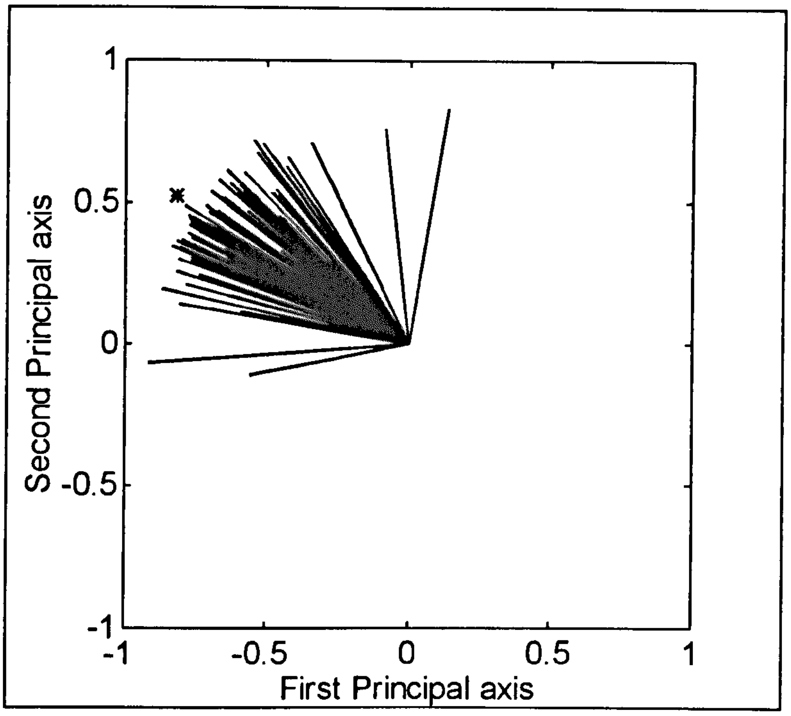

Figure 8.1 A Bootstrap Fan from Separate Biplot Analyses (Correlation Biplot)

8.2.3 Variable Selection and Bootstrapping

In this section we propose that if one of the variable selection methods described in Chapter Seven has been implemented, then it may be possible to make use of these selected variables when using bootstrapping to assess stability. This is particularly important if there are several subsets of variables which are able to distinguish between groups of observations, because that subset that consists of the most stable variables can be considered to be the most useful. When fitting a multivariate normal distribution to the data in preparation for bootstrapping, we believe that there are two choices.

Method One: Fit the distribution to the original variables and then delete those suggested by the variable selection method before implementing a biplot.

Method Two: Delete the variables suggested by the variable selection procedure and then fit the distribution to the resulting data before carrying out a biplot analysis.

Method two could be argued to be most sensible on the grounds that having selected a subset of variables, it is then that we need to confirm their stability. However, there may be an argument for using the bootstrap fans in the variable selection process, because unstable variables will be highlighted (i.e. those with wide confidence intervals for the true direction — see 8.3 and 8.4). We should bear in mind, however, that the stability of a particular variable varies according to which other variables are included in the analysis.

8.2.4 Reflection, Reordering and Procrustes Rotation

Because of the arbitrary nature of the singular value decomposition (which is unique only up to the sign of the eigenvectors, because the left and right eigenvectors are determined independently) and the problems discussed in 5.2.4, filtering must be applied to the co-ordinates obtained from each bootstrap. It is always necessary to apply reflection to biplot co-ordinates (and indeed any co-ordinates obtained from a singular value decomposition), but Milan & Whittaker (1995) suggest that reordering is only necessary if any of the singular values have changed order. It is unclear from Milan & Whittaker (1995) which form of filtering is appropriate for exploratory multivariate methods and the present author is unable to find any other relevant material. We apply filtering to the flint tools in the next section.

8.2.4.1 Application to Simpson Desert Flint Tools

We recall from Chapters One, Three and Seven that the Simpson Desert flint tool data (1.2.6) consist of measurements of six variables on tools from two sites. By assessing each variable for normality using the Kolmorgorov-Smirnov Test, we find that the distributions of all six variables exhibit departures from normality. Considering transformations of all the variables we find that the log transformation makes variables length, width, thickness and weight normally distributed and all variables are therefore subjected to this transformation. We replicate the data by fitting two multivariate normal distributions, one to each site, 100, 1000 and 5000 times, carrying out a biplot on the replicates for both sites combined and then applying reflection as discussed in 5.2.4.1. Figure 8.2 shows the results of reflection applied to 100 bootstraps, but for clarity only variables length and width are shown, although the analysis was carried out on all six variables. Again, the asterisks indicate the positions of the original variables.

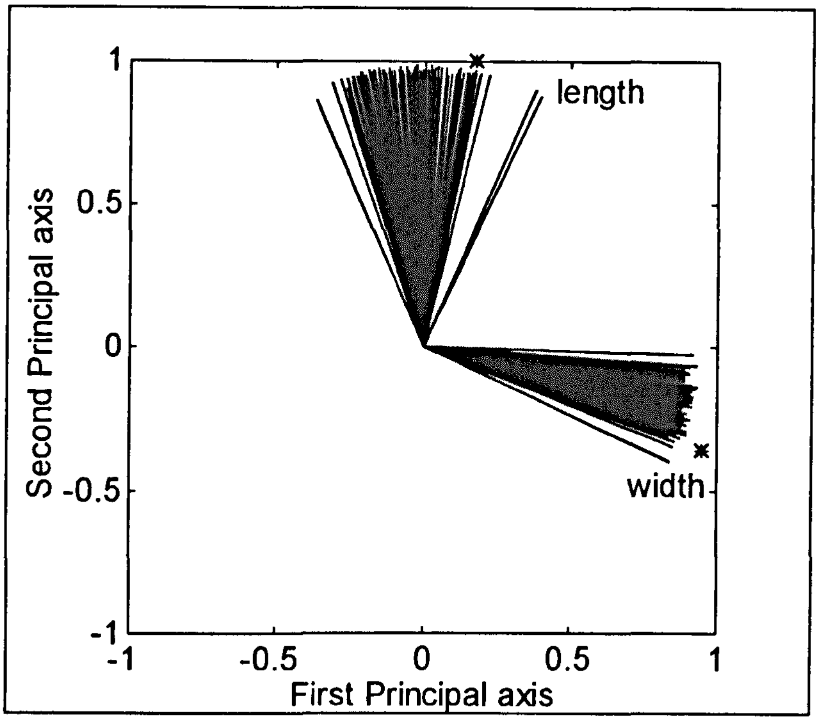

Figure 8.2 Bootstrap Fans for the Simpson Desert Flint Tool Variables (Correlation Biplot: All Data)

It is clear from Figure 8.2 that even after reflection the fans are not centred about the original co-ordinates; applying reordering produces an identical figure, but it is not clear from Milan & Whittaker (1995) whether this is necessary. The third and more stringent form of filtering is procrustes rotation, but we believe that this is unnecessary for two reasons. Firstly, because there can never be any translation of the biplot co-ordinates as a result of the singular value decomposition, a procrustes rotation would 'overcorrect' for nuisance variation that does not really exist. Secondly, the stretching incorporated into the procrustes rotations (see 5.2.4.3) would be problematic because the replicate vectors are of different lengths. We therefore apply only reflection throughout the remainder of this chapter.

Because the fans of Figure 8.2 are far from being centred we believe that there is a problem either with fitting the multivariate normal distribution, with the data set itself, or with the two-dimensional representation of the data. In order to examine the first possibility we generate univariate normally distributed variables and, separately, multivariate normal data. In the former case all variables have a low quality of representation and are virtually uncorrelated (as expected); in the latter case all are well represented and highly correlated (again, as expected). Additionally, the distributions of the variables do not exhibit departures from normality using the Kolmorgorov-Smirnov Test. This appears to confirm that the actual method of generating multivariate normal data is correct and suggests that it is either the data themselves, or the dimensions in which the fans are displayed, that determines whether the fans are centred.

To investigate the second and third of the possible causes of non-centring we consider the two sites separately. For site 9 alone all the variables except platform thickness are centred. We also calculate the quality of representation of each variable in two dimensions (Table 8.1) and the correlations between variables. This is done for the two sites separately and together and reveals that for site 9 alone, platform thickness has a quality of representation of only 11.5%, which is by far the lowest of all the variables. The corresponding bootstrap fan encompasses 360° and we believe that this is at least partly because the variable is uncorrelated with all the others, so that the two-dimensional display does not adequately capture this variable. Considering the quality of representation of the variables in the third dimension for site 9, we see that platform thickness has a value of 88.4% and is therefore well represented in this third dimension. Still considering site 9 alone and omitting platform thickness produces centred fans for the other variables.

Table 8.1 Quality of Representation of Simpson Desert Flint Tool Variables (%) for the Correlation Biplot

	Site(s)		
Variable	**8**	**9**	**8 & 9**
Length	94.6	93.6	87.9
Width	84.9	88.7	85.0
Thickness	81.5	65.7	79.3
Platform width	83.2	78.1	83.4
Platform thickness	78.7	11.5	69.4
Weight	83.8	92.3	89.0

Considering site 8 alone, we find that none of the bootstrap fans are centred. Using our knowledge of site 9 we omit platform thickness and obtain centred fans for the other variables. However, without this knowledge it is difficult to see how we would come to this decision: whilst platform thickness is the least well represented variable for site 8, it is still well represented and it has high positive correlations with the other variables. It turns out that the correlation structure of the variables is very different within each site.

Having examined the two sites separately, we again consider them together because we are interested in the data as a whole. However, omitting platform thickness no longer leads to centred fans for the other variables. Looking at the quality of representation of each variable in Table 8.1, we see that variable thickness also has a low value and omitting these two variables does lead to more centred fans (the fans for length and width are illustrated in Figure 8.3). Comparing Figure 8.3 with Figure 8.2 we see that the locations of the original variables have altered and this is because they are obtained from data consisting of different numbers of variables — their relative positions are unchanged.

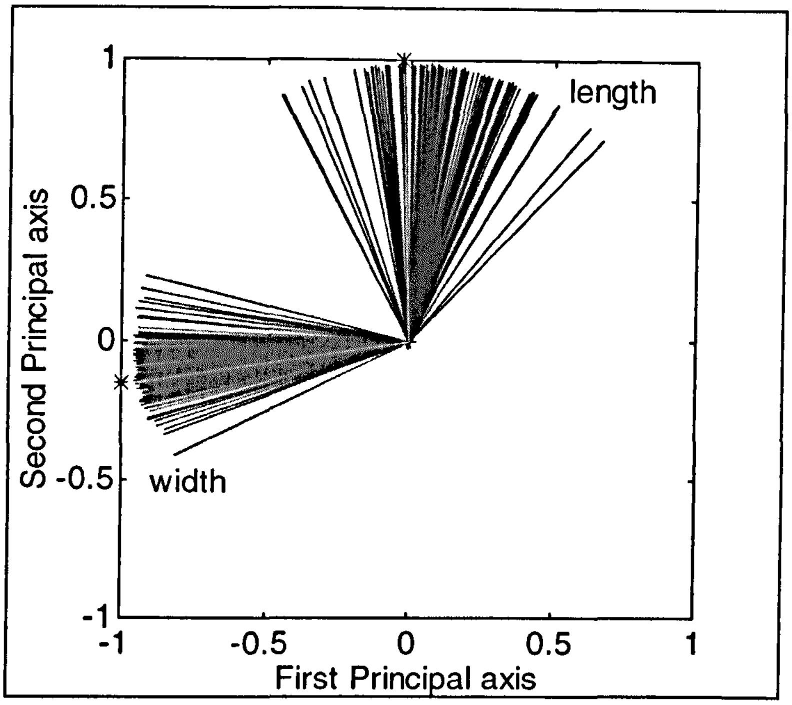

Figure 8.3 Bootstrap Fans for the Simpson Desert Flint Tool Variables (Correlation Biplot: Selected Variables)

The discussion of this section has focused on the correlation biplot, but considering the other forms of biplot (coefficient of variation, Spearman rank and principal component) we find that for all variables the fans are the same width, regardless of whether reflection or reordering is applied. We will therefore consider reflection to be the best means of correcting for the arbitrary sign changes of the singular value decomposition throughout this chapter.

8.2.4.2 Application to Ceramic Pots

The biplots for the ceramic pots (1.2.5) are not illustrated here, but the bootstrap fans are centred for all 13 variables for all biplots using both reflection and reordering; the correlation structure within each of the three pot groups is also very similar. Comparing the fans under reflection and reordering for the ceramic pots, the fan widths are found to be the same for the covariance, correlation, coefficient of variation and Spearman rank correlation biplots, but smaller under reordering for the principal component biplot. However, because there is no evidence of singular values changing order (Milan & Whittaker, 1995), we again only apply reflection to the co-ordinates.

8.3 Confidence Intervals for the True Directions of Biplot Variables using Standard Bootstrap Methods

In the previous sections we explained the methodology for bootstrapping a data matrix B times using the multivariate normal distribution and we also discussed the need to apply reflection to the replicate co-ordinates in order to correct for the arbitrary sign changes of the singular value decomposition. We would now like to estimate a confidence interval for the true direction of each variable (i.e. the direction that each variable would take if the whole population of data rather than a sample had been measured), based on these replicate vectors. However, the widths of the bootstrap fans (see 8.2.2.5) are related to the number of bootstraps that are generated (greater numbers of bootstraps result in wider fans). This is not really surprising because, as we generate more bootstraps, more and more unusual data sets are obtained, although, as we see below, sensible confidence intervals for the true direction of a particular variable do not depend on the number of bootstraps.

There are several well known confidence intervals for use with bootstrapped data, two of which we introduce below. We discuss their appropriateness for the flint tool and ceramic pot data.

8.3.1 The Standard Confidence Interval

In this section we describe the standard confidence interval and propose applying it to the biplot variables in order to obtain confidence intervals for their true directions. The literature contains little information on bootstrapping multivariate data and Efron & Tibshirani (1993) provide the most useful reference. Efron & Tibshirani (1993) discuss the singular value decomposition of a covariance matrix and use bootstrapping to measure the accuracy of $\hat{\theta}$, where $\hat{\theta}$ is the percentage of variation explained by the first p components in the particular data set they used. By sampling with replacement they obtain B bootstrap data sets and hence B values of $\hat{\theta}^*$, the bootstrap replication of $\hat{\theta}$. It is known that if the mean of the B replications, $\hat{\theta}^*(\cdot)=\sum_{b=1}^{B}\frac{\hat{\theta}^*(b)}{B}$, is close to $\hat{\theta}$, then $\hat{\theta}$ is close to unbiased. Efron & Tibshirani (1993) suggest using the standard confidence interval to assess the accuracy of the true value of $\hat{\theta}$, which is given by:

$$\theta \in \hat{\theta} \pm z^{(1-\alpha)} \cdot \hat{se}_B$$

with probability 1-2α, where $z^{(1-\alpha)}$ is the 100(1-α)-th percentile of a standard normal distribution. The standard error, se_B, is obtained from the B bootstraps by using:

$$\hat{se}_B = \left\{ \sum_{b=1}^{B} \frac{\left[\hat{\theta}*(b) - \hat{\theta}*(.)\right]^2}{B-1} \right\}^{\frac{1}{2}} .$$

We propose converting the co-ordinates of the vectors representing the biplot variables into angles that they make with the direction due east and using these in the calculations, because it is the directions of the variables that we are interested in. We therefore take $\hat{\theta}$ to be the angle of a particular variable in the original data and $\hat{\theta}^*(b)$ to be the angle of the vector representing this variable, obtained from the b-th bootstrap, where b=1,..., B. We must remember, however, that when implementing a biplot, arbitrary sign changes in the singular value decomposition can inflate the standard errors and so, as we saw in 8.2.4, reflection must be applied before confidence intervals are obtained.

8.3.1.1 Application to Simpson Desert Flint Tools

In this section we apply the standard confidence interval just described to the angles of the Simpson Desert flint tool (1.2.6) variables, in order to illustrate the importance of choosing an appropriate interval when assessing the variation in our estimate of the true direction of a variable. Table 8.2 lists the intervals obtained from 100 bootstraps for the correlation biplot. Because we are dealing with directional data (and considering the angles that the variables make with the direction due east), we should be aware that 360°≅ 0°. The confidence intervals are taken anti-clockwise.

Table 8.2 95% Standard Confidence Intervals (°) for the True Variable Directions for the Simpson Desert Flint Tools

Variable	Original Direction	Interval
Length	79	(62, 97)
Width	339	(330, 348)
Thickness	6	(357, 16)
Platform Width	325	(315, 336)
Platform Thickness	6	(348, 24)
Weight	26	(15, 38)

Because the intervals are symmetric they do not accurately represent the vectors in the bootstrap fans (we saw in Figure 8.2 that the fans are not centred about the original direction). We therefore look to an interval that does reflect this and we see in 8.3.2 that the BC$_a$ interval is more appropriate for these data.

8.3.1.2 Application to Ceramic Pots

In contrast to the flint tools, the bootstrap vectors of the ceramic pots (1.2.5) are centred about the original directions of each variable for all forms of biplot (see comments in 8.2.4.2) and so the standard confidence interval can be considered to be appropriate. Table 8.3 lists the intervals obtained from 100 bootstraps for the correlation biplot. The intervals are clearly much wider than those for the flint tools and all except that for variable 11 are of a similar range. This is interesting because variable 11 is the least well represented variable (see Table 3.1) and so there appears to be a connection between the quality of representation of a variable and its stability (when compared in the same dimensionality).

Table 8.3 95% Standard Confidence Intervals (°) for the True Variable Directions for the Ceramic Pots

Variable	Original Direction	Interval
1	305	(270, 341)
2	356	(318, 34)
3	203	(171, 234)
4	176	(141, 211)
5	350	(322, 18)
6	277	(241, 313)
7	306	(261, 351)
8	165	(135, 196)
9	60	(21, 98)
10	40	(4, 75)
11	86	(10, 162)
12	52	(19, 86)
13	148	(111, 185)

8.3.2 The BC$_a$ Method

In the previous section we considered the standard confidence interval for assessing the stability of biplot variables. In this section we propose using another method for calculating a confidence interval for the true direction of a variable — the BC$_a$ method. This method is also described in Efron & Tibshirani (1993) and works as follows. Let $\hat{\theta}^{*(\alpha)}$ indicate the 100α-th percentile of B bootstrap replications of the angle that a vector makes with the direction due east, $\hat{\theta}*(1),\ldots,\hat{\theta}*(B)$, where the BC$_a$ interval endpoints are given by percentiles of the bootstrap distribution. The percentiles used depend on two numbers $\hat{a}$ and $\hat{z}_o$, called the acceleration and bias-correction. The BC$_a$ interval of intended coverage, 1-2α, is given by:

$$(\hat{\theta}_{lo},\hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha_1)},\hat{\theta}^{*(\alpha_2)})$$

where $$\alpha_1 = \Phi\left(\hat{z}_o + \frac{\hat{z}_o + z^{(\alpha)}}{1-\hat{a}(\hat{z}_o + z^{(\alpha)})}\right)$$

and $$\alpha_2 = \Phi\left(\hat{z}_o + \frac{\hat{z}_o + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_o + z^{(1-\alpha)})}\right).$$

Here, $\Phi(\cdot)$ is the standard normal cumulative distribution function and $z^{(\alpha)}$ is the 100α-th percentile point of a standard normal distribution. The value of the bias-correction, $\hat{z}_o$, is obtained directly from the proportion of bootstrap replications less than the original estimate $\hat{\theta}$:

$$\hat{z}_o = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B}\right),$$

where $\Phi^{-1}(\cdot)$ indicates the inverse function of a standard normal cumulative distribution function. The acceleration, $\hat{a}$, can be computed in terms of the jack-knife values of a statistic $\hat{\theta} = s(x)$. Let $x_{(i)}$ be the original sample with the i-th point, x_i, deleted and let $\hat{\theta}_{(i)} = s(x_{(i)})$. Also, define:

$$\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \frac{\hat{\theta}_{(i)}}{n}$$

and $$\hat{a} = \frac{\sum_{i=1}^{n} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\left\{\sum_{i=1}^{n} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\right\}^{\frac{3}{2}}}.$$

We can now calculate α_1 and α_2. The BC_a method is known to have important theoretical advantages:

- It is transformation respecting. This means that the endpoints of the interval transform correctly if we change the parameter of interest from θ to some function of θ.

- A central $1-2\alpha$ confidence interval, $(\hat{\theta}_{lo}, \hat{\theta}_{up})$, is supposed to have probability α of not covering the true value of θ from above or below.

- It does not assume symmetry in the data.

However, the method also has one main disadvantage:

- A large number of bootstrap replications are required. At least B = 1000 replications are reported to be needed in order to sufficiently reduce the Monte Carlo sampling error.

We believe that a further problem with using this method (and also the standard confidence interval) in connection with biplots is that the differing vector lengths in the fan are ignored and only the angles that the vectors make with the direction due east are used. This is particularly relevant for the correlation and Spearman rank correlation biplots, where vector lengths represent the quality of representation of the variables (see 8.4.5).

8.3.2.1 Application to Simpson Desert Flint Tools

In this section we calculate the BC_a interval for the true directions of the flint tool variables (1.2.6), using the angles obtained from generating 100 and 1000 bootstrap vectors for the various forms of biplot. We display the results for the correlation biplot in Table 8.4. We see from the table that none of these intervals include the original directions of the variables, but they do accurately reflect the bootstrap fans (see figure 8.2) and this is because the fans are not centred (see earlier discussion in 8.2.4.1). The intervals are also much narrower than those for the standard confidence interval (see Table 8.2). Intervals of similar magnitudes are obtained for the other forms of biplot.

Table 8.4 95% BC$_a$ Confidence Intervals (°) for the True Variable Directions for the Simpson Desert Flint Tools

		Number of Bootstraps	
Variable	**Original Direction**	**100**	**1000**
Length	79	(98, 106)	(98, 108)
Width	339	(351, 358)	(351, 358)
Thickness	6	(16, 23)	(16, 24)
Platform Width	325	(338, 347)	(338, 345)
Platform Thickness	6	(353, 355)	(353, 355)
Weight	26	(38, 44)	(38, 46)

Following on from the discussion of 8.2.4.1, if we remove variables thickness and platform thickness and obtain intervals for the remaining four variables (see Figure 8.3), for which the corresponding bootstrap fans are slightly more centred, we obtain Table 8.5.

Table 8.5 95% BC$_a$ Confidence Intervals (°) for the True Variable Directions for the Simpson Desert Flint Tools (Selected Variables)

		Number of Bootstraps	
Variable	**Original Direction**	**100**	**1000**
Length	92	(56, 98)	(59, 97)
Width	189	(195, 203)	(194, 217)
Platform Width	203	(188, 207)	(188, 206)
Weight	143	(104, 152)	(97, 151)

Table 8.5 shows that the intervals are now slightly more centred about the original directions (which have altered because there are now only four variables to consider), but they are still a long way from being symmetric and the original direction is not included in the interval for the variable width. The intervals are also much wider than those for the corresponding variables in Table 8.4 (except for variable width). Despite this, the BC$_a$ method seems to be a reasonable method for calculating a confidence interval for the true direction of a variable.

8.4 Confidence Intervals for the True Directions of Biplot Variables using Directional Data Methods

In Section 8.3 we discussed two well known confidence intervals for use with bootstrap data and proposed using them to obtain confidence intervals for the true directions of biplot variables. However, because we are interested in variable directions it seems most appropriate to use methods specifically developed for directional data. We discuss and apply one such method in the following sections.

8.4.1 Lengths of Vectors in the Confidence Intervals

Before discussing and applying a confidence interval from Fisher (1993), we diverge slightly to make some general comments regarding confidence intervals. When calculating our confidence intervals we have only been using the angles that the vectors make with the horizontal and so we do not have a vector length to use when plotting the intervals. We therefore need to convert our lower and upper confidence limits into lower and upper values of x and y co-ordinates and we propose the following. If, for a particular variable, the lower and upper confidence intervals are given by a and c, then we have:

$$x_l = r\cos a$$

$$y_l = r\sin a$$

$$x_u = r\cos c$$

$$y_u = r\sin c$$

where x_l and x_u denote the lower and upper values of the x co-ordinate respectively, y_l and y_u denote the lower and upper values of the y co-ordinate and r denotes the length of the vector. Because both the x and y co-ordinates are multiplied by the same r, the width of the confidence interval is not altered. By taking r = R, the resultant length of the variable in the original data, the vectors representing the lower and upper confidence bands have similar lengths to the original vector.

We should remember that when calculating confidence intervals it is important to look at the bootstrap vectors as well as the interval, because if a large number of bootstraps means that the vectors (after reflection) are distributed over more than say, 180°, then finding a confidence interval for the true direction of a variable may not be sensible.

8.4.2 Confidence Interval from Fisher (1993)

If, having tested the angles for symmetry about the original direction, we find that this exists, we propose adapting the confidence interval in Chapter Five of Fisher (1993), using our B bootstraps instead of the n values in the sample and taking R to be the resultant of the original data (rather than the mean resultant of the n values in the sample). Instead of the definitions given in Fisher (1993), we calculate the circular dispersion, $\hat{\delta}_B$ and the circular standard error, $\hat{\sigma}_B$, of the B bootstraps using:

$$\hat{\sigma}_B^2 = \frac{\hat{\delta}_B}{B}$$

where $\hat{\delta}_B = \frac{(1-\hat{\rho}_2)}{2R^2}$;

$$\hat{\rho}_2 = \frac{1}{B}\sum_{i=1}^{B}\cos 2(\theta_i - \hat{\theta});$$

θ_i is the direction that the i-th bootstrap makes with the horizontal;

$\hat{\theta}$ is the original direction;

R is the resultant of the original variable direction.

An approximate 100(1-α)% confidence interval for the original direction, $\hat{\theta}$, is then given by:

$$\hat{\theta} \pm \sin^{-1}(z_{\frac{1}{2}\alpha}\,\hat{\sigma}_B) \qquad (8.3)$$

where $z_{\frac{1}{2}\alpha}$ is the upper $100\left(\frac{1}{2}\alpha\right)$% point of the N(0,1) distribution. However, because the denominator of $\hat{\sigma}_B^2$ is B, the number of bootstraps, the confidence intervals will be smaller the larger the number of bootstraps generated. We can only, therefore, compare these intervals across variables for a given number of bootstraps. We cannot compare them with the BC_a interval because this does not depend on the number of bootstraps.

8.4.2.1 Application to Simpson Desert Flint Tools

In this section we calculate confidence intervals for the true directions of all the original six variables for the flint tools and, additionally, intervals for the variables chosen under the backward elimination selection criteria of Chapter Seven (even though the vectors are not symmetric about the original variable direction — see 8.2.4.1). We generate 100, 1000 (and 5000) replicate matrices from the multivariate normal distribution and implement the principal component biplot. We list the original direction and values obtained from Fisher's confidence interval of (8.3) in Tables 8.6 and 8.7 below. Table 8.6 shows the 95% confidence intervals for each of the six variables.

Table 8.6 95% Fisher Confidence Intervals (°) for the True Variable Directions for the Simpson Desert Flint Tools

		Number of Bootstraps	
Variable	**Original Direction**	**100**	**1000**
Length	83	(69, 97)	(80, 86)
Width	329	(310, 348)	(322, 335)
Thickness	10	(353, 27)	(8, 12)
Platform Width	301	(293, 309)	(301, 301)
Platform Thickness	10	(344, 36)	(2, 17)
Weight	38	(20, 56)	(32, 43)

We see from the above table that the intervals based on 1000 bootstraps are smaller than those based on 100 bootstraps (see 8.4.2). However, we can still compare the relative widths of intervals for a given number of bootstraps between the original six variables and the three obtained from backward elimination variable selection. We can also compare the intervals for a given number of bootstraps across variables. Clearly, platform thickness has the widest interval, whereas platform width has a relatively small interval. We do not illustrate the other biplots here, but we find that the variables have confidence intervals of very similar magnitude to those in Table 8.6. Table 8.7 lists the intervals for the variables selected in Chapter Seven for the principal component biplot, using the backward elimination method.

Table 8.7 95% Fisher Confidence Intervals (°) for the True Variable Directions for the Simpson Desert Flint Tools (Selected Variables)

Variable	Original Direction	100 Bootstraps
Length	272	(262, 283)
Width	14	(4, 24)
Weight	349	(337, 2)

We should not place any emphasis on the differences between the original directions of the variables in Tables 8.6 and 8.7, because this is entirely due to the singular value decomposition. It is the relative directions of the variables within each table and the widths of the intervals that are relevant. It is evident that the intervals under variable selection are narrower than those for the same variables based on the original data and are all of similar width. The intervals based on variable selection are illustrated in Figure 8.4 — we note that none of them overlap.

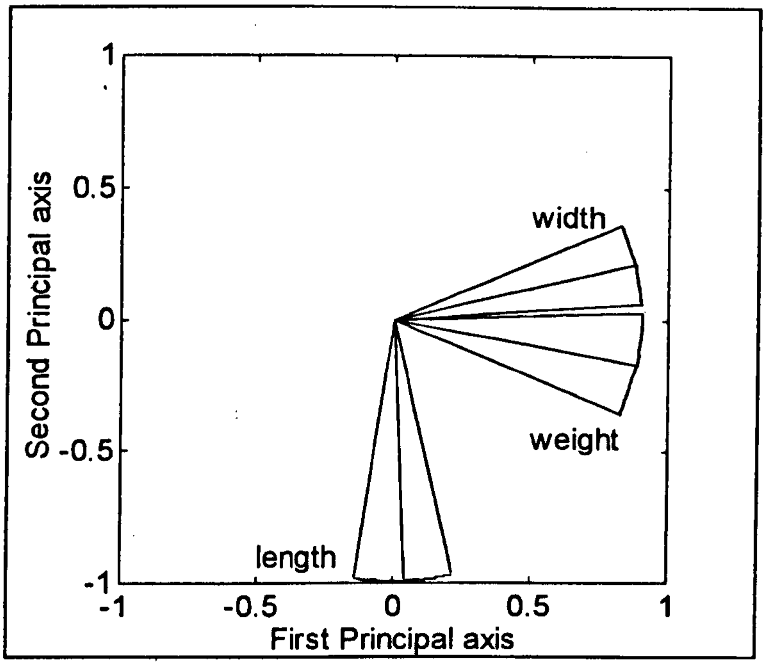

Figure 8.4 Original Directions and Fisher Confidence Intervals for the Simpson Desert Flint Tool Variables (Principal Component Biplot)

8.4.2.2 Application to Ceramic Pots

In this section we calculate confidence intervals for the true directions of the ceramic pot variables (1.2.5), using 100 replicate matrices. These are displayed in Table 8.8.

Table 8.8 95% Fisher Confidence Intervals (°) for the True Variable Directions for the Ceramic Pots

Variable	Original Direction	100 Bootstraps
1	305	(302, 309)
2	356	(352, 360)
3	203	(200, 205)
4	176	(173, 179)
5	350	(347, 353)
6	277	(274, 280)
7	306	(303, 310)
8	165	(162, 168)
9	60	(56, 63)
10	40	(37, 43)
11	86	(79, 93)
12	52	(49, 55)
13	148	(145, 152)

We see from the table that the intervals are all of a similar width, except for that of variable 11, which is roughly twice the size of the others; we recall that this is the least well represented variable in two dimensions (see the discussion in 8.3.1.2 and 3.7.1). It is also clear that the intervals are considerably narrower than those obtained from using the standard confidence interval (Table 8.3) — they are roughly $\frac{1}{10}$th of the size.

8.4.3 The von Mises Distribution

Given that we are dealing with the angles that the variables make with the horizontal axis and therefore with circular data, rather than use the standard normal distribution (as in the standard interval, the BC$_a$ method and Fisher's method) it may be possible to use a circular analogue. One such distribution is the von Mises distribution — Mardia (1972) says that the importance of the von Mises distribution on the circle is similar to that of the normal distribution on the line. We propose that the von Mises distribution, if it fits the data, can be used to obtain intervals for the true directions of the variables. To establish whether this distribution is appropriate, Fisher (1993) suggests using either a quantile-quantile plot or a formal test. We use the formal test in the next section.

8.4.3.1 Application to Ceramic Pots

We generate B = 100, 1000 and 5000 replicate matrices for the ceramic pots and test, for each variable, whether the angles that the vectors make with the direction due east follow the von Mises distribution. We discover that this is not the case for any of the variables from either the correlation, covariance, coefficient of variation, Spearman rank correlation or principal component biplot and so this is not an appropriate distribution to use to obtain confidence intervals for the true directions of the variables for these data.

8.4.4 Calculating the Mean Direction using Vectors of Equal Length

Although we are primarily concerned with estimating confidence intervals for the true directions of biplot variables, it is also of interest to consider the mean directions of the bootstrap fans. These can then be compared with the directions of the original variables and provide another indication of whether the fans are centred (a mean direction which is far from the original direction suggests non-centred fans). Mardia (1972) described how to calculate the mean direction of angular data and the details are given below. If all the vectors are considered to lie on the unit circle then, using the notation of Mardia (1972), we let P_i be the point on the circumference of the unit circle corresponding to the angle θ_i, where i = 1,..., n. The mean direction, $\overline{x}_o$, of $\theta_1,\ldots, \theta_n$, is defined to be the direction of the resultant of the unit vectors $\overline{OP_1},\ldots, \overline{OP_n}$. The cartesian co-ordinates of P_i are ($\cos \theta_i$, $\sin \theta_i$), where i = 1,..., n so that the centre of gravity of these points is $(\overline{C}, \overline{S})$ where:

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i \qquad \text{and} \qquad \overline{S} = \frac{1}{n}\sum_{i=1}^{n}\sin\theta_i .$$

Therefore, if:

$$\overline{R} = (\overline{C}^2 + \overline{S}^2)^{\frac{1}{2}}$$

then $R = n\overline{R}$ is the length of the resultant and $\overline{x}_o$ is the solution of the equations:

$$\overline{C} = \overline{R}\cos\overline{x}_o$$
$$\overline{S} = \overline{R}\sin\overline{x}_o .$$

8.4.5 Calculating the Mean Direction using Vectors of Unequal Length

In Chapter Three we explained that for the correlation and Spearman rank correlation biplots, the closer the length of a vector is to one, the better the corresponding variable is represented. Although we do not actually calculate mean directions here, we propose, for these two biplots, weighting the angles that the vectors make with the horizontal by their lengths, with the justification that a vector which is better represented should make a greater contribution to the mean direction of the vectors. If the vector making an angle θ_i with the horizontal is of length a_i, then we define:

$$\overline{C}_1 = \frac{\sum_{i=1}^{n}(a_i\cos\theta_i)}{\sum_{i=1}^{n}a_i} \qquad \text{and} \qquad \overline{S}_1 = \frac{\sum_{i=1}^{n}(a_i\sin\theta_i)}{\sum_{i=1}^{n}a_i}$$

with $\overline{R}_1 = (\overline{C}_1^2 + \overline{S}_1^2)^{\frac{1}{2}}$. The weighted mean direction, $\overline{x}_1$, is given by the solution of the equations:

$$\overline{C}_1 = \overline{R}_1\cos\overline{x}_1$$
$$\overline{S}_1 = \overline{R}_1\sin\overline{x}_1 .$$

8.5 Assessing Stability by using a Jack-knife Approach

So far we have proposed fitting the multivariate normal distribution to the data in order to help us to assess the stability of the biplot variables. In this section we suggest an alternative method, based on the jack-knife technique, although the results of the two methods are not directly comparable. We propose that each observation is omitted in turn, a biplot is implemented on each reduced data matrix and the co-ordinates of the vectors representing the variables are obtained. The width of these 'jack-knife fans' are then examined and as for bootstrap fans, the wider the fan the less stable the variable. We also suggest using this method to detect influential observations and we discuss this in Section 8.8. However, because with 'jack-knife fans' the vectors in the fans are not obtained from simulated data, but use the actual data (with one observation omitted), the fans are likely to be narrower than 'bootstrap fans', although are heavily dependent on both unusual observations and to some extent on the number of observations.

8.5.1 Application to Simpson Desert Flint Tools

We recall from Section 1.2.6 that there are 78 flint tools with six variables measured on each. Each of the tools is omitted in turn and a biplot is produced for each set of 77 tools. This results in 78 sets of co-ordinates for each set of six variables, which are overlaid on the same plot. We illustrate the correlation biplot in Figure 8.5, although the plots are very similar for the other types of biplot. Only three variables are illustrated because some of the vectors for different variables overlap, but all six variables are used in the analysis. Comparing these fans with those of Section 8.2, we see that the fans are much narrower under this jack-knife method.

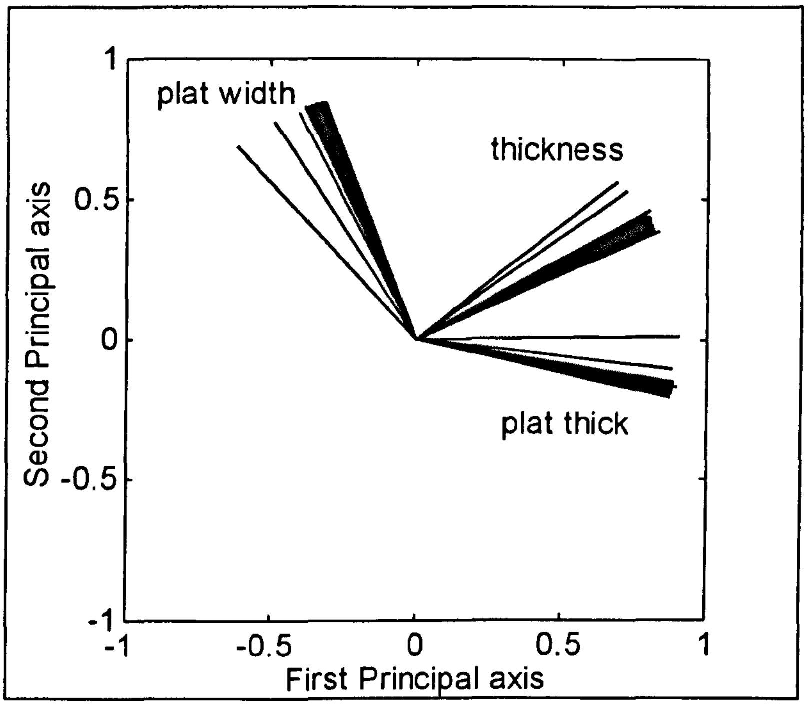

Figure 8.5 Simpson Desert Flint Tool Variables — Jack-knifing (Correlation Biplot)

8.6 The Influence of Sample Size

Having introduced bootstrapping and jack-knifing as methods of investigating stability and considered three different confidence intervals, we now investigate the influence of sample size on biplots. More specifically, we investigate how the number of observations measured affects the relationships between the variables. It may then be possible to make recommendations on the number of observations to measure in order to answer the questions posed by the archaeologists and ecologists. If we take many samples of size $h < n$, without replacement, where n is the original number of observations measured, then we can evaluate the stability of the variables as if we had originally taken a sample of size h. Sampling without replacement is not appropriate for evaluating the effect of larger samples than that obtained and so for this we must sample by using the multivariate normal distribution.

8.6.1 Sampling Without Replacement

When using sampling without replacement, observations can only be retained at most once and so this form of resampling is only suited to answering questions concerning smaller sample sizes than those actually obtained. We apply this method to the ceramic pots in the next section.

8.6.1.1 Application to Ceramic Pots

We recall from Chapter One that 13 measurements were taken on each of 30 ceramic pots, with 10 pots made by each of three potters. Therefore, when generating smaller samples of pots, it seems sensible to ensure that we obtain equal numbers of pots from each potter.

In 8.2.4.2 we revealed that for each of the 13 variables the bootstrap vectors were centred about the direction obtained from the original data. In this section we generate 100 bootstraps by sampling without replacement, implement the correlation biplot and apply the standard confidence interval. We do this for each of the sample sizes in Table 8.9 below. The BC_a interval is not appropriate for use with smaller samples than that actually collected, because the acceleration (see 8.3.2) cannot be calculated.

Table 8.9 95% Standard Confidence Intervals (°) for the True Variable Directions for the Ceramic Pots (Smaller Sample Sizes: Without Replacement)

	Sample Size		
Variable	21	15	9
1	(275, 336)	(230, 21)	(159, 92)
2	(325, 28)	(303, 409)	(266, 86)
3	(176, 229)	(157, 248)	(137, 268)
4	(152, 199)	(135, 217)	(113, 239)
5	(323, 16)	(303, 37)	(264, 75)
6	(244, 310)	(222, 332)	(192, 2)
7	(274, 338)	(252, 360)	(218, 35)
8	(142, 188)	(120, 210)	(96, 235)
9	(22, 98)	(340, 140)	(277, 202)
10	(6, 74)	(346, 94)	(321, 119)
11	(41, 132)	(335, 198)	(308, 224)
12	(21, 83)	(2, 103)	(341, 123)
13	(118, 178)	(97, 199)	(57, 239)

Table 8.9 shows that the smaller the sample size the wider the interval and so the less confident we are that the directions of the original variables are representative of those of the true population of data. Again (see Tables 8.3 and 8.8), variable 11 has the widest interval. We also need to consider larger sample sizes than that collected, but we cannot do this by sampling without replacement. Instead, we use the multivariate normal distribution.

8.6.2 Sampling using the Multivariate Normal Distribution

By fitting a multivariate normal distribution to the data as described in Section 8.2.1, we can generate a sample of any size. We can then look at the effects of sample size on the resulting groupings of the observations in the biplot, on the relationships between the variables and on the confidence intervals for the true directions of the variables.

8.6.2.1 Application to Ceramic Pots

In this section we generate 100 bootstraps of the ceramic pot data for the correlation biplot, using each of the sample sizes in Table 8.10. We then calculate the standard confidence interval of Section 8.3.1. An asterisk indicates that the intervals cover over 360° and are therefore not sensible.

Table 8.10 95% Standard Confidence Intervals (°) for the True Variable Directions for the Ceramic Pots (Smaller Sample Sizes: Multivariate Normal Distribution)

	Sample Size		
Variable	**21**	**15**	**9**
1	(211, 40)	(195, 56)	(93, 158)*
2	(298, 54)	(266, 86)	(242, 110)
3	(156, 249)	(147, 258)	(102, 304)
4	(132, 220)	(123, 229)	(80, 272)
5	(301, 39)	(298, 42)	(265, 74)
6	(226, 328)	(217, 337)	(177, 18)
7	(249, 3)	(239, 14)	(202, 50)
8	(125, 205)	(115, 215)	(78, 253)
9	(342, 137)	(319, 160)	(219, 260)*
10	(350, 91)	(339, 101)	(293, 146)
11	(343, 189)	(296, 236)	(238, 295)*
12	(2, 102)	(351, 113)	(308, 157)
13	(97, 200)	(72, 224)	(32, 264)

Comparing with Table 8.9, we see that the intervals in Table 8.10 are slightly wider and for samples of 9 pots, three variables have intervals exceeding 360° — 9 pots are clearly too few for any sensible conclusions to be drawn from the data. We can also use the multivariate normal distribution to generate larger samples than that actually obtained and we display the results of this for the correlation biplot in Table 8.11.

Table 8.11 95% Standard Confidence Intervals (°) for the True Variable Directions for the Ceramic Pots (Larger Sample Sizes: Multivariate Normal Distribution)

	Sample Size		
Variable	**75**	**60**	**45**
1	(285, 326)	(221, 30)	(243, 8)
2	(333, 20)	(319, 33)	(321, 32)
3	(185, 221)	(171, 234)	(175, 231)
4	(156, 196)	(145, 207)	(146, 206)
5	(332, 8)	(317, 22)	(322, 17)
6	(255, 299)	(243, 311)	(245, 309)
7	(280, 333)	(268, 345)	(269, 344)
8	(147, 183)	(134, 196)	(136, 194)
9	(37, 82)	(23, 96)	(25, 94)
10	(19, 61)	(5, 74)	(8, 72)
11	(34, 139)	(31, 142)	(16, 157)
12	(34, 71)	(19, 85)	(21, 84)
13	(125, 172)	(114, 182)	(113, 183)

The intervals in Table 8.11 are all smaller than those of Table 8.10, indicating that the larger the sample size the more stable the variables. For a sample of 75 pots the intervals span ≈ 40°, for 60 pots ≈ 60° and for 21 pots ≈ 100°; variable 11 has much wider intervals across all sample sizes.

8.7 Sample Size and Variable Selection

In the previous section we investigated the influence of sample size on the stability of variables and in Chapter Seven we discussed variable selection methods. We believe that recommendations on the size of the sample and on the number of variables measured on each observation should not be made independently, because one may influence the other. If an archaeologist has some idea of how many artefacts (e.g. tools or pot sherds) are potentially available, then it can be suggested in advance of the 'excavation' how many measurements should reasonably be taken on each artefact. However, if there is no clear idea of the number of artefacts available then it may be advisable for statistical analysis to be undertaken after an initial number of artefacts have been measured; it may then be possible to refine future recording by measuring fewer variables, particularly if the site is to be visited in future seasons. In the following section we consider the influence of sample size and variables selected together.

8.7.1 Application to Ceramic Pots

In this section we consider the sample sizes of Tables 8.10 and 8.11 and we combine these with the subset of variables obtained (8 and 9) using Krzanowski's backward elimination variable selection procedure of Chapter Seven for the correlation biplot (even though this method did not produce a good separation of pots into groups). Table 8.12 lists the intervals obtained from 100 bootstraps from the multivariate normal distribution.

Table 8.12 95% Standard Confidence Intervals (°) for the True Variable Directions for the Ceramic Pots (Selected Variables: Multivariate Normal Distribution)

	Variable	
Sample Size	**8**	**9**
75	(6,78)	(88,188)
60	(345,99)	(83,193)
45	(338,106)	(79,197)
21	(260,184)	(20,256)
15	(234,211)	(0,275)
9	(258,187)	(359,277)

Comparing Table 8.12 with Tables 8.10 and 8.11, we see that the intervals in the above table are clearly wider than those obtained for variables 8 and 9 when all 13 variables were used in the analysis; this is true across all the sample sizes. It therefore seems that when fewer variables are used in the analysis, the confidence intervals for the true directions become wider. Considering Table 8.12, we see that the intervals do not become much wider as we go from 15 pots to 9, but in Table 8.10 the intervals are still becoming much wider as the sample size is reduced.

8.8 Detecting Influential Observations by using a Jack-knife Approach

In this section we propose using the idea of 'jack-knifing', introduced to assess stability in Section 8.5, in order to identify influential observations in the data (which may also be causing instability in the biplot variables, leading to wider confidence intervals for the true directions of variables than would otherwise be the case). It is often the case that outlying observations are visible in biplots (see e.g. 3.7.2.1), but by omitting one observation at a time we are able to examine the influence of each individual observation. However, this is particularly time consuming when large numbers of observations have been collected and would not be viable for more than a few hundred.

We propose following the same methodology as in Section 8.5, but the interpretation of the resulting plot is different. The vectors representing each variable are plotted as before, but this time any that are clearly distinct from the majority are flagged as unusual and the corresponding omitted observation is examined further. We also discuss methods of establishing which observation is the 'most influential'.

8.8.1 Application to Simpson Desert Flint Tools

Implementing the jack-knife method on the flint tools (1.2.6) results in an identical plot to Figure 8.5, where we see that for each variable there are two vectors that are slightly removed from the remainder. Given that, for a particular variable, each vector results from omitting one observation (in this case tools), the plot suggests that there are two slightly unusual tools. When we refer back to the raw data we see that these tools are both from site 08. Considering univariate analyses, boxplots of individual variables show these tools to be particularly unusual on variables width, thickness and platform width, moderately different on variables platform thickness and weight, but not at all different on length. The jack-knife plot of variable length does, however, highlight two tools as being unusual. If these tools are removed from the data and the analysis repeated then it may be possible to obtain a better separation between tools according to sites, but we do not show this here. It is also likely that the confidence intervals for the true directions of the variables will become smaller.

We now discuss how to assess which observation is the 'most influential'. Potentially influential observations can be identified by eye from the jack-knife plots, as we have done above for the flint tools. Alternatively, we could make use of the procrustes statistic (see 6.8.2). Specifically, we could delete each observation in turn, implement a biplot and compare the resulting variable co-ordinates with those of the original data, obtaining a value of M^2 for each deleted observation. The observation with the largest procrustes M^2 can then be considered to be the 'most influential'.

8.9 Summary and Conclusions

This chapter has introduced the idea of assessing the stability of biplot variables by using the multivariate normal distribution and has identified several methods of obtaining confidence intervals for the true directions of the variables (i.e. the directions taken if the whole population of data, rather than a sample, had been measured). We also developed methods for projecting supplementary observations and variables onto the original biplot axes. In addition, we discovered that replicate co-ordinates cannot be directly projected onto the original co-ordinate system because, after fitting a multivariate normal distribution to the data, there is no reason why the first observation of a generated matrix should correspond to the first observation of the original data. Instead, the co-ordinates for the replicate matrices must be obtained by implementing a biplot on each matrix separately — the resulting vector co-ordinates for a particular variable are referred to as bootstrap vectors.

We have also developed the notion of a 'bootstrap fan' to describe a set of bootstrap vectors for a particular variable and we proposed using these fans to obtain confidence intervals for the true directions of the variables (i.e. for the whole population of data). It is clear that a biplot in two dimensions is not always appropriate, because two dimensions can be insufficient to both explain a high proportion of the variation in the data and also for each individual variable to have a high quality of representation. This can lead to the bootstrap fans not being centred about the original variable directions. It is also evident that variables that are poorly represented in the chosen dimensionality (typically two) have particularly wide confidence intervals for their true directions. In addition, we proposed using the jack-knife technique to assess the stability of the variables and this produces much narrower fans than those obtained from using the multivariate normal distribution.

Two traditional confidence intervals were considered for assessing the true direction of the variables — the standard interval (symmetric) and the BC_a method. These were implemented on the angles that the vectors representing each variable make with the direction due east and the choice of interval usually depends on whether the fans are centred or not. The BC_a method leads to smaller intervals and is considered to be the better method because it does not necessarily produce a symmetric interval (it uses percentiles that are based on the replicate vectors). We also proposed an adaption of the interval in Chapter Five of Fisher (1993), which was developed specifically for directional data, although this is only useful for comparing the relative widths of intervals across variables because it becomes narrower the more bootstraps that are generated. It can also be used to compare the interval widths of the original variables with those of the variables chosen using the variable selection methods of Chapter Seven. We suggested using the von Mises distribution to obtain confidence intervals because it was developed specifically for circular data, but for our data it was not appropriate. We also proposed accounting for the length of the bootstrap vectors when calculating mean variable directions for the correlation and Spearman rank correlation biplots, because for these biplots vector lengths represent the quality of representation of the corresponding variable (and therefore we believe that the longer vectors should be given greater weight). Mean directions are another means of assessing how close the replicate vector directions are to the directions of the original variables.

Sampling without replacement and sampling by fitting a multivariate normal distribution, in conjunction with confidence intervals, were used to investigate the influence of sample size on the stability of the variables. However, the BC_a method is not appropriate for smaller samples than that actually collected because the acceleration cannot be calculated. We also combined varying sample sizes with the variables selected under the backward elimination method of Chapter Seven, in order to investigate how selection affects the stability of the variables. It is clear that the fewer variables that are used in the analysis, the wider the confidence intervals for their true directions. Finally, we introduced the jack-knife technique into the biplot framework as a means of detecting influential observations and illustrated how such observations are highlighted on the resulting biplot. We suggested that the most influential observation can be identified by deleting each observation in turn, implementing a biplot and comparing the resulting variable co-ordinates with those of the original data, to obtain a value of the procrustes M^2 statistic for each deleted observation. The observation with the largest M^2 can be considered to be the most influential. This proved to be a very useful method.

Chapter Nine

Stability, Selection Methods, Sample Size and Canonical Correspondence Analysis

9.1 Introduction

In Chapter Four we explained the theory behind canonical correspondence analysis and illustrated its application to data on hunting spiders (1.2.8) and dune meadow vegetation (1.2.9). Chapter Four also raised questions regarding the effect of the number of categories (sites or species) and the size of the sample (number of sites or number of individuals recorded at each site) on the results of the analysis. In addition to the above, it was suggested that there may be a maximum number of environmental variables above which either the map becomes too cluttered for patterns in the data to be revealed, or where the multicollinearity between some of the variables is extremely high. This chapter uses techniques such as bootstrapping, procrustes analysis and jack-knifing, in combination with canonical correspondence analysis in order to answer questions such as those raised above. We also examine the stability of the sites obtained as a result of the analysis (i.e. how representative are the samples obtained at each site of the true population of data), by considering confidence regions based on convex hulls and concentration ellipses.

As we explained in Chapter Four, canonical correspondence analysis (CCA) is applied to data which consist of species abundances recorded at a number of sites, together with a set of environmental variables measured at each site. Thus, there is scope for applying variations on the techniques discussed in Chapters Five and Six (for correspondence analysis) to the species data and adapting the methods of Chapters Seven and Eight (which concerned biplots) to the environmental variables.

The structure of this chapter is as follows. In Section 9.2 we propose a method of assessing the stability of the sites in the CCA map and in 9.3 we discuss how the stability of the environmental variables might be investigated. Section 9.4 describes an existing technique for selecting a subset of environmental variables, before proposing an alternative method. One method of choosing which categories (usually sites) to delete from the analysis is introduced in 9.5 and the influence of sample size (i.e. the total species abundances) on the analysis is addressed in 9.6. Jack-knifing as a method of assessing stability is discussed in 9.7 and this same method is proposed in 9.8 for detecting influential categories. Connections between CCA and both biplots and correspondence analysis are discussed in 9.9 and the chapter is concluded in Section 9.10.

9.2 Site Stability

One of our main interests lies in investigating whether small differences in the species-by-sites matrix can produce relatively large differences in the CCA display, because this could suggest that our data sample is not representative of the true population of all possible data and therefore our inferences based on the analysis are of limited use (care must be taken in distinguishing between unrepresentative samples and influential categories — see the discussion in 9.8). As we explained in Chapter Five for correspondence analysis, if it were possible to obtain more data in exactly the same way as that already collected (i.e. by using the same sampling scheme), then we could repeat this process many times to obtain a set of replicate data matrices, each of which could be subjected to CCA to produce a new set of points (clouds), which give some indication of the stability of the data. However, because this repeated sampling is not usually possible (although we believe that it is more feasible for collecting species data as opposed to artefacts, because the former are likely to be more abundant), we treat the observed sample as a proxy for the underlying distribution and draw new samples from it. There are two main ways in which we can resample from the contingency table, namely by using the multinomial distribution or by sampling without replacement: these methods were described in Chapter Five. However, the latter is only useful for answering questions concerning smaller sample sizes than those actually obtained. We discuss the stability of both the dune meadow vegetation sites (1.2.9) and the hunting spider sites (1.2.8) in the following sections; both data sets raise different questions regarding the interpretation of the results obtained from the multinomial resampling.

Because a singular value decomposition (SVD) is involved in implementing CCA, it is possible that arbitrary reflection of the resulting co-ordinates may occur (it is known that the SVD is unique only up to sign changes in the eigenvectors — see Chapter Five) and this is corrected for throughout this chapter where necessary.

9.2.1 Application to Dune Meadow Vegetation

Using the species-by-sites dune meadow vegetation data (described in Chapter One and discussed in Chapter Four), which were collected using van der Maarel's scale (see Table 4.1) and treating each site as a separate multinomial sample (because this most closely resembles the original data collection strategy), we generate 200 replicate matrices. We leave the environmental variables unchanged so that the same environmental data are combined with each replicate species-by-sites matrix. The resulting bootstrap clouds for a subset of the sites (to avoid confusion) are illustrated in Figure 9.1. Because van der Maarel's scale was used for these data, fitting a multinomial distribution is not really appropriate, as the values on the scale are not frequencies but ordinal measures of abundance. However, in the absence of any other form of the data we use this method of resampling. Interest lies in assessing how representative the sample at each site is of the true population of species found at that site and also in determining which sites are similar in terms of the distribution of species that they contain.

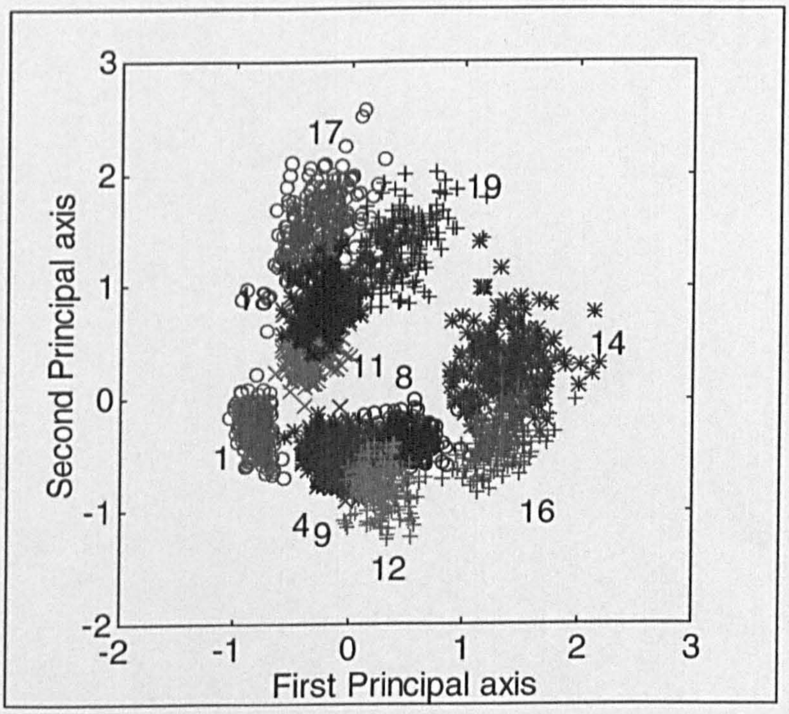

Figure 9.1 Two Hundred Bootstrap Points of Dune Meadow Sites

From Figure 9.1 we see that despite the fact that only a subset of the sites are displayed, there is still considerable overlap between the clouds of points from some sites. Of the sites represented, only site 1 is distinct (i.e. its cloud does not overlap with those of other sites), which suggests that this site consists of either different species, or of different proportions of the same species as compared with the other displayed sites. For sites with larger clouds (e.g. site 14), we are less certain that the samples obtained from these sites are representative of the true population of data, although both the degree of overlap and overall cloud size are related to the number of bootstraps generated (see Chapter Five). However, because we are making only informal inferences, we do not believe that this problem is too severe.

Although Figure 9.1 does not illustrate the environmental variables, because these were not our prime concern, it is interesting to note that the positions of these variables on the map also alter even though the environmental data have remained the same for each bootstrap. This is, of course, because of the algebra of CCA (see 4.3.3), which incorporates multiple regression of the site scores on the environmental variables into the iteration algorithm.

9.2.2 Application to Hunting Spiders

In this section, we treat each hunting spider site (1.2.8) as a separate multinomial sample and compare the results obtained from using both the original and transformed hunting spider data (see 4.5). This will nearly always be the appropriate sampling scheme to implement, because it is usually the case that a number of sites are chosen and the abundances of the species present at these sites are then recorded. Generating 500 bootstraps for the untransformed data (both species-by-sites and sites-by-environmental variables) and displaying a subset of the sites so as not to overcrowd the diagram, leads to Figure 9.2. We see from the figure that site 26 has a particularly large cloud associated with it, whereas sites 5 and 14, for example, have much smaller clouds.

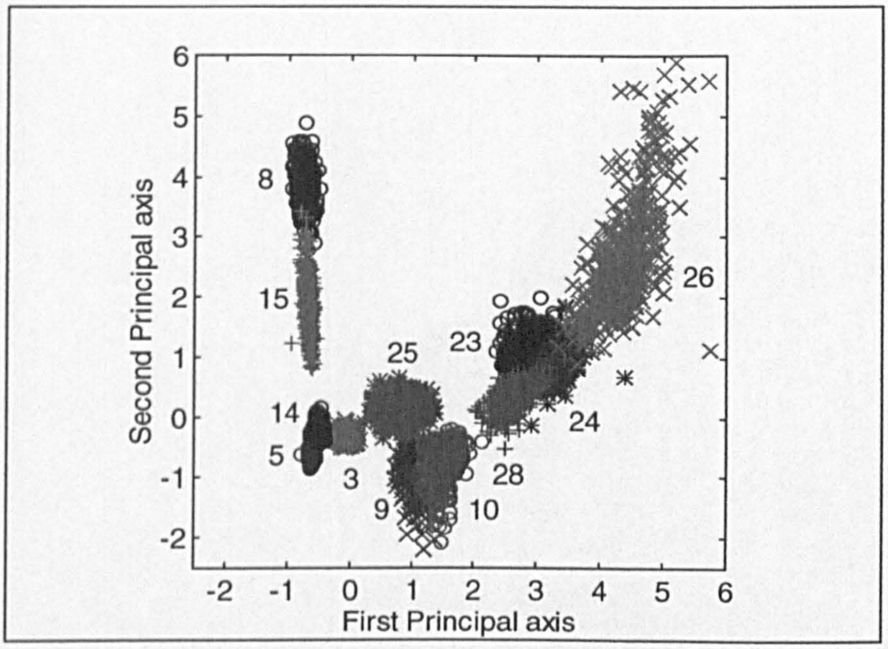

Figure 9.2 Five Hundred Bootstrap Points of Hunting Spider Sites

Transforming both the species data and the environmental data as described in Section 4.5 and generating 500 bootstraps produces Figure 9.3, where it is clear that the clouds are much larger than those of 9.2. On the basis of this figure we would therefore conclude that the samples obtained at the sites are much less stable i.e. less representative of the true population of data. Because of the large degree of cloud overlap, we also infer that there is a great deal of similarity between the species found at the corresponding sites. However, it is not satisfactory for the clouds to vary so much purely because of the form of the data and we therefore emphasise that careful consideration must be given to the data before any analysis is undertaken.

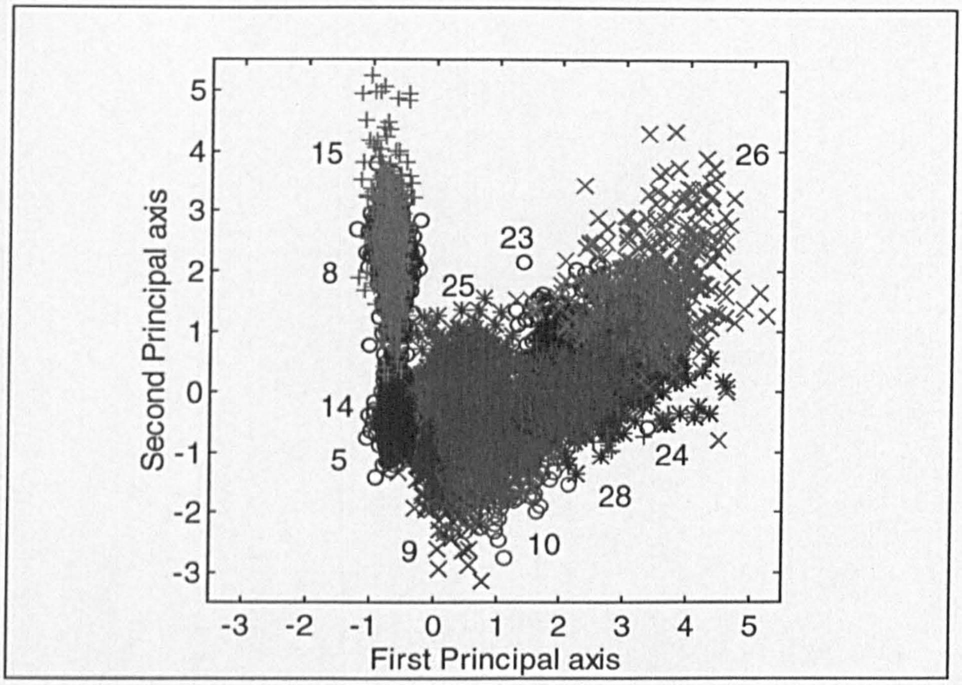

Figure 9.3 Five Hundred Bootstrap Points of Hunting Spider Sites (Transformed)

9.2.3 Convex Hulls and Concentration Ellipses

As described in Chapter Five, convex hulls and concentration ellipses can be used to summarise the points resulting from bootstrapping, although the sizes of the clouds depend on the number of bootstraps generated. We again propose calculating the areas of hulls and ellipses and therefore comparing the stability of each site (smaller areas indicate greater stability i.e. the sample obtained at the corresponding site is more representative of the true population of data). The methodology was described in Section 5.4 and is applied to the dune meadow vegetation data below.

9.2.3.1 Application of Concentration Ellipses to Dune Meadow Vegetation

In this section we obtain concentration ellipses and their areas for the dune meadow vegetation sites (1.2.9). Having generated 200 bootstraps, Figure 9.4 illustrates 95% concentration ellipses for the same subset of sites as in Figure 9.1. The relative sizes of the clouds are now much clearer and the ellipses of all the displayed sites, except that of site 1, show some degree of overlap. For example, the ellipses representing sites 4 and 9 overlap considerably and consulting Table A.15 of the Appendix, we see that these sites have fairly similar distributions of species. In contrast, site 14 is some distance away and it only has three species in common with site 9.

In Chapter Five we reported results due to Ringrose (1992), namely that when applying correspondence analysis, the sites with the largest clouds have similar numbers across the species and low numbers in each cell of the matrix. However, examining the cloud sizes reveals that this does not appear to be true for CCA and we suspect that this is because of the influence of the environmental variables.

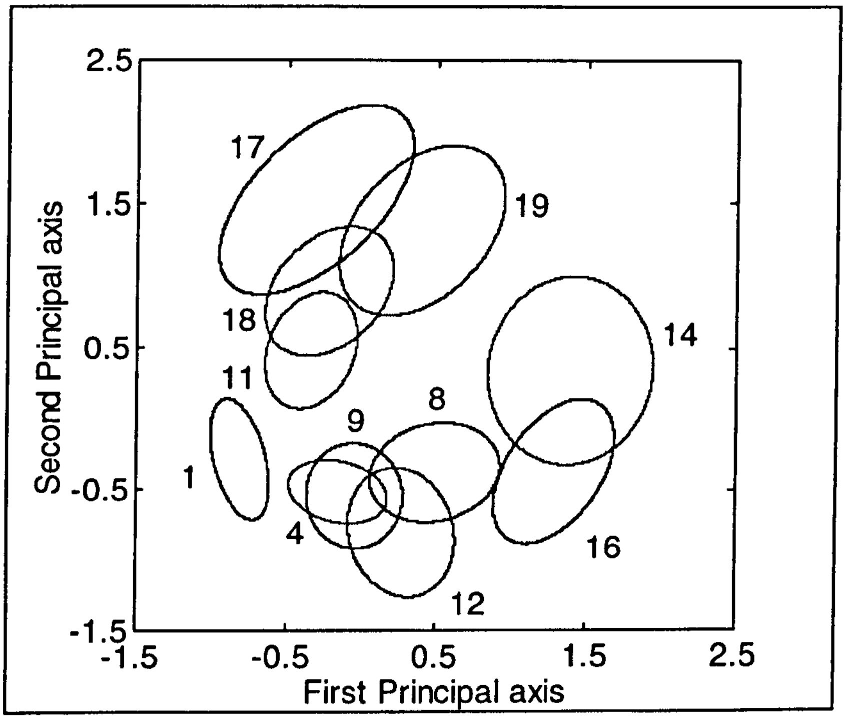

Figure 9.4 95% Concentration Ellipses for the Dune Meadow Sites

The areas of the 95% concentration ellipses for all the sites (referred to as A_4 in Chapter Five) are given in Table 9.1, where we see that site 14 is the most unstable and site 1 is the most stable. In Section 5.4.4.3 we suggested, as a rule of thumb, that if the centroid of a 95% ellipse representing a site is included in the 95% ellipse of another site, then these sites can be considered to be virtually indistinct in terms of their profile of species. Applying this to Figure 9.4, we infer that sites 4 and 9 are indistinct, although when more than two ellipses exhibit considerable overlap this rule of thumb will run into problems.

Table 9.1 The Measure A_4 for the Dune Meadow Sites

Site	Ellipse Area	Site	Ellipse Area
1	0.240	11	0.388
2	0.310	12	0.513
3	0.329	13	0.443
4	0.333	14	1.154
5	0.298	15	1.011
6	0.338	16	0.577
7	0.282	17	1.094
8	0.481	18	0.576
9	0.382	19	0.946
10	0.255	20	0.705

9.2.3.2 Application of Convex Hulls to Dune Meadow Vegetation

In this section we illustrate convex hull peeling for one of the dune meadow vegetation sites. Applying convex hull peeling, using the Green-Silverman algorithm (Green & Silverman, 1979) described in Section 5.4, to the bootstrap cloud for site 14 (illustrated in Figure 9.1), produces Figure 9.5. This figure shows that there are 18 peels in total and that the outer convex hull is some distance away from the remaining hulls. It is also apparent that considering the outer hull, as compared with the hull containing approximately 50% of the points, leads to a very different estimate of site stability.

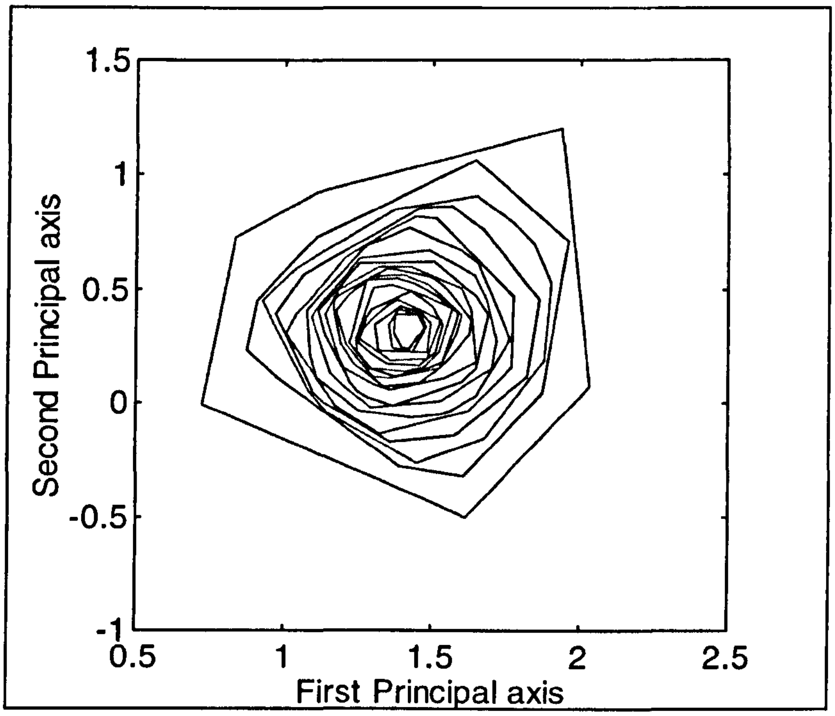

Figure 9.5 Convex Hull Peels of Dune Meadow Site 14

9.3 Environmental Variable Stability

Besides investigating the stability of the sites, it is also important to consider the stability of the environmental variables. Because the environmental data always consist of variables measured at a sample of all possible sites, we need to consider how representative these data are of the true population of data i.e. how stable are the variables. As explained in Chapter Eight for biplots, we need to first fit a distribution to the data matrix and then bootstrap from this distribution.

For quantitative environmental variables we need to fit a distribution to the data (perhaps after suitable transformations). However, there are often zeroes in the data and so the multivariate normal distribution is not really appropriate (by fitting this distribution we are ensuring that no zero values are generated). This is particularly problematic when a zero indicates that the combination of an environmental variable and site is not possible. For nominal variables the choice of which distribution to fit is even more difficult, because the levels of each variable are arbitrarily assigned a number e.g. 1, 2, 3, 4 for a variable with four levels, but there is no scaling involved (i.e. a value of 4 does not represent twice the value of 2). Additionally, because environmental variables are measured at sites (rather than variables measured on a sample of observations as for biplots) and because this information is then combined with species data, the relative magnitudes of each variable across the sites need to be retained and it is difficult to see how this can be achieved.

If we could decide on appropriate distributions to fit to the data, we could generate replicate environmental variable matrices, keeping the species-by-sites matrix fixed and implement CCA to obtain bootstrap fans (see Chapter Eight) and clouds for the quantitative and qualitative environmental variables respectively. These give an indication of the stability of the original environmental data and can be used to obtain confidence intervals and regions for the true directions and locations of the variables respectively, i.e. for the whole population of data. By examining the stability of the environmental variables it may then be possible to use this information in variable selection methods, so that only the more stable variables are retained. Selection methods are discussed in the following section.

9.4 Environmental Variable Selection Methods

Sometimes a large number of variables have been measured at each site and this can lead to both cluttering of the CCA map and to high multicollinearity between some of the environmental variables. It can also be the case that measuring 'inappropriate' variables hides the true patterns in the species data. We therefore believe that it is important to consider variable selection methods, both for analysing present data and for providing guidelines for future data collection. In the next section we discuss an existing method of variable selection, before proposing an alternative in Section 9.4.2.

9.4.1 An Existing Method of Variable Selection

Ter Braak & Verdonschot (1995) describe a method of selecting environmental variables which uses forward selection and this is an option in the package CANOCO, version 3.1. The method works as follows. In step one, CCA is implemented using each environmental variable on its own. The variables are then ranked on the basis of their fit, where the measure of fit is the eigenvalue of the CCA. The statistical significance of the effect of each variable is tested by a Monte Carlo permutation test (as in Manly, 1991) where, if a p-value of less than 0.05 is obtained, the variable is considered to be significantly related to the species data at the 5% level. At the end of the first step the variable with the greatest fit is selected. After this, all the remaining environmental variables are ranked on the basis of the fit that each separate variable gives in conjunction with the variable(s) already selected, where the measure of fit is now the sum of all the eigenvalues obtained from CCA, with each variable as the only additional environmental variable. CANOCO reports the 'extra fit', which is the change in the sum of all eigenvalues of CCA if the associated variable is selected. Later steps proceed in the same way and we stop adding variables when they cease to be significantly related to the species data. The Monte Carlo test replaces the usual F- or t-tests in forward selection multiple regression.

9.4.1.1 Application to Hunting Spiders

In this section, we apply the variable selection method just described to the hunting spider data (1.2.8). Originally, 26 environmental variables were measured at each site (see Figure 4.6) and so this provides considerable scope for applying variable selection methods. Using the untransformed data, taking $\alpha = 0$ as in Section 4.5 and applying the above method in CANOCO leads to the following table of results. We stop at step 17 because from this point onwards all the p-values exceed 0.13.

Table 9.2 Order of Variables Selected for the Hunting Spiders (from 26)

Step	Variable Selected	Extra Fit	P-Value	Step	Variable Selected	Extra Fit	P-Value
1	3	0.58	0.01	**10**	16	0.04	0.05
2	6	0.36	0.01	**11**	17	0.03	0.03
3	22	0.32	0.01	**12**	15	0.03	0.16
4	4	0.10	0.01	**13**	9	0.03	0.09
5	26	0.05	0.04	**14**	11	0.03	0.10
6	25	0.05	0.03	**15**	24	0.03	0.05
7	10	0.04	0.03	**16**	12	0.02	0.23
8	8	0.06	0.01	**17**	18	0.02	0.21
9	19	0.03	0.17				

On the basis of the above table, we stop the selection after step 8 and retain variables {3, 4, 6, 8, 10, 22, 25, 26}, because it is after this point that the first variable is not significantly related to the species data at the 5% level. We note that only three of these variables are included in the set of six selected by ter Braak (1986) and used in 4.5.2. Carrying out CCA on this reduced set of eight variables reveals that there are two variables with high variance inflation factors — 25 and 26. The resulting CCA map is illustrated in Figure 9.6, where we see that variables 3 & 4 are highly correlated (this is indicated by the small angle between the vectors representing them). It is also evident that variables {6, 25, 26} are highly correlated, although 6 & 26 are virtually uncorrelated in Figure 4.1. Given that the species data have remained the same for both figures, this apparent change in correlated variables must be entirely due to the effects of the included variables. We also note that the species and site points are located in similar positions in both Figure 4.1 and Figure 9.6.

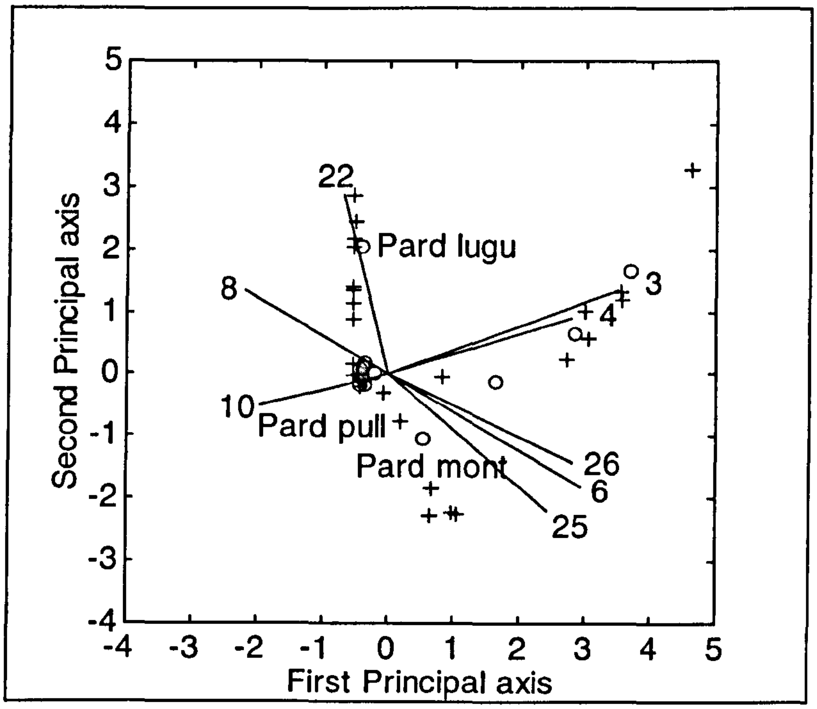

Figure 9.6 Canonical Correspondence Analysis Map of Selected Hunting Spider Variables (from 26)

For comparative purposes, we apply this forward selection method to the six variables of both the original and transformed data, discussed in Sections 4.5.1 and 4.5.2 respectively, which leads to the following table. The codes are those from Table 4.1.

Table 9.3 Order of Variables Selected for the Hunting Spiders (from 6)

	Original Data			Transformed Data		
Site	**Variable Selected**	**Extra Fit**	**P-Value**	**Variable Selected**	**Extra Fit**	**P-Value**
1	6	0.49	0.01	1	0.49	0.01
2	4	0.32	0.01	5	0.18	0.01
3	5	0.28	0.01	4	0.09	0.02
4	1	0.05	0.21	6	0.07	0.01
5	26	0.04	0.26	7	0.03	0.15
6	7	0.03	0.44	26	0.02	0.36

From the above table we see that when using the original data we only select three variables — {4, 5, 6}, although using the transformed data we select four (which includes the three selected in the untransformed data) — {1, 4, 5, 6}. A CCA map using these four variables is illustrated below in Figure 9.7 and all the variance inflation factors are less than 3.5. Again, the locations of the species and site points are similar to those in the previous figures. We also see that variables 1 & 4 are uncorrelated, but that 4 and 6 are located in similar positions to where they are in Figure 9.6.

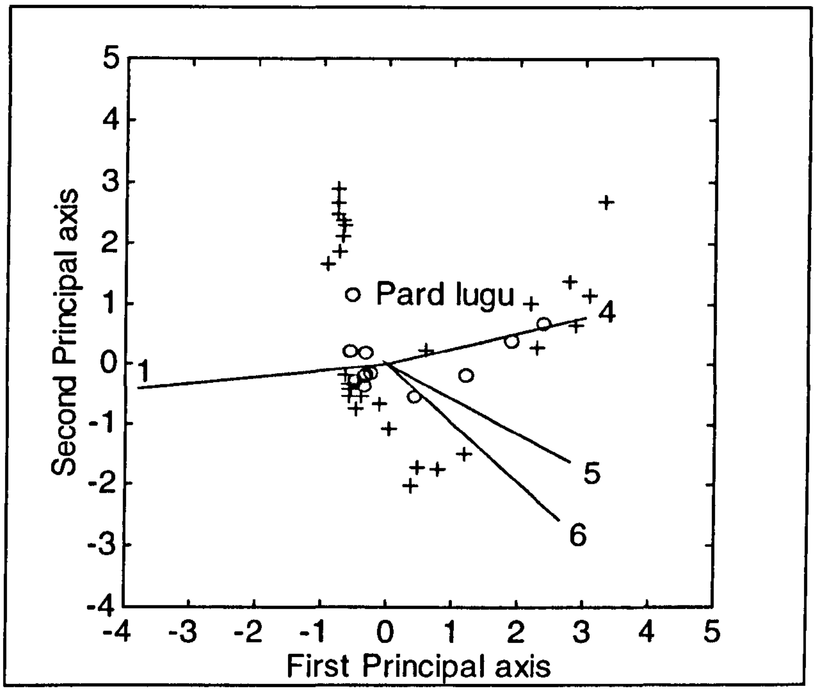

Figure 9.7 Canonical Correspondence Analysis Map of Selected Hunting Spider Variables (from 6)

In addition to the forward selection method implemented above, we could also use backward elimination, stepwise or all subsets methods, although these are not available in CANOCO 3.1 or in any other package as far as the author is aware. They are, however, relatively straightforward to implement in any standard programming language.

9.4.2 A Proposed Method of Variable Selection

In Chapter Seven we discussed variable selection methods for biplots and we believe that these can be adapted to CCA. For biplots, using the backward elimination method, each variable was omitted in turn and the resulting configurations of observation points were compared with that from all the original variables by using procrustes analysis. The reasoning behind this is that the removed variable, which corresponds to the least difference between configurations, has contributed least to the analysis and can therefore be permanently removed. However, in CCA we have two sets of points, one set representing the species and the other representing the sites of the species-by-sites matrix. We therefore propose the following backward elimination method.

Stage 1: Implement CCA on the original data and retain the co-ordinates of the sites followed by the co-ordinates of the species in the reference configuration X.

Stage 2: Delete each variable in turn and retain the site co-ordinates followed by the species co-ordinates in matrix Y.

Stage 3: Apply procrustes analysis to minimise trace$\{(X-Y)(X-Y)^T\}$ under translation, rotation and reflection of Y. This results in a residual sum of squares M^2. The smallest M^2 corresponds to the least important variable because deleting it results in a configuration that is the least different from the reference. Display the resulting M^2 values in a scree-plot or cumulative scree-plot (as explained in 6.3.2).

The variance inflation factors of each variable (see Section 4.3.5.5) are also important and we need to ensure that these are not too high for the selected variables. It is easy to see how this could be applied to a forward selection method, because we could ensure that both the variable corresponding to the minimum M^2 at each stage in the process and also the variables already selected, have variance inflation factors less than some specified value. It is more difficult to see how this could be implemented for a backward elimination method. As discussed in Chapter Seven for biplots, we can also apply stepwise and all subsets methods of variable selection.

9.4.2.1 Application to Hunting Spiders

Using the original 26 variables (and two dimensions for procrustes analysis) we apply the method of variable selection just described to the hunting spiders (1.2.8). The corresponding scree-plot is illustrated in Figure 9.8, where we can think of the vertical axis as a goodness of fit measure with the bottom being the best fit and the top the worst fit.

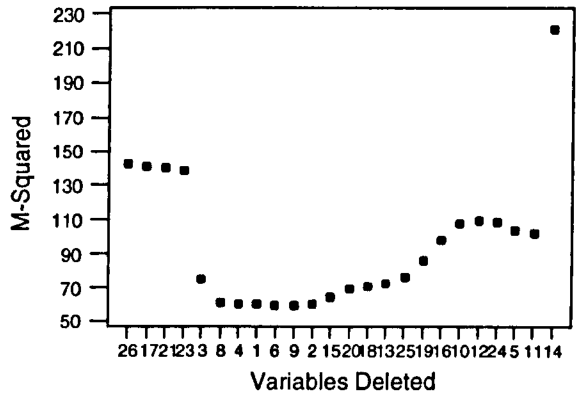

Figure 9.8 Scree-plot for the Hunting Spider Variables (Backward Elimination)

We can see from the plot that the fit improves as each of the first seven variables is deleted and we believe that this is because of the multicollinearity which is often present with large numbers of variables (and which was revealed in 4.5.4). After variable 2 is removed the fit worsens (the slope of the plot rises). We expect this to happen at some point in the selection process because we cannot delete variables indefinitely and expect the fit of those remaining to improve. We believe that the slight decline on the far right of the plot could be because only two dimensions have been used in the calculations. Removing the first 11 variables as suggested by the graph and implementing CCA leads to Figure 9.9 below. Only six variables have variance inflation factors of less than 20 and we note that only one of the 15 variables is the same as that selected from Table 9.2 using the method described in ter Braak & Verdonschot (1995). However, we see from Figure 9.9 that the pattern of species and site points is similar to that of the previous figures, with species *Pardosa lugubris* still occupying an aberrant position in the diagram.

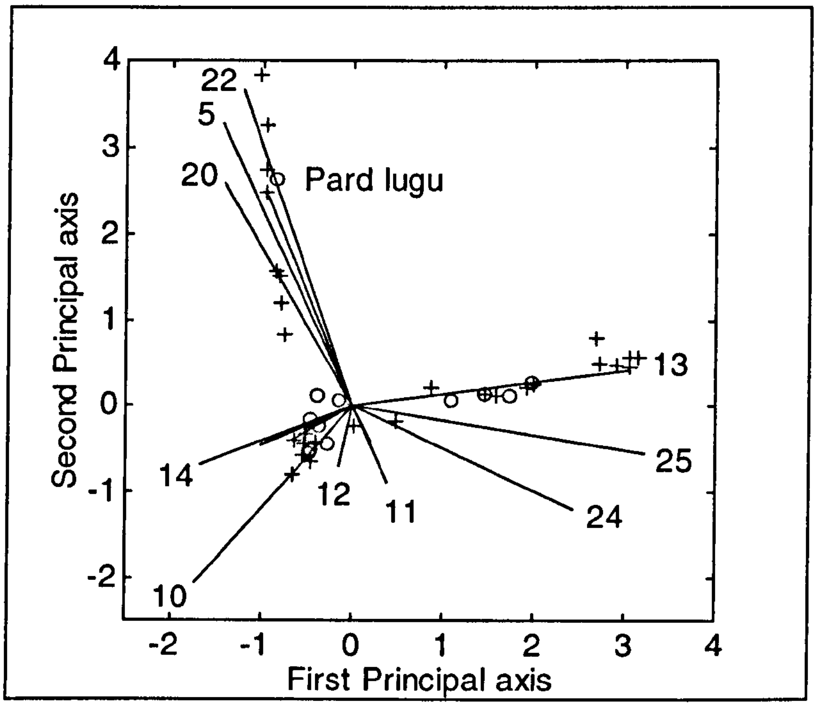

Figure 9.9 Canonical Correspondence Analysis Map of Selected Hunting Spider Variables (Procrustes Analysis)

If we sum the values of M^2 across the steps then we produce a cumulative scree-plot. The advantage of such a plot is that the vertical axis measures the 'total discrepancy' between the appropriate co-ordinates of the original data and those of the reduced data when successive variables have been deleted (rather than the marginal discrepancy resulting from the deletion of each variable). This is illustrated in Figure 9.10 (see 6.3.2), where we see that there is no clear change of slope. It is, therefore, difficult to decide where to stop deleting variables and in this situation the scree-plot appears to be more helpful.

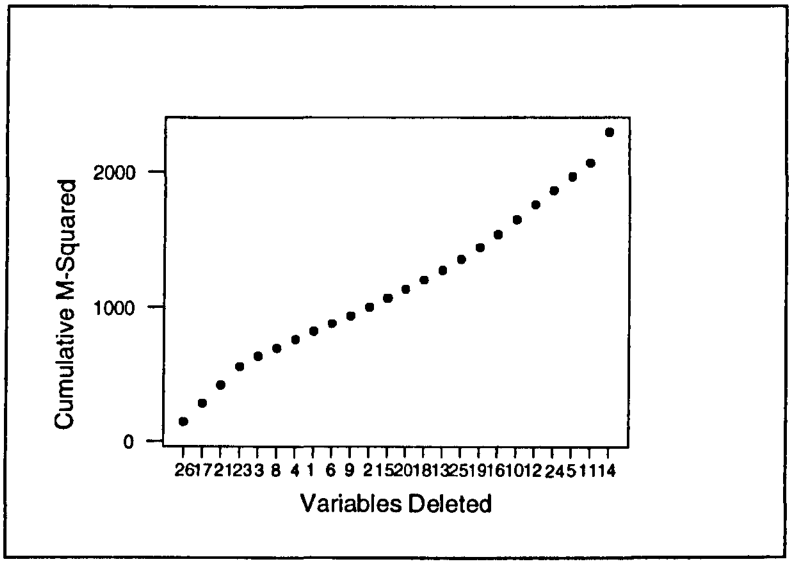

Figure 9.10 Cumulative Scree-plot for the Hunting Spider Variables (Backward Elimination)

9.5 Deleting Categories

Sometimes the data consist of large numbers of sites and/or species and it can be difficult to see any pattern in the resulting CCA map. It may, therefore, be of interest to remove some sites from the analysis. In Section 6.3 we described a method of deleting categories for use with correspondence analysis and we propose adapting this method for CCA:

Stage 1: Implement CCA on the original data matrix and retain the co-ordinates of the species, followed by the co-ordinates of the environmental variables, in the reference configuration X.

Stage 2: Omit each site in turn (from both the species-by-sites and the sites-by-environmental variables matrices) and retain the species co-ordinates of the new data set, followed by the environmental variable co-ordinates, in matrix Y.

Stage 3: Apply procrustes analysis to minimise trace$\{(X-Y)(X-Y)^T\}$ under translation, rotation and reflection of Y. This results in a residual sum of squares M^2. The smallest M^2 corresponds to the least important site because deleting it results in a configuration that is the least different from the reference.

Rather than applying a backward elimination algorithm, we can obtain M^2 for each possible combination of deleted sites (i.e. consider 'all subsets' of sites). We also suggest, for the backward elimination method, using a scree-plot or cumulative scree-plot to display the site deleted at each step against the corresponding M^2 (or cumulative M^2) value — we stop deleting sites where there is a large change in slope of the graph. We implement the method on both the transformed and untransformed hunting spider data in the next section.

9.5.1 Application to Hunting Spiders

Considering the untransformed spider data (see 4.5.1) and deleting each site in turn produces the following non-monotonic scree-plot (only the first 17 steps are shown). We see that as each of the first 12 sites is deleted, the fit improves. This could be because of the large number of sites — some sites are masking the effects of other sites. After site 16 has been removed the fit worsens and this is the point at which we stop deleting sites. However, on the far right of the plot we see that the slope again becomes negative. This could be because only two dimensions have been used in the calculations, or it could be that this method is not appropriate. Removing the first 12 sites in the plot and implementing CCA leads to a very similar ordination diagram to Figure 4.1 in terms of the locations of species points and environmental variables.

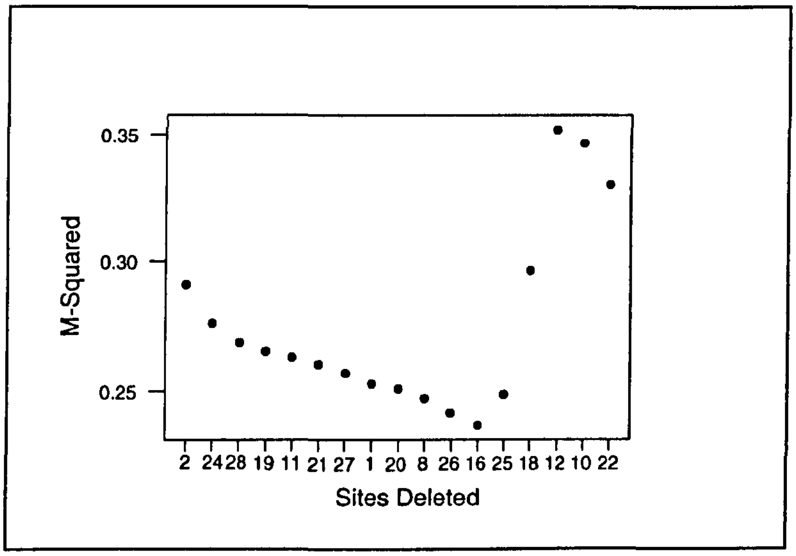

Figure 9.11 Scree-plot for the Hunting Spider Sites (Backward Elimination)

A cumulative scree-plot again shows no clear change of slope. Considering the transformed spider data (see 4.5.2) and deleting sites produces a non-monotonic scree-plot, which initially has a negative slope (i.e. as sites are deleted the fit improves), reaches a minimum M^2 and then undulates. We believe that this is also likely to be because only two dimensions have been used in the procrustes calculations.

9.6 The Influence of Sample Size

Having considered two methods of environmental variable selection, one method of category deletion and introduced bootstrapping as a means of investigating the stability of the CCA map, we now consider the influence of sample size (i.e. the species' abundances) on the ordination diagram. We investigate how the abundances of species measured at each site affect both the pattern of species and sites and the positions of the environmental variables. This is of interest because if a 'small' sample leads to widely differing inferences to those from a 'large' sample then we cannot be confident in our interpretation of the data.

9.6.1 Application to Hunting Spiders: The Multinomial Distribution

In this section we sample the species-by-sites hunting spiders matrix (1.2.8, Table A.12) by fitting a multinomial distribution, leaving the sites-by-environmental variables matrix unchanged. We generate 200 bootstraps with sample sizes consisting of varying proportions of the original spider counts obtained from each site, as indicated by the column headings of the table and calculate the ellipse area (A_4) for each of them. Reading from left to right across the table, we see that the smaller the sample size the larger the area of the ellipse and so the less stable the site i.e. the less confident we are that the sample obtained at the site is representative of the true population of data. The table also shows that site 26 is particularly unstable — referring to the raw data (Table A.12), this is probably due to the very large abundance of *Arctosa perita* at this site, compared with absence or small abundance at the other sites, but could also be due to the relatively unusual values of the environmental variables.

Table 9.4 The Measure A_4 for the Hunting Spider Sites

	Sample Size (proportion of original)						Sample Size (proportion of original)				
Site	**2**	**1**	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	**Site**	**2**	**1**	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
1	0.071	0.134	0.167	0.261	0.523	**15**	0.134	0.241	0.354	0.529	1.031
2	0.021	0.037	0.058	0.084	0.202	**16**	0.103	0.198	0.261	0.412	0.929
3	0.048	0.114	0.121	0.205	0.366	**17**	0.086	0.164	0.238	0.332	0.741
4	0.016	0.039	0.052	0.077	0.193	**18**	0.141	0.245	0.335	0.551	1.039
5	0.017	0.032	0.036	0.067	0.217	**19**	0.216	0.412	0.598	0.835	1.909
6	0.032	0.069	0.084	0.125	0.305	**20**	0.228	0.369	0.615	0.893	1.852
7	0.015	0.037	0.044	0.073	0.233	**21**	0.411	0.819	0.866	1.515	3.024
8	0.186	0.422	0.571	0.853	2.244	**22**	0.365	0.959	1.178	2.038	4.272
9	0.395	0.743	0.842	1.329	3.131	**23**	0.443	0.836	1.243	2.147	4.433
10	0.335	0.644	0.735	1.162	2.812	**24**	0.255	0.544	0.722	1.139	2.735
11	0.209	0.415	0.535	0.824	1.887	**25**	0.315	0.609	0.897	1.051	2.540
12	0.184	0.351	0.446	0.672	1.552	**26**	2.176	4.501	6.121	9.400	29.81
13	0.021	0.043	0.055	0.088	0.232	**27**	0.207	0.410	0.690	0.958	2.772
14	0.028	0.052	0.079	0.093	0.232	**28**	0.278	0.558	0.814	1.187	3.065

9.6.2 Minimum Sample Sizes

Because, for any particular data set, we have only a sample of all possible data, it is of interest to ascertain the minimum sample size required to estimate the proportion of species at a particular site to a certain level of accuracy, with a required probability. We use the notation of Chapter Five and consider the multinomial distribution.

9.6.2.1 Application to Hunting Spiders

By following the methodology of Section 5.6.2 and taking various values of the tolerance (d) and significance level (α), we estimate minimum sample sizes for the untransformed hunting spiders. We use equation (5.8) and list the values in Table 9.5. The estimated numbers will be the same for each site. In order to try to estimate the largest minimum sample size, we take $P_1 = P_2 = 0.49$ and $p_i = \frac{0.02}{(A-2)}$ for i = 3,..., A and A=12 species. Clearly, the values in the table increase as α decreases.

Table 9.5 Estimated Minimum Required Sample Sizes for the Hunting Spider Sites

	Tolerance (d)	
α	**0.1**	**0.05**
0.2	67	267
0.1	96	383
0.05	126	502

Comparing these estimated sample sizes with the actual abundances in Table A.12 of the Appendix, we see that for all combinations of d and α there are 16 sites (out of 28) that have total abundances less than the values in the table. For d=0.05 together with α=0.05 or α=0.1, the recommended sample sizes exceed those collected at all sites. If any data transformations are necessary then we believe that these should be carried out after the estimated minimum sizes above have been obtained.

9.7 Assessing Stability by using a Jack-knife Approach

In Sections 9.2 and 9.3 we discussed how the stability of the sites and the environmental variables could be assessed. An alternative method of assessing stability is to use a jack-knife approach, which we introduce here and which we discussed for correspondence analysis and biplots in Chapters Five and Eight respectively. We propose that the method works as follows. Each species or environmental variable is deleted in turn and a new CCA is implemented each time. The resulting clouds of sites points (one for each deleted species or variable) measure stability and can be compared with those obtained from bootstrapping.

9.7.1 Application to Dune Meadow Vegetation

In this section we apply the jack-knife method introduced above to the dune meadow vegetation (1.2.9). There appear to be two possible approaches to assessing site stability which incorporate the technique of jack-knifing and these arise because the data consist of both species-by-sites and sites-by-environmental variables matrices. We can either delete each species in turn and carry out CCA, before displaying the resulting site points or, alternatively, we can remove each environmental variable in turn before implementing CCA and again display the resulting site points. Removing each species results in Figure 9.12, which only displays three sites because otherwise the plot becomes overcrowded: site 17 is represented by circles (o), site 4 by asterisks (*) and site 19 by plusses (+). The main feature of this figure is that, for these particular sites, the points are not clustered together. Drawing 95% concentration ellipses for the points obtained from jack-knifing for the same subset of sites as in Figure 9.4, leads to Figure 9.13, where we see that there is an extremely high degree of overlap between sites — only four sites can be labelled without ambiguity. In this figure the sites appear to be more unstable and more similar in terms of the profiles of species that they contain, than was revealed by multinomial sampling in Figure 9.4.

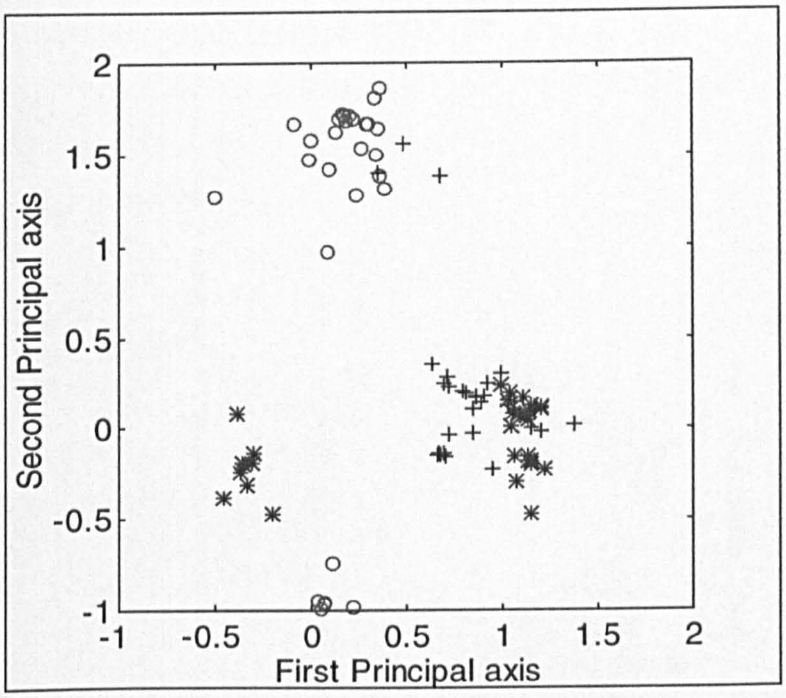

Figure 9.12 Dune Meadow Site Clouds (Jack-knifing Species)

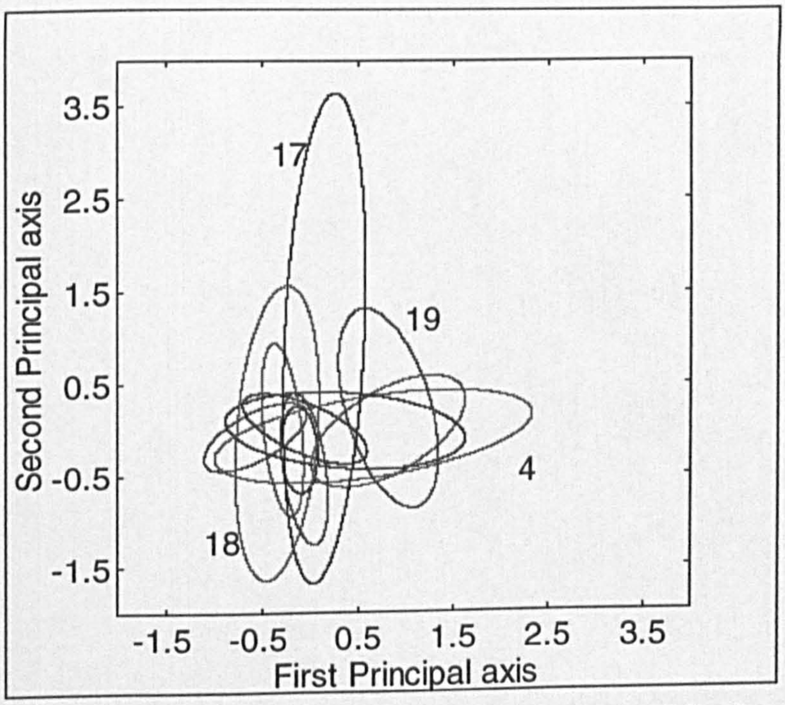

Figure 9.13 95% Concentration Ellipses for the Dune Meadow Sites

Removing each of the five environmental variables in turn and implementing CCA leads to Figure 9.14, where we again display a subset of the sites. It is clear that the clouds of points for each site are much smaller when environmental variables are deleted compared to when species are deleted. We therefore conclude that jack-knifing is not a good method of assessing the stability of the sites because we have conflicting information. It could, perhaps, also be argued that the points from deleting both species and environmental variables should be combined and that the resulting clouds should be taken as an indication of stability.

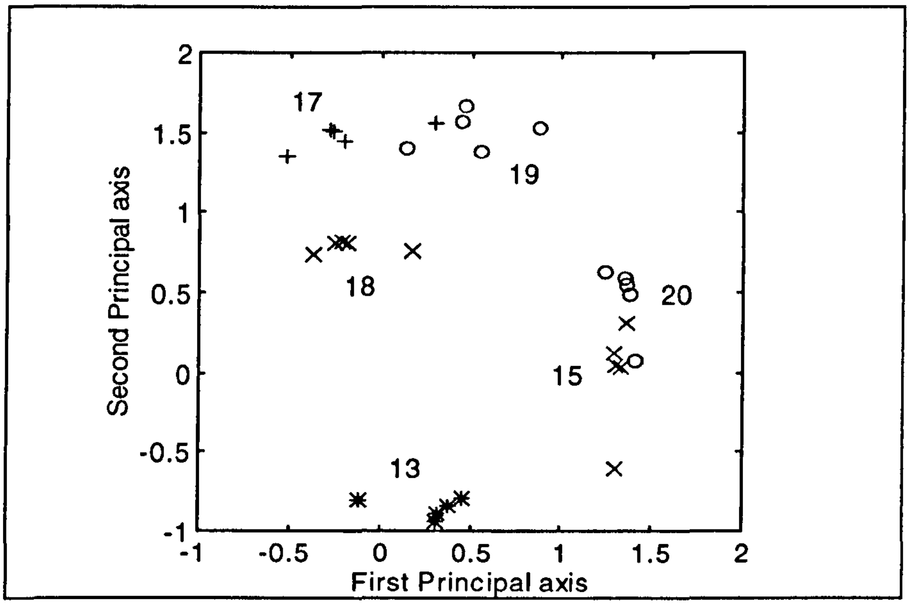

Figure 9.14 Dune Meadow Site Clouds (Jack-knifing Environmental Variables)

9.8 Detecting Influential Categories by using a Jack-knife Approach

In this section we propose using the technique of jack-knifing to identify species or environmental variables that are potentially influential in CCA (i.e. that cause a substantial change in the ordination diagram. This is apparent if a site point is located some distance away from the remainder of the points). The methodology is the same as that of Section 9.7, but the interpretation of the display is different. We also suggest how to ascertain which is the 'most influential' species or environmental variable.

9.8.1 Application to Dune Meadow Vegetation

Considering Figure 9.12 and site 19 (plusses), we see that there are three points that are located away from the majority (i.e. at the top of the diagram) and these are caused by the deletion of the species *Achi mill, Agro stol* and *Aira prae.* For site 17 (circles) there are 7 points away from the rest (at the bottom of the diagram), but these are not caused by the deletion of the three species listed above. For site 4 (asterisks) there are 9 points on the left and 21 on the right and so there do not seem to be any clear species that are influential. Similar patterns emerge for the other sites.

Considering Figure 9.14 we see that there is one point that is removed from the remainder for all sites except site 19 and in some cases there are two points located away from the rest (e.g. for site 17). The single point relates, in each case, to deletion of the variable moisture content and so this variable is potentially influential. Deleting this variable and implementing CCA leads to a different ordination diagram to that of Figure 4.7 (which used all five variables) — the nominal variables are clustered at the centre. When there are two points located some distance away from the rest, these are caused by deletion of the variables moisture and grassland management. Removing both these variables leads to Figure 9.15, where the quantitative variables are located in similar directions to in Figure 4.7, but where the distribution of the species points across the ordination map is quite different. There now appears to be a cluster of three species at the top of the diagram.

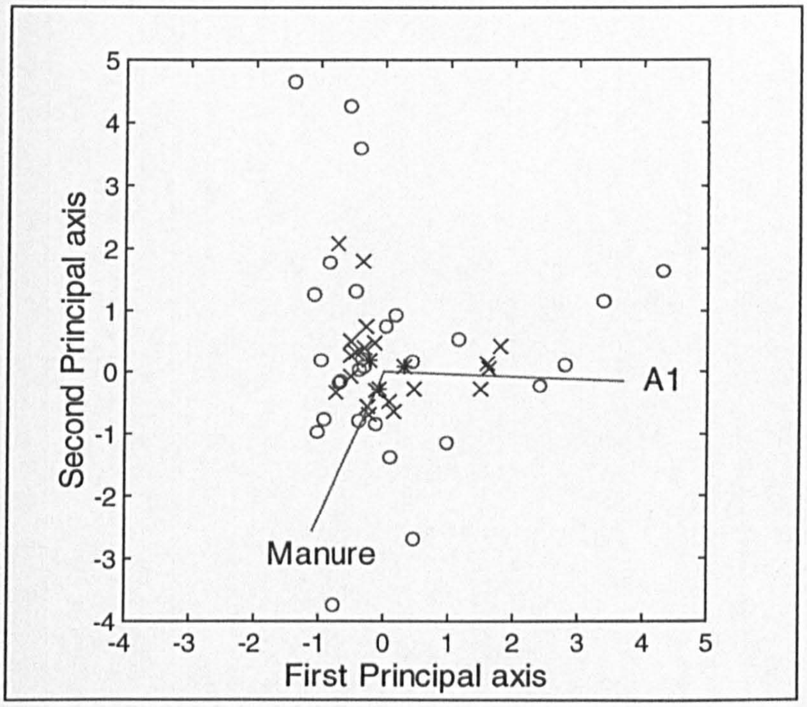

Figure 9.15 Canonical Correspondence Analysis Map of Dune Meadow Vegetation (Moisture and Grassland Management Deleted)

In order to identify the 'most influential' environmental variable, we propose using the backward elimination selection method, which we introduced in 9.4.2. However, in the first step, instead of looking for the deleted variable, which results in the smallest M^2, we look for the variable with the largest M^2. This variable is considered to be the most influential because it results in the biggest difference in co-ordinates from those of the reference configuration. Similarly, when seeking the most influential site, we propose using the method of Section 9.5 to identify, at the first step, which site deletion results in the largest M^2. To detect the most influential species we can apply the method in 9.5, but delete species instead of sites.

9.9 Connections with Other Techniques — Practical Application

In Section 4.4 we explained that in canonical correspondence analysis, each pair of {sites, species, environmental variables} form a biplot. Given that correspondence analysis is appropriate for species-by-sites matrices, that biplots are appropriate for sites-by-environmental variables matrices and that both biplots and correspondence analysis form major parts of this thesis, we believe that it is important to consider similarities between these three techniques as regards practical application. We concentrate on the hunting spider data in the following two sections.

9.9.1 Canonical Correspondence Analysis and Correspondence Analysis

In this section we implement correspondence analysis (CA) on the untransformed species-by-sites matrix for the hunting spiders (1.2.8) and compare the locations of the species (circles) and site points (plusses), plotted in Figure 9.16, with those obtained from CCA in Figure 4.1.

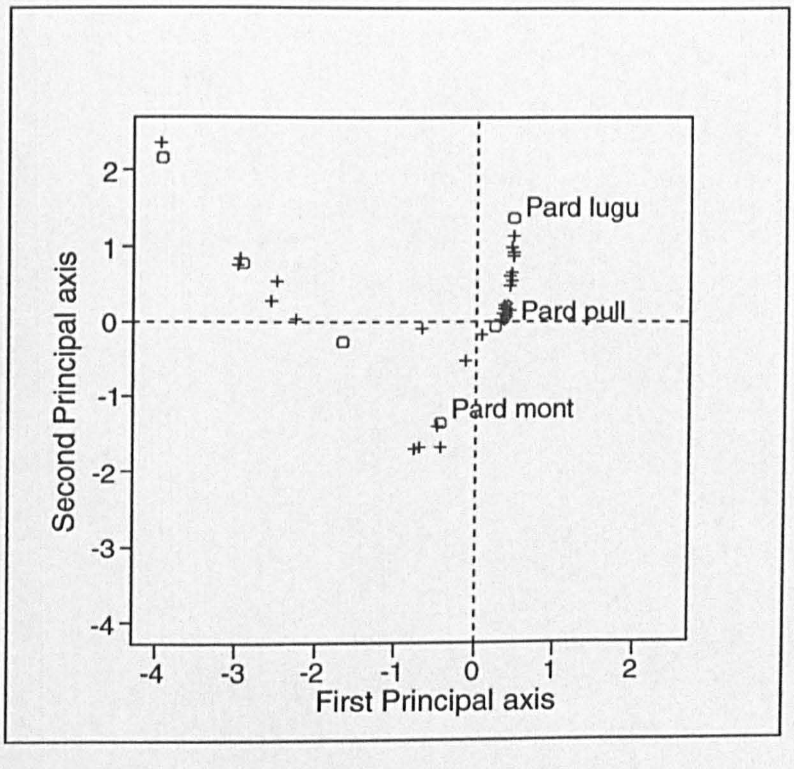

Figure 9.16 Correspondence Analysis Map of Hunting Spiders

Considering Figure 4.1, we see that there is one site located on the right of the diagram that is removed from the remainder of the sites and species points — site 26. This is the same site that is located on the top left of the above figure (the CA and CCA ordination diagrams are essentially reflections of each other), but this time it is located close to the species *Arctosa perita.* Looking at Table A.12 of the Appendix, it is clear that *Arctosa perita* is very abundant at site 26. It is also evident that both ordination diagrams show an 'arch effect' of both species and site points. In Section 4.7.1 we noted that when the number of environmental variables is close to the number of sites, then CCA is essentially CA, although in the above example there are only six variables and still the diagrams are very similar. When the data are transformed as discussed in 4.5.2 (but still with six variables), the ordination maps of CA and CCA are again very similar.

9.9.2 Canonical Correspondence Analysis and Biplots

In this section we implement the correlation biplot on the (untransformed) sites-by-environmental variables matrix for the hunting spiders and compare the locations of the variables with those of the variables obtained from a CCA on these data (Figure 4.1). The correlation biplot is illustrated in Figure 9.17.

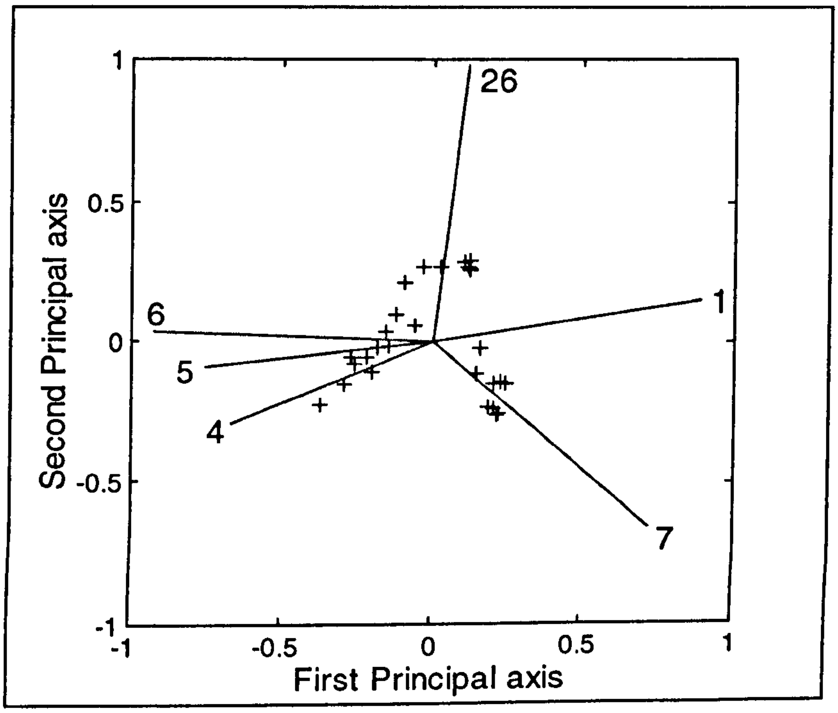

Figure 9.17 Correlation Biplot of Hunting Spiders

We see that the site points in the above diagram form an arch — this was also the case for both CCA and CA. It is clear that the relative positions of the environmental variables are the same in both Figure 4.1 and Figure 9.17; the two Figures are essentially reflections of each other. The correlation biplot and CCA produce similar ordination maps for these data and this is also the case when they are applied to the transformed hunting spider data.

9.10 Summary and Conclusions

This chapter has considered various methods of assessing the stability of the sites in canonical correspondence analysis maps and discussed the appropriateness of these methods, depending on the form of the data. The main method which we proposed involves resampling the species-by-sites matrix by using the multinomial distribution (i.e. bootstrapping): a multinomial distribution should usually be applied to each site separately, because this most closely resembles the method by which species data are obtained in the 'field'; we also suggested that the sites-by-environmental variables matrix should be left unchanged. We commented that particular problems arise when using the multinomial distribution to assess the stability of vegetation data measured on cover-abundance scales (such as those in Table 4.1). This is because the multinomial distribution treats the data as frequencies, whereas these scales tend to be of an ordinal nature, where the distances between units on the scale are not equal. However, we have ignored these problems which, incidentally, do not arise when the data consist of absolute abundances, as is usually the case for animal species data. An alternative method of investigating stability, which we proposed, is a jack-knife approach, although one problem with this method results from the fact that there are two data matrices to consider: species-by-sites and sites-by-environmental variables. It is not obvious how these two sets of information can be most effectively combined.

We concluded that as for correspondence analysis, (non-parametric) convex hulls and (parametric) concentration ellipses provide useful summaries of the clouds of points obtained from bootstrapping. By looking for overlapping 95% concentration ellipses, we can establish which sites are similar in terms of the types and distributions of species they contain. As we suggested in Chapter Five for correspondence analysis, a rough rule of thumb is that if the centroid of a 95% ellipse representing a site is included in the ellipse of another site, then these sites can be considered to be virtually indistinct in terms of their profiles of species. We also proposed using the area of hull peels and 95% ellipses to assess the relative stability across sites (larger areas mean greater instability). In correspondence analysis, the sites with the largest bootstrap clouds have similar numbers across the species and low numbers in each cell of the original data matrix, but this is not true for CCA and we suspect that this is because of the influence of the environmental variables.

The issues involved in assessing the stability of environmental variables were also discussed in some detail: problems arise partly because of zeroes in the environmental data and partly when there are nominal variables. It is, therefore, difficult to decide which distributions are most appropriate to fit to the data, although when this has been decided we can obtain confidence intervals for the true directions of the quantitative variables and confidence regions for the true locations of the nominal variables (i.e. in the whole population of data).

It is sometimes the case that many more variables are measured at a site than can be effectively displayed in the ordination diagram and it is not always known which variables are likely to be most effective in explaining the distributions of species (this is why we apply CCA). It is also known that multicollinearity often exists when large numbers of variables are measured and so we have considered variable selection methods in this chapter. In particular, an existing method of forward selection was implemented and the results compared with a new method that we introduced and which is based on the procrustes statistic. By displaying the results from our method in a scree-plot (and cumulative scree-plot) and looking for changes in slope of the graph, the selection process can be visualised. Both methods selected variables with high variance inflation factors, but both ordination maps are little changed from that of the original data. We could also have implemented stepwise and all subsets methods of variable selection and we could have considered using higher dimensionality in the calculations.

Sometimes there are too many categories (species or sites) to effectively display in the ordination diagram and so we proposed a method to reduce this number of categories. The method is an adaption of that introduced by Krzanowski (1993) into correspondence analysis and is based on the procrustes statistic — this was reasonably successful. However, we only used two dimensions in our calculations and it may be worth considering higher dimensions.

In addition, we used the multinomial distribution to calculate the minimum required sample sizes in order to estimate two or more categories of species simultaneously (by applying traditional sampling theory). It is clear that the actual sample sizes collected by ecologists at a particular site are often less than those required based on statistical criteria and are sometimes as little as $\frac{1}{30}$th of the size. In contrast, archaeologists often 'overcollect' artefacts (see 5.6.2), relative to the recommendations based on statistical calculations.

Because canonical correspondence analysis is applied to two data matrices, one of which is suitable for correspondence analysis and one of which is suitable for biplots, we compared the CCA ordination map with that obtained from a correspondence analysis implemented on the species-by-sites data and also with a correlation biplot of the sites-by-environmental variables data. For the data we considered, the three methods produce very similar ordination diagrams and interpretations. Finally, we introduced jack-knifing as a means of detecting potentially influential sites, species and variables in the CCA map and we found it to be a very useful technique. We suggested that the 'most influential' category or environmental variable can be identified by looking for the largest procrustes M^2 at the first step of the appropriate backward elimination selection method.

Chapter Ten

Summary and Conclusions

10.1 Introduction

This thesis has been concerned with three techniques of exploratory multivariate analysis — correspondence analysis (CA), biplots and canonical correspondence analysis (CCA) — and their application to 'field studies' — in particular archaeology and ecology. The main focus of Chapter One was to introduce the data sets (both published and new material), which we have returned to throughout to illustrate both existing and new methodology and which are listed in the Appendix. Chapters Two, Three and Four explained in detail the mathematical theory behind the three techniques, whilst also raising important questions driven by the data. These questions were the focus of Chapters Five to Nine. Similar issues were addressed using all three techniques and the purpose of this chapter is to summarise these, but also to explain the methods used to address them and the conclusions drawn. The remainder of this section explains the similarities, whereas Sections 10.2, 10.3 and 10.4 discuss issues specific to each of the techniques.

Correspondence analysis, biplots and canonical correspondence analysis are all informal, graphical, exploratory methods for displaying high-dimensional data in low-dimensional space. They all involve the singular value decomposition of a matrix. Correspondence analysis displays the rows and columns of a matrix of non-negative data as points in an ordination diagram and looks for patterns or associations in the data. In contrast, biplots are usually used to display data consisting of a series of variables measured on a number of observations, where the observations are represented by points and the variables by vectors. Canonical correspondence analysis is appropriate for data consisting of the abundances of a multitude of species at a number of sites, where environmental variables have also been measured at these sites and in some respects is a combination of the first two methods.

The questions that we have addressed in this thesis and the methodology that we have developed have, in part, been driven by the needs of the archaeologists who provided some of these data sets and who have been consulted extensively during the study. The main issues that we have investigated using the three techniques have followed a common theme and these are summarised as follows:

[1] **Stability**: Investigation of the stability of the data (sites, contexts and variables) i.e. how representative are they of the true population of data, within the framework of the multivariate technique that is used for analysis. This is of interest because our data sets are only samples of all the possible data that could be collected, if resources were unlimited.

[2] **Sample Size**: Assessment of the influence of sample size (number of artefacts collected or measured; abundance of species) on the ordination diagram. Quantities such as sample size, the number of categories of classification (e.g. of sites, wares, contexts) and the number of variables measured compete for fixed resources in archaeological and ecological applications and so if sample size can be reduced it may be that resources could be expended on other aspects of the study e.g. the number of variables measured could be increased or more sites could be visited. We investigated the influence of sample size for the Memphis (1.2.1) and Amarna pottery sherds (1.2.2), ceramic pots (1.2.5) and hunting spider data (1.2.8).

[3] **Selection Methods**: The development and implementation of both existing and new category selection and variable selection methods. Some categories and variables were shown to be redundant in the sense that the same patterns are revealed in the data if these are removed from the analysis. The resources that are spent collecting this 'excess' information could therefore be channelled into other parts of the study. Selection methods were successfully applied to the Memphis and Amarna wares, Amarna sites, Melanesian starch grains (1.2.3), Early Stone Age tools (1.2.4), ceramic pot variables, Simpson Desert flint tool variables (1.2.6) and hunting spider sites and variables. However, selection is not always appropriate and depends on the type of data and sampling scheme used to collect the data. For example, with the Memphis contexts, the scheme was not to collect a certain amount of pottery and then to cross-classify it into context and ware, but to collect all pottery within each context. Therefore, reducing the number of contexts before analysis is not sensible. Our analyses also suggested that even though the quantity of material recovered from each context is enormous, archaeological expertise is essential for its classification i.e. there is little scope for classification into broader categories by a less skilled person. Similarly with the hunting spider data, the aim is to identify which species characterise which sites and reducing the number of species would be inappropriate. The Amarna excavations are ongoing and so it is possible to alter the sampling strategy based on the results of our analyses i.e. to reduce the number of sherds collected at each 'site' and perhaps increase the number of 'sites' visited. We also recommend taking fewer measurements on each of the Simpson Desert flints — in particular, weight is expensive to measure and it may be worth omitting this variable.

[4] **Influential Categories, Variables and Observations**: The detection of influential categories (sites, wares, contexts, species, artefacts), variables and observations in terms of their effects on the ordination diagram. The influential categories identified by using statistical methods can be combined with the expertise of the archaeologist and conclusions can be formed based on this. Influential categories, variables and observations were investigated for the dune meadow vegetation (1.2.9), Simpson Desert flint tools and Amarna pottery sherds.

The methods used to address the issues described above were as follows:

[1] **Stability**: Bootstrapping was used to generate replicate matrices in order to assess the stability of the data (the multinomial distribution was used for both correspondence analysis and CCA and the multivariate normal distribution was used for biplots). Category stability for CA and CCA was summarised using both convex hulls and 68% and 95% concentration ellipses. Areas of hull peels were used to assess relative stability across categories (larger areas indicate greater instability) and to obtain comparable areas between clouds resulting from differing numbers of bootstraps. We also suggested using areas of concentration ellipses to measure stability and in particular, we suggested that if the centroid of a 95% ellipse representing one category is included in the 95% ellipse of another category, then the categories can be considered to be virtually indistinct. In correspondence analysis, the sites with the largest bootstrap clouds have similar numbers across the species and low numbers in each cell of the original data matrix, but we discovered that this is not true for CCA and we suspect that this is because of the influence of the environmental variables. The stability of the biplot variables was assessed by introducing the idea of 'bootstrap fans' and using the standard confidence interval and the BC_a method. We also proposed an adaption of the interval in Chapter Five of Fisher (1993), which was developed specifically for directional data. Jack-knifing was also used to assess both category and variable stability, although for CCA there are two data matrices to consider (species-by-sites and sites-by-environmental variables). Jack-knifing led to smaller clouds for CA and narrower 'fans' for biplots than bootstrapping, but for CCA cloud size depended on whether species or environmental variables were omitted. We concluded that a jack-knife approach provides a good benchmark against which other methods of assessing stability can be compared.

[2] **Sample Size**: For the abundance data of both CA and CCA we investigated how the actual numbers of artefacts collected by archaeologists and ecologists compare with recommendations based on statistical calculations, obtained by using traditional sampling theory. It is clear that the actual sample sizes collected by archaeologists tend to exceed those required based on statistical

criteria, sometimes by as much as 600% for any particular site (e.g. for the Amarna sherds), whereas the samples collected by ecologists at a particular site are often less than those required and are sometimes as little as $\frac{1}{16}$th of the size (e.g. for the hunting spider data). For CA, sampling from the multinomial distribution was compared with sampling without replacement and it is clear that the smaller the sample size the less stable the corresponding category, but also that sampling using the multinomial distribution produces greater instability than sampling without replacement. For biplots, sampling from the multivariate normal distribution was compared with sampling without replacement and we saw that the former method produces greater instability than the latter.

[3] **Selection Methods**: We used procrustes analysis with all three multivariate techniques to select categories (for CA and CCA) and variables (for CCA and biplots), focusing on the backward elimination method for CA and CCA, but also considering forward selection, all subsets and stepwise methods for biplots. It was clear that the higher the dimensionality used in the procrustes calculations, the fewer categories or variables that can be deleted. We introduced the use of a scree-plot and cumulative scree-plot in order to help identify which categories or environmental variables to delete and also to choose the correct scale. These are alternatives to critical values for CA and biplots and to a Monte Carlo permutation test for CCA. We believe that the aim of category reduction methods is to identify several subsets of categories rather than one unique set — the all subsets approach is closest to this ideal. For CA we introduced terminology for distinguishing between combining categories based on archaeological grounds as compared with on statistical grounds. It was clear that for some data sets (e.g. Early Stone Age tools and Memphis sherds) the expertise of an archaeologist is required before any amalgamation is undertaken (and that some categories should never be combined).

[4] **Influential Categories, Variables and Observations**: Jack-knifing was used to detect influential categories for CA and CCA, to detect influential variables for CCA and to detect influential observations for biplots. Procrustes analysis was used in combination with jack-knifing in order to detect the 'most influential' categories, variables and observations e.g. Amarna sites and wares, Simpson Desert flint tools and dune meadow vegetation.

As a result of our analyses of the Memphis sherds, Amarna sherds, Melanesian starch grains, ceramic pots, Simpson Desert flint tools and flake debitage data, we suggested that as a rule of thumb at least 50% of the variation in archaeological data should be explained in the first two dimensions of the ordination diagram for CA and biplots. Krzanowski (1993) suggested choosing the dimensionality for variable selection according to 80% of the variation explained, but we believe that this is too stringent.

The above methodology was appropriate for all three techniques. However, some issues were specific to a particular technique and these are discussed in the following three sections.

10.2 Correspondence Analysis

For correspondence analysis, we addressed the following questions:

- **Trace Zeroes**: Sometimes, large numbers of trace zero cells occur in archaeological data (i.e. the sampling technique is not adequate to detect rare artefacts). This can be a problem when generating replicate matrices based on the multinomial distribution, because each zero cell is allocated zero probability. We therefore developed two methods based on the binomial distribution to adjust the probabilities assigned to these cells. However, the sizes of the bootstrap clouds appear unchanged by these methods (e.g. for the Memphis sherds and Amarna sherds) unless the sample size is very small and this is because the probabilities assigned to the zero cells are also very small. We have therefore concluded that it is not worth accounting for trace zeroes in the data when assessing for stability. This has implications when deciding on an appropriate sampling scheme for data collection — the number of categories and sample size could be adjusted depending on the anticipation of trace zeroes.

- **Trace Zeroes and Selection Methods**: We also revealed that large numbers of zeroes in the data affect category deletion and clustering methods (e.g. Early Stone Age tools), partly because correspondence analysis requires non-zero row and column totals and partly because of the influence of zeroes on clustering methods (but not on the opinion of the archaeologist). In addition, we proposed using correspondence analysis to assess the effect of category division (which is based on external variables) and illustrated this using the Melanesian starch grains. We reiterated that sometimes no selection method is appropriate because the given categories are essential in testing a particular hypothesis of the investigator e.g. the Memphis contexts.

- **Combining Categories**: We proposed using clustering methods in archaeology in order to assess the effects of misidentifying (the Memphis) contexts when the stratigraphic method of excavation is used — the results showed that there are no serious consequences in terms of inferences based on the correspondence analysis map, if two neighbouring contexts are misidentified. We calculated the stability of and the influence of sample size on, combined categories and compared the results with those obtained from the original categories. It appears that when the data consist of smaller numbers of categories, there is little difference in the stability of these categories e.g. the Amarna sites.

10.3 Biplots

The following issues apply specifically to the various forms of biplot.

- **Diversity Biplot**: We introduced the diversity biplot into archaeology and this is clearly more useful than the numerous diversity indices in existence, mainly because it provides a graphical display of diversity. However, comparing the diversity biplot with CA for the bone engraving data (1.2.7) indicates that relationships between categories do seem to be more clearly displayed in the correspondence analysis map than in the diversity biplot.

- **Projection**: We developed methods for projecting supplementary observations and variables onto the original biplot axes.

- **Replicate Matrices**: We discovered that replicate matrices cannot be directly projected onto the original co-ordinate system. Instead, co-ordinates for these matrices must be obtained by implementing a biplot on each matrix separately. We illustrated this for the Simpson Desert flint tools and ceramic pots.

- **Dimensionality**: It was clear that a biplot in two dimensions is not always appropriate, because two dimensions can be insufficient to both explain a high proportion of the variation in the data and also for each individual variable to have a high quality of representation. This was the case for the Simpson Desert flint tools. It was also evident that variables that are poorly represented in the chosen dimensionality (typically two) have particularly wide confidence intervals for their true directions.

- **Mean Directions**: We proposed accounting for the length of the bootstrap vectors when calculating mean directions for the correlation and Spearman rank correlation biplots, because for these biplots vector lengths represent the quality of representation of the corresponding variable.

- **Sample Size and Variable Selection**: We also combined varying sample sizes with the variables selected from the backward elimination method in order to investigate how this affects the stability of the variables. It is clear that the fewer variables that are used in the analysis, the wider the confidence intervals for their true directions e.g. Simpson Desert flint tools.

- **Biplots and Correspondence Analysis**: Although biplots and correspondence analysis are distinct techniques, there might be some occasions where, for example, continuous variables could be categorised and correspondence analysis used instead of biplots, although we did not investigate this.

10.4 Canonical Correspondence Analysis

Canonical correspondence analysis is relatively underdeveloped in the statistical literature and there is considerable scope for further work in this area. We discussed the following points:

- **Data Transformations**: We investigated the effect of the form of the data (raw, transformed, or presence/absence) on the results of the analysis for the hunting spider data and concluded that it is not advisable to implement CCA on presence/absence data, although this was the only form of data for which there was no evidence of an 'arch effect'.

- **Cover-Abundance Scales**: We commented that particular problems arise when using the multinomial distribution to assess the stability of vegetation data measured on cover-abundance scales e.g. the dune meadow vegetation. This is because the multinomial distribution treats the data as frequencies, whereas these scales tend to be of an ordinal nature where the distances between units on the scale are not equal.

- **Comparisons between the Techniques**: Because CCA is applied to two data matrices, one of which is suitable for correspondence analysis and one of which is suitable for biplots, we compared the ordination diagram of CCA with that obtained from a correspondence analysis implemented on the species-by-sites data and also with a correlation biplot of the sites-by-environmental variables data for the hunting spiders. All three techniques produce similar ordination maps for all the data sets that we looked at e.g. hunting spiders.

Appendix

Data Sets

Table A.1 Weights of Memphis Pottery Sherds (kg)

		Context												
Code	Ware	377	465	509	476	289	690	716	739	740	707	761	758	749
1	A01.01	0.01	0.10	0.45	0.01	0.32	0.01	0.20	0.04	0.88	3.00	0.11	1.11	0.32
2	D01.01	0.01	0.05	0.08	0.01	0.54	0.47	0.08	0.30	0.19	3.71	0.10	0.40	0.72
3	E01.01	0.08	0.33	0.14	0.34	1.72	2.78	2.50	0.24	6.24	17.20	0.30	8.99	6.50
4	G01.01	3.90	3.14	2.98	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	G01.02.00.01	0.18	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	G01.06	0.12	0.07	0.50	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	G01.08	0.56	0.26	0.72	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	G01.29	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	G05.01	0.16	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	H01.05	0.04	0.30	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	NILEB2.01	0.00	0.00	0.00	0.00	0.00	8.01	6.32	0.86	7.76	10.64	1.30	12.50	7.04
12	UNKNOWN	0.00	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	D01.00.00.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
14	D01.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00
15	D01SMOKED	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00
16	E01SMOKED	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00
17	D01.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
18	NILEB2.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.03
19	NILEB2.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.34	0.00	0.00	0.26
20	P40.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
21	NILECHANDMADE	0.00	0.00	1.14	0.40	0.00	1.96	0.00	0.00	3.71	8.00	0.20	0.10	0.00
22	P16.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	G01.02	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	P33.01	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	NILEB2	0.00	0.00	0.00	0.00	3.72	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
26	NILEB2.06	0.00	0.00	0.00	0.00	0.40	0.05	0.25	0.30	1.48	7.00	0.00	0.90	0.70
27	NILEB2.07	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.34	0.00	0.02	0.20
28	NILEB2.08	0.00	0.00	0.00	0.00	1.28	0.81	1.04	0.00	0.87	2.24	0.02	0.94	1.68
29	NILEB2.15	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A.1 (continued)

		Context												
	Ware	**377**	**465**	**509**	**476**	**289**	**690**	**716**	**739**	**740**	**707**	**761**	**758**	**749**
30	**NILEB2SMOKED**	0.00	0.00	0.00	0.00	0.40	0.17	0.38	0.04	0.62	1.90	0.01	0.78	0.30
31	**H02.01**	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
32	**H08.01**	0.08	0.34	0.00	0.01	0.00	0.81	0.01	0.00	0.00	0.00	0.00	0.00	0.06
33	**H10.01**	2.50	0.06	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.04	0.00	0.00	0.00
34	**NILEC.01**	0.22	1.39	2.60	0.16	10.70	4.61	5.50	0.86	2.49	25.66	0.80	13.00	5.20
35	**P31.01**	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.20	0.00	0.01	0.03	0.08
36	**G01.04**	0.00	0.02	0.05	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
37	**G01.15**	0.00	0.01	0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
38	**G01SMOKED**	0.00	0.01	0.26	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
39	**H10.01.00.14**	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
40	**H24.01**	0.00	0.02	0.20	0.10	0.16	0.18	0.07	0.00	0.12	0.82	0.08	0.24	0.30
41	**BREADMOULDS**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.50	0..00	0.00	0.00
42	**D02.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
43	**NILEB2.06.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
44	**H24.01.00.02**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
45	**P34.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00
46	**E01.09**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
47	**D03HANDMADE**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
48	**D04.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00

Source: Janine Bourriau, Macdonald Institute, Cambridge, England.

Table A.2 Counts of Amarna Pottery Sherds

	Site											
Ware	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
1	30	130	12	12	0	0	7	58	69	351	4	11
2	21	100	0	3	4	5	60	49	230	1073	0	13
3	30	100	0	22	0	0	100	129	7	195	0	12
4	30	44	0	10	0	5	143	414	0	14	0	10
5	30	130	0	2	0	0	160	258	0	15	4	0
6	105	144	50	41	65	0	1108	1294	7	97	31	50
7	529	506	25	51	53	17	60	258	29	156	8	15
8	0	145	0	85	454	528	0	0	224	25	0	200
9	106	130	12	17	14	0	143	129	10	25	600	23
10	0	18	861	0	0	0	7	0	0	0	132	0

Source: Paul Nicholson, Cardiff University, Wales.

Table A.3 Counts of Melanesian Starch Grains

	Site														
Type	**1**	**2**	**5**	**9**	**11**	**12**	**15**	**16**	**17**	**19**	**20**	**21**	**24**	**25**	**26**
1	11	12	42	67	93	65	211	23	77	271	94	95	77	89	82
2	17	13	44	14	6	4	47	2	32	14	13	15	5	12	20
4	1	2	3	1	0	0	0	1	2	2	3	2	2	2	1
5	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
6	2	0	0	0	0	2	3	0	1	10	0	0	1	1	1
9	9	8	2	20	10	6	5	20	11	23	14	9	12	13	6
10	10	41	0	0	0	0	9	2	15	2	33	45	0	3	0
11	0	0	0	2	0	1	1	0	2	5	3	2	2	1	1
12	2	1	3	1	0	0	2	0	1	10	2	0	1	3	0
15	5	0	0	0	0	0	0	0	0	7	1	1	2	0	1
16	1	0	0	0	0	3	0	0	0	3	1	2	0	0	0
17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
18	2	0	0	0	0	1	0	4	1	0	0	0	0	1	1
19	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
20	0	0	0	1	8	2	3	6	5	5	3	3	4	5	5
21	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
22	6	14	22	8	1	1	15	0	6	4	6	5	3	7	5
23	11	9	12	11	6	50	7	0	12	9	4	5	0	29	9
26	2	0	0	0	0	1	3	9	3	0	6	1	5	2	2
27	3	1	4	3	2	1	8	63	3	1	2	2	1	1	7
28	0	0	1	1	1	0	2	0	0	2	1	1	0	0	0
30	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
31	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0
32	0	0	0	1	1	0	0	0	1	0	0	4	2	2	0
33	1	0	1	0	1	1	1	0	1	0	1	1	3	2	1
35	0	4	5	5	2	2	3	2	2	0	6	10	13	3	5
36	0	1	0	4	1	1	4	0	3	0	0	2	1	2	3
38	2	0	2	0	1	0	0	0	0	0	0	3	2	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
40	2	2	6	0	0	0	1	0	0	3	0	1	1	0	2
41	0	1	1	0	0	0	3	0	2	0	1	2	1	0	1
44	1	0	2	0	0	0	1	0	1	0	0	1	0	2	0
45	0	0	1	0	0	0	0	0	1	1	0	0	4	0	0
46	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0
47	2	1	1	3	2	1	2	2	0	0	0	1	24	3	4
48	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
49	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
51	3	4	1	0	2	0	4	1	3	2	0	1	3	5	3
53	4	1	1	2	0	0	6	0	1	0	0	0	1	2	0
55	1	1	4	9	4	2	4	0	0	2	3	1	2	2	4

Table A.3 (continued)

	Site														
Type	**1**	**2**	**5**	**9**	**11**	**12**	**15**	**16**	**17**	**19**	**20**	**21**	**24**	**25**	**26**
57	3	0	1	0	0	0	0	0	0	1	0	0	1	2	1
59	1	1	0	0	0	0	2	1	1	0	0	0	1	0	0
60	0	1	2	0	0	0	1	0	0	0	0	0	0	4	0
65	1	0	0	2	0	0	0	0	2	0	1	0	0	0	0
66	0	0	0	1	2	0	1	1	0	0	0	0	0	0	0
67	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
74	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
76	1	0	2	1	0	0	1	0	3	0	0	0	0	0	3
77	0	0	0	1	1	2	5	0	0	1	0	0	0	0	3
79	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4
81	1	1	0	0	0	0	0	0	0	0	0	0	0	0	3
84	0	0	0	0	1	0	0	0	4	0	0	0	0	0	1
86	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
87	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
88	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
89	0	1	2	1	0	0	1	0	1	0	0	0	0	0	0
91	0	0	0	0	1	0	1	0	2	0	0	0	0	0	1
92	1	2	1	5	1	2	1	1	7	1	0	0	0	0	0
97	0	0	2	1	0	0	1	0	2	0	0	0	0	0	0
98	1	0	3	4	1	1	4	10	0	1	0	0	0	0	0
99	2	2	0	2	0	0	0	1	1	0	2	0	0	0	2
112	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
113	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
114	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
116	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
118	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
119	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
122	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
123	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0
128	4	1	6	0	0	0	2	0	0	0	0	0	0	0	0
130	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
131	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
133	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
134	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
135	1	2	2	0	0	0	0	0	1	1	0	1	0	0	0
136	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
137	3	1	0	0	0	0	0	0	0	1	1	0	0	0	5
138	1	4	0	0	0	0	1	0	0	0	0	0	0	0	0

Table A.3 (continued)

	Site														
Type	**1**	**2**	**5**	**9**	**11**	**12**	**15**	**16**	**17**	**19**	**20**	**21**	**24**	**25**	**26**
139	21	5	0	0	0	0	0	0	0	0	0	0	1	1	0
140	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0
141	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1
142	6	4	1	0	0	0	0	0	0	0	4	1	0	1	6
143	1	2	0	0	0	0	0	0	0	0	0	0	0	1	0
144	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0
145	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0
146	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
147	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
148	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
149	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
150	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
152	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
153	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
154	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
156	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Source: Carol Lentfer, Southern Cross University, Australia.

Table A.4 Site Decriptions for the Melanesian Starch Grains

Number	Site	Description
1	Garala 1	Advanced regrowth forest on small rock island
2	Garala 2	Advanced regrowth forest on small rock island
5	Kaula 70-80m Transect	Advanced regrowth forest on small rock island
9	Mt Hamilton H3	Regrowth forest on old garden site, Garua Island
11	Garua FEK	Coconut plantation on strand plain
12	Garua Barge Landing	Coconut plantation on strand plain
15	Garu Garden 3	Old garden with cassava and bananas
16	Garu Garden 5	Old garden dominated by sweet potato
17	Garu Garden 7a	New garden planted with taro. 12 year old regrowth forest cleared for garden site.
19	Garu Garden 7b	New garden planted with taro. 12 year old regrowth forest cleared for garden site.
20	Garu Garden 7c	New garden to be planted with taro. 12 year old regrowth forest cleared for garden site.
21	Garu Garden 8	New garden adjacent to site 7 - cleared in 12 year old regrowth forest
24	Nave River	Heavily logged forest (looks like it was advanced regrowth forest before logging)
25	Imanuel's Garden	Garden
26	Garu, Swept Village	Bare ground in Garu village

Source: Carol Lentfer, Southern Cross University, Australia.

Table A.5 Counts of Early Stone Age Tools

	Tools															
	Arrows					Knives			Scrapers				Axes			
Site	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	2	0	0	0	1	0	0	0	12	0	0	4	0	2	0	0
2	0	0	0	0	0	1	0	0	2	0	8	0	0	0	0	0
3	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	3	0	0	1	0	1	0	0
5	0	0	0	0	0	0	0	0	2	0	0	3	0	0	0	0
6	0	0	2	1	0	0	1	0	0	0	1	1	0	0	0	0
7	0	0	0	1	0	0	1	0	2	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
9	2	0	0	0	0	1	0	0	4	0	1	6	0	0	0	0
10	4	2	0	0	0	0	0	0	2	1	1	0	0	0	0	0
11	2	2	1	0	1	0	0	0	3	1	5	1	0	0	2	1
12	10	0	7	0	4	6	1	0	5	3	1	7	1	0	0	0
13	1	0	1	0	2	1	1	0	2	0	1	2	0	0	0	3
14	1	0	0	0	0	4	0	0	0	0	6	4	0	0	0	4
15	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0
16	0	0	1	0	2	0	0	0	7	0	0	0	0	0	0	2
17	1	0	2	0	2	1	1	0	5	0	2	1	0	0	0	0
18	2	0	0	0	1	0	0	0	3	0	0	1	0	0	0	0
19	4	0	1	0	0	2	0	0	16	7	0	3	0	4	0	0
20	0	0	0	0	0	0	0	0	18	2	1	5	0	3	0	0
21	1	0	1	1	0	1	3	0	5	2	0	5	0	0	0	0
22	0	0	2	0	0	2	0	0	2	0	0	0	0	0	0	0
23	0	0	0	0	1	1	0	0	2	0	0	0	0	0	0	0
24	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
25	4	8	4	4	0	3	0	0	13	0	0	8	2	1	0	0
26	1	2	8	4	0	7	5	2	21	4	1	9	1	0	0	0
27	0	0	0	0	0	2	0	0	1	0	0	1	2	0	0	0
28	0	0	2	0	0	2	0	0	5	2	0	0	0	0	0	0
29	0	0	0	0	0	3	2	0	3	0	0	1	0	0	0	0

Table A.5 (continued)

	Tools															
	Arrows					Knives			Scrapers				Axes			
Site	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
30	0	0	0	0	0	1	0	0	5	0	0	2	0	0	0	0
31	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	0
32	1	0	0	1	0	1	1	0	0	0	0	3	0	0	0	0
33	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
34	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
35	0	0	1	0	1	0	1	0	2	0	0	1	0	0	0	0
36	0	0	4	1	0	0	0	0	3	0	0	1	1	0	0	0
37	0	0	2	0	0	0	0	0	3	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	7	0	5	0	0	0	0	0
39	5	0	3	0	1	0	1	0	6	0	3	3	0	0	0	1
40	24	0	5	0	6	9	2	0	12	2	1	13	1	0	0	0
41	16	1	10	0	9	25	11	0	32	2	3	19	5	0	0	3
42	2	0	0	0	0	0	0	0	1	1	3	1	0	0	0	0
43	18	0	8	0	14	26	11	0	30	2	8	26	4	0	0	4

Source: Bølviken *et al.* (1982).

Table A.6 Descriptions of Early Stone Age Tools

Code	Tool
1	Tanged arrows
2	Blade arrows
3	Transverse and oblique arrows
4	Atypical arrows
5	Microliths
6	Flake knives
7	Blade knives
8	Notched knives
9	Core and flake scrapers
10	Blade scrapers
11	Discscrapers
12	Burins
13	Axes
14	Chisels
15	Slate axes
16	Perforators

Source: Bølviken *et al.* (1982).

Table A.7 Weights of Memphis Pottery Sherds (kg): Merged Wares

		Context												
Code	**Ware**	**377**	**465**	**509**	**476**	**289**	**690**	**716**	**739**	**740**	**707**	**761**	**758**	**749**
1	**NILECHANDMADE**	0.00	0.00	1.14	0.40	0.00	1.96	0.00	0.00	3.71	8.00	0.20	0.10	0.00
2	**H10.01.00.14, H10.01**	2.50	0.08	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.04	0.00	0.00	0.00
3	**D04.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00
4	**A01.01**	0.01	0.10	0.45	0.01	0.32	0.01	0.20	0.04	0.88	3.00	0.11	1.11	0.32
5	**D01.04, D01.01, D01SMOKED, D01.00.00.01, D01.08**	0.01	0.05	0.08	0.01	0.54	0.47	0.08	0.39	0.20	3.71	0.10	0.40	0.72
6	**E01SMOKED, E01.09, E01.01**	0.08	0.33	0.14	0.34	1.72	2.78	2.50	0.28	6.24	17.20	0.30	9.00	6.50
7	**G01.01**	3.90	3.14	2.98	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	**G01.02, G01.02.00.01**	0.18	0.08	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	**G01.08, G01.06**	0.68	0.33	1.22	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	**G01.04, G01.29**	0.20	0.02	0.05	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
11	**G05.01**	0.16	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	**H01.05**	0.04	0.30	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	**NILE UNDATABLE**	0.22	1.39	2.60	0.16	16.61	13.65	13.49	2.06	13.42	48.12	2.14	28.16	15.41
14	**UNKNOWN**	0.00	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	**P40.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
16	**P16.01**	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	**P33.01**	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	**H02.01**	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	**H08.01**	0.08	0.34	0.00	0.01	0.00	0.81	0.01	0.00	0.00	0.00	0.00	0.00	0.06
20	**P31.01**	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.20	0.00	0.01	0.03	0.08
21	**NILEB2.15, G01.15**	0.00	0.01	0.42	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	**G01SMOKED**	0.00	0.01	0.26	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	**H24.01.00.02, H24.01**	0.00	0.02	0.20	0.10	0.16	0.18	0.07	0.00	0.12	0.83	0.08	0.24	0.30
24	**BREADMOULDS**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.00
25	**D02.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
26	**P34.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00
27	**D03HANDMADE**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08

Source: Janine Bourriau, Macdonald Institute, Cambridge, England.

Table A.8 Ceramic Pot Measurements

	Measurement (cm)												
Pot Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	6.50	7.85	5.65	6.70	7.30	7.00	6.30	6.55	4.80	4.10	0.15	0.30	0.30
2	6.85	7.40	5.85	6.80	7.10	7.55	6.70	6.75	4.10	3.50	0.30	0.25	0.30
3	6.85	7.30	5.80	6.80	6.60	7.55	6.75	6.65	4.30	3.55	0.15	0.35	0.35
4	5.90	6.80	5.70	6.75	6.30	7.00	6.35	6.75	4.40	3.60	0.30	0.40	0.25
5	6.30	7.30	4.70	6.55	6.75	6.95	6.35	6.35	4.60	4.15	0.30	0.55	0.30
6	6.15	7.15	6.20	7.20	6.20	6.85	6.15	7.00	4.75	3.90	0.25	0.40	0.45
7	6.50	7.40	6.25	6.85	6.70	7.40	6.75	6.15	4.00	3.45	0.25	0.25	0.30
8	6.35	7.60	5.25	6.75	6.95	6.95	6.40	6.40	4.65	3.95	0.35	0.50	0.30
9	6.60	7.70	6.55	7.05	6.80	7.40	6.70	6.75	4.20	3.55	0.25	0.20	0.30
10	6.60	7.35	6.40	7.00	7.00	7.45	6.80	6.60	4.05	3.40	0.30	0.25	0.28
11	6.80	7.40	6.30	7.00	6.85	7.40	6.60	6.70	4.20	3.65	0.15	0.20	0.30
12	6.50	7.40	5.10	6.60	6.80	7.15	6.55	6.45	4.50	4.00	0.15	0.55	0.30
13	5.90	7.25	6.55	7.25	6.30	6.90	6.05	7.05	4.20	3.50	0.25	0.30	0.45
14	5.85	7.10	6.55	7.20	6.35	7.05	6.15	7.10	4.40	3.55	0.30	0.35	0.35
15	6.10	7.15	6.50	7.15	6.25	7.10	6.35	7.10	4.45	3.50	0.25	0.40	0.40
16	6.65	7.45	5.85	6.90	6.60	7.35	6.60	6.70	4.50	3.85	0.10	0.35	0.35
17	6.55	7.45	5.25	6.60	6.70	7.10	7.10	6.55	4.65	4.10	0.20	0.50	0.35
18	6.15	7.20	6.70	7.20	6.20	7.15	6.65	6.85	4.35	3.50	0.30	0.30	0.35
19	6.00	7.00	6.50	7.10	6.15	7.05	6.30	6.90	4.10	3.30	0.25	0.35	0.45
20	6.00	7.20	6.55	7.15	6.30	7.05	6.25	7.00	4.35	3.70	0.30	0.30	0.45
21	6.75	7.50	5.95	6.90	6.75	7.45	6.80	6.65	4.30	3.55	0.15	0.35	0.35
22	6.30	6.80	6.50	6.90	5.95	7.30	6.30	6.80	4.30	3.55	0.20	0.25	0.35
23	6.10	7.25	6.65	7.40	6.40	7.00	6.10	7.25	4.40	3.70	0.15	0.30	0.50
24	6.75	7.55	6.60	6.95	6.80	7.35	6.55	6.75	4.20	3.60	0.20	0.30	0.35
25	6.30	7.60	5.20	6.55	6.80	7.10	6.50	6.55	4.40	3.95	0.30	0.50	0.35
26	6.90	7.60	6.25	6.80	6.65	7.55	6.85	6.55	4.10	3.55	0.15	0.25	0.30
27	6.40	7.65	5.55	6.70	6.85	7.20	6.50	6.40	4.75	4.15	0.20	0.55	0.40
28	6.20	7.55	5.75	6.85	6.95	7.05	6.45	6.65	4.75	4.10	0.40	0.50	0.25
29	6.15	8.05	5.65	7.05	7.15	6.85	6.20	6.70	4.50	3.85	0.30	0.55	0.40
30	6.50	7.55	5.15	6.85	6.90	7.05	6.60	6.65	4.75	4.05	0.35	0.60	0.35

Source: Impey & Pollard (1985).

Table A.9 Simpson Desert Flint Tool Measurements

	Measurement					
Site	**Length (mm)**	**Width (mm)**	**Thickness (mm)**	**Platform Width (mm)**	**Platform Thickness (mm)**	**Weight (grams)**
08	42	33	8	24	5	14.5
08	31	18	8	15	7	5.0
08	36	20	4	19	7	5.5
08	31	17	8	6	3	4.5
08	38	27	9	9	5	9.0
08	34	15	8	16	5	4.5
08	30	18	4	18	3	2.0
08	29	18	7	16	6	5.0
08	48	21	10	16	7	10.5
08	36	32	15	33	14	22.0
08	54	33	13	29	12	27.0
08	49	34	19	7	4	23.0
08	39	24	9	15	5	7.5
08	56	16	5	9	4	6.0
08	25	32	11	27	13	9.0
08	43	60	25	46	15	68.5
08	39	43	16	28	5	25.0
08	27	16	5	9	3	2.5
08	41	21	9	23	8	11.5
08	55	20	7	11	6	10.5
08	45	29	19	29	14	29.0
08	53	29	7	20	18	15.5
08	34	18	8	17	5	7.5
08	26	13	6	11	3	2.0
08	65	23	9	6	3	11.0
08	26	12	5	11	6	2.0
08	25	13	4	12	6	1.5
08	28	15	4	13	4	3.0
08	22	14	3	8	3	1.0
08	29	14	5	7	5	2.5
08	36	20	7	7	3	5.5
08	34	18	5	13	5	3.5
08	37	18	5	10	4	3.5
08	36	23	5	20	4	4.5
08	32	22	6	4	7	4.5
08	33	19	9	24	8	9.0
08	41	23	10	17	6	10.0
08	39	25	10	15	3	12.0
08	52	30	7	16	6	12.0
08	53	24	10	20	9	18.0

Table A.9 (continued)

	Measurement					
Site	*Length* (mm)	*Width* (mm)	*Thickness* (mm)	*Platform* Width (mm)	Platform Thickness (mm)	Weight (grams)
08	48	34	10	15	6	22.0
08	60	23	6	14	4	16.0
08	65	35	16	23	10	41.5
08	35	23	18	22	13	18.0
08	46	20	6	6	3	8.0
08	34	22	9	14	7	6.0
08	26	16	3	12	5	2.5
08	48	24	10	15	8	11.5
08	49	20	13	16	8	12.5
08	70	61	27	36	34	176.0
08	30	13	6	2	1	2.0
08	23	31	10	19	7	8.5
09	11	43	12	42	11	8.0
09	16	32	10	25	9	5.5
09	20	36	9	26	5	4.0
09	16	36	16	26	7	11.5
09	12	38	10	28	9	5.5
09	4	31	7	25	9	3.5
09	27	20	12	19	5	8.5
09	9	28	7	28	8	2.0
09	8	41	8	26	8	3.5
09	16	33	8	30	9	6.5
09	15	29	8	24	7	5.5
09	42	41	12	54	7	29.5
09	14	38	9	33	6	6.0
09	7	36	11	33	12	5.0
09	13	33	12	33	8	8.0
09	25	15	5	4	14	2.5
09	34	29	14	24	10	17.0
09	12	34	9	30	9	5.5
09	12	37	8	25	3	4.0
09	55	19	7	14	7	11.0
09	13	49	10	37	10	10.5
09	11	34	10	22	10	5.5
09	35	15	5	17	9	4.0
09	16	28	9	18	9	4.0
09	11	51	13	25	8	10.0
09	30	14	6	10	8	3.0

Source: Huw Barton, University of Sydney, Australia.

Table A.10 Site Descriptions for the Simpson Desert Flake Debitage

Site	Terrain	Landform	Water Permanency
13	Dunefield	Clayey Interdune	Ephemeral
14	Dunefield	Clayey Interdune	Ephemeral
15	Dunefield	Clayey Interdune	Ephemeral
16	Dunefield	Clayey Interdune	Ephemeral
19	Dunefield	Clayey Interdune	Semi-permanent
20	Dunefield	Clayey Interdune	Semi-permanent
10	Dunefield	Claypan	Ephemeral
11	Dunefield	Claypan	Ephemeral
17	Dunefield	Claypan	Ephemeral
18	Dunefield	Claypan	Ephemeral
27	Dunefield	Claypan	Ephemeral
30	Dunefield	Sand Sheet/Claypan	Semi-permanent
02	Dunefield	Sandy Interdune	Permanent
29	Dunefield	Sandy Interdune/Claypan	Semi-permanent
26	Dunefield	Spring	Permanent
01	Dunefield	Spring	Permanent
21	Dunefield	Stony Interdune	Ephemeral
22	Dunefield	Stony Interdune	Ephemeral
31	Dunefield	Swamp	Semi-permanent
05	Floodplain	Claypan	Semi-permanent
23	Floodplain	Dune Flank	Semi-permanent
12	Sandplain	Claypan	Ephemeral
25	Sandplain	Claypan	Ephemeral
06	Sandplain	Plain with drainage	Ephemeral
09	Sandplain	Plain with drainage	Ephemeral
24	Sandplain	Stony Interdune	Ephemeral
07	Dissected Residual	Escarpment	Ephemeral
08	Dissected Residual	Escarpment	Ephemeral
32	Gibberplain	Channel	Semi-permanent

Source: Huw Barton, University of Sydney, Australia.

Table A.11 Counts of Engraved Bone Design Elements

	Site				
Design	**Altamira**	**Cueto de la Mina**	**El Juyo**	**El Cierro**	**La Paloma**
1	2	1	0	0	0
2	12	12	8	5	4
3	7	2	2	1	0
4	1	0	1	2	0
5	0	1	0	0	0
6	3	0	0	0	0
7	12	0	0	0	0
8	15	3	12	7	1
9	0	1	3	3	2
10	3	5	9	2	2
11	1	0	0	1	0
12	1	1	0	0	0
13	12	4	2	4	3
14	7	3	1	0	0
15	3	1	1	2	0
16	11	0	0	0	0
17	3	0	0	0	0
18	1	1	1	0	1
19	7	2	1	2	0
20	2	4	0	0	0
21	4	0	1	0	0
22	3	1	0	0	0
23	3	1	2	1	0
24	1	0	0	0	0
25	5	1	1	0	1
26	1	0	0	0	0
27	1	0	1	0	0
28	1	2	1	0	0
29	0	2	0	0	0
30	2	0	0	0	0

Table A.11 (continued)

	Site				
Design	**Altamira**	**Cueto de la Mina**	**El Juyo**	**El Cierro**	**La Paloma**
31	0	1	0	0	0
32	1	0	0	0	0
33	0	7	0	0	0
34	1	0	0	1	0
35	1	0	0	0	0
36	1	0	0	0	0
37	3	0	0	0	0
38	0	1	0	0	0
39	2	1	1	1	1
40	1	2	0	0	0
41	4	2	2	0	1
42	5	6	0	2	4
43	4	1	3	1	1
44	5	0	0	0	2

Source: Kaufman (1998).

Table A.12 Counts of Hunting Spiders

	Species											
Site	Al. accentuata	Al. cuneata	Al. fabrilis	Ar. lutetiana	Ar. perita	Au. albimana	Pa. lugubris	Pa. monticola	Pa. nigriceps	Pa. pullata	Tr. terricola	Zo. spinimana
1	25	10	0	0	0	4	0	60	12	45	57	4
2	0	2	0	0	0	30	1	1	15	37	65	9
3	15	20	2	2	0	9	1	29	18	45	66	1
4	2	6	0	1	0	24	1	7	29	94	86	25
5	1	20	0	2	0	9	1	2	135	76	91	17
6	0	6	0	6	0	6	0	11	27	24	63	34
7	2	7	0	12	0	16	1	30	89	105	118	16
8	0	11	0	0	0	7	55	2.	2	1	30	3
9	1	1	0	0	0	0	0	26	1	1	2	0
10	3	0	1	0	0	0	0	22	0	0	1	0
11	15	1	2	0	0	1	0	95	0	1	4	0
12	16	13	0	0	0	0	0	96	1	8	13	0
13	3	43	1	2	0	18	1	24	53	72	97	22
14	0	2	0	1	0	4	3	14	15	72	94	32
15	0	0	0	0	0	0	6	0	0	0	25	3
16	0	3	0	0	0	0	6	0	2	0	28	4
17	0	0	0	0	0	0	2	0	0	0	23	2
18	0	1	0	0	0	0	5	0	0	0	25	0
19	0	1	0	0	0	0	12	0	1	0	22	3
20	0	2	0	0	0	0	13	0	0	0	22	2
21	0	1	0	0	0	0	16	1	0	1	18	2
22	7	0	16	0	4	0	0	2	0	0	1	0
23	17	0	15	0	7	0	2	6	0	0	1	0
24	11	0	20	0	5	0	0	3	0	0	0	0
25	9	1	9	0	0	2	1	11	6	0	16	6
26	3	0	6	0	18	0	0	0	0	0	1	0
27	29	0	11	0	4	0	0	1	0	0	0	0
28	15	0	14	0	1	0	0	6	0	0	2	0

Source: van der Aart & Smeenk-Enserink (1975).

Table A.13 Environmental Variable Measurements for the Hunting Spider Sites

	Environmental Variable												
Site	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	10.3	5.3	3.70	0	0	20	85	5	1	50	3	40	0
2	21.1	9.7	3.68	0	5	2	95	50	1	80	2	0	0
3	12.9	6.5	3.60	0	0	10	99	20	1	30	2	60	0
4	14.5	4.8	3.36	0	0	10	100	50	1	100	0	2	0
5	20.4	6.0	3.46	0	0	0	100	30	2	90	2	4	0
6	29.4	12.3	3.65	10	30	10	30	40	1	10	0	0	0
7	24.0	8.3	3.64	0	0	1	100	30	20	90	0	0	0
8	13.8	5.4	3.70	0	70	2	30	30	2	10	0	0	0
9	12.0	5.1	3.38	0	0	75	25	2	1	0	3	0	20
10	9.0	4.4	3.60	50	0	30	20	3	1	0	3	1	20
11	9.2	4.5	3.60	0	0	60	40	10	0	0	4	0	30
12	9.9	4.4	3.41	0	0	45	55	3	1	2	3	0	50
13	33.7	13.2	3.87	5	5	1	90	30	2	80	10	20	0
14	21.9	7.8	3.58	0	0	5	95	20	1	20	4	20	0
15	26.3	5.7	3.58	0	80	1	20	30	1	0	0	0	0
16	20.7	6.8	3.56	0	99	1	1	30	1	0	0	0	0
17	28.0	9.4	3.45	0	85	1	20	40	2	0	0	0	0
18	22.7	9.5	3.43	0	80	0	20	30	2	0	0	0	0
19	18.6	6.9	3.62	0	90	4	4	20	1	0	0	2	0
20	22.4	8.1	3.59	0	98	1	1	25	1	0	0	0	0
21	19.6	5.8	4.27	0	95	1	5	35	1	0	0	0	0
22	3.5	1.6	7.37	25	0	75	1	20	1	2	2	0	2
23	3.3	1.4	7.37	20	0	55	25	20	1	2	3	0	20
24	5.2	2.1	6.73	25	0	55	20	10	0	2	2	0	20
25	6.2	2.1	6.41	35	0	2	60	45	2	1	0	0	0
26	2.7	1.1	7.84	90	0	5	5	3	0	0	0	0	10
27	2.6	1.6	6.58	10	0	45	30	4	3	0	2	0	20
28	2.6	2.2	7.23	30	0	40	30	2	0	0	2	0	30

Table A.13 (continued)

	Environmental Variable												
Site	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
1	0	0	0	0	0	0	0	0	0	0	18	68	50
2	2	20	0	5	3	0	25	4	0	1	7	6	5
3	0	0	0	0	0	0	0	0	0	0	15	43	40
4	0	0	0	0	0	0	0	0	0	0	12	16	20
5	0	0	0	0	0	0	0	0	0	0	3	16	10
6	1	20	20	15	6	20	70	0	0	0	4	3	2
7	2	0	0	0	0	1	0	0	0	0	21	21	10
8	0	0	1	15	10	0	75	5	4	2	3	3	2
9	0	0	0	0	0	0	0	0	0	0	25	56	30
10	0	0	0	0	0	0	0	0	0	0	26	60	40
11	0	0	0	0	0	0	0	0	0	0	19	50	40
12	0	0	0	0	0	0	0	0	0	0	19	60	40
13	2	2	0	0	0	0	0	0	0	0	12	33	30
14	2	3	0	0	0	2	12	0	0	1	10	10	3
15	2	4	0	0	0	0	45	6	2	2	2	3	2
16	0	0	0	0	0	0	85	5	2	0	1	2	1
17	0	10	0	0	0	0	40	0	3	0	3	5	3
18	2	3	1	9	6	2	80	5	1	2	2	5	3
19	0	2	0	0	0	1	50	4	2	1	1	4	1
20	0	0	0	0	0	0	75	6	2	1	1	3	1
21	3	2	0	30	20	2	50	6	2	1	2	3	1
22	0	0	0	0	0	0	0	0	0	0	19	67	50
23	0	0	0	0	0	0	0	0	0	0	17	57	60
24	0	0	0	0	0	0	0	0	0	0	18	55	55
25	1	2	1	15	8	0	50	0	0	0	2	5	10
26	0	0	0	0	0	0	0	0	0	0	19	37	80
27	0	0	0	0	0	0	0	0	0	0	19	56	40
28	0	0	0	0	0	0	0	0	0	0	19	54	40

Source: van der Aart & Smeenk-Enserink (1975).

Table A.14 Environmental Variable Descriptions for the Hunting Spider Sites

Environmental Variable	Description
1	water content (percentage dry weight)
2	humus content (percentage dry weight)
3	acidity (pH-KCI)
4	percentage bare sand
5	cover by fallen leaves and twigs (percentage)
6	cover by moss layer (percentage)
7	cover by herb layer (percentage)
8	maximum height herb layer (centimetres)
9	minimum height herb layer (centimetres)
10	cover by *Calamagrostis epigejos* (percentage)
11	cover by *Carex arenaria* (percentage)
12	cover by *Festuca ovina* (percentage)
13	cover by *Corynephorus canescens* (percentage)
14	cover by *Urtica dioica* (percentage)
15	cover by *Moehringia trinervia* (percentage)
16	cover by shrub layer (percentage)
17	maximum height shrub layer (decimetres)
18	minimum height shrub layer (decimetres)
19	cover by *Ligustrum vulgare* (percentage)
20	cover by tree layer (percentage)
21	maximum height tree layer (metres)
22	cover by *Populus tremula* (five class scale)
23	cover by *Crataegus monogyna* (five class scale)
24	lux at equal grey sky (× 1000)
25	lux at cloudless sky (× 1000)
26	reflection of soil surface at cloudless sky (× 100)

Source: van der Aart & Smeenk-Enserink (1975).

Table A.15 Abundances of Dune Meadow Vegetation

	Site																			
Species	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Achi. mill.	1	3	0	0	2	2	2	0	0	4	0	0	0	0	0	0	2	0	0	0
Agro. stol.	0	0	4	8	0	0	0	4	3	0	0	4	5	4	4	7	0	0	0	5
Aira prae.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0
Alop. geni.	0	2	7	2	0	0	0	5	3	0	0	8	5	0	0	4	0	0	0	0
Anth. odor.	0	0	0	0	4	3	2	0	0	4	0	0	0	0	0	0	4	0	4	0
Bell. pere.	0	3	2	2	2	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0
Brom. hord.	0	4	0	3	2	0	2	0	0	4	0	0	0	0	0	0	0	0	0	0
Chen. albu.	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Cirs. arve.	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eleo. palu.	0	0	0	0	0	0	0	4	0	0	0	0	0	4	5	8	0	0	0	4
Elym. repe.	4	4	4	4	4	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
Empe. nigr.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Hypo. radi.	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	5	0
Junc. arti.	0	0	0	0	0	0	0	4	4	0	0	0	0	0	3	3	0	0	0	4
Junc. bufo.	0	0	0	0	0	0	2	0	4	0	0	4	3	0	0	0	0	0	0	0
Leon. autu.	0	5	2	2	3	3	3	3	2	3	5	2	2	2	2	0	2	5	6	2
Loli. pere.	7	5	6	5	2	6	6	4	2	6	7	0	0	0	0	0	0	2	0	0
Plan. lanc.	0	0	0	0	5	5	5	0	0	3	3	0	0	0	0	0	2	3	0	0
Poa prat.	4	4	5	4	2	3	4	4	4	4	4	0	2	0	0	0	1	3	0	0
Poa triv.	2	7	6	5	6	4	5	4	5	4	0	4	9	0	0	2	0	0	0	0
Pote. palu.	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0
Ranu. flam.	0	0	0	0	0	0	0	2	0	0	0	0	2	2	2	2	0	0	0	4
Rume. acet.	0	0	0	0	5	6	3	0	2	0	0	2	0	0	0	0	0	0	0	0
Sagi. proc.	0	0	0	5	0	0	0	2	2	0	2	4	2	0	0	0	0	0	3	0
Sali. repe.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	5
Trif. prat.	0	0	0	0	2	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Trif. repe.	0	5	2	1	2	5	2	2	3	6	3	3	2	6	1	0	0	2	2	0
Vici. lath.	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	1	0	0
Brac. ruta.	0	0	2	2	2	6	2	2	2	2	4	4	0	0	4	4	0	6	3	4
Call. cusp.	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	3	0	0	0	3

Source: Batterink & Wijffels (1983).

Table A.16 Environmental Variable Descriptions for the Dune Meadow Vegetation Sites

	Environmental Variable				
Site	**Thickness of A1 Horizon**	**Moisture Content**	**Quantity of Manure**	**Grassland Use**	**Grassland Management**
1	2.8	1	4	C	SF
2	3.5	1	2	C	BF
3	4.3	2	4	C	SF
4	4.2	2	4	C	SF
5	6.3	1	2	H	HF
6	4.3	1	2	C	HF
7	2.8	1	3	P	HF
8	4.2	5	3	P	HF
9	3.7	4	1	H	HF
10	3.3	2	1	H	BF
11	3.5	1	1	P	BF
12	5.8	4	2	C	SF
13	6.0	5	3	C	SF
14	9.3	5	0	P	NC
15	11.5	5	0	C	NC
16	5.7	5	3	P	SF
17	4.0	2	0	H	NC
18	4.6	1	0	H	NC
19	3.7	5	0	H	NC
20	3.5	5	0	H	NC

Source: Batterink & Wijffels (1983).

Bibliography

Barkman, J. J., Doing, H. & Segal, S. (1964). Kritische Bemerkungen und Vorschläge zur Quantitativen Vegetationsanalyse. *Acta Botanica Neerlandica*, **13**, 394-419.

Barnett, V. (1976). The Ordering of Multivariate Data (with discussion). *Journal of the Royal Statistical Society Series A*, **139**, 318-54.

Barnett, V. (1991). *Sample Survey Principles and Methods*. Edward Arnold, London.

Barr, G. D. I., Underhill, L. G. & Kahn, S. B. (1990). The Covariance Biplot for Graphical Display of Multivariate Time Series Data. *American Journal of Mathematical and Management Sciences*, **10**, 1-15.

Batterink, M. & Wijffels, G. (1983). *Een vergelijkend vegetatiekundig onderzoek naar de typologie en invloeden van het beheer van 1973 tot 1982 in de duinweilanden op Terschelling*. Report V.P.O., Agricultural University, Wageningen.

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

Bobrowsky, P. T. & Ball, B. F. (1989). The Theory and Mechanics of Ecological Diversity in Archaeology. In *Quantifying Diversity in Archaeology* (eds. R. D. Leonard, & G. T. Jones), pp 4-12. Cambridge University Press, Cambridge.

Bogaard, A., Palmer, C., Jones, G., Charles, M. & Hodgson, J. G. (1999). A FIBS Approach to the Use of Weed Ecology for the Archaeobotanical Recognition of Crop Rotation Regimes. *Journal of Archaeological Science*, **26**, 1211-24.

Bølviken, E., Helskog, E., Helskog, K., Holm-Olsen, I. M., Solheim, L. & Bertelsen, R. (1982). Correspondence Analysis: An Alternative to Principal Components. *World Archaeology*, **14**, 41-60.

Brillouin, L. (1962). *Science and Information theory*. Academic Press, New York.

Chessel, D., Lebreton, J. D. & Yoccoz, N. (1987). Propriétés de l'Analyse Canonique des Correspondances; une Illustration en Hydrobiologie. *Revue de Statistique Appliquée*, **35**, 55-72.

Conkey, M. W. (1980). The Identification of Prehistoric Hunter-Gatherer Aggregation Sites: The Case of Altamira. *Current Anthropology*, **21**, 609-30.

Eastment, H. T. & Krzanowski, W. J. (1982). Cross-Validatory Choice of the Number of Components from a Principal Component Analysis. *Technometrics*, **24**, 73-77.

Eckart, C. & Young, G. (1936). The Approximation of One Matrix by another of Lower Rank. *Psychometrika*, **1**, 211-18.

Efron, B. (1979). Bootstrap Methods: Another look at the Jackknife. *The Annals of Statistics*, 7, 1-26.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge.

Gabriel, K. R. (1971). The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*, **58**, 453-67.

Gabriel, K. R. (1972). Analysis of Meteorological Data by Means of Canonical Decomposition and Biplots. *Journal of Applied Meteorology*, **11**, 1071-77.

Gabriel, K. R. (1981). Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis. In *Interpreting Multivariate Data* (ed. V. Barnett), pp 147-73. John Wiley and Sons, New York.

Gabriel, K. R. (1995). Biplot Display of Multivariate Categorical Data, with Comments on Multiple Correspondence Analysis. In *Recent Advances in Descriptive Multivariate Analysis* (ed. W. J. Krzanowski), pp 190-226. Oxford University Press, New York.

Gauch Jr., H. G. (1982). *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge.

Gauch Jr., H. G., Whittaker, R. H. & Wentworth, T. R. (1977). A Comparative Study of Reciprocal Averaging and other Ordination Techniques. *Journal of Ecology*, **65**, 157-74.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* John Wiley and Sons, New York.

Gower, J. C. (1974). The Mediancentre. *Applied Statistics*, **23**, 466-70.

Gower, J. C. (1984). Multivariate Analysis: Ordination, Multidimensional Scaling and Allied Topics. In *Handbook of Applicable Mathematics Volume VI, Statistics, Part B* (eds. W. Ledermann & E. Lloyd), pp 727-81. Wiley, Chichester.

Gower, J. C. & Hand, D. J. (1996). *Biplots.* Chapman and Hall, London.

Grayson, D. K. (1984). *Quantitative Zooarchaeology: Topics in the Analysis of Archaeological Faunas.* Academic Press, New York.

Green, P. J. (1981). Peeling Bivariate Data. In *Interpreting Multivariate Data.* (ed. V. Barnett), pp 3-19. John Wiley and Sons, New York.

Green, P. J. & Silverman, B. W. (1979). Constructing the Convex Hull of a Set of Points in the Plane. *The Computer Journal*, **22**, 262-66.

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

Greenacre, M. J. (1988). Clustering the Rows and Columns of a Contingency Table. *Journal of Classification*, **5**, 39-51.

Greenacre, M. J. (1993a). Biplots in Correspondence Analysis. *Journal of Applied Statistics*, **20**, 251-69.

Greenacre, M. J. (1993b). *Correspondence Analysis in Practice.* Academic Press, San Diego.

Greenacre, M. J. & Hastie, T. (1987). The Geometric Interpretation of Correspondence Analysis. *Journal of the American Statistical Association*, **82**, 437-47.

Greenacre, M. J. & Underhill, L. G. (1982). Scaling a Data Matrix in Low-Dimensional Euclidean Space. In *Applied Multivariate Analysis* (ed. D. M. Hawkins), pp 183-268. Cambridge University Press, Cambridge.

Greenberg, J. H. (1956). The Measurement of Linguistic Diversity. *Language*, **32**, 109-15.

Hill, M. O. (1973). Reciprocal Averaging: An Eigenvector Method of Ordination. *Journal of Ecology*, **61**, 237-49.

Hill, M. O. & Gauch Jr., H. G. (1980). Detrended Correspondence Analysis: An Improved Ordination Technique. *Vegetatio*, **42**, 47-58.

Hirotsu, C. (1983). Defining the Pattern of Association in Two-Way Contingency Tables. *Biometrika*, **70**, 579-89.

Hirschfield, H. O. (1935). A Connection between Correlation and Contingency. *Cambridge Philosophical Society Proceedings*, **31**, 520-24.

Hoffman, D. L. & Franke, G. R. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research*, **23**, 213-27.

Iloni, K. (1991). Biplot Graphical Display Techniques. M.Sc. Thesis, University of Cape Town (unpublished).

Impey, O. R. (1979). The Earliest Japanese Porcelains; Styles and Techniques. In *Decorative techniques and styles in Asian Ceramics* (ed. M. Medley). Percival David Foundation of Chinese Art Colloquy on Art and Archaeology in Asia 8, London.

Impey, O. R. & Pollard, M. (1985). A Multivariate Metrical Study of Ceramics made by Three Potters. *Oxford Journal of Archaeology*, **4**, 157-64.

Jennrich, R. I. & Turner, F. B. (1969). Measurement of Non-circular Home Range. *Journal of Theoretical Biology*, **22**, 227-37.

Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial data. *Applied Statistics*, **21**, 160-73.

Jolliffe, I. T. (1973). Discarding Variables in a Principal Component Analysis. II: Real data. *Applied Statistics*, **22**, 21-31.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

Jongman, R. H. G., ter Braak, C. J. F. & van Tongeren, O. F. R. (eds, 1995). *Data Analysis in Community and Landscape Ecology*. Cambridge University Press, New York.

Kaufman, D. (1998). Measuring Archaeological Diversity: An Application of the Jack-knife Technique. *American Antiquity*, **63**, 73-85.

Kendall, D. G. (1971). Seriation from Abundance Matrices. In *Mathematics in the Archaeological and Historical Sciences* (eds. F. R. Hodson, D. G. Kendall & P. Tautu), pp 215-252. Edinburgh University Press, Edinburgh.

Kintigh, K. W. (1984). Measuring Archaeological Diversity by Comparison with Simulated Assemblages. *American Antiquity*, **49**, 44-54.

Kintigh, K. W. (1989). Sample Size, Significance, and Measures of Diversity. In *Quantifying Diversity in Archaeology*. (eds. R. D. Leonard & G. T. Jones), pp 25-36. Cambridge University Press, Cambridge.

Krzanowski, W. J. (1987). Selection of Variables to Preserve Multivariate Data Structure, using Principal Components. *Applied Statistics*, **36**, 22-33.

Krzanowski, W. J. (1993). Attribute Selection in Correspondence Analysis of Incidence Matrices. *Applied Statistics*, **42**, 529-41.

Krzanowski, W. J. (1996). A Stopping Rule for Structure-Preserving Variable Selection. *Statistics and Computing*, **6**, 51-56.

Lance, G. N. & Williams, W. T. (1967). A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems. *The Computer Journal*, **9**, 373-80.

Lebreton, J. D., Chessel, D., Prodon, R. & Yoccoz, N. (1988). L'Analyse des relations espèces-milieu par l'Analyse Canonique des Correspondances. I. Variables de milieu quantitatives. *Acta Oecologica-Oecologia Generalis*, **9**, 53-67.

Leonard, R. D. & Jones, G. T. (1984). The Concept and Measure of Archaeological Diversity. Paper presented at the 49th Annual Meeting of the Society for American Archaeology, Portland, Oregon.

Lock, G. & Wilcock, J. (1987). *Computer Archaeology*. Shire Publications Ltd, Princes Risborough.

Manly, B. F. J. (1991). *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall, London.

Mardia, K. V. (1972). *Statistics of Directional Data*. Academic Press, London.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

Margalef, D. R. (1958). Information Theory in Ecology. *General Systems*, **3**, 36-71.

McCabe, G. P. (1984). Principal Variables. *Technometrics*, **26**, 137-44.

McCartney, P. H. & Glass, M. F. (1990). Simulation Models and the Interpretation of Archaeological Diversity. *American Antiquity*, **55**, 521-36.

McCune, B. (1997). Influence of Noisy Environmental Data on Canonical Correspondence Analysis. *Ecology*, **78**, 2617-23.

McIntosh, R. P. (1967). An Index of Diversity and the Relation of Certain Concepts to Diversity. *Ecology*, **48**, 392-404.

Menhinick, E. F. (1964). A Comparison of some Species-Individuals Diversity Indices Applied to Samples of Field Insects. *Ecology*, **45**, 859-61.

Milan, L. & Whittaker, J. (1995). Application of the Parametric Bootstrap to Models that Incorporate a Singular Value Decomposition. *Applied Statistics*, **44**, 31-49.

Morgan, B. J. T. (1995). *Elements of Simulation*. Chapman and Hall, London.

Odner, K. (1966). Komsakulturen i Nesseby og Sør-Varanger. *Tromso Museums Skrifter*, **12**.

Odum, E. P. (1971). *Fundamentals of Ecology*. W.B. Saunders, Philadelphia.

Odum, H. T., Cantlon, J. E. & Kornicker, L. S. (1960). An Organizational Hierarchy

Postulate for the Interpretation of Species-Individual Distributions, Species Entropy, Ecosystem Evolution and the Meaning of a Species-Variety Index. *Ecology*, **41**, 395-99.

Oksanen, J. (1987). Problems of Joint Display of Species and Site Scores in Correspondence Analysis. *Vegetatio,* **72**, 51-7.

Orton, C., Tyers, P. & Vince, A. (1993). *Pottery in Archaeology.* Cambridge University Press, Cambridge.

Pack, P. & Jolliffe, I. T. (1992). Influence in Correspondence Analysis. *Applied Statistics*, **41**, 365-80.

Petrie, W. M. F. (1899). Sequences in prehistoric remains. *Journal of the Anthropological Institute of Great Britain and Ireland*, **29**, 295-301.

Pielou, E. C. (1975). *Ecological Diversity.* John Wiley and Sons, New York.

Pielou, E. C. (1977). *Mathematical Ecology.* John Wiley and Sons, New York.

Rhode, D. (1988). Measurement of Archaeological Diversity and the Sample-Size Effect. *American Antiquity*, **53**, 708-16.

Ringrose, T. J. (1990). The Statistical Analysis of Cave Palaeobiological Data. Ph.D. Thesis, University of Sheffield (unpublished).

Ringrose, T. J. (1992). Bootstrapping and Correspondence Analysis in Archaeology. *Journal of Archaeological Science,* **19**, 615-29.

Robinson, W. S. (1951). A Method for Chronologically Ordering Archaeological Deposits. *American Antiquity*, **16**, 293-301.

Schaafsma, W. & van Vark, G. N. (1979). Classification and Discrimination Problems with Applications, IIa. *Statistica Neerlandica*, **33**, 91-126.

Seheult, A. H., Diggle, P. J. & Evans, D. A. (1976). Contribution to the Discussion of Barnett (1976).

Shelford, V. E. (1911). Ecological Succession: Stream Fishes and the Method of Physiographic Analysis. *Biological Bulletin*, **21**, 9-34.

Simonsen, P. (1961). Varangerfunnene II. Fund og udgravninger pa fjordens sydkyst. *Tromsø Museums Skrifter*, 7.

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, **163**, 688.

Tenenhaus, M. & Young, F. W. (1985). An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and other methods for Quantifying Categorical Multivariate Data. *Psychometrika*, **50**, 91-119.

ter Braak, C. J. F. (1983). Principal Components Biplots and Alpha and Beta Diversity. *Ecology*, **64**, 454-62.

ter Braak, C. J. F. (1985). Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model. *Biometrics*, **41**, 859-73.

ter Braak, C. J. F. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, **67**, 1167-79.

ter Braak, C. J. F. (1987a). Unimodal Models to relate Species to Environment. Ph.D. Thesis, TNO Institute of Applied Computer Science, Wageninen.

ter Braak, C. J. F. (1987b). *CANOCO - A FORTRAN Program for Canonical Community Ordination by [Partial] [Detrended] [Canonical] Correspondence Analysis, Principal Components Analysis and Redundancy Analysis (Version 2.1)*. TNO Institute of Applied Computer Science, Wageningen.

ter Braak, C. J. F. & Verdonschot, P. F. M. (1995). Canonical Correspondence Analysis and Related Multivariate Methods in Aquatic Ecology. *Aquatic Sciences*, **57**, 255-89.

Underhill, L. G. (1990). The Coefficient of Variation Biplot. *Journal of Classification*, 7, 241-56.

van den Wollenberg, A. L. (1977). Redundancy Analysis. An Alternative for Canonical Correlation Analysis. *Psychometrika*, **42**, 207-19.

van der Aart, P. J. M. & Smeenk-Enserink, N. (1975). Correlations Between Distributions of Hunting Spiders (Lycosidae, Ctenidae) and Environmental Characteristics in a Dune Area. *Netherlands Journal of Zoology*, **25**, 1-45.

van der Maarel, E. (1979). Transformation of Cover-Abundance Values in Phytosociology and its Effects on Community Similarity. *Vegetatio*, **39**, 97-114.

Ward Jr., J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **58**, 236-44.

Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon*, **21**, 213-51.

Williams, C. B. (1964). *Patterns in the Balance of Nature and Related Problems in Quantitative Archaeology*. Academic Press, New York.

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, **20**, 397-405.

Zar, J. H. (1974). *Biostatistical Analysis*. Prentice Hall, New Jersey.