Notes on Quantitative Archaeology and R

Mike Baxter
Nottingham Trent University

Notes on Quantitative Archaeology and R

Mike Baxter

May 2015

Contents

List of Figures

List of Tables

Preface

These notes were started during an enforced period of idleness in late 2011. They were abandoned when I recovered enough to move on to other things (i.e. my day job). What was written up to that point, roughly Chapter 7 in the present format, was tidied up and made available on my academia.edu and web sites. It was always the intention to return to them, if only to complete and add a chapter on cluster analysis. This has only happened in the last year or so.

Things didn't go quite according to plan. To begin with, I was dissatisfied with aspects of what had been written and the existing chapters were subjected to major revision and/or expansion. Much of the `R` code for undertaking the analyses was also rewritten. Apart from the chapter on cluster analysis and some rearrangement, chapters on discriminant analysis, factor analysis and statistical inference (mainly hypothesis tests) were added. For reasons explained in the text – chiefly because I think they've been oversold and not delivered what was promised – I had second thoughts about including the last two but they have survived.

The original and rather unworthy genesis of these notes lay in dissatisfaction, possibly bordering on the irrational at times, with what I viewed as the widespread misuse of software such as `Excel` to produce inadequate graphs that disfigure archaeological publication. This remains manifest in the treatment of discrete data. The original intention, abandoned almost immediately, was to keep things short and restrict attention to a few topics traditionally regarded as 'simple'.

A main reason for expanding the coverage is that, in a sense I try to explain in the introduction, I'd regard the vast majority of statistical methods that have been found useful in archaeology as 'simple'. At the conceptual level, and regardless of any mathematical complexity, it is easy to explain in ordinary language what most methods are trying to do. Similarly, the compuations associated with the mathematics can be complex, and computational statistics is a field of study in its own right. However, those who have engaged in this kind of study have made the fruits of their labor widely available, and free, in software such as `R` so, from the point of view of the end-user, the execution of an analysis is also simple. Chapter 2 illustrates this by taking some of the standard methods of multivariate analysis used in archaeology – traditionally often presented as 'complex' – and shows how usefully interpretatble output can be produced in one line of code in R. The idea underlying most of these methods is the simple one of reducing a (possibly large) table of data to a two-dimensional, archaeologically interpretable, 'map'.

This is all made possible by the availability of powerful, open-source software such as R. This is often presented as something that is 'difficult' or 'forbidding' for users reared on menu-driven software, and if potential users do find R 'difficult' then I suppose this is true, though it's also a product of perceptions fuelled by some of the literature, or preconceptions. Writing introductory textbooks on R is something of a growth industry (and there's plenty of free material) so there is considerable assistance out there for aspirant users. Worth mentioning is the fact that it's possible to do some things more rapidly in R than in some popular statistical software such as SPSS; you don't have to wade through vast amounts of irrelevant and sometimes borderline incomprehensible output; and can customize your output (i.e. graphs) to your heart's content rather than being restricted to formats determined by anonymous programmers in the dim and distant past.

I want to emphasize that these notes are *not* intended as a textbook, either as an introduction to statistics or to R. I understand a textbook (as opposed to text) to be a piece of work written, at some length, for didactic purposes, on a topic systematically developed, intended to be read in a linear fashion, and as comprehensive as possible within its chosen remit. The present offering fails, I think, on every count.

The use of the word *Notes* in the title is meant to indicate this. As discussed in the introductory chapter, one or two sections need to be taken early and in some sort of sequence, but mostly the idea is that the text can be dipped into for code that enables you to get going with the kind of data you have to hand. This is, I'm claiming, 'simple', but this is not to be equated with the idea that statistics is 'easy'. Like any other subject worth studying effective use needs to be learned, and this occurs over time with experience and by accretion. The idea here is that you need to start somewhere, and a good case can be made for learning by doing something first and worrying about what you've done later[1].

On the 'learning-by-doing' principle the notes emphazise the importance of real data analysis, and to this end a lot of data sets from different archaeological specializations are analyzed, often in more than one way. These should be available on my academia.edu pages and website[2]. These provide a useful starting point for

[1] In *La Peste* Albert Camus creates a character, Joseph Grand, who is trying to write a novel but doesn't progress beyond the first sentence because he can't perfect it. I had friends at university who never completed, or even started writing, their PhDs for similar reasons. Much better is to get something written down and tidy it up later, possibly discarding much of it in the process. The important thing is to get started.

[2] https://nottinghamtrent.academia.edu/MikeBaxter and http://www.mikemetrics.com/ respectively.

beginning, though if your own data are immediately available in a suitable format analyzing them might be more fun.

Since R has become something of an industry standard in writings on applied statistics that has penetrated many areas of application I'm rather surprised it seems not to have gained much traction in archaeology, individual endeavour and possibly what's invisible and below the surface of publication apart. I don't know of any book length introduction to R for archaeologists, though David Carlson's website promises one[3] and includes R accompaniments to the examples in the standard introductory quantitative archaeology texts of Shennan (1997) and Drennan (2009). These texts together with Carlson's work provide a good entrèe to what is attempted here, which should be seen as complementary rather than a 'competitive' approach to the subject.

One final word of warning is that I don't regard computer programming as among my competencies – it was evident from my undergraduate days that I lacked any aptitude for it. This should encourage those with a similar view of their abilities; those who do have such an aptitude will find some of the code presented inelegant, inefficient, etc. and will be able to improve it. For my own part I have a pragmatic attitude and am usually happy if something works that does what I've asked for and I understand what's happening – that is, I treat R as an extremely useful practical tool for data analysis.

Mike Baxter, Nottingham, May 2015

[3] http://people.tamu.edu/ dcarlson/quant/index.html (Accessed May 2015)

Chapter 1

Introduction

1.1 Introduction

··· simple descriptive statistics and display techniques are indispensable preliminaries to the application of even the most basic inferential statistics or tests. To my knowledge, the vast majority of statistical analyses of archaeological data, published and unpublished, have been done without adequate scrutiny of the data with such elementary display techniques and descriptive statistics. From my experience, I will be so bold as to put forward the view that this lack of adequate scrutiny of the data renders every one of these analyses, and consequently the studies and interpretations based on them, suspect a priori. *··· However, there is another, more positive, pragmatic reason for so strongly advocating the cause of such humble displays and descriptions of one's data, which is that, in almost every instance, one can learn, more quickly, more clearly, and in more detail about one's data with these techniques than through the use of inferential statistics or tests.*

Whallon (1987: 135)

The above is from an article entitled *Simple Statistics*; there is nothing in it to take issue with. It was published more than 20 years after statistical methodology began to attract serious archaeological attention. Any statistical analysis carried out in anger should begin with 'simple descriptive and display techniques'. What has changed since Whallon wrote is what might legitimately be thought of as 'simple'. This is entirely down to the enormous increase in computing power now available. Many of the techniques discussed in these notes could not then be used easily and routinely, if at all.

Fast forward to the 2010s. Some of the software packages developed from the 1970s on survive and remain popular. This isn't necessarily a 'good thing'; practices developed then have become 'fossilized' in the software, not necessarily to good purpose (e.g., Chapter 8 for further comment). To misquote slightly 'the past is a foreign country: they [did] things differently [then]'; they do things differently now.

Software accessibility is important. It explains the understandable widespread use (and misuse) of `Excel` – not a purpose-built statistical software package – in archaeology. There is now no longer any excuse for this. High quality, open-source software means that powerful statistical resources are available to all. The software, `R`, is used in these notes. It is often described as a 'difficult' package to learn for anyone who is not a 'sophisticated' statistician and/or computer programmer; the thinking, simply put, seems to be that potential users who have been weaned on menu-driven software find the command-driven mode that characterizes `R` daunting. Some additional comment on this is provided in Section 1.3.

My own thought is that anyone who can write a sentence with due regard to spelling, syntax and capitalization can start analyzing their own data in `R` in a matter of minutes. Methods sometimes portrayed as 'complex', such as principal component analysis (PCA), are actually quicker to implement in `R` than the menu-driven implementations of software commonly used in teaching (Chapter 2). While `R` is undoubtedly 'sophisticated', matching sophistication on the part of a potential user is not essential. That is, by imitating existing code widely available in texts and online, analytical progress can be made without any great initial understanding of how what you can achieve has been done. Such understanding and greater 'sophistication' will come with practice; the important thing is to get started. Motivation, in the form of having your own data that you want to analyze efficiently, helps.

These views rest on the premise that a lot of statistical ideas are 'simple', in a sense somewhat broader than that used by Whallon. Many methods, traditionally thought of as 'complex', are *computationally* straightforward to execute. The idea of 'simplicity' motivates much of what follows in later chapters. It can be construed in various ways, as follows

- conceptual simplicity;
- computational simplicity (i.e. simplicity of execution);
- mathematical simplicity.

Here, 'computational simplicity' is being equated with simplicity of execution. The *application* of many 'standard' (and 'non-standard') statistical methods in `R` easily meets this criterion. This is not to claim that the computational details are necessarily simple, but the necessary work on this has been carried out by experts, leaving the average end-user free to enjoy the fruits of their labor.

To take full advantage of this an understanding of what any particular method is intended to do helps; that is, is it conceptually simple? If the answer to this is ‘yes’ then, allied to computational simplicity, and from the end-user’s viewpoint, a method may be regarded as ‘simple’. A fundamental thesis of these notes, argued in more detail in individual chapters, is that the statistical methods mostly used in archaeology are simple in this sense.

A simple idea can lead to mathematically complex developments but, as with the computational detail, experts have dealt with this, so it’s not usually necessary for the user to understand the details[1]. An exception to this generalization, as some might view it, is that some understanding of mathematical/statistical notation is desirable for a number of reasons, among them brevity and clarity of expression.

A fine but important distinction to make is that statistics is often simple in the sense described but this does not make it easy. Almost anyone who wishes to (and can afford it) can learn to drive a car without any deep understanding of the technology that makes this possible. The part that is less easy is learning to do this with facility and appropriately, and this only comes with practice.

The archaeological trowel provides another analogy. It’s a simple tool and what you do with it is simply explained; using it well does not, for most diggers, come immediately, effective use comes with practice. Most aspirant diggers can arrive on a site and be on their knees and using a trowel within a short space of time, but will do so under supervision or relegated to an area where they can do no damage. To be realistic, some remain on these margins or are transferred to other duties more suited to their aptitudes but, with practice and motivation, most graduate to a position where they can operate with minimal supervision. Some diggers will be better than others; it’s a shibboleth that ‘practice makes perfect’ but it certainly does no harm[2].

To summarize, employing statistics usefully can start from quite a limited knowledge base. You do eventually need to learn more about methods you find useful, a process of accretion, and be aware of when their use is appropriate. Statistics is no different, in this respect, from other subjects worthy of study; but, contrary to occasional misconceptions, you don’t need to be especially gifted, with

[1] Mathematics is not avoided in these notes but, for the most part, is quarantined so that it can be ignored by the reader if wished. Some appreciation of the mathematical distinction between principal component and factor analysis (Chapters 7. 8 and Appendix D) helps to appreciate the way in which they have been used and confused but, unless you are attracted by what factor analysis appears to offer, the mathematics can be ignored.

[2] When I did this sort of thing seasoned diggers, armed with their 4-inch cast-steel WHS pointing trowels (`R`) would look pityingly on aspirant excavators arriving with inappropriate oversized and over-flexible welded (they break easily) plasterers trowels (`Excel`) and condescendingly point them to the path of righteousness. The overly-seasoned digger would sport a second trowel, discreetly but visibly displayed and well-worn, testifying to their vast experience. These were very functional but equally important as symbols of superiority to be mutely admired by the cognoscenti. Except in the pub when conversation sometimes descended to the ‘mine is smaller that yours’ kind of boasting.

a strangely configured intellect, to engage in productive statistical analysis.

Chapter 2 illustrates by example what is intended by the term ‘simple’. The standard methods of multivariate analysis used in archaeology are each executed in one line of `R` code, and in most cases immediately useful and interpretable graphical output is produced. That’s the computationally simple bit. Conceptually, the methods are intended to reduce a large table of data to a picture, usually two-dimensional and often interpretable as a map showing the relationship between the rows and/or columns of the data table. The output can be examined for archaeological patterns in the data. Also a simple idea.

The simplicity conceals specific differences that distinguish between the methods that are discussed in the individual chapters devoted to them; Chapter 7 for principal component analysis (PCA), Chapter 9 for correspondence analysis (CA), Chapter 10 for cluster analysis and Chapter 11 for linear discriminant analysis (LDA). The chapter on LDA includes a brief treatment of classification trees, an attractive alternative to LDA that is non-linear and computationally intensive, but underpinned by a very simple idea.

It is worth emphasizing the important idea of ‘mapping’ the data in these methods. There’s an analogy with map projections in cartography where different projections can be chosen to emphasize particular features of interest. Methods such as PCA, CA and LDA measure the ‘distance’ between the rows of a data matrix in different ways (i.e. they project the data differently). The choice depends on the nature of the data available and the aim of an analysis. Results are ‘distorted’ in the sense that only a visually accessible *approximation* to the ‘reality’ is obtained, and the quality of approximation needs to be assessed. The niceties of this are discussed in the relevant chapters; other issues that require attention in practice are also covered, such as the choice of data transformation or specific methods (e.g. in cluster analysis) to use, and the many different ways of interrogating the output of an analysis that are available.

Chapter 8 on factor analysis follows that on PCA fairly naturally, since the two methods are often confused. I’ve suggested elsewhere that, despite the ‘historical’ importance of factor analysis in the development of quantitative archaeology it is past its ‘sell-by’ date. That I’ve devoted space to the subject is because of what I read as misleading advocacy of the method in a recent quantitative archaeology text. Anyone who thinks factor analysis might be for them ought to be acquainted with its ‘problematic’ aspects, even if they disagree with my take on the subject.

A systematic discussion of descriptive statistics, such as the mean, median, standard deviation etc. is not attempted. Read (almost) any introductory text on statistics for this kind of thing[3]. Chapters 3 and 4 look very selectively at

[3] Those that shy away from notation can get definitionally confused, particular in distinguishing between population parameters and their estimates – so read more than one text.

some of the more commonly used methods of graphical display for continuous and discrete data. Kernel density estimates, as an alternative to the histogram that is sometimes more useful, is given more space in Chapter 3 than is common in quantitative archaeology texts

Chapter 4 is primarily concerned with bar-charts (or barplots) and pie-charts. The former are sometimes confused with histograms or inappropriately presented as three-dimensional constructs. Pie-charts are often similarly misrepresented and, depite their popularity, a good case can be made for not using them at all. The chapter is more opinionated than I might allow myself if writing a textbook, but the opinions are not uniquely mine.

Chapter 6, which can be thought of as a continuation of Chapters 3 and 4, is a pot-pourri of graphical methods of analysis. Some are fairly standard; some are little-used but deserve more attention; and some are probably neglected for good reasons.

Chapter 5, on regression analysis, covers a topic invariably dealt with in texts on quantitative archaeology. At its simplest – fitting a straight line through a scatter of points – the ideas are straightforward. The chapter strays beyond this into the realms of non-parametric regression, not a topic much mentioned in archaeological textbooks. More notation is needed here than in most other chapters, to distinguish adequately between different models, and between models and their estimates.

I had second thoughts about including Chapter 12, on statistical inference. It's arguable that the development of formal theories of statistical inference was one of the greatest intellectual achievements of the first half of the twentieth century; unarguable that these ideas motivated the promotion of quantitative methodology in archaeology; and arguable that *formal* inferential methods have delivered much less than was originally promised. Quantitative archaeology texts sometimes present the methodology with enthusiasm but I suspect equally often out of a sense of duty. Having included the subject I've been equally dutiful and illustrated the use of `R` with sometimes quite extensive examples, but with added opinion that the reader is free to disagree with.

1.2 How (not) to read these notes

The point made in the preface that this is not intended as a textbook, should be reiterated. It's meant more as a 'dipping into' kind of text – as the use of the term *Notes* in the title indicates. It also indicates that coverage is selective, including methods I have practical experience of and excluding those where this is not the case. Some methods are excluded for other reasons (see below). It's assumed that anyone reading this will have been exposed to a systematic introductory treatment of statistics.

The material might be regarded as a kind of practical supplement to an introductory textbook development. The 'standard' texts for archaeological purposes will be taken to be Shennan (1997) and Drennan (2009) (see Section 1.3). Other than their treatment of sampling methods in archaeology, omitted here, much the same kind of methodology is covered, but with rather different emphases. As already noted some of the commoner descriptive and graphical methods used arise in context, rather than being systematically presented. There is considerably more emphasis on undertaking analysis than to be found elsewhere, with more analyses of real, large and, in some cases, quite complex data sets.

Some topics not usually covered in quantititative archaeology texts, or accorded only brief treatment, are introduced, including kernel density estimates (Chapter 3), non-parametric regression (Chapter 5), fuzzy cluster analysis (Chapter 10), classification trees (Chapter 11) and some of the graphical methods of Chapter 6 . These are present because they are useful and fulfil the criteria for 'simplicity' that is being applied.

In a few sections a more critical view is taken of the way some commonly used methods have been applied. This includes comment on the misuse of bar- and pie-charts (Chapter 4) and the real need for much of the standard inferential methodology of hypothesis tests (Chapter 12). Factor analysis is usually mentioned in the standard texts, but often in a cursory way or, where treated at any length, in a manner I think is unsatisfactory. It will be clear from Chapter 8 that I have my doubts about the value of the method for archaeological purposes and that others would disagree entirely; the chapter and associated Appendix D is possibly one of the fuller *critical* appraisals in the archaeological literature.

Faced with an academic text of any kind my normal practice is to skim it to see what's there then ignore the bits which don't interest me or are not suited to immediate purposes (which may be the entire text). Some experience is necessary to get away with this; teaching texts are usually intended to be read in a linear and systematics fashion, so you have to wade through some fairly boring, if essential, bits to get to the parts that might interest you. I'd never rely on a single text; just because something is published in some form doesn't make it sacrosanct (even if written by your instructor). It's as well to be aware that there's more than one way of tackling a subject, and it can come as surprise to learn that statisticians can have different views, sometimes vehemently expressed, about what does and doesn't constitute valid and useful methodology.

Having said this, a systematic reading of parts of these notes may be useful. Anyone entirely new to `R` would need to start with Appendix A and Section 2.6 and is then advised to do a few data analyses as soon as possible, either reproducing the analyses in the notes (e.g., those in Chapter 2) or using analyses found elsewhere in the notes as templates for looking at your own data. Chapter 7 can be usefully read as a prelude to those which follow, on other multivariate methods, as it includes discussion of data-analytic concerns of general relevance. If it seems like regular use of R is an attractive proposition an early acquaintance with user-defined functions (Section 3.2.1) is helpful.

I've suggested at various points in the notes that doing things first and then worrying about what you've done afterwards is a viable way of learning about R, and the statistics involved. This is not a flippant suggestion. Jump, or ease yourself, in at the shallow end and immerse yourself a little. Any discomfort induced by the temperature will rapidly disappear, and once you are convinced you are still breathing you can think about learning to swim. You may even come to enjoy it.

What has been omitted from these notes could, as they say, fill a book. Sampling methodology has already been mentioned. It's important and more-than-adequately covered in Orton (2000). It's not a topic I've had to engage with in recent years and I'm not sure how useful R might be for acquiring and analyzing data in a realistic archaeological setting.

Spatial analysis is an interesting area also largely ignored. It's fairly astonishing that Hodder and Orton (1976) still seems to be referred to as the standard archaeological text on the subject. Archaeologists don't, of course, ignore spatial analysis but it is, I suspect, largely associated with the use of geographical information systems (GIS). I know almost nothing about GIS so am not competent to write on the subject and probably shouldn't comment either. The subject is undoubtedly important and has been around for over 20 years; it was 'oversold' in its early days and, when I last looked, made surprisingly little effective use of what I'd regard as 'proper' statistics – things have probably changed without impinging on me.

Computer-intensive methodologies are important and will become increasingly so, but are not dealt with systematically. As mentioned in several places a review of the use of such methods, for statistical purposes, in archaeology would be useful, but is not attempted here. Archaeological simulation, thoroughly reviewed recently by Lake (2014), is computationally intensive but makes limited use of statistical methodology as I understand it. I have more in mind resampling methodologies such as bootstrapping, noted at several points in what follows.

A systematic treatment of non-parametric methodology is similary not attempted, though kernel desnity estimates fall into this class of methods. Had more detailed treatment been provided it would have been in Chapter 12, where the methods, can be viewed as more 'robust' alternatives to the hypothesis tests usually discussed in introductory texts. Given they are intended to avoid assumpions that trouble some users of statistics I was surprised, on reviewing the textbook literature, to see how little attention they have received.

Robust methods which, in some senses, are similarly motivated are similarly not much mentioned here. They are designed to minimize the effect on analytical outcomes of the violation of assumptions that underpin 'standard' methodologies, the deleterious impact of 'unusual' data on analyses having received particular attention. More exploration and exploitation of the use of the methods with archaeological data is probably needed to assess their merits. Almost needless to say, several `R` packages are available devoted entirely to robust methods.

It would have been possible to extend the treatment of regression analysis – more generally 'linear models' – to generalized linear models. These include log-linear models for contingency table analysis and logistic regression models. The beauty of the idea is that a lot of different models, some looking very complicated and non-linear, can be shown to have a common underlying structure that admits a common algorithmic approach to their estimation. Conceptually, the more 'complex' models inherit much of the interpretive machinery of linear models and – to the extent that the latter are 'simple' – it might be argued that generalized linear models are also 'simple'. In practice I don't find them as straightforward as other methods covered and have not attempted an exposition. Log-linear models had a brief vogue in archaeology in the 1980s and 1990s but I don't think are now widely used; logistic regression acquired something of a niche role in predictive modeling. Baxter (2003) provides examples of their use.

1.3 Suggested reading

As already mentioned, Shennan (1997) and Drennan (2009) are useful introductions to quantitative archaeology, referenced a lot in what follows. They can be viewed as complementary; both are worth looking at. I've seen nothing published since, in a similar vein, that I'd recommend. Don't neglect introductory statistics texts in subjects other than archaeology; they may suit your 'learning style' better.

In the past I'd recommend students to spend an hour or so browsing texts on the shelves of a university library if they had access to one. I retain an old-fasioned preference for written texts, finding them easier to dip into or exploit in a serendipitous fashion. The modern reality is to make a beeline for a search engine to see what can be found on the web, which can be very good but also highly variable in quality and reliability.

The first editions of Shennan and Drennan appeared in the late-1980s and mid-1990s and the current editions, with a few extra topics, don't differ in their conception. The style reflects their genesis; both authors are aware of the value of computational software but neither devote much space to practical implementation with specific software. This is perfectly understandable; both books were intended to sell and the software access available to purchasers was highly variable. This has remained the case until fairly recently, but the advent and development of R has effected a radical change in the publication of applied statistics texts, as well as practice. It can seem that every other statistics text now published has a title along the lines *Statistics for Something using* R[4].

Archaeology, unless I've missed something, stills need to catch-up with statistical developments. David Carlson's web pages promise a book on R and archaeology, and his notes on the subject provide a useful introduction, designed as R-based computational accompaniments to the texts of Shennan and Drennan[5]. His site also lists some of the free introductory material that is available.

Carlson makes use of the Rcmdr package, a graphical user interface (GUI) to R, designed to make R 'easier' to use and an entrée to the full command-driven version. The development of GUIs for R seems to be something of a growth industry with none having yet achieved market dominance (Valero-Mora and Ledesma, 2012). Within the command- or script-driven paradigm, much of fairly recent origin and likely to grow in importance, are developments that simplify the writing and use of code[6]. That is, R is developing at a rapid pace so any work produced over a period of time – this one included – is going to be 'out-of-date' by the time you see it.

That said, there's nothing wrong with older texts on the subject – the basics haven't changed, just the resources in terms of user-written contributions. As far as R goes, and its 'ancestor' S, I'm 'self-taught' and much of what I know was learned from Venables and Ripley (2002) and earlier editions. Its age notwithstanding, if I could only own one text on R this would be it; I still consult it regularly. In fact I own several; Dalgaard's introduction, now in its second edition (Dalgaard, 2008), was also useful. A number of, to some extent 'competing', texts emphasizing the powerful graphical facilities in R that develop the 'traditional' graphics that come with it are also available (e.g., Murrell, 2011; Sarkar, 2008; Wickham, 2009).

As with introductory statistics texts I think it's useful to own, or at least have access to, more than one treatment of R at introductory and intermediate levels. As I don't believe in the 'one-size-fits-all' principle, and if its not a requirement for a taught course, I'd hesitate to make definitive recommendations; intelligent 'shopping around' to suit your own requirements is what's needed[7].

[4] The site http://www.r-project.org/doc/bib/R-books.html lists about 150 texts on statistics and R, about a sixth at an introductory level. New texts appear on a regular basis and the list does not include free web-based material in plentiful supply.

[5] http://people.tamu.edu/ dcarlson/quant/index.html (Accessed May 2015)

[6] See Ben Marwick's brief review at http://www.r-bloggers.com/doing-quantitative-archaeology-with-open-source-software/ (accessed May 2015) and links provided there.

[7] Apart from anything else, I should confess to not having read a lot of what's available; it's nothing to be ashamed of.

On more specific statistical topics, the multivariate methods of Chapters 7 to 11, are covered in an archaeological context and in what is intended to be an accessible manner in Baxter (1994a) (reprinted as Baxter, 2015). The style of presentation, that is the graphics, shows its age but the basics haven't changed. Baxter (2003) is a wider-ranging, more recent, survey at an intermediate level, of archaeological usage of most of the methods covered in these notes, with the exception of the material in Chapter 12. In particular a fairly detailed review of the use of regression methods in archaeology is provided; Shennan (1997) includes a more conventional and quite lengthy introductory treatment that goes somewhat beyond what is often attempted in introductions, including a brief account of log-linear models avoided here.

More topic-specific reading is suggested in the relevant chapters; there are a lot of good texts out there on specific multivariate methods and regression analysis. As with other recommendations, that you may have noticed I'm not making, I hesitate to be too specific and, once you have gained a toe-hold in a subject of particular interest, read around to see what's there and suits you. Subject-specific statistical texts abound but sometimes need to be approached warily. Some domains of study have developed their own statistical idioms (I have in mind some of the psychological, social and life sciences, among others) that, beyond the introductory level, can result in texts with methodological biases and emphases not readily transferable to or appropriate for other domains. Quantitative archaeology publications are largely exempt from this comment since archaeological use of statistics is, for the most part, within the 'statistical mainstream'[8].

[8] This is notwithstanding the 'philosophical baggage' that came with some of the early promotion of statistics in archaeology, and a later mild 1980s obsession with the need for 'concordance' between archaeological theory and statistical method. The latter is a worthy aim that resulted in a body of sometimes impenetrable writing that reinvented existing statistical wheels in shapes not necessarily fit for purpose. Clive Orton (1992: 137), in the context of methods developed for intra-site spatial analysis, memorably characterized this as 'the Audrey syndrome ("an ill-favoured thing, sir, but mine own". *As You Like It*, Act V, Scene iv' (Baxter, 2003: 7–8). This has more general application – don't ignore statistical publications just because they have been produced by statisticians.

Chapter 2

Introductory examples

2.1 Introduction

The examples illustrate the use of R for implementing four methods commonly applied in quantitative archaeology applications. They are principal component analysis (PCA), correspondence analysis (CA), cluster analysis, and linear discriminant analysis (LDA). These are all examples of multivariate methods and are often considered to be at the 'complex' or 'advanced' end of the spectrum of quantitative methods used in archaeology. They usually feature in introductory treatments towards the end of the text, if at all.

Apart from introducing R the main aim of this chapter is to illustrate how computationally simple it is to implement such 'advanced' methodologies in R. It will also be argued that the ideas involved are easy to understand; this is discussed in more detail in the chapters devoted to each topic, where some of the finer points of application are covered.

As far as ideas go, given a table of data with n rows and p columns the aim of three of the methods discussed (PCA, CA, LDA) is often to produce a two-dimensional 'map' or bivariate plot of the data that shows (approximately) how similar rows are to each other in terms of the distance between them. Details depend on the type of data available and precise purpose of the analysis and are covered in later chapters. Cluster analysis is also a way of visualizing the similarity between the rows of the table, but uses a different form of graphical representation in the form of a tree-diagram or dendrogram.

The aim in general is to investigate structure in the data, not otherwise easily done for 'large' tables with $p > 3$. In order to concentrate on the practicalities of application, the data sets analyzed below have been chosen because their structure is fairly obvious. Complications arise in practice such as the occurrence of unusual data (outliers), the common need for some form of data pre-treatment (eg., data transformation), the absence of clear structure, and so on, dealt with in later examples.

2.2 Example – Principal component analysis

The data in Table B.1, from Tubb *et al.* (1980), consist of $n = 48$ rows and $p = 10$ columns to be referred to as *cases* and *variables*. The first nine columns, the data to be analyzed, are concentrations (%) of oxides in specimens of Romano-British pottery. This defines a 48×9 *data matrix*; the tenth column, coding the region of the kiln site where the pottery was found, is used to label plots.

In the original paper several questions were posed.

- Ignoring 'region' is there evidence of chemical grouping in the data?
- If grouping exists can it be associated with region?
- What variables contribute most to group separation, if groups exist?
- Can a subset of variables describe the data well?

With minor complications, to be discussed, these turn out to be fairly simple to answer. Figure 2.1 is a PCA biplot of standardized data based on the oxides (Section 7.2). A single command line was used to obtain the plot

```
biplot(prcomp(tubb.data, scale = TRUE))
```

where `tubb.data` is the name given to the 48×9 data matrix in `R`. The argument `scale = TRUE` standardizes the data so that all variables have zero mean and unit variance, giving them equal weight. Other options are possible (Section 7.2) but this scaling is often desirable (Venables and Ripley, 2002: 303). The appearance of the plot can be customized using other arguments – it is kept simple here.

In mathematical terminology, the data exist in a 9-dimensional space defined by the number of variables. The distance between cases in nine dimensions can be defined mathematically but can't easily be visualized . What PCA is often used for is to transform the data to new variables – *principal components* (PCs) – such that bivariate plots based on the first two of these *approximate* in two dimensions the distances between cases in nine dimensions. The points labeled 1–48 (the row numbers) identify the cases.

It is readily apparent that, apart from a few stray cases, there are three groups in the data, and is easily shown (see Figure 2.2) that these correspond to the three regions. The reasonably tight group to the left corresponds to Region 3, and the central group to Region 1. A biplot also provides information about

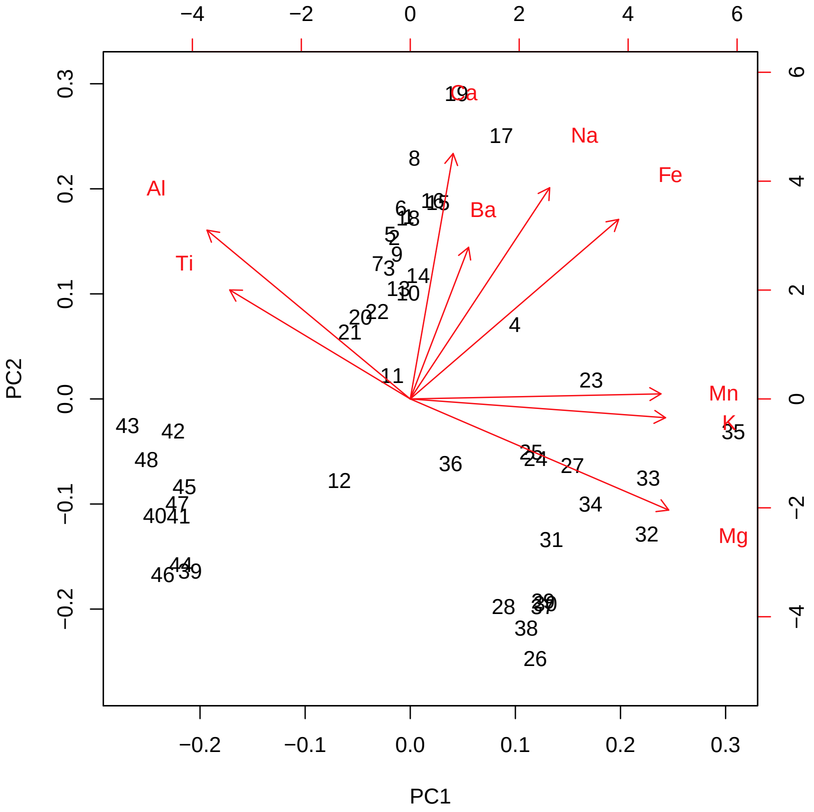

Figure 2.1: *PCA biplot of the standardized chemical data from Table B.1.*

the relationship between the variables. The arrowed lines (vectors) point to the variable markers. Angles between vectors approximate the correlations between variables[1], so we can infer that Al and Ti are strongly positively correlated with each other; negatively correlated with Mg, Mn and K; and poorly correlated with Na. The relative positions of row and column markers suggests that the group to the lower left, compared to other groups, has relatively low values for several variables that plot opposite it. This can, of course, be checked.

The previous code illustrates how simple it is to produce a PCA in a single line of code, and the default output obtained is very informative. There are, however, advantages in breaking the code up, to enhance readability and open up presentational possibilities not otherwise readily available. Thus

[1] Strictly speaking, it's the cosines of angles that approximate correlations.

```
tubb.pca <- prcomp(tubb.data, scale = TRUE)
biplot(tubb.pca)
```

will produce Figure 2.1, but also creates an *object*, `tubb.pca`, that holds information that can be manipulated for presentational and interpretive purposes. This is discussed in detail in Chapter 7, but an illustration of what can be done is shown in Figure 2.2, the code for which is given in Section 2.6. Suppose that only a plot based on cases is required, labeled according to region (remembering that this information is not otherwise used in the PCA). In the figure cases are labeled by number and colored plotting symbols that differentiate between regions.

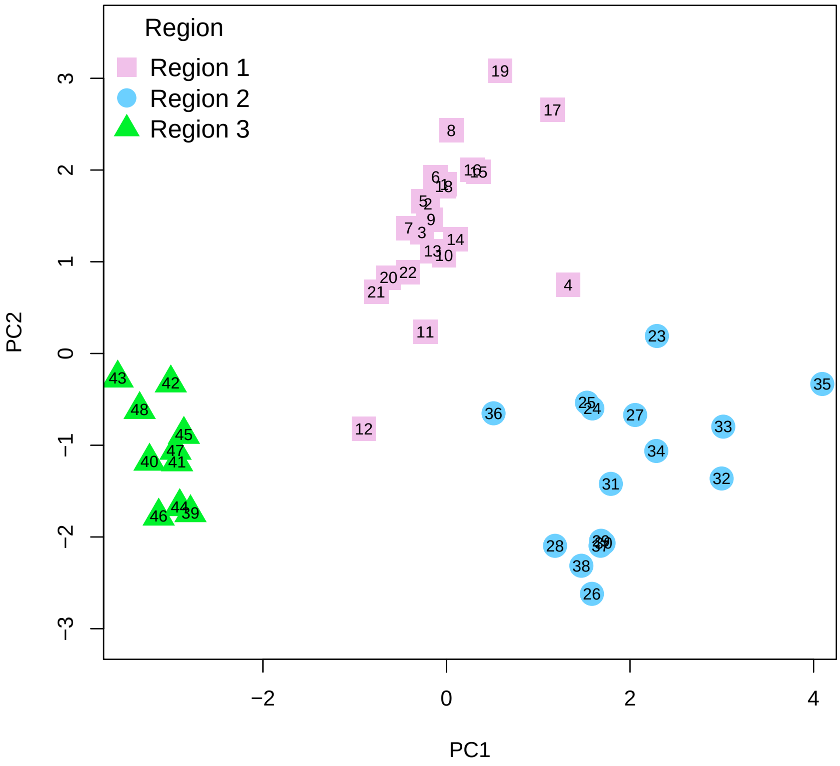

Figure 2.2: *An enhanced PCA row plot of the standardized oxide compositional data from Table B.1.*

An important point to note is that the axes of the plot are equally scaled. This allows the separation between cases to be interpreted as (approximate) distances. Equal scaling is not a feature readily available in some widely used statistical packages. The `biplot` function automatically provides this; for Figure 2.2

the `eqscplot` function produces equal scaling where this is not automatic (Section 2.6.3).

It is as easy to look at these kind of data using PCA as it is to use 'conventional' and 'simpler' graphical methods (which should not be ignored). If needed the PCA can be revisited and refined. Simply scanning the table of data can be informative. This is best done from hard copy, rather than looking at data on a terminal. Looking at the columns of Table B.1 reveals some obvious outliers, almost certainly typos. These are cases 4 (Ti = 0.03), 35 (Mn = 0.394) and 36 (K = 0.81). There is also some evidence of outliers within regions.

In Figure 2.1 the outliers stand a bit apart from the regional groupings to which they belong, with the suggestion of a small tight sub-group in Region 2 to the bottom right. Case 12 seems isolated relative to its regional group, but a chemical reason exists for this (not discussed here). Only a small number of variables are needed to show regional differences. Lacking regional information, sensible univariate or bivariate data inspection would reveal clusters. Two bivariate plots are shown in Figure 2.3 where a clear outlier for the variable K is obvious[2].

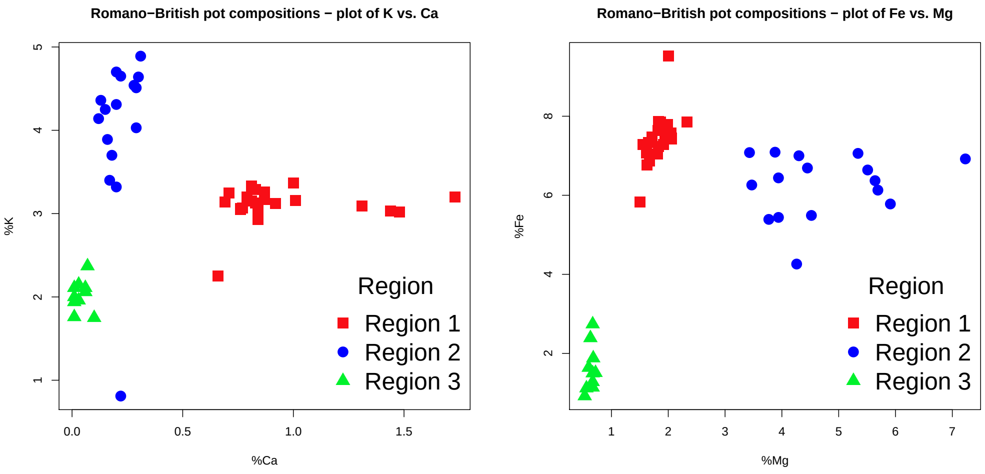

Figure 2.3: *Bivariate plots for selected variables from Table B.1. See the text for an explanation.*

The bivariate plots show that regional clusters are chemically distinct, that only two variables are needed to show this, and that these can be chosen in more than one way. The variable choice can be undertaken in several ways, but inspection

[2] Scatterplot matrices, or pairs plots, can be used to produce displays of all possible bivariate plots. These can be inspected to identify interesting pairs of variables that can be separately plotted in more detail. The `pairs` (Section 5.2) or `scatterplotMatrix` (Section 6.2) functions are available for this kind of analysis.

of Table B.1 is all that is needed here. For example, values for Mg are different between regions; those for Fe separate Region 3 from the other two regions; Ca separates Region 1 from the other two; and so on. In terms of the questions posed at the outset of this analysis intelligent inspection of Table B.1 may be all that is needed, which is what was meant when stated earlier that the data were 'easy to analyze'.

Once groups are established, identifying how they differ is of interest. This is often done by presenting a table of means, standard deviations and other summary statistics. The (arithmetic) `mean` is a *measure of location* usually used with the idea that it is, in some sense, 'typical' of the data. The `median` is an alternative if there are obvious outliers in the data. Both are inappropriate as a measure of what is typical if there is are clear sub-groups within those being summarized. That is, using a simple statistic such as the mean, as with the choices made in PCA, requires a consideration of the validity of the assumptions involved. The `standard deviation` is a `measure of dispersion` or spread often used in conjuction with the mean; the `interquartile range` (IQR) is a measure of dispersion often associated with the median. Table 2.1 shows various measures of typicality and spread for all the data for K, and for the regions. The clear regional differences, and that for other variables, could also be inferred from the biplot of Figure 2.1. The measures of dispersion are greater for Region 2 than the other two regions.

Region	Mean	Mean-	Median	SD	SD-	IQR
Region 1	3.11	-	3.13	0.22	-	0.15
Region 2	4.01	4.22	4.28	0.97	0.48	0.72
Region 3	2.02	-	2.03	0.19	-	0.17
All	3.18	3.22	3.16	0.92	0.86	0.96

Table 2.1: *Summary statistics for K from Table B.1. Mean- and SD- indicate that an outlier, case 36, has been omitted.*

2.3 Example – Correspondence analysis

It is not mandatory, but correspondence analysis (CA) is usually presented as a method appropriate for analyzing two-way cross-tabulations of categorical variables. This includes the special case where the data are recorded as presence or absence, coded as 0–1 (e.g., artifact types and contexts). At the mathematical level there are differences between CA and PCA (Section 9.2) but fundamentally their aims are identical, which is to display the approximate distances between rows, and in the case of CA possibly columns as well, in a 2-dimensional map.

To emphasize the difference in variable types the notation $I \times J$ will be used for the table of data to be analyzed, I and J being the number of categories for the two variables. The data as a whole are often displayed as a biplot - more commonly than with PCA where the emphasis is often on cases. This is not necessarily true of CA, and the roles of the variables can be reversed and a $J \times I$ table analyzed.

A common use of CA is for *seriation* (e.g., Madsen, 1988a; Section 9.5) where it is hoped that an ordering of the rows can be inferred from the row plot and that the order has a chronological interpretation. As an illustration of this kind of application data from McLellan (1979), in a study of the chronology of 'Philistine' burials, are used (Table B.2). Columns in the table correspond to tombs, and rows to counts of 52 different types of pottery found in the tombs. The main interest was on seriating the tombs, with a particular interest in sequencing tombs g to j. A 'horseshoe' shaped pattern to the plots is usually expected in a successful seriation, and the ordering is read around the horseshoe. Archaeological criteria are needed to determine the early and late ends of the sequence. The data have been used in Baxter (2003: 136–8) to illustrate the use of CA for seriation; other purposes to which CA can be put are illustrated in Chapter 9.

A biplot is obtained in a nearly identical fashion to the PCA biplot. With the data held in `burial.data` the `MASS` package needs to be loaded using

```
library(MASS)
```

then

```
biplot(corresp(burial.data, nf = 2))
```

does it. The `corresp` function from the `MASS` package has been used. This needs to be loaded using the `library` function (see Section A.3 for details). The other difference from the PCA analysis is the need for the argument `nf = 2` which specifies the number of 'components' to extract. Two are needed for bivariate plotting with the first two usual, though others can be extracted if wished.

While useful for initial exploratory purposes the resultant biplot is often too crowded to be easily read if there is a large number of rows and/or columns, and it is not shown here. For display purposes separate plots for the rows and columns, presented adjacently, are often to be preferred (see Sections 7.2 and 9.2 for a discussion of plotting issues). This is easily done; for example

```
biplot(corresp(burial, nf = 2), xlabs = rep("", 52))
```

suppresses the printing of labels for the rows. The `rep` function produces blank labels (using `""` for the 52 rows). Column labeling may be suppressed in a similar way using the `ylabs` argument[3].

Baxter's (2003) Figure 11.1 was produced in the manner just described and revealed an outlying tomb `"c"`, with somewhat larger numbers for four artifact types than other tombs, that cramped the rest of the plot, making it harder to read. As with the PCA analysis there are advantages to breaking down the computations. Specifically, `corresp(burial.data, nf = 2)` creates a CA object allowing access to information more easily manipulated for plotting purposes. This has been done, in much the same way as needed for Figures 2.2 and 2.3 to obtain Figure 2.4, where the outlying tomb has been omitted.

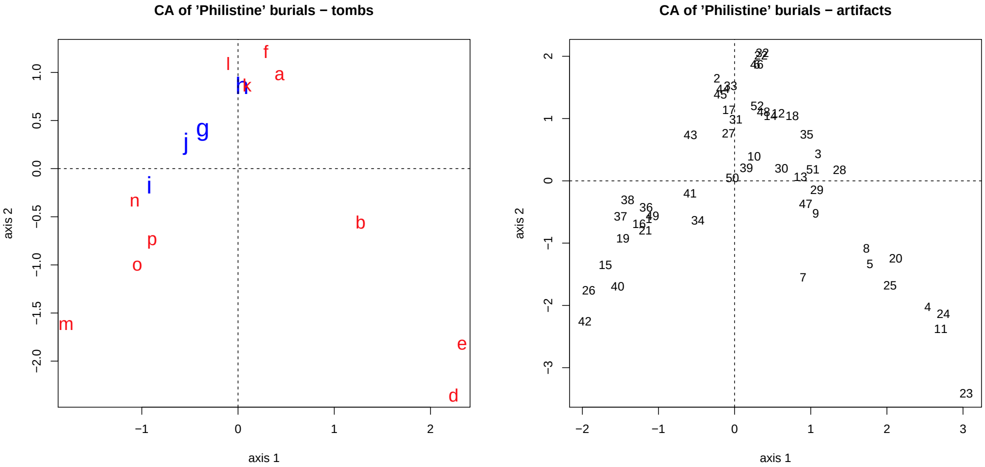

Figure 2.4: *Correspondence analysis plots of the data from Table B.2.*

The plot for tombs shows a good seriation of the data, reflected for the most part in that for the artifact types. Tombs `"g"` to `"j"`, which were of particular interest, are highlighted in blue and with a larger typeface. Their chronological ordering as inferred from the seriation corresponds to that hypothesized by McLellan (1979) using archaeological criteria.

2.4 Example – Cluster analysis

Cluster analysis is a generic term for a wide range of methods that can be implemented in different ways. Given an $n \times p$ matrix of continuous data – the situation assumed here – the common aim is to produce groups, or clusters, of cases such

[3] Several packages contain functions for undertaking CA, which differ in the defaults and the way plots can be labeled. The `ca` function from the `ca` package is used for some of the examples in Chapter 9. In addition to being loaded this package also needs to be imported; see Section A.3 for details.

that cases within a cluster have similar profiles that are distinct from those of clusters.

Obtaining a cluster analysis is straightforward; interpretation not always so. Some of the issues to be taken further in Chapter 10 are raised by the introductory examples presented here. Coding can be reduced to one line as follows.

```
plot(hclust(dist(scale(tubb.data)), method = "ward.D"))
```

The result is shown in Figure 2.5

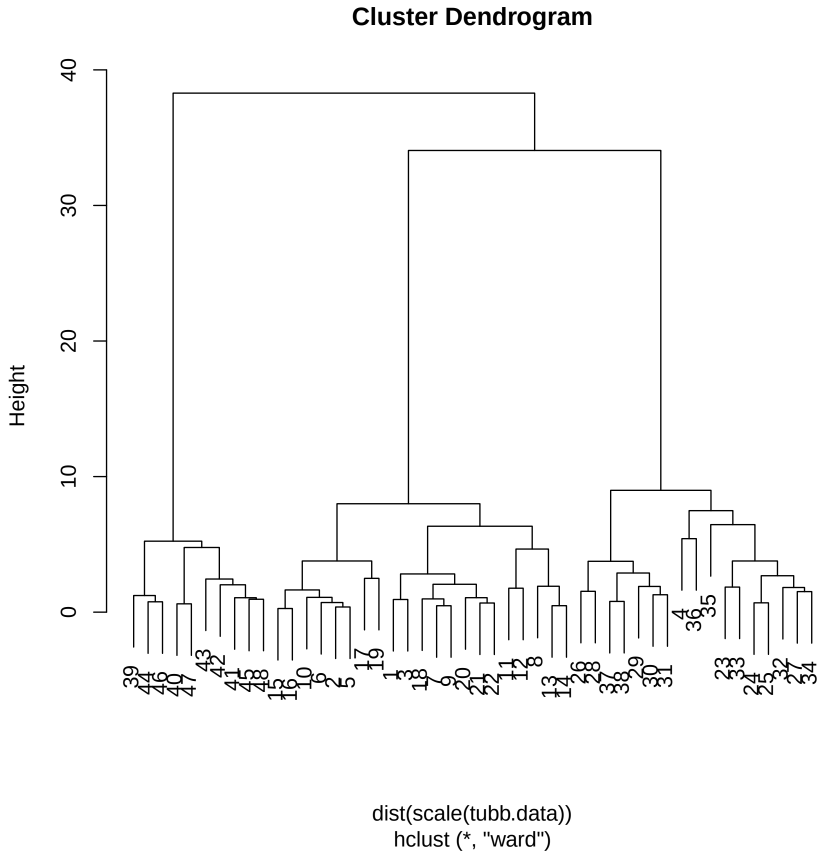

Figure 2.5: *The dendrogram for a Ward's method cluster analysis of the standardized oxide data from Table B.1. This is the default output*

The output produced is a tree-diagram (or *dendrogram*) that can be thought of as consisting of branches and leaves (corresponding to the cases). The idea is to cut the tree at some point to isolate distinct branches whose leaves define the clusters. There are, fairly obviously, three clear clusters in the figure that can be shown to correspond to the regions. Such identification is not usually as easy when there is less clear structure in the data.

There are three main steps in the methods of cluster analysis most widely used in archaeology. Firstly, the data usually needs to be transformed and this involves the same issues as in PCA (Section 7.2). Here the `scale` function standardizes the data to zero mean and unit variance. Secondly, a measure of (dis)similarity between rows needs to be defined and the `dist` function produces, by default, Euclidean distance[4]. Other choices are possible, but Euclidean distance is much the most common (Section 7.3). Finally, and this is where issues of interpretation arise, a method, or algorithm, for clustering the data needs to be chosen. Hierarchical clustering is common. Cases are initially treated as single clusters and successively merged until a single cluster, of all cases, results. The `hclust` function effects the clustering; there are different algorithms for this that depend on the criterion for merging clusters. The `method = "ward.D"` argument to `hclust` specifies that *Ward's method* is to be used (Section 10.3).

The choice of methods is simply effected using the `method` argument. Choices other than `"ward.D"`, such as `"s"` or `"a"`, produce *single-link* and *average-link* analyses of those illustrated in detail in Chapter 10. These will produce different output from Ward's method, and single-link is illustrated in Figure 2.6. Rather than simply replacing `"ward.D"` with `"s"` the opportunity is used to illustrate the use of the `plot` function, used in conjunction with `hclust`, to show how the default labeling can be 'tidied-up'. The code used is given below.

```
plot(hclust(dist(scale(tubb.data)), method = "s"),
labels = tubb.region, sub = " ", xlab = " ", cex = 0.8,
main = "Single-linkage cluster analysis - Romano-British pot compositions")
```

The `sub = " "` and `xlab = " "` arguments replace the labeling at the bottom of Figure 2.6 with blank space. Replacement text could be added if wished, with the `main` argument showing how a more informative title can be produced. The `labels` argument replaces the default labeling of leaves by row number with regional identifications held in `tubb.region` which needs to be created in advance of analysis. Finally the argument `cex = 0.8` controls the character expansion of the leaf labels, in this case to 0.8 of the default, to remove overlapping.

As far as interpretation goes the appearance of the dendrograms in the two analyses differs, but it is clear that there are three main groups in the single-link analysis, with the revised labeling making it clear that these are regional groups. the main difference is that single-link suggests three outliers (two from Region 2 and one from Region 1) not evident in the Ward's method analysis. This is fairly

[4] Mathematically this is just a generalization of distance as we measure it in two or three dimensions.

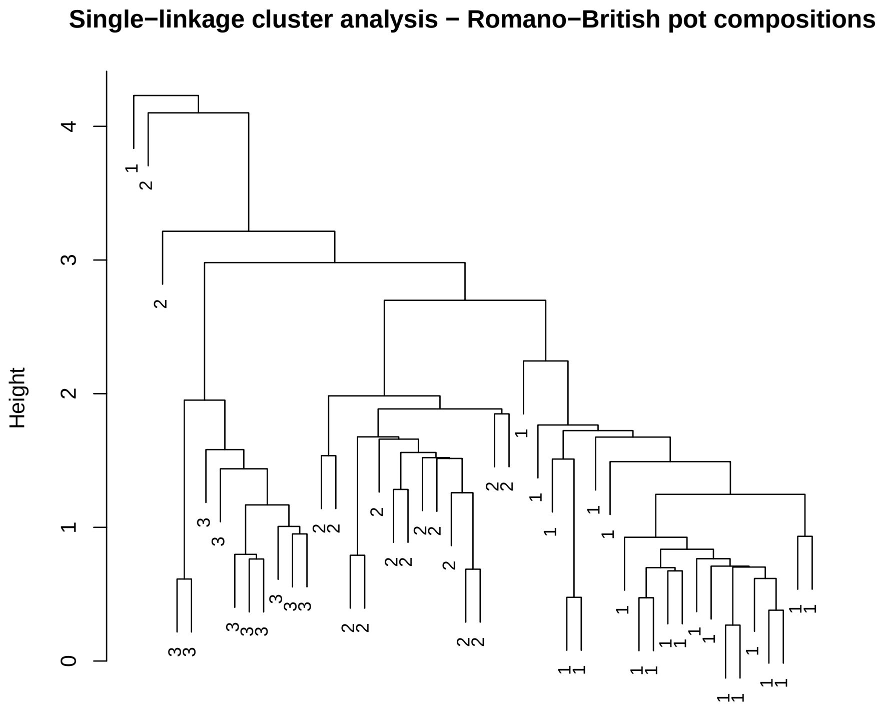

Figure 2.6: *A Single-link cluster analysis of the standardized oxide data from Table B.1.*

characteristic of the way these methods can differ; issues of method choice, cluster validation, and the comparison of analyses are considered in detail in Chapter 10.

2.5 Example – Linear discriminant analysis

The fundamental difference between LDA and PCA is that the former uses the information (or assumptions) about groups in the data in the analysis and, it is hoped, will show much better separation between groups than in the PCA. For what follows it is necessary to load the `MASS` package using `library(MASS)` in order to access the `eqscplot` function, which produces equal scaling of the axes and the `lda` function to carry out the LDA. Once this is done the following one-line command will produce the output of Figure 2.7.

```
eqscplot(predict(lda(tubb.data, tubb.region), dim = 2)$x)
```

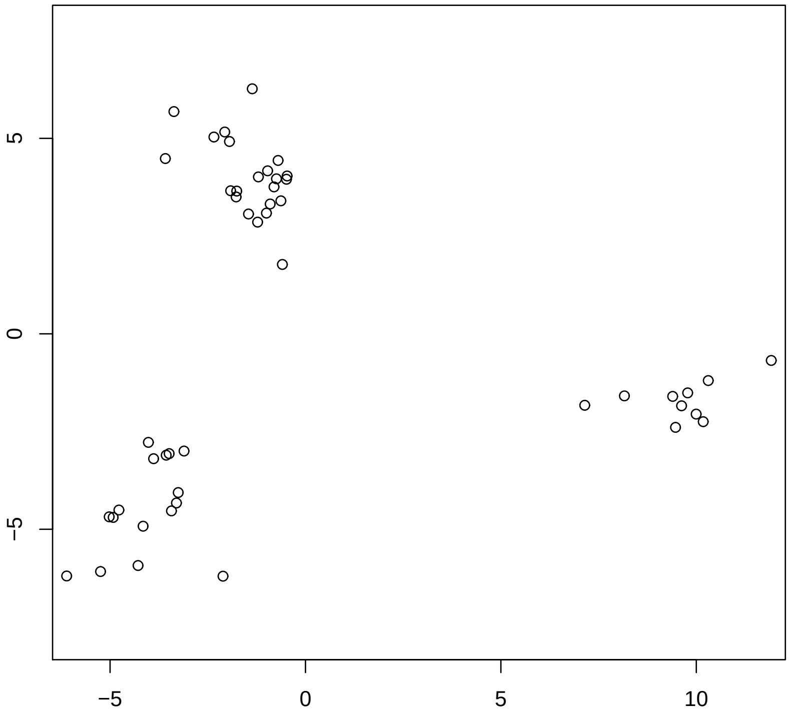

Figure 2.7: *A linear discriminant analysis plot of the data from Table B.1.*

The `lda` function carries out the LDA. The `predict` function with the argument `dim = 2` generates the scores for cases on the first two discriminant functions and the addition of `$x` extracts these for plotting purposes. To use the `eqscplot` function as shown the argument `dim = 2` is needed; the more general case is discussed in Section 2.6.2. More so than the other methods discussed the code benefits from being broken down into more than one line. The plot produced is not entirely satisfactory and would benefit from enhancement. The message is that there are three groups in the data with the separation between them much more evident than for the PCA in Figure 2.2 and this is what we hope to see. It can be assumed, given previous analyses, that the grouping displayed almost certainly corresponds to the regions, but for more general purposes, when the grouping is less clear-cut, we would like as a minimum to label the points to see what group a case is supposed to belong to. How this can be done, along with other aspects of labeling, is discussed at length in Section 2.6.2 as it introduces features of `R` used throughout these notes.

2.6 R notes

2.6.1 Introduction

The chapter concludes, as do other chapters, with notes on the R coding used to obtain the analyses in the chapter where these introduce new features of R. Some features are used on a regular basis and are covered here in the following sub-section to avoid repetition. This includes labeling and presentational options. Some other general features of R that are used regularly are covered in context in other sections. In particular, Section 3.2.2 introduces the idea of user-defined functions, not needed for the present chapter; Appendix A discusses aspects of data entry and accessing user-written packages in more detail than given here.

2.6.2 Aspects of labeling and presentation

The one-line coding used to generate the output for the examples illustrates the ease with which 'advanced' analyses can be undertaken. The plots obtained are useful for initial exploratory purposes; from the point of view of presentation and interpretation some form of enhancement is desirable and sometimes essential. The least satisfactory of the default ouputs presented in the examples was that for LDA, where the regional information is essential for plot construction and ought to be included in the plot in some way. Code for an enhanced version of Figure 2.7, with the associated output, is given below as a peg on which to hang a discussion of aspects of labeling. Without further ado an enhanced version of Figure 2.7 is shown in Figure 2.8 that uses the following code[5].

```
library(MASS)
tubb.lda <- lda(tubb.data, tubb.region)
tubb.ld <- predict(tubb.lda)$x
x1 <- tubb.ld[,1]
x2 <- tubb.ld[,2]

eqscplot(x1, x2, xlab = "first linear discriminant",
ylab = "second linear discriminant", col = Coltubb, pch = Symtubb,
main = "Enhanced LDA - Romano-British pot compositions",
cex = 1.3, cex.axis = 1.2, cex.lab = 1.3, cex.main = 1.3)

legend("topright", c("Region 1", "Region 2", "Region 3"),
col = c("red", "blue", "green2"), pch = c(15,16,17),
bty = "n", title = "Region", cex = 1.4)
```

[5] For display on the page individual directives that can occupy one line if typed at a terminal are sometimes run over two or more lines. Blank lines are used to make it clear where this is occurring. The # symbol comments out what follows it and can be used for annotation.

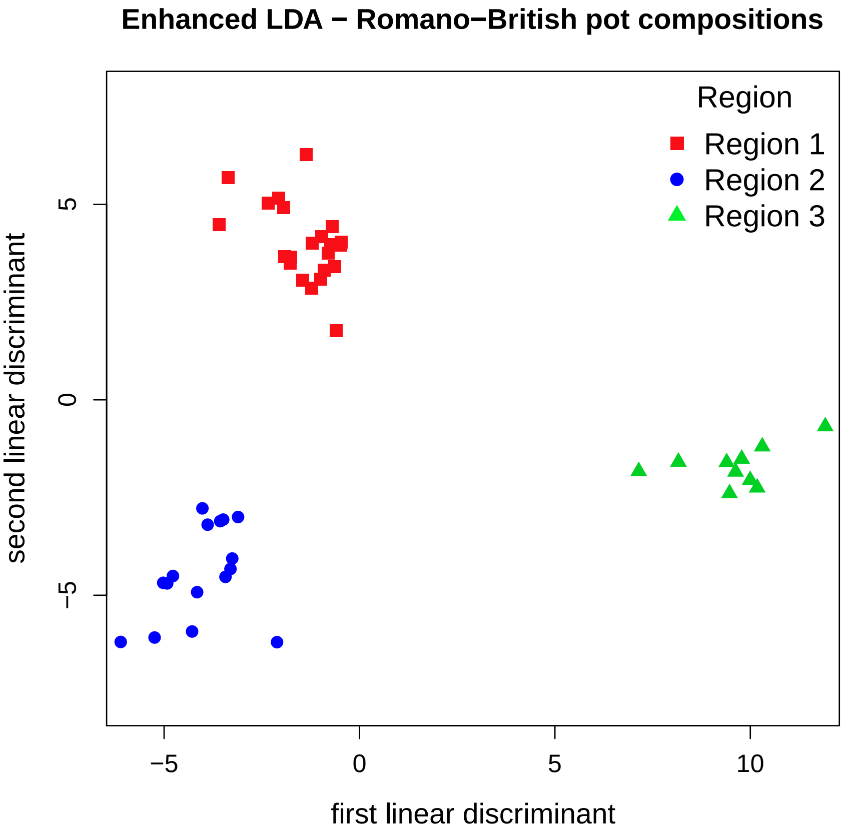

Figure 2.8: *Enhanced linear discriminant analysis plot of the data from Table B.1.*

The original one-line coding has been broken up to make it clearer, and there are slight modifications. The `dim` argument has been omitted from the `predict` function so that `tubb.id` holds information on all the discriminant functions; this requires a slightly different approach from that previously used for plotting. The first two discriminant functions are defined by `x1` and `x2` and these are used in the plotting function[6]. For this particular code to 'work' at least three groups are needed.

The more obvious differences from Figure 2.7 lie in the use of axis labels and titles, the use of different symbols and colors for labeling points, and the provision of a legend. Given the ubiquity of their usage in later examples they are discussed

[6] Note the use of `x1 = tubb.id[ , 1]` etc. to 'pick-out' the discriminant function to use. The `[ , ]` component is used to identify the rows and colums that are included or omitted. Thus, `x[-4, 1:2]` would extract the first two discriminant function omitting case 4. More complicated uses are illustrated in other chapters.

in turn, in some detail.

Labels and titles

Axis labels are produced using the `xlab` and `ylab` arguments in `plot` with the title supplied by the `main` argument. Note that, here and elsewhere, the text must be enclosed in double quotation marks, `" "`. If the content is left blank the axis labels and/or title are also blank. This can be useful for suppressing default labeling. The arguments `cex`, `cex.axis`, `cex.lab` and `cex.main` control the size of the plotting symbols, the size of the numbers on the axes, the size of the axis labels and the size of the main title. What is appropriate will depend on how the output will eventually be presented.

Plotting symbols

In general, points can be plotted using different symbols for individual cases and these are supplied by the plotting character argument, `pch` in the `plot` function (or `eqscplot` if equally scaled axes are needed). The default is to plot using open circles that are otherwise undifferentiated. The argument `pch = Symtubb` was used in the `eqscplot` function above. `Symtubb` is a list of plotting symbols that needs to be created, identified by numbers, of the same length as n. There are 22, 16 and 10 cases for the three regions and the rows are blocked by region. If `Symtubb` is defined as

```
Symtubb <- c(rep(15, 22), rep(16, 16), rep(17, 10))
```

this does the job, where 15, 16 and 17 correspond to solid squares, circles and triangles. The `rep` function replicates its first argument, the second argument defining the number of copies needed. Thus `rep(15, 22)` produces 22 replicates of the number 15; the function `c(...)` combines the arguments given in ... into a list of length 48. The available plotting symbols can be located using judicious Googling and are listed on several websites[7]. They are shown, for convenience of reference, in Figure 2.9. In the code given the character expansion `cex = 1.3` is used to produce more 'visible' symbols on the page than the default.

If only a single plotting symbol is needed, different from the default (e.g., solid circles) then `pch = 16` is sufficient. The graphs in these notes mostly use the symbols 15–17 if three or fewer are needed.

[7] http://research.stowers-institute.org/efg/R/Color/Chart/index.htm for one source for symbols and colors.

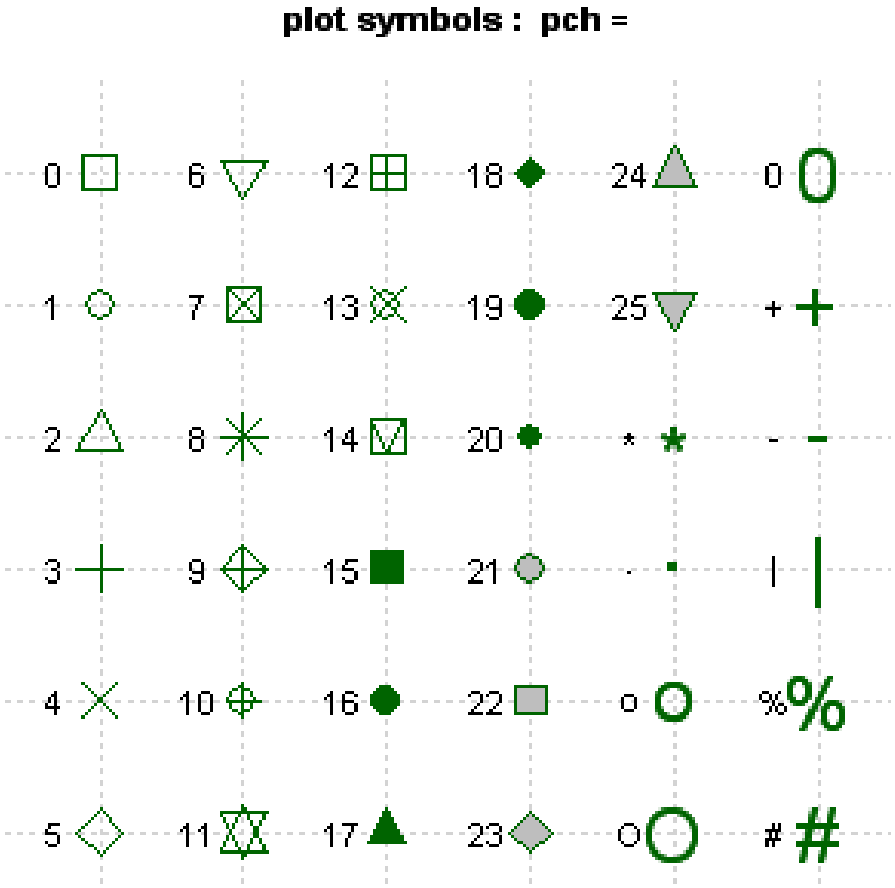

Figure 2.9: *Available symbols for plotting points in* R*.*

Colors

Colors are treated in much the same way as plotting symbols using the `col` argument to the plotting function. The argument `col = Coltubb` specifies the colors used in the example, where `Coltubb` has been defined as

```
  Coltubb <- c(rep("red", 22), rep("blue", 16), rep("green2",10))
```

A list of available colors can be obtained in R using the functions `colors()` or `colours()`; there are 657 in total. The colors `"red"`, `"blue"` and `"green2"` are numbers 552, 26 and 257 in the list. In the definition of `Coltubb` the text identifiers can be replaced with the numbers as follows, and if wished.

```
Coltubb2 <- c(rep(colors()[552], 22), rep(colors()[26], 16),
               rep(colors()[257],10))
```

This is not as 'neat' as the previous construction. It requires the numeric identifiers and these can be obtained from the listing produced by `colors`. An alternative is to use Figure 2.10 (from the same source as Figure 2.9) to identify color labels that seem suited to the purpose. Apart from displaying the colors on offer this is

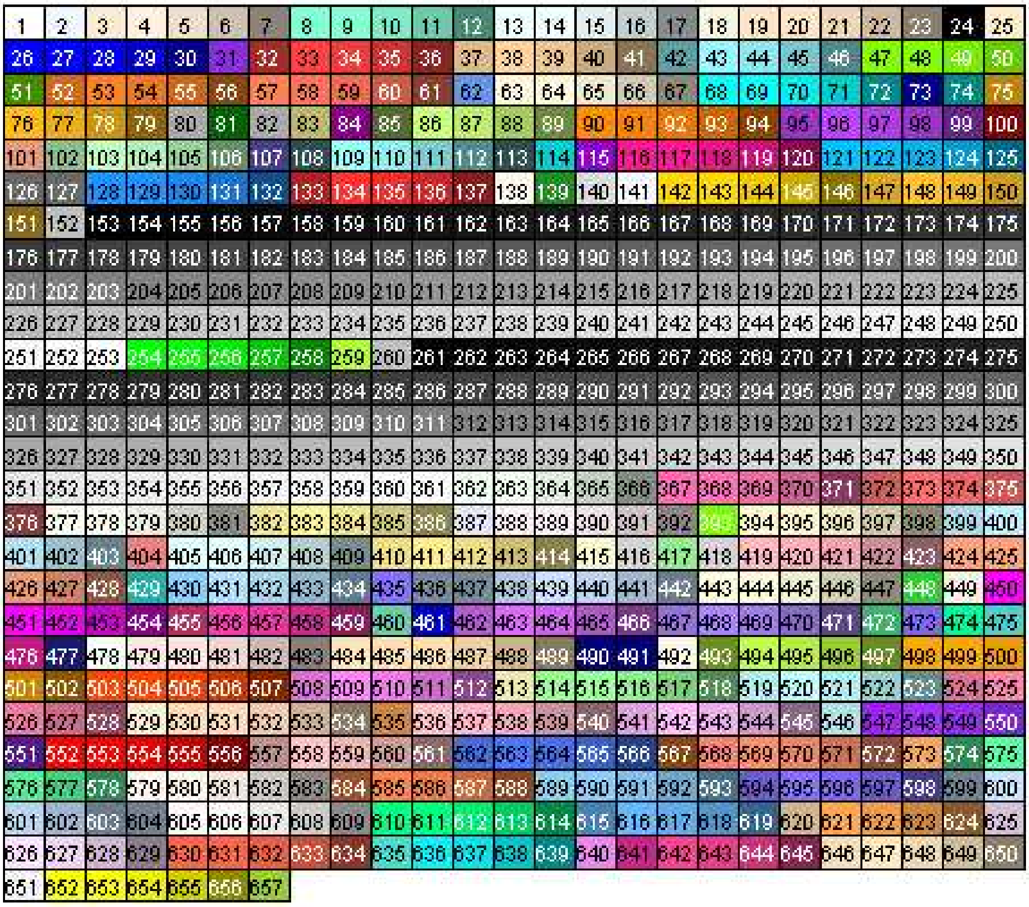

Figure 2.10: *Available colors for plotting in* R.

potentially useful for identifying adequately contrasting colors where several are needed – not always straightforward. In these notes, when two or three or colors are needed `"red"`, `"blue"` and `"green2"` are most used, the last tending to display more satisfactorily than `"green"` in many examples. For a single color something like `col = "blue"` will suffice.

Lines

Lines are not needed for Figure 2.8 but it is convenient to discuss them here. The `lty` and `lwd` arguments control the line type and width. The available line types are shown in Figure 2.11, from the same source as Figures 2.9 and 2.10. The default is `lty = 1`, a solid line, with `lwd = 1`; line color can be controlled using the `col` argument. Lines can be added to plots using the `abline` and `lines` functions, for which the same control is available.

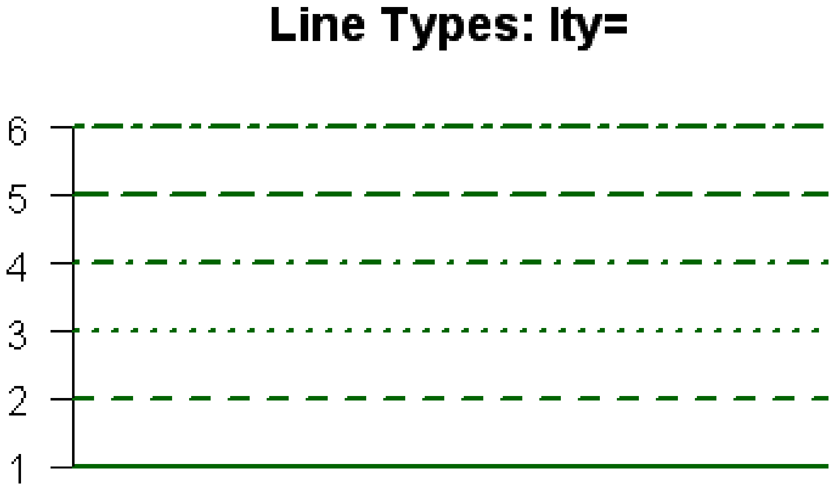

Figure 2.11: *Available lines for plotting in* `R`.

Legends

On first acquaintance the `legend` function can look quite forbidding – see `?legend`. Figure 2.8 is a fairly minimalist example that illustrates some of the more useful features.

In the first argument to the function given in the example code the position of the legend is specified. This can be done by providing the coordinates for the top left-hand corner of the 'box' that encloses the legend, but it is simpler, if space and aesthetics permit, to place the legend in one of the four corners of the figure using one of `"topright"`, as in the example, `"bottomleft"` etc.

The second argument lists the names to be used in the legend. The `col` and `pch` arguments are lists with the same length as that of the legend that indicate the colors and plotting characters associated with each component of the legend; `cex` controls the character expansion used in the legend with `pt.cex` available for addition control over point sizes. The default is to have a visible box enclosing the legend; `bty = "n"`, as used here, renders this invisible. A title for the legend can be added using the `title` argument. Other features will be introduced, when needed, in later examples.

The code for the examples in this chapter will be provided in detail, where this is not given in the text. In subsequent chapters arguments associated with the labeling and the legend will often be omitted unless it is useful to illustrate features not previously discussed. Thus the code for the example used here might be presented as

```
library(MASS)
tubb.lda <- lda(tubb.data, tubb.region)
tubb.ld <- predict(tubb.lda)$x
x1 <- tubb.ld[,1]
x2 <- tubb.ld[,2]
eqscplot(x1, x2)
```

2.6.3 Code used for analyses in the text

Figure 2.2

The main aim in this plot, apart from serving as a first illustration of plot construction, was to show the disposition of cases corresponding to different regions, while simultaneously labeling points by case number so that potential outliers could be identified. Additionally, the plotting of variable markers provided by the `biplot` default was not of interest, and greater control over labeling than that easily possible with `biplot` was needed.

```
library(MASS)
Coltubb <- c(rep("pink", 22), rep("skyblue", 16), rep("green2",10))
Symtubb <- c(rep(15, 22), rep(16, 16), rep(17, 10))
tubb.pca <- prcomp(tubb.data, scale = TRUE)
tubb.x <- tubb.pca$x
x1 <- tubb.x[,1]; x2 <- tubb.x[,2]

eqscplot(x1, x2, col = Coltubb, pch = Symtubb, xlab = "PC1",
ylab = "PC2", cex = 2.5)

text(x1, x2, 1:dim(tubb.x)[1], cex = 0.75)

legend("topleft", c("Region 1", "Region 2", "Region 3"),
col = c("pink", "skyblue", "green2"), pch = c(15, 16, 17),
title = "Region", bty = "n", cex = 1.2, pt.cex = 2)
```

The `prcomp` function carries out the PCA with the first argument specifying the data to be used, and the `scale = TRUE` argument standardizing the data.

Principal component scores are extracted using `tubb.pca$x` and stored in `tubb.x`. Next, `x1` and `x2` are defined to hold the scores for the first two components needed for plotting. The commands to generate these have been run together on the same line with a semi-colon needed to separate them.

The structure of the `eqscplot` function is as previously illustrated but `Coltubb` is constructed with lighter colors that the previous example, so that case numbers are more easily read when superimposed on them. The `text` function adds the case numbers to the plot. The first two arguments are the variables used and are the same as those used with the `eqscplot` function. The third argument supplies the labels to be used. In this example the function `dim` is the dimension of the data matrix, `tubb.x`, that holds the PC scores. It consists of two elements giving n and p, the number of rows and columns of the data matrix. The first of these is what we need and `dim(tubb.x)[1]` extracts it. It is known that $n = 48$ so the effect is to define the third argument as 1:48 which is a quick way of generating the numbers $(1, 2, \ldots, 48)$. Using 1:48 directly is simpler but less general.

The default text color is `"black"` but it can be changed using the `col` argument if preferred. The `cex` values vary between the `eqscplot` and `text` function. The idea is to arrange things so that the text fits within the symbols, and some experimentation was needed to obtain the effect shown.

Table 2.1

```
K <- tubb.data$K   # tubb.data[ , 6] could also be used
# Create new data omitting an outlier, case 36
K_Out <- K[-36]
tubb.region_Out <- tubb.region[-36]
m <- mean(K)
med <-median(K)
sd <- sd(K) # standard deviation
IQR <- IQR(K) # Inter-Quartile Range
statistics <- c(m, med, sd, IQR)
print(round(statistics, 2))
```

The code can be streamlined by writing it as a function (Section 3.2.2) but is adequate for immediate illustrative purposes. The variable K has a column heading of the same name and is the sixth variable in the data matrix; it can be extracted in either of the ways indicated. The code as given obtains the mean, median, standard deviation and interquartile range (IQR) of the data. The `print` function takes whatever is listed and returns it on the terminal. The `round` function takes the first argument – in this case `statistics`, a list of the summary statistics – and rounds it to the number of decimal places given by the second argument.

To do calculations for the regions replace `K` with `K[tubb.region == 1]`, for example, which extracts the data for Region 1 – note the use of square brackets and the 'double equal' symbol, `==`, to define the subset to be extracted. This can be done for each region in turn.

Figure 2.3 reveals a clear outlier for K in Region 2, and it is easily established that this is case 36 in the data table. The variables `K_Out` and `tubb.region_Out` remove this outlier from the data and regional classification. Only Region 2 is affected by the outlier and only the mean and standard deviation are of interest (the median and IQR will not be much affected, if at all, by the outlier). The mean, for example, can then be obtained by using

$$\mathtt{m} < -\mathtt{mean}(\mathtt{K\_Out}[\mathtt{tubb.region\_Out} == 2])$$

.

Figure 2.4

The data of Table B.2, named `burial` here, was analyzed in Baxter (2003: 137) where the third tomb was something of an outlier, plotting sensibly but cramping the rest of the display. In what follows, and for clarity of graphical presentation, `burial1` omits this tomb. The left-hand plot in the figure shows a seriation of the data, and the aim was to highlight tombs `"g"` to `"j"` in which there was a particular interest. This was done by modifying the plotting color and size of the labels for these tombs and introduces features not previously used.

```
library(MASS)     # Needed for corresp and eqscplot
burial1 <- burial[ ,-3]     # Omit tomb "c"
z <- corresp(burial1, nf = 2)

# Extract row and column coordinates for plotting purposes
x1 <- z$rscore[,1]; x2 <- z$rscore[,2]
y1 <- z$cscore[,1]; y2 <- z$cscore[,2]

# Row (artifact) plot
eqscplot(x1, x2, type = "n", xlab = "axis 1", ylab = "axis 2",
main = "CA of 'Philistine' burials - artifacts")

text(x1,x2, 1:52)
abline(h = 0, lty = 2); abline(v = 0, lty = 2)

# Column (tombs) plot
Labs <- letters[-c(3, 17:26)]    # Define text for point labels.
```

```
eqscplot(y1, y2, type = "n", xlab = "axis 1", ylab = "axis 2",
main = "CA of 'Philistine' burials - tombs")

Cex <- rep(1.5, 15); Cex[6:9] <- 2
Colburial <- rep("red", 15); Colburial[6:9] <- "blue"
text(y1,y2, Labs, cex = Cex, col = Colburial)
abline(h = 0, lty = 2); abline(v = 0, lty = 2)
```

Once the coordinates are extracted, as indicated, obtaining the row plot is straightforward. In both `eqscplot` commands, which produces equal scaling of the axes, the argument `type = "n"` produces a blank plot with the axis and titles shown. The `text` function adds text (the point labels) to the plots. For the rows the third argument, `1:52`, defines the row numbers 1 to 52. Labeling for the column plot is a little more involved.

The variables `Cex` and `Colburial` define the character expansion and color to be used, and are then modified to produce a larger size with a different color for tombs (columns) `"g"` to `"j"` (6 to 9 or `6:9` in `burial1`). Note that `cex = Cex` is now specified as has been previously the format for colors and symbols as it is now a variable quantity. Labels for points are defined by `Labs` and are the tomb identifiers `"a"` to `"p"`, omitting `"c"`. These are generated using the `letters` function which generates the 26 lower-case letters of the Roman alphabet (use `LETTERS` for uppercase). The letters from `"q"` onwards are not used and need to be omitted along with `"c"`. In `letters[-c(3, 17:26)]` note the use of `-c(3, 17:26)` to list letters to be omitted, specified by their position in the list generated by letters and that, via the minus sign, omission is intended.

Finally the `abline` function is used to add reference lines to the plots. Here the arguments `h = 0` and `v = 0` add horizontal and vertical lines to the plot. This is particularly useful for biplots presented as separate row and column plots since it aids the visual superimposition of plots when interpretating the results. The `lty = 2` argument specifies the line type `lty` to be used – dashed in this case.

Chapter 3

Continuous data

3.1 Continuous data

3.1.1 Introduction

This chapter deals with simple methods for the analysis of continuous or quantitative data. The distinction is often drawn between *interval-scaled* and *ratio-scaled* data, the latter type involving positive numbers and for which the concept of a ratio is meaningful. Temperatures on the Celsius (centigrade) and Farenheit scales are examples of interval scaled data. For example, we cannot say that a temperature of 40 degrees centigrade is 'twice as hot' (a ratio concept) as one of 20 degrees; using the Farenheit scale converts to 104 and 68 with a ratio 1.53. This differs from 2 and shows that the scales are not ratio-scales. On the other hand it makes sense to say that a pot that is 10 cm tall is twice as tall as one of 5 cm, so height is a ratio-scaled variable. The same is true of many variables used in archaeological data analysis (e.g., length, weight).

Continuous data are usually contrasted with discrete (or counted) data, the subject of Chapter 4. The nature of the data should reflect the method of analysis, including graphical presentation, and this is sometimes neglected. In deciding how to analyze a set of data it should be emphasized that the important issue is whether or not measurements are of a variable that is continuous, *in principle*. For example, (estimated) age at death is a continuous variable but may only be measured to the nearest year. In fact all continuous measurements are inevitably truncated/rounded. The height of a pot may be $10.2683310 \cdots$ cm but, depending on the measuring instrument and the accuracy that is realistic, may be recorded as 10 cm, 10.3 cm or 10.27 cm.

For illustration, unpublished data from Cool (1983) are used. They are the lengths (mm) of 90 copper alloy hairpins from southern Britain, 55 classified as early and 35 as late on archaeological grounds (see Cool, 1990, for a review of the use of such hairpins). The data are given in Table 3.1.

54	56	74	84	85	85	87	88	89	90
90	92	92	92	92	93	93	93	93	93
94	94	94	95	95	95	96	96	97	97
97	98	98	100	100	100	100	101	102	103
104	104	104	104	105	107	108	108	111	115
115	116	123	128	134					
51	52	54	56	57	58	60	60	61	62
62	63	63	63	65	65	66	67	68	68
70	70	70	70	71	74	75	77	78	78
80	80	82	82	87					

Table 3.1: *Lengths (mm) of Romano-British copper alloy hairpins from southern Britain (Cool,1990). The upper set are early and the lower set late.*

3.1.2 Histograms, dotplots and boxplots

Figure 3.1 shows examples of histograms, dotplots and boxplots for the early hairpins data. Much of this will be familiar. An aim here is to demonstrate the ease with which R can be used to get publication quality graphs. Code is given either following the figures or in Section 3.2.2. In terms of substantive interpretation the histograms and dotplot are unimodal, with a central concentration of data that tails off, and nothing too distressing in the way of outliers or long tails. The second histogram and dotplot draw attention to two possible outliers to the left that are smaller than the bulk of the data, but not enough to agonize about. We shall loosely refer to such data as (reasonably) *well-behaved*. Data that have a normal distribution are the ideal example of well-behaved data (Section 12.2.1).

A histogram is defined by counts of measurements in cells that have defined lower and upper limits; the default in software packages is to have cells of equal widths (bin-widths), with a default rule applied to determine the number of bins. With equal bin-widths the heights of the bars associated with the bins are proportional to frequencies. A probabilty density scale can be specified if preferred. If a case lies exactly on the boundary between two bins a rule of some kind is applied to determine whether such cases go into the left or right bin. In what follows `RBpins.early` is the name of the data file given to the early hairpins.

The appearance of default histograms can sometimes be improved by increasing the number of bins (i.e. the original is ‘oversmoothed’). To specify 20 bins use `hist(RBpins.early, 20)`. This is a guide since aesthetic considerations, namely

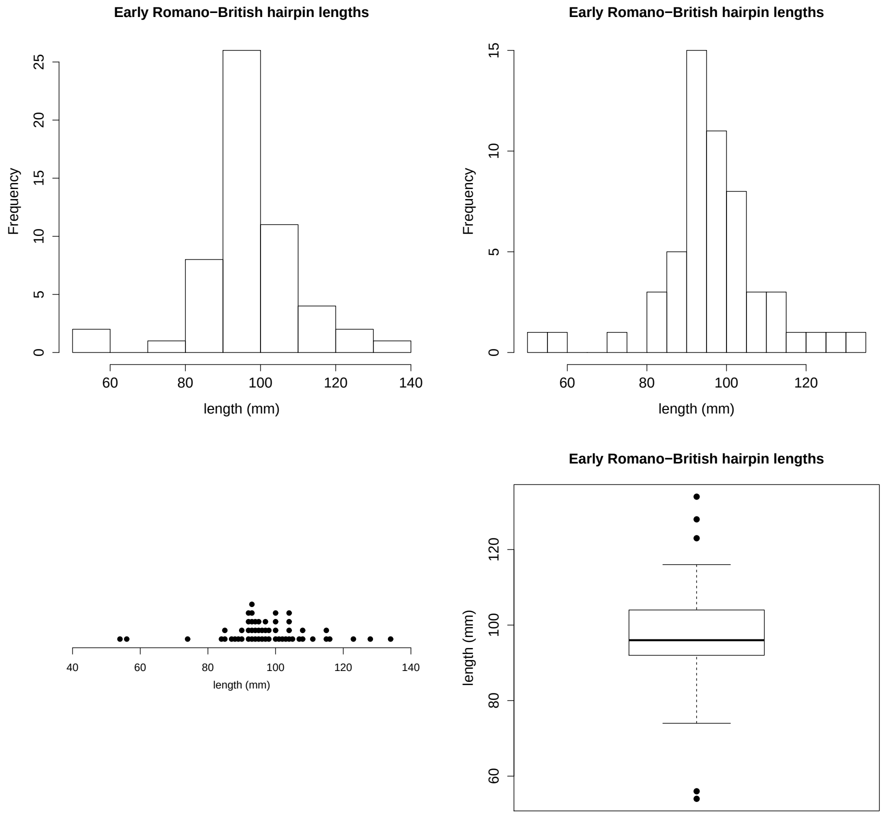

Figure 3.1: *Univariate graphical displays for the lengths of early Romano-British hairpins. The upper left histogram is the* R *default; that to its right specifies 20 bins. A dotplot and boxplot are shown beneath them.*

the desire for 'nice' numbers on the x-axis, may dictate a slightly different choice. It is legitimate to play around with the number of bins until the result is judged satisfactory, at which point labeling can be tidied up (Section 2.6.2).

If data are well-behaved it limits complications that may arise in further analysis. In the present instance it legitimizes the use of boxplots (sometimes called box-and-whisker plots) for data display, which are not well-suited to data with more than one mode. The way boxplots are drawn depends on the software used but typically, as here, the box covers the central 50% of the data with its width the interquartile range (IQR), and the line in the box the median. For very well-behaved data whiskers extend to the maximum and minimum of the data. Where, according to a default criterion, which is a bit arbitrary, 'unusual' values (or outliers) are detected the whiskers are broken and the unusual cases highlighted. In

R the default, which can be changed, is to highlight cases more than 1.5 × IQR from the limits of the box.

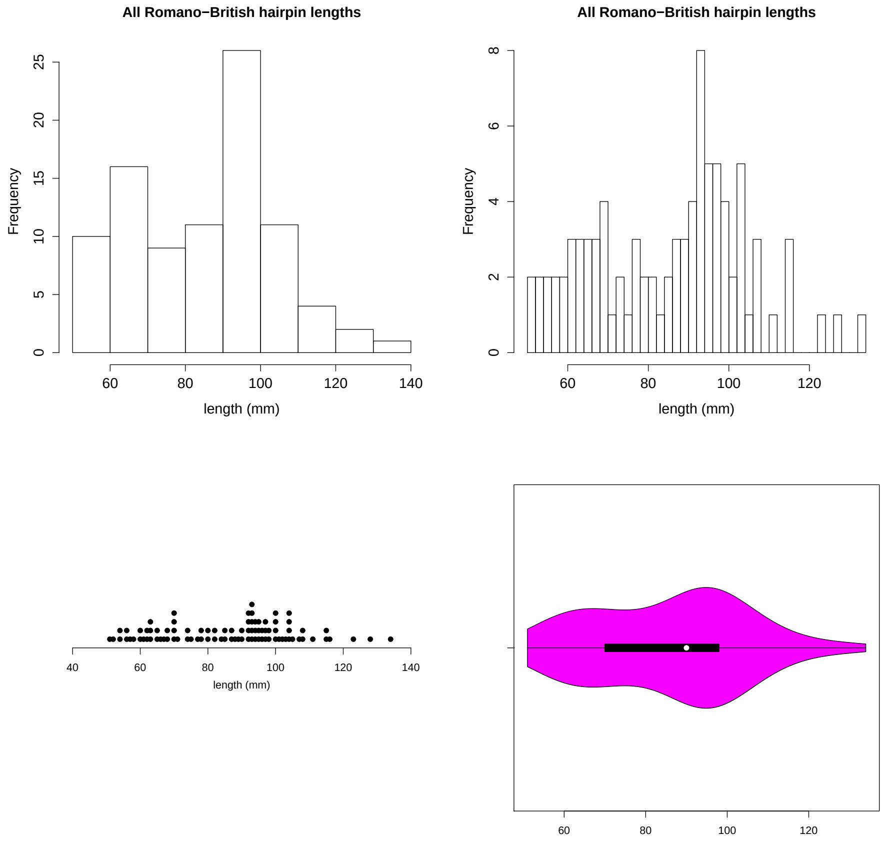

Figure 3.2: *Univariate graphical displays for the lengths of all Romano-British hairpins. See the text for discussion.*

Care needs to be exercised in interpretation, and the present plot exemplifies this. The highlighted cases in the lower part of the boxplot could be interpreted as outliers, whereas those in the upper part are more indicative of a 'slight' tail to the right. In more extreme cases than that illustrated it may be sensible to re-examine the data omitting outliers. If highlighted cases are indicative of a long tail, then the boxplot will not be symmetric. For some methods of statistical analysis long tails are not always welcome, and logarithmic transformation to improve symmetry is common.

Figure 3.2 is mostly as Figure 3.1 but uses all the hairpins, ignoring knowledge of their date. The main difference is the evidence of bimodality in the histograms and dotplot, rendering the boxplot unhelpful. It has been replaced by a violin plot

(Hintze and Nelson, 1998) which, as well as a boxplot, produces an estimate of the density along the boxplot, so that the bimodality can be seen. The package `vioplot` needs to be imported and loaded, with the function of the same name then being used. In the second histogram 30 rather than 20 bins were specified, and it might be thought a little ‘bitty’.

3.1.3 Kernel density estimates

Univariate data – one group

Kernel density estimates (KDEs) would be regarded as mathematically complex by many archaeologists (see Silverman, 1986; Wand and Jones, 1995; Bowman and Azzalini, 1997, to be convinced). Baxter and Beardah (1996) and Baxter *et al.* (1997) provide what counts as early expositions aimed at archaeologists, and Chapter 3 of Baxter (2003) reviews some uses of KDEs in archaeology, to that date. They can now be regarded as a standard method in archaeological data analysis. Conceptually, as illustrated in most of the examples here, KDEs can be used to produce smoothed histograms that overcome many of the problems of the latter illustrated in Whallon (1987). Computationally, `plot(density(RBpins.early))` using the `density` function gets you going for the early hairpins data. We proceed by illustration, starting with Figure 3.3.

The upper-left plot, for the early hairpins data, is the default plot for the `density` function. It is useful to look at because it gives the default bandwidth used. The bandwidth is analogous to the bin-width in a histogram, and controls the appearance of the KDE through the degree of smoothness. Various rules have been put forward for choosing ‘good’ bandwidths. These are mathematically based and, to my eye, sometimes result in KDEs that are too smooth. As with bin-widths, my preference is to experiment with different bandwidths and select the KDE subjectively. The upper-right plot uses a smaller bandwidth than the default, but it makes little difference to the appearance, with the tails being a bit bumpier. The labeling has also been modified from the default.

The two lower plots are for all the hairpins. That to the left is designed to show the bimodality in the data – remember, at this stage a difference between early and late hairpins is not being assumed. The plot to the right shows a KDE with a larger bandwidth, and therefore smoother.

The next example was provoked by examination of the data for Mg in Table B.1. It illustrates issues of smoothing and data transformation. Examination of the table shows quite clearly that there are regional differences with the values for Region 3 somewhat smaller than for other regions. This is not evident in the default KDE in the upper left of Figure 3.4. To show that Region 3 is separate in a KDE a rather drastic reduction in the bandwidth is needed, as shown in the

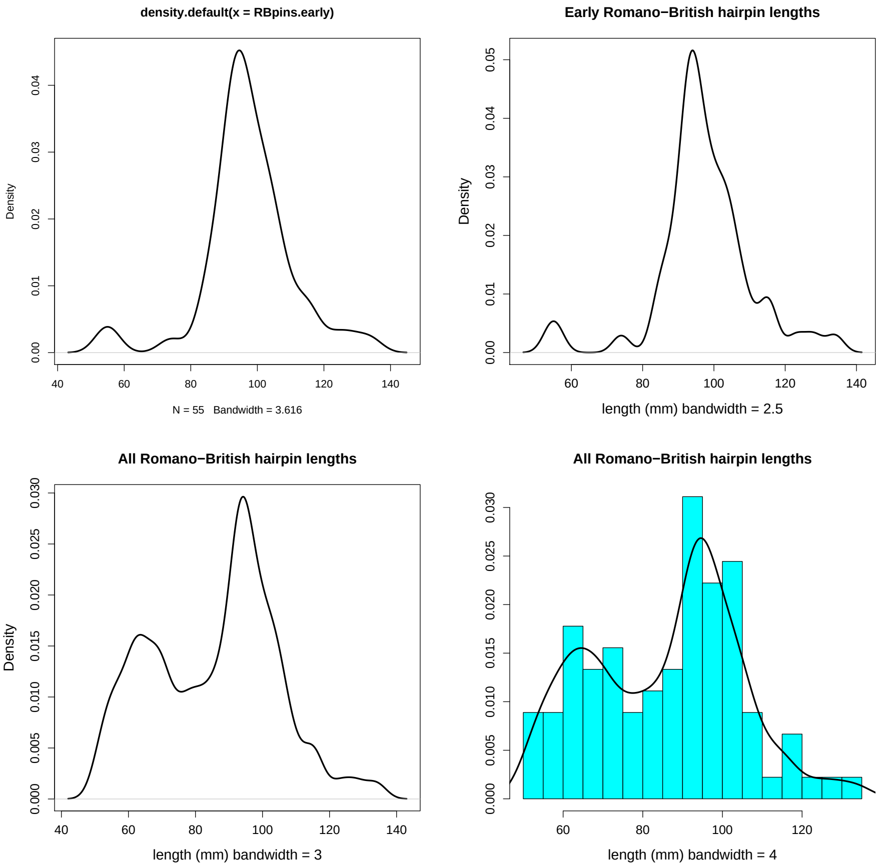

Figure 3.3: *Applications of KDEs to the Romano-British hairpin lengths of Table 3.1, discussed in the text.*

upper-right panel. The bandwidth was chosen to isolate Region 3, and it also does a good job of identifying Region 2 with the highest mode. My immediate reaction on seeing this was that the right-hand side of the plot was undersmoothed; in fact the lower left-hand plot of Figure 2.3 suggests it is reasonable. It is consistent with the fact that there is an extreme case in Region 2 and a small but separate group of five cases.

A common strategy in the analysis of artifactual chemical compositions, where measurements are strictly positive, is to transform to logarithms (Bieber *et al.*, 1976). If this is done the default KDE immediately suggests trimodality, with the mode for Region 3 to the right. Reducing the bandwidth emphasizes the grouping more, without introducing spurious detail (lower figures). The variation in Region 2, noted in the previous paragraph, is not picked up. In these examples some prior knowledge of the data structure, gained from tabular inspection, is needed

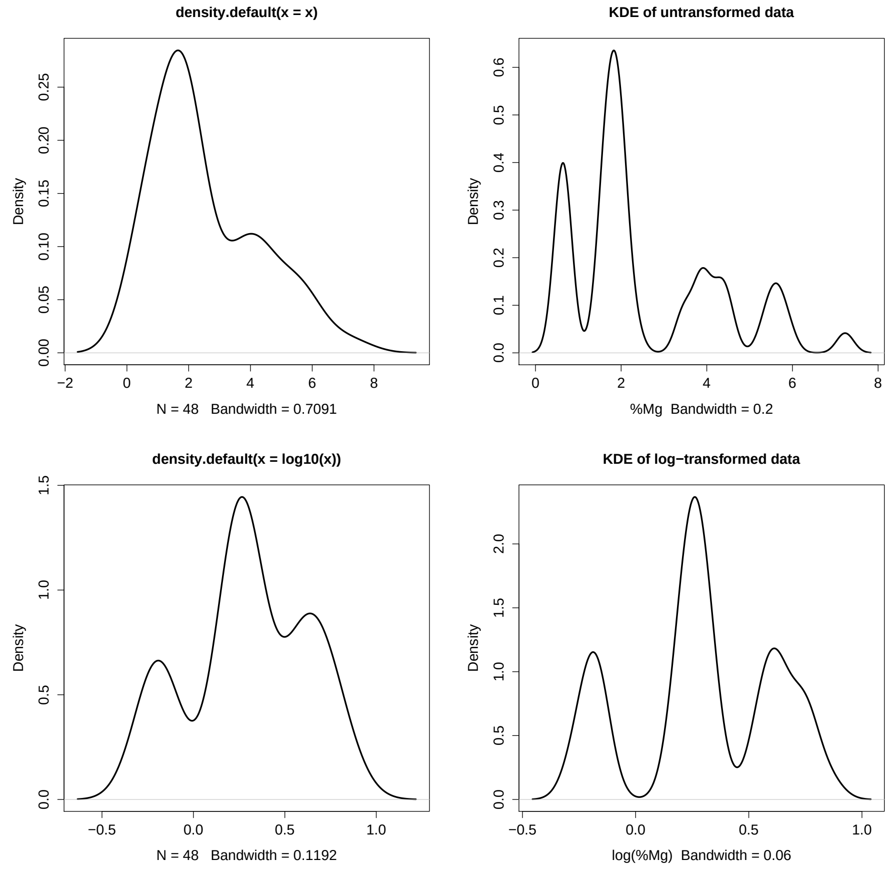

Figure 3.4: *The upper and lower plots contrast KDEs for untransformed and log-transformed data to base 10, for Mg from Table B.1. See the text for a discussion.*

to obtain sensible and informative displays. Almost invariably, it is useful to look at a set of data in more than one way.

Comparing two groups

The two left-hand panels in Figure 3.5 show histograms for the early and late hairpins. There are arranged vertically and the scales on the x-axis and bin-widths have been arranged to be the same. A probability scale is used so the comparison is of shape ignoring sample size. I am not generally a fan of this kind of display, but it works well enough here.

Other examples of this usage are given in Mellars and Wilkinson (1980) to compare the distribution of otolith lengths for samples from late Mesolithic shell-midden sites from Oronsay. Relatively few histograms are used (up to five) and the patterns are sufficiently clear to be informative (though the graphs use a lot of

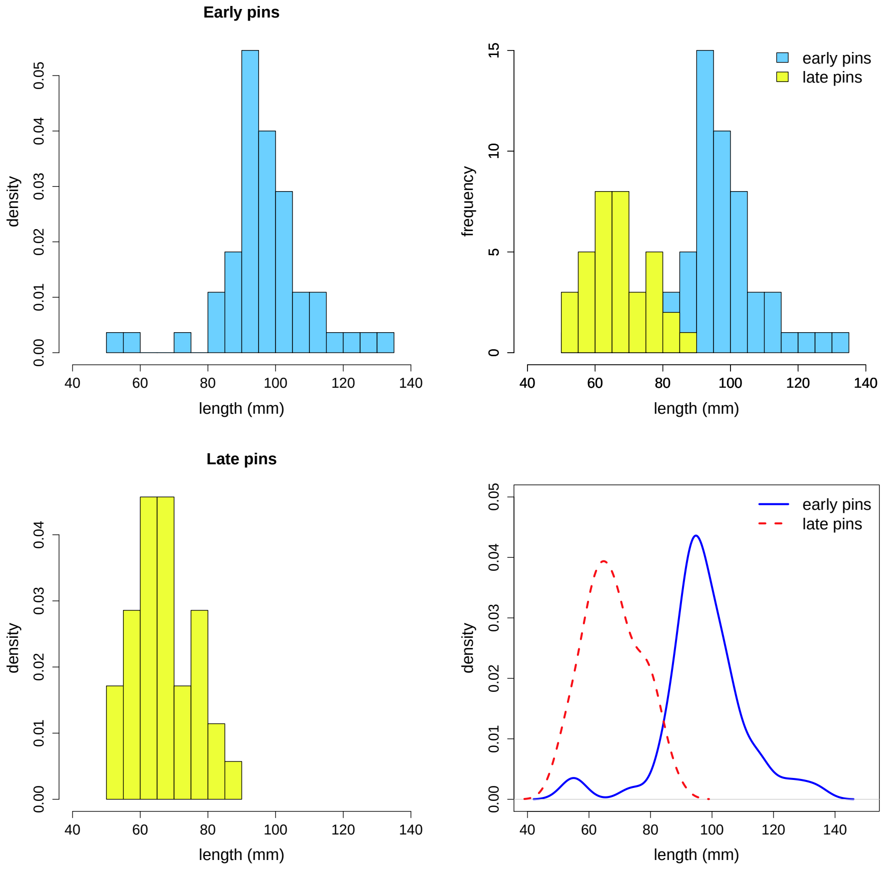

Figure 3.5: *Comparing two groups using histograms and KDEs. For the KDEs the solid and dashed lines are for early and late hairpins respectively. See the text and section notes for a discussion.*

space). I am less convinced by the numerous examples in Albarella *et al.* (2006) where 10/25 pages, most consisting of five histograms, are used to compare mostly pig lower-tooth measurements from a variety of Italian prehistoric contexts. Sample sizes are small for some contexts with, arguably, too many bins used; patterns are not easy to compare; and, with the exception of a clearly bimodal distribution, a table of summary statistics might have served as well or better.

The upper-right plot in Figure 3.5 superimposes the two histograms, and a frequency scale is used which allows the sample size difference to be seen. A problem with such plots is that one histogram will obscure part of the second histogram. In this instance the main message, that the groups have reasonably different locations, is not obscured but, in general, this kind of plot is not really suitable for histograms with much overlap, or for more than two histograms. My preference in this example would be to superimpose two KDEs, as shown in the lower-right figure. The separation between early and late hairpins is evident, without one KDE obscuring the other. The KDEs are on a density scale so provide no information on sample size differences.

Histograms and KDEs are not the only way of comparing groups. Figure 3.6 shows some possibilities using stripcharts, violin plots and boxplots. The stripchart for the %Fe data from Table B.1, by region, justifies the use of boxplots as there is no evidence of multimodality (also true of the violin plot). The stripchart suggests there are two outliers in Region 1, also suggested by the relevant boxplot. Boxplots for the Romano-British hairpins data are also shown in the lower-right plot and show that the early hairpins tend to be longer than the late hairpins.

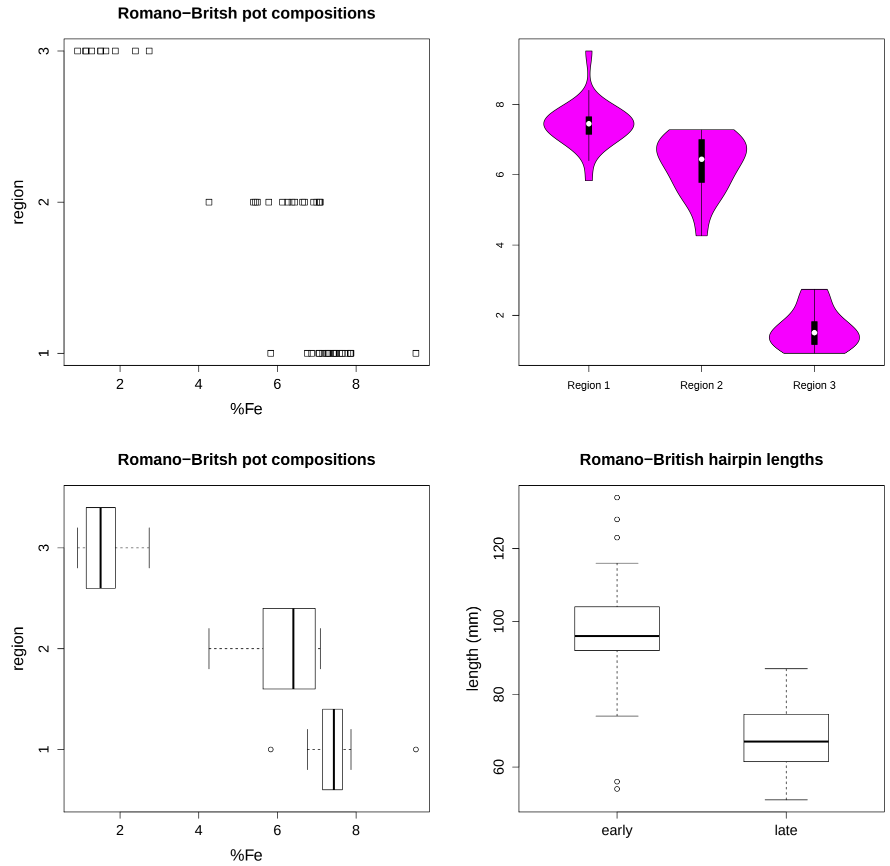

Figure 3.6: *The upper left-hand plot is a 'stripchart' showing the distribution of %Fe by region for the data from Table B.1. To its right, and below, the violin plot and boxplots effect similar comparisons. The final figure is based on the Romano-British hairpins data and shows the distinction between early and late hairpins.*

With software such as R it is sensible to explore different methods of display before selecting one for final publication that tells the story the data deserves. Sometimes the use of more than one display is merited, though publication constraints can militate against this.

3.2 R notes

3.2.1 Functions

In Section 2.6.3 the possibility of writing a *function* to ease the calculation of the statistics for Table 2.1 was mentioned in passing. User-defined functions are often introduced in the later stages of introductory texts on R, but it is useful to know about them at an early stage. With more than a few lines of code it can be more efficient to write a function to do the job. This can be tested and edited, using the edit function, before applying it 'in anger'[1]. One might also wish to reuse a function – for example, by applying the code used to generate the numbers in Table 2.1 for another variable. This will be taken as an initial illustration, then made more general.

The first line in the code that follows names the function RBpot.statistics1; function(x) defines the function with an argument, x, that will be replaced with the data to be used. The function code is enclosed within the braces { and } and consists mainly of code to calculate the statistics previously used. These are collected together in the object statistics, where the column bind function, cbind, treats each statistic as a 1×1 table, resulting in a 1×4 table. Doing it this way retains the statistics' names which makes the output, when printed, easier to read

```
RBpot.statistics1 <- function(x){
Mean <- mean(x)
Median <-median(x)
SD <-  sd(x) # standard deviation
IQR <- IQR(x) # Interquartile Range
statistics <- cbind(Mean, Median, SD, IQR)
list(stats = round(statistics, 2))
}
```

[1] For example, RBpot.statistics1 <- edit(RBpot.statistics1) will bring up an edit screen where the function in question can be edited. If mistakes are made then, when you exit from edit mode, you will get a message about this – not always that helpful. Typing the function name will return the original code. To recover the edited version for correction, mistakes and all, use edit().

The list function provides the names for and definitions of the objects of interest to be printed – in this case the rounded values of the statistics, named stats, are made available (more than one object can be listed). The rounding is achieved by the round function which takes the data in the first argument and rounds it to the number of decimal places given in the second argument. The contents of the list can be obtained by typing the function named. Should the output be needed for later use it can be saved using

```
Statistics <- RBpot.statistics1(tubb.data$K)
```

and viewed immediately by typing Statistics to get

```
$stats
     Mean  Median  SD   IQR
[1,] 3.18  3.16   0.92  0.96
```

There are several obvious advantages to creating even simple functions like this. One is that other statistics can be added (or subtracted) at will. The argument x can be varied to obtain statistics for different subsets of the data or different variables. For example, when invoking the function, replacing K with K[tubb.region == 1] will produce the statistics for Region 1 only; replacing it with Fe will calculate the statistics for that oxide etc.

A natural question to ask is if the regional calculations for all regions can be done with a single function. The following code is one possible way of doing this. The function includes a second argument, TypeId, which is where the list containing the group identifiers is entered.

```
RBpot.statistics2 <- function(x, TypeId){
Names <- names(table(TypeId))
Stats <- NULL

for(I in seq(1:length(Names))) {
   Mean <- mean(x[TypeId == Names[I]])
   Median <- median(x[TypeId == Names[I]])
   SD <- sd(x[TypeId == Names[I]]) # standard deviation
   IQR <- IQR(x[TypeId == Names[I]]) # Interquartile Range
  Stats <- rbind(Stats, round(cbind(Mean, Median, SD, IQR), 2))
}
row.names(Stats) <- Names
list(Stats = Stats)
}
```

The table function is a convenient way of finding how many categories are represented in the TypeId variable and their names are extracted using the names function. The extracted variable is here called Names, and the subsequent use of the function length provides the number of elements in Names (three in this instance). Group labels are integers here and will be listed accordingly; with text identifiers the ordering produced by the table function corresponds to alphabetical order.

A 'for loop' using the for function does the computations. The function seq, in conjunction with length as used here, generates the number of categories to loop through; each loop creates a table of the statistics with one row using the cbind function; and these are bound together using the row bind function, rbind, to produce, in this instance, a 3 × 4 table of statistics. The binding process needs to start from somewhere and an 'empty' object is created using Stats <- NULL that is subsequently filled in during the looping process. Finally, the row.names function adds row names to the table so that it is more readable when printed.

The function may be executed and results saved using

```
Statistics <- RBpot.statistics2(tubb.data$K, tubb.region)
```

and printed as in the previous example to obtain

```
$Stats
  Mean Median   SD  IQR
1 3.11   3.13 0.22 0.15
2 4.01   4.28 0.97 0.72
3 2.02   2.03 0.19 0.16
```

As a final example the calculation of statistics for all the variables is illustrated. This is not broken down by region, but this can be achieved, region-by-region, in the call to the function.

```
RBpot.statistics3 <- function(x) {
            Mean <- apply(x, 2, mean)
            Median <- apply(x, 2, median)
            SD <- apply(x, 2, sd)
            IQR <- apply(x, 2, IQR)
            Stats <- rbind(round(cbind(Mean, Median, SD,
                IQR), 2))
        list(Stats = Stats)
    }
```

The apply function takes the data matrix x as its first argument; the second argument dictates whether calculations are to be based on columns (2 as here) or rows (1); and the third argument is the function that is to be applied (which can be user-defined). Calling the function as previously illustrated produces the following results.

```
$Stats
    Mean Median   SD  IQR
Al 15.61  16.15 2.70 4.38
Fe  5.83   6.89 2.35 1.93
Mg  2.54   1.93 1.73 2.29
Ca  0.51   0.30 0.45 0.68
Na  0.25   0.21 0.17 0.27
K   3.18   3.16 0.92 0.96
Ti  0.85   0.90 0.21 0.25
Mn  0.08   0.08 0.07 0.05
Ba  0.02   0.02 0.00 0.00
```

If, for example, statistics are needed for Region 1 they can be obtained using

```
RBpot.statistics3(tubb.data[tubb.region == 1,])
```

3.2.2 Code used for analyses in the text

Figures 3.1 and 3.2

The following function might be used, omitting presentational arguments.

```
somegraphs <- function(x) {
win.graph()
hist(x)
# The default; look at the result and try 20 bins.
win.graph()
hist(x, 20)
library(plotrix) # Needed for the dotplot.
win.graph()
dotplot.mtb(x)
win.graph()
boxplot(x)
}
```

Print the graphs with `somegraphs(RBpins.early)` where `RBpins.early` contains the data on early hairpin lengths. The same graphs can be obtained, without using a function, as

```
hist(RBpins.early)
hist(RBpins.early, 20)
dotplot.mtb(RBpins.early)
boxplot(RBpins.early)
```

but the function can be used for other data, such as `RBpins.late`. In practice presentational arguments are included and it is easier to experiment with these by editing the function rather than re-typing the command every time. This can be done even more easily by expanding the function to add further arguments.

Thus, and for example, to label the axes

```
somegraphs <- function(x, Xlab = " ", Ylab = " ")
```

specifies the arguments;

```
hist(x, xlab = Xlab)
```

is included in the body of the function; and

```
somegraphs(RBpins.early, Xlab = "length (mm)", Ylab = "frequency")
```

labels the x- and y-axes accordingly.

Several graphs are produced; the `win.graph()` functions ensure that all plots are displayed on the terminal; if omitted only the final graph will be printed. The `hist` function produces the histograms. That to the left is the default histogram. For the second histogram the second argument, 20, specifies the number of bins preferred. To get sensible (i.e. short) labels on the axis `R` may modify this a little.

The `dotplot.mtb` function requires the `plotrix` package to be imported and loaded. It produces a dotplot that imitates what can be obtained in `MINITAB`; it is limited in the control that can be exercised over labeling. The function `stripchart` is the preferred alternative in `R`. The `boxplot` function produces the boxplot. Figure 3.1 has `x = RBpins.early` as its argument; Figure 3.2 replaces `boxplot` with `vioplot` and uses `x = RBpins.all`.

Figure 3.3

```
kde.plots1 <- function(x, y){
library(MASS)     # Needed for ‘truehist’ function.
win.graph(); plot(density(x))
win.graph(); plot(density(x, bw = 2.5))
win.graph(); plot(density(y, bw = 3))
win.graph(), truehist(y, nbins = 20), lines(density(y, bw = 4))
}
kde.plots1(RBpins.early, y = RBpins.all)
```

Presentational aguments are omitted. The semi-colons (;) allow commands to be placed on the same line.

The KDEs are produced using the `density` function . To produce the figure the function arguments, `x` and `y` were `RBpins.early` and `RBpins.all`. The first KDE is the default output for the former and provides a starting point for selecting the bandwidth. This is chosen by a default automatic bandwidth selection procedure; several options are available via the `bw` argument (see `?density` for details). My preference is to choose the bandwidth subjectively after experimentation starting from the default choice and the second plot uses `bw = 2.5` for the early pins data, producing a less smooth estimate.

For all the hairpins `bw = 3` is used in the third plot and `bw = 4` in the final plot, producing greater smoothing. The second of these shows how to overlay the KDE on a histogram using the `lines` function. This requires the `truehist` function from the `MASS` package. The KDE is on a density scale so the histogram must also be on this scale; the `truehist` function provides this by default.

Figure 3.4

The function `log10`, transforms the data to base 10 logarithms; for natural logarithms, $\log_e$, use the `log` function.

```
kde.plots2 <- function(x){
win.graph(); plot(density(x))
win.graph(); plot(density(x, bw = .2))
win.graph(); plot(density(log10(x)))
win.graph(); plot(density(log10(x), bw =.06))
}

kde.plots2(tubb.data$Mg)
```

Figure 3.5

In the following code the `col` argument for the `hist` function shows how to color the bars of the histogram. The `xlim` argument controls the range of the x-axis, with `ylim` doing the same for the y-axis. There are various reasons one might wish to do this; to restrict the range to 'magnify' parts of a plot, or to expand a plot to accommodate a legend, for example. In the present instance it is used to ensure that histograms have the same range on both the x- and y-axes. In the plots showing two histograms simultaneously this is essential. Here this is done by superimposing two plots; the first of these produces a histogram for the early hairpins and the histogram for the late hairpins is overlaid on this.

This is done using par(new = T) with limits given by xlim = c(40, 140) and ylim = c(0, 15) to ensure compatibility of the plots. Note that the argument freq = T is used in both cases, to provide a histogram on a frequency scale. This is the default and could be omitted; to get a probablity density scale, as in the separate histograms for the two groups, use freq = F. This removes the effect of the sample sizes in any comparison of the histograms, which the joint plot retains.

The par function can take numerous arguments that allow fine control of the graphics to be exercised. See Section 4.4 of Venables and Ripley (2002) for a discussion of this and some examples; ?par in R lists what arguments are available.

```
KDEHist <- function(x, y) {

win.graph()
hist(RBpins.early, n = 20, col = "skyblue", xlim = c(40, 140), freq = F)

win.graph()
hist(RBpins.early, n = 20, xlim = c(40, 140), ylim = c(0, 15), freq = T)

par(new = T)   # This superimposes the plot on the previous one
hist(RBpins.late, n = 10, xlim = c(40, 140), ylim = c(0, 15), freq = T)

legend("topright", c("early pins", "late pins"), fill =
c("skyblue", "yellow"), bty = "n", cex = 1.5)

win.graph()
hist(RBpins.late, col = "yellow", n = 10, xlim = c(40, 140), freq = F)

win.graph()
plot(density(RBpins.early, bw= 4),xlim = c(40, 150), ylim = c(0,.05))
lines(density(RBpins.late, bw= 4))
}

KDEHist(RBpins.early, RBpins.late)
```

Figure 3.6

Most of the features used below have already been discussed and illustrated in the figures in the text. This is the first use of the stripchart function, mentioned earlier as an alternative to the dotplot produced by dotplot.mtb. The vioplot function has fewer graphical capabilities than the other functions used.

The default in the boxplot function is to display the plots in a vertical array; the argument horizontal = T produces a horizontal array. The boxwex argument controls the width of the boxes and may be used to produce a more appealing appearance of the plot.

```
graphs <- function() {
library(vioplot)     # Needed for the ‘vioplot’ function

win.graph(); stripchart(tubb.data$Fe ~ tubb.region)

win.graph()
vioplot(tubb.data[1:21,2], tubb.data[22:38,2], tubb.data[39:48,2],
names = c("Region 1", "Region 2", "Region 3"))

win.graph()
boxplot(tubb.data$Fe ~ tubb.region, horizontal = T)

win.graph()
boxplot(RBpins.early, RBpins.late, boxwex = .5)
}

graphs()
```

Chapter 4

Discrete data

4.1 Discrete data, barplots and histograms

Histograms are appropriate for the presentation of continuous data. Such data are usually contrasted with discrete data which, at their simplest are counted data in different and disjoint categories. An example would be counts of distinct vessels by type in an assemblage of pot or glass. Most archaeologists will be familiar with the presentation of such data in the form of pie-charts or bar-charts (barplots). These are among the most visible of statistical methods used in the archeological literature; they are sometimes over-used, used unnecessarily, or misused (despite their apparent simplicity).

Discrete data may be ordinal, in that categories have a natural ordering but the 'distance' between categories is not known. Common archaeological examples may involve chronology; for example, counts of a single artifact type ordered chronologically on the basis of stratigraphy or phasing, without knowing the absolute chronology. Whether or not data are ordinal has implications for graphical presentation, and for the choice of analytical method.

Barplots (and pie-charts) for a single set of counts, possibly expressed as percentages, are often pretty boring. Unless there are a large number of categories, looking at the numbers in a table is often all that is really needed. Things get more interesting when tables of counted data arise from data analysis. If, for example, counts of different artifact types are available for a single context a simple barplot will do. If such data are available for different contexts they can be expressed in tabular form (examples follow) and questions can then be asked about the similarity of contexts in terms of the artifact assemblages that characterize them, or the similariy of artifacts in terms of their distribution across contexts.

Data in such a form are variously referred to as *cross-tabulated*, *cross-classified*, or *contingency tables*. Compared to data matrices for continuous data, where rows and columns (cases and variables) have a different 'status', in contingency tables the rows and columns have a similar status and a different notation is used here to reflect this. In general we refer to an $I \times J$ contingency table. Rows and columns can be interchanged, though in practice the emphasis may be on one or the other.

To herald later comment, it is convenient to focus on the difference between histograms and barplots. The examples in Figure 4.1 use the weights of loomweights data, from Tables B.3 and B.4. The lower-right plot shows the default `R` histogram using the `hist` function. Note that the bars associated with the bins 'touch' each other. Histograms are sometimes referred to as barplots or bar-charts in the archaeological literature and *vice-versa*. This is possibly understandable since the histogram is represented by bars whose *area* corresponds to the counts within bins, but should probably be regarded as incorrect.

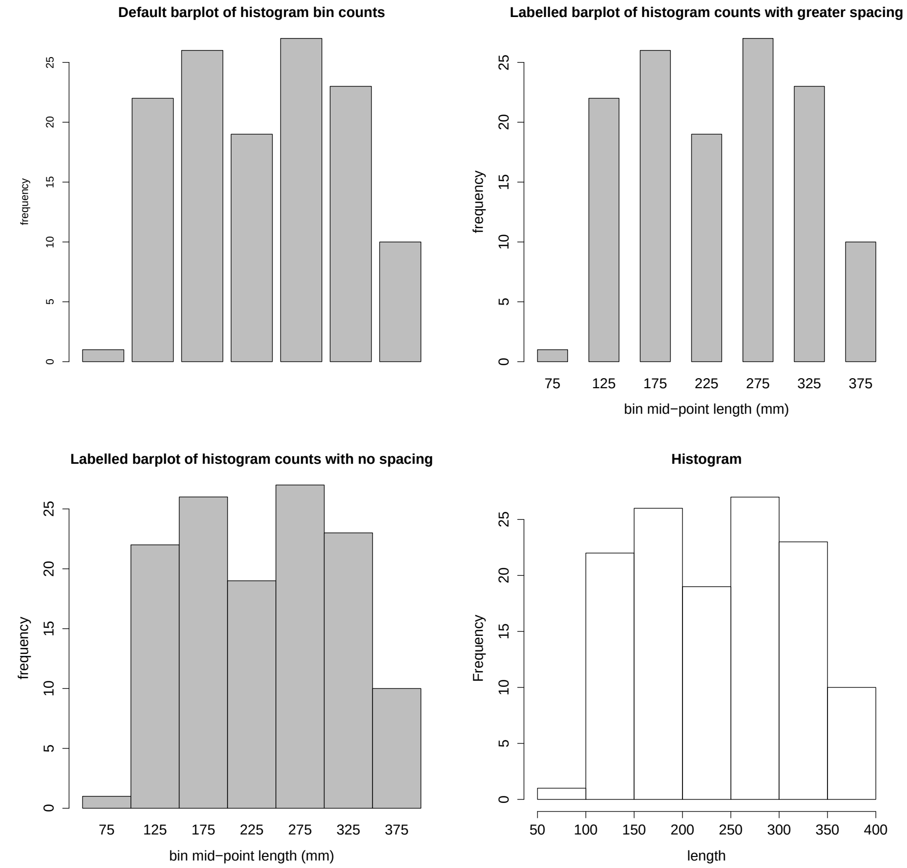

Figure 4.1: *Right and wrong ways of presenting a histogram, based on counts for bins in the default* `R` *histogram for the reduced loomweight weight data of Tables B.3 and B.4 – (1, 22, 26, 19, 27, 23, 10).*

It is easy enough to get the histogram if all the data are available. What if only counts within bins are given, in this instance (1, 22, 26, 19, 27, 23, 10)? It is impossible to tell by looking at the numbers alone whether they represent continuous or discrete data so the context, which would include the intervals, continuous or (usually) integers for discrete data, needs to be examined.

The upper plots incorrectly show the data as barplots; the gaps between bars indicate quite clearly (at least to my eye) that the data are discrete. This misuse is not uncommon. In `R` the `barplot` function allows control over the spacing used between bars and setting this to zero produces the histogram shown in the lower-left plot.

This kind of graphical/terminological misuse to be found in the literature is possibly not misinterpreted much, but may betray confusion about the distinctions involved, or a lack of adequate software. (The problem was common some years ago but may now be less prevalent because popular and widely used non-statistical software has belatedly caught up with the issue.)

Other issues concerning the use of barplots are illustrated later, but first a more interesting example is examined. The data relate to Roman pillar-moulded bowls found in excavations at Colchester. Eighteen periods, with a chronological sequence, were identified. Data are given in Table 4.1.

In the absence of knowledge of the date span of the periods it is legitimate to represent the data in the form of a barplot[1]. This is done in the top plot in Figure 4.2. On a technical note the width of a bar is arbitrary. What is shown is conventional, but vertical lines with heights corresponding to percentages would be as legitimate.

We can do better because information on the date span of the periods, in terms of actual dates and the width of the spans, is available. The earlier periods are more tightly defined than later ones. Pillar-moulded bowls are an early form and occur predominantly in the earlier periods. Knowing the dates and spans means that the data can be treated as continuous and represented in the form of a histogram. This is done in the right-hand plot in the figure, using period rather than mid-points to label the scale.

There is a complication. Most software for histograms produces equal bin-widths by default. Sarkar (2008: 39) goes so far as to say that unequal bin-widths are ‘rarely used outside introductory statistics textbooks’. This, and others in Cool and Price (1995), is a counter-example.

How the histogram was produced is discussed in Section 4.4. Briefly, the `barplot` function was used, with zero spacing between bars, and widths of bars

[1] The date span is being ignored here; so the sequence is ordered but the date-span of categories is not known. This is an example of *ordinal* data; this is quite common with chronological data – the situation here where the date-span can be specified is less often seen.

Period	Date	Width	Midpoint	%
I	43-50	8	46	11
II	51-60	10	55	22
III	61-70	10	65	9
IV	71-80	10	75	8
V	81-90	10	85	6
VI	91-100	10	95	5
VII	101-125	25	112	9
VIII	126-150	25	137	5
IX	151-175	25	162	4
X	176-200	25	187	3
XI	201-225	25	212	2
XII	226-250	25	237	2
XIII	251-275	25	262	2
XIV	276-300	25	287	3
XV	301-325	25	312	2
XVI	326-350	25	337	1
XVII	351-375	25	362	1
XVIII	376-400	25	387	1

Table 4.1: *Chronology of Roman glass pillar-moulded bowls found during excavations at Colchester, 1971-1985. The data are ordinal, but the periods are of different lengths. See Cool and Price (1995: 15–19) and the text for discussion.*

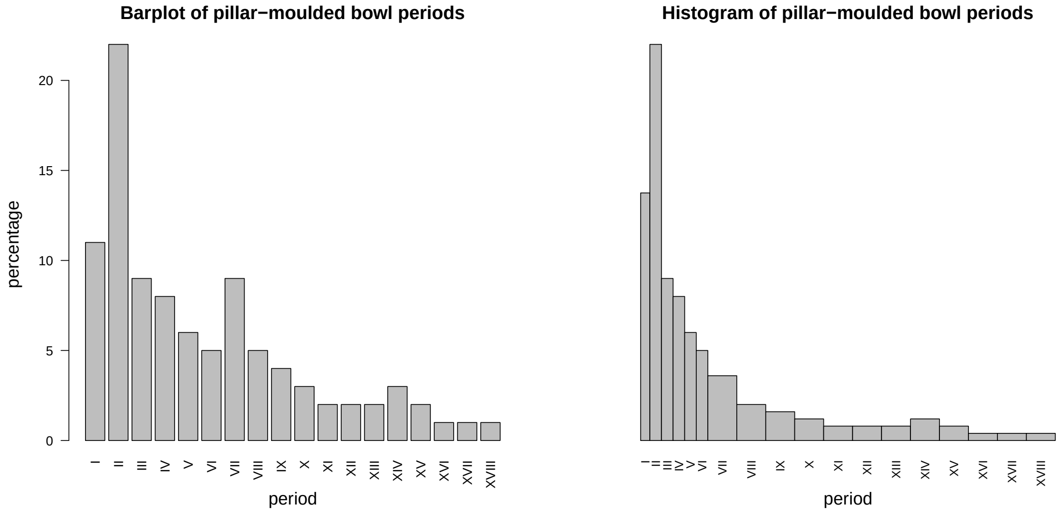

Figure 4.2: *Different ways of representing the data from Table 4.1. The barplot to the left respects the ordinal nature of the data, but not the fact that periods are of different lengths. The histogram to the right, constructed using the* `barplot` *function, does respect the differing lengths. Because of the unequal bin-widths a vertical scale is not appropriate.*

corresponding to the date span specified. Because of the unequal widths a sensible scale for the vertical axis is not available.

The advantage of the histogram compared to the barplot is that it emphasizes the decline in use over time more obviously. The quite sharp and steady decline in usage is readily apparent. The barplot does not do this, because it was constructed without use of the spans which are longer for the later periods. There is the suggestion in the barplot of a secondary peak in period VII that exists simply because the span is longer than earlier periods and, for the same reason, the sharp decline in usage over time is less apparent.

4.2 Barplots for two-way tables

The barplots in Figures 4.1 and 4.2 provide examples of what barplots for single sets of counts look like. This section deals with two-way tables of counts. For illustration, data adapted from Table 5 of Bailey *et al.* (1983) are used, showing the distribution by stratum of four classes of artifacts from excavations of the palaeolithic rockshelter of Kastritsa in north-west Greece. It is assumed that interest lies in a comparison of the distribution across strata of artifact types, expressed in Table 4.2 as percentages. The authors did not present analyses of the kind to follow – they are here to demonstrate 'technique' and coding.

Stratum	I	II	III	IV
1	13	7	21	59
3	12	9	19	60
5	17	14	16	53
7	16	14	19	52
9	11	9	8	72

Table 4.2: *Data on the distribution of artifact classes by strata (row percentages) from Table 5 of Bailey et al. (1983). The classes I–IV are cores, utilized flakes, unmodified flakes and waste.*

Figure 4.3 shows the types of plot available; to the left *stacked* barplots, and to the right *clustered* barplots. Choice of orientation is arbitrary. My impression is that stacked barplots are more prevalent than clustered barplots in the literature. Clearly Class IV, waste, dominates and is fairly similar in most strata, apart from Stratum 9 where there is a noticeable increase. As a generalization, with the exception of Stratum 9, the relative importance of other classes is mostly III, I and II. Re-ordering the first two columns of data would make this, and minor exceptions, even more clear. The sample sizes for Strata 7 and 9 are much smaller than for other strata.

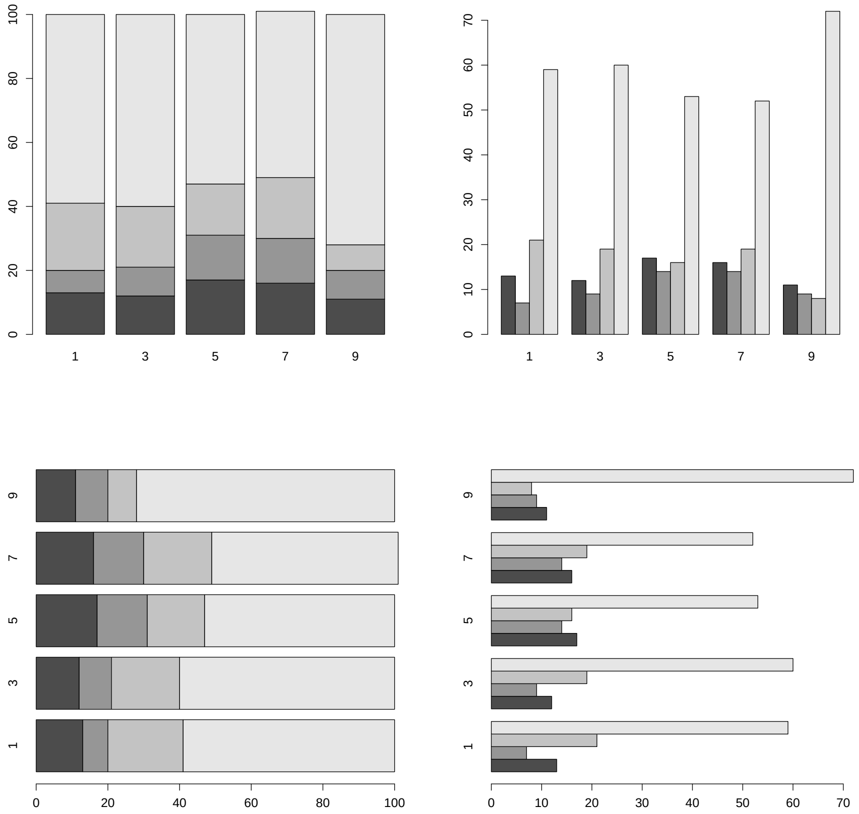

Figure 4.3: *Unannotated barplots for the data of Table 4.2; enhanced versions of the upper two charts are shown in Figure 4.4. Plots to the left are stacked and those to the right clustered; plots at the top and bottom are vertically and horizontally orientated respectively.*

Sarkar (2008: 61) suggests that the stacked barplot has limitations if one is interested in comparing proportions; if patterns are obvious a table will do; if not then making the necessary judgments may not be easy. In the barplots shown the fact that one class is dominant, and differentially so among strata, means that the less common classes are differentially 'squashed' at the bottom of the plot (in this example) so comparison across bars is not straightforward. The clustered barplot does a better job, since the dominant class can be mentally discounted and comparison among other classes is easier. If these are the main focus of interest the plot can always be produced omitting the dominant class.

By way of illustrating examples of barplot presentation that occur commonly but could be considered 'wanting', Figures 4.4 and 4.5 show, respectively, stacked and clustered barplots from R and two versions of stacked barplots from Excel.

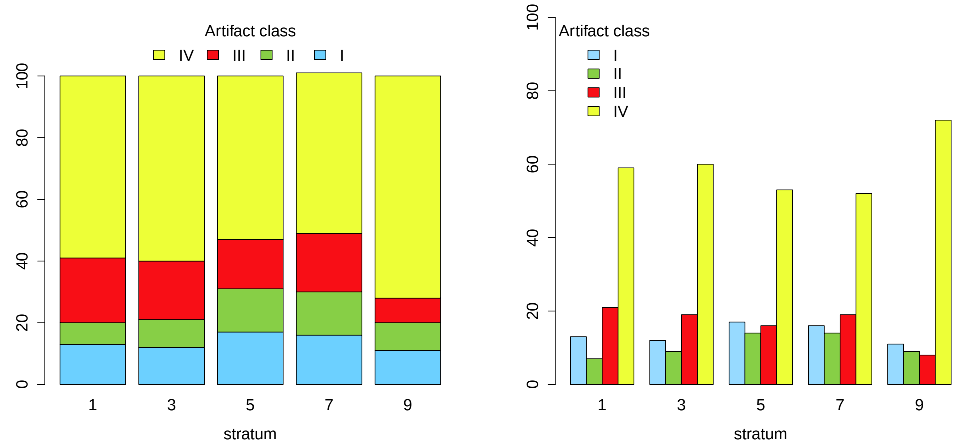

Figure 4.4: *Enhanced stacked and clustered barplots for the data of Table 4.2.*

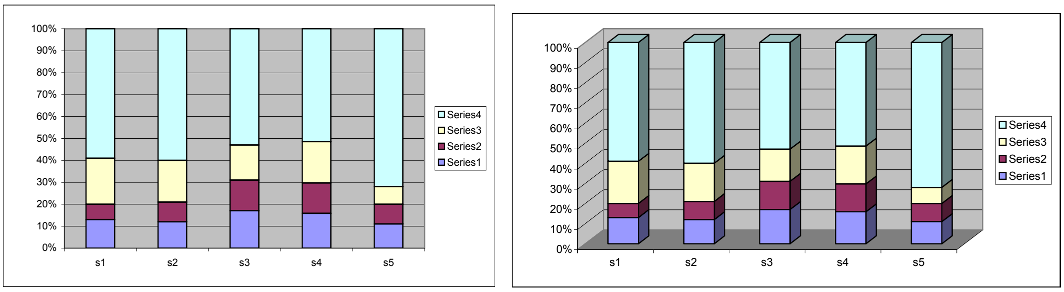

Figure 4.5: *Two- and three-dimensional* Excel *stacked barplots.*

There are several problems with the Excel barplots, which are defaults. In the two-dimensional plot the bars that should be there are innocent; the grid lines and background shading are the culprits. The lines are too prominent and there are too many of them. Coupled with the background the gridded parts also look like bars, and leap to the foreground if you stare long enough, distracting from the message of the plot. The three-dimensional plot is much worse. As with three-dimensional pie-charts, the third-dimension does not exist and should not be there. It adds, along with the grid and the 'three-dimensional' frame employed, to a variety of optical illusions. Even with the grid it is difficult to read off values from the scale. Such plots should not be allowed. If grids must be used they should be fewer than in the examples here and less obtrusive.

Barplots, of whatever variety, have probably been overused. Have users been seduced by myths such as 'every picture tells a story' or 'a picture paints a thousand words'? Tables tell stories too; in the context of tables that might be presented as barplots they require nothing like a thousand words; typically occupy less space than a graphic with commentary; and retain more precise information about actual numbers and their difference. I have particularly in mind single sets of counts with few categories (I've seen a barplot presented with just one category) and fairly small contingency tables. To see patterns, reordering of categories for both tables and plots, if the data are not ordinal, may be useful. With large numbers of categories barplots may be difficult to read, and correspondence analysis (Chapter 9) represents an alternative method of presentation.

Other methods of analyzing discrete data exist, attracting different degrees of archaeological attention. For moderate to large tables correspondence analysis is widely used as a graphical method of presentation (Chapter 9). A common method of assessing whether there is a significant association between the rows and columns of a table is the chi-squared test (Section 12.3.3). More formal modeling methods, *log-linear models*, analogous to the use of regression models for continuous data (Chapter 5) are available. Their use was explored in the 1980s and 90s and Shennan (1997: 201–13) and Baxter (203: 131–36) have brief sections on them, but I do not think they have been widely used and other than a very brief notice in Section 12.5 are not accorded further discussion.

4.3 The iniquitous pie-chart

Pie-charts are circular graphs that are divided into segments or slices whose areas are proportional to the percentages for a single set of counts. The R help notes state that 'Pie-charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data'. There is nothing to disagree with here and we proceed mainly by illustration.

Wainwright (1984: 8) shows a pie-chart of the regional distribution of scheduled monuments in England, by period. It is stated that the chart shows 49% of prehistoric date, 43% medieval, 7% Roman and 3% post-medieval. It is noted that the 'percentages are crudely derived' and given this and just four categories all that is needed is to say that prehistoric and medieval dates predominate with similar percentages, and that the other two periods are much less common.

A graphic is not needed, and if you must have one a barplot is better. The pie-chart in the paper is emulated in the top-left of Figure 4.6. There is nothing wrong with it, except that it is unnecessary and occupies over a third of a page in the original publication.

It is possible to do worse, and often is. A pie-chart is a two-dimensional construct; properly done, the areas provide a true representation of the numbers

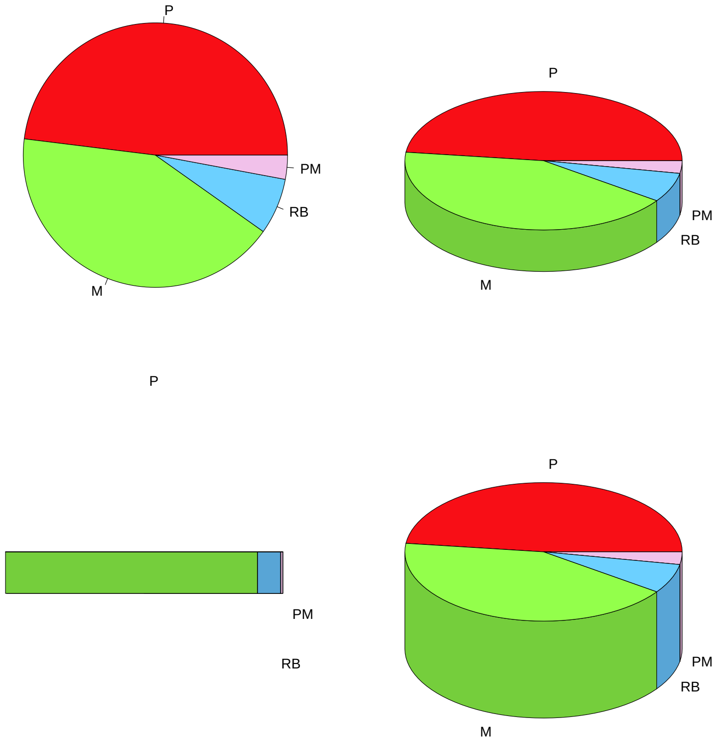

Figure 4.6: *Different 'pie-chart' presentations. The charts show the distribution of scheduled monuments by period; P = Prehistoric, M = Medieval, PM = Post-medieval, RB = Romano-British. (Source: Wainwright, 1984.)*

involved. The authors of too many publications have been seduced by the lure of 'business-type' presentations and seem to use default options in packages not designed for proper statistical use. These distort the 'truth' of data presentation and should not be allowed in academic publications. Similar comments apply to barplots, as illustrated in Figure 4.5. That the default options can be overridden is often neglected.

Some illustrations are provided in the rest of Figure 4.6. Plots like that in the upper-right are most commonly found. A meaningless third dimension, height, that is arbitrary, is added to the plot. To enable this to be seen it is necessary to tilt the plot, the degree of tilt also being arbitrary. Presumably this is done to make the chart look more 'exciting', but the price paid is to distort the information the chart is intended to display.

It might be claimed that this kind of misrepresentation does little harm, but take it to extremes! The bottom-left plot takes the tilt to its extreme and shows the chart from the side. It contains no useful information. The final plot uses a height that begins to distract from such information as is there. In using three-dimensional effects, incidentally, it can become difficult to judge the size of the smaller categories. No-one, of course, would be stupid enough to use plots like the bottom two, but if you depart from the presentation given in the top-left figure you are on the slippery slope that leads you there.

Shennan (1997: 23) suggests that the 'pie-chart is a very helpful mode of data presentation when the aim is to illustrate relative proportions of unordered categories, especially when making comparisons'. This, particularly the comment about making comparisons, is highly questionable. His following comment that pie-charts can be confusing if there are numerous categories or categories with small or zero entries is to the point.

A good case can be made for abandoning the pie-chart as a method of archaeological data presentation.

4.4 R notes

Figure 4.1

Most of the labeling arguments in the code for this and the following example are omitted.

```
histbar <- function() {
wt <- loomweights$weight
wt <- wt[wt > 90 & wt < 400]
win.graph() # histogram
hist.wt <- hist(wt))

count <- hist.wt$counts #histogram counts
mid <- hist.wt$mids     #mid-points of bins

win.graph()
barplot(count))
win.graph()
barplot(count, names.arg = mid, space = 0.7)
win.graph()
barplot(count, names.arg = mid, space = 0)
}
histbar()
```

The loomweight data of Tables B.3 and B.4, from the file `loomweights`, are used. The weights are extracted, in `wt`, using the appropriate column heading from that file and outliers, determined by prior data exploration, are then omitted. This is done by selecting weights greater than 90 g and less than 400 g, excluding six outliers outside this range[2].

As coded here the histogram needs to come first since an object `hist.wt` is created from which the bin counts and mid-points, used in the subsequent barplot constructions, are extracted. The first barplot is the default presentation; the subsequent barplots use the `names.arg` and `space` arguments to add the mid-points as axis labels and control the spacing between bars. The use of `space = 0` closes up the gaps and a histogram results. A more complicated example of this kind of usage follows.

Figure 4.2

```
hist.varbins <- function() {
z <- pillarmoulded

win.graph()     # barplot
barplot(z$Percentage, names.arg = z$Period, las = 2)

win.graph()     # histogram with unequal bin-widths
barplot(z$Percentage/z$Width, space = 0, width = z$Width, las = 2,
names.arg = z$Period, axes = FALSE)
}

hist.varbins()
```

The file `pillarmoulded` used is that based on Table 4.1 with the column headings as given there. Both plots use `names.arg = z$Period` to supply axis labels and this causes some problems in fitting all the labels on the plot, particularly with the small bar widths evident to the left of the histogram in the figure. Using `cex.names` to reduce the expansion factor is unsatisfactory because labels become too small to be easily legible. The 'solution' adopted here was to use the `las = 2` argument to produce labels perpendicular to the axis (other options can be found in the help for the `par` function).

Other than this, the first argument in the second plot 'adjusts' the percentages, dividing by width to compensate for the different duration of the periods; uses the `space = 0` argument (as illustrated in the previous example), in conjunction with the `width` argument that specifies the bin-widths to use, to produce the histogram

[2] Available logical operators include `<=` for 'less that or equal to' and `>=` for 'greater than or equal to'. Thus `wt <- wt[wt > 90 & wt < 400]` or `wt <- wt[wt >= 91 & wt <= 399]` could be used. Other operators include `==` for equality, and `!=` for lack of equality which can also be used with character variables.

desired; and uses the axes = FALSE argument to remove, in particular, the default y-axis which is meaningless given the different bin-widths.

Figure 4.3

```
Kastritsa.barplot <- function() {
z <- t(Kast) # Interchange rows and columns
win.graph(); barplot(z, beside = F)
win.graph(); barplot(z, beside = T)
win.graph(); barplot(z, beside = F, horiz = T)
win.graph(); barplot(z, beside = T, horiz = T)
}

Kastritsa.barplot()
```

The final four columns of Table 4.2 were imported into `R` as the object `Kast`. For the purpose of the analysis, which operates on the columns of the data table, the rows and columns need to be interchanged so that the columns correspond to strata. This is achieved by the `transpose` function `t()`. The default, `beside = F`, produces a vertically arrayed stacked barplot; the `beside = T` argument produces a clustered barplot; the argument `horiz = T` produces a horizontal array.

Figure 4.4

This is the code for the plot to the left. That for the plot to the right is identical except that the `beside = T` argument is used; the legend is displayed vertically at the top-left; and `ylim = c(0,100)` is used.

```
stacked <- function() {
barplot(t(Kast), names.arg = c("1", "3", "5", "7", "9"),
xlab = "stratum", legend.text = TRUE, args.legend = list(x = "top",
horiz = T, bty = "n", title = "Artifact class", cex = 1.3),
ylim = c(0,119), col = c("skyblue","yellowgreen","red","yellow"),
cex.lab = 1.3, cex.axis = 1.3, cex.names = 1.3)
}

stacked()
```

The `names.arg` supplies the names to be used for the x-axis. Note that quotation marks are used (e.g., `"1"`) so that the names are character rather than numeric variables. The legend is supplied as an argument to the `barplot` function, rather than externally, using `legend.text = TRUE`; the `list` function supplied to the args.legend controls the placement and appearance of the legend. Most of the arguments are familar from previous uses of the legend function. Placement is more explicitly declared using x = "top" which puts it at the top center with a horizontal alignment using horiz = T. The ylim = c(0,119) argument expands the range of the y-axis slightly to accommodate the legend. The R help, ?barplot, gives far more detail about the construction of barplots that could be used.

Figure 4.6

```
wainpie <- c(49,43,7,3)
wainpie.name <- c("P", "M", "RB", "PM")
Col <- c("red", "greenyellow", "skyblue", "pink")

Pie <- function() {}
library(plotrix)

win.graph()
pie(wainpie, wainpie.name, col = Col , radius = 1)
win.graph()
pie3D(wainpie, labels = wainpie.name, col = Col)
win.graph()
pie3D(wainpie, labels = wainpie.name, theta = pi/2, col = Col)
win.graph()
pie3D(wainpie, labels = wainpie.name, height = .7, col = Col)
}

Pie()
```

The pie function in R very properly does not encourage the use of three-dimensional pie-charts (or pie-charts at all, for that matter), for which the pie3D function from the plotrix package was used. The first three lines provide the four numbers from which the chart is constructed, shortened names indicating the period of construction of the monuments that are the focus of analysis, and colors to be used for plotting. The arguments theta and height control the angle of view and the ‘depth’ of the artificial third dimension.

Chapter 5

Regression analysis

5.1 Linear regression analysis

5.1.1 Introduction – an example

The methods that have been examined so far are mostly descriptive and/or exploratory. All the methods covered have been widely used in archaeological applications. Regression methods – the focus of this chapter – have also been widely used (Baxter, 2003: 50–65). Such methods require a model to be formuated for the data, representing an important departure from most of what has gone before.

At its simplest, and stripped of context, the starting point in treatments in introductory texts is that of finding a 'best-fitting' straight line through a 'cloud' of points displayed in a bi-variate scatterplot. This is mathematically easy but, from the point of view of the average end-user, detailed knowledge of the mathematics is unnecessary. Other than in taught courses based on texts that illustrate hand calculations it is doubtful that anyone does anything other than use software to obtain results. This, once a data file is created, can be accomplished almost immediately. This frees the user to concentrate on the more interesting and challenging problems of model formulation and model interpretation.

These matters are discussed below and involve more use of mathematical notation than has hitherto been the case. To fix some initial ideas, an example is first presented. The data used are those of Table B.5, named `pmedwine` in `R` (from Robertson 1976), and are for six variables descriptive of the morphology of 49 post-medieval sealed wine bottles of known date.

It is clear that morphology changes over time and we shall suppose that interest lies in developing a model that can predict the date of undated bottles from their morphology. We shall further suppose that a simple (linear) regression model with just one variable as a *predictor* is sought. The full data set is examined in more detail in Example 1 of Section 5.2 where it is clear that body height, `BH` is likely

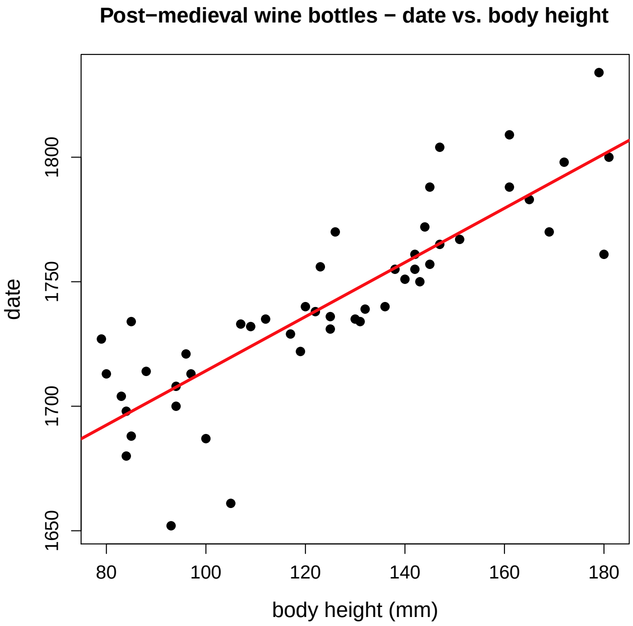

Figure 5.1: *A linear regression fit superimposed on a plot of date against body height for the data of Table B.5.*

to be the best single linear predictor of date.

Figure 5.1 is a plot of date against body height with a linear fit superimposed. It is straightforward to produce this. Omitting presentational arguments, and assuming that variables `BH` and `date` have been previously created, use

```
plot(BH, date)
fit <- lm(date ~ BH)
abline(fit)
```

where `lm` is the linear modelling function that fits the model required and saves the result, in this example, in the object `fit`. The `abline` function adds the fitted line to the plot. These are discussed in more detail in Section 5.4.

Once this is done, we would minimally like to know what the fitted line is and how well it fits the data. Execute the `summary` function, using `summary(fit)` to get the following (deleting some of the output).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.605e+03  1.282e+01  125.18  < 2e-16 ***
BH          1.088e+00  9.977e-02   10.91  1.8e-14 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.46 on 47 degrees of freedom
Multiple R-squared: 0.7168
F-statistic:   119 on 1 and 47 DF,  p-value: 1.801e-14
```

Under `Coefficients`, the estimated model for predicting date is given as

$$1605 + 1.088\ \texttt{BH}$$

with a goodness-of-fit of $R^2 = 71.7\%$, which is the `Multiple R-squared` expressed in percentage terms and rounded to one decimal place (See Section C.2.1 for more on R^2 and Sections 12.2.2 and 12.3.2 for the interpretation of p-values and the `F-statistic` that appear in the output.). This anticipates the fuller discussion provided in Sections 5.1.3, but in this context would usually be regarded as 'reasonable'. The main point about introducing the idea here is to show how easily such basic information is extracted.

With simple linear regression, other than provision of the line of best fit, inspection of the graph is often as or more informative than the basic output. For example:

- It is clear that a reasonable, though not perfect, linear fit will be obtained.
- It is evident that at the lower (less than 105) and higher (greater than 155) body heights prediction is less good than at the intermediate heights; that is, the variation about the fitted line is greater.
- The two earliest dates are badly over-predicted by the model and stand out as potential *outliers*.

Most of this is not evident from the numerical analysis to date. It can be taken further to address some of the issues involved. This is dealt with in Section 5.1.3 after a discussion of models, terminology and notation.

5.1.2 Regression models and notation

Begin with n observations on a *dependent* variable, y, and a single *independent* variable, x. A plot of y against x will suggest whether or not there is a relationship between y and x, its nature, and whether or not any of the data are unusual. Typically there will be some deviation from an exact mathematical relationship, attributable to what is conceived of as random *error* or *variation*. The most familiar model, the *simple linear regression model*, can be written

$$y = \alpha + \beta x + \varepsilon \tag{5.1}$$

where α and β are unknown *parameters* and ε is an unobserved *error* term[1].

The model specified is an *additive* one. By this it is meant that the model is *linear in the parameters* and the error term added as shown. The parameters, α and β, are the intercept and gradient (or slope) of the line. The intercept is the value of y when $x = 0$. The gradient is the change in y that occurs when there is a unit change in x, and is dependent on the units of measurement.

The simple linear model looks restrictive, but it can be extended in various ways, for example, to a two-variable regression model of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{5.2}$$

One special case is when $x_2 = x^2$, a quadratic term, so that model (5.2) becomes

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{5.3}$$

where the systematic component is a non-linear quadratic function that can be used to model this kind of non-linear pattern. The regression model is linear because of the linearity in the parameters. These are examples of *multiple regression* models. It can obviously be extended by adding cubic, quadratic terms and so on.

Another extension is when the data are split into two groups, and interest lies in seeing if and how relationships vary between groups. Define a variable z to have the value 0 for one group and 1 for the second group, an example of what is called a *dummy* or *indicator* variable. The model

$$y = \alpha + \beta_1 x + \beta_2 z + \varepsilon \tag{5.4}$$

has the effect of fitting two parallel lines through the data for the two groups. If a further term $(xz) = x \times z$ is defined the model

$$y = \alpha + \beta_1 x + \beta_2 z + \beta_3 (xz) + \varepsilon \tag{5.5}$$

simultaneously fits separate regressions for the two groups. The use of these models is illustrated, with further discussion in Example 3 of Section 5.2.

[1] The expression $\alpha + \beta x$ is the mathematical expression for a perfect straight line and is the *systematic* component of the model, in contrast to the *random* error; real data almost never follow such a line and to express this the error term is added to model the scatter about the line. The terms *dependent*, *independent* and *error* are hallowed by usage, and used here, but the terminology has been queried. The use of *dependent* might be taken to imply that the relationship is a causal one. This is sometimes the case, but if regression is used for description or prediction, for example, there is no implication of causality, and terms such as *regressor* and *regressand* have been used as an alternative. The term *predictor* has also been used above for the independent variable. Similarly, the 'error' need not be an error (of measurement or model mis-specification, for example) but may represent natural random variation about the dependent variable. The more neutral term, *disturbance*, is sometimes preferred.

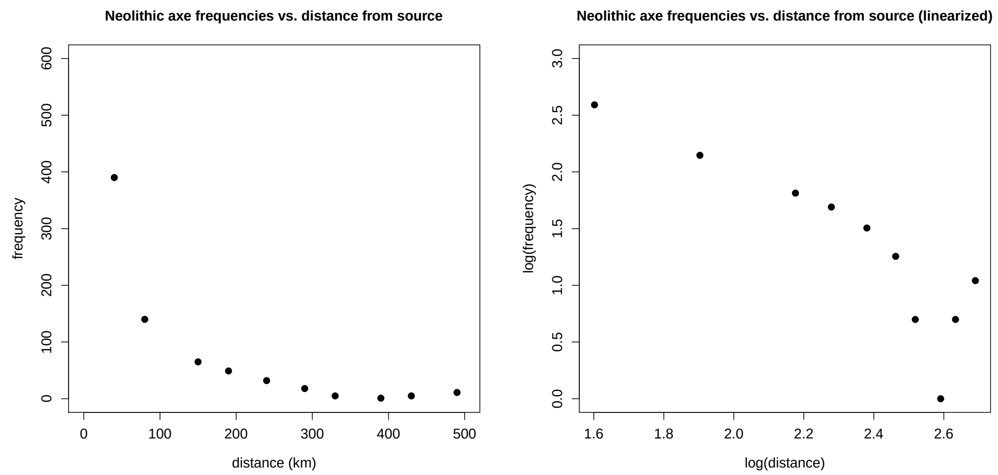

Figure 5.2: *The stone axe distance-decay data of Table 5.1 before and after a log-log transformation.*

To motivate further discussion two small data sets are shown in Tables 5.1 and 5.2. These use data of a similar kind, both measuring the frequency of artifact types found at different distances from a source of production or distribution center. Plots of frequency against distance typically show a *distance-decay* pattern with frequency declining to zero at some distance from the source. Linear models are inappropriate as they will result in negative predictions at some point. Two simple distance-decay models are explored below.

Table 5.1 is based on Cummins (1980) and shows the frequency of Neolithic stone axes at different distances from a distribution center. The data have been reconstructed from Figure 7 in Cummins, which is on a log-log scale.

Frequency	390	140	65	49	32	18	11	5	5	1
Distance	40	80	150	190	240	290	490	330	430	390

Table 5.1: *Frequency of neolithic stone axes at different distances (km) from a distribution center (Cummins, 1980).*

A linear regression model (5.1) will be fitted after data transformation where y is the logarithm of frequency and x is the logarithm of distance (Section 5.2, Example 2). Figure 5.2 shows plots of the data before and after the log-log transformation, and before undertaking the regression.

Although Cummins does not discuss this, the implicitly assumed model for the untransformed data has the form

$$y' = \alpha' x'^{\beta} \varepsilon' \tag{5.6}$$

where $y = \log y'$, $x = \log x'$, $\alpha = \log \alpha'$ and $\varepsilon = \log \varepsilon'$ show the correspondence with the notation of model (5.1). This is called a *power-law* model and is an example of a multiplicative model where ε' is a multiplicative error term. It is also an example of a *linearizable* model.

The appearance of the plot for the untransformed data suggests that a smooth model of distance-decay is reasonable. After the 'linearization' it is noticeable that there are departures from linearity at the longer logged distances. There is a clear outlier that may cause problems in fitting a linear model to the transformed data. This is discussed in detail in the continuation of the example in Section 5.2.

As a second and similar example that poses different problems in analysis, data from Morris (1994) are used. These are based on Figure 2A of that paper and are given in Table 5.2. The table shows the frequency of Late Iron Age pottery found at different distances from a production source. Morris did not examine the data in the way it is to be treated here.

Distance	4	18	21	22	23	27	30	34	36	43	52	62
Frequency	80	61	41	17	18	8	6	43	2	3	3	1

Table 5.2: *Frequency of Middle-Late Iron Age pottery at different distances (km) from a source (Morris 1994).*

In contrast to the power-law model, an *exponential decay* model will be used which has the form

$$y' = \alpha' \exp x^{\beta} \varepsilon' \tag{5.7}$$

which is linearizable after a (natural) logarithmic transformation where, in the simple linear regression model (5.1), $y = \log y'$, $\alpha = \log \alpha'$ and $\varepsilon = \log \varepsilon'$. As with the data from Cummins (1980) the data before and after transformation are shown in Figure 5.3.

The most obvious feature of this is a clear outlier in both plots (that Morris does not discuss). It would be legitimate from a model-fitting perspective to omit this from the outset and seek an explanation for it, but it is retained in some later analyses for illustrative purposes. The transformation to 'linearity' is not especially impressive, and this will be explored further in the continuation of the example in Section 5.2.

Finally, note that models of the kind, with a non-linear systematic component and additive error, such as

$$y = \alpha' \exp x^{\beta} + \varepsilon \tag{5.8}$$

cannot be simply linearized.[2]

[2] For those unfamiliar with logarithms, terms of the form ab can be transformed as $\log ab = \log a + \log b$ but this is not possible for $a + b$. The systematic component *can* be linearized, and this is an example of a *generalized linear model*, which is beyond the scope of the present notes.

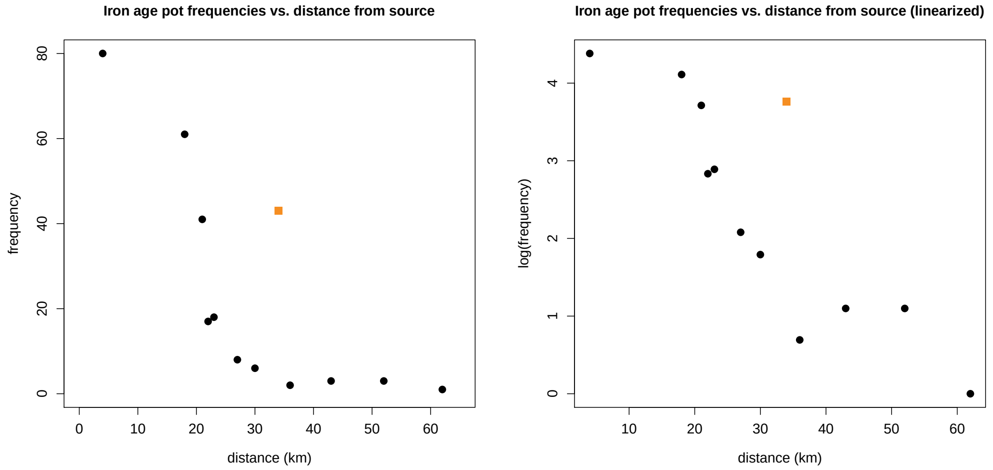

Figure 5.3: *The pottery distance-decay data of Table 5.2 before and after a log-transformation of the frequency. An obvious outlier is highlighted.*

5.1.3 Model checking

More notation and terminology

The general aim in regression analysis is to say something useful about the unknown systematic component in the model with, in simple linear regression, the focus often being on β. This requires *estimation* of the parameters that, in the light of the unknown errors, is 'sensible'. This can be done in more than one way.

It is important to distinguish between the 'theoretical' model, in which the parameters and error term are unknown, and the fitted model. The latter can be written as

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{5.9}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated parameters and $\hat{y}$ is the fitted/predicted value of y using these estimates. With this in place $\hat{\varepsilon}$ defined as

$$\hat{\varepsilon} = y - \hat{y}. \tag{5.10}$$

is the estimated error or *residual*. The latter term is used to distinguish the observable estimated errors from the unknown true errors[3].

Parameters must be estimated. The default, often the only one in the texts and software used by archaeologists, is the *method of least squares*. This involves determining the estimates to minimize the sum of squared residuals. These estimates

[3] The use of Greek letters for unknown parameters, surmounted by a circumflex or 'hat' for their estimates, is a common convention; other conventions can be used.

have ‘optimal’ properties under certain error assumptions – most importantly that they have a normal distribution with constant variance and are independent. It is not, however, necessary for linear regression to be useful for the errors to be normally distributed, as is sometimes incorrectly asserted, so long as they are reasonably ‘well-behaved’.

It is convenient here to introduce the idea of *correlation*. The data are to be regarded as a sample from a population, and the correlation coefficient in the population, ρ, is estimated as r (see Appendix C, and texts such as Shennan (1997: 140) for the mechanics of calculation). The correlation is a measure of the strength of the *linear* relationship between x and y, with 1 showing a perfect positive linear relationship and -1 a perfect negative one. In the context of simple linear regression the square of r, perhaps confusingly called R^2, often expressed as a percentage, is used as a measure of the ‘success’ of the regression, with values close to 1 ‘good’ and close to zero ‘bad’. It is called the *coefficient of determination* with the interpretation that it is the amount of variation in y ‘explained’ by variation in x. One reason for the notational distinction is that the definition of R^2 extends to linear models with more than one independent variable, and with the same interpretation.

Diagnostic statistics

It is desirable to check the validity of the error assumptions in assessing the fit of the model. While R^2 provides a global meaures of fit it can be more helpful to identify unusual observations or patterns in the data that compromise the adequacy of the model. A plethora of statistics have been developed for this purpose; many are just variations on similar themes and only some have garnered reasonably widespread use. The reader should be warned that for simple linear reqression the use of these statistics can be superfluous, since what they tell you can be obvious from graphical examination. They do, however, extend to more complex models where potential problems may be much less evident.

The error terms are unobservable but their properties are ‘mimicked’ by the observable residuals. Cases can be unusual becauses the value of the dependent variable is an *outlier*, having a ‘large’ residual. They can also be unusual because the value of the independent variable is ‘extreme’; such cases are said to have high *leverage*. Outliers generally need some attention; cases with high leverage can actually be beneficial in fitting a model, but this depends on their other characteristics. Neither need affect the fitted model much; cases whose removal has an undue effect on the parameter estimates are said to have a large *influence*.

Dealing with leverage first, often denoted by h_i or m_i and lying between $1/n$ and 1, it measures how distant a case is from the centroid of the ‘point’ cloud of independent variables. For simple linear regression where the number of independent variables $p = 1$, points of high leverage will stand out at the extremes of the scatter. Mathematically it is sensible to base a formal measure of the leverage of case i on the distance of the case from the mean, $(x_i - \bar{x})$, and the mathematics leads to a measure based on the square of this scaled to have a maximum of 1. For $p = 2$ recourse is needed to matrix algebra, but the measure obtained does a similar job. Rules-of-thumb exist for deciding what is an extreme leverage, but an index plot (of h_i against i) is often most useful.

Cases with high leverage are sometimes called *outliers*, but the term is best reserved for cases with large residuals. These can be defined in various ways. The *raw* or *ordinary* residuals, $\hat{\varepsilon}$, form a sensible starting point for detecting outliers, but are scale dependent. Rescaling is achieved by dividing $\hat{\varepsilon}$ by an estimate of its standard deviation s, where s^2 is an estimate of the assumed common error variance σ^2.

Let s^2 be this estimate using all the data, with $s^2_{(i)}$ an estimate omitting the ith case (which will differ for each i). Terminology is confusing. What have been called standardized residuals can be defined as $\hat{\varepsilon}/s$, and were what was available in older software. We follow the usage of Venables and Ripley (2002: 151) and define *standardized* residuals as

$$r_i = \hat{\varepsilon}/s\sqrt{(1 - h_i)}.$$

The term *studentized* residuals will be used for

$$t_i = \hat{\varepsilon}/s_{(i)}\sqrt{(1 - h_i)}.$$

See Cook and Weisberg (1982: 20) for the mathematical relationship between r_i and t_i. The r_i have also been called *internally studentized* residuals; the t_i have variously rejoiced in the names *studentized*, *externally studentized*, *jacknife* or *deleted-t* residuals, the last being used in the MINITAB package.

Although residuals mimic the properties of the errors they do not have exactly the same properties. In particular their standard errors depend on h_i in the manner indicated. If a case is an outlier it will inflate the estimate of s and $s_{(i)}$ will be somewhat smaller, so $t_i > r_i$. The general idea is that, for large enough samples, the distribution of the scaled residuals should mimic the assumed (usually) normal distribution of the errors. This means that, keeping the numbers simple, values in excess of 2 (in absolute value) can be regarded as 'unusual', and values in excess of 2.5 or 2.6 as 'very unusual'. For small samples the t-distribution can be used to define such 'rules-of-thumb'. For specialized situations more exact theory exists, but in practice it is generally more useful to inspect a plot of the scaled residuals against the fitted values rather than relying on rules-of-thumb.

An illustrative example for the post-medieval bottle body heights and date, following the regression illustrated in Figure 5.1, is provided in Figure 5.4, where standardized and studentized residuals are contrasted with reference lines at ±2.5 shown (only the negative value being relevant here).

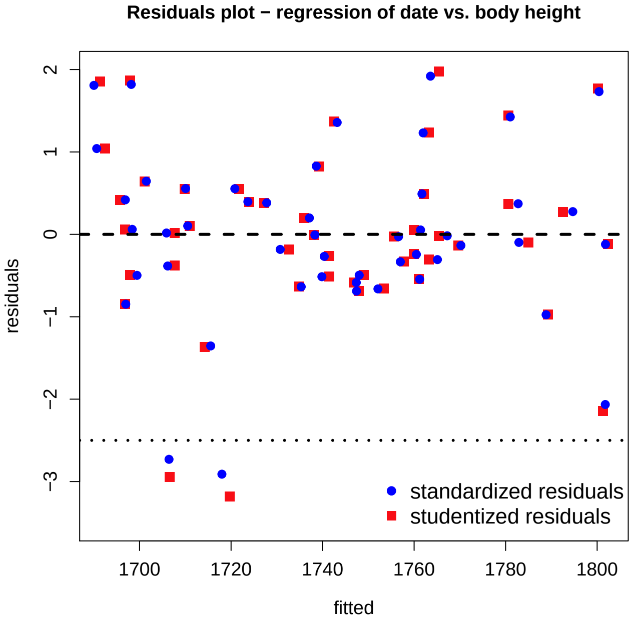

Figure 5.4: *Residuals from the regression fit of Figure 5.1.*

The two kinds of residual are largely coincident in their values and the standardized residuals have been 'jittered' to avoid overwriting the studentized residuals (see Section 5.4). The exceptions to this are two cases at the bottom left of the plot where the studentized residuals are larger than the standardized residuals, though both residuals suggest the cases are outliers. These are just the two bottles with the earliest dates, which are predicted to have somewhat later dates than the known values.

Of measures of influence available Cook's statistic

$$d_i = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}$$

is the most commonly used. It is a function of residual and leverage statistics and large values will often have large values of one or both statistics, though this is not inevitable (remember h_i is bounded above by 1). Conversely, large values of one statistic will not necessarily give rise to a large d_i if the other is small. The statistic has an interpretation as the distance between parameter estimates with and without case i, or, equivalently, the distance between predicted values. An index plot is useful for making judgments about what is 'large'.

Figure 5.5 shows index plots of the leverage statistic h_i, and Cook's statistic for the regression of body height against date.

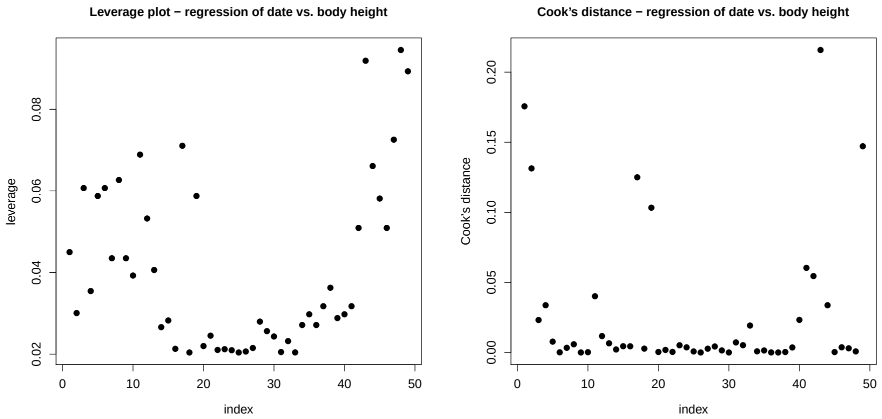

Figure 5.5: *Diagnostic index plots from the regression of* `Date` *against* `BH`*. Cook's distance to the right and leverage, h_i, to the left.*

The plots don't give too much cause for concern. The shape for h_i is to be expected as the index corresponds to a natural ordering (by date) with the extremes at either end. The plot for d_i has some values larger than others (some have to be), but nothing 'shouts out' as seriously extreme. Omitting the first two cases that were suggested as outlying in Figure 5.4 increases R^2 by about 5% without introducing further noticeable problems.

Other than the last case, the gap between the second and third earliest dates, of 19 years, is noticeably larger than for other adjacent pairs (ordered by date). The last case has nothing else unusual about it. Reporting results omitting the first two cases, on the basis that they are early and untypical, is a sensible option. A plot of the residuals against *Date*, which is sensible though not shown here, suggests even more starkly that the first two cases are untypical.

5.1.4 Inference

Intentionally, not too much has been said about traditional methods of statistical inference at this point. There are differing points of view about its value for archaeological data analysis; mine is that its importance has been exaggerated, partly for historical reasons (see Chapter 12). Some engagement with ideas is needed with model-based methods, however, if only to interpret output provided by software.

As far as regression goes, two concepts need to be distinguished, that of *statistical significance* of the regression fit, and *goodness-of-fit* which has been covered in the foregoing discussion. For $p = 1$ the hypothesis that $\beta = 0$ (i.e. there is no linear relationship) is notionally of interest but formally testing this, using significance tests, is often a waste of time, since it will either be obvious that the regression is significant, or that it is too poor to be of substantive interest. Some acquaintance with *p-values* is desirable and discussed in context below and in Chapter 12. Repeating the analysis previously reported, with some editing

```
           Estimate SE  t-value  Pr(>|t|)
Intercept   1.605   13    125   0.0000 ***
BH          1.088  .01   10.9   0.0000 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared: 0.7168
F-statistic:   119 on 1 and 47 DF,  p-value: 0.0000
```

allows us to infer from the very small p-values that, with the t-distribution as a reference, the hypotheses that $\alpha = 0$ and $\beta = 0$ are unsustainable. The F-statistic does the same job as the t-statistic for β (`BH)` in this instance but this is not the case if there are two independent variables or more. The conclusions are obvious from graphical inspection alone.

For $p > 1$ things are more complicated and interesting[4]. The significance of the regression is often obvious, but it is not always clear which variables are important, so testing the importance of subsets of variables is of interest. An illustration is provided in Example 3 in the next section.

5.2 Examples

Example 1 – Post-Medieval wine bottle dimensions

A *pairs plot*, also sometimes called a *scatterplot matrix*, of the post-medieval wine bottle data is shown in Figure 5.6 omitting `Type`, and `Height` which is the sum of neck height, `NH`, and body height, `BH`. This was obtained using the `pairs` function and was part of the preliminary data analysis that informed the choice of body height as the independent variable in the analyses that are the subject of Figures 5.1, 5.4 and 5.5.

It seems clear that `BH` will be the best single linear predictor of date, and a reasonable fit can be anticipated. The relationship of `Base` to `Date` is interesting; the pattern is distinctly non-linear but suggests a positive relationship for earlier dates and a negative one for later dates. The issues are pursued in Section 5.3.

[4] p as notation for the number of independent variable and p-value need to be distinguished.

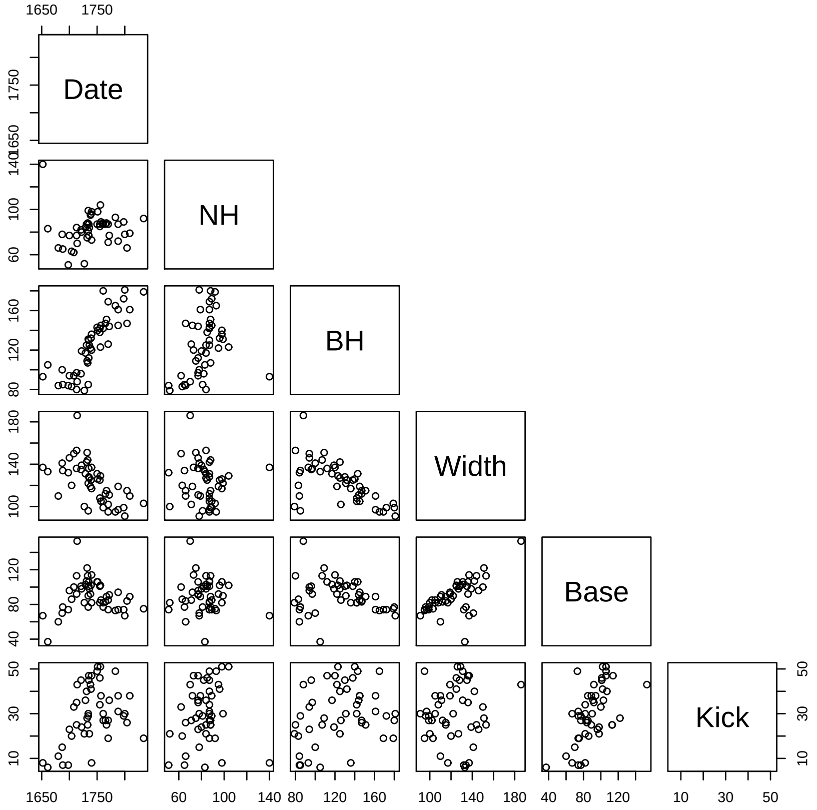

Figure 5.6: *Post-medieval wine bottles - a pairs plot for selected variables.*

Example 2 – Linearizable models

This continues the analyses of the data from Cummins (1980) and Morris (1994), begun in Figures 5.2 and 5.3. Specifically, a variety of fitted models are added to the plots previously presented, with commentary.

Cummins (1980) preferred log-log model, chosen presumably on the basis of visual interpretation, was to use just the first five observations. This was interpreted as a change in 'regime' at the associated distance and is shown as the dotted blue line in Figure 5.7. The dashed red line shows the fit using all the data.

Apart from the small number of observations used to fit the preferred model the interpretation is questionable on statistical grounds. Transforming the model using all the data back to the original scale results in a fit in the left-hand plot that is virtually indistinguishable from the data. On the log-scale, with all the data, the tenth observation for which the frequency is 1 is highlighted, with $t_i = -4.1$. If 1 is changed to 2 then $t_i = -2.1$. That is, the evidence for a 'boundary effect' resides

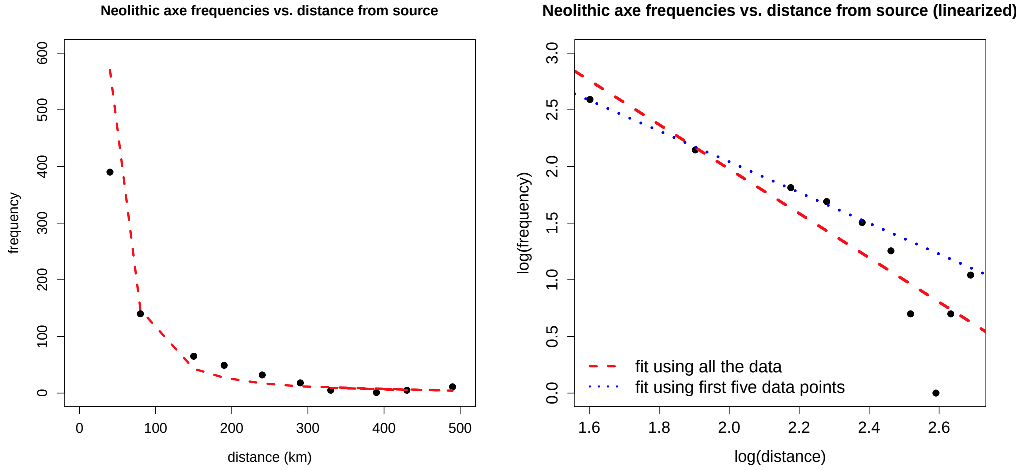

Figure 5.7: *The plot to the shows linearized model fits for the Cummins (1980) data, one using all the data and the other the first five observations. The plot to the left transforms the model fit using all the data back to the original scale. See the text for a full discussion.*

with a single observation, and its importance is almost certainly attributable to the small frequency and use of a log-transformation. The estimates of β differ by 0.24 with standard errors of 0.38 and 0.22. The estimates are not independent, but this is convincing evidence that they do not differ significantly as the standard errors would need to be much smaller to suggest a significant difference. The conclusion has to be that there is little statistical evidence for a boundary effect, with the attendant lesson that relying on visual interpretation in the presence of small samples is unsafe.

Turning to Morris (1994), following the initial presentation of the data in Figure 5.3, the right-hand plot of Figure 5.8 shows the fits for the exponential model, linearized as in model 5.7, with and without the outlier. Case 8 has $|t_i| = 2.79$. Omitting the outlier changes the fit from 73% to 84%, but has virtually no influence on $\hat{\beta}$, which can been seen visually. This is because the omitted case has very low leverage as it is centrally placed along the x-axis.

The exponential model was used for illustration. The linear transformation does not do a good job of 'straightening out' the plot, suggesting that a power-law model might have been better. If, omitting the outlier, this and the exponential model are fitted, the left-hand plot suggests the power-law model does a better job of fitting the data other than the first observation. Interestingly, for the linearized power-law model, case 1 has a larger value of $|t_i|$ than case 8, 2.46 compared to 2.36.

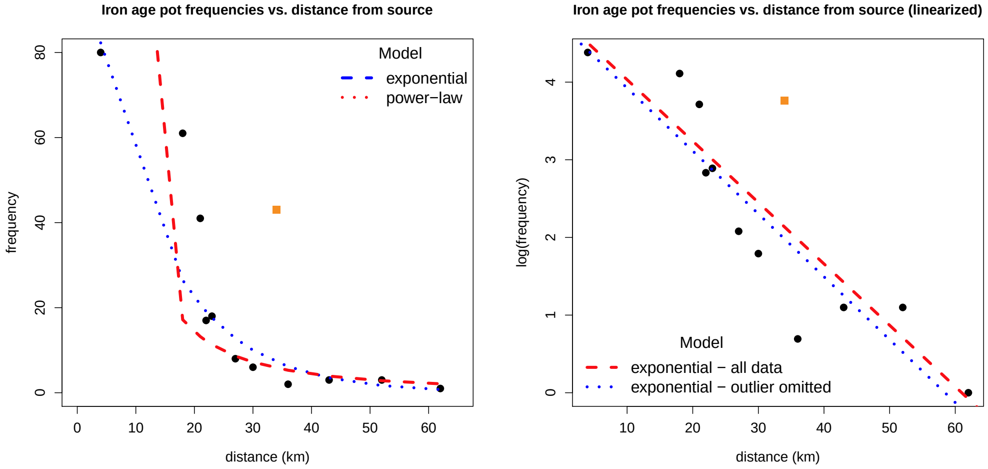

Figure 5.8: *The plot to the right shows linearized model fits for the Morris (1994) data, both omitting and including the outlier. The plot to the left transforms the model fits for the exponential decay and power-law models, omitting the outlier, back to the original scale. See the text for a full discussion.*

There are, perhaps, two lessons here. One is that a large residual with low leverage need not affect the fitted model much. The other is that the two models can give rise to noticeably different results, theory not always providing a guide to choice .

Example 3 – Stone 'circle' dimensions

Barnatt and Moir (1984) present and analyze data from Thom (1967) on the dimensions of stone 'circles'. These are not usually exact circles, some deviations from true circularity being greater than others, The difference between maximum and minimum diameters is used as a measure of deviation. Figure 5.9 reproduces Figure 3 of Barnatt and Moir using 69 circles with diameters less than 160 ft and deviations less than 20 ft (Table B.6). A distinction is made between northern circles (open triangles) and southern circles (closed triangles). Separate regressions are fitted to each subset, that for the northern data having the steeper slope.

Barnett and Moir (1984: 210) draw several conclusions from this. They claim, presumably on their interpretation of the visual evidence, that the regression lines are distinct, and that in southern England circles tend to be constructed more accurately. These conclusions need to be qualified. It is questionable whether linear regression should be used at all and this is pursued in Section 5.3.

To begin, however, the analysis shown in Figure 5.9 is examined more closely. Model (5.5) is used, defining a dummy variable $z = 1$ for southern circles and 0 otherwise. Bearing in mind that z can only take these values it can be seen that for northern circles the intercept and slope are (α, β_1) and for southern circles $(\alpha + \beta_2, \beta_1 + \beta_3)$. A test of the hypothesis that $\beta_3 = 0$ tests whether regressions have the same slope. If this is not rejected, drop (xz) from the model and refit it; a test of the hypothesis that $\beta_2 = 0$ then indicates whether the separate intercept is needed or not. As opposed to this sequential testing, a simultaneous test of the hypothesis $\beta_2 = \beta_3 = 0$ using ANOVA (analysis of variance) methods is also possible (Section 12.3.4).

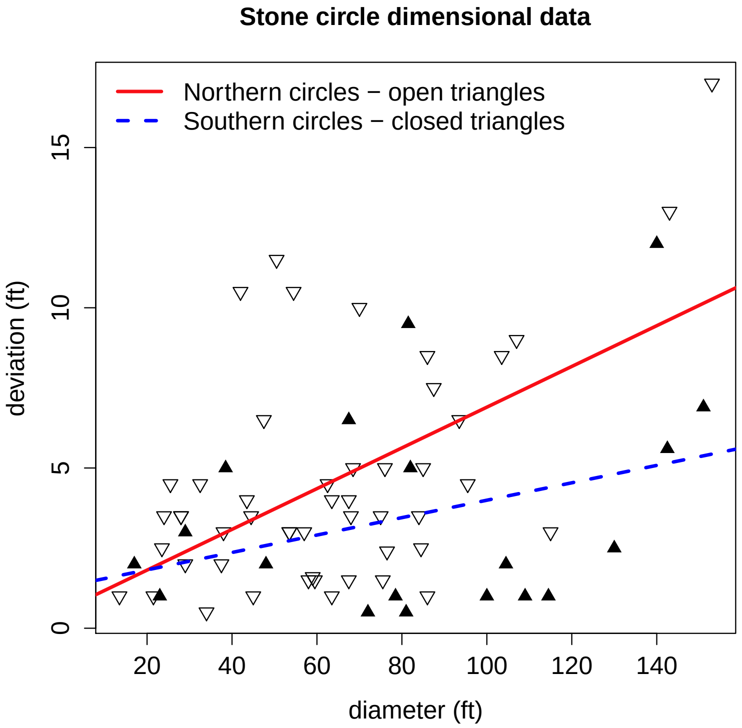

Figure 5.9: *The plots are based on data for 69 stones circles. Circles are divided into northern (open triangles) and southern circles (closed triangles) with simple linear regressions fitted separately to the two regions. See the text for a full discussion.*

Whichever approach is adopted there is no real evidence at the levels usually used to suggest, on the basis of the methodology in the paper, that regressions for the two regions differ significantly. This is shown by the tests just described (details not presented); it can be seen more informally as follows. Barnatt and Moir (1984) fit their regressions separately (i.e. they use two simple linear regressions, rather than the interaction model used here). This gives a difference in slope estimates of

0.036. The standard error of $\hat{\beta}$ for the southern data is 0.018 so that the northern estimate lies within two standard errors of the southern estimate ignoring error in the latter. It would need to be larger to show a significant difference (neither intercept estimate differs significantly from 0, as is to be hoped for). If allowance is also made for the standard error of the slope estimate for the northern data of 0.014 it becomes even clearer that the regressions are not significantly different.

That this conclusion differs markedly from that in the paper is because no allowance was made there for the uncertainty of estimation. This is relatively large because of the noticeable variability in the data. The example is pursued using non-parametric regression methods, in the next section.

5.3 Non-parametric regression

The models used so far have the form $y = f(x) + \varepsilon$ where $f(x)$ is a function linear in the parameters. Linearizable models that reduce to this form have been illustrated. It is possible to define models with rather simple functions $f(x)$ that cannot be linearized so that non-linear least squares estimation, for example, is required. Such models use explicitly defined forms of $f(x)$ usually dependent on just a few parameters.

An alternative approach, rather than assuming a parametric model, is to allow the data itself to determine a form for $f(x)$. This gives rise to the idea of *non-parametric regression* or *scatterplot smoothing*. These methods have been used much less than linear methods in archaeology; Baxter (2003: 63–65) lists some examples available at that date; this section provides a discussion of some possibilities, with application.

The methodology can be viewed as conceptually simple and mathematically complicated but, with the aid of R, relatively straightforward to implement. This qualifies the methodology as 'simple' in the sense defined in Section 1.1 but caveats need to be entered. There are a lot of approaches that have been developed – Venables and Ripley (2002: 229) illustrate six. Choices within each approach need to be made that determine the degree of smoothing so that many different estimates are possible and, with much variation in the data, selection and interpretation of an estimate is not straightforward. Accounts of non-parametric smoothing geared to R include Venables and Ripley (2002: 228–232) and Faraway (2006: 211–230); Simonoff (1996: 134–214) provides a general presentation. Ambitions here do not extend beyond providing an intuitive account of the method of *loess* smoothing, with examples. Faraway (2006: 228) suggests that the loess smoother is a good all-purpose smoother.

Example 4 – Non-parametric regression of stone 'circle' dimensions

The idea is illustrated in Figure 5.10. For the circle data and the two regions separately a smooth, using the defaults in the `loess.smooth` function, has been fitted. This is discussed in more detail in Section 5.4. The departure from linearity is sufficient to suggest that the original fitting of linear models is not appropriate (a conclusion that might be reached by simply looking at the pattern of scatter involved in the two plots).

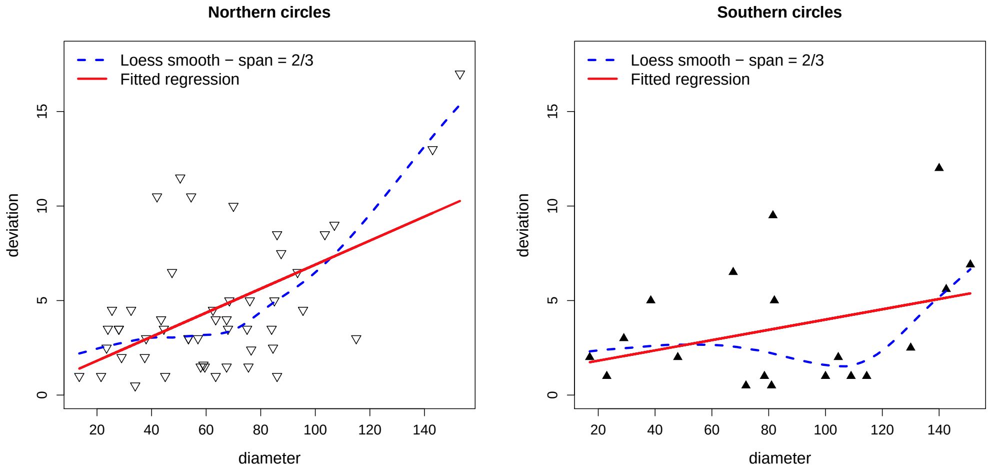

Figure 5.10: *The plots are for the northern and southern circles fitted separately and showing loess smooths as well as the fitted regression.*

The loess method works by defining a *neighborhood* of points (or *window*) about each value of x and fitting a linear or quadratic regression model within each neighborhood to predict the smoothed values at x. Cleveland (1979) provides a detailed technical account. A complex iterative weighted regression procedure is used for fitting (Baxter 2003: 65). Loosely speaking, a complicated kind of average is computed for each neighborhood, and the resultant points are 'joined up' to get the smooth. The appearance and smoothness of an estimate is dictated by the size of the neighborhood, determined by the `span` in the `loess.smooth` function, and precise fitting procedure. Section 5.4 discusses other arguments available.

Example 5 – Varying neigborhood size in non-parametric regression

Figure 5.11 uses the stone circle data for the southern region. The main purpose is to illustrate the variation that can arise with different levels of smoothing; defaults in the `loess.smooth` function have been used, other than that the span is varied.

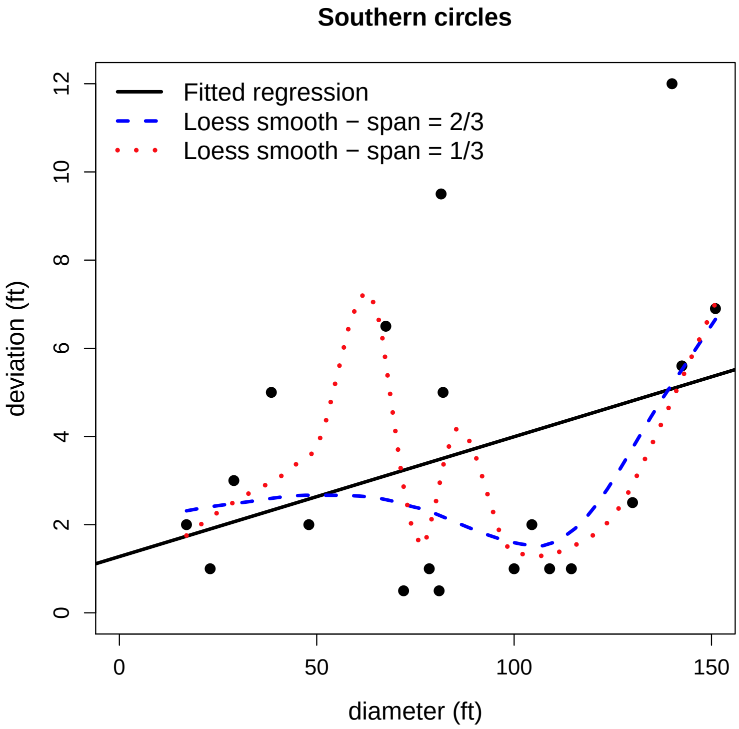

Figure 5.11: *Non-parametric regressions for the southern circle data showing the effect of varying the span.*

The solid line shows the linear regression fit. The dashed line shows the fit based on the loess smoother with a span of 2/3; the dotted line uses a span of 1.3. The larger values of a span produce larger neighbourhoods and hence greater smoothing. The choice of the degree of smoothing has a limited effect for larger diameters of more than 100 ft; for smaller diameters the results are highly sensitive to the choice. Both spans suggest that simplifying to a liner model is inappropriate.

Example 6 – Non-parametric regressions of the post-medieval wine bottle data

For a final example, and to make some slightly different points, we return to the post-medieval wine bottle data, the pairs plot for which was shown in Figure 5.6. With *body height* as the independent variable, smoothed fits are presented using four different dependent variables.[5]

In the analyes the `loess.smooth` function with defaults is used. This explains

[5] Other than for date, treating body height as the 'independent' variable is a convenience. The interest lies more in description than prediction.

why some obvious outliers have little effect on the smooths since their effect is downweighted (see the notes for Figure 5.10 in Section 5.4). Varying the level of smoothing between 1/3 and 2/3 has little effect on the fitted models.

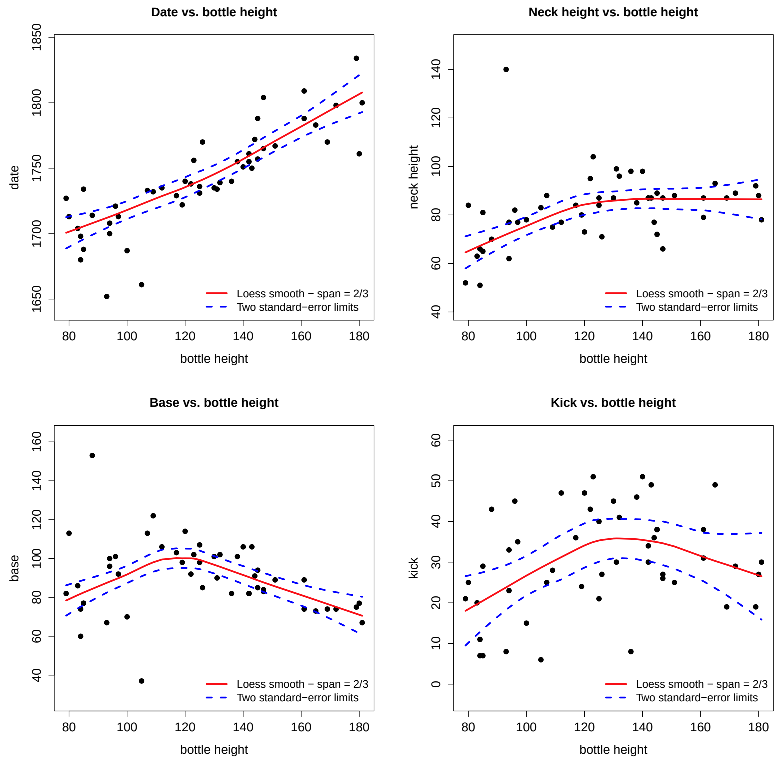

Figure 5.12: *Non-parametric regressions for the post-medieval wine bottle data with approximate two standard-error limits. See the text for an explanation.*

For *date* as the dependent variable there is little evidence of any serious departure from linearity, justifying the use of a linear fit; for other variables there is clear evidence of non-linearity with two 'regimes' either side of a body height of about 120 mm. For *date* as the dependent variable the two outliers previously noted depart from the general linear pattern; if the `family` and/or `degree` arguments are varied in the `loess.smooth` function these result in a departure from linearity at the lower end of the smooth. In summary, and notwithstanding the fact that variability is appreciable in most examples as the standard-error limits show, there is generally a fairly stable and simple underlying non-linear pattern, differentiated by bottles with small (less than 120 mm) and large body heights.

5.4 R notes

Figures 5.1, 5.4, 5.5, and 5.6

Code for Figure 5.1 is in the text where an object `fit` was created that is needed for other plots. The preliminaries are

```
date <- pmedwine$Date
BH <- pmedwine$BH
index <- 1:dim(pmedwine)[1] # create list of row numbers
fit <- lm(date ~ BH)
```

The residual plot, 5.4, is obtained with

```
library(MASS)
win.graph()
plot(fitted(fit), studres(fit))
points(jitter(fitted(fit),amount = 0), jitter(stdres(fit)))
abline(h = 0); abline(h = 2.5); abline(h = -2.5)
```

where labeling and legend commands etc. are omitted. The `MASS` package is needed to obtain the standardized (`sddres`) and studentized (`studres`) residuals. The `points` function operates in the same way as the `lines` function and adds points, with the specified coordinates, to the graph previously created. The usual color, character expansion etc. arguments are available. The `jitter` function is used to displace points slightly to avoid overwriting; see `?jitter` for details of control.

The plots put together for Figure 5.5 are obtained with

```
plot(index, lm.influence(fit)$hat)     # leverage statistic
plot(index, cooks.distance(fit))
```

The `lm.influence` function using `lm.influence(fit)$hat` extracts the *leverage* statistic. The `?lm.influence` query helps direct you to a lot of other diagnostic statistics, not illustrated here, obtained via the `influence.measures` function. The `cooks.distance` function extracts, as the name suggests, Cook's statistic.

For Figure 5.6

```
pairs(pmedwine[ , -c(1,3)])
```

removes the first and third variables not used in the plot.

Figures 5.2, 5.7, 5.3 and 5.8

Data from Tables 5.1 and 5.2 are in the files `cummins` and `morris`. Presentational arguments and the legends are omitted. As written, the appropriate models, discussed in the text, are added to the initial plot; if only the former is needed use the first lines after the `win.graph()` directives.

```
cummins.plots <- function(){
win.graph()
plot(Frequency ~ Distance, data = cummins) # Figure 5.2

pred <- lm(log(Frequency) ~ log(Distance),
data = cummins)$fitted.values

lines(cummins$Distance, exp(pred))         # Add for Figure 5.7

win.graph()
# Figure 5.2
plot(log10(cummins$Distance), log10(cummins$Frequency))

# Add for Figure 5.7
abline(lm(log10(cummins$Frequency) ~ log10(cummins$Distance)))
abline(lm(log10(cummins$Frequency[1:5]) ~ log10(cummins$Distance[1:5])))
}
cummins.plots()
```

The above is for the data from Cummins (1980). Logarithms to base 10 in the second plot emulate Cummins (1980). The code for the data from Morris (1994) is similar, allowing for the difference in the treatment of outliers and the models fitted. The `exponential` function, `exp`, transforms the fit for the linearized models back to the original scale.

```
morris.plots <- function() {
Col <- rep("black", 12); Col[8] <- "darkorange"
Sym <- rep(16,12); Sym[8] <- 15  # Case 8 is the outlier

win.graph()
# Use first line for Figure 5.3
plot(Frequency ~ Distance, data = morris, pch = Sym, col = Col)
# Add for Figure 5.8
# Exponential decay omit outlier
ze <- lm(log(Frequency) ~ Distance, data = morris[-8,])
lines(morris$Distance[-8], exp(ze$fitted.values))

# power-law model omitting outlier
zp <- lm(log(Frequency) ~ log(Distance), data = morris[-8,])
lines(morris$Distance[-8], exp(zp$fitted.values))

win.graph()
# Use first line for Figure 5.3
plot(log(Frequency) ~ Distance, data = morris, pch = Sym, col = Col)
abline(lm(log(Frequency) ~ Distance, data = morris))      # All data
# Outlier omitted
abline(lm(log(Frequency) ~ Distance, data = morris[-8,]))
}
morris.plots()
```

Figure 5.9

The original data are split into northern and southern data for the stone circles, called `circlesN` and `circlesS`. Most of the presentational arguments are omitted.

```
circles.plots <- function() {
N <- circlesN; S <- circlesS
regN <- lm(N$Deviation ~ N$Diameter)
regS <- lm(S$Deviation ~ S$Diameter)
plot(N$Diameter, N$Deviation, pch = 6)
points(S$Diameter, S$Deviation, pch = 17)
abline(regN)
abline(regS)
}
circles.plots()
```

In general the `xlim` and `ylim` arguments, not shown here, may need setting so that when the `points` function is invoked all the points are included on the plot previously created.

Figure 5.10

In the following function presentational arguments, other than `pch` and `main` which are specified in the call to the function, have been omitted. Its use is illustrated for the Northern circles `circlesN`; use `circlesS` for a plot of the Southern circles with the appropriate modification of `Pch` and `Main`. Legend specifications have been omitted from the function but can be supplied either within the function or after it is invoked. For both plots `ylim = c(0,18)` was used to ensure comparability.

```
region.plots <- function(Data, Pch = 16, Main = ""){
reg <- lm(Data$Deviation ~ Data$Diameter)$fitted.values
plot(Data$Diameter, Data$Deviation, pch = Pch, main = Main)
lines(loess.smooth(Data$Diameter, Data$Deviation))
lines(Data$Diameter, reg)
}
region.plots(circlesN, 6, Main = "Northern circles")
```

Loess smoothing is described briefly in the text; the loess.smooth function is a convenience for adding the fit produced by the loess function to a plot; the latter function is more flexible for some purposes. Users need to be aware that the defaults in the two implementations vary. The default span, for example, is 2/3 for loess.smooth and 3/4 for loess. The degree and family arguments control other aspects of a smooth with, respectively, options 1, 2 and "s", "g". The first option in each case is the default for loess.smooth; the second options are the defaults for loess.

The method works by fitting a (weighted) regression to points within a neighborhood, using the fit to predict a representative point for the neighborhood. A local linear fit is applied if degree = 1, otherwise degree = 2 produces a quadratic fit. If family = "s" is used a robust regression fit that downweights the influence of outliers is used; otherwise family = "g" assuming Gaussian errors (i.e. normally distributed) is applied. If spans are specified to be the same then the defaults in loess.smooth will produce a smoother fit than the loess defaults – which may not be what is sought.

Although the defaults for loess.smooth have been used in the example, for illustration and the message is clear enough, the choices can have a noticeable effect on the smooth and are worth experimenting with. This is pursued, with regard to the span, in the next example.

Figure 5.11

Presentational arguments other than span, and the legend, are omitted.

```
varyspan <- function() {
S <- circlesS
regS <- lm(S$Deviation ~ S$Diameter)

win.graph()
plot(S$Diameter, S$Deviation)
abline(regS)
lines(loess.smooth(S$Diameter, S$Deviation, span = 2/3))
lines(loess.smooth(S$Diameter, S$Deviation, span = 1/3))
}
varyspan()
```

Figure 5.12

Define `BH, Date, NH, Base, Kick` using `BH <- pmedwine$BH` etc. Some presentational arguments and the legend are omitted from the code.

```
bottlesloess <- function(x, y, Xlab = "", Ylab = "", Main = "")
{
pred <- predict(loess(y ~ x, span = 2/3, family = "s", degree = 1),
se = TRUE)
upper <- pred$fit + 2*pred$se
lower <- pred$fit - 2*pred$se

plot(x, y, xlab = Xlab, ylab = Ylab, main = Main)
lines(loess.smooth(x, y, span = 2/3))
lines(loess.smooth(x, upper, span = 2/3))
lines(loess.smooth(x, lower, span = 2/3))
}

win.graph()
bottlesloess(BH, date, "bottle height", "date", Main = "Date vs.
bottle height")
```

The `predict` function needs to be applied to a fit from the `loess` function to obtain (approximate) standard errors using the `se = TRUE` argument. Arguments to `loess` produce the same fit as the `loess.smooth` default; `predict` does not work with `loess.smooth`, hence the need for `loess` if standard errors are needed. In the call to the function replace `Date` and labeling arguments with those for `NH` etc. to get the other plots.

Chapter 6

Graphs – a miscellany

6.1 Introduction

In this chapter a miscellany of graphical applications are presented. Usage in the literature is variable with some of the approaches illustrated having had little, if any, archaeological application. Nevertheless implementation is usually straightforward and the methods can be quickly applied for exploratory purposes. Labeled pairs plots and two-dimensional contour plots have general application, while ternary diagrams have a specialist niche. Confidence ellipsoids are widely used for display purposes in the archaeometric literature; convex hulls have had limited use. I do not recall seeing any uses of correlation diagrams or of Chernoff faces.

6.2 Enhanced pairs plots

Pairs plots, or scatterplot matrices, are illustrated elsewhere (e.g., Figure 5.6). They provide a tool that merits routine use. Here an enhanced pairs plot, obtained with the `scatterplotMatrix` function from the `car` package, is illustrated.

The chemical compositional data from Table B.1, used in Section 2.2 are revisited. It was shown there that the variables Ca, Fe, K and Mg in the file `tubb.data` did a good job of distinguishing between the regions held in the variable `tubb.region` (Figure 6.1).

```
library(car)   # load package car
scatterplotMatrix( ~Ca+Fe+K+Mg, data = tubb.data, smooth = F,
by.group = F, group = tubb.region,reg.line = F)
```

That the regions are chemically separate is evident, as is the outlying value for K in Region 2 previously noted. The KDEs down the diagonal are optional, and other choices of graphical display are possible, or none at all. The plots are useful in this instance for emphasizing the multi-modality of the data. The *rug* at the bottom of each KDE displays the individual data points.

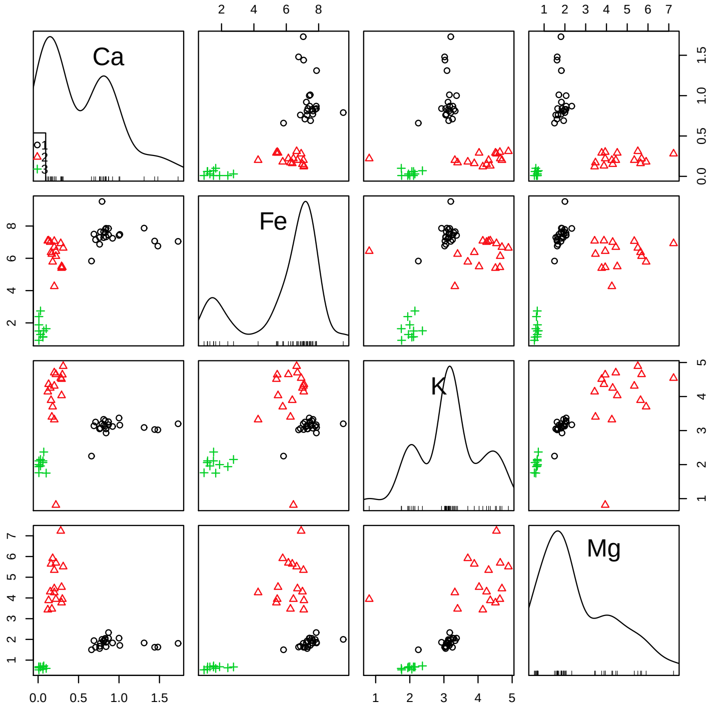

Figure 6.1: *A labeled pairs plot, or scatterplot matrix, for a subset of the oxide compositional data from Table B.1.*

6.3 Graphics with more than one variable

6.3.1 Two-dimensional KDEs

The examples in this section use data on the dimensions of loomweights, from Tables B.3 and B.4. Using KDEs, Baxter and Cool (2008) and Baxter *et al.* (2010) showed that the weights of the loomweights had an unexpected bimodal distribution; the latter paper explores reasons for and some of the implications of this. Baxter and Cool (2010a), using formal statistical tests, establish that the bimodality is not an accidental by-product of sampling variability. The bimodality is reflected in the distribution of `height` and `volume` which, along with `weight`, are the variables used in Figures 6.2 and 6.3. A reduced data set omitting outlying weights of more than 400g or less than 90g was used.

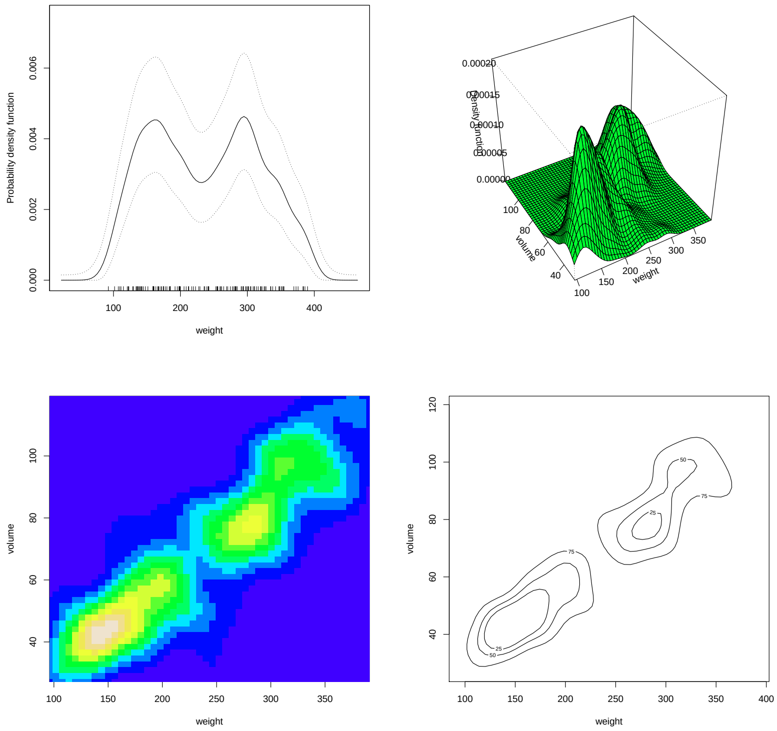

Figure 6.2: *One and two-dimensional KDEs for the loomweight data of Tables B.3 and B.4 for weight volume.*

All the plots were produced using defaults in the `sm` package except that bandwidths were subjectively chosen. They are otherwise estimated separately using automatic methods. The package is associated with the book of Bowman and Azzalini (1997). Other than loading other packages needed and setting up the data appropriately only four lines of code are needed to produce the four graphs.

The upper-left plot, a KDE for `weight`, establishes the bimodality of the variable. The added confidence band provides reassurance that the bimodality is genuine. Remaining plots shows different methods of displaying output based on a two-dimensional kernel density estimate for `weight` and `volume`/10000. An alternative would be to use the cube-root of volume, but essentially the same results are obtained. Perspective, image and contour plots are shown, with all clearly demonstrating bimodality.

6.3.2 Ellipses, convex hulls, contours – one group

Two-dimensional KDEs can also be obtained using the `kde2d` function from the `MASS` package. Similar display methods can be used (Venables and Ripley, 2002: 131), and the contour plot in Figure 6.3 was obtained using this function. In Figure 6.3 different summary displays of a mass of points are illustrated. They are potentially useful if the amount of data makes perception of any pattern difficult, or if the distribution of two or more large groups is to be compared.

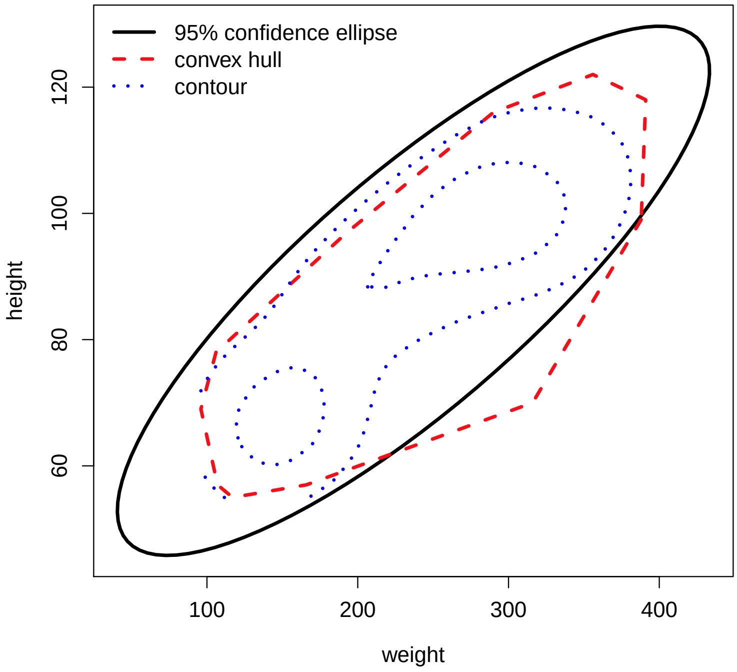

Figure 6.3: *Confidence ellipse (95%), convex hull and contour plot at two levels for the reduced loomweight data using weight and height.*

Confidence ellipsoids are quite widely used in studies of artifact provenance based on chemical compositional data, and elsewhere. In provenance studies, based on an $n \times p$ data matrix where n and p can be quite large, information on groups within the data is often available. The groups may be defined independently of the chemical composition, as in the regional data in Table B.1, or may be derived from some statistical procedure such as cluster analysis (Chapter 10). Bivariate plot may be based on a selection of a pair of variables from the original p, or derived variables such as the linear combinations obtained from a PCA (Chapter 7).

Group separation can be investigated by comparing confidence ellipsoids for the groups on the plot. Where the groups are defined independently of the chemistry visible separation then implies that different provenances are chemically distinct. Where groups are derived from the chemistry the first task is to establish that they are clearly chemically distinct (cluster analyses will produce groups whether they are 'real' or not). Interpreting any grouping in terms of provenance is a separate task, undertaken independently of the chemistry.

Confidence ellipsoids are based on the assumption that within groups the data are sampled from a population with a (bivariate) normal distribution, so that the ellipsoids are *estimates* of the extent of the groups in the population and typically extend beyond the observed limits of the data. The assumption is rarely tested and is not always self-evidently true. Some applications drop, sometimes silently, outlying points that violate the assumption, so that the appearance of a nicely bivariate normal sample of data becomes a 'self-fulfilling prophecy'.

Convex hulls are purely descriptive and summarize the data in the form of an envelope based on the minimum bounding set of points; they can be obtained with the `chull` function. They have been used relatively infrequently, except possibly in the GIS literature (which I am not very familiar with) where delineation of the spatial extent of a set of features or artifacts with something in common is an obvious application. Ringrose (1992) provides an interesting application. A correspondence analysis (Chapter 9) of an $r \times c$ table of data can be used as the basis for a bivariate plot showing the relationship between columns based on coordinates for the first two components. The stabilty of the plot is an issue. Using computer intensive methodology, bootstrapping, Ringrose replicates the data N times, producing a set of N distinct coordinates for each column marker. Convex hulls are used to display the distribution of points for each marker, and these can be compared for overlap or its lack to see how distinct each column marker really is.

6.3.3 Ellipses, convex hulls, contours – several groups

In Figure 6.3, for illustration, only one group was used, displaying a 95% confidence ellipsoid and convex hull for the data, which may be contrasted with the contour plot also shown. Contour plots can reveal sub-regions of dense point scatters that the first two cannot capture. The inner contours contain about 42% of the data.

To illustrate the use of confidence ellipsoids and convex hulls when the comparison of two or more groups is of interest, three groups were defined with Ward's method of cluster analysis (Section 10.3) using three variables `weight`, `height` and `volume`. A PCA was undertaken with these variables and Figure 6.4 shows a bivariate plot based on the first two components, with points labeled by group membership and confidence ellipsoids added for each group.

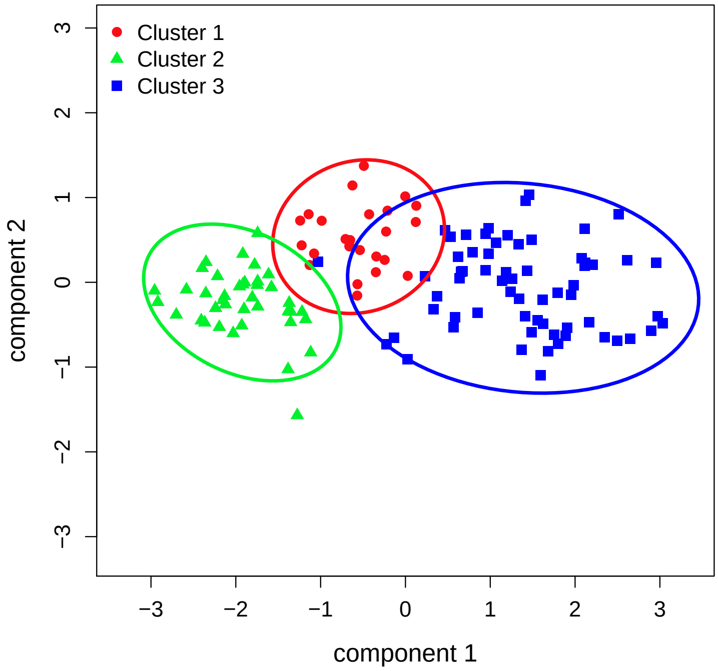

Figure 6.4: *Confidence ellipses (95%) for the reduced loomweight data using weight and height based on a partitions into three groups, determined by a Ward's method cluster analysis. See the text for fuller details.*

With only three variables and three groups defined by cluster analysis – Ward's method in particular – good group separation is to be expected, and this is what we get. There is some overlap between the central group and the other two groups, owing to the extent of the ellipsoids which go beyond the limits of the data. Visually it can be seen that the groups are largely distinct and the ellipsoid representation does not show this as effectively as one might wish.

The comparable plot, Figure 6.5, using convex hulls for the three groups is more satisfactory. The idea of peeling is illustrated. The outer hull is stripped away and a second hull calculated for the remaining data. This can be repeated so long as sufficient data remains. Only one peel is needed here to completely separate groups, the original overlap being attributable to one case.

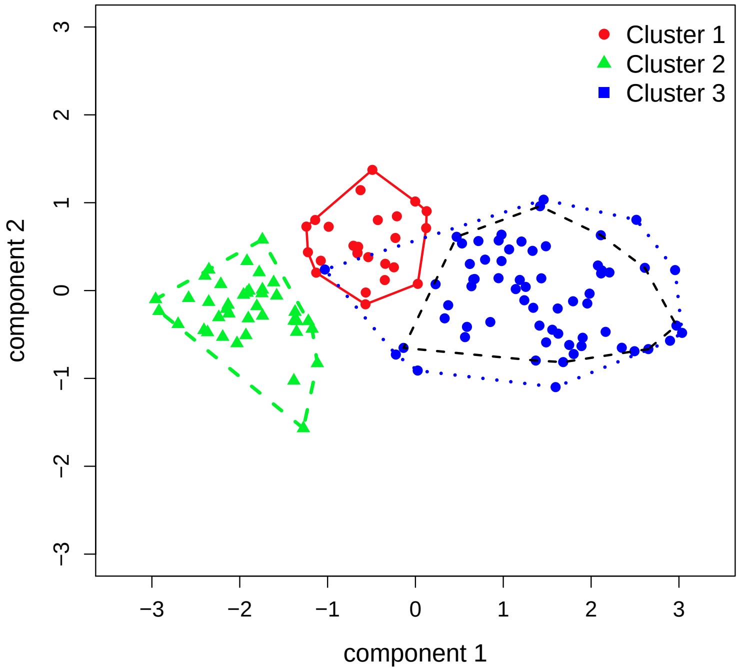

Figure 6.5: *Convex hulls for the reduced loomweight data using weight and height, based on a partition into three groups, determined by a Ward's method cluster analysis. See the text for fuller details.*

6.4 Correlation diagrams

The correlation between two variables can be represented as an ellipse. An elongated ellipse that slopes to the right is associated with a positive correlation; to the left indicates a negative correlation; (near) circularity shows a weak correlation; correlations close to a limit of -1 or +1 show a nearly linear pattern. Correlations can be presented numerically as a table of the correlation matrix; alternatively a correlation diagram, which is a visual summary of the information in the matrix, can be used. Figure 6.6, using the data in Tables B.9 to B.11 on the dimensions of polished Neolithic stone axes (O'Hare, 1990) illustrates the idea.

The general pattern is one of positive correlations and leads us to expect that the first component in the PCA will be interpretable as a *size* component (Section 7.4.1). The thickness variables are very strongly correlated, as are the breadth variables together with the width of the cutting edge, `WC`, while the depth of the cutting edge, `DC`, shows a relatively weaker correlation with most other variables.

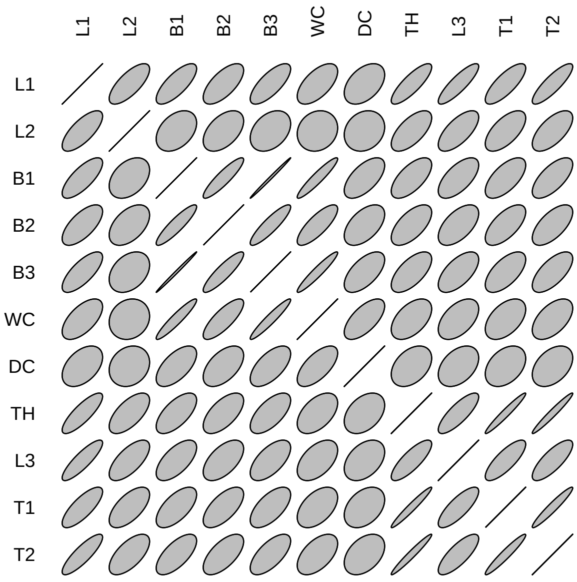

Figure 6.6: *Correlations represented as ellipses for stone axe dimensions.*

It is to be expected, and turns out to be the case, that this will be reflected in components other that the first, which have a *shape* interpretation. To obtain the correlation diagram all that is needed is the following.

```
library(ellipse)   # load package ellipse
plotcorr(cor(stoneaxe.data))
```

6.5 Ternary diagrams

If data are available in the form of an $n \times 3$ data matrix where each row sums to 100% (or 1) they may be represented in the form of a *ternary diagram*[1]. Each row is represented as a point within an equilateral triangle, with proportions represented by the perpendicular distances to the axes of the triangular coordinate system. The geometry is illustrated in Greenacre (2007: 13). To illustrate, data from Doran and Hodson (1975) given in Table 6.1 are used.

[1] Also called triangular, tri-polar or trinary diagrams or plots.

Levels	Cores	Blanks	Tools
25	21	12	70
24	36	52	115
23	126	650	549
22	159	2342	1633
21	75	487	511
20	176	1090	912
19	132	713	578
18	46	374	266
17	550	6182	1541
16	76	846	349
16	17	182	51
14	4	21	14
13	29	228	130
12	133	2227	729

Table 6.1: *Counts of cores, blanks and tools from middle levels of the palaeolithic site at Ksar Akil (Lebanon). This is Table 9.12 from Doran and Hodson (1975).*

Before plotting, the artifact counts in the final three columns must be converted to row proportions so that sample size is ignored. Where $p > 3$, in some applications, a subset of $r = 3$ columns is selected and rescaled, or $r = 3$ new variables that are linear combinations of subsets of the original columns are defined. Several different R packages contain functions to plot ternary diagrams and Figure 6.7 shows some of the different plotting options available. The packages used are listed in the caption; there are several others that might equally well have been chosen.

Because the proportions must sum to 1, given any two the third is readily calculated so the data are two-dimensional. The upper-left plot in the figure is three-dimensional, but a view has been chosen to show that the data can be ‘captured’ in a two-dimensional slice. There are choices that can be made about positions of labels, the inclusion of a grid or not and so on, with different packages differing in the options allowed. What is chosen may be partly a matter of preference. The lower- right plot, for example, only uses the smallest triangle needed to include all the data and this can help in reading a plot as it minimizes ‘bunching’ of points. It is common in practice to label the vertices at the apexes of the triangle as in the lower-left plot, but placing labels as in the lower-right plot makes it easier to see which axis a label refers to. The plots show some evidence of a seriation (not perfect) with blades tending to increase from levels 25 to 12, while tools decrease. Seriation is an early use to which ternary diagrams were put in archaeology (Meighan, 1959).

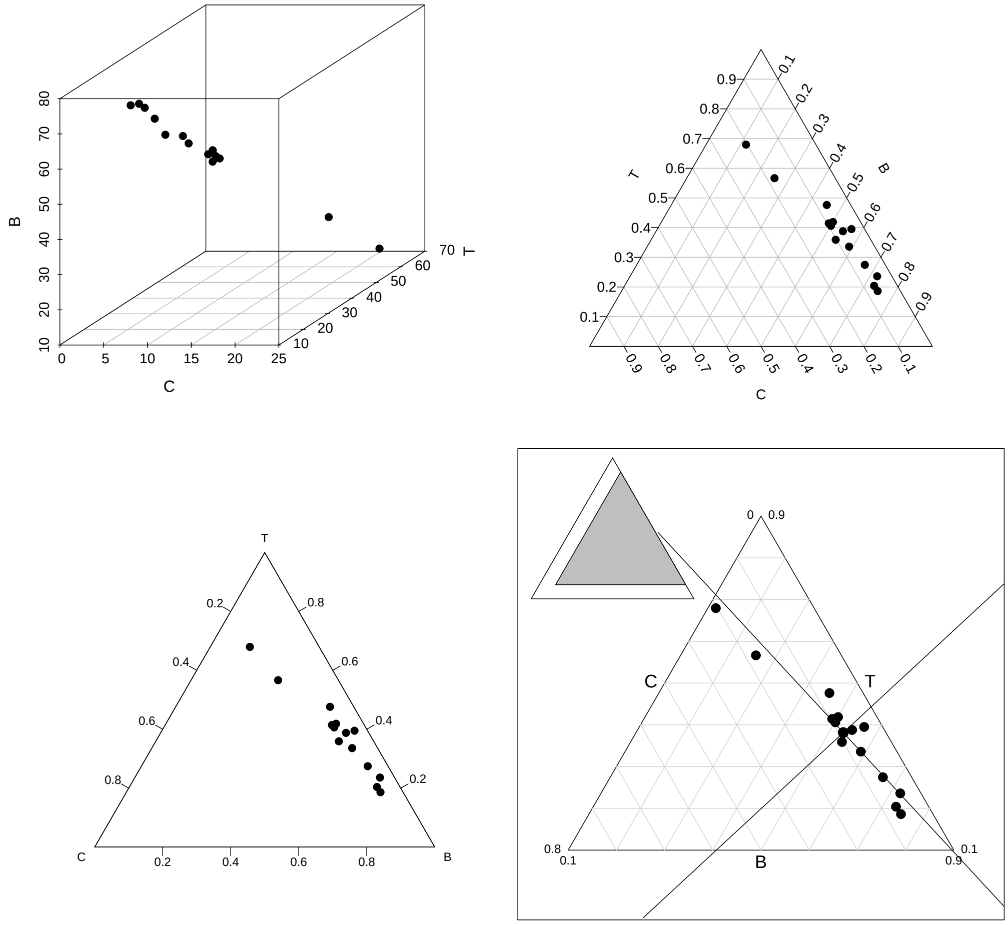

Figure 6.7: *Ternary plots from different packages for the data of Table 6.1. Reading clockwise from the upper-left the functions* `scatterplot3d`*,* `triax.plot`*,* `triangle.plot` *and* `plot.acomp` *from the packages* `scatterplot3d`*,* `plotrix`*,* `ade4` *and* `compositions` *were used.*

The use of ternary diagrams is scattered, often appearing in 'specialist' publications that draw on the traditions of the specialization involved. One such area is zooarchaeology where, for example the relative proportions of three species such as cattle, sheep/goats and pigs are displayed as points within a ternary diagram.

The data used for illustration are based on Figures 1–4 in Hesse (2011). For four different regions up to six clustered barplots for different site types show the proportions of cattle, sheep/goat and pigs in each assemblage, of which there are 20. Hesse does not use ternary diagrams, but the data are based on King (1999) who does make extensive use of them. The information contained in the 20 clustered barplots shown in Hesse can be recast as in Table B.12 and, minimally, displayed in a single ternary diagram as in the upper-left plot of Figure 6.8.

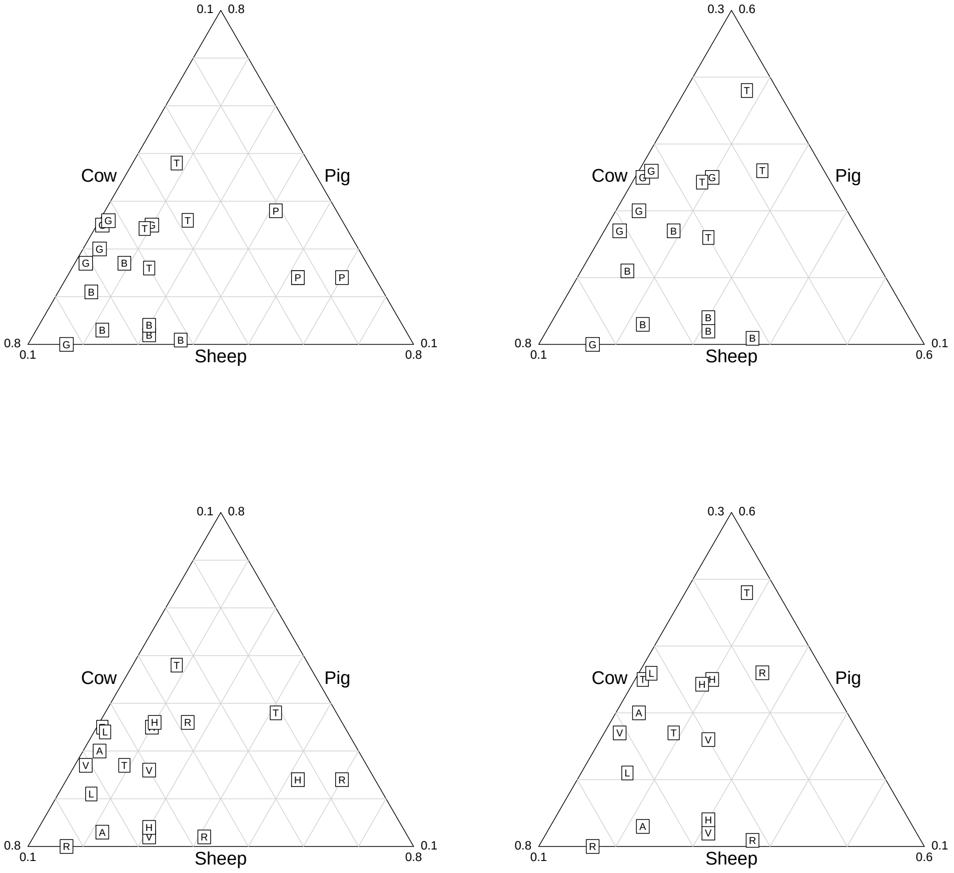

Figure 6.8: *Ternary plots, using different labels, for data derived from Hesse (2011) (Table B.12) based on King (1999). The upper plots are labeled according to the region of the site type, the site type being used to label the lower plots. Plots to the left omit the data from Roman Provence.*

Labeling is by region and two of the assemblages from Roman Provence (P) are identical, so only 19 points are visible. The `triangle.plot` function from the `ade4` package was used. The most obvious feature of the upper-left plot is the complete separation of assemblages from Roman Provence from other regions. Other regions separate reasonably well, though not perfectly. This is a bit clearer in the upper-right plot omitting data for Roman Provence. Relative to other regions Roman

Provence has a higher proportion of sheep/goats and lower proportion of cattle; Roman Britain tends to have a higher proportion of cattle and Roman Germany a lower proportion of sheep; ignoring Roman Provence the Three Gauls tend to have lower proportions of cattle and sheep/goats and hence higher proportions of pig. Variation with respect to site type is harder to discern, and numbers are probably too small to admit generalization.

Hesse (2011: 217–218) reaches essentially these conclusions based on clustered barplots[2]. The ternary diagram(s) are a more economical way of presenting the data.

Another use of ternary diagrams in zooarchaeology is for comparing mortality patterns by plotting age profiles such as juvenile, prime and old. Steele and Weaver (2002) list several papers that have used this approach of which Stiner (1990) is the earliest. Stiner illustrates the division of a ternary diagram into regions corresponding to general types of mortality pattern, allowing the type of an individual assemblage to be characterized. In other contexts such divisions are called phase diagrams. Geoarchaeology (silt/sand/clay diagrams to characterize soils) and archaeometallurgy are other areas of specialist application. Googling 'ternary diagram' with an appropriate choice of other terms will produces plenty of examples. Steele and Weaver (2002: 319) note that ternary diagrams take no account of sample sizes for assemblages, making statistical comparisons between different sets of data problematic. They propose the use of resampling methods (bootstrapping) to simulate the distribution from which a set of data is sampled.

6.6 Chernoff faces

No text of this kind is complete without an illustration of Chernoff faces. In fact I've never seen them used in anger, though they appear quite often in texts on multivariate analysis and are fun. Figure 6.9 illustrates, using the pottery chemical compositional data of Table B.1. To get this use the `apl` package, `library(aplpack)` and `faces` function, `faces(tubb.data)`.

[2] Attention is drawn to the fact that data are averaged across provinces, concealing variation *within* regions.

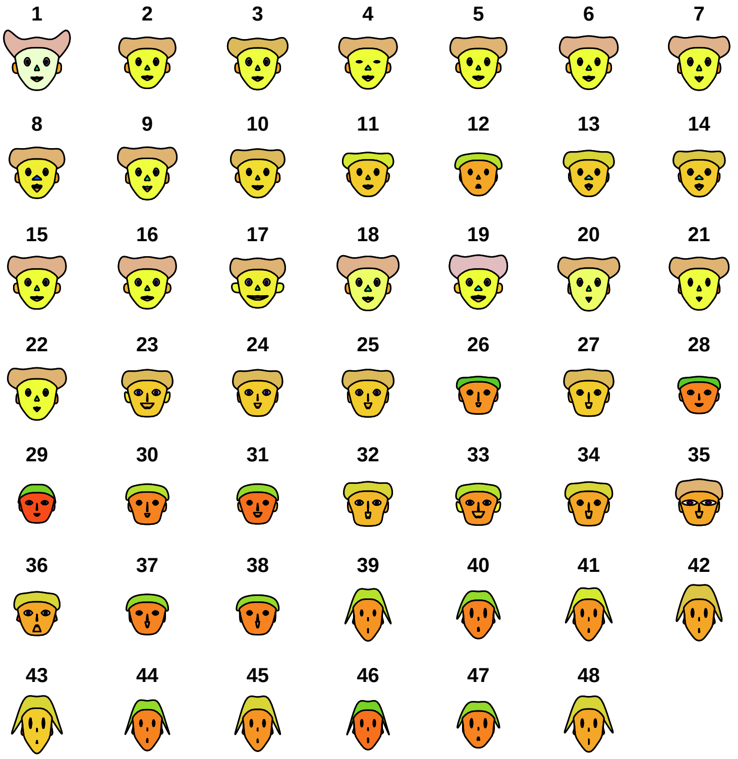

Figure 6.9: *Chernoff faces for the chemical compositional data of Table B.1.*

A list showing how variables correspond to features is returned, but it is the overall impression that is most useful. The three regions from which the pottery comes correspond to cases 1-22, 23-38 and 39-48. The Region 3 faces are distinctive – they have a lean and hungry look and appear to be rather surprised and concerned, perhaps worried about the fact they are growing pigtails.

6.7 R notes

Figure 6.2

```
kde2.plots <- function() {
library(sm)  #load sm package
wt <- loomweights$weight
x <- loomweights[wt > 90 & wt < 400,]
weight <- x[,6]; volume <- x[,7]/10000
win.graph()
sm.density(weight, display = "se", h = 15, lwd = 8)
win.graph()
sm.density(cbind(weight, volume), display = "persp", h = c(15,6))
win.graph()
sm.density(cbind(weight, volume), display = "image", h = c(15,6))
win.graph()
sm.density(cbind(weight, volume), display = "slice", h = c(15,6))
}
kde2.plots()
```

The third and fourth lines select loomweights with weights between 90g and 400g.

Figure 6.3

```
EllipseEtc <- function() {
library(ellipse)
library(MASS)
wt <- loomweights$weight
z <- loomweights[wt > 90 & wt < 400,]
height <- z$height; weight <- z$weight
X <- cbind(weight, height)
m1 <- mean(X[,1]); m2 <- mean(X[,2])

Z <- ellipse(cov(X), centre = c(m1, m2))   # ellipse
plot(Z)

hpts <- chull(X)        # convex hull
hpts <- c(hpts,hpts[1])
lines(X[hpts,])

K <- kde2d(weight, height)   # contour plot
contour(K, add = T, drawlabels = F, nlevels = 5)
}
EllipseEtc()
```

Some presentational arguments and the legend are omitted. Line type and width, color and character expansion can be controlled in all the plotting functions. The ellipse is plotted first as it is the most extensive of the various plots and determines the scale on which they are superimposed. Otherwise the `xlim` and `ylim` arguments need to be experimented with.

Figure 6.4

```
multiple.ellipse <- function() {
library(ellipse)
library(MASS)
wt <- loomweights$weight
z <- loomweights[wt > 90 & wt < 400,]
height <- z[,1]; weight <- z[,6]; volume <- z[,7]
X <- cbind(weight, height, volume)
nclust <- cutree(hclust(dist(scale(X)), method = "w"), k=3)

# Set-up plotting characters and colors
Symbol <- rep(16, length(nclust))
Symbol <- ifelse(nclust == 2, 17, Symbol)
Symbol <- ifelse(nclust == 3, 15, Symbol)

Col <- rep("red", length(nclust))
Col <- ifelse(nclust == 2, "green2", Col)
Col <- ifelse(nclust == 3, "blue", Col)

PCA <- prcomp(scale(X))$x[, 1:2]

eqscplot(PCA[,1], PCA[,2], pch = Symbol, col = Col,
xlim = c(-3.5, 3.5))

X1 <- PCA[nclust == 1, ]
X2 <- PCA[nclust == 2, ]
X3 <- PCA[nclust == 3, ]
m11 <- mean(X1[,1]); m12 <- mean(X1[,2])
m21 <- mean(X2[,1]); m22 <- mean(X2[,2])
m31 <- mean(X3[,1]); m32 <- mean(X3[,2])

Z1 <- ellipse(cov(X1), centre = c(m11, m12)); lines(Z1)
Z2 <- ellipse(cov(X2), centre = c(m21, m22)); lines(Z2)
Z3 <- ellipse(cov(X3), centre = c(m31, m32)); lines(Z3)
}
multiple.ellipse()
```

Figure 6.5

```
chull.loomweights <- function(){
library(MASS)

wt <- loomweights$weight
z <- loomweights[wt > 90 & wt < 400,]
height <- z[,1]
weight <- z[,6]
volume <- z[,7]
X <- cbind(weight, height, volume)

nclust <- cutree(hclust(dist(scale(X)), method = "w"), k=3)
PCA <- prcomp(scale(X))$x[, 1:2]

eqscplot(PCA[,1], PCA[,2], type = "n")
X1 <- PCA[nclust == 1, ]
X2 <- PCA[nclust == 2, ]
X3 <- PCA[nclust == 3, ]
points(X1[,1], X1[,2], pch = 16, col = "red", cex = 1.2)
points(X2[,1], X2[,2], pch = 17, col = "green2", cex = 1.2)
points(X3[,1], X3[,2], pch = 16, col = "blue", cex = 1.2)

hpts <- chull(X1)       #plot for cluster 1
hpts <- c(hpts,hpts[1])
lines(X1[hpts,], lwd = 2, col = "red")

hpts <- chull(X2)       #plot for cluster 2
hpts <- c(hpts,hpts[1])
lines(X2[hpts,], lty = 2, lwd = 3, col = "green2")

hpts <- chull(X3)      #plot for cluster 3
hpts <- c(hpts,hpts[1])
lines(X3[hpts,], lty = 3, lwd = 3, col = "blue")

points = chull(X3)     # One peel of the data for the cluster
X3 <- X3[-points,]
hpts <- chull(X3)
hpts <- c(hpts,hpts[1])
lines(X3[hpts,], lty = 2, lwd = 2, col = "black")
}
chull.loomweights()
```

Figure 6.7

```
Ksar.ternary <- function() {
# Set up data
Ksar <- Ksar_Akil[,-1]
Ksarsum <- apply(Ksar, 1, sum)
Ksar <- Ksar*100/Ksarsum

win.graph(); library(scatterplot3d)
scatterplot3d(Ksar$C, Ksar$T, Ksar$B, xlab = "C", ylab = "T",
zlab = "B")

win.graph(); library(plotrix)
triax.plot(Ksar, cex.ticks = 1.2, show.grid = TRUE)

win.graph(); library(compositions)
plot.acomp(Ksar, axes = TRUE)

win.graph(); library(ade4)
triangle.plot(Ksar, addaxes = TRUE, box = TRUE, cpoi = 2.5)
}
Ksar.ternary()
```

For other than `triangle.plot` the usual presentational arguments are available and not shown. See the `?help` facility for the many variations possible with each type of plot.

Figure 6.8

```
library(ade4)
triangle.plot(king.data, show.position = F,
label = king.region, clabel = .8)

triangle.plot(king.data[-c(13:16),], show.position = F,
label = king.region[-c(13:16)], clabel = .8)

triangle.plot(king.data, show.position = F,
label = king.type, clabel = .8)

triangle.plot(king.data[-c(13:16),], show.position = F,
label = king.type[-c(13:16)], clabel = .8)}
```

Chapter 7

Principal component analysis

7.1 Introduction

In Chapter 2 principal component analysis (PCA) and correspondence analysis were introduced to show how easy it is, using `R`, to undertake such analyses. It is (almost) as simple as calculating a mean. To remind the reader, given a suitable $n \times p$ table of data, $\mathbf{Y}$ entered as `Y` in `R`,, a basic PCA can be carried out using `prcomp(Y)`.

Calculation of a mean should not be carried out without thought. It is useless if the data are seriously multi-modal; can be compromised if there are obvious outliers in the data; and is not sufficient as a single summary measure of location if the data are highly skewed. Preliminary data inspection is called for, and PCA is no exception. Principal component analysis is conceptually, as well as computationally, simple. As usually applied, it takes an $n \times p$ data set and reduces it to a 'picture' that allows patterns in the data to be investigated using conventional two- and three-dimensional plots. There are practicalities to be addressed in applications and interpretation, and this chapter discusses and illustrates these.

Assuming $n > p$, the table of data is p-dimensional and not susceptible to conventional methods of data exploration if $p > 3$. It is possible to define the *distance* between the rows of data cases (Section 7.3); one way of thinking about PCA is that it is designed to approximate these distances in two- or three-dimensional space using the first two or three principal components (PCs). Inevitably information is lost in the approximation so some means of measuring its quality is desirable.

Sections 7.2 to 7.4 are centered on examples, used to introduce the ideas that underpin PCA, implementation and interpretation. The algebra that underpins PCA is covered in Appendix D. It is helpful to have some acquaintance with the associated terminology that finds its way into software output. It is perfectly possible to apply PCA usefully without needing to master the algebra. There is a focus in Section 7.2 on data standardization and transformation, and Section 7.3 interpolates a brief discussion of the important concept of distance before the example of Section 7.4.

The earliest uses of PCA in archaeology date back to the 1960s. This coincided with the start of a period when the method of *factor analysis* was in vogue. The two methods were, and still are, confused. Chapter 8 attempts to explain what these differences are and why they can be considered fundamental. The opportunity is taken, in Section 8.4, to recount, briefly, some of the history of the use of the methods in archaeology, along with critical comment on more recent advocacy of factor analysis.

7.2 Example 1 – Roman glass compositions

Standardization and log-transformation

The data consist of 105 specimens of Romano-British waste glass, measured with respect to nine major and minor oxides, excavated from sites at Leicester and Mancetter in the UK (Tables B.7 and B.8)[1]. One question is whether or not the glass from the two sites is chemically distinct. The analysis will be used as a peg on which to hang a discussion of *data transformation* in PCA.

The initial data table (or matrix), $\mathbf{X}$, has a typical entry x_{ij}, with $\bar{x}_j$ and s_j denoting the estimated mean and standard deviation of variable (column) j, $j = (1, 2, \ldots, p)$. Invariably, PCA is carried out after some transformation of the original data to a data matrix $\mathbf{Y}$.

The simplest form of transformation is *centering* where y_{ij} is defined as

$$y_{ij} = (x_{ij} - \bar{x}_j).$$

This is the first stage in packages where PCA is available, but usually analysis goes beyond this for two reasons. One is that, if variables are measured in different units, the use of centered data alone is inappropriate as the variables are not comparable. The second reason, illustrated in Figures 7.2 and 7.3, is that even if the variables are in the same units a PCA of centered data will be dominated by the variables with the larger variances, so potential information provided by other variables is lost. Thus, unless variables have similar variances to begin with, some further transformation beyond centering is called for.

[1] The data were collected by Dr. Caroline Jackson of Sheffield University, UK, as part of an unpublished PhD thesis. Some analyses are undertaken in Baxter (1994a).

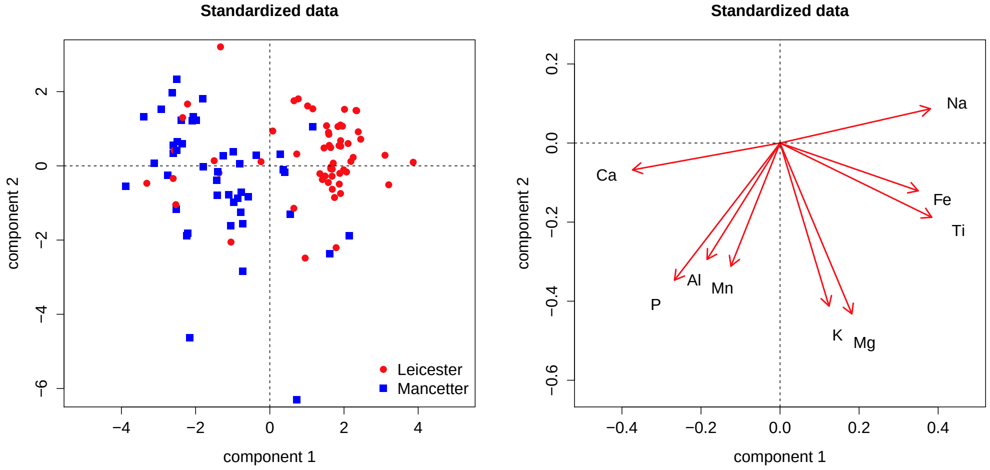

Figure 7.1: *PCA score and variable plots for the standardized Romano-British glass data of Tables B.7 and B.8.*

The most common form of data pre-treatment, and the default in many software packages though not `R`, is to *standardize* the data as

$$y_{ij} = (x_{ij} - \bar{x}_j)/s_j$$

producing variables with a mean of 0 and standard deviation (and variance) of 1. This gives each variable equal 'importance' so that each has an 'equal chance' of influencing the PCA. A biplot for the PCA of standardized data is shown in Figure 7.1 presented as adjacent score and variable plots.

This can be contrasted with the output obtained using log-transformed (to base 10) data (Figure 7.2)

$$y_{ij} = \log x_{ij}$$

which is the other approach to have found widespread use. This is applied before any subsequent centering and standardization [2].

It is quite common to standardize the data after log-transformation, though this then often produces results very similar to standardization of the original data (Baxter, 1995). Values recorded exactly as zero cannot be log-transformed and the problem is usually resolved by adding a small value to the data (see Section 8.3.2 for an example).

[2] Terminology in the literature is confusing. Standardization is sometimes called *normalization*, implying that the transformed variables have a normal distribution. They will not, unless the untransformed data begin with a normal distribution. To confuse matters further log-transformation is sometimes referred to as standardization. It is also sometimes used with the hope that it will induce normality.

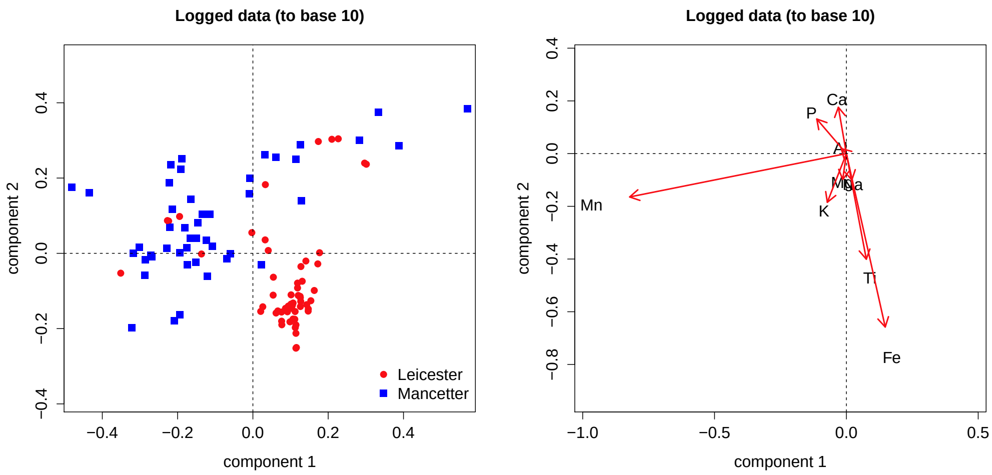

Figure 7.2: *PCA score and variable plots for the unstandardized log-transformed Romano-British glass data of Tables B.7 and B.8.*

Some of the differences that can arise between the use of standardized and unstandardized log-transformed data are illustrated in Figures 7.1 and 7.2. The two sites largely separate out, with the grouping for Leicester the more compact. There are about 10 cases that plot more closely with the Mancetter data. The score plot using log-transformed data suggests a possible sub-division among the Mancetter data.

The variable plots are more obviously distinct. The variable markers for the standardized data lie (very roughly) round a circle, equidistant from the origin. This reflects the equal weighting induced by standardization, though it is still possible for some variables to have little effect on the PCA, or to play a lesser role (e.g., Al, Fe, Mn). If row (cases) and column (variable) plots are compared it can be seen for the row plot that the Leicester data lie largely to the right and the Mancetter data to the left. For the variables, (Fe, Na, Ti) plot to the right, while (Al, Ca, Mn, P) plot to the left. It can be inferred that the Mancetter glasses are richer in the latter group and poorer in the former group, relative to Leicester. This is readily checked, where it can be seen that the variable means for the two sites support this inference (Table 7.1).

Site	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn
Leicester	2.38	0.70	0.55	6.59	18.20	0.71	0.10	0.12	0.27
Mancetter	2.47	0.48	0.53	7.19	17.20	0.72	0.08	0.14	0.41

Table 7.1: *Variable means (%) for the Romano-British glass and the two sites.*

The variable plot for the log-transformed data is rather different, being dominated by Mn and Fe. This implies that these two variables have much the largest variances on the log-transformed scale, and this too is readily checked. The variances for Mn and Fe are 0.0338 and 0.0175 respectively, with the next largest that for Ti of 0.007. The pattern in the row plot for the log-transformed data is thus dominated by the effect of Mn and Fe, in contrast to that for standardized data. The row plots, while showing differences, are not incompatible with each other, pointing to the fact that different transformations can give rise to similar patterns in the data, even though different variables may be responsible. Similarly, it is possible for row plots to differ, but admit equally valid archaeological interpretations. The message is that in exploratory work the examination of different transformations is worthwhile.

The interpretation of biplots was discussed briefly in Chapter 2, and some elaboration is provided here. Further discussion is provided in Section 9.2 after their use in correspondence analysis has been introduced. What is understood by the term ‘biplot’ varies – I err on the side of a more informal usage, allowing the term to embrace the joint presentation of row and column plots, rather than superimposing them, for example. It is easy enough to imagine the two plots being superimposed, provided a common origin is indicated and correct aspect ratios are used, though issues of axis scaling arise (see Section 9.2).

With the caveat that the PCA should be of reasonable quality, variable markers that lie opposite to each other on the plot should have a negative correlation (e.g., Ca and Na in the plot for standardized data); variables at right angles should show weak correlation (e.g., K, Na); variables plotting close to each other, with an acute angle at the origin, should exhibit strong positive correlation (e.g., Fe, Ti). This can be seen to be broadly the case from Table 7.2, which shows correlations to one significant digit, with the ordering based on reading clockwise from Na on the variable plot for standardized data. It can be seen from Figure 7.2 that the Mancetter glass has a comparatively higher concentration of Mn than Leicester where the concentration of Fe stands out. As already noted the coincidence of interpretation of the score plots can be attributed to different subsets of variables and is not uncommon in applications that contrast standardized with unstandardized log-transformed data. It is not inevitable; situations where the latter approach highlights variables with a low absolute presence leading to no useful interpretation are also not uncommon.

Other forms of transformation have been proposed but little used. Baxter (1995) suggested rank-transformation as a possibility, and ‘standardizing’ variables to the range [0,1] has occasionally been seen. More needs to be said about log-ratio transformations.

	Na	Fe	Ti	Mg	K	Mn	Al	P	Ca
Na	1	0.5	0.6	0.3	0.2	-0.2	-0.4	-0.6	-0.8
Fe	0.5	1	0.8	0.4	0.2	-0.2	-0.1	-0.3	-0.6
Ti	0.6	0.8	1	0.5	0.3	-0.2	-0.1	-0.3	-0.7
Mg	0.3	0.4	0.5	1	0.4	0.1	0	0.1	-0.1
K	0.2	0.2	0.3	0.4	1	0.2	0.1	0.1	-0.2
Mn	-0.2	-0.2	-0.2	0.1	0.2	1	0.2	0.4	0.1
Al	-0.4	-0.1	-0.1	0	0.1	0.2	1	0.3	0.4
P	-0.6	-0.3	-0.3	0.1	0.1	0.4	0.3	1	0.5
Ca	-0.8	-0.6	-0.7	-0.1	-0.2	0.1	0.4	0.5	1

Table 7.2: *Correlations between variables for the Romano-British glass data.*

Log-ratio transformation

Data of the kind used here are sub-compositional. They would be fully compositional, adding to 100%, if all the naturally occurring elements were measured. Only a subset are ever measured and hence the data are *sub-compositional*. It is possible to convert such data to fully compositional form, either by rescaling to 100% or defining a 'residual' as the sum of the measured data subtracted from 100%. In the present case the 'residual' will largely coincide with the silica content of the glass (not measured by the instrumentation used).

The use of ratio transformations has been debated in archaeology from time to time and was explored more generally in the seminal text of Aitchison (1986). He advocated the use of log-ratio transformation, a symmetric version being

$$y_{ij} = log(x_{ij}/g(\mathbf{x}_i))$$

where $g(\mathbf{x}_i)$ is the geometric mean of row i, defined as the $1/p$th root of the product of the elements of row i. An argument for this is that the raw compositional data are positive and constrained to lie in a $(p-1)$-dimensional space, for which the more common methods of analysis, including the use of standardization, are inappropriate. The log-ratio transformation removes the constraints on the data, allowing standard methods to be applied. This is illustrated in Figure 7.3, where the original composition is augmented by the constructed 'silica' (Si) variable.

The analysis is practically indistinguishable from that of log-transformed data. This is precisely the point. Aitchison's advocacy of the methodology he developed for analyzing compositional data is theoretically compelling. Baxter (1989) used the methodology for analyzing glass compositional data and was initially enthusiatic, but later exploration suggested that, regardless of theory, the more usual methods of analysis often produced equivalent or more satisfactory and archaeologically interpretable results. This is because log-ratio transformed data are not

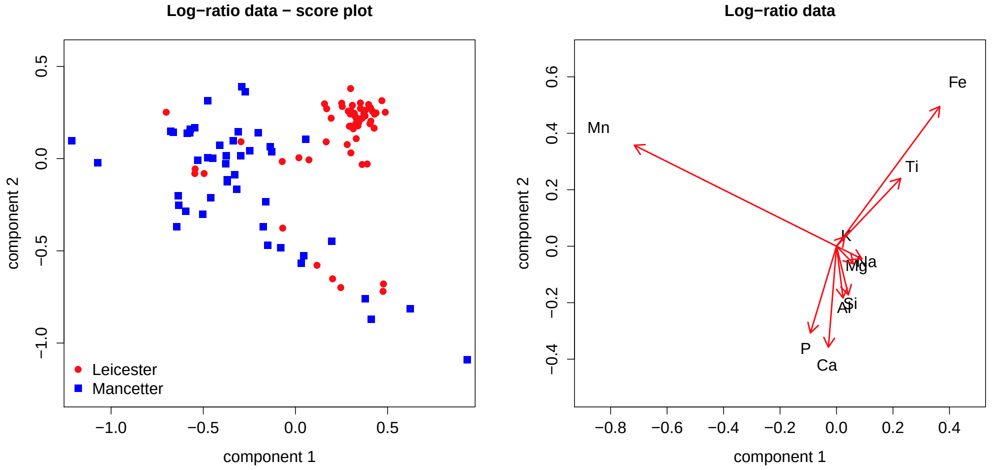

Figure 7.3: *PCA score and variable plots for the unstandardized log-ratio-transformed Romano-British glass data of Tables B.7 and B.8.*

usually standardized in subsequent analysis. This means that variables with a low absolute presence and high relative variance are emphasized, at the expense of variables with a greater presence that may be more important for understanding the production processes that produced the glass.

The dominance of the variables with low absolute presence is what is illustrated in Figure 7.3. It happens to make sense in that interpretable results, in the form of as good a separation between sites as can be reasonably be expected, is attained. This is also, in a sense, 'accidental' since, as noted, variables with smaller typical values can dominate log-ratio (and log-transformed) analyses to no good effect. Usually this can be 'corrected' for by adopting pragmatic measures, omitting such variables from an analysis for example, but this often then leads to outcomes similar to the simpler (if not necessarily 'theoretically correct') methods that are prevalent in the literature. Log-ratio analysis has not really caught on in archaeological applications but anyone with compositional data to analyze should be aware of the issues involved; a detailed account of what these are is provided in Baxter and Freestone (2006)[3].

[3] A lot of work has been done on developing a rigorous mathematical framework for compositional data analysis (Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn *et al.*, 2015). Advocates of log-ratio methodology can be dismissive of analytical approaches that do not conform to their theoretical *dicta*. My practice, when confronted with compositional data, is to examine a range of analyses including the use of log-ratios, so my opinions are driven by practical experience (relating primarily to multivariate data analysis) rather than theoretical agonizing. An R package, `compositions`, is available along with a book devoted to its use (van den Boogaart and Tolosana-Delgado, 2013) for those wishing to explore the ideas and application.

7.3 The idea of distance

A fundamental idea behind PCA is that, using all the components (PCs), the distance between cases on the scale of the data used is reproduced. A subset of the first few PCs allows distances to be approximated in a low-dimensional space that can be more readily interrogated using standard graphical methods. The distances approximated depend on the standardization/tranformation used. As will be seen in later chapters, different definitions of distance underpin, and distinguish between, different methods of multivariate analysis. The (mathematically) simplest application of the ideas arises in the context of PCA, and a general discussion is provided here.

Given two rows of a data matrix – call them $\mathbf{y}_i$ and $\mathbf{y}_k$ – it is possible to measure exactly the distance between them, d_{ik}, in p dimensions. Several multivariate methods work by defining new variables, in which the rows are $\mathbf{z}_i$ and $\mathbf{z}_k$. For the first $r < p$ columns the distance between the rows can be defined, but will only approximate the true distances. This begs the questions of how to define the new variables and how the quality of approximation is judged. These will be dealt with in the context of the example in Section 7.4.

The point to stress is that distance, d_{ik}, is not a uniquely defined concept. Any proposed measure of distance qualifies as such if it satisfies a particular set of mathematical rules. Different measures of distance are appropriate for different kinds of data and problem specification. Of the 'standard' methods of multivariate analysis PCA is the easiest to understand since it is based on Euclidean distance which we are familiar with from everyday experience.

We can judge distances between points, and measure them exactly if needed. At the scale we normally operate on this is *Euclidean* distance. It can be defined mathematically in a way that generalizes to p-dimensions[4]. In p-dimensional space a point is defined by a set of values for the variables $(Y_1\ Y_2\ \ldots\ Y_p)$. After PCA the distances between cases are approximated by the distances between new variables, $(Z_1\ Z_2\ \ldots\ Z_r)$, where r is usually 2 or 3.

[4] The definition, of squared Euclidean distance, is

$$d_{ik}^2 = \sum_{j=1}^{p} (y_{ij} - y_{kj})^2$$

with the square-root of this giving d_{ik}.

7.4 Example 2 – Stone axe morphology

In this section ideas previously introduced are elaborated on, along with the practicalities of application and interpretation. The data used, from O'Hare (1990), are dimensional meaurements on 11 variables for 181 Neolithic polished stone axes from southern Italy, classified into three types according to their butt shape – pointed, rounded or square. The data are a subset of a larger sample of 209 axes, two small groups of intermediate butt types having been omitted for the purposes of our analyses. The full data set, with a definition of the variables, is given in Tables B.9 to B.11 in Appendix B.

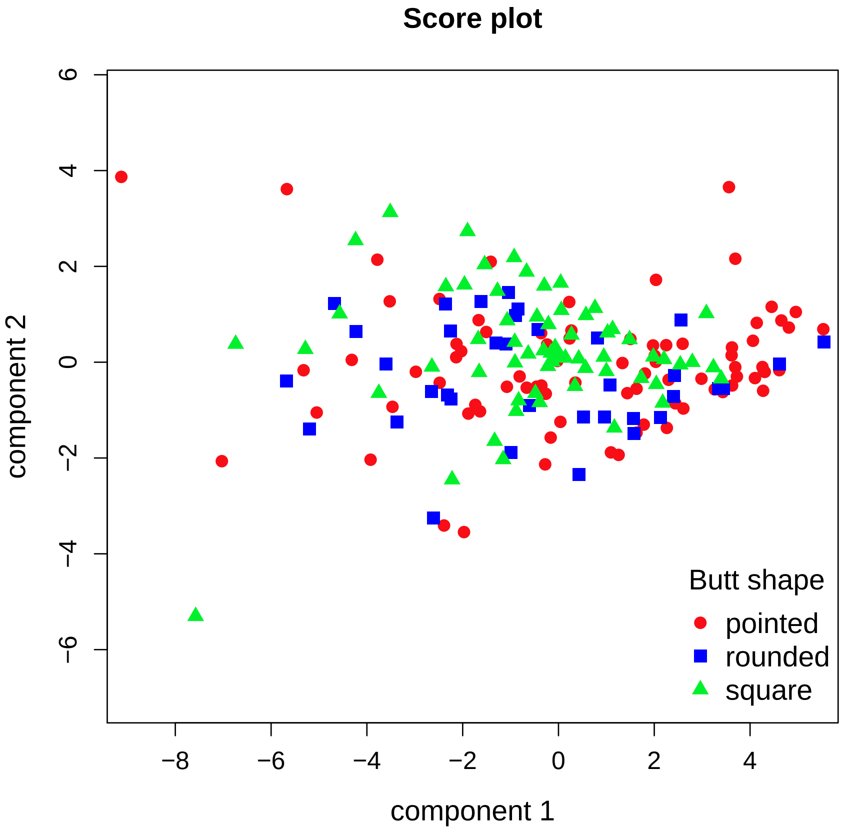

Figure 7.4: *A score plot for components 1 and 2 from a PCA of the standardized stone axe data labeled by butt shape.*

One of the research questions was whether the dimensional data revealed patterns that could be associated with butt type. One of the analytical tools used was PCA and this is emulated here, using butt type for labeling and interpretive purposes only. An example is provided first, before spelling out some of the detail. Figure 7.4 plots scores on the first two components from a PCA of the standardized data. There is no obvious grouping in the data and no outliers to cause undue concern.

7.4.1 Definition and properties of principal components

For a more detailed account of the notation introduced immediately below see Appendix D. The $n \times p$ data matrix $\mathbf{Y}$ has typical element y_{ij}, for $j = 1, \ldots p$ and $i = 1, \ldots, n$; principal components (PCs) are *defined* to be linear combinations of the form

$$Z_j = a_{j1}Y_1 + a_{j2}Y_2 + \ldots + a_{jp}Y_p$$

for component j.

The a_{ji} are *coefficients* to be determined[5]. Given the a_{ji} principal component *scores*, z_{ij}, held in the matrix $\mathbf{Z}$ with the same dimensions as $\mathbf{Y}$, can be calculated.

Principal components are *defined* by the following criteria.

(a) The components are uncorrelated.

(b) The first component, Z_1, has maximum variance; subject to the lack of correlation Z_2 has the second largest variance; and so on. The variances are called *eigenvalues* in some software and will be denoted by λ_i.

(c) There is a complication in that the variances are unbounded unless a constraint is imposed on the coefficients. Often this has the form

$$a_{j1}^2 + a_{j2}^2 + \ldots + a_{jp}^2 = 1 \tag{7.1}$$

but

$$a_{j1}^2 + a_{j2}^2 + \ldots + a_{jp}^2 = \lambda_j. \tag{7.2}$$

is also used. Constraint (7.1) is usual in software where PCA and factor analysis are clearly distinguished; constraint (7.2) is usual in implementations of factor analysis[6].

Given these conditions we let the mathematics do the work of obtaining the 'solution' to the problem (of determining the a_{ji}) (Appendix D). Computational aspects are embedded in `R` functions such as `prcomp`, the fine detail of which the average user may remain blissfully unaware.

[5] Also sometimes called *loadings*, a term more commonly used in factor analysis (Chapter 8).

[6] In the widely-used `SPSS` package PCA is treated as a special case of factor analysis. This has led to a lot of confusion among users and will be discussed further in Chapter 8.

The idea behind (a) is that it can be easier to work with uncorrelated variables. It should be emphasized that all that is involved is a *mathematical transformation* of the data. There is no requirement that the components be interpretable other than as a linear combination of the original variables, though they often will be. That is, assignment of a 'meaning' to components is not a fundamental issue. This is a source of confusion between PCA and factor analysis where the definition of 'meaningful' factors is a central concern (see Chapter 8).

The idea behind (b) is that the components with the larger variances are likely to be *structure carrying* in the sense that plots based on them will reveal patterns in the data, if they exist. This idea is empirically rather than theoretically based but it frequently 'works'.

The correlation diagram in Figure 6.6 shows that there are generally high correlations among the variables. This leads to the expectation that PCA will be an effective dimension-reduction method. Furthermore, all the correlations are positive, which leads to the expectation that the coefficients for the first PC will be of the same sign and similar order of magnitude[7]. This has an interpretation as a '*size*' component, literally in the present example. Components with a mixture of signs among the coefficients can be interpreted as '*shape*' components, further interpretability concerning aspects of shape depending on the context.

The importance of a variable in defining a component depends on the value of $|a_{ji}|$. Where this, or its square, is close to 1 (if constraint (7.1) is used) the component is effectively the same as variable i. Various strategies exist for aiding interpretation, for example by ignoring coefficients for which $|a_{ji}|$ is less than some predetermined value (Section 8.3). It can help if a coefficient is either 'large' or 'close' to zero[8]. This can be achieved more formally, and mathematically, by subjecting components to *rotation*, more common in applications of factor analysis than PCA and discussed further in Sections 8.1 and D.3.2.

7.4.2 Interrogating PCA output

To interrogate numerical information the PCA of standardized data is first undertaken using

```
axe.pca <- prcomp(stoneaxe135.data, scale = T)
```

where `stoneaxe135.data` is the subset of Tables B.9 to B.11 containing the three butt types under investigation. Component scores are held in `axe.pca$x` and

[7] The signs of the PCs are arbitrary so if, for example, all signs are positive or all negative the interpretation is unaffected.

[8] This is not essential; 'size' components are readily interpretable without satisfying this criterion.

coefficients in `axe.pca$rotation`. The former are used as the basis for the plot of Figure 7.4. The latter can be viewed to a sensible number of significant digits using `round(axe.pca$rotation, 1)` with the result shown in Table 7.3.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
L1	-0.3	0.2	0.2	-0.1	0.4	0.4	-0.6	0	-0.3	0	0
L2	-0.2	0.4	0.5	-0.4	-0.3	0.2	0.4	0.1	0	-0.1	0
B1	-0.3	-0.3	0	-0.2	0.1	0.1	0.2	-0.1	0	0.4	-0.7
B2	-0.3	-0.2	0	-0.4	-0.5	-0.4	-0.5	-0.1	0.1	-0.1	0
B3	-0.3	-0.3	-0.1	-0.2	0.1	0.1	0.2	0	0	0.4	0.7
WC	-0.3	-0.4	-0.1	-0.1	0.2	0.1	0.3	0.2	0	-0.7	0
DC	-0.2	-0.3	0.6	0.7	-0.2	0	0	0	0	0.1	0
TH	-0.3	0.3	-0.3	0.2	-0.1	0.1	0.1	-0.8	0	-0.2	0
L3	-0.3	0.2	0.2	0	0.6	-0.6	0.1	0	0.2	0	0
T1	-0.3	0.3	-0.3	0.2	-0.2	-0.2	0.2	0.4	-0.6	0.1	0
T2	-0.3	0.2	-0.3	0.2	-0.1	0.3	-0.1	0.4	0.7	0.1	0

Table 7.3: *PC coefficients, rounded to one decimal place, from the PCA of the stone axe data.*

As expected, the coefficients for the first component have the same sign and similar magnitude and can be interpreted as a size component – essentially it averages the standardized measurements of all the variables. The second component is a shape component that contrasts the length and thickness variables with the breadth and cutting-edge variables. The pattern is displayed in a readily appreciated form in the variable plot for the first two components, in the left-hand plot of Figure 7.5.

Size may or may not be of intrinsic interest; for the axe data the focus in O'Hare (1990) was on typology as revealed by shape, so size is of less interest and the plot based on the second and third components in Figure 7.5 is of potentially greater interest. It shows three distinct clusters based on length, thickness and breadth variables, the last of these also associated with the width of the cutting-edge. Depth of cutting-edge (`DC`) is isolated from the other variables, and also dominates the fourth PC. The graphs tell the same story as those in O'Hare (1990).

Output concerning the 'importance' of the PCs can be investigated in several ways. Commonly in software packages the variances (eigenvalues), both individual and cumulative, are presented. In `R`, if `prcomp` is used, the standard deviations, as opposed to variances, are stored in `axe.pca$sd`. These can be manipulated to produce Table 7.4, emulating what is to be seen in other software.

Section 6.1 of Jolliffe (2002) discusses a large number of criteria that have been used for selection; the most commonly used, and the only ones considered here,

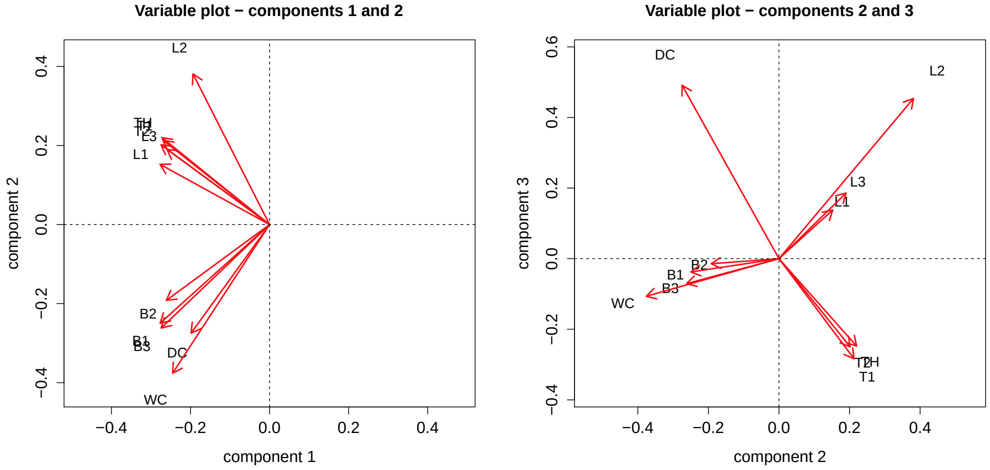

Figure 7.5: *Coefficient plots for components 1 and 2, and components 2 and 3, from a PCA of the stone axe data.*

Component	1	2	3	4	5	6	7	8	9	10	11
St.Dev.	2.8	1.3	0.7	0.7	0.6	0.4	0.2	0.2	0.1	0.1	0
Variance (%)	71.0	14.6	4.7	4.4	2.8	1.3	0.6	0.2	0.2	0.1	0
Cumulative (%)	71.0	85.6	90.3	94.7	97.5	98.9	99.4	99.7	99.9	100	100

Table 7.4: *Standard deviations, variances (%) and cumulative variances (%) for the PCs for the stone axe data.*

are described as *ad-hoc*. Jolliffe (2002: 112) comments that they are 'intuitively plausible' and 'work in practice'. One simple rule is to require that the cumulative percentage of variance accounted for by the components exceeds some threshold. Thus, 80% would lead to a choice of two components here. Choice of the threshold is arbitrary; 70% is sometimes mentioned as a reasonable lower limit, but I have seen examples (mostly archaeometric) where the first two components only account for 50–60% of the variance, but are useful. Another common criterion is to require the variances (when standardized data are used) to exceed some value such as 1 (Kaiser's rule) or 0.7. Both lead to a choice of two components here.

The information in Table 7.4 can be represented in a *scree plot*, the R version of which is shown in the left-hand plot of Figure 7.6. The idea is to identify the point at which the plot 'levels off' or, as it is sometimes expressed, an 'elbow' is evident. Once again two components are suggested. Such plots can be quite difficult to interpret; roughly, it is not unusual for them to exhibit something looking like 'exponential decay' so an elbow is not apparent.

Given that the size component is of limited interest in O'Hare (1990) it might be ignored and the scree plot rescaled, as in the right-hand plot in Figure 7.6. It

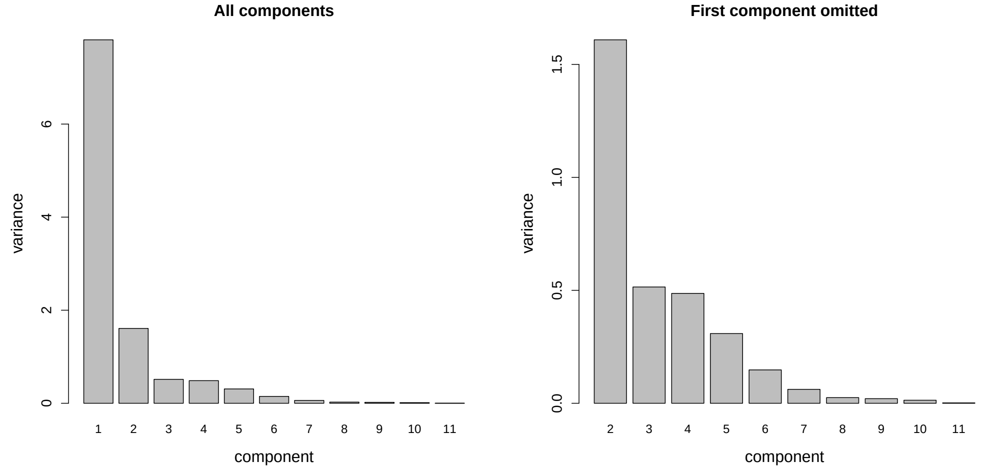

Figure 7.6: *Scree plots for the PCA of the stone axe data, with and without the first component.*

can now be seen that the second component is clearly dominant, with the next two or three contributing information.

Some of the rules just described, as Jolliffe notes, derive from studies in factor analysis, where the 'correct' choice of the number of factors can be critical. The concept of a 'correct' choice in PCA, if it has a meaning, has possibly been exaggerated in importance. My reason for thinking this is that, without committing oneself to a choice, an inspection of pairs plots of all the PCs that seem useful is perfectly possible – those that are useful being subject to more detailed scrutiny. It can happen that, for example, plots involving the fourth component reveal useful information whereas those based on the third component do not. Also, analysis often proceeds iteratively, with outliers removed and analysis repeated, for instance. Under these circumstances attempting to select a 'correct' number of components seems pointless.

Figure 7.7 is a pairs plot based on PCs 2, 3 and 4, using the version available in the `car` package. The plots are best viewed in color but, however approached, can be 'messy'. Concentrating on the plot for the second and third components, there is no discernible separation of types. Pointed axes are perhaps more evident on the periphery of the plot, but are all over the place, and are the most numerous class. Similar observations can be made about the other plots.

Given the obvious overlap of butt-types, confidence ellipses or convex hulls (Figures 6.3 and 6.5) will be of no value for separating types. Experimenting with contouring is an idea since denser regions at the centers of the scatter for each type might separate, but this did not happen and the plots are not shown. The

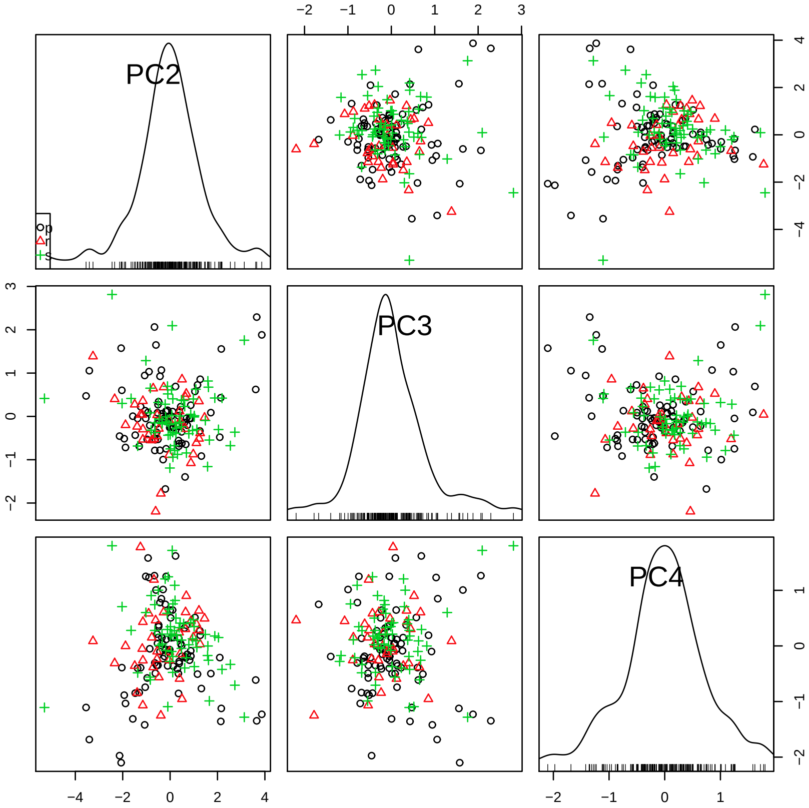

Figure 7.7: *Pairs plot for components 2-4 from a PCA of the stone axe data.*

main outcome of this analysis has been to show that there are clear patterns of correlation in the variables, but this is not reflected in any structure in the score plots related to butt type. Analysis continues in Section 8.3 where the idea of rotation of components is illustrated.

7.5 R notes

Figures 7.1 to 7.6

Some of the presentational arguments and legend have been omitted. Data for the oxides are held in `oxides` and `site` respectively. The colors (`Col`) and plotting characters (`Sym`) for the site were created separately and the relevant code is not shown. The following basic code is for Figure 7.1.

```
library(MASS)
```

```
pcO <- prcomp(oxides, scale = T)
x1 <- pcO$x[ , 1]; x2 <- pcO$x[ , 2]
y1 <- pcO$rotation[ , 1]; y2 <- pcO$rotation[ , 2]

# Score plot
eqscplot(x1, x2, main = "Standardized data")
abline(h = 0); abline(v = 0)

# Coefficient plot
eqscplot(y1, y2, type = "n")
text(y1, y2, names(oxides))
arrows (0, 0, y1 * .85, y2 * .85, code = 2, length = .15)
abline(h = 0); abline(v = 0)
```

For Figure 7.1 define `pcO <- prcomp(log10(oxides), scale = F)`; for Figure 7.3 replace `log10(oxides)` with `LR` where `LR` is the (centered) log-ratio transformation that can be obtained from the `clr` function in the `Hotelling` package

```
library(Hotelling)
    Si <- 100 - apply(oxides, 1, sum)
    oxidesSi <- as.data.frame(cbind(oxides, Si))
    LR <- clr(oxidesSi)
```

where `Si` is the 'residual', which can be equated with silica, that is used to augment the data set so that it is fully compositional.

Other than the data used, Figures 7.4 and 7.5 introduce nothing new. The scree plot to the left of Figure 7.6 might be obtained using the `screeplot` function, `screeplot(PCA, xlab = "component")`, where `PCA` is the object obtained on using `prcomp` in the first stage of analysis. This draws on the `barplot` function, which was used directly here in order to obtain the plot to the right of Figure 7.6 with `barplot(PCA$sd^2, names.arg = 1:11)` producing the plot to the left and `barplot(PCA$sd[2:11]^2, names.arg = 1:11)` that to the right.

Chapter 8

Factor analysis and PCA

8.1 Factor analysis

The history of the use of factor analysis in archaeology is recounted in Baxter (1994a, 2003). A summary is that 'factor analysis' was widely used (or abused) in applications to the mid-1980s, but has not been prominent since. It was widely confused with PCA and confusion still exists. The need for a separate chapter on factor analysis is questionable, but it has been 'promoted' in the fairly recent undergraduate text of VanPool and Leonard (2010) – at the apparent expense of PCA, though this may not be intentional – who perpetuate some of the misconceptions surrounding the distinction between the methods. This, with a summary of the historical background, is discussed in Section 8.4.

The view almost invariably expressed in texts written by statisticians, and adopted here, is that PCA and factor analysis are *different* methods and that maintaining this distinction is important. The view expressed in some archaeological writing (e.g., Drennan, 2009: 299–300) that the methods are conceptually different but that this is of no practical importance is only tenable if, having accepted there is a conceptual distinction, the consequences are then ignored. The concluding section returns to this argument.

Section 8.2 summarizes some of the main differences between PCA and factor analysis and 'problematic' aspects of application – that is, how the methods differ and why. The example in Section 8.3.1 illustrates the effects of *rotation* and the choice of coefficient constraints on the outcome of a PCA. The idea of rotation is 'borrowed' from factor analysis where it is fundamental. The intent is to take an initial 'solution' and modify it in the interests of 'interpretability'. This can be done in many different ways, so there is an unavoidable 'indeterminacy' involved in any application of factor analysis. The example in Section 8.3.2 shows how numerical results can vary as a consequence of the choices that have to be made in implemention. This is, regrettably, a subject where understanding of the mathematics that underpins the methods helps to clarify why they are different, and a succinct account is provided in Appendix D.

8.2 Theory - a brief summary

Remember that PCs are constructed as linear combinations of Y_i

$$Z_j = a_{j1}Y_1 + a_{j2}Y_2 + \ldots + a_{jp}Y_p \tag{8.1}$$

that are uncorrelated and, subject to this, account for successively decreasing amounts of the variance in the data. Determining the a_{ji} is a purely mathematical operation that depends, additionally, on constraints that it is necessary to impose for a unique solution (constraint 7.1 or 7.2)[1].

Constraint (7.1) is used in `R` and other software, where PCA and factor analysis are clearly distinguished. An important exception, and a source of confusion, is the widely-used `SPSS` package, where PCA is treated as a particular case of factor analysis and constraint (7.2) is used.

Equation (8.1), subject to the chosen constraint, has a unique solution determined mathematically. The relationship can be 'inverted' to obtain an expression for the variables as a function of the components (Section D.3.1).

$$Y_j = a_{1j}Z_1 + a_{2j}Z_2 + \ldots + a_{pj}Z_p \tag{8.2}$$

This does not involves any notion of random variation such as might be represented by an 'error' term, and does not require estimation of the coefficients with associated measures of uncertainty[2].

In contrast, and it is important, factor analysis requires that a statistical model be specified for the data. This has the form

$$Y_j = b_{1j}F_1 + b_{2j}F_2 + \ldots + b_{qj}F_q + \varepsilon_j \tag{8.3}$$

where the final term is a random component ('error' term). The *loadings* b_{ij} that determine the factors must be estimated. In PCA there are p components, of which a subset, q, may be selected for presenting results; in factor analysis there are $q < p$ factors. The hope is that these *latent variables* can be assigned a 'meaning' at the interpretive stage as unobservable variables that explain the observed covariance

[1] A distinction, not always made in the literature, will be maintained between the use of the term *coefficients* for PCA and *loadings* for factor analysis. Similarly, *component* and *factor* rotation are distinguished.

[2] To be clear about this, the way in which component coefficients are determined using mathematical methods is being distinguished from statistical estimation.

structure of the data. In PCA no such 'meaning' is necessarily attributed to the components, nor need there be for productive analysis.

The covariance matrix of the data is **S**. In PCA the coefficients are extracted via the singular value decomposition (SVD) of **S**, or some procedure with equivalent effect. This is a mathematical operation that produces PCs with the required properties. In factor analysis **S** is broken down into the sum of two components which are the contributions of the random terms and the systematic components represented by the factors as in equation (D.6). A distinction often made is that, in contrast to PCA where the aim is to account for as much variance as possible with a small number of components, the emphasis in factor analysis is on the covariance structure of the data rather than the variances, with the effort directed at modeling this in terms of *common factors* that explain this structure.

Another way of stating this is that PCA and factor analysis have different aims. Factor analysis implies a belief that unobservable variables – with a 'meaning' that can be articulated – explain the covariance structure of the data. Factor analyis results are unavoidably affected by analytical choices for which definitive statistical theoretical guidance does not exist. Some commentators are uneasy about the 'flexibility' of interpretation this allows; what is not in doubt is the 'indeterminacy' in the results (i.e. factors identified) that can be obtained. Factor *rotation* is at the heart of this.

Conditionally on the data pre-treatment used, PCA provides a unique solution to the problem it is designed for. Any attempt to modify the PCA solution destroys its optimality properties. Not so factor analysis, since it does not attempt to optimize any well defined criterion that the factors should satisfy. An initial solution (i.e. determination of the b_{ij}) is not unique. Rotation has the aim of achieving simple and interpretable structure. The idea is that it is easier to attach a 'meaningful' label to rotated factors than it is to the initial solution[3].

To summarize, in factor analysis factor rotation is *de riguer*. Formally, in PCA, components can be rotated but the results then do not have the optimal properties PCA is designed to achieve. The witting use of rotation of PCs to enhance interpretation is sometimes seen; the confusion of rotated PCA with factor analysis is more pernicious. This is discussed further, in the context of the archaeological literature, in Section 8.4. Apart from the choice of a rotation method (Section D.3.2), many methods of factor extraction are possible, contributing to the variety of solutions possible. Section D.3.3 describes some of these and Section 8.3.2 provides illustrative applications.

[3] It seems to be implicit, in this approach, that factors should have simple structure, with loadings either 'high' or 'close' to zero. It is not obvious that there is a logical reason why latent variables should have simple structure, so the requirement is really one of interpretive convenience.

8.3 Examples

8.3.1 PCA and rotation

The data used for illustration are the stone axe data of Tables B.9 to B.11 for butt types 1, 3 and 5, already analyzed in some detail in Section 7.4. The sequence of commands that follows provides the basis for the results given in the upper-part of Table 8.1, and uses the constraint in equation (7.1).

```
PC1 <- prcomp(SAst)
PC1L <- PC1$rotation[, 1:4]
PC1R <- varimax(PC1L)$loadings
```

	Components				Components			
	1	2	3	4	1	2	3	4
	PCA (`prcomp`)				Varimax rotation (cutoff = 0.3)			
L1	-0.33	0.18	0.16	0.09	-0.39			
L2	-0.23	0.45	0.53	0.40	-0.80			
B1	-0.33	-0.29	-0.04	0.21		-0.49		
B2	-0.31	-0.23	-0.02	0.37		-0.49		
B3	-0.32	-0.31	-0.08	0.20		-0.50		
WC	-0.29	-0.44	-0.13	0.06		-0.51		
DC	-0.23	-0.32	0.58	-0.67				-0.96
TH	-0.32	0.26	-0.29	-0.24			-0.55	
L3	-0.30	0.22	0.22	-0.04	-0.39			
T1	-0.32	0.25	-0.33	-0.23			-0.57	
T2	-0.32	0.24	-0.29	-0.21			-0.54	
	PCA (`principal`)				Varimax rotation (cutoff = 0.45)			
L1	0.91	0.23	0.12	-0.06	0.60		0.56	
L2	0.64	0.57	0.38	-0.28			0.89	
B1	0.91	-0.37	-0.03	-0.15		-0.88		
B2	0.86	-0.29	-0.01	-0.26		-0.83		
B3	0.90	-0.39	-0.06	-0.14		-0.89		
WC	0.81	-0.56	-0.09	-0.04		-0.91		
DC	0.65	-0.41	0.41	0.47				0.86
TH	0.89	0.33	-0.21	0.17	0.88			
L3	0.85	0.28	0.16	0.03	0.61		0.55	
T1	0.88	0.32	-0.24	0.16	0.88			
T2	0.90	0.30	-0.21	0.15	0.87			

Table 8.1: *Coefficients and varimax rotated PCs for different treatments of the stone axe data of Tables B.9 to B.11.*

Here, `PC1L <- PC1$rotation[, 1:4]` extracts the coefficients for the first four components. These are shown in the upper-left of Table 8.1 and, apart from rounding, are the same as those in Table 7.3. The eigenvalues for the first two PCs are 2.79 and 1.27; for the third and fourth PCs they are 0.72 and 0.70. The first two components account separately for 71.0% and 14.6% of the total variance, and cumulatively for 85.6%; the next two account for about 9% cumulatively so about 95% of the variance is attributable to the first four components (Table 7.4).

The interpretation in terms of size and shape components is discussed in Section 7.4.2. While several criteria for component selection in advance of rotation, such as Kaiser's rule, would lead to a choice of two, more forgiving criteria would lead to different choices (e.g., a modified Kaiser's rule using 0.7, or what VanPool and Leonard (2010: Chapter 15) state is a 'common cutoff' for the cumulative percentage used to select the number of components of 95%). Four components will be rotated for illustrative purposes.

Jolliffe's (2002, pp. 112–133) account of component selection in PCA concludes that rules having a 'sound statistical foundation' seem 'to offer little advantage over the simpler methods in most cirumstances'. Since he also notes that the simpler methods are 'very much ad-hoc rules-of-thumb' (page 112), and that the choice of the number of components to rotate 'can have a large effect on the results after rotation' (page 271) this would appear to leave the aspirant rotator of principle components in something of a quandary when deciding what to do!

The command `PC1R <- varimax(PC1L)$loadings` uses the `varimax` function to rotate the four components with results shown in the upper-right table. The common convention of suppressing rotated coefficient values below some cutoff, is followed. The default is to suppress values for which $|a_{ij}| < 0.1$. Here,

```
print(PC1R, digits = 2, cutoff = 0.3).
```

which rounds the a_{ij} to two digits and prints those for which $a_{ij} > 0.3$, is used, as in the table to the upper-right. The cutoff, 0.3, is arbitrary (as is the default) but designed to emphasize the most important clusters of variables that characterize each rotated component.

How does interpretation differ from the unrotated solution, if at all? The plots in Figure 7.5, based on the first three components, can be clearly interpreted in terms of three clusters of variables corresponding to 'length', 'breadth' and 'thickness' with depth of cutting-edge as an isolated variable. The rotated solution to the upper-right of Table 8.1 does not add to this, but loses the direct interpretation in terms of 'size' and 'shape' components evident in the unrotated analysis. We might, if the fancy takes us, call these variable clusters 'factors', but the analysis is not a factor analysis.

The lower set of tables repeats the analysis obtained with the `principal` function from the `psych` package using constraint (7.2). The number of components (called 'factors') to extract needs to be specified explicitly and varimax rotation is applied by default. To get a rotated PCA the following can be used.

```
library(psych)
PC2 <- principal(SAst, nfactors = 11, rotate = "none")
```

```
PC2L <- PC2$loadings[, 1:4]
PC2R <- varimax(PC2L)$loadings
```

The argument `nfactors` specifies the number of components/factors to extract – the maximum of 11 in this case – and the `rotate = "none"` argument suppresses rotation. Thereafter things proceed as previously. Because of the different constraints, coefficients in the two unrotated analyses differ by a constant factor. The outcome of rotation depends on the constraint used (Jolliffe, 2002: 272–74) but both lose the variance-maximization properties of the unrotated solution and the property that component scores are uncorrelated.

After rotation coefficients no longer differ by a constant factor. The results in the lower part of Table 8.1 for `principal` can be compared with those from `prcomp`. A cutoff of 0.3 was used for the varimax rotated components in the latter case as it divided the variables neatly into different types; not such a neat division was possible for the results from `principal`, a cutoff of 0.45 eventually being chosen[4].

8.3.2 Variants of factor analysis

The data used are a 'classic' set of measurements on 30 La Tène Bronze Age fibulae from Münsingen, Switzerland. These were used in several early experimental studies of applications of multivariate methods in archaeology, in the late 1960s and early 1970s. They are published as Table 9.1 in Doran and Hodson (1975) and reproduced in Table B.15. There are three angular measurements and one variable of counts, with the other dimensions measured as millimeters. One of these variables has some values that could not be ascertained, and one fibula has missing data. Doran and Hodson replace these with estimates; here the offending row and column have been omitted in the analyses to follow so a 29 × 12 data matrix is used.

Doran and Hodson (1975: 225) stress that the data set, which is small, was intended to 'test out alternative methods and *not* to provide a useful archaeological classification' (their emphasis). This is the spirit in which the data are used here.

Other than the angular data, and following Doran and Hodson, variables are transformed to logarithms before analysis. There are some zero values; following Doran and Hodson 0.1 was added to all the non-angular data before taking

[4] With the `principal` function, rotated components do not necessarily retain the same ordering when compared to the original components they most resemble. Rotated components 2 and 4 in both analyses have a similar interpretation; component 1 in the `prcomp` rotation resembles the third rotated component for `principal`; and the third rotated component for `prcomp` and first component fro `principal` may be interpreted as 'thickness'. Broadly, though, the analyses lead to similar interpretations.

(natural) logarithms[5].

Following suggestions in Section D.4 for comparisons that might be of interest the upper part of Table 8.2 contrasts the results of orthogonal varimax rotation using principal axis and maximum-likelihood factor analysis; the lower part of the table provides a similar contrast using oblique oblimin rotation. The default in the `fa` function from the `psych` package was used for all analyses.

	Factors				Factors			
	A	B	C	D	A	B	C	D
	Principal axis factor analysis, varimax rotation				Maximum-likelihood, varimax rotation			
FL		0.83				0.81		
BH	0.59	-0.51		-0.45	0.58	-0.47		-0.50
CD	0.65				0.60			
ED	0.49		0.67		0.46		0.73	
FEL		0.88				0.94		
C	0.59	0.54		0.53	0.53	0.42	0.43	0.59
BW								
BT		-0.46	0.54	-0.43		-0.45	0.57	-0.59
Coils			0.68				0.48	
BFA	-0.74			0.40	-0.83			
FA				0.77				0.73
BRA	-0.85				-0.82			
	Principal axis factor analysis, oblimin rotation				Maximum-likelihood, oblimin rotation			
FL			0.50	0.54	0.41		0.51	0.54
BH	-0.59	0.43			-0.55	0.45		
CD	-0.44			-0.41				
ED		0.65				0.71	0.45	
FEL				0.91				0.93
C			1.01				0.99	
BW		0.45						
BT		0.69				0.88		
Coils	0.42	0.67				0.46		
BFA	0.84				0.93			
FA	0.44		0.44	-0.46		-0.41		-0.44
BRA	0.84				0.80			

Table 8.2: *Factor loadings from different analyses of the Bronze Age fibulae data of Table B.15, treated as described in the text and extracting four factors.*

As in Doran and Hodson (1975: 200–01), using Kaiser's rule, four factors were extracted. The code that follows is for principal axis factor analysis with varimax rotation.

```
FA <- fa(y, nfactors = 4, fm = "pa", rotate = "varimax")
```

where `y` is the data matrix determined as previously described.

[5] I have been unable to reproduce the results of this transformation given in Table 9.2 of Doran and Hodson, other than for the counted variable. My numbers don't differ too much from theirs, and I get similar results when reproducing their analyses.

The method selection argument `fm = "pa"` selects principal axis factoring, while `rotate = "varimax"` gives the rotation method to use. Change these to `fm = "ml"` and `rotate = "oblimin"` for maximum-likelihood estimation and oblimin rotation. There are other options available; `"oblimin"` is the default rotation and `"minres"` the default estimation method. The latter produces an ordinary least squares solution; the documention states it will produce results very similar to maximum-likelihood and it did so here (not shown).

The loadings were obtained with a cutoff of 0.4 for all the analyses reported, using the `loadings` function and the `print` method associated with it.

```
print(loadings(FA), digits = 2, cut = .4)
```

In Table 8.2 factors are ordered to make comparisons more readily between different analyses, rather than in the order that occurs in `fa` output.With relatively minor variations the two varimax analyses show fairly similar results to each other, and the two oblique rotations are also quite similar to each other.

The more obvious differences that can be seen, which do not depend on the cutoff used, arise in the comparison of the orthogonal and oblique rotations. Most obviously, perhaps, the catchplate dimension (C) does not stand out in any of the factors obtained using the varimax rotation, but dominates the third factor in the oblimin rotations. Similarly, factor D is dominated by the foot extension length (FEL) which has a high loading for factor B in the varimax rotations but does not dominate (or define) that factor. It is, in fact, difficult to discern much correspondence between the varimax and oblimin rotations. There is some similarity between the third factor for the former and the second factor for the latter.

For this example at least the results are sensitive to the choice of rotation method. The other potentially important source of difference concerns the choice of numbers of factors to rotate (Section D.4). For illustration Table 8.3 presents the results from a maximum-likelihood analysis using both varimax and oblimin rotation and extracting three factors. This may be compared with the output from Table 8.2.

For the varimax rotation factor C in Table 8.2 and factor B in Table 8.3 compare reasonably well; there is also some correspondence between factor A in the different analyses. Factor C in the three-factor analysis has no clear relationship to either factor B or D in the four-factor analysis or to any simple combination of them.

For the oblimin rotation Factor C in all analyses, which is dominated by the catchplate variable, corresponds well. The bow-angle variables are the most important in defining factor A in all analyses but the contributions of foot length (FL) and coil diameter (CD) make somewhat greater and non-trivial contributions in the three-factor analysis.

This leaves factor B in the three-factor analysis to be compared with factors B and D from the four-factor analyses. There is a quite good correspondence with

	Factor A	Factor B	Factor C	Factor A	Factor B	Factor C
	Maximum-likelihood, varimax			Maximum-likelihood, oblimin		
FL	-0.42		0.74	0.61		0.58
BH	0.72		-0.56	-0.66	0.44	
CD	0.70			-0.62		
ED	0.51	0.72			0.72	0.43
FEL	-0.47		0.44	0.45		
C	0.41		0.83			0.99
BW						
BT		0.62	-0.68		0.91	
Coils		0.50			0.50	
BFA	-0.83			0.87		
FA			0.50			
BRA	-0.84			0.85		

Table 8.3: *Factor loadings from maximum-likelihood factor analyses of the Bronze Age fibulae data of Table B.15, using varimax and oblimin rotation for three factors.*

factor B from the four-factor oblimin rotation; less so with the varimax rotation. Factor D in the four-factor analysis, largely determined by foot extension length, does not correspond to anything in the three-factor analysis.

The sole intention here has been to demonstrate that the choices to be made in conducting a factor analysis can have a non-trivial effect on the numerical output obtained. This Tables 8.2 and 8.3 do, particularly with respect to the rotation method used and numbers of factors rotated.

8.4 Factor analysis in archaeology

The importance of the paper by Binford and Binford (1966), which popularized the use of factor analysis in archaeology in the 1970s, is widely recognized (e.g., Doran and Hodson, 1975: 203–05; Orton, 1980: 136–39; Read, 1989) even when commentators disagree with aspects of it. Read (1989: 6–7) commented to the effect that the Binford's application of 'factor analysis', which was actually PCA with rotation, could serve 'as an example of incorrectly applied statistical methods'.

Vierra and Carlson (1981) listed over 70 studies between 1970 and 1978 that called their methodology 'factor analysis'. Their Table 1 gives details of 43 applications; more than half (27) were PCA, mainly with varimax rotation. Baxter (1994a: 277–79) notes about 20 additional analyses, mostly to the mid-1980s, that were dominated by PCA with varimax rotation. That is, more than 20 years after the introduction of factor analysis into the archaeological literature, confusion still existed between it and PCA.

There was a visible decline in archaeological uses of factor analysis and other multivariate methods from the mid-1980s or so arising from a backlash against the (mis)use of these methods. Baxter(1994a: 8) suggested that a consequence of this was that ‘methodologically useful babies were unfairly thrown out with the theoretical bathwater’. Happily most of these babies survived to grow to maturity and become useful citizens in the world of quantitative archaeology; however, it was also suggested that factor analysis had ‘possibly been lost with the bathwater, but the loss [was] not necessarily one to mourn’ (Baxter, 1994a: 86).

Doran and Hodson (1975: 197–205) summarize their attitude towards uses of factor analysis up to that data as ‘not very favourable’. Not all quantitatively able archaeologists have been so troubled. Cowgill (1977a), in his review of Doran and Hodson (1975), thought they exaggerated the distinction between PCA and factor analysis. An obvious comment is that if the techniques really are so similar, why bother with anything other than the simpler PCA methodology?

We return to this after first discussing VanPool and Leonard’s (2010) treatment of the subject which, in my view, unwittingly encapaulates many of the reasons for the confusion between PCA and factor analysis in archaeological usage[6]. Put bluntly, I think they ‘oversell’ factor analysis, avoiding contentious issues that arise in using it. The distinction made between factor analysis and PCA is confused and the subject of what follows.

There is an absence of reference to texts that might be viewed as forerunners or ‘competitors’ (Doran and Hodson, 1975; Orton, 1980; Shennan, 1997; Drennan, 2009; Fletcher and Lock, 2005) so the reader is not exposed to more qualified assessments of factor analysis that have been voiced within some of these texts. This comment also applies to the wider journal literature. References to practical applications are very limited and hardly calculated to persuade the reader that factor analysis is a ‘live’ topic in archaeology.

The exception to this general lack of acknowledgment of a critical literature is Jolliffe (2002) who is cited in support of views expressed by the authors that arguably misrepresent what he says. It is stated that ‘[p]rincipal component and factor analysis are *very similar* but differ in the way they measure variation’ (my emphasis) citing Jolliffe (2002: 180–96) as the authority for this. Jolliffe says no such thing; both PCA and factor analyses measure (co)variance in the same way, but the emphasis in PCA is on accounting for the variance (the diagonal elements of the covariance matrix) whereas factor analysis concentrates on modeling, or ‘explaining’, the covariance structure (off-diagonal elements) of the covariance matrix.

More worryingly, contrast the claim that the two methods are ‘very similar’ with what Jolliffe writes. To wit, ‘the view [that PCA is a special case of factor analysis] is misguided since PCA and factor analysis, as usually defined, are really quite distinct techniques’ (p. 150); ‘a major distinction between factor analysis and PCA is that there is a definite model underlying factor analysis, but for most

[6] I’m working from the Kindle edition of the book, so can’t give exact page references.

purposes no model is assumed in PCA' (p. 158); and 'there are many ways in which PCA and factor analysis differ from one another' (p. 160). This is not a ringing endorsement of the claim that the methods are 'very similar'.

VanPool and Leonard state that '[p]erhaps it isn't necessary to go so far as to say one should never use principal component analysis, but it is fair to say that it should only be used when the researcher can be reasonably sure that specific variance and error is small'. This follows an apparently supportive quote from Jolliffe. It can be read as asserting a preference for factor analysis over PCA and can charitably be described as misleading. To impose sense on it, it needs be interpreted as saying that *if* factor analysis is the appropriate method of analysis then PCA, as a surrogate for a proper factor analysis, should not be used.

This is important and worth spelling out in detail since it is at the heart of the confusion between factor analysis and PCA. The supposedly 'supportive' remark of Jolliffe is prefaced with another quote to the effect that '*various authors*' (my emphasis) have concluded that '... principal component analysis should not be used if a researcher wishes to obtain parameters reflecting latent constructs or factors'. This is from a single author, Widaman (1993), quoted exactly as above by Jolliffe (2002: 161). Widaman's remark was made in the context of an earlier 1990s disussion which, in Jolliffe's words, was underpinned by the 'assumption that unobservable factors are being sought from which the observed behavioural variables can be derived'. Jolliffe concludes that 'Factor analysis is clearly designed with this objective in mind, whereas PCA does not directly address it. Thus, at best, PCA provides an approximation to what is truly required'. Only this last sentence is referenced by VanPool and Leonard as a 'supportive quote'; without very careful reading and reference back to the original source it is all too easily understood as a fairly general view of Jolliffe, rather than specific comment on the view of a particular scholar made in the context of a focused 1990s debate in the behavioral sciences literature.

The fundamental problem with VanPool and Leonard's treatment is that it appears to be founded on the 'PCA as a special case of factor analysis' philosophy, with PCA then found wanting, rather than the 'PCA and factor analysis as distinct methods' philosophy. This perpetuates the crop of confusion between PCA and factor analysis sown by Binford and Binford (1966) that Read (1989) suggested was an exemplar of 'incorrectly applied' methodology.

Another problematic assertion is the statement that 'From the perspective of both techniques there are three "types" of variation in a data set; common, specific, and error ... Both [techiques] are excellent means of measuring common variance, but they differ in their treatment of specific variance and error.', followed by factor analysis 'only measures common variance' whereas PCA 'doesn't mathematically discard the specific variance and error as factor analysis does'. The PCA formulation does not involve any conception of specific or error variance; the statement only makes sense if PCA is treated as an inferior way of undertaking a factor analysis, rather than as a method in its own right.

This view that PCA and factor analysis are properly treated as distinct methods is overwhelmingly that of texts dealing with factor analysis and PCA writen by statisticians. Chatfield and Collins' (1980: 89) comment, in their introductory text on multivariate analysis, that 'we recommend that factor analysis should not be used in most practical situations' is at one extreme, but not untypical. Jolliffe (2002) provides a more dispassionate account. Other works in the same vein as Jolliffe's are listed in Section D.4. All are agreed that factor analysis and PCA are different methods; that the former involves a model for the data; and that *if* this model is appropriate PCA (with or without rotation) is not an optimal method for extracting the factors of interest. Claims that the two methods are 'very similar' or that they typically lead to very similar results (which can be queried[7]) fail to acknowledge what many scholars regard as fundamental differences between the methods.

One can take an entirely pragmatic view of this and ask how useful the methods have proved to be for archaeological data analysis. It would be straightforward to put together a book – albeit repetitive in places – consisting solely of uncontentious applications of PCA that produce readily understood results that have been found to be useful; such a book would be populated with examples from the mid-1960s to the present day. It would, I suspect, be a major challenge to do this for applications of factor analysis, avoiding studies where PCA masquerades as factor analysis. If the relative merits of the two methods must be discussed this might be seen as an acid test. Perhaps it isn't necessary to go so far as to say one should never use factor analysis (and indeed would be foolish to do so) but it is fair to say that it should only be 'promoted' if it has been demonstrably useful, and where the distinction between it and PCA is clearly maintained.

[7] The *raison dêtre* of factor analysis is that rotation, and the choice of method of rotation, *does* make a difference. It would seem to follow logically that claims to the effect that PCA and factor analysis lead to very similar results are only tenable if you envisage a PCA solution being subjected to the same rotational procedures as the factor analysis to which it is being compared. Even if it then turns out that claims about similarity are valid the thinking is predicated on the 'PCA as a special case of factor analysis' philosophy.

Chapter 9

Correspondence analysis

9.1 Introduction

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

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

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

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

9.2 CA and PCA – similarities and differences

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

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

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

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

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

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

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

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

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

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

Thus, X^2 can be written as

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

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

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

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

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

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

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

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

9.3 Romano-British glass assemblages

9.3.1 First- to third-century vessel glass

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

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

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

Figure 9.1 shows the outcome of a CA using the `ca` function from the `ca` package. The inertias are shown on the axes. The obvious interpretation is, again, chronological. Three of the later sites sit within the region occupied by the earlier sites, or are on the same side of the plot. This is explored in the paper, where the decision was made to analyze the two chronological groups separately, as is done in the next two subsections. Some of the diagnostic information available is illustrated in Table 9.1. Code for carrying out the analysis is given in Section 9.6

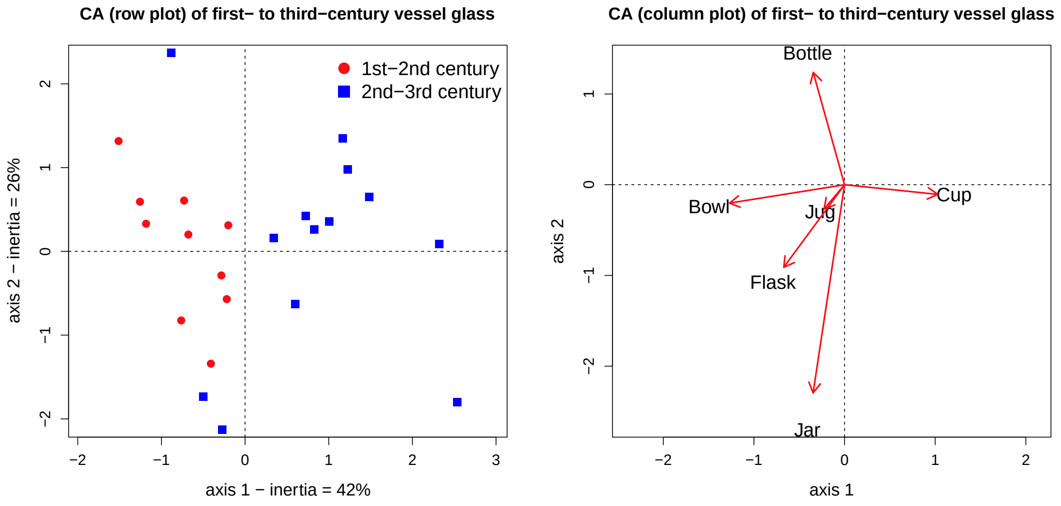

Figure 9.1: *Row and column plots for a correspondence analysisof the first- to third-century AD Romano-British vessel glass assemblages of Table B.13.*

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

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

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

9.3.2 First- to second-century vessel glass

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

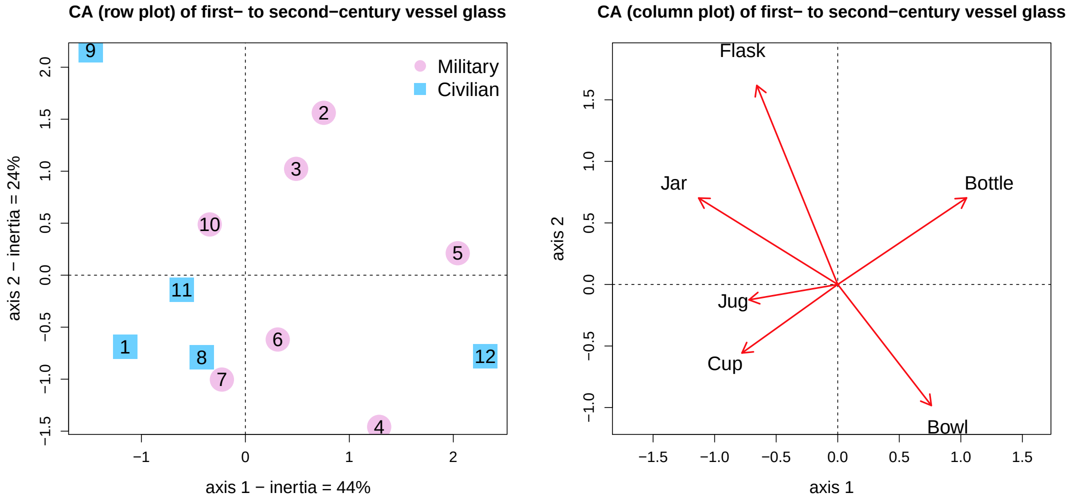

Figure 9.2: *Row and column plots for a correspondence of the first- to second-century AD Romano-British vessel glass assemblages of Table B.13.*

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

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

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

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

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

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

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

9.3.3 Second- to third-century vessel glass

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

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

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

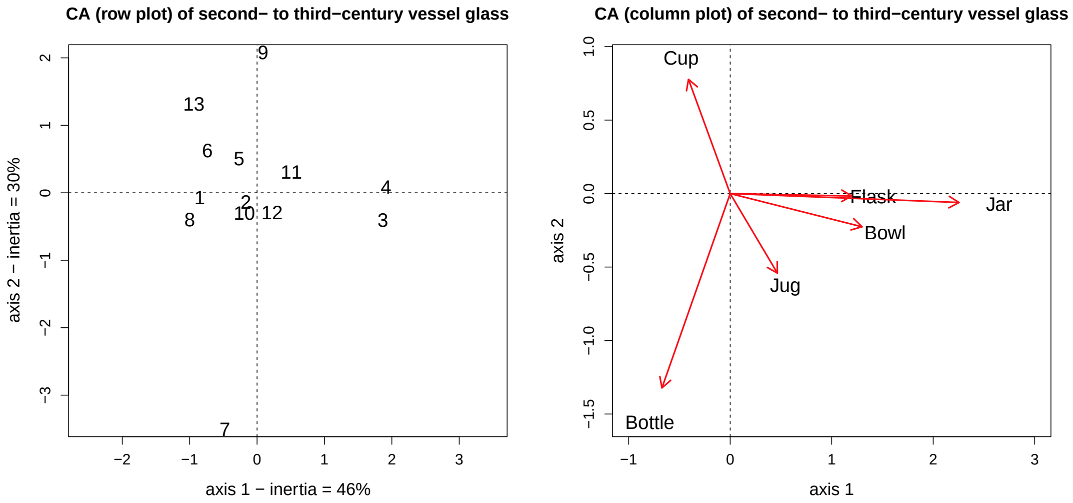

Figure 9.3: *Row and column plots for a correspondence analysis of the second- to third-century AD Romano-British vessel glass assemblages of Table B.13.*

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

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

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

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

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

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

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

9.4 Flavian drinking-vessels

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

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

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

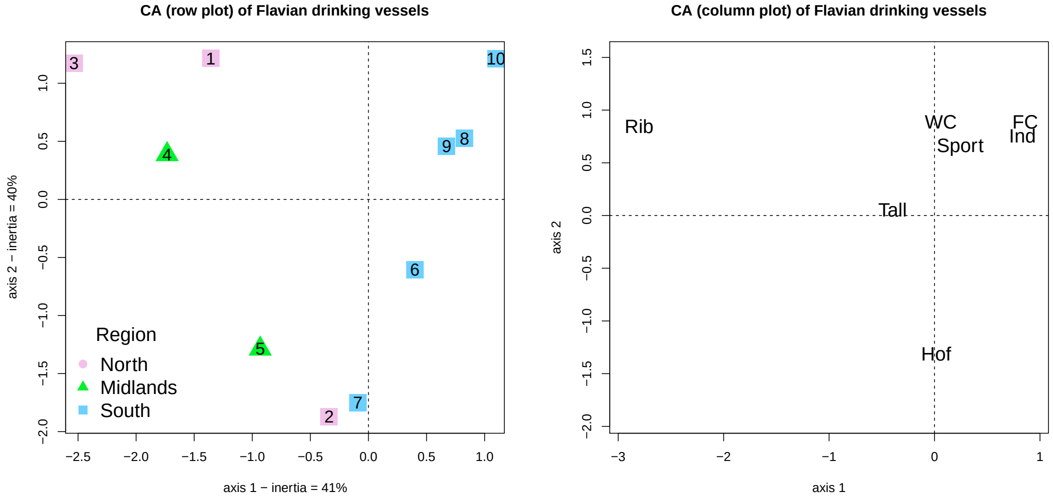

Figure 9.4: *Correspondence analysis row and column plots for the Flavian drinking-vessels of Table B.14.*

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

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

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

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

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

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

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

The thought, though it can be differently articulated, is not terribly original; something of the kind is expressed in Section 3.15 of Doran and Hodson (1975), and others, around at the time when quantitative methodology began to be widely applied, grappled with the conflicting requirements of archaeological and statistical inference. In the context of debates that were taking place at the time it might, fancifully, be seen as a choice between the Scylla of unquestioning acceptance of the rigor and utility of statistical analysis for archaeological purposes, and the Charybdis of complete rejection of statistical methodology. This depends, of course, on the conception one has of what statistics is 'about', and there was particular concern with the merits of 'classical' methods of statistical inference (Chapter 12). The extremes of both positions had their proponents; the issues raised have not disappeared, though they can be discussed in a more sober fashion that was sometimes evident. The strait between the two is not so narrow that it can't be negotiated.

9.5 Anglo-Saxon male graves and seriation

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

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

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

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

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

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

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

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

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

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

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

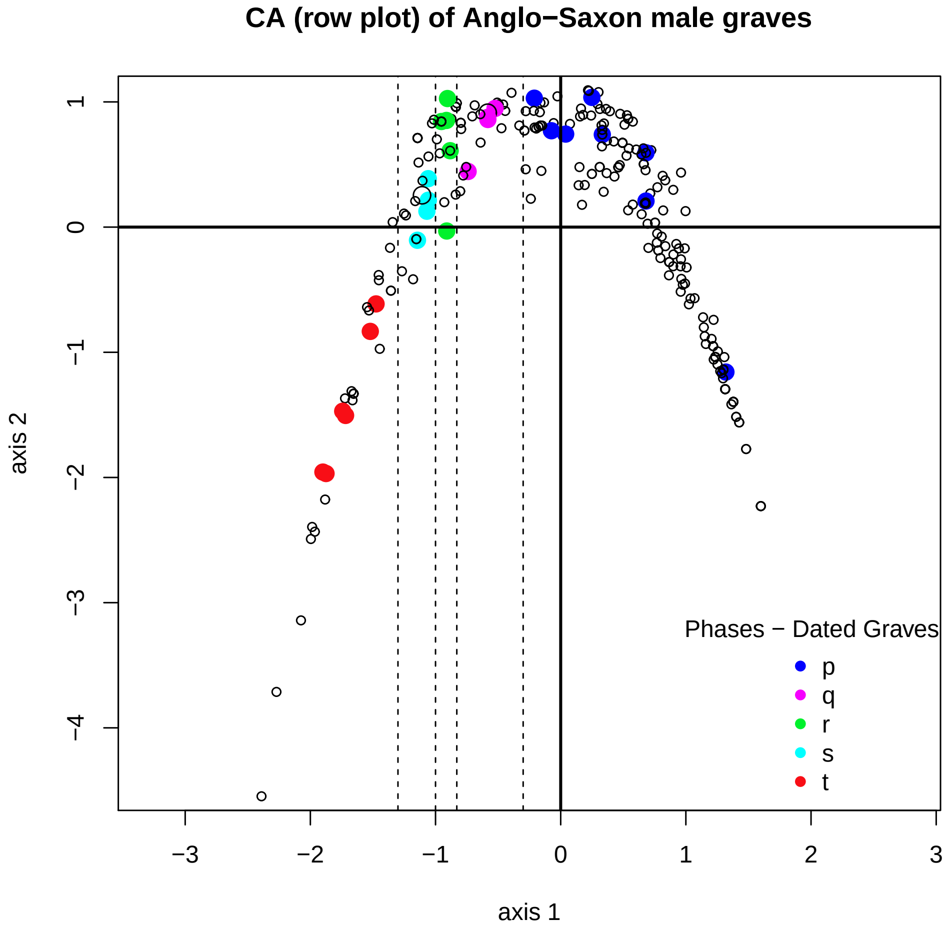

Figure 9.5: *A seriation of 6th and 7th century Anglo-Saxon graves from Britain. Dashed vertical lines are phase boundaries. (Data source: Bayliss* et al.*, 2013.)*

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

Overall the analysis is a thorough attempt to reconcile relative with 'absolute' dates. Previous work on these lines I have seen has been hampered by a paucity of dates, and is not always applied 'in anger'. From the perspective of these notes Chapters 6 and 7 make considerable use of statistics and are central to the book. By common consent they are also a very 'difficult read', particulary for nonstatisticians; Baxter (2014a, b) provides a commnentary that attempts to separate the essential material from the considerable and less essential detail .

9.6 R Notes

Figures 9.1 to 9.5

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

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

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

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

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

Chapter 10

Cluster analysis

10.1 Introduction

10.1.1 Main ideas

Cluster analysis is a generic term for a range of methods aimed at identifying groups in a set of data. It is probably the most widely used multivariate method in archaeology. To give only a few examples, cluster analysis has been used to group artifacts on the basis of their dimensions or chemical compositions; assemblages on the basis of the similarity of their profiles; and to spatial clustering on the basis of the location of artifacts in space.

Many methods of cluster analysis result in the identification of G groups, with the hope that cases in a group are similar to each other and dissimilar from cases in other groups. This introduces the idea of (dis)similarity, which is crucial to an understanding of how many methods of cluster analysis work. Many measures of (dis)similarity can be defined, contributing to the many methods of cluster analysis available. Another reason for this proliferation is that, given a measure of (dis)similarity, a large number of clustering algorithms have been proposed for the subsequent grouping exercise. This chapter discusses the most common methods used in archaeological practice.

In some ways these have not changed much since the earlier days of exploration in the 1960s and 1970s (Doran and Hodson, 1975; Hodson, 1969, 1970; Bieber *et al.*, 1976), which are often more interesting in the way cluster analysis was exploited that it commonly is now. A lot of research has subsequently been undertaken on more complex methods but most have, as yet, found limited archaeological application. Some of these are discussed in Section 10.3. Before discussing methods in more detail a small example is provided to illustrate ideas and issues.

10.1.2 Example – Blue medieval window glass

The data given in Table B.16 in Appendix B were originally published by Cox and Gillies (1986) and have subsequently been reanalyzed by Baxter (1989), Bell and Croson (1998) and others, usually for the purpose of methodological illustration. The measurements, the chemical composition in percentages for 11 oxides, are for 27 specimens of blue medieval glass from the windows of York Minster and elsewhere. It was of interest to see if the Minster glass was distinct from other sources. A basic cluster analysis is shown in Figure 10.1 and can be obtained in one line of code using

```
plot(hclust(dist(scale(york)), method = "a"))
```

though it makes for easier reading if broken down into its component parts. To herald discussion in Section 10.2 some of these components are discussed now.

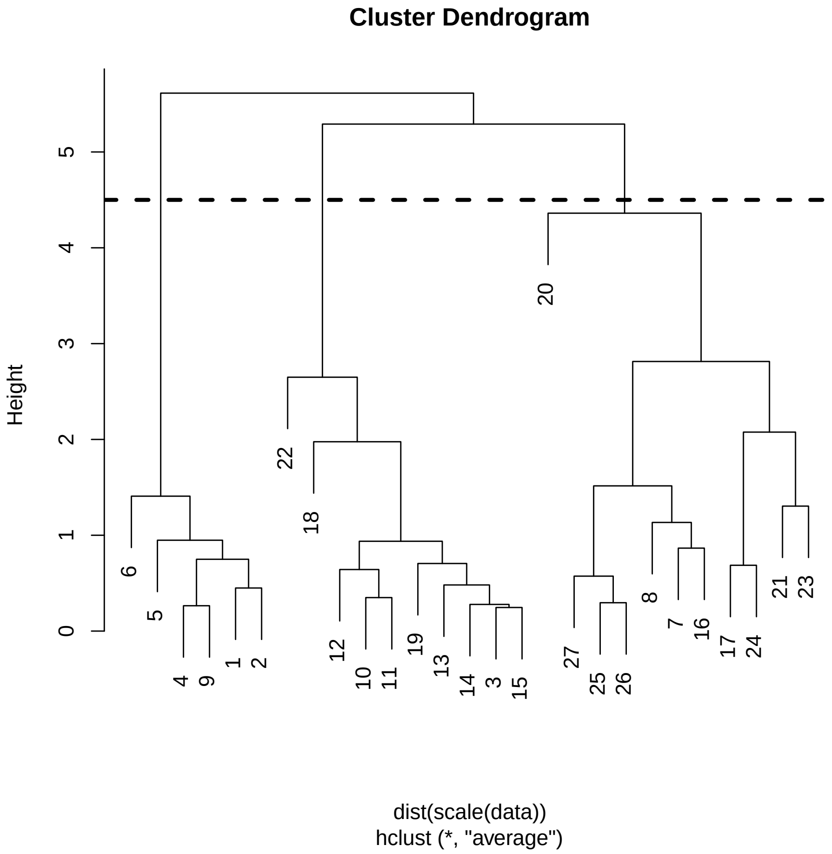

Figure 10.1: *An average-link cluster analysis for the standardized medieval glass compositional data from Table B.16.*

The `method = "a"` argument specifies that the average-link method of clustering is to be used. This is a *hierarchical* clustering method of which several alternatives are available in `R`. The same considerations concerning data transformation occur with cluster analysis as with principal component analysis (Section 7.2); thus `scale(york)` standardizes the 27 × 11 data matrix (named `york`). The `dist` argument computes Euclidean distances (Section 7.3) between the rows of scaled data; `hclust` executes the analysis; and `plot` displays the result in the form of a *dendrogram*, as in Figure 10.1.

This can be thought of as a tree consisting of branches and leaves which are the individual cases. The idea is to identify branches that contain leaves that are similar to each other, in terms of the distance between them, and at some distance from leaves associated with other branches. This is not always easy and not always possible. The most common practice is to cut the tree at some chosen height to identify distinct branches. A cut that gives three clusters is shown at a height of 4.5. The choice is subjective; if a cut is made at a slightly lower height four clusters are obtained, once containing a single case, 20, that may be an outlier. Splitting the coding above as

```
clus <- hclust(dist(scale(data)), method = "a"); plot(clus)
```

allows the object `clus` to be interrogated. Thus, `cutree(clus, h = 4.5)` identifies cluster membership for the cut at a height of 4.5, as follows,

```
 1 1 2 1 1 1 3 3 1 2 2 2 2 2 2 3 3 2 2 3 3 2 3 3 3 3 3
```

which is useful for labeling purposes in further analysis. It is possible to specify the number of clusters required, using `k = 3` rather than `h = 4.5`.

Checking the validity of a proposed cluster is not always easy. Clustering algorithms are designed to identify clusters even when they do not exist. Given the clusters, a simple method of assessing their integrity is to use labeled principal component plots. This is shown for a scatterplot matrix of the first three components in Figure 10.2, where the three clusters are mostly clearly distinct. Clusters 1 and 2 are very tightly defined, the former in particular, apart from one case that is outlying relative to the rest of the cluster. Cluster 3 is rather more dispersed but plots coherently on the first two components.

The default output from the `plot` command is invaluable for a quick look at the data, but some sort of enhancement is desirable for presentational and interpretive purposes. Figure 10.3 illustrates some possibilities. Default titles have been removed or replaced so that the figure is more informative. Readers familar with applications of cluster analysis may not be familiar with the default style of presentation used in `R` in Figure 10.1. Figure 10.3 may be a more familiar representation, where all the leaves ‘descend’ to a base of zero. This is obtained by including the argument `hang = -1` in the `plot` command.

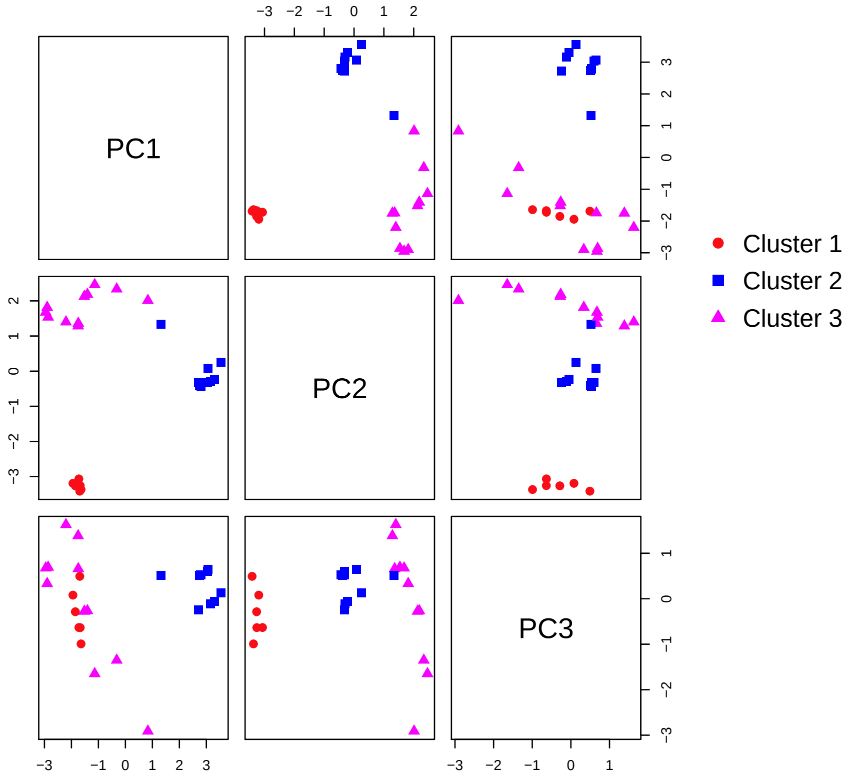

Figure 10.2: *A scatterplot matrix for the first three components of a PCA of the medieval glass compositional data from Table B.16, showing cluster labeling from an average-link cluster analysis of the data.*

The coloring requires more explanation. The default labeling in R is by case number. This can be replaced by other text, such as cluster identifications (1, 2, 3). With many observations the labeling can be unreadable unless corrective action is taken, by splitting the dendrogram into component parts, for example. The use of different colors and symbols as labels can make it easier to read the dendrogram. Color labeling is illustrated in Figure 10.3. It is uninformative in this example because the dendrogram itself defines the labeling. It is useful, however, for looking at the extent to which other methods of clustering reproduce the results. This is illustrated in Section 10.2.2 .

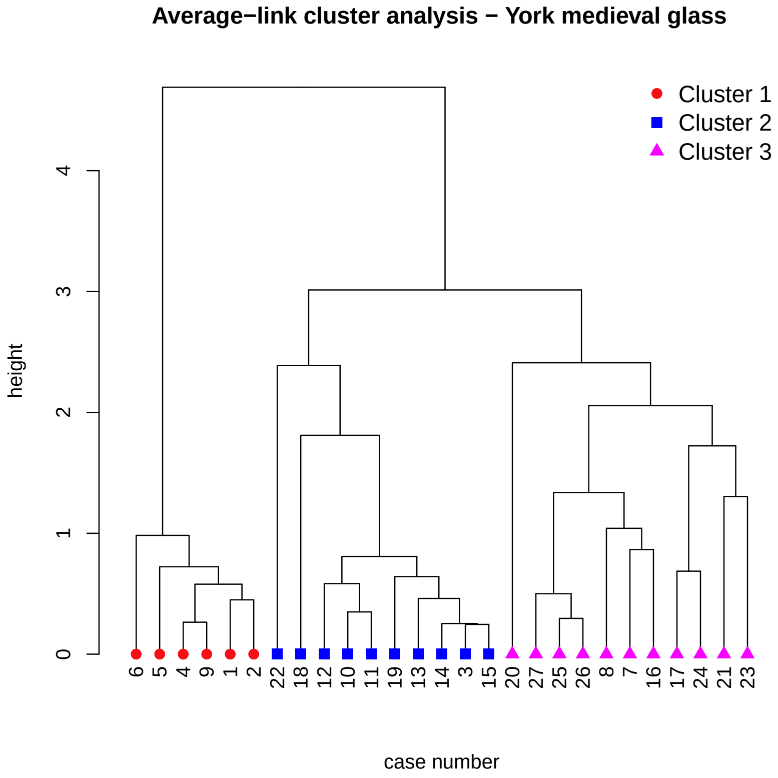

Figure 10.3: *An average-link cluster analysis for the standardized medieval glass compositional data from Table B.16 - an enhanced version of Figure 10.1.*

10.2 Hierarchical Clustering

10.2.1 The most commonly used methods

Hierarchical agglomerative methods of cluster analysis are those most commonly used in practice. Each case is initially treated as a single cluster so there are n in all. The two most similar cases are merged to form a cluster of two cases, giving $(n-1)$ clusters. Thereafter, clusters are successively merged (treating single cases as clusters) on the basis of which pair is most similar at any stage. Eventually all cases are merged into a single cluster. It is possible to start by assuming that all cases belong to a single cluster and then successively split clusters up, one case at a time, until all cases are distinct. This method, hierarchical divisive clustering, has had comparatively limited use, and will not be considered further.

To merge clusters a measure to determine how similar clusters are is needed. Similarity can be defined in different ways. In *single-link* cluster analysis the similarity of two clusters is measured by the smallest distance between two cases, one from each cluster. The two clusters merged are those for which this smallest distance is smallest. In *complete-link* cluster analysis, similarity is defined by the largest distance between two cases, one from each cluster, the clusters being merged for which this largest distance is smallest.

Single-link cluster analysis is rarely used because it tends to produce uninterpretable results unless the structure is obvious. It is sometimes useful for detecting outliers. A criticism of both single- and complete-link clustering is that the measure of similarity between clusters depends only on two cases, and fails to take account of group structure. *Average-link* cluster analysis attempts to overcome this problem by defining similarity between clusters as the average distance between all possible pairs of cases, one from each cluster. It has probably been the most widely used method of cluster analysis in archaeology. Ward's method (Section 10.3) also takes group structure into account.

The results from a hierarchical cluster analysis need to be validated and interpreted. This is usually done using a dendrogram, useful in conjunction with PCA. Cases that merge at a low level (e.g., 4 and 9 in Figure 10.1) show a high level of chemical similarity. The appearance of a dendrogram depends on the style of presentation, choice of method, and the distance measure used.

10.2.2 Example – Levantine glass compositions

The York data used so far is not especially suitable for exploring the issues raised above. The structure is obvious and recovered by all the methods mentioned. For further illustration a 67 × 5 data matrix showing the compositions of Levantine glass found at primary glass-production sites in Israel, from the first centuries AD, is used (Table B.17).

A Ward's method analysis using standardized data is shown in Figure 10.4. The appearance is 'cleaner' compared to the average-link analyses previously presented, making it easier to select a level of clustering to work with. A cut at a height of 7 produces seven reasonably convincing looking clusters. A six-cluster 'solution' is also defensible; a three-cluster solution ignores some of the structure in the lower part of the plot. The default dendrogram configuration from `R` is used but this is not very evident from the plot. This is a function of the way Ward's method can tend to work, where the depth to which the leaves hang look mostly the same.

The exception to this comment is one case from Cluster 4 which, as will be seen later, is an extreme outlier. Ward's method can be poor for outlier detection. This is a consequence of the 'model' that implicitly underpins the method and is discussed further in Section 10.3.

In a sense single-link analysis is at the opposite pole to Ward's method. The cluster analysis for single-link for the Levantine data is shown in Figure 10.5.

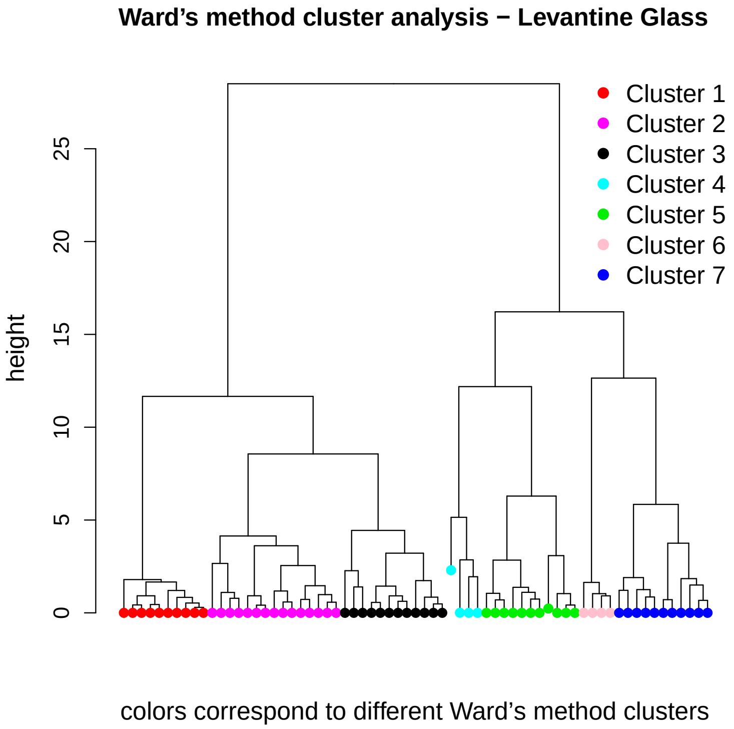

Figure 10.4: *A Ward's method cluster analysis for the standardized Levantine glass compositional data.*

It should be emphasized that the cluster identifications are those derived from the Ward's method analysis. A reason for the lack of use of single link in the archaeological literature is evident from the plot, and this is the phenomenon called 'chaining'. This arises because otherwise distinct clusters can be linked because of the effect of a small number of cases that are intermediate between the otherwise disparate clusters. In its purest manifestation the dendrogram will have a 'staircase' like appearance that makes cluster identification impossible without the use of externally derived cues. The dendrogram in the figure is not quite that bad but apart from some outliers to the left and two small groups there is no obvious structure. If the 'cues' provided by the clustering from the Ward's method analysis are used it can be seen that apart from Cluster 6, and the partial exception of Cluster 1, there is no real match, with cases from different clusters scattered throughout the dendrogram.

The contrast with the average-link cluster analysis in Figure 10.6 is more interesting. Average-link is possibly the most widely used, in archaeology, of the available methods. There are several reasons for this; is is the default in several

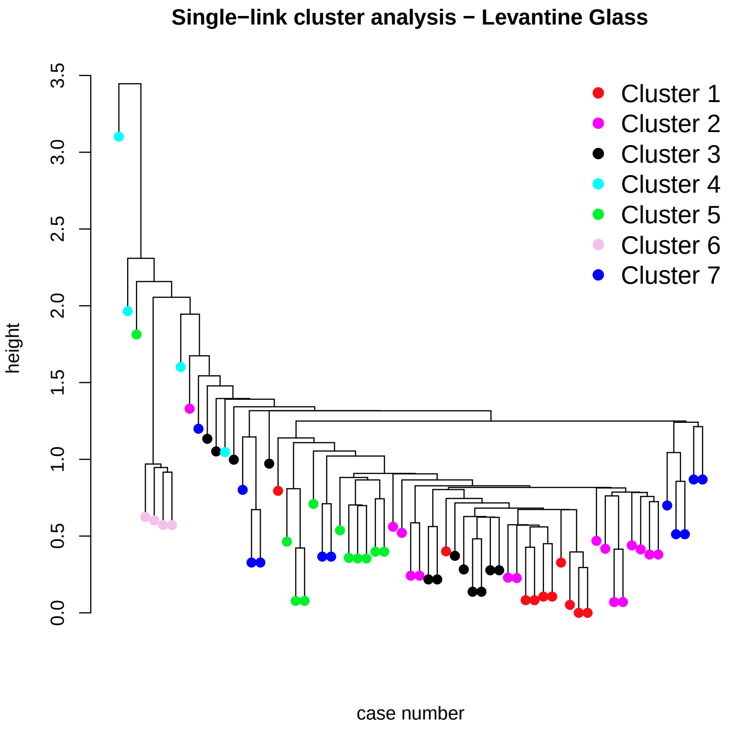

Figure 10.5: *A single-link cluster analysis for the standardized Levantine glass compositional data.*

popular software packages; it was 'promoted' from an early stage once archaeology engaged with quantitative ideas (e.g., see Doran and Hodson, 1975: 177); and (one hopes) practitioners have found it useful.

Coming to the dendrogram 'cold', without additional cues, interpretation is not straightforward. The eleven cases to the left can be treated as 'outlying' and include the two small Clusters 4 and 6. Visually cutting the dendrogram at between 2 and 3 (allowing the cut-height to vary) suggests three clusters. Stray cases apart, that to the right can be identified with cluster 5 from the Ward's method analysis, and that to the left with Cluster 7. The larger central cluster could be cut at just below a height of 2 to give a subdivision of three clusters, one of which can be identified with Cluster 1 from the Ward's method analysis. There is a small cluster of six cases, all from Cluster 2, with the remaining cluster mixing cases from the Ward's method Clusters 2 and 3.

Thus the correspondence between the Ward's method and average-link results is reasonable, with 5/7 of the clusters from the former method blocking together on the dendrogram for the latter method. The average-link analysis might thus

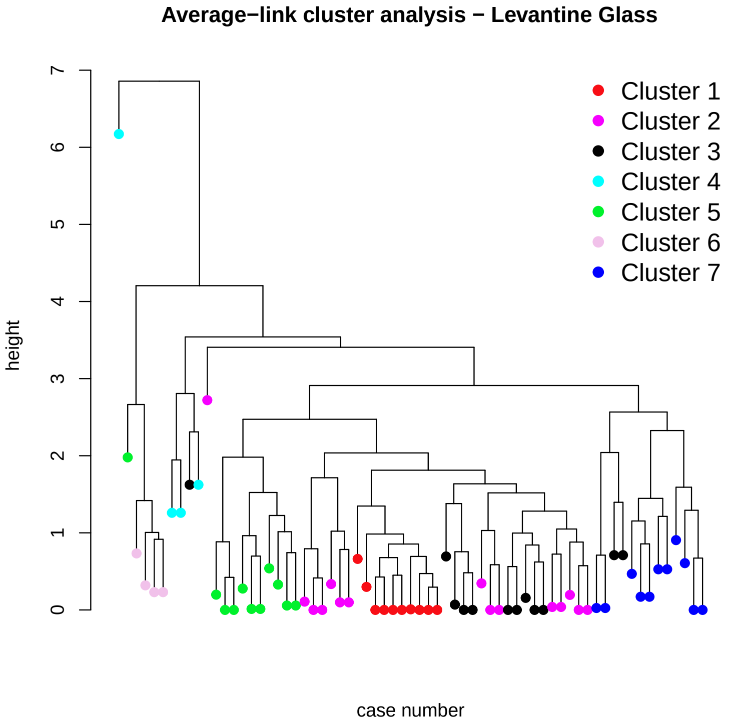

Figure 10.6: *An average-link cluster analysis for the standardized Levantine glass compositional data.*

be regarded as producing a more nuanced analysis than Ward' method.

Just because the different methods produce moderately similar results doesn't mean they are 'right'. It is advisable to check on this and PCA is one way of doing so[1]. Figure 10.7 shows plots based on the first three PCs.

In the plot of the first two components two outlying points below the bottom of the plot, which were from Cluster 4, are not shown, for easier reading of the rest of the plot. Essentially Cluster 4 consists of outliers, so that it plots separately but not coherently. Cluster 6 is a small group that plots coherently on both plots and is 'extreme' relative to other clusters. A stray case apart, Cluster 1 plots separately on the plot for the first and third components, and much the same can be said for Cluster 7 on the first two components. Cluster 5 is rather less compact than either cluster analysis might suggest. It can be largely separated from other clusters, though not perfectly. It does, however, separate out on a plot of the third

[1] In practice exploratory analyses would be carried out before a cluster analysis; PCA can be used for this. It may reveal outliers that one could consider omitting from the cluster analysis, or obvious clusters that can be separated out before undertaking further analysis.

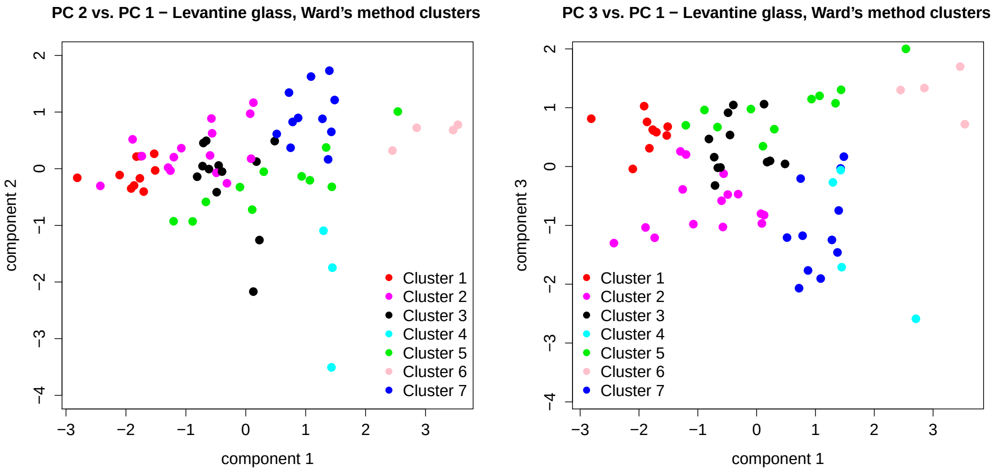

Figure 10.7: *Principal components plots for the standardized Levantine glass compositional data, labeled after the Ward's method clustering. Two outlying values from Cluster 4 are not shown in the left-hand plot.*

and fourth components (not shown).

This leaves Clusters 3 and 6 which were not especially well-separated on the average-link dendrogram, the same being true for the plot on the first two components. One case from each cluster apart they separate out on a plot of the first and third components, though they are contiguous. The plot on the first two components suggests that two cases from Cluster 3 are fairly clear outliers.

Overall the PCA suggests that the clustering produced by Ward's method is acceptable, provided one examines more than the first two components. The average-link cluster analysis, while confirming much of what can be inferred from the Ward's method analysis, fails to distinguish between two Ward's method clusters that the PCA shows can be distinguished. Both methods of cluster analysis provide little information on the coherence or otherwise of clusters. Ward's method is of little use for detecting outliers; average-link is much more satisfactory for this. The overall message is that a combination of both methods of cluster analysis, allied to checking using PCA, is a much better way of interrogating the data than relying on a single method of cluster analysis (which many publications give the impression of having done so).

It remains to ask if the sites can be distinguished. There is an imbalance between the sample sizes of 53 and 14. They are not readily separated; analyses are not shown, but the best that can be done is with the second and third PCs which separate out 9/14 of the smaller sample. This can be examined using PCA without recourse to cluster analysis which can, however, be useful for identifying sub-groups within site assemblages if sites plot separately. The small Cluster 6 comes from the site with the smaller sample size.

10.3 Ward's method and model-based methods

10.3.1 Ward's method

Ward's method is an exception to the generalization that most commonly used methods of cluster analysis in archaeology are just grouping algorithms with no firm basis in statistical theory. These methods have been widely used because they have seemed sensible to the people who devised them, and have found favor with practitioners. Statisticians have been less impressed (see Cormack, 1971, for an early and damning review from a statisticians perspective) and this has led more recently to the development of *model-based* clustering methods. Ward's method is discussed in further detail here, partly to introduce some of the ideas used in model-based and other methods.

In contrast to the other linkage methods, Ward's method attempts to optimize an explicit objective function, S_G. All the agglomerative methods discussed suffer from the drawback that once a merge is made it cannot be undone. Ward's method, as usually applied, is no exception and the word 'attempts' was used above because often S_G will not be optimized. That is, given any specific partition into G clusters produced by Ward's method, it may be possible to improve S_G by relocating cases between clusters. This is the basis of *k-means* methodology (Section 10.3.3).

Ward's method was popular in the 1970s and 80s, partly because it was the default in CLUSTAN, one of the earliest software packages designed specifically for cluster analysis. Applications of the method often use squared Euclidean distance as a dissimilarity measure. This distance forms the basis of measuring the variability *within* a cluster. At any given stage of clustering let G be the number of clusters. For a single case, i, and single cluster the overall closeness to the cluster centroid, $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p)$, can be measured by

$$\sum_j (y_{ij} - \bar{y}_j)^2$$

where summation is over the p variables. Summing this over the n_g cases in cluster g gives S_g, a measure of the 'compactness' of the cluster, and summing over g gives S_G a measure of the compactness of the clustering[2]. Any merge in the clustering process will increase S_G and the merge is chosen for which this increase is least.

[2] For a singleton cluster (i.e. consisting of a single case) which predominate in the early stages of clustering, $n_g = 1$ and $S_g = 0$. It is easy to see that merging two non-identical cases will increase S_G and this is generally true regardless of cluster size.

10.3.2 Model-based clustering

It has already been noted that Wards method can suggest clusters quite clearly, even when none exist. This behaviour can be understood by viewing Ward's method as a special case of a *model-based* method. Specifically, Ward's method will tend to produce (hyper-)spherical clusters of the same size. It can be shown to be an 'optimal' method if the assumption that clusters have a spherical normal distribution of equal size is correct. The method will tend to impose this kind of structure on clusters even if the assumption is not true.

Model-based methods were developed partly in response to the lack of theoretical justification for more heuristic methods. Such models depend on assumptions; it is unfortunate that archaeological data rarely satisfy the assumptions for Ward's method to be optimal, but the theory also explains the typical appearance of a Ward's method dendrogram.

Only a very brief account of model-based methods is attempted here – the mathematics is beyond the level of these notes. Banfield and Raftery (1993) provide a technical account that led on to the development of a package in `R`, `mclust`, to implement some of the methods; Papageorgiou *et al.* (2001) provide a detailed archaeological application, but the methodology has not been much used. Some possible reasons for this are discussed after describing the ideas involved.

What follows is less general than would be possible but covers the most common uses. Assume that clusters are (hyper)-ellipsoidal in p-dimensional space; a special case is when the clusters are (hyper)-spherical. Assume the sample is from K sub-populations and that the data for a single cluster, k, are sampled from a multivariate normal distribution. In principle the size and orientation of the clusters can vary. Given these assumptions it is possible to define a probabilty density function for the data, and parameters can be estimated using the method of maximum-likelihood; that is, an objective function is optimized but the optimization depends on the assumptions just listed. The parameters are those associated with the underlying probability densities and cluster labels, the estimation of which is the object of the exercise.

In its most general form the model is rarely used, as the number of parameters allied to the size of the data make estimation impractical. Banfield and Raftery (1993) simplify matters by developing models that impose constraints on the size, orientation and shape of the clusters. Thus, clusters may be constrained to have the same orientation, but allowed to vary in size, or *vice-versa*. Note that all this involves moving away from the theoretically preferred model. The most extreme simplification assumes that clusters are spherical and of the same size (orientation is irrelevant) and this leads to what is essentially Ward's method.

Thus Ward's method approximates the solution to a model that *assumes* that clusters have identical multivariate normal distributions other than their variation in location. This results in a method that 'looks for', and tends to find, spherical clusters of the same size, resulting in a typical dendrogram having the appearance illustrated in Figures 10.4. The issue of choosing K remains. Banfield and Raftery (1993) develop an approximate Bayesian method, rather complicated mathematically, that involves investigation of a range of values for K. In passing it can be noted that fully-fledged Bayesian approaches to clustering have been developed that are model-based and even more complex (Buck *et al.*, 1996). They have been little used in archaeology and such papers I have seen mostly use data for illustrative purposes where the cluster structure is obvious.

That model-based methods, other than in their simplest form, are rarely used is (apart from mathematical complexity) possibly because data sets are often too small or of too high a dimension to exploit the power of the methodology and/or because the structure of the data invalidates the assumptions of the methods. One area of application where model-based methodology is more practical is in spatial clustering, where the (usually) two-dimensional nature of the data (i.e. its low dimensionality) allows the use of more complex models.

Having said all this, Ward's method and other, *ad-hoc*, methods have been widely used and this is likely to continue. Ward's method is interesting in that it only approximates an ideal solution, and it may be possible to improve on this. This leads into the idea of k-means clustering.

10.3.3 K-means clustering

At its simplest the idea behind k-means clustering is straightforward. Newer and more complex methods that extend the ideas have been developed (e.g., Hastie *et al.* 2009). Attention here is confined to Ward's method. The basic idea is to take an initial starting position for the group centroids, either randomly chosen or from an initial Ward's method analysis (involving a choice of G), and reallocate cases between clusters until the optimum is, hopefully, attained.

For the seven-cluster Ward's method solution we follow Venables and Ripley (2002) and take the centroids of each cluster as a starting point. The distance from each case to each centroid can be calculated along with the criterion to be optimized, and cases can be reallocated if this improves the optimization. Table 10.1 compares the clustering from the original analysis with that resulting after reallocation.

Summing down the diagonal. 59/67 (88%) of cases are allocated to the same cluster as produced by Ward's method with three clusters remaining unaltered. Of the reallocated cases 5/8 were originally in Cluster 2; this result perhaps is not surprising given the earlier evidence from the average-link analysis and PCA. This kind of analysis is less common in the archaeological literature than one might expect; a possible reason is that such applications as exist often show very little

Ward' method	K-means clusters						
Cluster	1	2	3	4	5	6	7
1	10	0	0	0	0	0	0
2	2	10	0	0	0	0	3
3	0	0	11	0	1	0	0
4	0	0	0	3	1	0	0
5	0	0	0	0	10	1	0
6	0	0	0	0	0	4	0
7	0	0	0	0	0	0	11

Table 10.1: *A comparison of the clusterings obtained by Ward's method and subsequent reallocation using the k-means algorithm from the* `kmeans` *function in* `R`*.*

difference, if any, in the clusterings obtained. The results are not readily presented in a simple form such as the dendrogram.

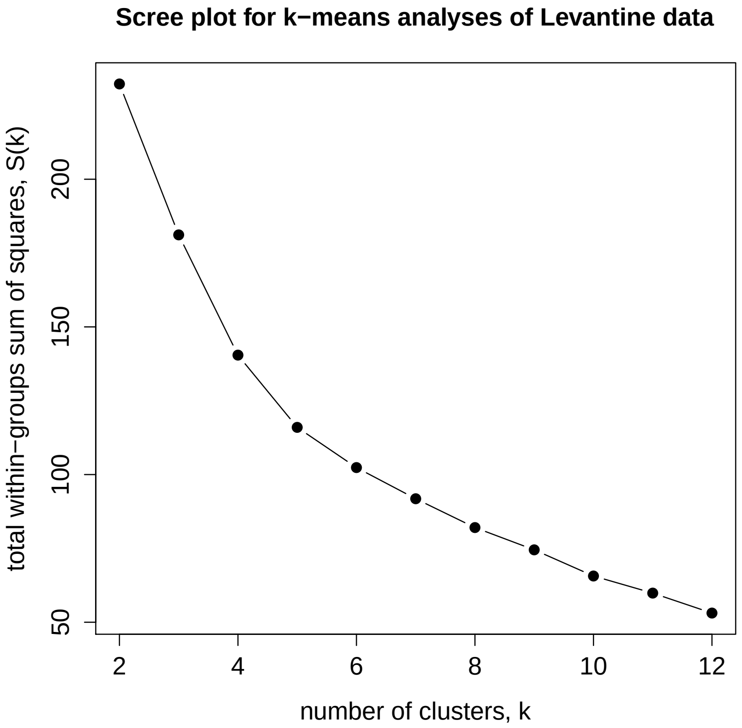

Figure 10.8: *A scree plot of $S(k)$ against k for the Levantine data k-means analysis.*

The question 'what is the correct number of clusters' remains; it is not necessarily an easy question to answer. One approach that has been suggested is to look at a scree plot of the total within sums of squares for the clusters against the number of clusters, or $S(k)$ against k, where $S(k)$ is the criterion that k-means optimizes for k clusters (Section 10.3). The idea then is to look for a clear 'elbow' that suggests the appropriate number of clusters. A scree plot for the Levantine data is shown in Figure 10.8.

The idea is simple but, as with other uses of the scree plot (for example, determining the appropriate number of components in a PCA in Section 7.4), it is frequently of little use in practice. This is because a clear elbow is often not apparent. This is the case for Figure 10.8. It is inevitable that $S(k)$ will decrease as k increases. In the figure the decay is greater up to $k = 5$ compared to larger values of k where the decrease is almost linear. There is not a clear elbow; it could be argued that a value of $k = 6$ seems appropriate, but the graphical evidence is not especially compelling.

10.3.4 Fuzzy clustering

Levantine glass compositions

The methods discussed so far produce what are called *hard clusters* where each case is assigned to a single cluster. In *fuzzy clustering* cluster membership is distributed across all clusters. Baxter (2009) discusses the ideas involved with some technical detail that is not repeated here. Implementation in `R` is straightforward and the function used here, `cmeans` from the `e1071` package, is almost identical in structure to the `kmeans` function. An alternative implementation, not explored here, is the function `fanny` from the `cluster` package.

The one difference between the command line for `cmeans` and `kmeans` is the inclusion of a *fuzzification factor* for which the default in `cmeans` is the argument `m = 2`. This controls the 'crispness' of the clustering; as `m` approaches 1 a hard clustering is obtained; as it becomes large a totally fuzzy clustering results. The use of `m = 2` is arbitrary but claimed to work in a generally satisfactory way. Baxter (2009) provides examples where values of `m` less than 2 were judged to produce better results. Output from the two functions differs in various ways, the most important for present purposes being a table of membership values.

An illustration is provided in Table 10.2 for a subset of the Levantine data. Cluster 2 from the original Ward's method analysis contained 15 cases; 5/15 cases were reallocated using k-means analysis. Fuzzy clustering provides further insight into this and Table 10.2 shows how c-means distributes the membership of this cluster across the seven suggested by the Ward's method analysis.

The c-means analysis has the highest membership in Cluster 2 for 11/15 cases, though not always dramatically so. This is fairly similar to what is suggested by the k-means analysis but with values of the membership for 10 of these 11 cases below 50 it suggests, as might be inferred from earlier analyses, that the clustering is not a very crisp one. Of the four cases where Cluster 2 does not have the highest

	Cluster						
	1	2	3	4	5	6	7
Case	Membership						
2	23	48	9	8	1	10	2
3	21	33	11	22	3	8	3
5	17	41	22	8	2	7	2
6	22	42	10	8	2	14	2
15	26	49	7	8	1	8	2
25	32	22	4	32	1	8	1
29	38	17	3	36	1	4	1
34	22	32	11	23	3	7	2
35	21	48	10	12	2	6	2
43	26	43	5	17	1	6	1
45	3	94	1	1	0	1	0
46	18	38	27	6	2	8	2
63	19	26	12	21	6	12	4
66	12	26	47	5	1	7	2
67	12	23	48	5	2	8	3

Table 10.2: *Results from a c-means clustering of the Levantine data showing membership values for cluster 2 from the Ward's method analysis with seven clusters.*

membership, two could plausibly be associated with Clusters 2 or 4 and the other two with Cluster 3.

Medieval glass compositions

For a different and simpler illustration the York medieval glass data are subjected to a similar analysis in Table 10.3. The analysis is simpler in the sense that the clustering is quite a crisp one with 25/27 cases clearly associated with a single cluster (taking, a little arbitrarily, 'clearly associated' to imply that the smallest membership value is 65 or greater).

The two exceptions to this observation are cases 20 and 22 which were the most outlying in Figure 10.1, particularly case 20. Case 22 can be seen from Table 10.3 to have a much higher membership value for Cluster 2 than 3 (60 compared to 29). The impression given by the plot for the first two PCs does not suggest this very clearly; further investigation reveals that the case is much closer to Cluster 2 if plots using the third component are examined[3]. It is clear, though, from the dendrogram and the PCA of the first two components, that this case is best regarded as an outlier. Case 20 is allocated to Cluster 3 in the original analysis; it has the highest membership value for this group but the difference compared to Cluster 2 is marginal (42 compared to 40). Inspection of plots using the third

[3] The plots of Figure 10.2 need to be labeled by case number to see this.

component show it is at some distance from the rest of Cluster 3, so the conclusion is that this is also a clear outlier.

Id.	Cluster		
	1	2	3
	Membership		
1	100	0	0
2	98	1	1
3	1	98	1
4	98	1	1
5	95	2	2
6	92	4	5
7	8	7	85
8	4	3	93
9	99	1	1
10	1	97	2
11	1	97	1
12	1	97	2
13	1	98	1
14	1	98	1
15	1	99	1
16	7	7	86
17	5	6	89
18	6	86	8
19	1	98	1
20	19	40	42
21	12	21	67
22	11	60	29
23	11	14	76
24	6	7	86
25	5	4	91
26	5	4	91
27	5	3	92

Table 10.3: *Results from a c-means clustering of the York medieval glass data showing membership values.*

.

10.4 Summary

Given the difficulties of applying more sophisticated model-based methods to typical archaeological data (with the possible exception of spatial clustering), it is not surprising that practitioners continue to rely largely on the older and more *ad hoc* methods. Presumably they are commonly found to give archaeologically interpretable results, with the caveat that results judged to be unsatisfactory usually never get published.

Some space has been devoted to Ward's method, but not because it is a preferred choice. It is good practice to compare more than one clustering method, and Ward's method is a useful starting point because it usually suggests clear, if possibly illusory, clusters. Given an initial and provisional identification the results can then be compared with those from other methods, as illustrated above.

Ward's method is also a useful peg on which to hang a discussion of model-based clustering methods. Although, for the reasons outlined, they have had limited archaeological use the underlying theory explains why Ward's method can be a potentially misleading method. Other methods lack a theoretical basis but can be subject to analogous or other problems. The method also provides an entrée into k-means and c-means clustering. Both have probably been underused by archaeologists (fuzzy clustering in particular). Both are very easily implemented in R. Fuzzy clustering is capable of producing a more nuanced view of the data than hard clusterings afford, and deserves more attention than it has received.

Books written specifically for archaeologists, that discuss cluster analysis, include, in ascending order of difficulty, Shennan (1997), Baxter (1994) and Baxter (2003). Not all the methods discussed in this chapter are covered in the last two books. General statistical texts, with a wider coverage, include Everitt *et al.* (2011), which is devoted to cluster analysis, accessible, and includes some archaeological examples. Good statistical texts on multivariate analysis, with treatments of CA, abound. They include Everitt and Dunn (2001), Krzanowski and Marriott (1995), and Seber (1984). This is in rough order of difficulty.

For 'newer' approaches to cluster analysis Everitt *et al.* (2011) is probably the most accessible statistical text for a non-statistical readership and includes material on mixture models and fuzzy cluster analysis. Hastie *et al.* (2009) covers several newer methods at a more advanced level. They are primarily concerned with methods of supervised pattern recognition (e.g., discriminant analysis, classification trees, neural networks), but have chapters on unsupervised pattern recognition that cover many more recent methods. Banfield and Raftery (1993) is a useful starting point for a statistical treatment of model-based clustering. Buck *et al.* (1996) is the best starting point for an exposition of the uses of Bayesian methods in archaeology, with references to, and applications of, cluster analysis to archaeological data; their pioneering work has not been emulated much.

10.5 R notes

Figure 10.2

```
pca <- prcomp(scale(york))
pairs(pca$x[,1:3], oma=c(4,4,6,12))
par(xpd=TRUE)
legend(0.85, 0.7)
```

This shows how, if you wish, placement of the pairs plot and a legend can be controlled. The oma argument in the call to the pairs function controls the placement of the plots by adjusting the outer margins; par(xpd = TRUE) clips the plot to the figure region with the effect of emsuring the legend is visible. See the help on par for details. The first two arguments in the call to legend specify its location. The arguments col, pch and cex are available for the pairs function but have been omitted, as they and other arguments have in legend.

Figures 10.3 to 10.6

The code for Figure 10.4 is given first since it is a prerequisite for Figures 10.5 and 10.6; Figure 10.3 is produced in a similar way. Normally this would be best written as a function, but it is convenient for expository purposes, to split the code that would be in the function into blocks. It is assumed that the standardized data for the Levantine glass compositions are held in Levantine. Version 3.1.2 of R was used for this; earlier work on these notes used R 2.13.1. For the most part it doesn't matter much, but in the later version, in hclust, it is necessary to select the version of Ward's method (there are now two possibilities); method = "ward.D", based on Euclidean distance, was used here.

A Ward's method analysis is undertaken first since the clusters identified determine the labeling used in later analyses. The first block of code produces and plots the dendrogram, from which it was decided that a seven-cluster solution woud be used as a starting point.

```
data <- Levantine
clus <- hclust(dist(scale(data)), method = "ward.D")
plot(clus)
```

Next the cutree function is used to associate each case with a cluster label, 1 to 7, and these, in turn, are used to define a character variable, Colour, that associates each cluster with a color label.

```
clus.id <- cutree(clus, h = 7)
Colour <- c(rep("black",dim(data)[1]))
Colour <- ifelse(clus.id == 2, "magenta", Colour)
Colour <- ifelse(clus.id == 3, "blue", Colour)
Colour <- ifelse(clus.id == 4, "red", Colour)
Colour <- ifelse(clus.id == 5, "cyan", Colour)
Colour <- ifelse(clus.id == 6, "green2", Colour)
Colour <- ifelse(clus.id == 7, "pink", Colour)
```

Folllowing this the as.dendrogram function converts the previously defined object clus to a dendrogram structure that can then be manipulated to achieve the desired effect using other functions. These include `order.dendrogram` which extracts the ordering (left to right) of case labels as they appear on the dendrogram.. The small function that follows then associates each of the re-ordered labels with its appropriate color as defined from the initial clustering. A plotting character (`pch = 16`) is also defined for the colored labels that are to be used subsequently in plotting the dendrogram. The main reason for doing this was so that dendrograms using clustering methods other than Ward's have the same colors associated with case labels, so the similarity to Ward's method results is more apparent[4].

```
hcd <- as.dendrogram(clus, hang = 0.1)
newindex <- order.dendrogram(hcd)
i <- 0
colLab <- function(n) {
    if (is.leaf(n)) {
     i <<- i + 1
     a <- attributes(n)
     attr(n, "label") <- NULL
     attr(n, "nodePar") <- c(a$nodePar, list(lab.col = "black",
     pch = 16, col = Colour[newindex[i]], cex = 1.2))
     attr(n, "frame.plot") <- TRUE
    }
    n
}
```

Having done all this the desired dendrogram can be plotted. The function `dendrapply` uses the function `colLab` defined above to color the symbols used for each label according to their cluster. The presentational arguments for `plot` have been omitted, as has the legend.

```
clusDendro = dendrapply(hcd, colLab)
plot(clusDendro)
```

To obtain single- and average-link clusterings replace `"ward.D"` wih `"s"` or `"a"` at the start of the above code, omitting the second block where the coloring is defined.

[4] I'd assumed that this would be a fairly simple thing to do, but based on what I found on the web apparently not. I eventually hit on what I used on a couple of sites, but did not make a note of the source. At least two packages, `dendroextras` and `dendextend`, have been written to facilitate dendrogram manipulation. I suspect the latter can do what I wanted, but how to do it is not especially obvious.

Figure 10.8

The kmeans function is used to obtain the total within-groups sums-of-squares for $k = 2, 3, \ldots, 12$ cluster solutions. The second argument centers = initial specifies centers of the initial clusters to use. These are the centroids of k clusters obtained from an initial Ward's method analysis using code based on Venables and Ripley (2002: 318). It is simpler to use centers = k which will randomly sample k (distinct) rows from the data matrix as starting centroids. If this is done the appearance of the scree plot is at the mercy of the random selection, and the scree plot may vary if an analysis is repeated.

```
WithinSS <- NULL
k = 2
hh <- hclust(dist(Levantine), method = "ward.D")

while(k < 13) {
initial <- tapply(Levantine,
           list(rep(cutree(hh, k), ncol(Levantine)), col(Levantine)),
           mean)

km <- kmeans(Levantine, centers = initial)
WithinSS <- c(WithinSS, km$tot.withinss)
k <- k + 1
}
 plot(2:12, WithinSS, type = "b")
```

Table 10.2

This proceeds much as the code for the k-means analysis above. In the call to cmeans the argument m = 2 is the (default) 'fuzzification' factor, and can be varied. The object cmmembership can be printed and edited to get Table 10.2.

```
hh <- hclust(dist(Levantine), method = "ward.D")
cluslev.id <- cutree(hh, 7)

initial <- tapply(Levantine,
           list(rep(cutree(hh,7), ncol(Levantine)),
           col(Levantine)), mean)

cm <- cmeans(Levantine, initial, m = 2)
cmmembership <- 100 * round(cm$membership[cluslev.id == 2,], 2)
```

Chapter 11

Discrimination and classification

11.1 Introduction

Principal component, correspondence and cluster analysis are examples of *unsupervised learning* methods. Discriminant analysis is, by contrast, an example of a *supervised learning* method; only linear discriminant analysis (LDA) is considered in any detail here. Unsupervised learning methods are geared towards the discovery of structure in data, often in the form of distinct groups. Information may exist about suspected grouping, but this does not feed into the analysis except that it may be used for labeling graphs to aid interpretation (Chapters 7, 9, 10). By contrast, supervised learning methods include information on (suspected) groups in the data which underpins, and is incorporated into, the mathematics of the methods. The aims of analysis may be to confirm that the groups are genuinely distinct; to display group differences graphically; to identify variables that best discriminate betwen groups, or to allocate cases not in the analysis to an appropriate group. This last aim, often called 'classification', motivates much of the more recent methodological development (Hastie *et al.*, 2009).

Mathematically, PCA, CA and LDA, can be viewed in terms of their increasing order of complexity related to the measure of distance used in analysis. Euclidean distance is used in PCA, chi-squared distance distance in CA, and *Mahalanobis distance* in LDA. Mahalanobis distance is the most complex (Sections 11.2 and C.2.3).

The three methods have, in common, the derivation of linear combinations of the original variables, ordered by importance, the first two of which are used for display purposes. Provided $n > p$ (or $I > J$ for CA) there are p (or J) linear combinations that can be derived. In LDA, where G groups are assumed, $(G - 1)$ linear combinations can be defined, so that bivariate graphical display is not available when $G = 2$ and univariate display is needed (e.g., Figure 11.6). A brief methodological account of LDA is provided in (Section C.2.3).

11.2 Mahalanobis distance

Mahalanobis distance (MD) has several uses in archaeology other than for discrimination. Mathematical details are provided in Baxter and Buck (2000), Baxter (2003: 69–72) and Section C.2.3. An important feature is that MD takes account of group structure, in particular allowing for the possibility that groups may have an ellipsoidal shape.

11.2.1 MD and confidence ellipsoids

To illustrate, Figure 11.1 reproduces, with enhancement, Figure 6.1 from Baxter (2003) that was based on the analysis of lead isotope ratio data.

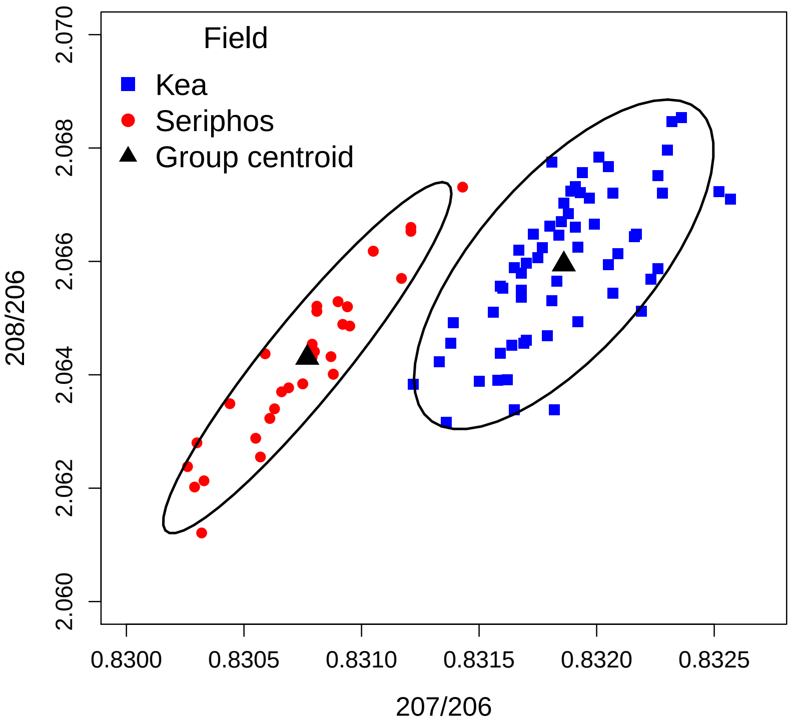

Figure 11.1: *90% confidence ellipsoids for the Kea and Seriphos lead isotope fields using the* $^{208}Pb/^{206}Pb$ *and* $^{207}Pb/^{206}Pb$ *ratios.*

Measures on ore bodies (fields), mined in antiquity for copper, can be characterized by three lead isotope ratios ^{208}Pb/^{206}Pb, ^{207}Pb/^{206}Pb and ^{206}Pb/^{204}Pb. The idea is that different lead isotope fields can be distinguished on the basis of these ratios. Apart from looking at all pairwise possible plots (e.g., Figure 11.3) they have sometimes been embellished with confidence ellipsoids, as shown. The boundary of an ellipsoid is determined by points equidistant from the centroid in terms of MD (Section C.2.3).

Ellipsoids can be used to delineate groups defined using archaeological criteria, as above, or identified using statistical methods (e.g., PCA or cluster analysis) plotted on pairs of PCs. Figures 6.4 and 6.5 illustrate this for the Pompeiian loomweight data of Tables B.3 and B.4 where this was contrasted with the use of convex hulls in Figure 6.5.

The rationale for using confidence ellipsoids is that the data are samples from populations where the true extent of the field extends beyond that observed. Convex hulls do not allow for this; confidence ellipsoids represent an attempt to estimate the true extent. A potential drawback of their use is that it needs to be assumed that the population has a (multivariate) normal distribution and this is sometimes obviously dubious. The construction of confidence ellipsoids, when the assumption of normality is valid, is discussed in Section C.2.3.

11.2.2 MD, outliers, and allocation to groups

For the Seriphos field there is a case, 33, at the upper extreme of the ellipsoid and lying just outside it. Visually it seems to belong with the Seriphos field, but also seems further away from the centroid of that field than that of Kea. This is confirmed if the Euclidean distance of the case to the two centroids is calculated, after standardization. The distances are 2.33 and 1.20 for Seriphos and Kea.

If MD is calculated the conclusions are reversed and conform more closely with the visual assessment of group assignment. The MDs to the centroids of Seriphos and Kea are 5.76 and 10.29; these are squared quantities so their square-roots of 2.40 and 3.21 may be compared more directly with Euclidean distance. The MD calculations depend on both the centroid of a group and its covariance matrix (Section C.2.3). If the latter is diagonal MD reduces to squared Euclidean distance. Where the data exhibit strong covariances/correlations MD and Euclidean distance calculations can produce rather different results, as shown, because MD but not Euclidean distance allows for the elliptical nature of the groups.

A complication is that MD calculations depend on estimates of the centroid and covariance matrix, the calculation of which is influenced by the case whose membership is being assessed. An obvious idea here is to base calculations on what have been called *leave-one-out* (LOO) methods, where calculations omit the case of interest. Applying this idea to case 33 from the Seriphos field, the square-rooted LOO MD value is 2.66 which exceeds the original value of 2.40, as expected, but is still closer than the distance to the Kea centroid of 3.21 (calculations of which are unaffected). In R, and in the context of LDA, the term leave-one-out *cross-validation* is used.

11.3 Linear discriminant analysis – examples

11.3.1 Lead isotope-ratio data – three groups

Data for three lead isotope fields are given in Table B.18 for three ratios. With $G = 3$ two linear discriminant functions are defined which lends itself to the display of results in two-dimensional plots. Figure 11.2 contrasts the results of applying PCA and LDA to the three groups using all the ratio data. Remember that LDA uses the information about groups to maximize the separation between them.

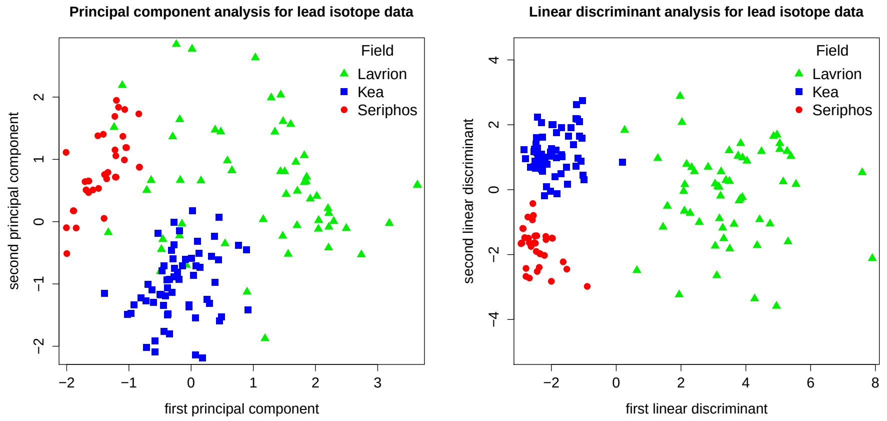

Figure 11.2: *Principal component and linear discriminant analyses of the lead isotope-ratio data of Table B.18 using all three fields.*

From Figure 11.1 it was seen that a plot based on two ratios is more than adequate to establish the fact that the Seriphos and Kea fields are distinct. This is also a feature of the PCA and DA plots. The presence of the Lavrion field complicates matters; it is widely spread out on the PCA plot, showing some overlap with the Seriphos field and more with that for Kea. The LDA, by contrast successfully separates the three fields apart from one case each for the Kea and Lavrion fields whose group membership is in doubt. This is a further clear illustration that PCA and LDA can produce noticeably different results.

In practice, preliminary data inspection is sensible. A pairs plot of the ratios is shown in Figure 11.3. It is clear that all three fields can be separated using just the first two ratios; plots involving the third ratio confuse matters. This is reflected in Figure 11.2 where PCA fails to show the field separation evident from the simple

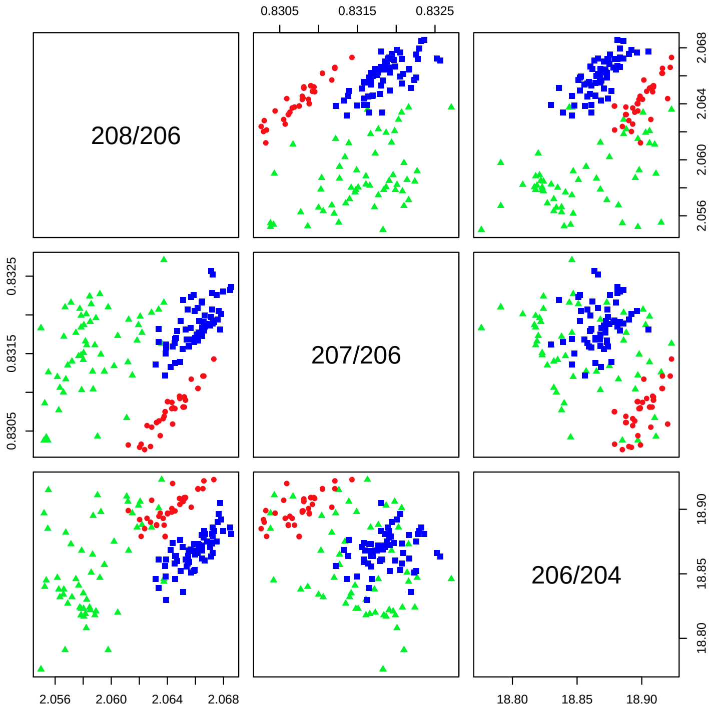

Figure 11.3: *A pairs plot for the lead isotope ratio data. Labeling is as in Figure 11.2, green triangles for Lavrion, blue squares for Kea, red circles for Seriphos.*

analyses. This points to the potential importance of variable selection when using multivariate methods; the problem is that 'non-structure-carrying' variables can obscure the lessons to be learned from variables that are revelatory of structure.

This will not be pursued in detail here; some of the issues are discussed in Baxter (1994b). Examples there show that variable selection can improve the success of allocation, though selection of the best discriminating variables is not guaranteed. With two groups LDA can be formulated in terms of linear regression, with methods based on stepwise selection procedures widely available in software packages. Analogous stepwise procedures for LDA with more than two groups are available in software packages such as SPSS and SAS. Stepwise selection methods have been subjected to considerable criticism (failure to find 'optimal' solutions; lack of generalizability; etc.). More modern methods are discussed in Chapter 3 of Hastie *et al.* (2009) but have had few archaeological applications.

11.3.2 Neolithic pot dimensions

For a second example data from Table 1 of Madsen (1988b: 18), reproduced in Tables B.19 and B.20, are used. They consist of 16 measurements on eight profile points from pottery vessels from the the Early and earlier Middle Neolithic TRB culture in Denmark.

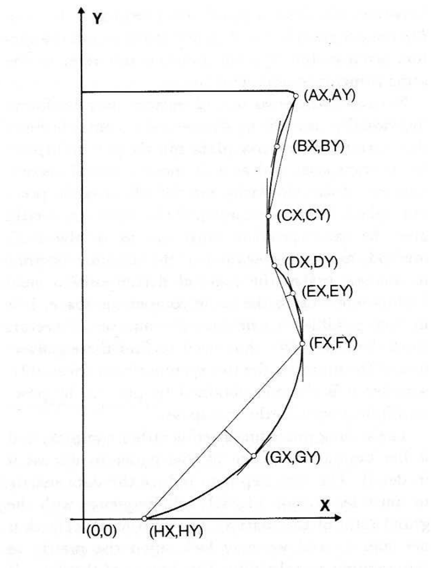

Figure 11.4: *Measurement points of Danish Neolithic pot profiles. (Source: Madsen 1988b: 16, after E.K.Nielsen.)*

Each profile point is represented by vertical and horizontal components, so there are 16 measurements in total (see Figure 11.4). The plots were classified into three vessel forms, funnel beakers, bowls and flasks, with sample sizes of 81, 21 and 16. Classification was based on archaeological criteria and it was of interest to see if this could be reproduced using multivariate methods. It is obvious from simple bivariate plots (e.g., Madsen 1988b: 17) that flasks are dimensionally distinct, and sensible to separate these out at an initial stage of analysis, though they are retained for illustrative purposes in some of the examples to follow.

Madsen (1988b) concentrates on analyses of the funnel-beakers using PCA, whereas we use the data to illustrate aspects of LDA. Initial analysis was dominated by a size component – undesirable in the context of the aims of the analysis. Madsen discusses various ways of removing this and we follow him in scaling vertical measurements to pot height and horizontal measurements to rim width.

Three groups

Figure 11.5 contrasts PCA and LDA analyses of the data for the three types[1]. Both analyses separate out the flasks from the other forms, the separation being much clearer for LDA. Neither analysis separates out the other two forms, though LDA is a bit better. Both analyses suggest a small but distinctive group of six or seven bowls.

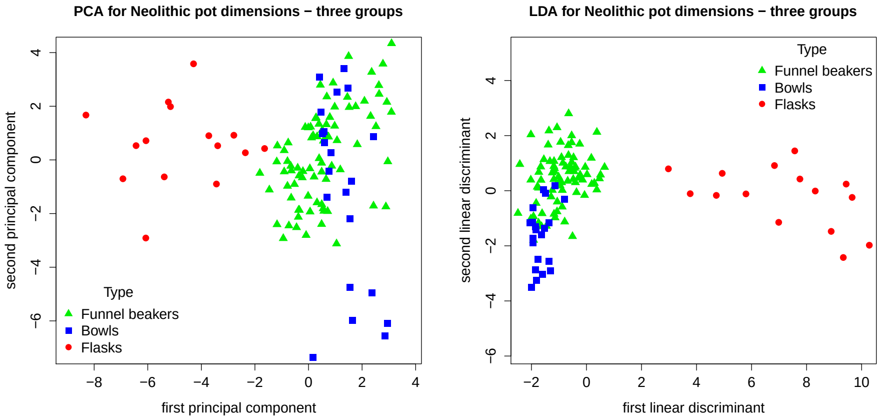

Figure 11.5: *Principal component and linear discriminant analyses of the Neolithic pot dimensions of Tables B.19 and B.20 using all three vessel types.*

Two groups

It is often easier to discern pattern using a small number of groups for any one analysis, and Figure 11.6 repeats the previous analysis after omitting the flasks. There is only one discriminant function, so graphical display must be accomplished using methods other than bivariate plots. Other options than that used, such as boxplots, are available but don't show the individual data points. The message is the same as that derived from the three-group analysis, namely that with the measurements used funnel beakers and bowls are not well discriminated. Using LOO classification 85/102 (83%) of cases are successfully classified. The *resubstitution* method,where statistics are influenced by the case to be classified, produces over-optimistic assessments.

A more informative way of assessing success, though more time-consuming to digest, is to examine the posterior probabilities of group membership, using leave-one-out calculations. These are obtained from the `lda` function in `R` with the

[1] Case 112 (Id. 237 in Table B.20) was found to be a clear outlier in preliminary analysis and omitted from this analysis.

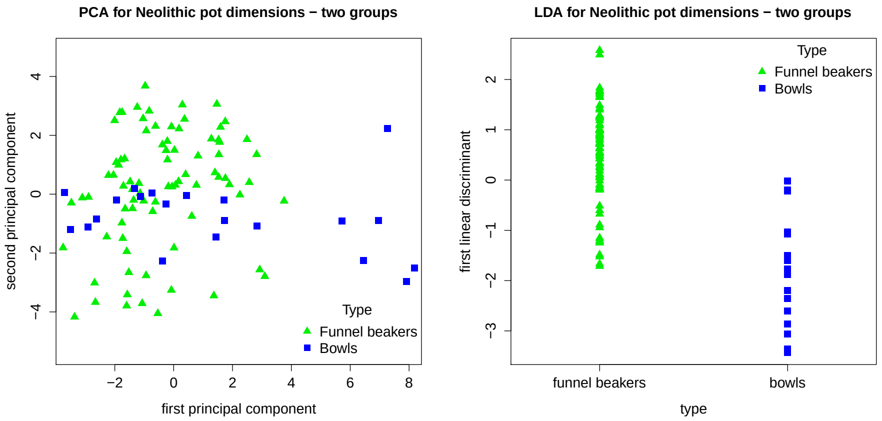

Figure 11.6: *Principal component and linear discriminant analyses of the Neolithic pot dimensions of Tables B.19 and B.20 using the first two types, beakers and bowls.*

argument `CV = TRUE`. The outcome for a subset of the funnel-beaker and bowl data is shown in Table 11.1.

If a ‘hard’ classification is needed a case would be assigned to the type with highest probability. With the exception of vessels 307, 308, 311 and 329 the funnel-beakers would mostly be convincingly classified as such, with probabilities close to 1. Four beakers are classified as bowls, beaker 307 with a probability of 0.45 of being a beaker and 0.55 of being a bowl only marginally so; the others are more convincing. Of the six bowls shown, only two are classified as such, the remaining four being classified as beakers with high or quite high probabilities.

A limitation of this kind of analysis is that results are presented in terms of relative probabilities and assume that cases must belong to a group in the analysis. This need not be the case; for example outliers may occur, or classes not recognized in the analysis may be represented. In these circumstances a case will be assigned to one of the assumed groups, possibly with high probability, even though the reality is that it belongs to none.

Often this kind of issue will be recognized in preliminary data analysis, or from the LDA itself. More formal methods exist; using normality assumptions MDs can be converted to absolute probabilities that allow a more realistic assessment of likely group membership. This is discussed in Section C.2.3.

			Predictions	
Id.	Actual	Predicted	Beaker	Bowl
⋮	⋮	⋮	⋮	⋮
307	Beaker	Bowl	0.45	0.55
308	Beaker	Bowl	0.25	0.75
311	Beaker	Bowl	0.14	0.86
312	Beaker	Beaker	0.98	0.02
320	Beaker	Beaker	0.93	0.07
321	Beaker	Beaker	0.89	0.11
323	Beaker	Beaker	0.98	0.02
324	Beaker	Beaker	1	0
325	Beaker	Beaker	0.92	0.08
326	Beaker	Beaker	0.95	0.05
327	Beaker	Beaker	0.99	0.01
328	Beaker	Beaker	0.99	0.01
329	Beaker	Bowl	0.33	0.67
336	Beaker	Beaker	0.87	0.13
3	Bowl	Bowl	0.03	0.97
33	Bowl	Beaker	0.77	0.23
59	Bowl	Beaker	0.78	0.22
60	Bowl	Beaker	0.97	0.03
61	Bowl	Bowl	0.01	0.99
131	Bowl	Beaker	0.98	0.02
⋮	⋮	⋮	⋮	⋮

Table 11.1: *Cross-validated estimates of the relative probabilities of belonging to a group, after a two-group LDA for a subset of the Danish Neolithic pot data.*

11.3.3 Practicalities

The normality assumption

The examples touch on a number of practical issues. Among them is the question of normality – it is sometimes incorrectly asserted that LDA requires the assumption that groups have a multivariate normal distribution. In fact Fisher's (1936) original derivation of LDA, subsequently developed by Rao (1948), made no such assumption. A sensible measure of group separation is defined and optimized mathematically to determine the discriminant functions. For descriptive, graphical analysis normality is not assumed. If normality can be assumed then LDA has some optimal properties, but the lack of normality does not necessarily compromise its practical utility. Normality does need to be assumed for probability calculations of the kind described in the previous section and Section C.2.3 but, as some of the examples suggest, it can be questionable.

The equal covariance assumption – quadratic discriminant analysis

Applications so far have assumed that groups are sampled from separate populations having equal covariance matrices, allowing their estimates to be pooled and leading to LDA. If the assumption is relaxed this gives rise to quadratic discriminant analysis (QDA) (Venables and Ripley, 2002: 333–334) which, as the name suggests, leads to quadratic boundaries between groups. A practical drawback of QDA is the need to estimate separate covariance matrices within groups, which requires more parameters and may be problematic with a reasonable number of variables to deal with, so demands more data than LDA.

11.3.4 Example - Steatite compositions

The extended example in this section illustrates the kind of output produced by R. It is based on compositional data for samples of steatite (soapstone) analyzed by Truncer *et al.* (1998) in order to see if the method of chemical analysis used succeeded in distinguishing between quarry sources more effectively than had previously been the case. The data are given in Tables B.21 to B.23.

There are six quarry sources used here with sample sizes between 24 and 31; there was a considerable number of measurements below the level of detection and only 6 of the 17 variables that were measured, for which complete information was available, are used. The analysis in the original paper is not emulated, the data are useful for illustrating aspects of LDA in this section, and classification trees in Section 11.4.2. Data are logarithmically transformed, to base 10, before analysis.

The transformation `St.log <- log10(steatite)` is applied to the 159 × 6 data matrix, `steatite`. The vector of quarry identifiers is named `St.type`. With this in place

```
library(MASS)
St.lda <- lda(St.log, St.type, CV = FALSE)
St.ld <- predict(St.lda)$x
```

sets up an object `St.lda` that can be used for interrogating the analysis, and `St.ld` which is used for plotting purposes. The obvious thing to do, and one of the main purposes of the exercise, is to look at the data graphically using a plot based on the first two functions, as in Figure 11.7[2]. This shows that discrimination is far from perfect. Visually Lawrenceville seems most distinct, but there is overlap between the predictions for all the quarries, particularly Susquehanna.

This can be investigated more closely using the 'confusion' table, which shows the predictions for each quarry. In the `lda` function the argument `CV = FALSE`

[2] This is obtained using `eqscplot(St.ld[,1], St.ld[,2])` where the arguments governing the labeling have been omitted.

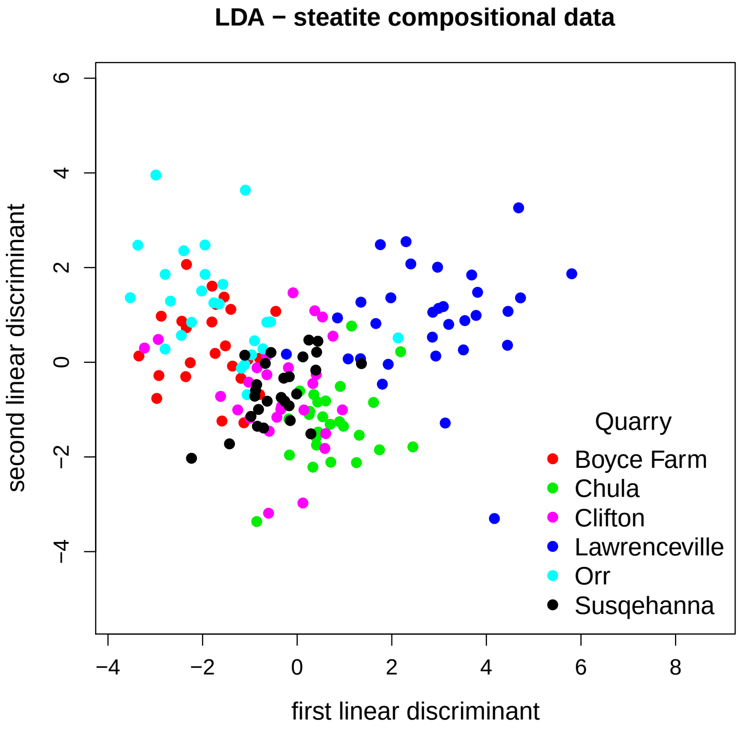

Figure 11.7: *LDA of the steatite compositional data of Tables B.21 to B.23.*

was used. This is the default; if `CV = TRUE` is used LOO cross-validation is implemented. This provides output in a different form; in particular `St.lda$class` provides the LOO predictions and this can be cross-tabulated with `St.type` to get the 'confusion' table (Table 11.2). The overall 'success' of classification is 113/159 = 71%, where 113 is the sum of the diagonal cells. If `St.lda$posterior` is accessed this provides posterior probabilities for the classes, of the kind shown in Table 11.1. The success of classification is summarized in percentage terms for each quarry in the third column of Table 11.3. This suggests that, for example, the classification for Boyce Farm and Orr is not very impressive.

In fact these assessments may be rather pessimistic. In attempting to discriminate between all six groups simultaneously the discriminant functions can be thought of as 'averaging' over all the groups, producing results that fail to show that good discrimination between pairs of groups may be possible. An alternative approach is to undertake LDAs on pairs of quarries. This is illustrated in Table 11.3 where the LOO and resubstitution success rates for each possible pair are given.

For the pairs the success rates are mostly noticeably greater than the global

Quarry	Predicted quarry						
	B	Ch	Cl	L	O	S	n
B	13	1	2	0	4	4	24
Ch	0	23	0	2	0	1	26
Cl	2	0	19	0	0	5	26
L	0	5	0	24	0	2	31
O	1	0	2	1	15	6	25
S	2	2	4	0	0	19	27

Table 11.2: *The 'confusion' table for LDA of the steatite data.*

			Pairwise comparisons					
Quarry	n	CV (%)	B	Ch	Cl	L	O	S
Boyce Farm	24	54		98	90	96	82	88
Chula	26	88	98		90	84	94	89
Clifton	26	73	92	94		93	88	85
Lawrenceville	31	77	100	90	95		96	90
Orr	25	60	86	98	92	96		90
Susquehanna	27	70	90	96	87	93	94	

Table 11.3: *Classifications from LDAs of the steatite compositional data. The CV is the percentage correctly classified using LOO methodology. The second part of the table shows the success rate for pairwise comparisons; the upper triangle is for LOO calculations, the lower triangle for the resubstitution approach.*

rates and in excess of 90%. Although not pursued in detail here, this is because the variables are weighted rather differently in the discriminant functions for pairs, both from each other and from the global analysis.

11.4 Classification trees

11.4.1 Basic ideas

Classification trees have been used intermittently in archaeological applications (Baxter, 2003: 116–118). They are an attractive alternative supervised learning method for problems often tackled using (LDA), if sample sizes are large enough. Venables and Ripley (2002: 331) state that 'classical methods of multivariate analysis [including LDA] have largely been superseded by methods from pattern recognition', and classification trees are one such 'modern' alternative.

Many of these modern methods are computationally complex, rather 'black-boxy', and difficult for the non-specialist to understand. Classification trees, by contrast, are conceptually simple and result in economical and elegant displays of the results that – with some caveats to be entered – can be readily understood. Output is typically in the form of a tree diagram. The data are initially treated as a single group that is successively sub-divided until some stopping criterion is satisfied. The aim is to identify those variables that best separate the groups. This is best illustrated by example in Section 11.4.2. Technical aspects of the method are discussed in Section 11.5 with a further example in Section 11.6.

11.4.2 Example – Steatite compositions (continued)

The data on steatite compositions from Tables B.21 to B.23, used to illustrate LDA in Section 11.3.4, are used. Figure 11.8, the same as Figure 9.2 in Baxter (2003), shows the outcome of a classification tree analysis of the data.

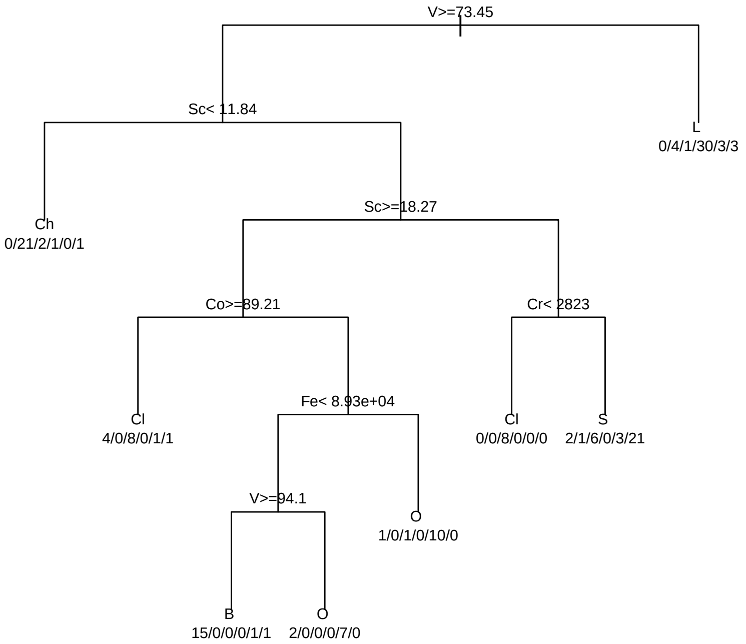

Figure 11.8: *A classification tree for the steatite compositional data.*

The starting point, the *root node*, based on all the data before group separation, is associated with a measure of *purity* (Section 11.5). The root node is split into two groups/nodes on the basis of one variable, in order to increase the purity of the tree by as much as possible. This (binary) splitting continues until some stopping criterion is met.

Using the variable V (Vanadium) cases having V $>$ 73.45 go to the left in the first split. It happens that the node to the right, according to the criterion used, is a *terminal node* and no further attempt is made to split it. The numbers below a terminal node show how many cases from each group are assigned to that node; 30/41 come from the Lawrenceville quarry which is the dominant group identified by the label L for the node. A *pure* terminal node is one where all the cases come from a single group, and the ideal is that all terminal nodes are pure.

To the left, the second split is based on values of Scandium (Sc) that are less than 11.844 and produces a second terminal node dominated by the Chula (Ch) quarry. The third split is also based on Scandium, this time using the higher values, Sc $>$ 18.2695, without producing any further terminal nodes. As well as the fact that a variable can be used more than once in the splitting process, note that there are more terminal nodes (eight) than there are groups in the data. The quarries Clifton (Cl) and Orr (O) each dominate two terminal nodes; this can indicate either that there are distinct sub-compositions within the groups, or that they are not compositionally well-defined.

Most of the terminal nodes are not pure, so that the classification is less than perfect. The classification success can be assessed at 78% which compares favorably with that for LDA, using leave-one-out classification, of 72%.

11.5 Methodology

The preceding section describes the bare outlines of the methodology, which is conceptually straightforward. Implementation is computationally intensive, requiring a number of choices to be made on the way. A brief, but more technical, discussion of further aspects is provided in this section.

If the sample of size n is partitioned into G classes, with n_i cases in class i, the root node is defined by G probabilities $(p_1, p_2, \ldots, p_G)$ where $p_i = n_i/n$. A node is labeled according to the most dominant class. The misclassification rate at a node is the number of cases not in the dominant class. For the root node call this R_0. Splitting occurs as discussed in the previous section and criteria for what constitutes a terminal node need to be defined. In the example no attempt has been made to split a node with fewer than 20 cases and a terminal node must contain at least 7 cases.

The *impurity* of a node can be defined in different ways; in the example the

Gini index

$$1 - \sum_i p_i^2$$

was used. For a pure node this takes the value zero, and the smaller the value of the index the purer the node. Assume continuous variables are used. Any one of these, X say, can be ordered as $(x_{(1)}\ x_{(2)}\ \ldots\ x_{(n)})$. For $i = 2, \ldots, n$, splits may be contemplated such that cases with values less than $x_{(i)}$ are separated out from cases with values greater than or equal to $x_{(i)}$. New nodes are defined by the two subsets of cases thus defined, and the impurity of these, and hence their average impurity, measured. All possible splits for all variables are examined and the split that most reduces the average impurity is chosen. Each new node is treated in a similar way and the process continues until terminal nodes are reached.

Clearly the appearance of the final tree depends on a variety of choices, including that of the smallest node for which a split is allowed and the minimum size of the terminal node. An understanding of this, and other aspects of the final appearance that can be controlled by the user, is aided by looking at the code used to obtain Figure 11.8.

```
library(rpart)
z.rp <- rpart(St.quarry ~ Co + Cr + Fe + Mn + Sc + V,
data = St.data, cp = .03, minsplit = 20, minbucket = 7)

plot(z.rp, uniform = T, margin = .05)
text(z.rp, use.n = T, cex = .7)
```

The `rpart` package needs to be loaded. In the arguments to the `rpart` function `minsplit` and `minbucket` specify the minimum node size that can be split, and the minimal terminal node size. The values given are the defaults; varying them will change the appearance of the tree. The argument `cp` is a *complexity parameter* (of which more below) for which the default is 0.01; reducing this will result in a larger tree relative to the default, and increasing it will produce a smaller tree. For more detailed information see the `R` help on `rpart.control`. The `plot` function prints the tree. For large trees and/or those with many groups obtaining a satisfactory appearance usually requires some experimentation. The last two lines of code use arguments that affect the appearance; see the `R` help on `plot.rpart` for more detail.

It is possible to get satisfactory results by 'playing around' with the arguments in the `rpart` function, but different analysts may arrive at different trees. A more 'principled' approach to determining tree size is available in the `rpart` package. If `cp` is too small the data may be 'over-fitted' and give an over-optimistic assessment of the success of the classification. The tree in Figure 11.8 originally had 11 terminal nodes using `cp = .001` rather than the eight terminal nodes using `cp = 0.03`.

Cost-complexity pruning can be used to reduce the size of the tree. Let $R(T)$ be the number of misclassifications in a tree of size T as measured by the number of terminal nodes, and let C be the cost-complexity parameter defined by `cp`. A cost-complexity measure of the form

$$R_C = R(T) + CTR_0$$

can be defined. As C increases, and leaves are pruned, T will decrease and $R(T)$ will increase. The degree of pruning is chosen to minimize R_C.

To choose C cross-validation is used. The data are split into 10 groups of roughly equal size. Nine of the groups are used to grow a tree and it is tested out on the remaining group to obtain a measure of the error involved. This can be done in 10 ways and the results averaged to get an estimate of the error and its standard deviation. Results may be summarized graphically, as in Figure 11.9 based on the unpruned tree which, after pruning, led to Figure 11.8, obtained using `plotcp(z.rp)`.

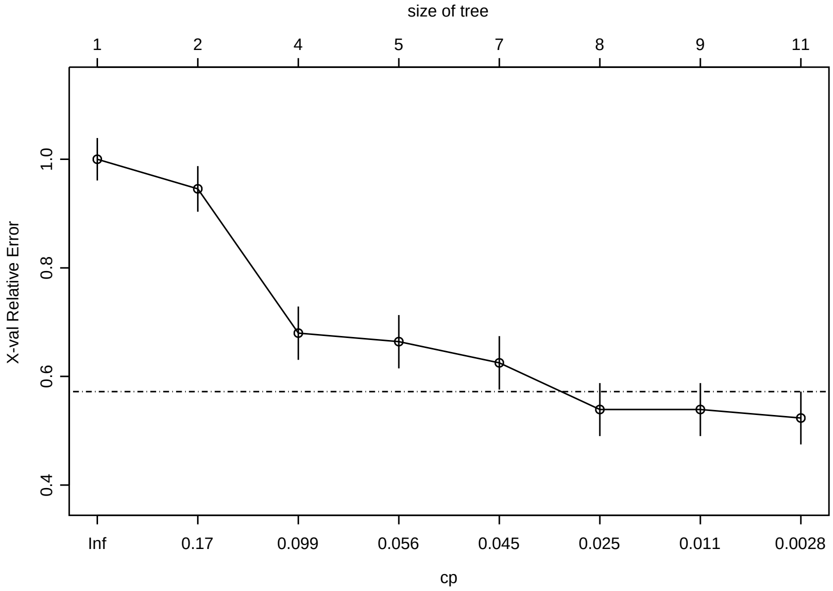

Figure 11.9: *A plot of error-rate against tree size as determined by* `cp`*, based on 10-fold cross-validation. It leads to a pruned tree of size 8 which is that shown in Figure 11.8.*

Plotted points show the mean number of errors across the 10 analyses divided by R_0 (i.e. the vertical axis is scaled to lie between 0 and 1), with 'one standard error' bars for each mean. The horizontal reference line is located at the mean error for the largest tree plus one standard error. The tree selected is the smallest one whose mean error lies below and within one standard error of the line. In this case this rule results in the unpruned tree of size 11 being pruned to one of size 8. Incidentally, the choice of `cp` in the original call to `rpart` can be thought of as controlling pruning of the tree that takes place as the tree is grown; larger values of `cp` have the effect of preventing the growth of branches likely to be 'lopped-off' when using the above selection procedure.

To the best of my knowledge this 'rule' is not justified theoretically. Therneau and Atkinson (1997: 13), the originators of the `rpart` package, observe that this method has proved good at screening out 'pure noise' variables. That is, the justification for its use is empirical. Classification trees can be viewed as an alternative to variable selection methods in LDA, with several potential advantages. One is the transparency of the method. Others are the fact that the method can handle missing data (illustrated in Baxter and Jackson, 2001) or mixed data, and that it is not necessary to worry about data transformation because of the monotonic relationship between raw and log-transformed data.

11.6 Example 2 - North Apulian pottery

This analysis is based on that shown (with presentational modifications) in Figure 5 of Gliozzo *et al.* (2013). Chemical compositional data were available for pottery from four late antique/early medieval sites (4th to early 7th centuries AD) in northern Apulia, Italy. The main interest lay in seeing if two of the sites, *Herdonia* and *Canusium*, produced pottery that was chemically distinguishable, but similar data from two other sites, *Posta Crusta* and San Giusto, had previously been investigated and the opportunity was taken to compare these with the newer data sets. Data on two kinds of pottery were available, coarse table wares and fine painted wares, and a 65 × 36 table of results for the latter type – split into groups defined by the four sites – is used here for further illustration. The data are similar in nature to the previous example but far more variables are involved (Tables B.24 and B.25).

A classification tree is shown in Figure 11.10. Sample sizes for some of the sites are quite small so splits of nodes with more than 10 cases, as opposed to the default of 20, were allowed. This had the effect of making it much clearer that the productions of the sites could largely be distinguished in terms of their chemistry.

What is worth noting about this is that only five splits are needed, involving five of the 36 variables, a considerable 'saving' in terms of economy of description and comprehension. There are six terminal nodes of which four are pure; two sites, *Herdonia* and San Giusto split into two mostly quite small groups which are, however, almost completely distinct from other sites. Of the 65 cases just 3 are misclassified, so the success of the classification is 95%.

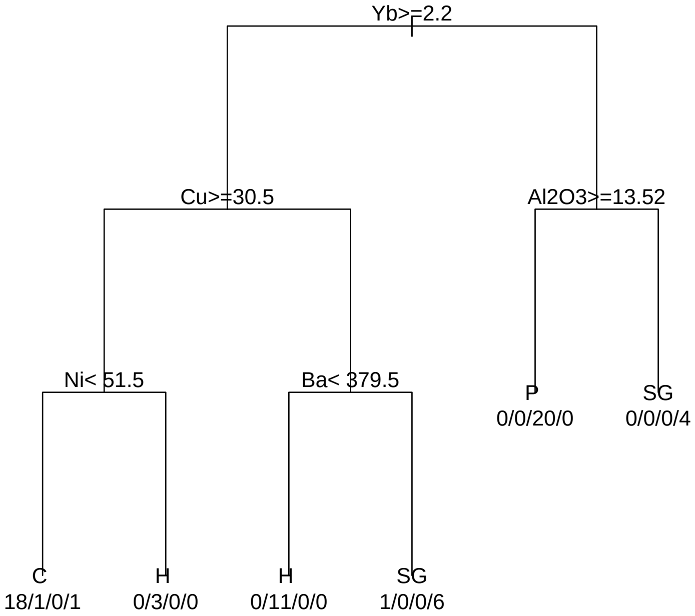

Figure 11.10: *A classification tree for the northern Apulian fine ware pottery data.*

Given the success of classification there is little point in comparing it with the results from pairwise comparisons. Results reported in Gliozzo *et al.* (2013) for the coarse ware data noted that the success rate for classification with all four sites in the analysis was 87%. For pairwise analyses the success rate varied between 90% and 96%, with only one or two variables needed to achieve this, except for the *Posta Crusta* and San Giusto comparison, which needed three. As with the fine wares, therefore, chemical separation between the sites for the coarse wares wss good; that in the latter case pairwise comparisons improved on the global analysis mirrors what was observed for the LDA of the steatite compositional data, and for analogous reasons.

11.7 R notes

Figure 11.1

The data sets Kea, Seriphos and Lavrion need to have been created; the last of these is not used for this figure, but is used in later examples.

```
library(car)
K <- Kea; S <- Seriphos; L <- Lavrion

plot(K[, 2], K[, 1], type = "n")
points(K[, 2], K[, 1]); points(S[, 2], S[, 1])

lines(dataEllipse(K[, 2], K[, 1], levels = c(0.9), plot.points = F,
      center.pch = 17, center.cex = 2))

lines(dataEllipse(S[, 2], S[, 1], levels = c(0.9), plot.points = F,
      col = "black", center.pch = 17, center.cex = 2))
```

If the dataEllipse function is used separately from the lines function it may be necessary to include other arguments to get the desired effect: see the help facility. The levels argument specifies that a 90% confidence ellipsoid is required; other arguments dictate the color, plotting character and size of the group centroids and are optional.

Figure 11.2

The three data sets are combined into LKS using the rbind function. The object SymLKS needs to be created and defines the plotting symbols used in the pch argument in eqsc but also doubles as the group identifier needed for the second argument of lda (presentational arguments are not shown). In the predict function the argument dim = 2 extracts two functions (the maximum possible with three groups). This allows a single argument, LKS.ld, to define the plotting positions; in general, if more than two functions are extracted, the two columns to be used for plotting need to be listed separately.

```
library(MASS)
LKS <- rbind(L, K, S)

LKS.lda <- lda(LKS, SymLKS, CV = F)
LKS.ld <- predict(LKS.lda, dim = 2)$x
eqscplot(LKS.ld)
```

```
LKS.pca <- prcomp(scale(LKS))$x
x1 <- LKS.pca[, 1]; x2 <- LKS.pca[, 2]
eqscplot(x1, x2)
```

Figure 11.5

The data of Tables B.19 and B.20, excluding `Id.` and `Type`, are held in `TRB.data`. They need to be transformed in the manner described in the text, as follows, with the result in `TRB.trans`.

```
z1 <- TRB.data[, c(1, 3, 5, 7, 9, 11, 13, 15)]
z1 <- z1/z1[, 1]
z2 <- TRB.data[, c(2, 4, 6, 8, 10, 12, 14, 16)]
z2 <- z2/z2[, 1]
TRB.trans <- cbind(z1[, -1], z2[, -1])
```

Nothing really new is illustrated in the PCA below. An obvious outlier, case 112, was identified in preliminary analysis, and is omitted as shown. This needs to be done, also, for the previously created colors (`ColTRB`) and plotting characters (`SymTRB`) and this is shown in the call to `eqsc` for emphasis; other presentational arguments and the legend are omitted.

```
TRB.pca <- prcomp(scale(TRB.trans[-112, ]))$x
x1 <- TRB.pca[, 1], x2 <- TRB.pca[, 2]
eqscplot(x1, x2, col = ColTRB[-112], pch = SymTRB[-112])
```

The commands for LDA are shown for completeness, but introduce nothing new.

```
TRB.lda <- lda(TRB.trans[-112, ], SymTRB[-112], CV = F)
TRB.ld <- predict(TRB.lda, dim = 2)$x
eqscplot(TRB.ld, col = ColTRB[-112], pch = SymTRB[-112])
```

Figure 11.6

The left-hand graph introduces nothing new; all that needed is to select the first two types (rows 1 to 102) from the original data and labeling variables, so analysis is based on `TRB.trans[c(1:102), ]` for the PCA. The first two lines in the following code set up the results for the LDA plot, but the usual bivariate plots are not available because there are only two groups and one discriminant function. The remaining code produces the plot shown in the right-hand side of the figure.

```
library(MASS)
TRB.lda <- lda(TRB.trans[c(1:102), ], SymTRB[c(1:102)], CV = F)
TRB.ld <- predict(TRB.lda, dim = 1)$x

id <- ifelse(SymTRB2 == 15, 2, 1)
plot(id, TRB.ld, xaxt = "n")
axis(1, at = c(1:2), labels = c("funnel beakers", "bowls"))
```

The variable `id` provides a convenient label for the two groups using the `ifelse` function. Most of the plotting arguments are omitted. The `xaxt = "n"` suppresses printing of the x-axis since its appearance is not as needed. It is replaced using the `axis` function. The first argument, 1, specifies the side at which the new axis is to be placed (bottom – see the help for `axis` for other placements); the `at` argument provides the locations of new labels (replacing 1 and 2 in the original plot) whose names are given in the `labels` argument.

Table 11.1

Replace the argument `CV = F` with `CV = T` in the above code; then

```
posteriors <- round(cbind(id, TRB.lda$class, TRB.lda$posterior), 2)
```

will produce the original and predicted groupings (`id` and `TRB.lda$class`), with the posterior probabilities, `TRB.lda$posterior`, rounded to two decimal places. This is for all the data; a certain amount of (straightforward) editing is needed to get it in the format presented in the text.

The analyses for the steatite data use more groups but involve nothing new in what is needed for the plot.

Chapter 12

Statistical inference

12.1 Introduction

The development of statistical ideas in the first half of the 20th century, including the 'classical' theory of statistical inference and hypothesis testing, was arguably one of the greatest and possibly undersung intellectual achievements of that period. The powerful ideas developed – endowing the treatment of problems involving uncertainty with precision – are seductive and the 'New Archaeologists', from the 1960s on who promoted quantitative methodology, were accordingly seduced (see Chapter 1 of Baxter, 2003).

The statistical landscape has changed since statistics acquired its cachet in some archaeological circles, particularly with respect to the development of computing power and the exploitation of computer-intensive methodologies. The effect of these more recent computer-intensive techniques on *statistical* practice in archaeology merits a review, though this is not attempted here.

The main aim of this chapter is to provide a brief review of the ideas that underpin hypothesis testing which, with variation in emphasis, form one of the staples of quantitative archaeology texts. Their application using `R` is illustrated.

Central to the statistical theory is the idea of drawing inferences about populations from random samples. As discussed in many texts on quantitative methods in archaeology, serious questions arise about the nature of the 'population' sampled, the extent to which samples can be treated as 'random', and the implications this has for the applicability of formal methods of statistical inference to archaeological data. Attitudes have ranged from enthusiastic promotion of statistical inferential methods to scepticism about its practical value – this latter among archaeologists otherwise sympathetic to what statistics has to offer.

This ambivalence can be traced back to the early flourishing of quantitative methodology in archaeology. Archaeologists who embraced statistical inferential ideas sometimes explicitly embedded their thinking within an overtly positivist framework, promoting the approach as 'scientific' and 'objective'. 'Over-selling' of the ideas predictably generated negative reactions. At one extreme, rejection of the New Archaeology led to wholesale rejection of statistical methodology. This is not a logically defensible position; the use of statistical analysis can be decoupled from any overarching philosophy attached to it. As Brandon later commented, in the preface to Westcott and Brandon (2000), 'during its heyday, statistics had been waved above archaeologists' heads as an "answer" to dealing with a multitude of archaeological problems' but 'after much yelling and arm-waving, most agreed that statistics were not an answer in themselves but ... an extremely important tool available for archaeological use'.

Doran and Hodson (1975), in a text sympathetic to the exploration of statistical ideas, explicitly distanced themselves from New Archaeology, finding its claims 'greatly exaggerated and therefore dangerous' and 'a bizarre mixture of naivety and dogmatism' (Doran and Hodson 1975: 5). Their treatment of statistical inference was fairly short and 'theoretical' with few examples of applications; they suggested that, compared to the classical statistical approach, 'a rather different rationale for disciplined inference in archaeology is required' and that this 'remains to be devised' (Doran and Hodson, 1975: 94–95).

Later textbooks have accorded more space to hypothesis testing and inference. Shennan (1988, 1997) and Drennan (1996, 2009) are the most widely used. The more recent text of VanPool and Leonard (2010) contains the most extensive treatment of inferential procedures in introductory archaeology texts I know of but is flawed in some respects (footnote 10).

The normal distribution is fundamental to the statistical theory. Though not used directly for hypothesis testing as much as one might think, it leads to the development of other distributions (t, F, chi-squared etc.) that are widely used in practice. These are the subject of Sections 12.2.3 and 12.2.4. Sections 12.2.1 and 12.2.2 use the normal distribution to introduce ideas central to statistical inference. The chapter misses out a lot; some omissions are noted in Section 12.4

12.2 Common hypothesis tests

12.2.1 The normal distribution

The intention here is to provide a condensed summary of some common hypothesis tests, and associated ideas, along with their implementation in `R`. It is assumed that the reader is acquainted with the idea of the normal distribution. Computational formulae are not provided; they are widely available if needed. Software such as `R` does the computation, leaving users to concentrate on underlying ideas and interpretation.

Mathematically, the normal distribution defines a curve with a total area of 1 beneath it. Distributions are characterized by *parameters*; the normal depends on only two, its *mean*, μ, and *standard deviation*, σ. The distribution is symmetrical about its mean, and 'bell-shaped', so the mean is also the median and mode. About 95% of the distribution lies within two standard deviations from the mean, and 68% within one standard deviation.

If a random variable X has a normal distribution the notation

$$X \sim N(\mu, \sigma^2)$$

is used to express this, where σ^2 is the *variance*. Given a *random sample* of size n from a normal distribution denote the *estimated* mean and standard deviation by $\bar{x}$ and s; the estimated standard error of the mean is $s/\sqrt{n}$. If σ is known $\sigma/\sqrt{n}$ can be used rather than its estimate.

With this notation in place, assuming known σ,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

has the *standard normal distribution* with mean zero and unit variance.

Many applications involve finding the probability in the tail of the distribution, defined by some value of z. Output in `R` for analyses based on the normal and other distributions often include the information needed, but functions exist allowing their separate calculation if necessary.

Thus, the `pnorm` function, for $z = -2.36$ using `pnorm(-2.36)`, returns the value 0.009137468, the probability in the lower-tail of the distribution written as P(z < -2.36) $= 0.009$. Replacing z with + 2.36 gives 0.9908625, or P(z $< +2.36$) $= 0.991$. These are examples of a p-values.To get the upper-tail probability P(z $> +2.36$) either subtract this from 1 or, more directly, use `pnorm(2.36, lower.tail = FALSE)`. In either case, given the symmetry of the distribution, 0.009137468 is obtained[1].

These are examples of lower-tail and upper-tail probabilities. Often interest lies in assessing how extreme the observed result is; that is, what is the (combined) tail probability of getting a value either more negative than -2.36 or more positive than +2.36. This is obtained by doubling the lower- or upper-tailed probability calculated and is written as the two-sided p-value P(|z| > 2.36) $= 0.018$, or slightly less than 2%. The p-value needs to be interpreted and this brings us on to the subject of inference.

[1] This renders redundant the detailed tables of the normal and other distributions – once essential – still to be found in the appendices of introductory texts.

12.2.2 Inference

The value of z derived from a sample can be used in two ways. If μ and σ are known for some population of interest, assumed to be normal, the p-value can be used to assess if it is plausible that the sample is selected from the population specified. If the p-value is 'small' it would be concluded that the sample is unlikely to be from the population. In practice this kind of application is limited.

More usually, and the main idea in the tests used here, a value is *assumed* for the mean μ, and the observed data are used, via the calculated z statistic, to assess the plausibility of this assumption. A simple way of thinking about the more basic inferential procedures is that they are employed in order to say something useful about the true values of the unknown parameters; this includes the use of confidence intervals (Section:12.2.5).

The *null hypothesis* states that μ equals some specified value, μ_0. The notation used to express this is $\mathrm{H}_0 : \mu = \mu_0$. The z-statistic can be calculated using this assumption. If it is 'unusual' as measured by the p-value then we *reject* the null hypothesis, otherwise we do not reject it[2].

Decisions need to be made on the part of the user. One is whether to use one- or two-sided p-values. The choice amounts to specifying an *alternative hypothesis*; for example, if $\mathrm{H}_0 : \mu = 0$ possible alternative hypotheses are $\mathrm{H}_1 : \mu < 0$, $\mathrm{H}_1 : \mu > 0$ or $\mathrm{H}_1 : \mu \neq 0$. The first two of these are examples of one-sided alternative hypotheses where it is believed that, if the null hypothesis is incorrect, the true value differs in a particular direction from that assumed; the third possibility, an instance of a two-sided test, specifies that if the null hypothesis is incorrect the true value of μ could lie either side of that hypothesized. Two-sided tests will be used below unless otherwise stated.

The important question is 'what constitutes a 'small' or 'unusual' p-value?'. Some researchers are content to report the exact p-value, allowing others to decide on their own interpretation. Commonly, though, a *decision rule* is used. These are arbitrary; conventionally the use of rules based on p-values of 0.05 and 0.01 are widespread. If the p-value is less than 0.05, using a two-sided test, the null hypothesis is rejected at the 5% level of significance or, more concisely is *significant at the 5% level.* If the p-value is also less than 0.01 the null hypothesis is rejected at the 1% level, providing stronger evidence against the null hypothesis.

[2] It is tempting to say, in the latter case, that the null hypothesis is 'accepted' and the temptation should be resisted. A null hypothesis may not be rejected for a host of reasons, of which the possibility that it is 'true' is only one. Small sample sizes and/or large sample variability can both lead to non-rejection, even when the null hypothesis is false. Non-rejection implies that there is not enough evidence with the data to hand to reject the null hypothesis; the term 'accepted' carries connotations of establishing that the null hypothesis is 'correct' which is not the case.

As a 'rule-of-thumb', using p-values based on 0.05, 0.01, and 0.000, and rejecting the null hypothesis, implies 'strong', 'very strong' and 'overwhelming' evidence against the null hypothesis. For the normal distribution the 5% and 1% levels of significance correspond to values of $|z| > 1.96$ and $|z| > 2.575$. This means that, given a z-score one looks to see if its modulus exceeds 1.96 or 1.575. These are called *critical values*.

As a notational aside, α can be used for the associated p-values, with $z_{\alpha/2}$ being the z-score that 'cuts off' $\alpha/2$ of the probability in the upper-tail. Thus, $z_{.025} = 1.96$ defines a tail probability of 0.025 which, for a two-sided test, is doubled to get the significance level of 0.05 (5%), that is, $\mathrm{P}(|z| > 1.96) = 0.05$. Some help is available in choosing α. It is the probability of incorrectly rejecting a correct null hypothesis; this is a *Type I error* and the user controls the probability of this in setting the decision rule. If it is important not to make this sort of error a relatively small value of α would be set.

A *Type II error*, occurs when an incorrect null hypothesis is not rejected. Call the probability of this β; the *power* of a test is defined as $(1 - \beta)$ and is the probability of correctly rejecting an incorrect null hypothesis. A balancing act between the two types of error is implicit since reducing the size of α increases the size of β and hence also reduces the power. Usually the focus of interest is on controlling α and one 'lives with' the associated power; increasing the sample size n is one way of improving the power. Other things (i.e. α) being equal, competing tests that do the same job can be compared in terms of their relative power.

12.2.3 Tests of means – t-tests

The z-test is of limited use for practical purposes because of the 'σ known' assumption; it usually isn't. This is remedied in a straightforward way at the expense of a little extra complexity. The z-statistic can be expressed in a general form as

$$z = \frac{\text{difference}}{\text{SE}}$$

where 'difference' is the observed difference between $\bar{x}$ and μ_0, and SE is the standard error which scales the statistic so it is dimensionless. The 'obvious' modification if σ is unknown is to use the estimated standard error, $s/\sqrt{n}$ resulting in the t-statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

The distribution of this (assuming independent random sampling from a normal distribution) looks much like the normal but with thicker tails, the thickness of which depends on the sample size, n, through a quantity known as the *degrees of* *freedom* (DF) which is $(n-1)$ for single-sample tests. As n gets larger $s \to \sigma$ and the t-distribution approaches the normal distribution.

The t-statistic is therefore most valuable when dealing with small samples. It was for precisely this problem that the statistic was developed by W. G. Gosset, published under the pseudonym 'Student' in the journal *Biometrika* in 1908 (hence the terminology 'Student's t-test' sometimes seen). The ideas of the previous section carry through, but the critical value that varies with n is now denoted $t_{\alpha/2}$.

In R suppose that, for a sample size $n = 7$ with 6 DF, $t = -2.36$ is obtained. Using the `pt` function, `pt(-2.31, 6)` returns 0.0281, where the second argument in the call to the function, 6, is the degrees of freedom. Doubling this to get the two-sided p-value gives 0.056 so the null hypothesis would not be rejected at the 5% level of significance; for $n = 17$, 37 and 137 the respective probabilities are 0.031. 0.024, and 0.020. If a critical value, $t_{\alpha/2}$ is required use the `qt` function; for a sample size of 7 and $\alpha = .025$, `qt(0.025, 6)` gives $|t_{\alpha/2}| = 2.45$.

One-sample hypothesis tests are often uninteresting. The specification of a realistic null hypothesis can be 'artificial' (where does it come from?). It is difficalt to find examples in the literature where real data (i.e. with contextual information) is confronted with a realistic problem requiring a one-sample test (see, also, Section 12.2.5). Practically, the comparison of samples from two populations is more likely to be of interest. Here there is a 'natural' null hypothesis to test.

The two-sample t-test has a similar structure to those already discussed. The difference between the two sample means is $(\bar{x_1} - \bar{x_2})$ and the natural specification of the null hypothesis is that the difference in population means is $(\mu_1 - \mu_2) = 0$ (note the use of subscripts to distinguish between the two populations/samples). Extra complexity is introduced because the SE can be estimated in two ways. The first of these – that usually described in introductory texts and the default in many software packages – assumes that the population variances are equal, $\sigma_1^2 = \sigma_2^2$, and the estimate of the SE is based on a weighted average of their estimates, with DF $= (n_1 + n_2 - 2)$. In the second case no such assumption is made and the SE is estimated as the square-root of the sum of the squared SE estimates of the two samples. In this case the DF needs to be approximated, and an approximate t-test is carried out using this.

It's possible to test the hypothesis that the variances are equal in advance of a t-test – the subject of the following section. More simply, it is straightforward to carry out tests with and without the assumption (Section 12.3.2). Minor numerical differences apart results, in terms of the conclusions to be drawn, will often be the same; if not there may be problems with the samples used that need addressing before applying any formal test about the means. The difference in the definitions of the estimated SEs is spelled out in the footnote; it is highly improbable that you will ever need to calculate these 'by hand'[3].

12.2.4 Tests of variances

The sum of squares of n independent identically distributed normal random variables has the chi-squared (χ^2) distribution, which is asymmetrical and bounded below by zero. The ratio of two chi-squared variables has the F-distribution; these can be used directly for one- and two-sample hypothesis tests but arise in other contexts as well.

The null hypothesis $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$ can be tested using a chi-squared statistic, but I cannot recall seeing any realistic application of this in the archaeological literature and will not discuss it further. Chi-squared tests do have a useful role in testing the hypothesis of no association between two categorical variables and this is illustrated in Section 12.3.3.

The F-test is based on the ratio of two chi-squared random variables and may be used for testing hypotheses about two population variances where the 'natural' null hypothesis is $\mathrm{H}_0 \colon \sigma_1^2 = \sigma_2^2$, or $\mathrm{H}_0 \colon \sigma_1^2/\sigma_2^2 = 1$. This can be applied in advance of a two-sample t-test to see if the equal variances assumption is reasonable. The hypothesis is tested using the statistic

$$\mathrm{F} = s_1^2/s_2^2$$

which under the null hypothesis, follows the F-distribution with $(n_1 - 1), (n_2 - 1)$ DF, a little more complicated than the previous tests since it depends on two separate degrees of freedom.

[3] In the first case the estimated SE is

$$s\sqrt{(1/n_1 + 1/n_2)}$$

where

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

. In the second case the estimated SE is

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

None of the quantitative archaeology texts I am familiar with give the formula for the approximate DF in the second case so, for the record, here it is.

$$DF = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_1 - 1)}.$$

As commented you ought never need to have to calculate this yourself; you should know of the two possibilities since the second case is the default in `R`.

It is convenient to label the samples so that $s_1^2 > s_2^2$ and F > 1 so only the upper-tail of the distribution is of concern. The decision rule for a 5% level of significance requires that the critical values of $F_{0.025}$ or $F_{0.05}$ be determined, depending on whether a two- or one-sided test is being used. For a problem where F $= 2$, $n_1 = 7$ and $n_2 = 9$ there are (6, 8) DF. Using the `qf` function, `qf(.025, 6,8, lower.tail = F)` gives $F_{0.025} = 4.652$, so the null hypothesis is not rejected at the 5% level of significance using a two-sided test. Replacing 0.025 with 0.05 gives $F_{0.05} = 3.581$, leading to the same conclusion as for a one-sided test. Replace 0.025 and 0.05 with 0.005 and 0.01 for tests at the 1% level.

The function `pf`, analogous to `pnorm` and `pt` is also available if one wishes to avoid strict adherence to a decision rule. Thus, supposing F $= 4.2$ with (6, 8) DF, `pf(4.2, 6,8, lower.tail = F)` gives a p-value of 0.033, which is significant at the 5% level, but not at 1%, for a one-sided test. On doubling it, it is not significant at either level for a two-sided test.

12.2.5 Confidence intervals

An 'objection' of sorts to the tests that have been described, already voiced – even when the assumptions needed are valid – is that they are 'uninteresting'. Ultimately all they do is tell you whether the single value defined by the null hypothesis is 'plausible' or not. No information is provided about what other values are, or are not, plausible, or how precisely the sample statistics estimate the population values. Confidence intervals provide the same information as an hypothesis test, and much more besides. Attention is confined to problems involving means using t-tests; similar ideas can be applied to problems concerning variances.

In a slight extension of notation let $t_{crit} = |t_{df,\alpha/2}|$ be the critical value of t for a (two-sided) test at the α (or $100\alpha\%$) level of significance, with *df* the degrees of freedom. The formulae for both one- and two-sample tests can be rearranged to get an expression for μ or $\mu_1 - \mu_2$ which can be evaluated at the positive and negative values of t_{crit} to obtain a $100(1-\alpha)\%$ confidence interval.

For the single-sample problem this leads to

$$\bar{x} - SE \times t_{crit} \leq \mu \leq \bar{x} - SE \times t_{crit}$$

and for the two-sample problem

$$(\bar{x}_1 - \bar{x}_2) - SE \times t_{crit} \leq ((\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + SE \times t_{crit}$$

where SE and *df* are the appropriate standard error and degrees of freedom that would be used for the associated hypothesis test. The following may be noted.

1. There is no commitment to a particular null hypothesis.

2. Any value contained in a $100(1-\alpha)\%$ confidence interval would not be rejected at the $100\alpha\%$ level of significance; if a value of μ is contained in a 95% confidence interval it would not be rejected as a null hypothesis at the 5% level of significance[4].

3. A consequence of the above is that conclusions about any null hypothesis of interest can be 'read off' from the confidence interval, so testing the null hypothesis in the manner described in earlier sections is redundant.

4. The width of a confidence interval provides some indication of how good an estimate the sample mean (or difference in means) is of the population means (or their difference). A very narrow confidence interval, for example, 'pin-points' the true value of the parameter of interest very well. An hypothesis test, without elaboration, provides no such information.

Together these amount to a powerful argument for preferring confidence intervals rather than hypothesis tests in many practical applications.

12.3 Examples of R use

12.3.1 Data

The data used are from Shennan (1997: 103). Twenty-four observations are available on the areas (square meters) of *marae* ceremonial enclosures of the Society Islands in the Pacific, located in two different valleys. Interest lies in whether or not there are differences between the valleys in terms of the mean size of the enclosures as measured by area.

The data are given in Table 12.1 and have been reconstructed from Shennan who provides log-transformed values because 'a preliminary check indicated that [area] was very skewed so it has been logged' (Shennan, 1997: 102).

The reconstructed data for the two valleys are shown in the stripcharts/dotplots of Figure 12.1, along with those for the log-transformed data.

[4] This is possibly the simplest way of thinking about the interpretation of confidence intervals. Another way is that in a hypothetically large number of repeated samples the confidence intervals will cover the true value 95% of the time; this is illustrative of the *frequentist* way of thinking about probability which not all statisticians like (to put it mildly!). In practice one has a single confidence interval to deal with, and it either does or doesn't cover the true value – you don't know. It is tempting, and understandable, to say that you are 95% 'confident' that your interval contains the true value, and is wrong. This use of the term 'confidence' can be taken to imply a conception of probability as a 'degree of belief'; this is outwith the frequentist paradigm though employed in other approaches to inference.

	Area (square meters)												
Valley 1	2.87	2.52	2.35	1.48	1.92	1.99	1.84	2.01	2.16	2.56	1.95	2.00	2.35	1.87
Valley 2	1.94	2.17	2.05	2.22	2.31	2.14	1.95	2.40	2.81	2.74				

Table 12.1: *The area (square meters) of* marae *ceremonial enclosures from two regions (adapted from Shennan, 1997: 103).*

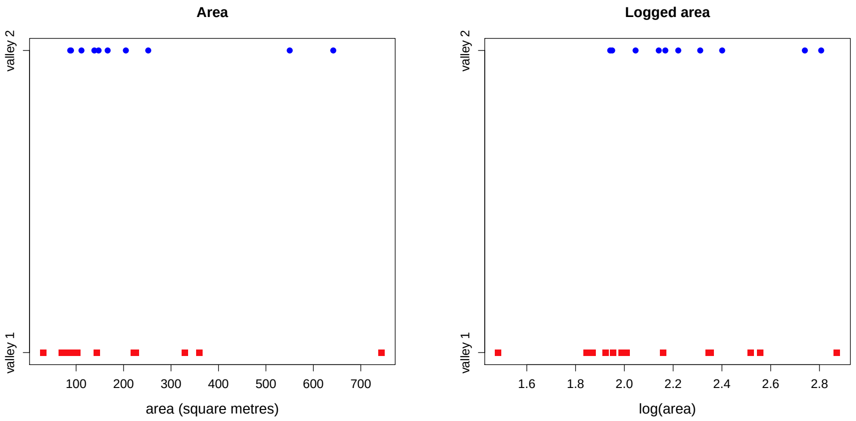

Figure 12.1: *Stripcharts of the area and log-transformed area of* marae *enclosures from two separate locations.*

The most obvious feature of the plot for the untransformed data is the two clear outliers for Valley 2, and an outlier for Valley 1 which might be interpreted as evidence of skewness; the log-transform appears to 'cure' this problem for the first valley, but not entirely so for the second[5].

Looking at the plot for the log-transformed areas there is little obvious evidence of serious differences in the location and spread of the data for the two valleys and an experienced analyst might stop at this point; we shall proceed with more formal tests by way of illustration. The estimated means and variances for the log-transformed data are $\bar{x}_1 = 2.13$, $\bar{x}_1 = 2.27$, $s_1^2 = 0.128$ and $s_2^2 = 0.091$. For illustrating one-sample t-tests the data for Valley 2 is used and it is assumed that $\mathrm{H}_0 \colon \mu = 2.10$ is of interest[6].

[5] Since area is a squared quantity a square-root transformation is another possibility and produces results very similar to the log-transformation.

[6] This kind of assumption is sometimes motivated in texts by assuming that previous study, involving a large sample or a population, has suggested the null hypothesis.

12.3.2 *t*-tests

For both one- and two-sample *t*-tests in `R` the `t.test` function is used. Defaults are to carry out a two-sided test, reporting a 95% confidence interval. For tests about the equality of variances the `var.test` function is used. The following six tests were applied; the first three involve one-sample *t*-tests and illustrate typical output and the use of arguments to the function. The fourth and fifth examples illustrate the two versions of the two-sample *t*-test, and the sixth example involves the F-test.

```
t.test(valley_2, mu = 2.1)
t.test(valley_2, mu = 2.0)
t.test(valley_2, mu = 2.0, conf.level = 0.99)
t.test(valley_1, valley_2)
t.test(valley_1, valley_2, var.equal = TRUE)
var.test(valley_1, valley_2)
```

The log-transformed data for the two sites are contained in the objects `valley_1` and `valley_2`. For the one-sample tests the null hypothesis is specified using the argument `mu`; for the two-sample *t*-tests equality of the population means is the default null hypothesis, without assuming equality of population variances; for `var.test` equality of population variances is the default null hypothesis (these can be varied if necessary – use the help facility for details). The confidence interval can be varied using the `conf.level` argument and will be illustrated; one-sided tests are not shown but the arguments `alternative = "less"` or `alternative = "greater"` are available, depending on which of $\mu < \mu_0$ or $\mu > \mu_0$ is of interest.

For the first test shown, where $\mathrm{H}_0 \colon \mu = \mu_0$ with $\mu_0 = 2.1$, the following output is obtained.

```
        One Sample t-test

data:  valley_2
t = 1.8119, df = 9, p-value = 0.1034
alternative hypothesis: true mean is not equal to 2.1
95 percent confidence interval:
 2.057083 2.488317
sample estimates:
mean of x
   2.2727
```

The *p*-value 0.1034 is greater than 0.05 so is not significant at the 5% or even 10% level (if it is not significant at 5% it cannot be significant at 1%). The 95% confidence interval is (2.06, 2.49). Any value ouside this range will be rejected as a null hypothesis at the 5% level; this is illustrated in the second example where $\mu_0 = 2.0$ is assumed for the null hypothesis. The following output excludes some of the information identical to that from the first analysis.

```
t = 2.861, df = 9, p-value = 0.01875
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 2.057083 2.488317
```

The t- and p-values change, the latter to 0.019. This can be reported by simply noting it; observing that the test is (just) significant at the 2% level; or that the null hypothesis is rejected at the 5% but not the 1% level. Information on the confidence interval does not change and is shown here precisely to emphasize this point; it does not depend on the value used for the null hypothesis. The third example is identical to the second except for the confidence level

```
99 percent confidence interval:
 1.962942 2.582458
```

where the 99% confidence interval, (1.96. 2.58) is (inevitably) wider than the 95% interval. Note that the interval contains the value 2.0, another way of showing that the null hypothesis $\mathrm{H}_0 : \mu_0 = 2.0$ would not be rejected at the 1% level of significance.

For the first of the two-sample t-tests, not assuming equal variances, the output is

```
        Welch Two Sample t-test

data:  valley_1 and valley_2
t = -1.026, df = 21.309, p-value = 0.3164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4193637  0.1421065
sample estimates:
mean of x mean of y
 2.134071  2.272700
```

In the call to the `t.test` function separate listing of the data for the two groups as the first two arguments alerts the function to the fact that a two-sample test is intended.

The additional argument `var.equal = TRUE` in the fifth analysis specifies that equal variances are to be assumed. The differences from the results from the analysis that does not make this assumption are

```
        Two Sample t-test

data:  valley_1 and valley_2
t = -0.9959, df = 22, p-value = 0.3301
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4273128  0.1500556
```

There are minor differences in the numerical output but the conclusion that there is no evidence of a difference in population means, with *p*-values of more than 0.3, is the same in both cases. The 'Welch' in the heading to the output in the first analysis pays tribute to one of the scholars responsible for the theory associated with the unequal variances problem.

The results are so similar that the simpler and more widely advertized equal-variances version is clearly acceptable for reporting purposes, and implies there is no evidence to contradict this assumption. An experienced analyst familiar with F-tests would recognize this by just looking at the variance estimates. More formally, using the `var.test` function we get

```
        F test to compare two variances

data:  valley_1 and valley_2
F = 1.4132, num df = 13, denom df = 9, p-value = 0.6124
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3689233 4.6805505
sample estimates:
ratio of variances
          1.413196
```

The *p*-value is so 'large' that the conclusion would often be expressed – a little loosely – as 'not significant', with specific levels such as at 10%, 5% and 1% left implicit.

12.3.3 Chi-squared tests

The chi-squared test arises naturally in the context of testing hypotheses about single variances, but these are of limited interest and not discussed. The test is of greater importance for testing the hypothesis of no association between categorical variables displayed in a two-way contingency table. The left-hand side of Table 12.2 (Table 2 in VanPool *et al.*, 2000), is an example. The entries show the frequencies

	Observed				Expected				
	Room 1	Room 2	Room 3	Plaza	Room 1	Room 2	Room 3	Plaza	Total
Chert	86	21	38	97	81	23	31	106	242
Chalcedony	39	10	12	13	25	7	10	33	74
Obsidian	4	3	3	8	6	2	2	8	14
Quartzite	16	8	7	6	12	4	4	16	37
Igneous	217	63	81	353	238	69	93	314	714
Total	362	105	143	477	362	105	143	477	1085

Table 12.2: *The observed and expected frequencies of flaked stone artifacts from three room blocks and a plaza area at Galeana, a large pueblo-like settlement in north-western Mexico. (Source: Table 2 in VanPool* et al.*, 2000.)*

of flake stone artifacts, categorized by material and location, found within a pre-Hispanic settlement site in Mexico.

The chi-squared test is almost invariably covered in introductory texts so only an outline is given here, the focus being on practical application. The table is of size $I \times J$, in this instance 5×4. There are thus 20 observations or *cells* in the table with entries that will be denoted by O_{ij} for row i and column j. Another way of expressing the null hypothesis of no association between material and location is that the distribution of frequencies across location (expressed as a percentage) should be similar apart from random variation. Exploratory methods for investigating this include barplots (Chapter 4) and correspondence analysis (Chapter 9); the chi-squared test is a more formal means of investgation.

To proceed, the *expected values*, E_{ij}, under the null hypothesis are needed and these are given (rounded to integers for easier comparison) to the right-hand side of Table 12.2, by

$$\frac{\text{row total} \times \text{column total}}{\text{overall total}}.$$

The chi-squared test statistic is then calculated as

$$X^2 = \sum \frac{(O-E)^2}{E}$$

which follows the chi-squared distribution (χ^2) (approximately) with $(I-1)(J-1)$ DF if the null hypothesis is true[7]. If the null hypothesis is incorrect some of the O will differ noticeably from E so X^2 will be 'large' and a one-sided test is called for in using χ^2 to make this assessment. It is fairly obvious from tabular inspection (particularly if row percentages are used) that the distribution of materials across locations varies, so a 'significant' result rejecting the hypothesis of no association is expected.

[7] I use X^2 for the test statistic to distinguish it from the theoretical χ^2 value – this usage is not universal.

Using `chisq.test(flakes, correct = F)`, where `flakes` is the name of the data set, confirms this

```
        Pearson's Chi-squared test

data:  flakes
X-squared = 49.1023, df = 12, p-value = 2.007e-06

Warning message:
In chisq.test(flakes) : Chi-squared approximation may be incorrect
```

The small p-value leads to clear rejection of the null hypothesis of no association. This can also be obtained using `pchisq(49.103, 12, lower.tail = F)`. If a decision rule needs to be explicitly stated (as some journals require) `qchisq(0.01, 12, lower.tail = F)` gives the 1% critical value as 26.22.

Rather than relying on the p-value alone more detailed inspection of output can be helpful. Use `chisq.test(flakes, correct = F)$expected` to obtain expected values; for residuals replace `expected` with `residuals` and for standardized residuals use `stdres` instead of `expected`. The (Pearson) residuals are defined as $(O - E)/\sqrt{E}$ and the standardized residuals as $(O - E)/s$ where s is an estimate of the standard deviation of $(O - E)$.

Values for the residuals, with and without standardization, are given in Table 12.3. As a rule-of-thumb, absolute values of the standardized residuals in excess of 2 draw attention to cells where departure from the null hypothesis is most obvious. Several such values stand out in the table; these are mostly associated with the 'Chalcedony' and 'Igneous' categories, the 'Plaza' also standing out. Examining Table 12.2 shows that 'Chalcedony' is under-represented in the Plaza and over-represented in the rooms compared to the predictions of the hypothesis of no association; the opposite is true for the 'Igneous' material.

	Residuals				Standardized residuals			
	Room 1	Room 2	Room 3	Plaza	Room 1	Room 2	Room 3	Plaza
Chert	0.59	-0.5	1.17	-0.91	0.81	-0.6	1.42	-1.38
Chalcedony	2.88	1.06	0.77	-3.42	3.66	1.16	0.85	-4.74
Obsidian	-0.82	0.95	0.43	0.03	-1.01	1.01	0.47	0.04
Quartzite	1.04	2.34	1	-2.55	1.3	2.5	1.09	-3.46
Igneous	-1.37	-0.73	-1.22	2.21	-2.88	-1.32	-2.24	5.04

Table 12.3: *Residuals, raw and standardized, from a chi-squared test of the data in Table 12.2.*

The warning message alerts you to the fact that some of the expected values are small (less than 5) which may invalidate the approximation of the distribution of the test statistic to χ^2. The sample size for obsidian of 14 is quite small and two of the expected values are noticeably less than 5. It is not a problem here; the cells affected do not contribute greatly to the highly significant value of X^2 as the residuals in Table 12.3 show. This is not an inevitable outcome and small expected values can be particularly problematic with small overall sample sizes[8].

Small samples can be dealt with in a variety of ways. The example so far used the argument `correct = F` in the call to `chisq.test`, which calculates X^2 as defined. The default for 2 × 2 tables is actually `correct = T` which applies Yates's *continuity correction*, replacing $(O - E)$ in the definition with $|O - E| - 0.5$ with the intention of improving the approximation to χ^2. Another alternative, particularly useful for 2 × 2 tables, is the *Fiser exact test* the details of which are not entered into here. For illustrative purposes the data from the rows for 'Chert' and 'Igneous' and columns for 'Room 2' and 'Room 3' from Table 12.2, divided by 10 and rounded, will be used, giving

	Room 2	Room 3	Total
Chert	4	8	12
Igneous	12	16	28
Total	16	24	40

Table 12.4: *Artificial data. The observed frequencies of flaked stone artifacts have been adapted from Table 12.2 using two rows and columns only and rescaling by dividing by 10.*

The p-values for the chi-squared test with and without the continuity correction gives p-values of 0.83 and 0.57, both tests issuing a warning about the validity of the chi-squared approximation. The Fisher test, implemented with the `fisher.test` function, `fisher.test(flakes)`, returns a p-value of 0.67. None of these values are significant at the levels usually employed, but the example is sufficient to demonstrate that they can give rise to different p-values so can potentially lead to different conclusions about the null hypothesis.

It is difficult to advise on which test is 'best'; opinions differ. For something as apparently simple as a 2 × 2 table there is a considerable statistical literature on the subject[9]. In `R` it is possible to use simulation to obtain a p-value without assuming a chi-square distribution with the `simulation.p.value = T` argument.

[8] Some of the test statistics described so far are obviously dependent on sample size and increase as n gets larger, resulting eventually and inevitably in the rejection of any null hypothesis. The same is true, less obviously, for X^2. If the observed values of O are scaled by a factor of k the expected values follow suit and $X^2 \to kX^2$. An implication of this is that such tests are of limited value for very large samples, the 'significance' of any differences being a matter of expert judgment rather than formal testing.

[9] Some of this is to do with practical performance; some to do with theoretical matters con-

For the 5×4 Table 12.2 a p-value is 0.0005, which is different from that previously obtained; for the 2×2 example with and without continuity correction the p-values are 0.72 and 0.74. If simulation is used, continuity correction is only available for the 2×2 case; the p-value will vary slightly if simulation is repeated, hence the use of the term 'a p-value'.

12.3.4 F-tests and ANOVA

Analysis of variance (ANOVA) techniques are important in a wide range of statistical applications. They are not mentioned much, if at all, in quantitative archaeology texts (VanPool and Leonard, 2010, is an exception), possibly because the applications are often 'complex' enough to be beyond the ambitions of such texts. It is also the case that applications in the archaeological literature are not profuse (though this is also true of less complex tests that are accorded space).

The general idea underpinning ANOVA is discussed briefly; the only application considered in any detail is the problem of comparing more than two sample means using one-way ANOVA. Models for data, where ANOVA is relevant, typically assume that the data can be modeled as the sum of systematic and random (or error) components, that is

$$\text{Data} = \text{Systematic} + \text{Random}$$

with a focus on whether or not the systematic component is, in some sense, 'important' compared to the random component. In general the systematic element can itself consist of component parts, and interest may focus on whether or not only a subset of these are required to explain variation in the data.

To test this, the total variation in the data is broken down into the contributions of systematic variation and random variation or sums of squares (SS); the SS are converted to variances by division by the appropriate degrees of freedom so MS = SS/DF where MS is the *mean square* – call these MS_S and MS_R for the systematic and random components – and their ratio is calculated as

$$\text{F} = \text{MS}_S/\text{MS}_R$$

where F is an F-statistic with degrees of freedom determined by the context. Under the null hypothesis, which will also depend on context, and assuming normality of the error term this should follow an F-distribution and the approach outlined for

cerning 'experimental design'. Commonly Fisher's test has been recommended when observed values are small, but it has also been noticed that it often performs similarly to a chi-squared test with a continuity correction. A preference for chi-squared without the continuity correction is possibly a minority opinion among statisticians but sometimes the only version presented in introductory texts. The continuity-corrected version is the default in R but other software can differ.

comparing two sample variances in Section 12.2.4 and at the end of Section 12.3.2 can be used.

To show what R output looks like, we revisit the two-sample t-test of Sections 12.2.3 and 12.3.2, reformulated as an ANOVA problem. The log-transformed area data for both valleys need to be stacked as a single column of data (log_area), with a second column of equal length providing information on site location (valley).

Different ways of conducting the ANOVA are available of which the function oneway.test is simplest. The test may be carried out with or without assuming equal error variances within samples. The latter is the default. In the following output the last two lines were obtained using var.equal = TRUE in the function call. The p-values and their DF are identical to those from the t.test analysis.

```
oneway.test(log_area ~ valley, var.equal = FALSE)

        One-way analysis of means (not assuming equal variances)
data:  log_area and valley
F = 1.0527, num df = 1.000, denom df = 21.309, p-value = 0.3164

# Using var.equal = TRUE
        One-way analysis of means
F = 0.9918, num df = 1, denom df = 22, p-value = 0.3301
```

The more general problem of comparing $p > 2$ means involves the null hypothesis

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_p.$$

For a second example data on the maximum flake length (mm) of 10 unbroken flakes for each of four raw material types from Cerro del Diablo, a Late Archaic Mexican site, given in Table 10.1 of VanPool and Leonard (2010), are reproduced in Table 12.5.

Assuming these can be treated as random samples for the four material types the null hypothesis is $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (i.e. the mean population lengths are the same) [10]. It is obvious that H_0 will be rejected (compare the lengths for obsidian and rhyolite). A formal test of this is not really needed; the comparison

[10] It was noted in the introduction that VanPool and Leonard (2010) deal with hypothesis tests at greater length than competing texts. It was also suggested that the treatment is flawed, and this is an apposite point at which to elaborate. The fundamental problem is that, having alerted the reader to the importance of distinguishing between population and sample quantities, the authors ignore their own advice. The treatment of the way null hypotheses are expressed is wrong. Thus, in the context of two-sample t-tests, the usual null hypothesis that the population means are the same, $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$ in the example, is expressed incorrectly in terms of equality of the *sample means*, $\mathrm{H}_0 : \bar{X}_1 - \bar{X}_2 = 0$. This is a statement that the sample means are the same; the observed data are used to test hypotheses about unknown population quantities and it does

Chert	Obsidian	Rhyolite	Silicified wood
41	30	135	113
110	53	141	111
73	45	138	97
52	34	175	70
176	105	143	117
61	102	132	48
69	51	130	134
40	47	109	115
64	71	125	103
48	58	120	106

Table 12.5: *The maximum flake length (mm) of unbroken flakes for four raw material types (Source: Table 10.1 in Van Pool and Leonard, 2010.)*

between pairs of material types is of more interest and can be achieved using the `aov` function, which is more general than `oneway.test`.

In the first instance it is sensible to look at the data graphically, which can be done using boxplots as in Figure 12.2. That there are differences between material types, implying that the null hypothesis will be rejected, is obvious. Flakes made of rhyolite and silicified wood clearly tend to be longer than those of chert and obsidian; it is less clear if chert and obsidian differ significantly (though probably not) or if rhyolite and wood differ. There are some clear outliers within material types that will be temporarily ignored in order to illustrate the mechanics of application.

The basic analysis is as follows. The data are held in a data frame `aov.data` with the columns labeled `length` and `material`. It is obvious from the p-values (and the cues provided in the output) that the null hypothesis is comprehensively rejected.

```
aov.example <- aov(length ~ material, data = aov.data)
summary(aov.example)
            Df Sum Sq Mean Sq F value    Pr(>F)
material     3  33156 11051.9  13.373 5.085e-06 ***
Residuals   36  29751   826.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

not makes sense to express the null hypothesis about unknowns in terms of known sample values. This notational problem, arguably a conceptual one as well, is pervasive in the several chapters dealing with hypothesis testing problems, including those on ANOVA and chi-squared tests.

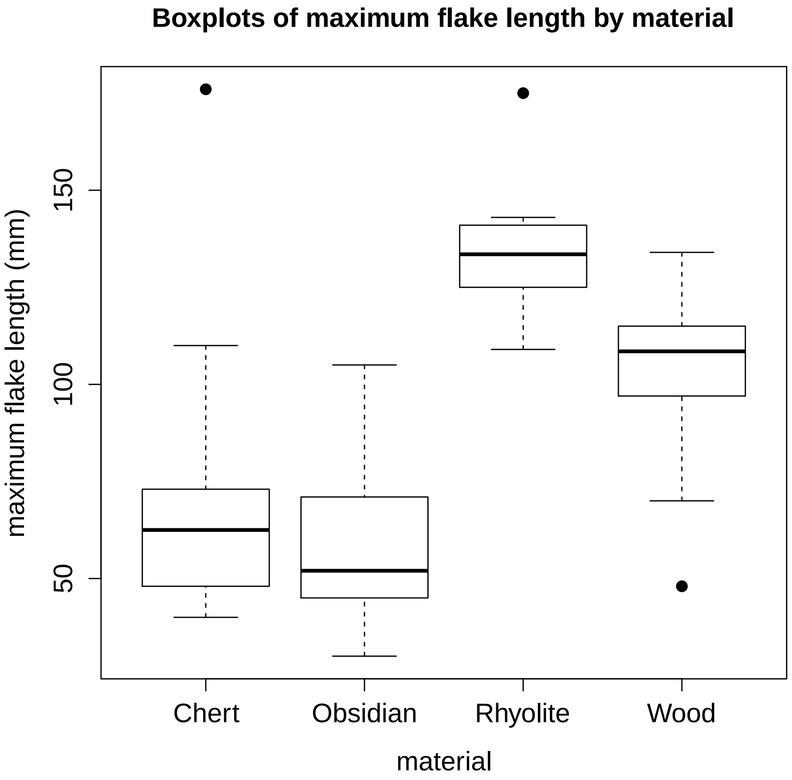

Figure 12.2: *Boxplots of maximum flake lengths by material using the data of Table 12.5.*

To investigate which material types differ in their typical length in a formal way the obvious thing to do is to undertake all possible pairwise comparisons; this involves a *multiple comparison test*. The two-sample *t*-test is a candidate for such testing but inflates the probability of finding a significant difference. There are different approaches to 'correcting' for this; a popular one is Tukey's HSD (Honestly Significant Difference) test which, given the fitted `aov` model, can be implemented using the `TukeyHSD` function so `TukeyHSD(aov.example)` returns

```
$material
                   diff        lwr        upr     p adj
Obsidian-Chert    -13.8 -48.424703  20.824703 0.7076807
Rhyolite-Chert     61.4  26.775297  96.024703 0.0001685
Wood-Chert         28.0  -6.624703  62.624703 0.1488159
Rhyolite-Obsidian  75.2  40.575297 109.824703 0.0000064
Wood-Obsidian      41.8   7.175297  76.424703 0.0127440
Wood-Rhyolite     -33.4 -68.024703   1.224703 0.0620233
```

Here `diff` is the difference in means and `lwr` and `upr` are lower and upper values of a 95% confidence interval (the default). Superficially the *p*-values do not entirely concur with the expectations raised by preliminary data analysis. In particular, the wood-chert comparison is not significant at the 10% level, the wood-obsidian comparison is significant at the 5% but not the 1% level, and the wood-rhyolite comparison, with the least 'predictable' outcome, is not significant at the 5% level.

The term 'superficially' is used because the ANOVA assumes that the samples within material types are from a normal distribution, and equal variances are assumed. So far nothing has been done about the outliers, that for chert at 176 mm being particularly prominent. It is possible to test for equal variances using `bartlett.test(length ∼ material, data = aov.data)` which gives a *p*-value of 0.09 so, if a decision rule of 5% is used, the variances do not differ significantly at this level. The problem here is that outliers will inflate the variances, possibly quite considerably, and their effect on the test outcome is unpredictable.

If nothing else, the outliers also call into question the normality assumption. This can, if one wishes, be tested for formally. Many such tests have been proposed a large number of which are rarely used, if at all; the Shapiro-Wilk test, implemented with the `shapiro.test` function in `R`, is widely-regarded as one of the best. For chert, the first ten values in the stacked data set, produces a *p*-value of 0.005 with `shapiro.test(aov.data$length[1:10])` which is very strong evidence against the normality assumption; repeating this for other materials gives non-significant results at the 10% level, so chert is the main problem.

In the absence of archaeological knowledge that might dictate that the chert outlier be treated separately, the sensible thing is to omit it to see if it affects substantive conclusions. The Bartlett test now gives a *p*-value of 0.68 – much larger than that originally used, so no hint that the equal variances assumption may be wrong. For Tukey's HSD test all the *p*-values change as follows.

```
$material
                   diff       lwr        upr     p adj
Obsidian-Chert     -2.4 -30.54352  25.743516 0.9956328
Rhyolite-Chert     72.8  44.65648 100.943516 0.0000002
Wood-Chert         39.4  11.25648  67.543516 0.0031873
Rhyolite-Obsidian  75.2  47.80711 102.592887 0.0000001
Wood-Obsidian      41.8  14.40711  69.192887 0.0012254
Wood-Rhyolite     -33.4 -60.79289  -6.007113 0.0117777
```

Other than the obsidian-chert comparison the *p*-values are reduced; most noticeably the wood-chert comparison is now significant at the usual levels compared to its previous non-significance, and wood-rhyolite now differs significantly at the 5% level.

If the more radical, and possibly less justifiable, omission of all three outliers suggested by Figure 12.2 is contemplated there is relatively little change, the only one worth noting being the fact that the wood-rhyolite comparison is now only significant at the 9% level. This is not surprising since outliers that inflate the mean for rhyolite and deflate it for wood are now omitted, so the mean difference is reduced by about 10 mm to 20 mm.

This mixture of formal and informal analysis shows that there are highly significant differences between length for material types; you don't need the formal testing to arrive at this conclusion. An outlier apart you can accept, quite happily, a conclusion that chert and obsidian could be sampled from populations with the same mean length, and the boxplots show that their dispersion is also similar. All other pairwise comparisons, silicified wood and rhyolite apart, suggest highly *statistically significant* differences. Conclusions about the wood-rhyolite comparison are equivocal – it depends on the attitude towards outliers, but the evidence for a difference is not overwhelmingly significant whatever treatment is adopted. The more important issue is whether or not the observed differences matter much in terms of the archaeological aims driving the analysis. Is a difference of 20–30 mm in the mean length of silified wood and rhyolite of any consequence, regardless of statistical significance. If not, there is little point in worrying about the statistical significance and you are spared the effort of further data collection, necessary if differences of this magnitude are potentially important and you want to assert that they are 'real'.

12.4 Some omitted topics

As already noted, this chapter misses out a considerable amount. The focus has been on 'classical' statistical inference, the 'objectivity' and 'scientific rigor' of which first attracted the New Archaeologists and which, at an introductory level, is the approach most archaeologists will have been exposed to in texts on quantitative methodology currently in use.

Arguments about 'theory' in statistics have raged as much, and as fiercely, as they have have in archaeology. Competing 'theories', of which Bayesian inference is the most prominent, reject much of the conceptual machinery that underpins classical theory, in the way that probability is to be understood, for example, and how data should be interrogated and inferences drawn from them.

Bayesian thinking had its early advocates in archaeology (Cowgill, 1977b: 361–62; Orton, 1980: 220; Orton, 1992: 139) but, except at a basic level, very little was done about exploring the methodology. There is a good reason why this state of affairs arose, which is that for the practical application of Bayesian ideas the necessary computational power needed to be developed. This happened; Buck *et al.* (1996) was the first book-length treatment of the *Bayesian Approach to Interpreting Archaeological Data* (to give the book its full title) and remains the standard text. Nevertheless – with one major exception – Bayesian methods have not yet come to be routinely used in statistical analyses of archaeological data.

The major exception – and it is difficult to over-emphasize its importance – is in application to dating problems, and especially the calibration of radiocarbon dates and their interpretation. Without going into detail, the software in common use to provide these dates typically depends on Bayesian calculations, even if the user does not always appreciate this. As examples of what has been achieved, recent programmes of dating using Bayesian methods have produced important revisions of the previously accepted chronology for the British Neolithic and Anglo-Saxon periods (Whittle *et al.*, 2011; Bayliss *et al.*, 2013; see, also, Section 9.5). Despite its ubiquity there is the suspicion that not all 'consumers' of radiocarbon dates fully understand how they are to be interpreted, notwithstanding the literature that exists to explain this; it is something that might usefully be included in future texts on quantitative methodology.

Within the 'classical' paradigm more could have been said about significance testing in the context of regression models – noted briefly in Section 5.1.1 and 5.1.4 and with a more detailed illustration in the third example of Section 5.2 – and more complex ANOVA models (though these have had little archaeological use). Other modeling methodologies, within the class of generalized linear models that depend on inferential ideas and have attracted some archaeological use, have also not been discussed. These include log-linear models (Lewis, 1986; Shennan, 1997: 201-13; Baxter, 2003: 131–36) and logistic regression (Baxter, 2003: 60–62, 162–63).

Another topic omitted – on grounds of length rather than complexity – is that of non-parametric hypothesis testing methods. Such methods do not assume an underlying probability distribution for the sampled population, removing the dependency on the normality assumption. Although parametric tests, such as the t-test, are more powerful if the normality assumption is valid, non-parametric tests that can be used as an alternative can also be competitive in terms of power and would seem to be an attractive alternative when normality is in doubt.

Given this, I am a little surprised at the relative lack of space given to such methods in the standard quantitative archaeology textbooks. For those wishing to explore this further the `R` function `wilcox.test` does a similar job to `t-test` using the Wilcoxon one- or two-sample tests (the latter also known as the Mann-Whitney test). The Kruskal-Wallis test, `kruskal.test`, is the non-parametric analog of `oneway.test`.

12.5 Discussion

The question that motivated this chapter was 'how useful are the 'classical' methods of statistical inference for archaeological purposes?'. I take it as axiomatic that the ideas and methodologies of statistical inference are important. On the 'horses for courses' principle, however, it is not axiomatic that the methodologies as originally developed are equally applicable to different domains of study.

There are at least two aspects involved; one concerns the correct use of methodology and the other its usefulness. Much of the critical commentary that emerged in the 1970s and 1980s falls into the former category. Cowgill (1977b), for example, identified several areas of mis-use and mis-understandings and offered corrective advice. His appraisal of the value of statistical inferential methods in archaeology was less negative than that of, for example, Doran and Hodson (1975).

In a spirit of what might seem deliberate provocation, Cowgill used the adjectives 'mind-boggling' and 'riduculous' to characterize *some* uses of significance testing. He commented (Cowgill, 1977b: 365) that it 'seems so much more useful that it seems incredible that the estimation approach [i.e. confidence intervals] is not used more often [than significance] tests' – a still pertinent view, endorsed in this chapter. The emphasis on significance testing at the expense of estimation was attributed to 'tradition' in the social sciences, and an 'uncritical' acceptance of the hypothesis testing framework. This, also, remains pertinent[11].

The view that significance testing is frequently uninteresting and not of much use has already been expressed. It is 'usefulness' that I would use as the main, and pragmatic, criterion for judging the merits, in practice, of any particular method of data analysis. Judgments need to be divorced from any irritation with abuse of methodology, and not confused with 'philospophical stances' that involve the wholesale rejection of 'scientific' methodology.

Generalization should be approached with trepidation. Mine would be that, with some exceptions, I have been struck, whenever I have reviewed the literature, by the paucity of widespread and what I'd regard as convincing uses of the methods discussed in this chapter[12]. Some methods have been used hardly at all, or not to the extent you might assume from their textbook treatments.

For example, VanPool and Leonard (2010: Chapter 10) observe that ANOVA,

[11] As an aside, at the time of writing, the journal *Basic and Applied Social Psychology* has just 'banned' the use of the null hypothesis significance test (and confidence intervals) from its pages on the grounds that they are 'invalid'. What is meant by 'invalid' is not very clear; it seems to stem more from concerns about the misuse and misinterpretation of significance tests, a concern many statisticians share.

[12] An exception that springs to mind is sampling theory, where random sampling (of regions using test-pits, for example) leads naturally to the use of standard inferential ideas. This is not typical of the manner in which archaeological 'samples' are usually acquired. The subject is quite a specialized one; Orton (2000) is a thorough and accessible account for archaeologists.

one of the ‘most powerful tools in the statistician’s toolkit’, ‘hasn’t been applied as widely as it deserves to be in archaeological analysis’. Two-sample *t*-tests are simpler but still not extensively used; graphical methods will often make it clear whether there are substantively important differences or not, obviating the need for formal tests. Should the data merit further scrutiny after graphical analysis then, even if one has no qualms about the assumptions involved, sample-size effect need to be borne in mind. With large enough samples any difference, however small, will be found to be significant. The main merit of testing is for small samples where a non-significant result guards against reading too much into apparent differences.

None of these observations are new; commentators such as Cowgill (1997b) were saying this kind of thing from an early stage. I would add that it is probably difficult to put together a convincing collection of case-studies based on the *t*-test that lead to insights not more readily attainable by other means. The same is true, only more so, of one-sample tests. It is difficult to think of examples, particularly those in textbooks, that are other than illustrative and artificial.

Chi-squared tests for no association in contingency tables have probably been more widely used than tests of means. Analyses often do not go beyond reporting a statistically significant association or its lack. If detection of ‘significance’ is the sole reason for analyzing a table it is tempting to suggest that much can be achieved by judicious tabular inspection. This includes the scaling of rows or columns to percentages, with the ordering chosen to highlight similarities or differences where there is not a natural ordering. This is, of course, done, but not necessarily consistently. The suspicion exists that, despite the ubiquity of tabular presentation in archaeology, the widespread use of ill-chosen `Excel` bar- and pie-charts in preference to such direct interpretation (Sections 4.2 and 4.3) testifies to a certain discomfort with the latter.

The last few paragraphs express a view about the value of the standard and ‘classical’ methods of statistical methods for archaeological purposes. The importance of the theory *per se* is unquestionable, as is the beneficial impact it has had in many areas of practical application. Whether archaeology is one of these areas is the interesting question.

References

Aitchison, J. 1986: *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Albarella, U., Tagliacozzo, A, Dobney, K. and Rowley-Conwy, P. 2006: Pig hunting and husbandry in prehistoric Italy: a contribution to the domestication debate. *Proceedings of the Prehistoric Society* **72**, 193–227.

Alberti, G. 2013: An `R` script to facilitate Correspondence Analysis. A guide to the use and the interpretation of results from an archaeological perspective. *Archeologia e Calcolatori* **24**, 25–53.

Anderson, T.W. 1958: *An Introduction to Multivariate Analysis.* New York: Wiley

Bailey, G.N., Carter, P.L., Gamble, C.S., and Higgs, H.P. 1983: Asprochalika and Kastritsa: further investigations of palaeolithic settlement in Epirus (North-west Greece). *Proceedings of the Prehistoric Society* **49**, 15–42.

Banfield, J.D. and Raftery, A.E. 1993: Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Barnatt, J. and Moir, G. 1984: Stone circles and megalithic mathematics. *Proceedings of the Prehistoric Society* **50**, 197–216.

Baxter, M.J. 1988: The morphology and evolution of post-medieval wine bottles revisited. *Science and Archaeology* **30**, 10–14.

Baxter, M.J. 1989: Multivariate analysis of data on glass compositions: a methodological note. *Archaeometry* **31**,45–53.

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

Baxter, M.J. 1994b: Stepwise discriminant analysis in archaeometry: a critique. Journal of Archaeological Science **21**, 659–666.

Baxter, M.J. 1995: Standardisation and transformation in principal component analysis, with applications to archaeometry. *Applied Statistics* **44**, 513–527.

Baxter, M.J. 1999: Detecting multivariate outliers in artefact compositional data. *Archaeometry* **41**, 321–338.

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

Baxter, M.J. 2008: Cluster analysis. In Liritzis I. (ed.), *New Technologies in the Archaeognostic Sciences*, Gutenberg Press, Athens, Greece, 445–481. (Paper and book in Greek)

Baxter, M.J. 2009: Archaeological data analysis and fuzzy clustering. *Archaeometry* **51**, 1035–1054.

Baxter, M.J. 2014a: Anglo-Saxon Chronology I - the male graves: A commentary on Chapter 6 of *Anglo-Saxon Graves and Grave Goods of the 6th and 7th centuries AD: A Chronological Framework*
https://nottinghamtrent.academia.edu/MikeBaxter/Papers

Baxter, M.J. 2014b: Anglo-Saxon Chronology II - the female graves: A commentary on Chapter 7 of *Anglo-Saxon Graves and Grave Goods of the 6th and 7th centuries AD: A Chronological Framework*
https://nottinghamtrent.academia.edu/MikeBaxter/Papers

Baxter, M.J. 2015: *Exploratory Multivariate Analysis in Archaeology.* New York: Percheron Press. (This is a reprint of Baxter (1994a) with a new introduction,)

Baxter, M.J. and Beardah, C.C. 1996: Beyond the histogram: improved approaches to simple data display in archaeology using kernel density estimates. *Archeologia e Calcolatori* **7**, 397–408.

Baxter, M.J. and Buck, C.E. 2000: Data handling and statistical analysis. In Ciliberto, E. and Spoto, G. (eds.), *Modern Analytical Methods in Art and Archaeology.* New York: Wiley, 681–746.

Baxter, M.J. and Cool, H.E.M. 2008: Notes on the statistical analysis of some loomweights from Pompeii. *Archeologia e Calcolatori* **19**, 49–66.

Baxter, M.J. and Cool, H.E.M. 2010a: Detecting modes in low-dimensional archaeological data. *Journal of Archaeological Science* **37**, 2379–2385.

Baxter, M.J. and Cool, H.E.M. 2010b: Correspondence Analysis in R for archaeologists: an educational account. *Archeologia e Calcolatori* **21**, 211–28.

Baxter, M.J. and Freestone, I.C. 2006: Log-ratio compositional data analysis in archaeometry. *Archaeometry* **48**, 511–531.

Baxter, M.J. and Jackson, C.M. 2001: Variable selection in artefact compositional studies. *Archaeometry* **43**, 253–268.

Baxter, M.J., Cool, H.E.M. and Anderson, M.A. 2010: Statistical analysis of some loomweights from Pompeii: a postscript. *Archeologia e Calcolatori* **21**, 185–200.

Bayliss, A., Hines, J., Høilund Nielsen, K.H., McCormac, G. and Scull, C. 2013: *Anglo-Saxon Graves and Grave Goods of the 6th and 7th Centuries AD: A Chronological Framework.* London: Society for Medieval Archaeology.

Bell, S. and Croson, C. 1998: Artificial neural networks as a tool for archaeological data analysis. *Archaeometry* **40**, 139–151.

Bieber, A.M., Brooks, D.W., Harbottle, G. and Sayre, E.V. 1976: Application of multivariate techniques to analytical data on Aegean ceramics. *Archaeometry* **18**, 59–74.

Binford, L.R.B. and Binford, S.R. 1966: A preliminary analysis of functional variability in the Mousterian of Levallois facies. *American Anthropologist* **68**, 239–295.

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

Bowman, A.W. and Azzalini, A. 1997: *Applied Smoothing Techniques for Data Analysis.* Oxford: Clarendon Press.

Buck, C.E., Cavanagh, W.G. and Litton, C.D. 1996: *Bayesian Approach to Interpreting Archaeological Data.* Chichester: Wiley.

Cattell, R.B. 1978: *The Scientific Use of Factor Analysis in Behavioral and Life Sciences.* New York: Plenum Press.

Chatfield, C. and Collins, A.J. 1980: *Introduction to Multivariate Analysis.* London: Chapman and Hall.

Cleveland, W.S. 1979: Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Cook, R.D. and Weisberg, S. 1982: *Residuals and Influence in Regression.* London: Chapman and Hall.

Cool, H.E.M. 1983: *Roman Personal Ornaments Made of Metal, Excluding Brooches, from Southern Britain.* Cardiff: University of Wales Ph.D, Thesis.

Cool, H.E.M. 1990: Roman metal hair pins from southern Britain. *Archaeological Journal* **147**, 148–182.

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

Cool, H.E.M. and Price, J. 1995: *Roman Vessel Glass from Excavations in Colchester, 1971-85.* Colchester: Colchester Archaeological Trust.

Cormack, R.M. 1971: A review of classification. *Journal of the Royal Statistical Society A* **134**, 321–367.

Cowgill, G.L. 1977a: Review of *Mathematics and Computers in Archaeology* by J.E. Doran and F.R. Hodson. *American Antiquity* **42**, 126–29.

Cowgill, G.L. 1977b: The trouble with significance tests and what we can do about it. *American Antiquity* **42**, 350–68.

Cox, G.A. and Gillies, K.J.S. 1986: The X-ray fluorescence analysis of medieval durable blue soda glass from York Minster. *Archaeometry* **28**, 57–68.

Cox, D.R. and Hinkley, D.V. 1974: *Theoretical Statistics.* London: Chapman and Hall.

Cummins, W.A. 1980: Stone axes as a guide to neolithic communications and boundaries in England and Wales. *Proceedings of the Prehistoric Society* **46**, 45–60.

Dalgaard, P. 2008: *Introductory Statistics with* R*:, Second Edition.* New York: Springer.

Doran, J.E. and Hodson, F.R. 1975: *Mathematics and Computers in Archaeology.* Edinburgh: Edinburgh University Press.

Drennan, R.D. 1996: *Statistics for Archaeologists.* New York: Plenum Press.

Drennan, R.D. 2009: *Statistics for Archaeologists, Second Edition.* New York: Springer.

Everitt, B.S. and Dunn, G. 2001: *Applied Multivariate Data Analysis: Second Edition.* London: Arnold.

Everitt, B.S., Landau, S,, Leese, M. and Stahl, D. 2011: *Cluster Analysis: Fifth Edition.* Chichester, Wiley.

Faraway, J.J. 2005: *Linear Models with* R. Boca Raton: Chapman and Hall/CRC.

Faraway, J.J. 2006: *Extending the Linear Model with* R. Boca Raton: Chapman and Hall/CRC.

Fisher, R.A. 1936: The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Fletcher, M. and Lock, G. 2005: *Digging Numbers: Elementary Statistics for Archaeologists.* Oxford: Oxford University Committee for Archaeology,

Freestone, I. C., Gorin-Rosen, Y. and Hughes, M. J. 2000: Composition of primary glass from Israel. In Nenna M.-D. (ed.), *La Route du Verre: Ateliers Primaires et Secondaires de Verriers du Second Millénaire av. J.-C. au Moyen-Age.* Travaux de la Maison de l'Orient Méditerranéen no. 33, 65–104.

Gliozzo E., Fortina C., Turbanti Memmi I., Turchiano M. and Volpe G. 2005: Cooking and painted ware from San Giusto (Lucera, Foggia): the production cycle, from the supply of raw materials to the commercialisation of products. *Archaeometry* **47**, 13–29.

Gliozzo E., Leone D., Origlia F., Turbanti Memmi I. and Volpe G. 2010: Archaeometric characterisation of coarse and painted fine ware from *Posta Crusta* (Foggia, Italy): technology and provenance. *Archaeological and Anthropological Sciences* **2**, 175–189.

Gliozzo, E., Turchiano, M., Lombardi, M., Turbanti Memmo, I. and Baxter, M.J. 2013: North Apulian coarse wares and fine painted wares: A reappraisal according to new data from *Herdonia* and *Canusium. Archaeometry* **55**, 423–448.

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

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

Greenacre, M.J. 2007: *Correspondence Analysis in Practice: Second Edition.* Boca Raton (FL): Chapman and Hall/CRC.

Greenacre, M.J. 1993: *Correspondence Analysis in Practice.* London: Academic Press.

Hastie, T., Tibshirani, R. and Freedman, J, 2009: *The Elements of Statistical learning: Second Edition.* New York; Springer.

Hesse, R. 2011: Reconsidering animal husbandry and diet in the northwest provinces. *Journal of Roman Archaeology* **24**, 215–248.

Hill, M.O. 1974: Correspondence analysis: a neglected multivariate technique. *Applied Statistics* **23**, 340–354.

Hintze, J.L. and Nelson, R.D. 1998: Violin plots: A boxlot-density trace synergism. *The American Statistician* **52**, 181–184.

Hodder, I. and Orton, C. 1976: *Spatial Analysis in Archaeology.* Cambridge: Cambridge University Press.

Hodson, F.R. 1969: Searching for structure within multivariate archaeological data. *World Archaeology* **1**, 90–105.

Hodson, F.R. 1970: Cluster analysis and archaeology: some new developments and applications. *World Archaeology* **1**, 299–320.

Hume, I.N. 1970: *A Guide to Artifacts of Colonial America.* New York: Random House.

Jensen, C.K. and Høilund Nielsen, K. 1997: Burial data and correspondence analysis. In Jensen, C.K. and Høilund Nielsen, K. (eds.), *Burials & Society: the Chronological and Social Analysis of Archaeological Burial Data.* Aarhus: Aarhus University Press, 29–61.

Jolliffe, I.T. 2002: *Principal Component Analysis: Second Edition.* New York: Springer-Verlag.

King, A.C. 1999: Diet in the Roman world: a regional inter-site comparison of the mammal bones. *Journal of Roman Archaeology* **24**, 168–202.

Krzanowski, W.J. 1988: *Principles of Multivariate Analysis.* Oxford: Clarendon Press.

Krzanowski, W.J. and Marriott, F.H.C. 1995: *Multivariate Analysis: Classification, Covariance Structures and Repeated Measurements.* London: Edward Arnold.

Lake, M. 2014: Trends in archaeological simulation. *Journal of Archaeological Method and Theory* **21**, 251–287.

Lawley, D.N. and Maxwell, A.E. 1963: *Factor Analysis as a Statistical Method.* London: Butterworths

Lewis, B. 1986: The analysis of contingency tables in archaeology. In Schiffer, M.B. (ed.), *Advances in Archaeological method and Theory 9.* New York: Academic Press, 277–310.

Lockyear, K. 2013: Applying bootstrapped correspondence analysis to archaeological data. *Journal of Archaeological Science* **40**, 4744–4753.

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

Madsen, T. 1988b: Multivariate distances and archaeology. In Madsen, T. (ed.), *Multivariate Archaeology.* Aarhus: Aarhus University Press, 7–27.

McClellan, T.L. 1979: Chronology of the 'Philistine' burials at Tell el-Farah (South). *Journal of Field Archaeology* **6**, 57–73.

Meighan, C.W. 1959: A new method for the seriation of archaeological collections. *American Antiquity* **25**, 203–211.

Mellars, P.A. and Wilkinson, M.R. 1980: Fish otoliths as indicators of seasonality in prehistoric shell middens: The evidence for Oronsay (Inner Hebrides). *Proceedings of the Prehistoric Society* **46**, 19–44.

Miller, A.J. 1990: *Subset Selection in Regression.* London: Chapman and Hall.

Moreno-Garcia, M., Orton, C. and Rackham, J. 1996: A new statistical tool for comparing animal bone assemblages. *Journal of Archaeological Science* **23**, 437–453.

Morris, E.L. 1994: Production and distribution of salt in Iron Age Britain: A review. *Proceedings of the Prehistoric Society* **60**, 371–393.

Morrison, D.F. 1967: *Multivariate Statistical Methods.* New York: McGraw-Hill.

Murrell, P. 2011: R *Graphics: Second Edition.* Boca Raton: Chapman and Hall/CRC.

O'Hare, G.B. 1990: A preliminary study of polished stone artefacts in prehistoric southern Italy, *Proceedings of the Prehistoric Society* **56**, 125–132.

Orton, C. 1975: Quantitative pottery studies: Some progress, problems and prospects. *Science and Archaeology* **16**, 30–35.

Orton, C. 1980: *Mathematics in Archaeology.* London: Collins.

Orton, C. 1992: Quantitative methods in the 1990s. In Lock, G. and Moffett, J. (eds.), *Computer Applications and Quantitative Methods in Archaeology 1991.* Oxford: Tempus Reparatum, 137–40.

Orton, C. 2000: *Sampling in Archaeology.* Cambridge: Cambridge University Press.

Papageorgiou, I., Baxter, M.J. and Cau, M.A. 2001: Model-based cluster analysis of artefact compositional data. *Archaeometry* **43**, 571–588.

Pawlowsky-Glahn, V. and Buccianti. A. (eds.) 2011: *Compositional Data Analysis: Theory and Applications.* Chichester: Wiley.

Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. 2015: *Modeling and Analysis of Compositional Data.* Chichester: Wiley.

Peeples, M.A., and Schachner, G. 2012: Refining correspondence analysis-based ceramic seriation of regional data sets. *Journal of Archaeological Science* **39**, 2818–2827.

Rao, C.R. 1948: The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society B* **10**, 159–203.

Read, D.W. 1989: Statistical method and reasoning in archaeological research: a review of praxis and promise. *Journal of Quantitative Archaeology* **1**, 5–78.

Ringrose, T. J. 1992: Bootstrapping and correspondence analysis in archaeology. *Journal of Archaeological Science* **19**, 615–629.

Robertson, W.S. 1976: A quantitative morphological study of the evolution of some post-medieval wine bottles. *Science and Archaeology* **17**, 13–20.

Sarkar, D. 2008: *Lattice: Multivariate Data Visualization with* R. New York: Springer.

Seber, G.A.F. 1984: *Multivariate Observations.* New York: Wiley.

Shennan, S. 1988: *Quantifying Archaeology.* Edinburgh: Edinburgh University Press.

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

Silverman, B.W. 1986: *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Simonoff, J.J. 1996: *Smoothing Methods in Statistics.* New York: Springer.

Spearman, C. 1904: "General intelligence," objectively determined and measured. *American Journal of Psychology* **15**, 201–293.

Steele, T.E. and Weaver, T.D. 2002: The modified triangular graph: A refined method for comparing mortality profiles in archaeological samples, *Journal of Archaeological Science* **29**, 317–322.

Stiner, M.C. 1990: The use of mortality patterns in archaeological studies of hominid predatory adaptations. *Journal of Anthropological Archaeology* **9**, 305-351.

Stos-Gale, Z.A. , Gale, N.H. and Annetts, N. 1996: Lead isotope data from the Isotrace Laboratory, Oxford: *Archaeometry* data base 3, ores from the Aegean, part 1. *Archaeometry* **38**, 381–390.

Therneau, T.M., and Atkinson, E.J. 1997: An introduction to recursive partitioning using the RPART routines, Technical Report, Mayo Foundation.

Thom, A. 1967: *Megalithic Sites in Britain.* Oxford: Clarendon Press.

Truncer, J,, Glascock, M.D. and Neff, H. 1998: Steatite source characterization in eastern North America: New results using instrumental neutron activation analysis. *Archaeometry* **40**, 23–44.

Tubb, A., Parker, A.J. and Nickless, G. 1980: The analysis of Romano-British pottery by atomic absorption spectrophotometry. *Archaeometry* **22**, 153–171.

Valero-Mora, P.M. amd Ledesma, R.D. 2012. Graphical User Interfaces for R. *Journal of Statistical Software* **49**, Issue 1.

van den Boogaart, K.G. and Tolosana-Delgado, R. 2013. *Analyzing Compositional Data with* R. New York: Springer.

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

Vierra, R.K. and Carlson, D.L. 1981: Factor analysis, random data, and patterned results. *American Antiquity* **46**, 272–283.

Wainwright, G.J. 1984: The pressure of the past. *Proceedings of the Prehistoric Society* **50**, 1–22.

Wand, M.P. and Jones, M.C. 1995: *Kernel Smoothing.* London: Chapman and Hall.

Westcott, K.L. and Brandon, R.J. (eds.) 2000: *Practical Applications of GIS for Archaeologists: A Predictive Modelling Kit.* London: Taylor and Francis.

Whallon, R. 1987: Simple statistics. In Aldenderfer, M.S. (ed.), *Quantitative Research in Archaeology.* Newbury Park, California: Sage Publications, 135–150.

Whittle, A., Healy, F. and Bayliss, A. 2011: *Gathering Time: Dating the Early Neolithic Enclosures of Southern Britain and Ireland.* Oxford: Oxbow Books.

Wickham, H. 2009: *ggplot2.* New York: Springer.

Widaman, K.F. 1993: Common factor analysis versus principal component analysis: Differential bias in representing model parameters. *Multivariate Behavioral Research* **28**, 263–322.

Appendix A

Getting R, getting started

A.1 Finding R

Either google CRAN R (the **C**omprehensive **R** **A**rchive **N**etwork) or use

```
http://cran.r-project.org/
```

which is where to begin. You are directed to pages that tell you how to install R on various platforms, and more information, if needed, is provided in the FAQs. R is updated frequently.

Documentation on R is comprehensive and much of it is free. It is well worth looking at what is available in CRAN at an early stage.

A.2 Data entry

For other than very small data sets it is best to import data from an external source. This can be done in different ways. Although not the preferred method of the developers of R, many users may find it simplest, if starting from scratch, to create and import an Excel file.

Create the data file; it is assumed below that headers naming the columns are given. Spaces in headers must be avoided (and if other illegal characters are used an error message in R will inform you). Next, highlight the data you want to import; copy it (to the clipboard) and go into R. The data file is named in R and for illustration the data of Table B.1 is used, which will be named tubb in R. Type

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

and type tubb to see the result. Here <- is the assigment operator, and note that clipboard must be enclosed in quotation marks or an error message results.

If data are missing R requires that the offending cell be filled with NA. To use read.table a rectangular table of data is expected. Commands are preceded by the R prompt >.

It is best to keep headers informative but short (in writing-up an analysis or captioning a figure a key can always be provided). Headers beginning with a number are allowed, but column names in R may not be quite as you expect.

The writers of the R manual for data import/export prefer you to write the Excel file to a Tab or comma-separated file and use read.delim or read.csv. For most of the data used in these notes row names were not supplied, and the default then produces case numbers as row names. For alternatives to this see ?read.table in R.

Once entered into R you may need to work with subsets of the data. Selecting subsets is discussed in Section 2.6.

A.3 Packages

Packages are collections of *functions* that, together with *arguments* provided to them, control the analyses undertaken[1]. Some packages are loaded automatically with R and the functions in them are immediately accessible. Others are bundled with R and must be loaded before use. Yet others need to be imported before they can be loaded.

For illustration the bundled MASS package, associated with the book *Modern Applied Statistics with S* (Venables and Ripley, 2002), is used. In the following code comments follow #.

```
library(MASS)            # loads MASS
library(help = MASS)     # lists available functions
?kde2d                   # information on the function kde2
kde2d                    # prints source code for kde2
```

A list of available functions is provided by library(help = MASS) and ? lists the documentation for the function specified, in the example kde2d for 2-dimensional kernel density estimation. It is sometimes useful to look at the code to see what is going on. It can be edited to suit individual requirements. The '?' facility does not always work as you wish. For example the function biplot is contained in the stats package which is automatically loaded. If, however, ?biplot is typed then the following is obtained.

```
function (x, ...)
```

[1] There are numerous examples in the text. Code can be typed in directly but it is often more convenient to construct a function first. Further detail is given in Section 3.2.1.

```
UseMethod("biplot")
<environment: namespace:stats>
```

That is, the code is invisible; typing `stats:::biplot.default` should reveal the code. More generally, this requires the package where the function is located and the function name to be specified as indicated.

There are an enormous number of user-contributed packages available. These need to be imported before they can be used and this is done in two stages. In `R` from the `Packages` menu select `Set CRAN mirror` to choose a site to import from then, from the same menu, select `Install package(s)`. The package then needs to be loaded, using the `library` function as shown in the example above.

Appendix B

Data sets

Romano-British pottery compositions

Al	Fe	Mg	Ca	Na	K	Ti	Mn	Ba	Region
18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015	1
16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018	1
18.2	7.64	1.82	0.77	0.40	3.07	0.98	0.087	0.014	1
17.4	7.48	1.71	1.01	0.40	3.16	0.03	0.084	0.017	1
16.9	7.29	1.56	0.76	0.40	3.05	1.00	0.063	0.019	1
17.8	7.24	1.83	0.92	0.43	3.12	0.93	0.061	0.019	1
18.8	7.45	2.06	0.87	0.25	3.26	0.98	0.072	0.017	1
16.5	7.05	1.81	1.73	0.33	3.20	0.95	0.066	0.019	1
18.0	7.42	2.06	1.00	0.28	3.37	0.96	0.072	0.017	1
15.8	7.15	1.62	0.71	0.38	3.25	0.93	0.062	0.017	1
14.6	6.87	1.67	0.76	0.33	3.06	0.91	0.055	0.012	1
13.7	5.83	1.50	0.66	0.13	2.25	0.75	0.034	0.012	1
14.6	6.76	1.63	1.48	0.20	3.02	0.87	0.055	0.016	1
14.8	7.07	1.62	1.44	0.24	3.03	0.86	0.080	0.016	1
17.1	7.79	1.99	0.83	0.46	3.13	0.93	0.090	0.020	1
16.8	7.86	1.86	0.84	0.46	2.93	0.94	0.094	0.020	1
15.8	7.65	1.94	0.81	0.83	3.33	0.96	0.112	0.019	1
18.6	7.85	2.33	0.87	0.38	3.17	0.98	0.081	0.018	1
16.9	7.87	1.83	1.31	0.53	3.09	0.95	0.092	0.023	1
18.9	7.58	2.05	0.83	0.13	3.29	0.98	0.072	0.015	1
18.0	7.50	1.94	0.69	0.12	3.14	0.93	0.035	0.017	1
17.8	7.28	1.92	0.81	0.18	3.15	0.90	0.067	0.017	1
14.4	7.00	4.30	0.15	0.51	4.25	0.79	0.160	0.019	2
13.8	7.08	3.43	0.12	0.17	4.14	0.77	0.144	0.020	2
14.6	7.09	3.88	0.13	0.20	4.36	0.81	0.124	0.019	2
11.5	6.37	5.64	0.16	0.14	3.89	0.69	0.087	0.009	2
13.8	7.06	5.34	0.20	0.20	4.31	0.71	0.101	0.021	2
10.9	6.26	3.47	0.17	0.22	3.40	0.66	0.109	0.010	2
10.1	4.26	4.26	0.20	0.18	3.32	0.59	0.149	0.017	2
11.6	5.78	5.91	0.18	0.16	3.70	0.65	0.082	0.015	2
11.1	5.49	4.52	0.29	0.30	4.03	0.63	0.080	0.016	2
13.4	6.92	7.23	0.28	0.20	4.54	0.69	0.163	0.017	2
12.4	6.13	5.69	0.22	0.54	4.65	0.70	0.159	0.015	2
13.1	6.64	5.51	0.31	0.24	4.89	0.72	0.094	0.017	2
12.7	6.69	4.45	0.20	0.22	4.70	0.73	0.394	0.024	2
12.5	6.44	3.94	0.22	0.23	0.81	0.75	0.177	0.019	2
11.6	5.39	3.77	0.29	0.06	4.51	0.56	0.110	0.015	2
11.8	5.44	3.94	0.30	0.04	4.64	0.59	0.085	0.013	2
18.3	1.28	0.67	0.03	0.03	1.96	0.65	0.001	0.014	3
15.8	2.39	0.63	0.01	0.04	1.94	1.29	0.001	0.014	3
18.0	1.50	0.67	0.01	0.06	2.11	0.92	0.001	0.016	3
18.0	1.88	0.68	0.01	0.04	2.00	1.11	0.006	0.022	3
20.8	1.51	0.72	0.07	0.10	2.37	1.26	0.002	0.016	3
17.7	1.12	0.56	0.06	0.06	2.06	0.79	0.001	0.013	3
18.3	1.14	0.67	0.06	0.05	2.11	0.89	0.006	0.019	3
16.7	0.92	0.53	0.01	0.05	1.76	0.91	0.004	0.013	3
14.8	2.74	0.67	0.03	0.05	2.15	1.34	0.003	0.015	3
19.1	1.64	0.60	0.10	0.03	1.75	1.04	0.007	0.018	3

Table B.1: *Data on the chemical composition of a sample of Romano-British pottery and region of origin (Source: Tubb et al., 1980.)*

Artefact counts in Early Iron Age tombs.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
0	0	0	0	0	0	2	2	1	1	0	0	0	1	5	3
0	0	0	0	0	0	1	0	0	1	1	7	0	0	0	0
0	2	0	0	0	1	0	1	0	0	0	0	0	0	0	0
0	8	0	2	3	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
0	4	0	0	2	0	0	1	0	0	0	0	1	0	2	0
0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
0	7	6	1	2	0	0	0	0	1	3	3	0	0	1	1
3	6	1	1	3	3	2	6	2	2	12	5	2	1	2	1
0	8	0	4	4	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	2	0	0	1	0	0	0	0	0
0	2	12	0	0	0	1	0	0	0	1	0	0	0	0	0
0	1	0	0	0	0	0	2	0	0	2	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1	1	0	7	3	4	2
1	1	0	0	0	1	2	0	1	0	4	1	9	4	0	0
0	0	0	0	0	0	0	0	0	0	7	0	0	0	1	0
2	2	0	0	0	1	0	0	0	0	3	0	0	0	0	0
0	0	0	0	0	0	0	1	3	1	0	2	3	2	7	0
0	5	0	0	2	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	2	0	1	0	4	0
1	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0
0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0
0	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0
1	8	26	3	1	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	2	1	2	0	0	13	1	4	4
1	4	0	0	0	2	0	1	2	4	14	2	1	3	0	0
1	1	13	1	0	2	0	0	0	0	0	0	0	0	0	0
0	3	14	0	0	0	1	0	0	0	1	0	0	0	0	0
1	6	0	0	0	1	1	1	0	0	4	0	0	0	1	1
10	14	2	0	0	11	6	16	3	5	24	27	4	5	2	2
1	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	1	5	2	0	0	0	0
0	4	0	0	0	1	0	0	0	0	4	1	2	1	7	0
4	5	0	0	0	1	0	1	0	0	3	0	0	0	0	0
0	0	0	0	0	0	1	0	1	1	1	0	0	2	2	1
0	0	0	0	0	0	0	0	3	0	0	1	1	4	0	1
0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
2	13	0	2	2	5	7	5	3	5	15	5	0	4	6	3
0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1
5	9	3	4	0	5	7	7	11	9	13	13	14	8	11	12
0	1	0	0	0	0	0	0	0	1	0	0	8	0	1	1
0	0	0	0	0	0	1	2	1	2	0	0	0	0	0	0
0	0	0	0	0	0	1	5	0	1	0	3	0	0	0	0
0	0	1	0	0	1	0	2	1	1	2	2	0	0	0	0
0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
0	5	0	0	1	2	0	0	0	0	2	0	0	0	2	0
1	3	1	0	1	5	2	5	0	0	9	2	0	0	0	1
0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
1	7	0	1	1	3	1	4	3	3	9	2	1	2	4	4
3	12	0	0	2	1	0	7	1	1	1	2	0	0	0	0
0	1	0	0	0	1	1	2	0	0	3	0	0	0	0	0

Table B.2: *Counts of pottery types in Early Iron Age tomb assemblages. Column headings identify tombs; rows to pottery types; and table entries to counts (Source: McClellan, 1979).*

Loomweights from Insula VI.1, Pompeii, I

height	topmax	topmin	bottommax	bottommin	weight	volume
93	23	23	35	35	222	475974
69	30	30	45	40	204	548550
98	30	30	40	40	260	725200
100	20	20	50	50	343	780000
107	25	20	55	55	283	1019175
70	38	25	48	35	173	545300
102	28	25	48	40	276	771120
98	20	20	40	35	267	499800
100	32	30	45	40	266	815000
100	25	25	40	40	257	645000
70	35	25	40	30	140	434000
99	30	30	57	50	318	1060290
88	30	28	48	48	242	798336
95	30	30	55	52	337	1019350
118	26	26	50	40	313	907656
127	47	47	80	72	737	2931414
97	30	25	40	40	306	669300
89	30	30	40	40	269	658600
65	30	30	35	35	137	412750
110	25	25	50	46	350	907500
71	21	21	50	50	166	566722
89	30	26	50	43	258	752050
66	30	30	38	38	170	459888
68	30	26	46	33	152	461176
100	27	27	50	50	324	915800
107	27	27	50	37	291	803249
64	32	24	48	27	118	393216
94	35	30	53	46	355	956544
103	25	20	45	45	297	728725
103	30	30	45	45	371	880650
72	20	20	45	45	159	478800
104	25	25	50	35	281	715000
107	26	26	45	40	286	766334
107	25	20	50	40	261	749000
94	30	30	45	40	273	747300
119	30	30	70	70	618	1880200
87	30	30	50	50	260	852600
107	30	30	55	55	342	1193050
64	35	20	50	30	139	412800
81	27	26	40	38	164	527310
69	35	30	42	42	171	576702
84	30	20	46	46	187	649488
64	50	23	50	23	163	441600
72	31	27	45	29	143	460656
70	33	31	51	46	316	688590
110	38	25	42	38	288	834460
104	39	33	48	46	327	1078272
87	34	34	46	17	202	523566
122	26	26	44	42	356	888648
128	25	25	56	47	420	1163392
100	26	26	42	39	256	673400
95	28	25	55	55	306	984675
99	28	28	46	45	233	817344
101	26	26	50	44	290	827796
71	33	25	43	34	112	480741
55	32	26	40	22	117	284240
100	28	28	30	30	239	504800
88	30	27	35	33	201	516120
86	29	26	33	26	163	415896
97	14	14	37	27	247	318742
74	23	23	38	38	142	421356
58	28	28	40	40	133	406464
56	44	24	44	24	119	354816
59	37	33	44	32	160	465746
57	38	21	46	17	107	272004
79	35	35	48	48	199	823022
58	35	35	34	34	136	414236
57	48	48	48	48	166	787968
97	22	22	32	32	193	429128
104	32	32	53	53	348	1150032
105	33	33	59	59	467	1368570

Table B.3: *Loomweight dimensions from Pompeii. Continued – see Table B.4 for details.*

Loomweights from Insula VI.1, Pompeii, II

height	topmax	topmin	bottommax	bottommin	weight	volume
81	27	27	45	45	202	642978
100	25	25	52	45	301	835500
61	32	27	41	32	146	395463
69	39	24	35	25	96	375153
95	30	30	36	33	202	593370
87	25	21	32	32	178	397590
96	23	13	44	37	228	506592
72	19	14	38	34	288	309168
116	36	36	57	55	484	1495704
111	31	31	52	52	383	1171494
67	26	21	36	36	134	360192
98	26	26	52	49	304	889252
118	30	30	50	41	391	1018340
112	23	17	58	47	328	929712
111	30	30	48	41	349	933066
101	20	20	54	54	374	887992
72	30	25	44	42	166	544032
78	32	26	55	47	235	761904
72	27	27	48	43	175	579096
91	23	23	45	45	198	653198
112	33	29	52	38	321	966336
116	29	24	58	52	420	1197584
83	24	21	44	34	159	476420
98	27	27	51	46	300	859362
70	35	25	45	28	133	446250
61	47	25	54	25	166	462075
89	20	20	49	42	211	599504
100	26	26	56	55	322	1039800
64	29	29	39	37	136	433408
68	24	21	42	36	182	392904
87	29	28	52	46	280	800226
114	24	24	48	34	296	727776
82	22	22	40	40	180	486096
103	23	23	48	40	255	712966
74	25	22	37	32	138	376068
36	12	9	20	20	15	51696
78	30	26	40	40	106	546000
62	30	23	40	30	111	347200
91	35	30	50	42	284	843570
87	35	35	53	53	343	1024686
118	25	25	50	50	340	1032500
97	28	28	45	45	314	789386
87	25	25	44	44	207	637014
112	29	29	53	53	373	1161888
108	28	28	55	55	383	1155384
99	31	31	51	51	296	1018314
116	22	22	42	42	290	735904
69	17	17	39	39	113	341274
88	20	20	37	37	180	441584
99	31	31	43	43	230	820314
79	21	21	34	34	143	365138
84	14	14	42	42	240	428064
99	23	23	58	58	388	1034946
60	33	33	39	39	133	467640
90	28	26	44	44	226	693360
91	32	32	52	52	351	981344
120	27	27	57	57	449	1324080
98	31	31	50	50	303	982156
100	24	24	40	40	216	627200
91	35	35	50	50	304	996450
114	36	36	66	66	564	1830384
87	28	28	36	36	214	537312
60	35	35	48	48	191	625080
88	25	25	39	39	189	549296
69	38	38	45	45	158	714702
97	30	30	50	50	298	950600

Table B.4: *Data on loomweight dimensions from excavations at Pompeii, Insula VI.1. Unpublished, but see Baxter and Cool (2008, 2010) and Baxter et al. (2010) for previous analyses of these data. The first five variables are measured as mm; weight as g. The terms 'max' and 'min' refer to the maximum and minimum side lengths of the rectangular tops and bottoms of the loomweights.*

Post-medieval wine bottle dimensions

Date	Height	NH	BH	Width	Base	Kick
1652	233	140	93	137	67	8
1661	188	83	105	133	37	6
1680	150	66	84	110	60	11
1687	178	78	100	141	70	15
1688	150	65	85	134	77	7
1698	135	51	84	132	74	7
1700	170	77	94	146	96	23
1704	146	63	83	120	86	20
1708	156	62	94	150	100	33
1713	174	77	97	136	92	35
1713	164	84	80	153	113	25
1714	158	70	88	186	153	43
1721	178	82	96	135	101	45
1732	184	75	109	151	122	28
1733	195	88	107	144	113	25
1722	199	80	119	139	98	24
1727	131	52	79	100	82	21
1731	212	87	125	142	107	40
1734	166	81	85	96	77	29
1729	201	84	117	131	103	36
1735	189	77	112	136	106	47
1740	193	73	120	137	114	47
1734	230	99	131	122	90	30
1735	217	87	130	128	101	45
1736	209	84	125	127	98	21
1738	217	95	122	119	92	43
1739	228	96	132	125	102	41
1750	230	87	143	131	106	49
1751	238	98	140	126	106	51
1755	223	85	138	125	101	46
1756	227	104	123	129	102	51
1740	234	98	136	117	82	8
1770	197	71	126	102	85	27
1755	229	87	142	108	82	34
1757	234	89	145	105	85	38
1761	229	87	142	105	82	30
1765	234	87	147	112	83	27
1767	239	88	151	115	89	25
1772	221	77	144	111	91	36
1788	217	72	145	119	94	38
1804	213	66	147	115	84	26
1809	240	79	161	110	89	38
1761	268	88	180	99	77	27
1770	256	87	169	95	74	19
1783	258	93	165	95	73	49
1788	248	87	161	97	74	31
1798	261	89	172	99	74	29
1800	259	78	181	91	67	30
1834	271	92	179	103	75	19

Table B.5: *Dimensions of post-medieval wine bottle data (mm). Height = NH + BH, where NH is neck height and BH body height (Source: Robertson, 1976).*

Stone ‘circle’ diameters

Northern Britain		Southern England	
Diameter (ft)	Deviation (ft)	Diameter (ft)	Deviation (ft)
86.0	1.0	81.0	0.5
59.5	1.5	104.5	2.0
58.0	1.5	114.5	1.0
45.0	1.0	72.0	0.5
59.0	1.6	100.0	1.0
63.5	1.0	109.0	1.0
84.5	2.5	78.5	1.0
75.5	1.5	130.0	2.5
67.5	1.5	67.5	6.5
34.0	0.5	142.5	5.6
115.0	3.0	81.5	9.5
76.5	2.4	140.0	12.0
47.5	6.5	48.0	2.0
43.5	4.0	151.0	6.9
63.5	4.0	23.0	1.0
68.5	5.0	17.0	2.0
84.0	3.5	38.5	5.0
21.5	1.0	29.0	3.0
75.0	3.5	82.0	5.0
76.0	5.0		
53.5	3.0		
107.0	9.0		
29.0	2.0		
95.5	4.5		
37.5	2.0		
103.5	8.5		
50.5	11.5		
62.5	4.5		
23.5	2.5		
42.0	10.5		
25.5	4.5		
53.5	3.0		
143.0	13.0		
28.0	3.5		
38.0	3.0		
70.0	10.0		
32.5	4.5		
13.5	1.0		
93.5	6.5		
28.0	3.5		
57.0	3.0		
24.0	3.5		
87.5	7.5		
86.0	8.5		
44.5	3.5		
54.5	10.5		
85.0	5.0		
68.0	3.5		
67.5	4.0		
153.0	17.0		

Table B.6: *Neolithic stone ‘circle’ diameters from northern Britain and southern England. The ‘circles’ are not ‘true’ circles; ‘deviation’ is the difference between the maximum and minimum diameters (Source: Barnatt and Moir, 1984).*

Waste glass compositions I

Site	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn
Mancetter	2.51	0.53	0.56	6.98	17.44	0.73	0.09	0.15	0.58
Mancetter	2.36	0.49	0.53	6.71	17.69	0.68	0.09	0.13	0.40
Mancetter	2.30	0.36	0.49	8.10	15.94	0.68	0.07	0.13	0.77
Mancetter	2.42	0.52	0.56	6.93	17.59	0.72	0.09	0.14	0.47
Mancetter	2.32	0.37	0.51	7.51	16.27	0.69	0.07	0.13	0.21
Mancetter	2.34	0.56	0.52	6.10	18.61	0.69	0.10	0.11	0.30
Mancetter	2.50	0.46	0.50	6.83	17.46	0.79	0.08	0.15	0.40
Mancetter	2.47	0.53	0.55	6.55	18.55	0.75	0.09	0.12	0.35
Mancetter	2.41	0.67	0.62	6.18	18.33	0.81	0.12	0.14	0.52
Mancetter	2.64	0.50	0.63	7.76	15.66	0.63	0.08	0.16	0.21
Mancetter	2.77	0.58	0.50	7.33	16.10	0.68	0.08	0.14	0.57
Mancetter	2.43	0.69	0.72	6.27	17.84	0.98	0.12	0.22	0.63
Mancetter	2.50	0.36	0.53	8.51	15.46	0.60	0.07	0.16	0.45
Mancetter	2.63	0.46	0.47	7.25	16.26	0.59	0.07	0.12	0.30
Mancetter	2.66	0.41	0.50	7.35	17.12	0.63	0.07	0.15	0.11
Mancetter	2.43	0.62	0.52	6.89	17.17	0.69	0.08	0.13	0.44
Mancetter	2.55	0.53	0.52	7.91	16.20	0.62	0.07	0.15	0.38
Mancetter	2.44	0.54	0.56	6.65	17.68	0.97	0.10	0.12	0.40
Mancetter	2.22	0.34	0.46	7.08	16.14	0.63	0.06	0.15	0.12
Mancetter	2.59	0.37	0.46	7.57	15.71	0.56	0.07	0.16	0.07
Mancetter	2.45	0.48	0.55	6.84	17.73	0.76	0.09	0.14	0.62
Mancetter	2.42	0.49	0.51	7.00	16.32	0.93	0.08	0.14	0.42
Mancetter	2.27	0.38	0.48	7.88	16.28	0.52	0.07	0.14	0.26
Mancetter	2.48	0.55	0.55	6.64	18.76	0.75	0.09	0.12	0.36
Leicester	2.27	0.32	0.39	6.75	17.95	0.75	0.07	0.12	0.18
Leicester	2.32	0.84	0.55	6.19	19.78	0.70	0.10	0.11	0.24
Mancetter	2.46	0.49	0.54	6.82	18.07	0.75	0.08	0.13	0.60
Mancetter	2.67	0.34	0.49	6.94	18.04	0.54	0.06	0.11	0.44
Mancetter	2.47	0.42	0.51	7.57	17.94	0.76	0.07	0.14	0.41
Mancetter	2.40	0.45	0.54	7.62	17.76	0.64	0.08	0.13	0.40
Mancetter	2.41	0.36	0.54	8.15	16.65	0.54	0.07	0.13	0.44
Mancetter	2.68	0.38	0.59	8.47	16.14	1.54	0.07	0.14	0.42
Mancetter	2.41	0.63	0.53	6.84	17.77	0.76	0.08	0.16	0.45
Mancetter	2.38	0.55	0.55	6.73	17.37	0.76	0.08	0.16	0.44
Leicester	2.50	0.78	0.56	6.40	18.35	0.73	0.11	0.11	0.27
Leicester	2.38	0.84	0.54	6.17	18.05	0.70	0.10	0.11	0.26
Mancetter	2.50	0.54	0.58	7.21	16.86	1.05	0.09	0.14	0.57
Mancetter	2.35	0.43	0.51	8.02	17.52	0.56	0.07	0.14	0.29
Leicester	2.31	0.74	0.54	6.26	18.59	0.69	0.10	0.10	0.25
Mancetter	2.42	0.36	0.47	7.31	17.76	0.62	0.07	0.13	0.24
Leicester	2.34	0.54	0.54	6.76	17.62	0.68	0.09	0.13	0.42
Leicester	2.21	0.85	0.56	6.21	19.64	0.71	0.09	0.11	0.25
Leicester	2.17	0.56	0.56	6.22	20.03	0.69	0.10	0.10	0.23
Mancetter	2.40	0.54	0.54	7.14	16.87	0.79	0.08	0.13	0.56
Mancetter	2.58	0.37	0.49	7.36	16.58	0.65	0.07	0.13	0.47
Leicester	2.45	0.89	0.55	6.19	18.30	0.71	0.11	0.12	0.26
Leicester	2.24	0.52	0.52	6.36	18.69	0.60	0.09	0.11	0.29
Mancetter	2.49	0.48	0.55	7.32	18.14	1.00	0.08	0.14	0.40
Mancetter	2.40	0.50	0.54	6.70	18.85	0.70	0.09	0.12	0.39
Leicester	2.27	0.75	0.55	6.24	19.53	0.67	0.09	0.11	0.25
Leicester	2.27	0.87	0.56	6.39	18.98	0.68	0.09	0.11	0.29
Leicester	2.34	0.43	0.58	9.42	15.72	0.59	0.08	0.13	0.14
Leicester	2.49	0.85	0.54	6.36	18.01	0.73	0.11	0.11	0.27
Leicester	2.43	0.44	0.50	6.77	17.70	0.74	0.08	0.15	0.48
Leicester	2.25	0.59	0.56	5.52	20.55	1.01	0.12	0.09	0.26

Table B.7: *Romano-British waste glass major oxide compositions (%) from two sites. Continued – see Table B.8 for details.*

Waste glass compositions II

Site	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn
Mancetter	2.33	0.37	0.52	7.31	16.75	0.49	0.06	0.11	0.90
Mancetter	2.46	0.47	0.52	7.03	17.48	0.67	0.08	0.14	0.49
Leicester	2.55	0.56	0.58	7.17	17.34	0.72	0.09	0.14	0.69
Mancetter	2.38	0.64	0.61	5.99	19.63	0.79	0.12	0.14	0.50
Mancetter	2.38	0.47	0.52	6.79	17.36	0.67	0.08	0.13	0.45
Mancetter	2.71	0.37	0.47	7.50	16.57	0.48	0.07	0.13	0.21
Leicester	2.23	0.73	0.54	6.07	18.58	0.64	0.10	0.10	0.23
Leicester	2.45	0.77	0.56	6.41	19.07	0.73	0.11	0.11	0.28
Mancetter	2.58	0.37	0.54	7.57	16.11	0.61	0.07	0.14	0.14
Leicester	2.46	0.35	0.51	7.72	16.51	0.56	0.07	0.12	0.17
Leicester	2.17	0.54	0.57	6.23	19.98	0.67	0.10	0.10	0.21
Mancetter	2.59	0.58	0.56	7.61	16.74	0.68	0.08	0.17	0.50
Leicester	2.22	0.48	0.52	6.44	18.66	0.62	0.09	0.11	0.31
Leicester	2.52	0.86	0.56	6.45	18.32	0.74	0.12	0.12	0.26
Leicester	2.34	0.78	0.58	6.37	19.34	0.73	0.10	0.11	0.26
Leicester	2.64	1.11	0.59	7.89	17.78	0.75	0.12	0.15	0.26
Leicester	2.32	0.64	0.58	5.66	20.08	0.79	0.13	0.12	0.31
Leicester	2.73	0.74	0.55	6.12	18.83	0.77	0.11	0.10	0.29
Leicester	2.51	0.78	0.55	6.44	18.30	0.73	0.11	0.12	0.26
Leicester	2.37	0.81	0.55	6.38	19.03	0.70	0.10	0.11	0.24
Leicester	2.31	0.88	0.57	6.42	18.90	0.76	0.10	0.12	0.28
Leicester	2.50	0.78	0.56	6.46	18.57	0.73	0.11	0.12	0.26
Leicester	2.57	0.80	0.56	6.43	18.41	0.75	0.12	0.12	0.26
Leicester	2.24	0.84	0.56	6.26	19.49	0.73	0.09	0.12	0.23
Leicester	2.37	0.44	0.50	6.78	17.15	0.70	0.08	0.15	0.45
Leicester	2.48	0.77	0.55	6.36	18.30	0.73	0.11	0.12	0.26
Leicester	2.26	0.58	0.61	6.16	19.47	0.74	0.10	0.11	0.21
Leicester	2.59	0.48	0.60	8.76	14.50	0.51	0.07	0.13	0.27
Leicester	2.25	0.66	0.52	6.20	18.06	0.64	0.09	0.11	0.24
Leicester	2.43	0.48	0.56	7.60	15.57	0.62	0.08	0.16	0.49
Leicester	2.49	0.93	0.55	6.18	16.54	1.10	0.12	0.13	0.25
Leicester	2.46	0.76	0.55	6.37	17.95	0.72	0.11	0.12	0.26
Leicester	2.47	1.05	0.56	7.62	17.02	0.70	0.11	0.14	0.26
Leicester	2.16	0.74	0.53	6.09	17.25	0.65	0.09	0.11	0.25
Leicester	2.26	0.58	0.52	6.41	17.28	0.67	0.09	0.13	0.28
Leicester	2.29	0.78	0.56	6.24	18.45	0.70	0.10	0.11	0.26
Leicester	2.30	0.78	0.53	6.28	18.20	0.65	0.10	0.11	0.25
Leicester	2.52	0.65	0.55	6.16	18.69	0.74	0.11	0.10	0.29
Leicester	2.28	0.68	0.55	6.37	18.60	0.68	0.10	0.11	0.24
Leicester	2.25	0.62	0.56	5.55	19.47	0.74	0.13	0.11	0.31
Leicester	2.32	0.80	0.54	6.34	18.25	0.66	0.10	0.11	0.25
Leicester	2.35	0.74	0.55	6.54	18.44	0.71	0.10	0.11	0.26
Leicester	2.45	0.42	0.61	9.79	16.22	0.62	0.08	0.13	0.14
Leicester	2.19	0.84	0.54	6.13	17.99	0.69	0.10	0.11	0.24
Leicester	2.62	0.82	0.54	6.25	17.79	0.73	0.11	0.09	0.29
Leicester	2.35	0.65	0.54	6.73	17.91	0.72	0.10	0.12	0.28
Leicester	2.44	0.35	0.51	7.70	16.27	0.62	0.07	0.13	0.16
Leicester	2.42	0.68	0.53	6.15	17.19	0.77	0.13	0.11	0.27
Leicester	2.52	0.79	0.56	6.37	18.11	0.74	0.12	0.11	0.26
Leicester	2.37	0.75	0.55	6.33	18.55	0.69	0.10	0.11	0.25

Table B.8: *Romano-British waste glass major oxide compositions (%) from two sites. The source is an unpublished University of Sheffield, UK, PhD thesis by Dr. Caroline Jackson. They are given in Table A.1 of Baxter (1994).*

Stone axe dimensions I

L1	L2	B1	B2	B3	WC	DC	TH	L3	T1	T2	Type
164.0	53.0	54.7	40.5	53.4	43.7	12.6	36.2	43.2	32.9	33.3	3
42.3	1.0	34.3	20.6	33.0	34.3	1.0	10.4	12.9	6.5	7.7	5
48.1	5.2	36.7	26.6	36.3	36.7	5.2	13.4	12.2	10.7	10.6	5
40.6	10.1	23.0	17.0	22.8	19.8	2.8	9.5	14.0	6.9	8.0	5
65.6	2.7	43.7	26.1	42.2	43.7	2.7	18.0	30.5	15.2	12.3	5
105.7	7.7	47.9	29.3	46.9	47.9	7.7	22.1	50.9	20.0	17.3	5
105.0	43.1	76.0	63.3	71.4	63.4	13.3	38.4	54.3	35.1	29.3	5
75.1	27.1	38.0	28.4	36.9	31.2	5.3	28.4	35.4	25.5	22.4	1
56.4	16.7	36.0	26.0	35.6	31.9	3.9	21.3	28.8	18.4	15.6	1
55.0	6.1	32.3	23.4	32.2	32.3	6.1	19.0	26.0	15.9	13.7	2
108.4	15.1	53.6	30.3	53.0	53.6	15.1	24.4	67.2	20.8	20.7	1
60.0	29.6	31.4	27.9	26.0	15.1	1.8	20.9	29.3	18.0	14.4	3
57.0	3.0	42.9	35.8	42.6	42.9	3.0	13.9	13.6	11.9	10.4	5
83.0	25.9	44.8	27.5	43.1	42.1	16.9	19.0	51.6	17.4	12.0	1
59.2	6.4	36.9	30.0	36.7	36.9	6.4	17.4	24.0	15.1	12.8	5
46.7	8.5	21.2	16.6	21.5	19.0	0.1	10.3	22.3	7.1	7.2	5
23.9	7.3	18.2	15.9	18.4	16.2	0.0	9.0	10.5	6.9	6.3	5
24.5	2.9	23.2	15.0	23.2	23.2	2.9	5.9	9.8	4.6	5.1	5
29.1	3.2	26.9	17.0	26.7	26.9	3.2	7.8	13.5	6.3	4.9	4
51.5	2.0	34.0	20.9	32.3	34.0	2.0	9.8	13.5	8.3	7.5	5
44.5	5.2	41.6	28.2	41.2	41.6	5.2	11.0	16.6	8.9	7.9	5
42.2	7.0	34.0	17.6	33.3	32.4	4.4	11.8	19.3	10.2	8.9	1
150.0	38.8	65.3	42.5	65.0	61.0	12.0	40.7	52.5	32.7	33.1	1
36.9	2.3	25.6	15.5	25.4	25.6	2.3	6.0	16.8	4.7	4.8	1
31.7	5.9	27.9	22.9	27.8	27.0	2.8	8.6	10.4	7.6	7.8	5
37.0	9.7	28.3	20.9	28.3	26.4	2.5	7.9	15.4	6.9	6.9	5
39.7	18.6	32.2	28.0	28.6	24.2	3.7	14.8	20.3	13.9	10.3	4
48.6	2.7	25.0	20.2	24.7	25.0	2.7	11.7	14.5	9.3	9.8	3
23.2	1.1	13.2	10.8	12.9	13.2	1.1	6.4	7.2	4.8	4.9	5
280.0	98.5	75.1	49.6	72.8	57.0	10.7	55.1	126.3	43.9	48.7	1
151.5	57.7	92.0	66.4	91.3	88.0	14.3	32.0	91.9	25.8	26.5	5
189.5	12.9	68.1	38.5	68.1	68.1	12.9	44.8	80.4	33.8	41.3	1
189.0	60.0	59.8	41.9	57.4	40.5	5.0	38.9	57.9	25.3	37.4	1
128.7	7.5	60.5	34.4	59.4	60.5	7.5	41.5	51.0	30.7	35.0	1
235.0	73.0	56.4	38.7	55.5	46.2	6.8	50.1	82.9	37.0	44.3	1
116.1	25.6	74.6	56.0	74.3	70.9	12.6	14.7	27.3	11.7	13.8	3
88.2	11.8	72.8	51.2	71.9	72.8	11.8	15.9	39.1	12.0	14.1	5
117.5	33.3	62.4	42.7	62.4	57.7	7.9	19.4	56.4	15.4	16.1	4
123.6	42.7	43.2	28.7	42.8	39.3	3.7	37.4	40.2	27.4	31.8	1
54.5	3.7	50.9	37.6	50.9	50.9	3.7	14.4	17.0	11.0	12.1	5
96.4	21.5	53.3	31.3	52.9	48.9	3.1	36.8	44.0	27.8	28.3	1
82.8	8.7	48.8	28.9	48.6	48.8	8.7	33.6	32.2	25.1	26.3	1
66.0	26.3	49.5	40.7	47.7	44.0	10.0	26.6	25.1	23.1	20.0	5
55.0	22.9	34.5	28.9	33.3	30.2	3.6	16.7	17.2	13.7	14.7	5
108.6	28.6	47.5	34.4	46.7	42.8	16.1	35.2	41.3	26.9	29.6	1
76.1	21.0	40.4	33.1	40.3	39.0	4.2	26.0	30.8	22.2	21.4	3
78.8	25.3	43.8	31.3	44.6	35.3	7.5	28.7	37.5	24.6	22.3	3
106.2	17.9	54.4	37.2	54.9	47.8	6.6	26.4	31.8	20.8	24.2	1
73.4	11.4	60.0	47.9	59.7	58.8	3.0	14.0	34.8	11.6	12.4	5
104.1	27.3	52.7	36.4	52.4	48.0	11.4	34.1	49.2	27.4	26.6	1
39.5	15.4	37.0	30.7	36.0	33.8	5.2	16.7	24.4	15.3	12.0	5
98.8	1.4	34.8	23.5	33.4	34.8	1.4	12.6	43.7	10.8	11.8	3
61.8	3.4	46.6	32.2	45.0	46.6	3.4	11.5	21.4	10.1	10.2	5
38.6	9.0	19.8	17.1	19.7	8.1	1.5	7.2	14.0	6.4	6.5	5
33.0	11.4	13.8	11.1	13.3	11.8	3.3	7.8	13.2	5.8	6.7	2
33.3	1.8	24.6	18.4	24.0	24.6	1.8	7.5	13.2	5.6	6.5	4
35.8	9.8	22.9	16.2	21.0	21.1	11.0	3.3	8.1	3.2	3.0	5
72.5	19.7	46.0	29.3	45.8	41.4	5.3	29.1	37.6	24.0	22.2	1
46.2	2.1	40.0	24.8	38.7	40.0	2.1	11.1	18.4	8.9	9.1	5
35.1	5.9	28.7	17.9	28.4	28.7	5.9	9.0	11.3	6.7	7.8	1
45.3	10.7	34.0	17.5	32.2	34.0	9.4	10.7	42.8	10.4	8.7	1
55.1	21.2	17.9	16.3	17.0	15.2	1.0	6.1	14.5	5.5	5.5	3
97.8	69.1	13.9	13.8	10.4	3.9	0.5	9.9	18.9	10.0	8.1	5
69.3	40.0	20.2	19.3	16.5	9.2	0.4	7.0	35.7	6.4	5.6	5
37.3	10.0	27.7	21.3	25.4	25.2	0.9	10.8	16.7	10.4	7.3	5
76.9	7.1	40.2	34.2	40.0	40.2	7.1	23.0	57.8	22.2	18.7	4
72.1	19.9	47.2	41.4	45.9	39.4	10.1	23.9	36.2	20.6	19.2	3
80.5	18.1	41.2	33.7	41.2	32.6	3.7	26.4	36.1	22.7	20.0	3
112.0	22.8	51.4	44.0	51.0	47.6	6.1	32.0	56.1	28.1	27.8	3
123.0	36.2	56.2	38.7	55.3	43.0	3.1	38.7	51.2	30.3	30.7	1
73.8	20.7	55.1	44.4	54.8	51.8	6.1	16.9	22.4	14.0	14.5	4
100.0	29.0	51.7	31.3	51.3	39.0	9.4	33.6	35.8	27.1	26.0	1
52.3	12.7	35.5	27.2	35.5	33.0	3.2	16.8	18.0	13.6	13.0	5

Table B.9: *Stone axe dimensions. Continued – see TableB.11 for details.*

Stone axe dimensions II

L1	L2	B1	B2	B3	WC	DC	TH	L3	T1	T2	Type
55.4	5.7	45.5	30.5	44.4	45.5	5.7	14.1	12.9	11.4	12.0	5
71.8	4.3	47.5	26.8	47.5	43.5	4.0	18.8	42.7	18.4	14.8	2
34.1	6.4	32.4	18.2	32.1	29.5	8.0	8.1	9.2	5.7	7.7	3
51.0	8.0	27.7	23.2	27.3	26.2	2.2	11.9	17.5	10.5	10.7	3
65.0	16.0	32.8	20.3	32.7	30.8	2.3	16.4	28.5	14.6	13.9	1
68.2	10.5	53.7	36.4	52.2	53.7	9.2	18.0	31.3	14.7	12.5	5
63.8	6.5	45.9	27.9	45.2	45.9	6.5	11.2	27.3	7.7	9.7	3
57.1	10.3	53.8	37.9	53.0	53.8	10.3	14.4	20.7	11.3	10.9	5
61.6	5.5	46.8	30.4	46.4	46.8	5.5	17.0	33.5	13.4	12.8	5
90.5	28.7	44.4	30.1	43.7	37.5	6.4	33.9	33.5	26.9	29.3	4
133.1	36.4	78.8	57.0	78.5	73.2	5.7	39.7	53.8	44.1	39.7	3
41.2	3.8	31.1	21.7	31.2	31.1	3.8	8.8	17.5	7.0	7.9	3
27.5	1.4	22.5	17.0	22.3	22.5	1.4	6.4	12.2	4.0	5.9	5
78.9	37.6	42.1	37.4	39.0	25.8	6.0	16.7	22.6	16.2	14.8	5
92.0	18.8	59.1	44.9	58.8	59.1	18.8	37.7	40.3	34.0	30.3	3
125.7	37.2	77.5	61.3	76.0	69.9	12.3	37.7	49.6	32.8	32.1	5
42.4	7.0	42.7	34.0	42.5	39.8	2.1	12.1	15.1	9.6	9.8	5
120.2	12.8	55.5	36.1	54.2	55.5	12.8	35.1	35.2	27.7	30.5	3
96.9	28.2	43.3	24.8	41.8	36.6	10.5	30.3	38.4	19.4	24.6	1
169.0	35.2	61.9	40.9	61.2	53.8	9.0	46.6	73.4	36.8	36.8	1
132.4	30.9	65.6	39.3	65.6	62.0	7.6	39.1	47.0	31.0	33.7	1
52.4	15.6	38.1	24.8	37.8	35.3	5.9	10.1	21.0	7.6	7.4	5
67.9	8.0	31.1	21.7	30.7	31.1	8.0	12.7	27.4	9.1	10.9	1
134.5	17.6	74.1	39.5	71.2	74.1	17.6	19.3	37.0	12.8	17.0	1
68.0	10.7	43.4	35.6	42.9	43.4	10.7	19.2	26.4	14.1	16.5	5
33.0	2.4	31.4	23.1	31.1	31.4	2.4	11.4	9.2	7.6	9.1	5
64.9	7.1	42.3	30.2	40.5	42.3	7.1	14.5	17.1	10.0	13.6	5
15.3	3.0	16.4	14.2	16.4	13.9	0.0	4.6	4.0	3.8	3.5	5
31.8	10.0	27.4	22.6	27.0	24.4	2.4	8.5	13.9	23.0	26.6	5
79.5	7.1	58.2	46.4	58.1	58.2	7.1	41.5	36.3	35.7	32.6	3
92.3	10.0	49.2	35.7	47.9	49.2	10.0	25.9	24.2	22.3	22.5	3
60.5	13.0	42.1	34.7	42.0	39.2	4.2	18.6	29.9	13.3	15.4	5
126.5	34.8	50.5	32.0	49.2	41.0	10.1	37.4	62.4	27.6	29.2	1
141.0	44.8	62.5	38.6	62.0	56.7	12.2	41.5	66.6	31.3	33.4	1
121.8	23.6	50.1	38.1	50.0	42.9	9.5	36.4	53.8	29.6	28.9	1
122.3	14.6	59.3	39.0	58.8	56.2	9.2	30.6	44.3	27.1	24.5	1
111.9	33.5	48.7	33.1	47.5	39.4	6.1	36.7	45.5	26.8	29.0	1
109.7	33.9	54.5	32.0	53.9	44.8	6.0	28.1	45.8	21.8	22.3	1
108.0	31.2	43.1	27.5	42.4	36.4	7.1	35.6	50.7	25.3	25.8	1
94.2	21.1	45.5	32.6	45.1	38.0	4.7	32.0	45.4	27.8	27.0	1
95.2	24.9	48.7	35.1	48.3	39.6	8.2	31.9	40.4	25.1	23.5	3
97.6	10.3	62.2	30.0	61.6	62.2	9.5	22.1	31.2	17.0	18.7	1
81.3	32.9	47.7	38.0	46.7	42.7	10.3	26.8	32.4	25.0	21.2	5
77.3	24.9	45.7	37.9	44.8	37.2	4.4	26.0	45.8	24.9	18.0	3
78.9	22.7	47.5	34.1	47.3	41.1	5.7	35.2	34.2	29.0	25.9	3
74.0	6.8	37.7	26.5	37.4	37.7	6.8	17.1	32.0	14.4	13.9	5
71.6	7.1	39.5	25.5	39.3	39.5	7.1	22.5	35.0	18.9	17.9	1
185.0	56.6	58.1	41.4	57.4	79.1	6.9	43.0	77.2	31.7	36.4	1
120.8	51.2	42.4	34.8	40.0	28.1	4.8	32.5	64.8	29.2	25.0	5
110.5	53.1	50.5	42.5	41.2	26.6	3.2	27.2	43.3	23.9	23.1	3
121.3	39.5	50.4	36.0	48.7	39.3	5.5	36.9	62.3	30.9	29.7	1
117.3	48.0	48.3	36.8	45.1	38.1	9.7	37.1	44.7	29.8	31.1	1
117.9	45.2	47.6	36.0	44.3	28.1	1.4	38.2	59.8	33.1	26.9	2
113.1	28.7	49.4	39.2	48.8	42.0	3.5	41.7	61.2	36.2	27.2	5
71.9	19.1	45.1	29.4	44.6	42.4	4.9	29.3	29.3	24.2	22.5	1
137.0	47.6	54.5	44.4	52.9	42.8	4.8	45.8	83.2	42.3	36.5	2
131.7	47.8	45.5	34.0	43.5	31.4	2.0	37.8	57.2	30.4	31.3	1
101.9	10.7	48.4	39.8	48.0	45.8	4.0	29.1	39.9	25.6	23.6	4
85.8	18.0	42.5	27.6	42.4	38.4	7.0	22.9	33.0	16.9	19.8	1
182.5	47.3	76.7	48.1	75.7	67.6	14.2	50.6	67.8	38.1	46.4	1
177.5	19.1	112.7	61.6	112.1	112.7	19.1	30.0	61.9	24.3	26.0	1
162.0	18.0	66.2	39.2	65.8	63.3	10.7	36.9	59.5	26.4	31.2	1
139.0	43.6	60.1	45.9	58.8	53.0	11.6	45.0	66.9	37.9	35.1	2
191.0	72.7	53.3	36.3	49.0	33.9	5.9	37.9	86.5	24.3	28.9	1
115.2	34.6	44.9	28.9	44.2	38.0	6.4	34.6	47.7	25.4	28.2	1
101.6	29.1	51.7	36.0	51.0	45.2	7.3	31.5	48.3	26.6	26.1	1
100.1	12.0	46.8	30.0	46.4	45.5	5.7	28.1	48.3	24.6	21.4	2
100.6	13.3	44.7	26.7	44.1	43.3	6.7	28.1	44.6	21.3	21.8	1

Table B.10: *Stone axe dimensions. Continued – see Table B.11 for details.*

Stone axe dimensions III

L1	L2	B1	B2	B3	WC	DC	TH	L3	T1	T2	Type
96.7	33.2	43.9	29.0	41.6	30.8	0.0	27.8	42.4	21.8	22.2	1
85.2	26.9	36.7	24.5	35.8	31.3	7.4	25.1	37.1	18.7	20.3	1
84.2	7.9	41.7	27.0	41.5	41.7	7.9	30.0	36.4	23.6	25.7	1
90.2	6.9	43.1	25.8	43.0	43.1	9.5	28.8	39.2	23.0	22.7	1
85.8	26.3	56.0	45.8	53.3	56.0	26.3	20.1	46.7	18.0	17.3	3
74.0	25.1	51.3	40.8	47.5	43.0	15.2	22.6	40.5	18.8	16.9	3
54.3	23.7	34.5	31.0	32.1	27.5	6.4	15.4	11.0	13.6	14.5	3
126.4	36.2	47.0	35.5	46.1	36.7	6.4	28.5	86.0	26.4	23.8	2
94.0	42.3	38.8	33.3	35.5	26.8	7.8	28.9	41.8	21.5	22.3	3
108.6	46.6	41.0	34.1	37.9	28.0	5.0	31.6	51.5	28.9	21.4	3
99.5	25.7	45.9	30.5	46.1	43.5	8.5	29.6	41.6	23.4	24.3	1
81.4	16.5	42.7	26.6	42.7	38.9	8.7	28.3	35.3	21.5	21.8	1
47.0	21.6	32.9	28.9	30.9	28.4	6.5	14.7	20.4	11.2	12.1	5
52.0	9.5	38.0	24.4	37.7	36.0	5.2	16.5	23.4	11.9	11.2	2
47.4	15.8	29.6	20.6	28.2	24.0	7.1	10.4	6.1	8.0	10.2	3
51.2	6.0	30.7	20.9	30.2	30.7	6.0	15.7	16.5	13.5	14.8	2
94.4	47.6	60.7	50.3	57.3	50.9	9.0	32.0	43.6	23.5	25.1	5
165.0	36.3	67.3	48.7	67.1	61.4	12.0	45.2	66.0	36.2	37.4	2
105.0	14.0	47.0	37.0	47.0	46.0	13.0	31.0	58.0	26.0	23.0	4
97.0	43.0	37.0	31.0	35.0	33.0	7.0	30.0	45.0	26.0	26.5	2
114.0	37.0	45.0	30.0	40.0	37.0	5.0	38.0	57.0	28.0	30.0	1
98.0	18.0	40.0	15.0	39.0	39.0	15.0	17.0	78.0	16.0	12.0	1
113.0	7.3	49.6	35.0	49.2	49.6	7.3	16.8	39.3	12.7	13.3	2
82.6	17.9	45.9	30.5	45.7	45.9	17.9	14.8	15.2	12.2	13.0	5
86.4	11.7	50.7	30.5	50.0	48.6	11.3	29.0	38.4	23.3	23.4	2
140.0	44.8	64.4	50.7	62.9	51.7	5.1	14.0	19.6	10.7	13.0	5
78.4	9.1	36.0	25.8	35.8	36.0	9.1	15.6	19.2	10.8	13.7	4
92.3	24.5	44.3	28.1	44.4	39.6	7.0	25.4	34.1	17.8	19.9	1
88.8	12.3	41.9	24.0	41.5	36.5	3.6	26.8	37.1	21.5	22.1	1
76.5	22.8	52.0	39.0	51.6	47.3	4.1	23.5	44.3	19.0	18.8	5
121.1	7.3	71.4	45.2	69.1	71.4	7.3	16.6	33.3	13.5	15.7	5
143.0	20.3	61.3	42.6	61.5	51.2	3.5	31.8	37.2	26.7	29.3	2
44.0	5.8	31.3	21.8	31.0	29.7	3.5	16.3	17.9	12.7	11.5	1
42.4	3.8	31.7	15.9	30.3	31.7	3.8	11.6	14.2	8.7	8.4	1
82.8	10.8	42.5	25.8	41.9	42.5	10.8	22.7	35.7	16.7	18.6	1
71.0	21.6	34.8	24.8	34.3	31.5	6.0	19.1	30.4	14.8	15.0	1
52.4	10.0	41.0	26.7	40.9	37.6	4.2	11.5	23.1	8.8	9.0	5
117.0	12.1	57.7	35.4	57.6	57.7	12.1	23.9	38.7	15.4	19.5	1
80.0	14.5	58.6	35.8	57.4	58.6	14.5	23.0	22.2	18.1	22.5	5
73.5	27.0	39.4	29.8	38.3	34.1	8.4	26.0	35.3	20.6	18.9	3
104.5	23.4	54.0	29.4	54.8	51.2	7.5	16.5	43.9	15.8	13.5	1
96.0	32.1	46.0	28.7	45.4	41.9	8.2	32.9	43.9	26.1	23.8	1
67.6	12.1	39.0	24.9	38.8	25.5	0.0	14.7	23.9	13.6	12.2	3
127.8	16.3	63.0	34.7	60.2	63.0	16.3	36.7	56.2	26.9	33.7	1
79.8	24.4	44.1	34.7	43.2	37.7	9.6	17.5	34.7	14.5	15.2	3
31.9	1.9	26.3	15.4	25.5	26.3	1.9	7.0	14.4	5.2	5.1	2
58.0	4.2	52.1	31.5	52.2	52.1	4.2	11.2	17.4	10.9	11.2	5
78.9	8.7	40.1	23.5	39.7	37.8	3.8	18.9	38.3	13.0	15.4	5
107.5	37.1	51.0	33.7	49.7	44.1	6.9	33.2	45.2	27.1	25.2	1
91.9	13.9	50.9	41.2	51.1	49.1	14.9	30.4	27.1	26.2	26.7	3
153.0	25.6	70.8	54.2	70.9	67.3	14.8	28.9	42.2	22.3	25.9	1
100.4	12.3	55.5	40.6	54.3	55.0	10.3	36.3	54.0	30.0	28.3	3
88.1	9.8	48.1	30.0	48.0	48.1	9.8	25.9	40.7	20.3	20.7	1
83.5	14.2	43.5	23.1	42.9	40.8	5.2	18.7	33.7	14.3	15.1	1
81.6	10.4	41.5	25.7	40.9	40.3	7.5	23.1	36.8	16.7	19.5	1
80.3	19.4	44.3	27.8	43.6	35.1	3.7	21.2	20.8	14.8	19.2	5
78.7	28.0	38.0	27.8	37.4	33.8	4.6	20.2	28.3	15.9	17.2	3
75.2	13.9	39.1	24.3	39.4	33.7	1.6	24.7	32.7	17.4	19.4	1
75.5	24.9	37.5	26.8	36.5	29.8	3.9	26.7	32.9	20.0	20.5	5
57.5	9.0	38.8	25.7	38.6	35.7	5.1	17.0	28.0	14.5	12.3	2
56.7	4.0	49.7	28.3	48.5	49.7	4.0	17.0	24.4	13.4	12.0	1
46.5	17.8	28.5	21.9	26.2	18.0	1.3	14.6	25.6	12.6	10.0	4
98.6	22.3	42.8	29.8	42.5	37.0	5.0	29.3	38.3	20.7	23.8	1
97.7	6.0	44.0	24.1	42.2	44.0	6.0	28.5	34.0	21.8	23.9	1
89.5	34.2	39.5	28.4	38.0	32.6	7.3	25.3	42.6	20.3	19.8	1
88.8	12.0	46.1	28.3	46.2	42.7	6.0	28.0	33.3	22.5	21.6	1
87.3	9.5	50.9	32.5	49.2	49.0	10.9	28.5	34.8	22.5	22.7	5
89.3	30.8	42.8	29.1	42.6	39.4	6.0	31.7	31.5	23.0	25.9	1

Table B.11: *Dimensions (mm) of polished Neolithic stone axes from southern Italy. ‘L‘, ‘B’ and ‘T’ are length, breadth and thickness variables; WC and DC are the width and depth of the cutting edge. ‘Type’ is a classification by butt shape; 1 = pointed, 3 = rounded, 5 = square, 2 and 4 are intermediate types (Source: O’Hare, 1990).*

Zooarchaeological species assemblages

Type	Region	Cow	Sheep	Pig
T	B	54	19	27
V	B	57	31	12
H	B	56	30	14
R	B	47	41	12
L	B	63	16	21
A	B	65	22	13
T	G	54	11	35
V	G	61	12	27
H	G	45	20	35
R	G	73	17	10
L	G	54	12	34
A	G	57	13	30
T	P	21	41	38
V	P	24	52	24
H	P	24	52	24
R	P	16	60	24
T	T	34	18	48
V	T	50	24	26
H	T	44	20	36
R	T	38	26	36

Table B.12: *Average percentages of species from four regions (B = Roman Britain, G = Roman Germany, P = Roman Provence, T = The Three Gauls) and six site types (T = Town, V = Vici, H = Villae, R = Rural settlements, L = legionary sites, A = Auxiliary sites. The data are derived from Figures 1-4 in Hesse (2011), who presented the data in the form of clustered barplots, based on numerical information from King (1999).*

Romano-British vessel glass of the 1st-3rd centuries A.D.

Site	Date (AD)	Cup	Bowl	Jar	Flask	Jug	Bottle	EVE total
London	75-90	36	20	1	14	17	12	21.04
Castleford	70-95	19	21	4	18	6	33	16.06
Gloucester	70-98	24	17	3	13	10	33	13.92
Caerleon TS	74-100	25	36	0	7	6	25	7.14
Carlisle	70-105	15	29	2	7	7	40	18.32
Chester	70-120	28	24	5	6	11	26	14.47
York	70-120	32	30	7	11	4	15	19.96
Colchester	65-150	34	21	7	6	11	21	26.84
Dorchester	70-150	27	9	11	22	11	19	13.11
Wroxeter	80-150	31	13	3	11	14	27	16.62
Castleford	140-80	54	3	0	8	6	30	26.54
Verulamium	150-60	47	11	3	4	10	24	10.57
Towcester	155-65	28	19	16	7	17	13	8.45
Harlow	160-70	33	14	16	19	6	12	17.10
Pentre Farm	120-200	56	14	0	6	5	19	10.32
Caerleon	130-200	62	9	0	4	3	22	4.49
Rocester	140-200	9	3	9	3	11	65	6.52
Catterick	150-200	52	5	2	1	5	35	10.48
Housesteads	150-500	76	0	14	0	8	2	7.10
Lincoln	160-230	45	4	4	7	15	25	9.33
Wroxeter	175-225	47	11	5	12	11	15	7.30
York	175-250	42	14	4	6	11	23	9.99
York	160-280	73	0	3	4	3	17	9.90
Caersws	70-130	39	19	7	10	17	27	15.83
Wilcote	70-150	7	37	0	3	18	37	2.76

Table B.13: *Site by type data for Romano-British vessel glass. Numbers are expressed as percentages and define, in percentages, the* profile *of a row. The EVE totals allow these to be converted to EVEs, and this has been done for analyses in the text. The first 10 and last two sites are classified as first- to (mid) second century and labels 1 to 12 in the figures in the text; remaining sites are (mid) second- to third century and labeled 1 to 13 in the relevant figures (Source: Cool and Baxter, 1999).*

Romano-British Flavian drinking vessels

Site	Order	Sport	Tall	Rib	Hof	Ind	FC	WC	EVEs
		Percentage							Total
Carlisle	1	29	7	36	0	7	7	14	2.8
York	2	0	0	8	92	0	0	0	2.4
Castleford	3	0	17	58	0	8	8	8	2.4
Wroxeter	4	0	0	43	29	0	14	14	1.4
Caersws	5	0	7	20	73	0	0	0	3.0
Colchester	6	18	0	0	55	14	9	5	4.4
Gloucester	7	0	18	0	82	0	0	0	3.4
Caerleon	8	0	0	0	25	21	38	17	4.8
London	9	15	8	0	22	23	25	8	13.0
Fishbourne	10	0	14	0	0	29	57	0	1.4

Table B.14: *Assemblage profiles for seven drinking vessel types, from the Flavian period in England. 'Order' is geographical from north to south. 'Sport', 'Rib' and 'Hof' refer to mould-blown sport cups, mould-blown ribbed cups and Hofheim cups; 'Tall', 'Ind', 'FC' and 'WC' are beaker types, tall mould-blown, indented, facet-cut and wheel-cut respectively (Source: Cool and Baxter, 1999).*

La Tène fibulae from Münsingen

FL	BH	BFA	FA	CD	BRA	ED	FEL	C	BW	BT	FEW	Coils	Length
93	24	7	10	16	1	13	31	47	3.5	3.5	*	4	114
21	7	6	9	6	5	2	11	10	3.5	1.7	*	12	35
33	15	2	8	7	3	8	10	20	3.9	3.2	*	4	60
23	26	4	7	9	5	12	1	16	6.2	7.7	2.8	4	74
20	23	2	8	7	1	8	5	16	7.7	5.2	2.6	6	68
27	15	6	8	7	5	3	11	11	3.7	3.5	1.8	4	55
10	16	1	10	9	1	7	0	11	6.1	4.1	0	4	45
15	18	1	10	10	1	5	0	15	3.5	3.5	0	4	40
31	13	4	9	7	4	5	11	18	17.6	1.4	3.6	6	54
19	17	1	7	6	2	6	10	12	9.2	6.6	3.9	6	39
41	23	3	8	11	3	14	15	24	7.3	5.8	8.6	6	71
47	17	5	9	10	4	8	14	26	5.8	4.7	6	6	78
29	15	3	8	6	3	6	10	17	11.7	3.9	6.4	6	47
23	13	3	8	6	2	10	7	15	5.2	2.7	5.4	12	41
20	15	1	7	5	1	12	4	12	4.7	4.8	3.5	6	38
17	16	1	7	7	1	8	3	11	5.1	3.5	2.2	6	44
20	15	2	7	7	3	6	10	12	5.5	3.8	3.9	6	50
20	13	5	8	5	2	10	5	10	4.4	4.4	5.1	6	36
21	18	2	9	9	1	5	6	15	8.1	2.3	1.9	4	49
28	17	1	10	10	2	8	6	20	2.5	2.6	2.2	4	53
94	15	7	10	12	5	11	31	50	4.3	4.3	*	6	128
22	18	1	8	7	1	5	8	17	8.8	3	2.4	6	59
20	14	1	8	6	1	3	4	14	14.3	1.4	1.7	6	44
22	15	3	8	7	3	13	1	17	5	4.6	2.5	10	47
12	22	1	6	9	1	9	0	11	6.8	6.4	0	4	45
27	15	1	8	10	2	9	11	19	8.2	4	7.6	4	53
15	19	2	8	7	3	3	4	12	3.7	3.5	1.9	4	56
10	10	2	10	6	2	2	-	9	2	2.3	2.2	3	26
9	13	3	10	4	4	9	0	8	9.6	5	0	22	28
68	18	7	9	9	7	3	50	18	9.3	6.5	*	4	110

Table B.15: *Measurements on Bronze Age fibulae from Münsingen, Switzerland. FL = foot length, BH = bow height, BFA = bow foot angle, FA = foot angle, CD = coil diameter, BRA = bow rear angle, ED = element diameter, FEL = foot extension length, C = catchplate, BW = bow width, BT = bow thickness, FEW = foot extension width, Coils = Number of coils. Angles in intervals of* 10^o *and dimensions are in millimetres. The data are from Table 9.1 of Doran and Hodson (1975) with fibulae illustrated in their Figure 9.1.*

Medieval glass compositions

Na	Mg	Al	P	K	Ca	Mn	Fe	Cu	Zn	Pb
16.2	2.1	2.9	0.3	2.8	5.2	0.42	0.47	0	0.01	0.13
16.3	2.2	2.9	0.3	2.7	5.2	0.42	0.46	0	0.01	0.12
12.3	3.1	1.7	0.8	4.7	9.5	0.60	0.40	0.11	0.02	0.14
16.6	2.2	2.7	0.4	2.7	5.1	0.42	0.46	0	0.01	0.14
16.2	2.1	2.9	0.3	2.7	5.3	0.43	0.49	0	0.01	0.15
15.2	2.5	3.1	0.4	2.6	5.3	0.40	0.47	0.03	0.01	0.11
17.0	1.1	1.1	0.2	1.1	7.8	0.50	0.35	0.21	0.01	0.09
16.5	1.1	1.2	0.2	1.3	8.0	0.51	0.36	0.22	0.01	0.12
16.5	2.3	2.8	0.4	2.7	5.1	0.42	0.47	0	0.01	0.14
12.2	3.2	1.5	0.9	5.0	9.7	0.61	0.39	0.11	0.02	0.16
12.6	3.4	1.5	0.9	4.8	9.5	0.60	0.39	0.11	0.02	0.16
12.5	3.1	1.6	0.8	4.5	9.2	0.58	0.39	0.11	0.02	0.16
12.5	3.2	1.6	0.9	4.9	9.8	0.60	0.39	0.11	0.02	0.14
12.8	3.2	1.6	0.8	4.7	9.5	0.60	0.40	0.11	0.02	0.14
12.5	3.1	1.6	0.8	4.5	9.5	0.61	0.40	0.11	0.02	0.14
15.8	1.2	1.3	0.3	1.1	7.7	0.50	0.34	0.21	0.01	0.10
17.3	0.8	1.0	0.3	0.5	7.2	0.51	0.34	0.23	0.02	0.13
12.1	3.0	1.5	0.9	4.0	9.7	0.62	0.41	0.14	0.03	0.14
12.2	3.2	1.5	0.9	4.2	9.9	0.62	0.39	0.14	0.02	0.14
15.0	1.5	0.9	0.5	1.6	8.7	0.54	0.36	0.23	0.02	0.22
14.8	0.8	0.8	0.3	1.1	8.6	0.50	0.32	0.19	0.02	0.17
13.9	2.2	0.9	0.7	2.4	10.1	0.54	0.38	0.20	0.02	0.13
16.5	0.6	0.8	0.3	0.4	7.9	0.48	0.34	0.24	0.02	0.17
17.4	0.6	0.8	0.3	0.4	7.4	0.54	0.36	0.22	0.02	0.13
16.8	0.5	1.1	0	0.2	7.1	0.47	0.34	0.21	0.01	0.11
16.6	0.5	1.2	0	0.2	7.1	0.48	0.35	0.21	0.01	0.11
16.8	0.3	1.1	0	0.2	7.3	0.47	0.33	0.20	0.01	0.12

Table B.16: *The compositions are for blue medieval glass from York Minster and various archaeological excavations. The full data are in Cox and Gillies (1986) of which the above is used in Baxter (1989).*

Levantine glass compositional data

SiO_2	Al_2O_3	FeO	MgO	CaO	Site
70.27	2.74	0.33	0.51	8.52	1
69.37	2.64	0.31	0.66	8.77	1
71.35	2.60	0.32	0.70	7.60	1
69.55	2.91	0.42	0.90	8.90	1
70.48	2.64	0.39	0.75	8.90	1
68.52	2.74	0.41	0.64	8.07	1
71.08	2.78	0.32	0.59	8.83	1
69.38	2.96	0.35	0.66	10.50	1
69.68	2.65	0.36	0.54	9.15	1
71.53	2.86	0.35	0.52	8.62	1
70.39	2.73	0.39	0.57	8.70	1
71.13	3.00	0.39	0.57	9.56	1
71.85	2.72	0.29	0.46	8.42	1
68.34	2.83	0.44	0.67	10.30	1
69.44	2.79	0.38	0.63	8.06	1
71.54	2.64	0.27	0.41	8.14	1
71.23	2.63	0.28	0.45	8.02	1
69.48	2.87	0.38	0.80	10.50	1
68.74	2.97	0.36	0.92	8.95	1
72.81	2.44	0.28	0.43	8.17	1
68.47	2.76	0.50	0.87	9.25	1
71.75	2.56	0.29	0.50	8.68	1
67.44	2.84	0.52	0.35	8.07	1
71.63	2.60	0.28	0.53	8.27	1
70.13	2.55	0.33	0.56	8.48	1
69.50	2.75	0.34	0.46	8.72	1
66.63	2.93	0.89	0.38	8.42	1
71.22	2.64	0.32	0.46	8.10	1
70.73	2.64	0.32	0.56	8.53	1
69.45	2.81	0.46	0.64	9.81	1
71.27	2.54	0.32	0.48	8.47	1
71.76	2.82	0.36	0.60	8.77	1
69.62	2.65	0.34	0.48	9.38	1
72.07	2.60	0.42	0.68	7.97	1
71.09	2.62	0.42	0.69	8.48	1
68.87	2.97	0.50	0.48	8.10	1
67.28	3.14	1.42	0.49	8.27	1
68.52	2.66	0.35	0.50	8.29	1
69.65	3.00	0.32	0.57	8.50	1
68.73	2.44	0.42	0.48	8.43	1
72.36	2.65	0.32	0.52	7.81	1
71.12	2.65	0.25	0.52	8.36	1
70.04	2.62	0.36	0.62	7.97	1
68.11	2.88	0.86	0.66	8.92	1
69.98	2.69	0.40	0.66	8.53	1
69.76	2.87	0.44	0.69	8.51	1
67.76	3.01	0.50	0.78	8.67	1
68.19	2.75	0.36	0.55	9.18	1
67.12	2.71	0.39	0.52	9.91	1
67.73	2.64	0.34	0.51	8.96	1
68.50	2.56	0.39	0.45	8.41	1
67.20	2.61	0.46	0.56	8.80	1
66.26	2.76	0.61	0.61	9.11	1
69.75	2.43	0.37	0.71	6.61	2
70.60	3.03	0.43	0.72	8.81	2
70.80	3.16	0.39	0.72	8.81	2
69.89	2.76	0.37	0.74	9.17	2
69.48	2.72	0.35	0.77	9.25	2
70.46	2.92	0.54	0.76	10.20	2
66.23	2.83	0.32	0.52	9.29	2
64.85	2.81	0.24	0.57	11.28	2
65.93	2.91	0.24	0.58	9.29	2
65.80	2.85	0.35	0.51	9.54	2
66.25	3.19	0.34	0.52	11.47	2
66.83	3.28	0.40	0.61	11.11	2
67.42	3.22	0.32	0.53	10.90	2
67.30	3.26	0.30	0.50	10.08	2

Table B.17: *Roman Levantine glass compositions. This is a subset of data analyzed, but not published, in Baxter and Freestone (2006). The intention there was to contrast log-ratio analysis with other methods; the purpose to which the data are put here is different. Professor Ian Freestone of University College London is thanked for making the data available; other analyses and the archaeological background are discussed in Freestone et al. (2000).*

Lead isotope ratio data

	Kea			Lavrion			Seriphos		
Id.	208/206	207/206	206/204	208/206	207/206	206/204	208/206	207/206	206/204
1	2.06847	0.83232	18.886	2.05822	0.83201	18.808	2.06377	0.83069	18.888
2	2.06854	0.83236	18.881	2.05748	0.83177	18.846	2.06431	0.83079	18.898
3	2.06666	0.83199	18.860	2.05659	0.83100	18.834	2.06432	0.83087	18.901
4	2.06732	0.83191	18.883	2.05785	0.83148	18.823	2.06323	0.83061	18.888
5	2.06703	0.83186	18.871	2.05690	0.83135	18.827	2.06370	0.83066	18.893
6	2.06614	0.83209	18.866	2.05803	0.83151	18.823	2.06441	0.83080	18.899
7	2.06796	0.83230	18.883	2.05768	0.83147	18.841	2.06653	0.83121	18.916
8	2.06553	0.83160	18.868	2.05713	0.83216	18.873	2.06401	0.83088	18.897
9	2.06556	0.83159	18.859	2.05805	0.83187	18.817	2.06384	0.83075	18.879
10	2.06469	0.83179	18.860	2.05720	0.83140	18.832	2.06520	0.83094	18.909
11	2.06389	0.83150	18.848	2.05785	0.83184	18.818	2.06340	0.83063	18.895
12	2.06383	0.83122	18.856	2.05662	0.83172	18.838	2.06213	0.83033	18.879
13	2.06456	0.83169	18.846	2.05634	0.83106	18.832	2.06255	0.83057	18.893
14	2.06594	0.83205	18.853	2.05816	0.83166	18.819	2.06121	0.83032	18.899
15	2.06494	0.83192	18.852	2.05883	0.83161	18.818	2.06486	0.83095	18.908
16	2.06510	0.83156	18.871	2.05801	0.83142	18.835	2.06512	0.83081	18.906
17	2.06767	0.83205	18.896	2.06375	0.83216	18.844	2.06489	0.83092	18.904
18	2.06492	0.83139	18.876	2.05526	0.83086	18.840	2.06521	0.83081	18.908
19	2.06423	0.83133	18.868	2.05776	0.83208	18.824	2.06349	0.83044	18.897
20	2.06338	0.83165	18.839	2.05785	0.83199	18.818	2.06202	0.83029	18.892
21	2.06544	0.83207	18.858	2.06287	0.83203	18.886	2.06238	0.83026	18.885
22	2.06456	0.83138	18.864	2.06375	0.83271	18.846	2.06280	0.83030	18.890
23	2.06662	0.83180	18.883	2.05850	0.83191	18.822	2.06529	0.83090	18.909
24	2.06712	0.83197	18.874	2.05499	0.83183	18.776	2.06618	0.83105	18.916
25	2.06724	0.83189	18.886	2.05676	0.83117	18.882	2.06454	0.83079	18.899
26	2.06720	0.83207	18.873	2.05951	0.83127	18.857	2.06323	0.83061	18.888
27	2.06620	0.83167	18.873	2.05902	0.83043	18.911	2.06512	0.83081	18.906
28	2.06660	0.83191	18.878	2.05857	0.83214	18.851	2.06529	0.83090	18.909
29	2.06452	0.83164	18.858	2.06124	0.83194	18.868	2.06618	0.83105	18.916
30	2.06775	0.83181	18.905	2.05671	0.83210	18.791	2.06520	0.83094	18.909
31	2.06644	0.83216	18.880	2.05891	0.83196	18.821	2.06288	0.83055	18.907
32	2.06720	0.83228	18.881	2.05978	0.83210	18.791	2.06437	0.83059	18.920
33	2.06587	0.83226	18.852	2.05846	0.83224	18.824	2.06731	0.83143	18.923
34	2.06438	0.83159	18.874	2.05537	0.83042	18.845	2.06660	0.83121	18.922
35	2.06723	0.83252	18.866	2.05521	0.83038	18.897	2.06570	0.83117	18.902
36	2.06537	0.83168	18.856	2.05616	0.83120	18.847	2.06401	0.83088	18.897
37	2.06569	0.83223	18.851	2.05789	0.83103	18.868	2.06384	0.83075	18.879
38	2.06549	0.83168	18.864	2.06219	0.83177	18.888			
39	2.06316	0.83136	18.846	2.05626	0.83077	18.838			
40	2.06648	0.83217	18.862	2.05850	0.83191	18.822			
41	2.06579	0.83168	18.868	2.05499	0.83183	18.776			
42	2.06624	0.83177	18.872	2.06020	0.83134	18.875			
43	2.06607	0.83175	18.868	2.06184	0.83167	18.886			
44	2.06684	0.83188	18.875	2.06110	0.83067	18.910			
45	2.06721	0.83193	18.879	2.05548	0.83038	18.885			
46	2.06648	0.83173	18.868	2.05825	0.83161	18.830			
47	2.06646	0.83184	18.872	2.06046	0.83173	18.820			
48	2.06670	0.83185	18.881	2.06359	0.83163	18.923			
49	2.06338	0.83182	18.861	2.06150	0.83122	18.897			
50	2.06461	0.83170	18.864	2.06194	0.83187	18.903			
51	2.06390	0.83158	18.861	2.05951	0.83127	18.857			
52	2.06531	0.83181	18.861	2.05926	0.83149	18.898			
53	2.06565	0.83183	18.869	2.05919	0.83227	18.847			
54	2.06757	0.83194	18.890	2.06119	0.83139	18.906			
55	2.06589	0.83165	18.873	2.06338	0.83207	18.901			
56	2.06784	0.83201	18.892	2.05867	0.83127	18.865			
57	2.06597	0.83170	18.865	2.05871	0.83104	18.895			
58	2.06512	0.83219	18.836	2.05552	0.83126	18.915			
59	2.06710	0.83257	18.864	2.06206	0.83198	18.906			
60	2.06751	0.83226	18.875						
61	2.06625	0.83192	18.872						
62	2.06391	0.83162	18.830						

Table B.18: *Lead isotope-ratio data for three sources in the Aegean (Source: Stos-Gale et al., 1996).*

Neolithic pot dimensions I

Id.	Type	AX	AY	BX	BY	CX	CY	DX	DY	EX	EY	FX	FY	GX	GY	HX	HY
15	A	4.32	8.65	3.92	7.25	3.72	5.75	3.82	4.95	4.22	4.07	4.32	3.22	3.10	0.92	0	0
23	A	11.40	20.00	10.37	17.42	9.60	15.20	9.60	15.20	10.20	13.95	10.37	11.87	8.77	5.35	4.37	0.10
32	A	8.15	17.00	7.35	15.70	6.65	12.82	7.02	11.42	7.92	9.67	8.20	7.87	5.80	1.95	0	0
37	A	10.60	19.20	9.50	17.05	9.03	14.07	9.33	12.33	10.27	10.97	10.43	9.37	7.83	2.67	0	0
38	A	7.53	14.20	6.67	11.40	6.40	8.72	6.45	7.90	7.00	6.90	7.17	5.92	5.97	2.75	2.32	0.05
43	A	6.80	12.10	6.27	10.33	5.97	7.83	6.10	7.30	6.40	6.20	6.55	5.10	5.70	2.00	3.50	0.30
45	A	8.05	15.60	7.52	12.67	7.20	10.20	7.35	9.52	8.30	7.97	8.60	6.65	7.17	2.55	3.97	0.22
46	A	8.80	13.77	8.22	12.45	7.70	11.15	7.70	11.15	7.80	10.87	7.90	10.55	6.25	4.37	2.47	0.35
52	A	7.53	17.20	6.67	15.13	6.00	10.02	6.42	9.02	7.10	8.05	7.40	6.55	6.10	2.95	2.37	0.10
65	A	8.40	14.82	7.97	13.10	6.85	10.35	6.85	10.35	7.52	8.70	7.75	6.97	6.45	2.80	4.07	0.20
82	A	9.30	18.70	8.37	16.20	8.00	13.27	8.62	10.27	9.42	9.37	9.70	8.20	7.65	3.75	2.10	0.10
86	A	6.97	15.90	6.40	14.17	5.87	11.93	5.93	11.77	6.83	10.00	7.30	6.73	6.17	2.70	3.60	0.20
87	A	6.35	14.20	5.95	11.65	5.63	9.17	5.77	8.77	6.37	7.77	6.65	6.07	5.45	2.52	2.75	0.07
88	A	10.43	18.20	9.27	15.50	8.93	13.93	9.23	12.47	9.67	11.50	9.90	9.27	8.50	4.05	4.43	0.07
91	A	8.25	16.50	7.70	14.55	7.25	12.05	7.85	8.95	8.65	7.70	8.95	6.15	7.40	2.20	3.75	0.05
93	A	7.90	16.20	7.10	14.00	6.80	11.33	7.22	9.27	7.85	8.15	8.12	6.27	6.40	2.40	2.40	0
94	A	8.50	19.07	8.00	16.93	7.73	14.63	8.10	10.87	8.90	9.60	9.27	7.77	7.53	3.10	2.82	0.30
99	A	4.50	10.20	4.07	7.95	3.90	6.57	4.00	6.10	4.45	5.20	4.62	4.17	3.92	1.77	2.10	0
181	A	7.82	16.30	7.05	14.35	6.75	12.80	6.85	12.07	7.37	10.35	7.57	8.27	6.22	3.37	3.02	0.20
182	A	5.70	13.20	5.12	10.65	5.00	9.02	5.32	7.50	5.82	6.75	6.00	5.70	4.75	2.10	2.07	0.20
183	A	8.47	18.90	7.63	15.77	7.40	14.10	7.47	13.43	8.63	11.73	9.12	8.82	7.45	3.82	3.57	0.25
184	A	6.77	14.20	5.90	11.77	5.50	10.27	5.62	9.77	6.40	8.20	6.62	6.47	5.37	2.52	2.80	0.03
185	A	5.10	8.10	4.32	5.92	4.10	5.15	4.10	5.15	4.10	5.15	4.10	5.15	2.90	1.47	0	0
186	A	6.60	16.50	6.10	14.00	5.90	11.60	6.07	11.40	6.97	10.00	7.40	7.50	6.02	2.77	2.95	0
187	A	9.23	20.00	8.53	18.03	8.03	15.33	8.03	15.33	9.27	13.40	9.60	10.87	7.95	5.02	3.80	0.12
188	A	8.20	14.10	7.67	12.10	7.22	9.92	7.47	9.42	7.80	8.65	7.92	7.32	6.63	3.00	3.00	0.02
189	A	3.70	8.30	3.30	6.95	3.10	5.40	3.25	4.20	3.65	3.65	3.85	2.90	2.95	0.90	1.20	0
190	A	5.33	10.20	5.07	8.77	4.87	7.40	5.00	4.90	5.30	4.33	5.43	3.70	4.10	1.27	1.30	0.03
196	A	8.92	16.10	7.60	13.10	6.80	9.95	7.05	9.22	7.72	7.97	7.95	6.62	5.40	1.75	0.00	0
200	A	5.43	8.20	4.50	6.80	4.20	5.83	4.17	5.60	4.30	5.15	4.32	4.57	3.67	1.97	1.55	0.30
201	A	6.30	14.00	6.00	12.20	5.17	9.80	5.17	9.80	6.20	8.63	6.83	6.43	5.60	2.80	2.57	0.12
202	A	5.00	7.60	4.40	6.50	4.00	5.70	4.00	5.70	4.10	5.60	4.10	5.40	3.30	2.30	1.70	0.00
205	A	14.10	28.30	12.70	23.40	12.40	20.70	12.53	19.70	13.67	17.73	14.05	14.82	11.10	5.77	5.32	0.40
207	A	4.70	10.00	4.15	8.00	4.00	6.55	4.10	4.80	4.45	4.40	4.55	3.60	3.85	1.45	1.80	0.10
209	A	12.42	20.40	11.12	16.95	10.60	14.52	10.87	13.17	11.87	11.80	12.27	9.67	8.52	2.52	0	0
210	A	4.60	9.70	3.90	8.00	3.62	6.45	3.72	5.67	4.07	4.97	4.20	4.12	3.55	1.85	1.70	0.02
214	A	6.25	11.00	5.72	9.75	5.47	8.97	5.55	8.77	5.57	8.40	5.60	8.15	4.32	3.17	1.52	0.15
221	A	12.67	23.10	11.57	20.10	11.02	16.75	11.70	13.92	12.97	12.02	13.23	10.37	10.72	3.90	5.37	0.05
222	A	13.67	27.00	12.10	22.67	11.57	19.20	11.95	16.87	12.87	14.60	13.13	12.27	10.40	4.90	4.60	0
223	A	12.83	22.00	11.87	18.87	11.57	16.40	11.77	14.97	12.63	13.60	12.87	11.87	10.03	4.73	4.80	0.13
225	A	4.97	9.00	4.35	7.70	4.10	6.72	4.10	6.70	4.27	6.42	4.35	5.82	3.65	2.57	2.02	0.02
227	A	7.33	13.90	6.47	11.53	6.10	8.53	6.17	7.93	6.63	7.23	6.85	5.92	5.70	2.85	2.52	0.05
229	A	4.62	9.60	4.12	8.17	3.92	6.67	3.97	6.00	4.45	5.07	4.65	4.02	3.85	1.70	1.77	0.02
230	A	10.50	23.50	9.15	19.22	8.50	16.55	8.50	16.55	10.07	14.47	10.65	11.47	8.72	4.85	3.82	0.05
249	A	8.37	16.40	7.37	13.17	6.92	10.15	7.00	9.57	7.52	8.52	7.72	7.22	6.50	3.25	3.12	0.13
250	A	18.20	32.50	16.35	30.40	14.75	26.10	14.90	25.30	15.40	23.70	15.55	22.10	12.10	8.00	5.10	0.40
251	A	13.05	25.20	11.85	21.75	10.70	19.25	11.15	18.75	11.40	17.85	11.45	16.30	8.75	8.40	6.00	0.50
255	A	17.55	32.50	15.30	29.20	14.60	25.95	14.80	25.35	15.55	23.60	16.00	21.10	12.55	7.60	5.60	0.05
259	A	9.10	18.40	8.20	15.35	7.85	13.95	8.20	13.55	8.75	12.20	9.00	10.20	6.95	4.15	3.40	0
260	A	6.10	11.30	5.70	9.90	5.20	8.50	5.25	8.30	5.60	7.55	5.80	6.90	4.60	3.05	1.85	0
262	A	8.60	13.80	7.55	11.90	7.20	10.70	7.20	10.40	7.35	9.90	7.40	9.20	5.75	4.15	2.30	0.20
267	A	6.43	14.50	6.00	11.97	5.80	10.47	6.10	8.97	8.60	7.73	6.15	6.62	5.35	2.57	2.52	0.20
268	A	9.87	18.10	9.30	16.30	9.07	14.80	9.17	14.07	9.33	13.37	9.43	12.20	7.97	5.23	4.03	0.17
271	A	14.83	26.60	13.17	23.87	12.57	22.33	12.57	22.20	13.53	19.87	13.73	18.33	11.10	7.50	5.17	0.07
272	A	4.30	9.40	3.70	7.60	3.40	5.70	3.80	5.00	4.20	4.50	4.40	3.60	3.60	1.80	1.85	0
273	A	11.00	20.60	10.40	18.30	9.60	15.40	9.80	14.20	10.50	13.00	10.90	10.00	8.40	3.60	4.30	0
274	A	5.90	11.70	5.60	10.10	5.07	8.63	5.17	8.40	5.47	7.17	5.63	5.80	4.67	2.27	2.12	0.20
278	A	15.50	29.50	13.90	24.25	13.40	20.20	13.90	19.10	15.60	16.45	16.25	13.25	12.70	4.80	6.00	0.20
279	A	10.60	22.40	9.37	18.85	8.92	16.50	9.22	16.10	9.80	14.43	10.20	11.45	8.25	4.55	3.85	0.25
281	A	4.50	8.60	3.90	7.20	3.65	5.82	3.70	5.12	4.20	4.47	4.45	3.40	3.75	1.45	2.00	0.10
283	A	10.80	22.40	9.73	19.87	9.07	17.93	9.37	16.90	10.10	15.57	10.32	13.22	8.82	5.85	4.52	0.22
288	A	13.85	22.60	12.10	19.10	11.30	17.07	11.40	16.13	11.90	13.70	12.03	11.60	10.50	5.80	4.07	0.30

Table B.19: *Dimensions of Early and early Middle Neolithic pot vessels. Continued – see Table B.20 for details.*

Neolithic pot dimensions II

Id.	Type	AX	AY	BX	BY	CX	CY	DX	DY	EX	EY	FX	FY	GX	GY	HX	HY
290	A	12.80	22.90	11.80	19.65	11.47	17.72	11.92	15.37	12.95	13.30	13.25	11.37	10.57	4.67	4.90	0.25
293	A	12.90	22.00	11.20	17.70	10.70	14.90	11.00	13.30	12.10	11.20	12.40	7.50	10.90	3.40	5.10	0
294	A	10.70	19.90	9.20	16.30	8.80	14.10	8.90	13.75	9.65	12.20	9.85	10.40	7.70	3.55	3.75	0.05
300	A	10.20	15.10	9.70	13.10	9.37	11.57	9.45	10.02	9.62	9.70	9.75	9.05	6.87	3.70	3.12	0.07
305	A	9.20	15.10	7.57	11.50	6.72	8.17	15.75	8.10	7.35	7.07	7.51	5.87	6.75	3.30	3.60	0.00
307	A	12.77	21.10	11.17	17.13	10.27	13.10	10.53	11.10	11.03	10.00	11.23	8.80	9.20	3.93	4.02	0.12
308	A	8.35	14.10	7.50	10.90	6.95	8.60	7.10	7.80	7.45	7.05	7.45	6.05	6.35	3.10	3.45	0.20
311	A	9.53	15.00	8.47	12.05	8.07	8.87	8.10	8.30	8.22	7.80	8.25	7.17	6.50	2.67	2.62	0.05
312	A	6.95	15.90	6.45	12.87	6.17	10.87	6.27	9.92	7.25	8.20	7.42	6.90	6.30	3.17	3.17	0.15
320	A	14.77	28.00	13.47	23.20	12.92	19.55	13.15	16.67	14.20	14.87	14.57	12.52	12.22	5.97	5.00	0.27
321	A	4.40	10.20	3.92	8.10	3.75	6.07	3.77	5.57	4.20	5.02	4.45	3.72	3.07	1.05	0	0
323	A	16.20	31.20	15.40	26.53	15.03	23.03	15.13	22.27	16.17	19.71	16.53	16.23	13.20	6.60	3.87	0
324	A	6.37	11.60	5.83	9.80	5.57	8.65	5.72	8.15	6.10	7.20	6.30	6.10	5.10	2.12	2.65	0.17
325	A	20.30	31.50	18.00	27.35	16.15	23.90	16.20	23.45	16.75	21.35	16.97	18.60	13.12	8.05	5.95	0.27
326	A	4.80	11.20	4.53	9.33	4.40	7.55	4.70	6.52	4.97	6.00	5.13	4.73	4.23	2.03	2.22	0.05
327	A	11.05	24.80	10.17	21.67	9.92	19.97	10.42	18.60	11.22	16.75	11.57	14.62	9.80	7.22	4.20	0.32
328	A	10.47	23.00	9.77	19.83	9.27	17.30	9.50	16.62	10.40	14.70	10.85	11.45	9.02	5.32	3.87	0.07
329	A	12.35	22.20	12.10	19.97	11.85	16.62	12.05	15.17	12.50	14.70	12.67	14.10	9.10	5.35	3.25	0
336	A	7.00	12.10	5.70	9.45	5.32	7.95	5.37	7.80	5.42	6.95	5.45	5.95	4.50	2.60	2.50	0.27
3	B	9.63	18.73	6.87	14.40	5.52	8.97	5.52	8.97	6.52	8.10	6.90	6.90	5.62	3.72	2.10	0.15
33	B	14.40	29.20	13.75	26.35	13.50	23.20	13.95	20.20	14.55	19.25	14.75	17.55	12.50	9.05	5.85	0
59	B	11.92	22.10	10.45	17.30	9.90	14.22	9.90	14.22	10.32	12.97	10.45	10.70	8.00	4.57	3.20	0.05
60	B	8.40	14.40	7.85	11.85	7.60	9.65	7.65	9.15	7.85	8.65	7.90	8.00	6.35	2.85	3.50	0.25
61	B	13.20	23.10	12.50	19.90	12.10	16.90	12.10	16.90	12.70	15.40	12.80	13.60	11.40	9.20	4.20	0
131	B	11.60	23.00	10.77	19.77	9.90	16.67	10.25	15.27	11.22	13.87	11.70	11.82	9.12	5.22	4.97	0.30
195	B	9.40	16.30	8.32	13.25	8.05	11.47	8.17	10.37	8.57	9.37	8.70	8.00	7.17	3.52	2.97	0
211	B	7.27	12.80	5.62	10.30	4.52	7.02	4.55	6.60	5.12	5.97	5.35	5.17	4.30	2.30	1.67	0.05
212	B	8.27	14.70	7.97	13.13	7.60	10.47	7.65	9.27	7.77	8.97	7.82	8.60	6.52	3.92	3.10	0.03
213	B	6.50	14.00	5.00	11.00	4.00	6.83	4.00	6.70	4.77	6.07	5.07	4.97	4.10	2.10	2.17	0
216	B	7.00	14.70	5.95	12.10	4.57	7.15	4.57	7.15	5.17	6.15	5.42	5.17	4.67	3.07	1.45	0.05
218	B	8.70	16.00	8.43	13.70	8.30	12.43	8.37	11.90	8.60	11.27	8.73	10.63	6.60	5.13	2.75	0.05
226	B	6.22	15.30	5.07	9.60	4.60	6.25	4.62	6.17	5.10	5.12	5.30	4.22	4.42	2.12	2.37	0.02
233	B	9.77	18.90	9.15	15.70	9.00	13.42	9.15	11.85	9.30	11.55	9.37	11.15	7.02	3.72	3.85	0.27
234	B	10.80	18.80	9.70	15.90	9.03	10.93	9.10	10.87	9.43	10.27	9.60	9.00	7.27	3.53	3.23	0.03
236	B	9.90	17.20	8.40	13.55	7.70	10.33	7.83	9.97	8.10	9.03	8.20	8.22	6.92	3.80	3.52	0.20
286	B	12.02	22.50	11.42	19.65	11.27	17.20	11.37	14.97	11.57	14.60	11.62	14.20	9.47	6.80	3.50	0.15
292	B	15.10	24.20	13.17	19.97	12.27	15.87	12.43	14.97	12.77	14.03	12.92	12.45	10.32	5.72	4.55	0.20
306	B	20.23	39.40	18.17	34.20	17.53	27.93	18.17	24.03	19.10	22.63	19.50	21.23	14.63	9.47	4.33	0.07
309	B	11.43	14.40	10.17	11.27	9.77	10.00	9.80	9.83	9.90	9.47	9.97	9.07	8.93	5.17	3.60	0.07
322	B	11.00	21.20	8.00	16.90	6.20	10.57	6.20	10.57	7.15	9.20	7.45	7.87	6.42	4.72	2.77	0.07
17	C	7.90	32.80	7.35	27.80	7.02	23.17	7.37	22.75	11.57	19.32	13.47	12.25	10.92	4.15	4.82	0.17
26	C	4.80	17.65	4.50	15.47	4.25	11.70	4.97	10.72	6.67	9.12	7.22	6.87	6.07	3.07	2.77	0.18
28	C	3.70	14.30	3.23	11.77	3.20	8.90	3.20	8.90	5.33	6.83	5.93	4.83	5.07	2.30	2.53	0.20
42	C	2.13	14.60	2.07	14.00	2.32	8.87	3.02	7.50	4.50	6.85	5.20	5.35	3.55	1.37	0	0
56	C	7.60	26.70	7.30	23.00	7.00	20.10	7.10	16.90	9.80	13.70	10.60	10.00	8.70	4.30	4.05	0.20
110	C	4.10	16.50	4.00	14.40	4.20	11.30	5.10	10.42	6.52	9.10	7.05	6.85	5.62	2.50	2.80	0.10
113	C	4.00	18.50	3.40	18.50	3.70	16.55	3.80	14.75	6.35	11.85	7.00	7.35	5.60	2.30	3.70	0.15
117	C	6.70	21.50	6.05	19.45	5.70	14.85	5.70	14.85	8.65	12.25	9.30	9.30	7.05	3.20	2.15	0
219	C	4.95	17.60	4.35	14.82	4.12	13.05	4.65	11.87	6.55	9.77	7.27	7.07	6.10	3.12	2.82	0.07
237	C	1.27	12.90	1.17	11.20	1.57	9.07	2.55	8.22	5.05	6.65	6.10	4.97	4.25	1.42	0.72	0
238	C	4.45	21.10	4.15	18.97	3.95	15.02	5.00	11.30	7.92	8.97	8.72	6.80	7.10	3.37	3.37	0.12
239	C	3.60	19.90	3.20	16.50	3.10	13.63	4.53	10.60	6.40	8.40	6.93	5.93	5.83	2.20	3.13	0
240	C	4.47	24.50	3.90	21.77	3.67	18.50	4.25	13.55	8.00	11.05	9.40	7.82	6.95	2.47	1.22	0
241	C	4.42	23.40	3.70	19.45	3.42	13.50	4.90	9.95	7.75	8.65	8.52	7.07	6.92	2.87	2.95	0.03
242	C	4.55	22.00	4.32	19.75	4.10	16.20	4.60	12.40	7.60	10.30	8.72	6.85	7.17	2.57	3.40	0.10
245	C	4.95	22.00	4.65	20.10	4.37	17.85	5.60	14.00	8.17	11.40	9.17	8.63	7.37	3.73	3.20	0.03

Table B.20: *Dimensions of Early and early Middle Neolithic pot vessels. Types are A = funnel beakers, B = bowls and C = flasks. The data are given in Madsen (1988b: 18) from an unpublished thesis by Eva Koch Nielsen.*

Steatite (soapstone) compositions I

Co	Cr	Fe	Mn	Sc	V	source
79.922	3651.09	85046.8	1974.9	28.718	136.6	B
85.987	3048.68	81645.0	1080.8	18.842	98.3	B
55.996	3715.52	67993.0	1086.1	19.929	87.8	B
84.972	2692.54	87749.5	1994.5	44.837	90.3	B
54.895	4377.54	85050.6	765.5	25.098	114.9	B
86.328	3126.59	76897.8	1723.8	28.278	109.5	B
98.446	3161.21	97033.0	1059.3	44.330	183.7	B
72.584	2498.93	93137.6	1477.9	31.388	179.4	B
85.537	3377.98	87725.7	1106.2	43.713	140.1	B
88.970	2496.70	85346.6	1449.3	36.250	94.3	B
81.500	2806.32	83937.6	1452.2	20.607	106.7	B
87.819	3639.65	72122.3	1570.2	38.688	108.4	B
95.089	2433.46	78117.3	2736.2	18.656	94.6	B
66.747	4140.51	81671.2	1628.6	17.693	109.5	B
82.174	2869.83	84889.0	1271.2	22.632	123.4	B
74.482	3773.00	77696.4	1034.3	38.099	134.4	B
186.279	4241.28	82515.4	769.8	35.676	116.6	B
77.931	3190.86	73053.4	984.5	33.541	120.8	B
88.912	2560.53	78735.9	1366.1	35.176	101.1	B
82.011	2234.10	71537.6	1719.1	32.274	125.2	B
120.568	2751.49	88574.6	1005.1	40.898	168.7	B
81.017	3988.95	77190.7	1589.8	19.510	135.0	B
82.411	2421.41	79235.4	1319.4	31.036	103.9	B
77.020	4472.66	72434.2	960.2	16.496	133.6	B
87.675	3035.39	60636.0	986.2	8.351	77.7	Ch
81.600	4407.28	66807.5	1043.3	9.330	86.0	Ch
73.759	3717.79	68003.5	1168.6	10.115	78.3	Ch
70.814	4113.90	63823.5	760.9	13.223	97.0	Ch
72.729	3281.03	74038.8	1030.3	10.991	101.9	Ch
66.888	4433.68	71142.0	1034.8	11.607	108.1	Ch
86.635	3653.33	58785.4	919.1	9.768	76.5	Ch
89.540	3708.10	65341.5	1075.8	10.352	76.5	Ch
64.606	4291.03	66308.3	832.8	10.733	103.9	Ch
73.728	2874.23	71408.8	1431.5	9.603	92.8	Ch
82.703	2126.38	68577.2	1222.0	7.957	48.2	Ch
61.868	2775.08	70254.7	1127.7	11.205	63.3	Ch
82.426	4421.77	74533.8	1221.2	10.009	104.3	Ch
67.676	3639.89	67706.8	1199.0	11.457	87.7	Ch
73.770	3866.04	65228.5	1106.9	9.965	94.7	Ch
69.097	4225.50	60646.9	707.9	8.454	114.8	Ch
77.361	3802.93	72153.7	1149.0	8.656	111.0	Ch
83.165	3654.83	61353.1	877.6	8.991	79.5	Ch
67.761	3829.01	64021.9	840.3	9.928	118.1	Ch
63.633	2778.38	77672.0	1047.6	9.415	90.1	Ch
88.000	3120.50	54605.0	801.9	8.925	63.6	Ch
85.000	3170.20	43014.2	657.9	6.816	66.7	Ch
66.410	3462.40	59068.0	974.8	6.432	84.0	Ch
65.170	3655.30	60649.6	1077.0	7.123	98.5	Ch
72.310	3739.40	63223.7	1153.5	11.400	84.3	Ch
138.890	3913.20	71468.7	1826.8	11.812	114.2	Ch

Table B.21: *Steatite compositions. Continued – see Table B.23 for details.*

Steatite (soapstone) compositions II

Co	Cr	Fe	Mn	Sc	V	Source
92.040	2451.40	76451.8	819.1	20.923	108.6	Cl
82.840	2378.00	64386.4	1938.3	14.768	75.3	Cl
100.680	3079.80	103244.6	528.2	14.278	108.9	Cl
89.230	2195.50	60465.3	559.9	19.088	78.0	Cl
98.410	2761.70	76687.8	1010.4	20.677	102.2	Cl
98.830	2346.40	80684.0	2425.2	14.867	91.1	Cl
89.800	2789.30	75676.5	464.7	16.385	111.5	Cl
84.430	2042.30	75668.8	409.2	17.883	93.5	Cl
95.360	1819.20	67577.3	486.3	26.087	91.9	Cl
102.830	2341.40	91330.0	1612.7	23.380	136.0	Cl
100.500	3051.90	90766.7	1348.6	17.624	119.8	Cl
107.310	2390.90	93048.6	913.6	8.369	101.6	Cl
116.080	3001.10	54015.5	1999.2	11.415	51.4	Cl
117.490	2600.20	85841.7	1155.5	14.136	99.1	Cl
98.350	3049.80	99653.6	2276.3	14.120	100.5	Cl
115.950	2159.90	75418.0	2458.6	23.301	91.2	Cl
95.470	1333.70	62553.7	1293.0	19.199	79.5	Cl
120.050	2959.00	88593.6	861.9	13.152	90.7	Cl
98.750	1974.40	61930.7	1168.1	16.894	90.7	Cl
88.270	2321.60	108413.8	1283.6	40.643	183.2	Cl
99.790	2128.10	105636.3	1900.5	46.700	178.7	Cl
101.750	2353.50	83656.8	1704.0	15.008	117.3	Cl
82.990	1486.70	73714.6	1579.1	12.377	129.3	Cl
107.740	3850.20	67945.4	464.7	11.876	184.2	Cl
114.660	3711.90	67114.3	483.7	9.490	136.3	Cl
105.360	4002.20	65773.2	661.8	12.216	81.2	Cl
68.220	1778.00	43286.4	766.3	7.514	48.0	L
65.820	1392.40	72872.7	3023.4	9.179	64.3	L
79.400	1817.20	56212.5	2463.8	13.139	25.7	L
78.510	1849.40	44407.9	727.7	7.786	29.8	L
77.280	2650.30	60557.9	468.4	13.422	60.0	L
79.950	1752.60	61094.1	483.8	7.880	44.6	L
41.240	2396.20	59655.5	318.0	5.032	26.8	L
67.640	3652.90	43790.2	466.3	2.577	67.6	L
80.700	2238.20	53133.4	562.5	11.908	37.2	L
60.680	2227.30	47506.0	688.1	8.508	54.1	L
74.620	2784.60	56947.0	730.9	8.826	68.2	L
67.240	1616.50	39345.7	1072.4	12.101	47.4	L
45.560	1554.10	40783.0	444.4	6.076	44.7	L
75.700	2140.80	51960.6	523.9	8.533	42.4	L
77.280	1926.40	47027.3	666.2	8.111	28.8	L
78.220	1842.80	48426.6	496.0	6.923	37.5	L
130.330	3833.90	78915.2	1764.1	14.447	45.6	L
68.710	2566.90	57777.7	824.5	11.586	74.6	L
97.450	3108.50	67163.9	993.5	10.638	52.5	L
65.400	1543.00	44103.7	776.8	21.566	54.7	L
73.800	1108.50	28853.5	591.0	6.678	22.9	L
78.560	2316.90	56192.8	380.9	8.993	35.6	L
67.730	2507.30	55955.9	830.8	16.484	72.7	L
56.530	2243.80	55195.2	695.1	11.890	60.6	L
55.810	2769.20	39754.2	432.0	4.772	37.9	L
66.300	1942.10	39032.9	492.5	6.234	33.4	L
71.530	1917.60	45460.6	768.2	8.581	43.0	L

Table B.22: *Steatite compositions. Continued – see Table B.23 for details.*

Steatite (soapstone) compositions III

Co	Cr	Fe	Mn	Sc	V	Source
58.040	2527.40	47353.1	762.1	5.538	70.1	L
54.350	2701.00	63554.1	736.9	10.207	68.1	L
70.950	1766.00	47650.6	393.7	7.719	48.6	L
75.100	2132.80	42379.7	407.1	8.985	59.2	L
90.072	3824.31	76071.8	1433.6	18.106	72.2	O
89.183	4005.69	73403.6	1280.0	20.187	74.4	O
71.795	3756.80	107389.4	857.7	53.524	132.3	O
80.309	4404.39	94794.9	810.2	15.956	102.1	O
86.367	3888.40	72182.4	1164.0	20.318	86.7	O
77.380	2470.15	90857.9	2771.1	40.485	75.0	O
85.727	3963.24	82954.1	1004.9	33.330	90.5	O
76.624	4254.53	96869.4	846.0	40.523	114.5	O
84.134	4265.30	93321.0	1217.5	28.727	102.4	O
83.055	4451.62	69344.3	941.6	30.615	84.7	O
67.739	1430.23	108572.9	2662.6	39.545	130.5	O
87.746	3894.78	78481.5	1571.3	30.218	77.6	O
79.231	3062.89	73564.9	845.1	39.042	55.2	O
59.658	3444.40	107507.3	3490.2	47.306	143.8	O
85.141	3849.75	90248.4	1346.4	17.641	95.2	O
65.271	2043.88	80796.1	1715.7	90.449	106.9	O
75.200	2858.29	86502.4	1798.2	47.077	93.9	O
83.711	3838.21	91770.1	1002.2	38.903	119.3	O
75.153	2402.87	91478.2	877.1	19.886	105.8	O
96.053	2180.93	91864.7	828.6	33.932	104.3	O
76.352	4399.33	90425.5	806.6	14.782	121.1	O
77.250	2185.84	81638.4	1652.4	21.609	89.4	O
75.982	585.32	136444.8	3842.4	51.701	326.6	O
88.929	2261.14	61838.2	956.3	9.122	45.2	O
78.656	4167.74	110293.6	912.0	26.532	133.5	O
72.955	3197.77	58609.2	714.9	15.859	78.8	S
82.164	3337.16	78199.5	982.0	17.649	98.2	S
73.678	2856.82	71380.8	1250.2	15.000	74.2	S
91.021	2927.90	70001.5	888.1	15.403	77.6	S
91.118	2879.10	65949.1	2236.0	20.177	111.0	S
78.231	3568.37	76089.6	1100.5	12.244	90.2	S
80.324	3709.52	85105.9	1266.1	15.387	108.8	S
112.139	3442.62	83422.4	994.8	13.954	104.5	S
78.521	3010.26	73176.3	1096.7	15.206	95.9	S
86.799	3324.53	78478.9	1217.0	13.850	94.5	S
81.734	3208.94	77612.1	1566.9	21.609	96.6	S
138.492	3706.58	90452.5	813.6	16.296	120.8	S
96.085	3528.38	81847.7	811.2	16.824	107.9	S
102.407	3245.50	66971.5	596.5	14.161	67.5	S
84.873	3149.49	68530.2	1643.1	15.034	91.1	S
98.672	2796.14	71192.6	1214.7	11.339	94.0	S
89.922	2852.13	70818.1	1371.5	13.107	68.9	S
92.321	3059.89	83270.9	1290.7	15.167	103.2	S
106.716	3549.32	73392.0	1154.1	16.797	97.1	S
93.940	3150.30	65742.7	1107.3	15.070	94.2	S
64.510	3653.20	81583.4	962.6	13.279	107.0	S
95.450	3000.50	63977.9	1387.6	14.778	95.7	S
87.350	3723.30	83794.5	909.8	16.130	104.9	S
111.510	5760.00	96060.2	1665.9	16.351	121.9	S
86.610	3439.30	76397.3	619.9	17.872	97.1	S
113.980	3485.30	72656.0	1129.4	16.418	102.6	S
71.410	2433.50	59932.4	1465.1	11.051	64.4	S

Table B.23: *Chemical composition (in parts per million) of Steatite from six quarry sites. The data are a subset of those published in Truncer et al. (1998).*

North Apulian fineware compositions I

Sample	SiO2	Al2O3	Fe2O3	MnO	MgO	CaO	Na2O	K2O	TiO2	P2O5	Ba	Sr	Y	Zr	V
C 01	57.347	16.314	6.272	0.106	2.830	11.393	0.919	3.421	0.770	0.401	452	390	26	150	125
C 02	56.468	15.066	5.835	0.089	2.285	14.722	0.743	3.062	0.742	0.810	396	387	26	147	122
C 03	58.536	15.542	6.004	0.064	2.374	12.252	0.894	2.863	0.762	0.554	353	308	30	166	106
C 04	58.596	15.649	5.824	0.064	2.295	11.992	1.019	3.035	0.765	0.601	324	299	26	175	109
C 05	61.276	16.750	6.479	0.072	2.402	7.757	0.858	3.127	0.821	0.296	361	263	30	183	133
C 06	58.692	15.956	6.005	0.097	2.189	11.083	0.841	3.471	0.739	0.755	401	411	27	142	122
C 07	56.531	15.193	5.992	0.109	2.349	14.595	0.859	2.969	0.744	0.489	389	376	27	151	122
C 08	56.951	15.833	6.606	0.102	2.642	12.566	1.014	3.011	0.753	0.358	349	376	29	158	123
C 09	58.455	15.766	5.891	0.054	2.358	11.997	1.002	2.962	0.748	0.603	330	296	30	171	108
C 10	56.811	16.462	6.443	0.105	2.803	12.038	0.868	3.242	0.773	0.293	355	394	28	151	118
C 11	59.130	15.787	6.118	0.085	2.343	10.942	0.838	3.160	0.790	0.636	399	353	29	165	106
C 12	56.546	15.899	6.245	0.095	2.689	12.861	0.921	3.303	0.749	0.498	350	432	27	142	111
C 13	61.925	14.104	5.284	0.096	2.283	10.825	1.040	3.076	0.648	0.536	455	356	23	146	105
C 14	59.927	15.469	6.707	0.085	2.267	8.720	1.430	3.835	0.716	0.636	488	427	26	165	121
C 15	61.156	14.639	5.157	0.096	2.078	10.644	1.225	3.409	0.684	0.714	449	437	24	167	122
C 16	58.029	16.235	6.198	0.097	2.488	11.281	0.887	3.440	0.760	0.389	431	408	25	140	133
C 17	60.285	15.878	5.994	0.065	2.278	10.178	0.806	3.280	0.766	0.294	362	290	28	164	124
C 18	62.364	15.580	5.972	0.095	2.273	7.811	1.131	3.351	0.739	0.497	454	336	28	168	140
C 20	56.071	15.302	5.857	0.087	2.247	15.204	0.796	2.890	0.748	0.611	396	397	27	153	134
H 01	56.356	14.827	5.966	0.087	2.589	15.165	0.852	2.994	0.686	0.306	333	356	24	144	98
H 02	57.289	15.545	6.088	0.087	2.370	12.849	1.098	3.479	0.754	0.283	280	327	26	144	128
H 03	57.593	15.543	6.035	0.086	2.369	12.713	1.168	3.259	0.742	0.332	293	349	27	153	111
H 04	56.850	14.198	5.463	0.112	2.245	15.930	1.039	3.005	0.658	0.346	285	332	23	134	99
H 05	54.835	16.191	6.338	0.096	3.355	12.942	0.937	3.877	0.750	0.490	384	484	26	141	118
H 06	54.839	16.001	6.333	0.105	3.125	13.956	1.269	3.083	0.739	0.377	343	387	26	141	128
H 07	65.580	14.952	5.509	0.063	1.857	6.188	1.096	3.527	0.712	0.376	304	240	22	168	96
H 08	57.332	15.820	5.983	0.089	2.139	12.600	1.070	3.688	0.733	0.390	288	303	24	140	122
H 09	57.870	13.800	4.571	0.086	1.718	16.253	1.135	3.307	0.603	0.486	372	343	23	163	94
H 10	57.535	15.208	5.528	0.096	2.167	13.255	1.323	3.533	0.689	0.491	360	390	25	153	112
H 11	56.495	16.783	6.199	0.074	2.498	12.271	0.928	3.489	0.759	0.337	329	392	24	129	111
H 12	58.757	15.427	5.968	0.097	2.378	12.011	0.854	3.222	0.688	0.432	339	309	24	144	112
H 13	57.693	14.568	5.359	0.087	1.947	15.159	1.017	3.008	0.656	0.350	277	336	23	137	105
H 14	59.372	17.173	6.201	0.085	2.645	8.867	0.825	3.513	0.750	0.402	307	268	24	135	133
H 15	59.779	13.771	4.381	0.087	1.659	14.422	1.149	3.448	0.596	0.542	364	305	22	165	99
P 21	59.785	14.210	5.262	0.098	2.131	13.329	1.316	2.816	0.675	0.217	405	316	27	179	84
P 23	55.016	15.491	5.944	0.113	2.409	15.479	1.160	3.141	0.733	0.315	590	418	24	139	120
P 24	53.680	16.591	6.358	0.107	2.772	15.146	1.049	2.997	0.766	0.343	462	475	26	130	127
P 25	54.625	14.375	5.512	0.102	2.094	18.076	1.087	2.920	0.677	0.351	541	321	24	140	102
P 26	54.686	13.903	5.219	0.101	2.414	18.776	0.984	2.761	0.665	0.313	459	439	23	120	109
P 27	55.787	14.561	5.180	0.111	2.634	16.117	1.089	3.146	0.674	0.511	503	440	24	139	116
P 28	55.904	14.163	5.997	0.101	2.495	16.166	1.096	2.875	0.654	0.347	671	420	23	128	101
P 29	56.912	15.864	5.884	0.110	2.316	13.525	1.230	3.041	0.748	0.209	351	321	27	143	118
P 30	53.852	14.344	5.538	0.100	2.116	18.453	1.418	2.957	0.682	0.354	603	373	25	137	93
P 31	57.757	14.495	5.487	0.089	2.488	14.684	0.900	2.899	0.704	0.322	509	364	24	116	115
P 32	57.669	14.546	5.455	0.120	2.629	13.823	1.413	3.089	0.694	0.372	499	457	25	157	110
P 33	58.216	15.326	6.546	0.103	2.563	12.012	1.184	2.789	0.732	0.360	409	418	27	141	116
P 34	54.723	16.728	6.417	0.127	2.535	13.556	1.199	3.331	0.780	0.414	511	409	27	149	129
P 35	52.996	15.142	6.214	0.129	2.800	16.886	1.292	3.123	0.709	0.517	537	481	25	120	122
P 36	57.209	14.520	5.222	0.099	2.038	16.150	1.102	2.556	0.691	0.242	453	361	24	154	77
P 37	55.477	14.252	5.819	0.109	2.352	16.669	1.367	2.811	0.668	0.306	457	402	25	134	96
P 38	53.480	15.387	6.636	0.141	2.819	15.788	1.030	3.470	0.719	0.336	517	453	24	120	112
P 39	56.429	15.086	6.191	0.097	2.436	14.604	1.084	2.908	0.717	0.279	398	400	25	136	105
P 40	57.682	16.499	6.281	0.104	2.271	11.489	1.094	3.281	0.784	0.333	476	397	27	155	135
P 41	53.188	14.437	6.174	0.113	3.522	16.649	1.242	3.477	0.680	0.327	500	464	24	124	104
S 51	51.956	12.121	4.398	0.106	2.790	23.938	1.084	2.656	0.535	0.232	377	497	18	125	77
S 52	55.694	12.506	4.451	0.127	1.825	20.734	1.081	2.510	0.547	0.325	578	442	19	120	82
S 53	54.773	13.146	4.838	0.106	2.380	19.983	1.066	2.627	0.583	0.315	434	413	21	122	86
S 54	51.959	12.554	4.387	0.120	2.582	23.576	1.063	2.657	0.557	0.318	649	558	20	124	75
S 55	54.218	14.746	5.677	0.102	2.383	17.578	1.208	2.789	0.674	0.440	387	431	23	133	115
S 56	55.651	13.913	5.245	0.127	2.068	17.943	1.145	2.792	0.618	0.292	607	397	24	130	91
S 57	57.742	15.415	5.543	0.097	2.414	13.228	1.223	3.042	0.698	0.400	516	366	25	157	126
S 58	54.126	14.466	5.540	0.122	2.638	17.840	1.022	3.078	0.643	0.342	452	421	23	127	81
S 59	58.797	15.586	5.955	0.104	2.621	11.410	1.151	3.100	0.707	0.395	424	329	26	153	115
S 60	55.082	16.255	5.825	0.080	2.554	14.631	1.079	3.238	0.730	0.309	635	480	24	129	126
S 61	59.287	14.503	5.274	0.111	2.105	13.727	1.170	2.754	0.655	0.245	416	360	26	163	90

Table B.24: *North Apulian fineware compositions. Continued – see Table B.25 for details.*

North Apulian fineware compositions II

Sample	As	Co	Cr	Cs	Hf	Rb	Sb	Sc	Th	U	La	Ce	Nd	Sm	Eu	Yb	Lu	Cu	Ni	Pb	Zn
C 01	12	13	83	5.8	3.2	100	0.6	10.3	8.0	3.5	28.5	41	22	2.9	1.0	2.5	0.20	519	42	17	96
C 02	14	12	94	4.8	3.9	80	0.6	10.9	9.2	9.9	29.8	45	22	3.6	1.2	2.7	0.20	39	37	13	93
C 03	9	10	80	4.4	3.8	80	0.4	9.8	7.8	8.5	27.1	42	18	3.3	1.2	2.7	0.36	36	40	15	93
C 04	12	11	95	4.3	4.4	70	0.7	11.0	9.6	9.6	30.6	49	26	3.7	1.2	3.0	0.27	52	41	15	106
C 05	7	15	122	7.7	5.1	90	0.9	13.5	11.6	6.2	37.6	61	24	4.5	1.5	3.4	0.30	38	48	14	109
C 06	10	10	79	4.6	3.0	70	0.5	9.8	8.0	3.2	26.9	45	22	3.0	1.0	2.3	0.19	41	41	16	106
C 07	10	12	86	5.0	3.9	70	0.4	9.9	8.1	6.7	27.1	43	17	3.3	1.0	2.3	0.18	38	39	17	91
C 08	10	15	103	5.9	4.2	90	0.9	12.2	10.1	4.1	34.3	54	30	4.0	1.4	3.0	0.22	31	44	12	99
C 09	21	14	99	6.3	4.6	80	0.9	12.8	10.8	8.7	35.2	57	27	4.1	1.5	2.9	0.26	46	41	16	109
C 10	12	12	88	6.0	3.5	80	0.6	11.0	9.0	3.5	31.0	47	22	3.3	1.2	2.6	0.18	34	40	17	90
C 11	14	14	101	6.8	4.7	80	0.6	11.8	10.2	6.4	34.0	54	28	3.9	1.3	3.2	0.26	43	39	14	99
C 12	14	18	149	10.9	5.9	150	1.1	17.7	14.8	6.3	49.9	80	33	5.7	2.0	4.1	0.58	53	48	15	102
C 13	18	16	116	8.1	6.2	110	0.9	14.7	11.9	5.4	43.1	68	26	5.0	1.9	3.8	0.30	32	36	16	94
C 14	30	16	143	8.5	6.6	130	1.3	16.3	14.8	5.9	46.9	66	27	5.5	1.9	4.2	0.59	45	42	12	105
C 15	33	16	112	7.8	6.2	120	0.8	14.5	13.7	6.0	42.4	74	28	5.0	1.7	4.0	0.28	35	39	15	89
C 16	17	17	121	7.7	5.6	130	0.8	15.5	13.2	4.3	42.5	67	27	4.8	1.6	3.6	0.49	37	40	16	94
C 17	15	18	114	5.7	5.4	90	0.8	14.5	11.9	10.4	41.1	64	41	4.5	1.6	3.4	0.30	57	40	15	93
C 18	36	15	114	8.4	6.6	100	1.3	14.5	11.9	5.4	41.4	62	35	4.6	1.6	3.9	0.36	29	41	16	98
C 20	28	15	108	6.7	4.8	90	1.1	14.0	11.5	12.2	37.5	62	26	4.5	1.6	2.9	0.50	44	37	16	87
H 01	8	16	130	8.6	5.0	110	0.8	13.2	11.9	5.3	39.3	55	32	4.2	1.5	3.4	0.20	25	51	16	80
H 02	15	13	120	7.0	4.2	70	0.4	13.7	11.2	3.1	37.4	62	34	4.1	1.7	2.9	0.29	26	38	15	80
H 03	11	14	102	6.6	4.6	100	0.7	12.8	11.0	4.3	36.3	60	27	4.2	1.4	2.7	0.26	27	40	13	91
H 04	13	12	88	5.3	4.0	80	0.5	10.7	9.2	3.7	29.5	46	28	3.5	1.2	2.6	0.13	29	36	13	88
H 05	11	14	116	7.6	3.6	100	0.4	12.1	9.4	5.4	32.1	51	20	3.4	1.2	2.7	0.20	49	58	18	112
H 06	11	15	137	9.6	4.7	100	0.8	14.5	11.7	5.8	40.2	63	30	4.6	1.6	3.6	0.19	29	49	10	91
H 07	10	12	94	5.4	4.9	100	0.7	11.0	9.3	3.1	30.3	47	23	3.5	1.3	2.9	0.31	24	38	26	82
H 08	11	14	91	5.1	3.8	70	0.7	11.1	9.4	3.6	30.5	50	22	3.3	1.3	2.5	0.21	32	38	18	112
H 09	12	14	98	7.6	5.5	120	0.7	12.1	11.3	3.8	38.2	61	29	4.4	1.5	3.4	0.46	30	33	11	79
H 10	17	14	113	7.2	4.4	90	0.9	12.6	11.4	3.3	37.7	61	31	4.4	1.4	3.0	0.20	26	42	33	80
H 11	11	14	111	8.7	4.3	110	0.7	13.3	11.1	5.1	35.9	50	29	3.9	1.4	2.8	0.24	23	44	14	93
H 12	8	14	119	7.5	3.4	120	0.4	11.5	9.8	3.9	33.2	46	18	3.6	1.4	2.6	0.12	39	55	12	97
H 13	12	14	100	6.3	4.7	80	0.9	13.2	9.9	5.5	36.4	56	29	4.2	1.5	3.5	0.24	26	36	17	92
H 14	9	19	136	8.9	4.3	140	0.5	15.2	12.5	4.5	40.6	56	31	4.4	1.6	3.4	0.23	37	61	17	104
H 15	14	13	108	7.5	5.9	100	0.6	12.4	12.6	5.2	40.6	59	32	4.6	1.6	3.3	0.24	28	31	24	75
P 21	9	12	71	4.4	4.0	80	0.6	9.8	7.2	2.9	27.1	48	18	4.1	0.9	2.1	0.31	25	36	32	76
P 23	11	11	89	4.4	3.1	70	0.6	10.4	6.6	2.5	26.5	47	17	4.0	0.9	1.9	0.32	24	43	14	91
P 24	10	12	102	5.9	2.9	100	0.5	11.8	7.3	2.8	29.8	52	17	4.2	0.9	2.0	0.32	26	49	15	100
P 25	10	10	80	4.5	2.9	70	0.6	9.7	6.7	1.9	25.6	45	15	3.8	0.8	1.8	0.27	31	41	15	87
P 26	13	10	82	4.2	3.0	60	0.5	9.4	6.5	2.8	24.9	44	14	3.7	0.8	1.6	0.27	22	43	17	78
P 27	20	11	85	4.4	2.9	80	0.4	9.9	6.6	2.6	25.7	44	17	3.8	0.9	1.8	0.30	23	44	18	84
P 28	14	9	76	4.1	3.1	80	0.7	9.2	6.3	3.5	25.6	45	15	3.7	0.8	1.7	0.28	24	38	17	80
P 29	8	12	85	4.6	3.1	80	0.7	11.1	7.4	2.2	28.2	48	16	4.2	0.9	1.9	0.32	31	47	25	99
P 30	13	11	83	4.3	2.9	70	0.6	9.8	6.8	2.4	26.0	44	14	3.9	0.9	1.9	0.29	25	42	15	78
P 31	11	9	87	4.7	2.7	70	0.5	9.9	6.7	2.6	25.5	44	14	3.8	0.9	1.8	0.30	20	37	16	91
P 32	16	10	82	4.6	3.2	80	0.5	9.9	6.8	2.0	25.9	45	16	3.9	0.9	2.0	0.28	21	43	23	92
P 33	11	11	95	5.4	3.2	100	0.7	10.9	7.6	1.9	28.9	50	18	4.3	0.9	2.0	0.31	26	45	11	98
P 34	10	14	99	5.3	2.9	90	0.7	12.0	8.1	2.1	31.8	55	20	4.5	1.0	2.0	0.31	33	54	19	111
P 35	16	11	98	5.0	2.7	80	0.7	10.7	7.0	2.5	27.2	49	17	4.1	0.9	1.8	0.28	22	46	13	89
P 36	15	11	76	4.5	3.5	70	0.7	9.4	6.9	2.0	26.0	47	16	3.9	0.9	1.8	0.31	35	40	31	90
P 37	15	10	80	4.0	2.8	50	0.6	9.5	6.4	2.1	25.5	45	15	3.8	0.8	1.7	0.27	25	41	11	84
P 38	16	13	105	5.3	2.7	90	0.8	11.0	7.4	1.6	26.5	47	18	4.0	0.9	1.8	0.30	31	60	18	99
P 39	11	11	92	5.1	3.0	90	0.7	10.6	7.2	2.5	27.9	49	18	4.1	0.9	2.0	0.30	28	51	19	90
P 40	24	12	100	5.4	3.5	90	0.8	11.6	8.0	2.3	31.6	54	20	4.7	1.0	2.0	0.33	37	50	15	93
P 41	13	12	84	4.4	2.7	80	0.6	9.9	6.9	2.3	24.9	45	16	3.7	0.8	1.8	0.31	25	53	18	82
S 51	13	10	61	4.0	3.0	59	0.6	8.7	8.1	2.8	24.8	46	20	3.8	1.0	1.8	0.27	22.3	29.7	31.8	68.9
S 52	22	13	64	4.5	3.2	100	0.7	9.4	8.8	2.8	26.2	49	18	4.1	1.0	1.9	0.30	25.6	33.0	19.7	73.3
S 53	71	12	82	5.6	2.7	92	0.4	10.5	8.7	2.9	28.3	50	20	4.3	1.0	2.0	0.31	28.5	39.7	16.6	77.5
S 54	13	11	59	4.2	2.6	81	0.5	8.9	7.5	3.6	24.8	44	17	3.8	0.8	1.7	0.25	25.6	32.8	18.0	77.6
S 55	23	13	95	6.9	3.3	104	0.8	12.5	10.7	6.3	33.9	59	30	5.2	1.2	2.4	0.35	28.7	48.6	19.2	103.4
S 56	12	13	81	4.7	3.3	88	0.7	11.0	9.9	4.0	30.7	57	26	4.6	1.0	2.3	0.35	28.7	41.6	19.7	81.3
S 57	26	13	98	6.8	3.7	128	0.7	13.1	12.0	2.8	35.3	64	33	5.3	1.0	2.6	0.36	25.2	42.0	22.0	99.4
S 58	15	13	94	5.5	2.9	108	0.7	12.2	9.8	2.5	30.4	55	22	4.7	1.1	2.3	0.35	24.5	49.4	15.3	88.4
S 59	15	14	92	6.7	3.4	101	0.6	13.1	11.4	2.8	34.6	62	28	5.3	1.2	2.3	0.35	27.9	42.3	10.5	103.3
S 60	10	13	107	7.8	3.1	142	0.9	13.9	11.7	3.1	35.6	64	28	5.3	1.1	2.3	0.35	28.0	46.1	18.1	97.7
S 61	9	15	84	5.8	4.1	103	0.6	12.0	11.1	3.0	33.4	61	26	5.2	1.1	2.3	0.35	31.4	42.6	8.2	90.3

Table B.25: *North Apulian fineware compositions. Sample site identifiers are C =* Canusium, *H =* Herdonium, *P =* Posta Crusta *and S = Santa Giusto. Sources, with analyses and discussion, are Gliozzo et al. (2013) for* Canusium *and* Herdonium, *Gliozzo et al. (2010) for* Posta Crusta*, and Gliozzo et al. (2005) for Santa Giusto.*

Appendix C

Covariance and correlation

The ideas of correlation and covariance are important in the use of several methods covered elsewhere in these notes (e.g., regression, principal component analysis and linear discriminant analysis). Some of the ideas involved are summarized here to avoid repetition in the relevant chapters. Any good intermediate/advanced text that deals with the topics involved will provide a more thorough treatment of the mathematics involved.

Measures of covariance and correlation are designed to provide information about the strength of a *linear* relationship between two variables. More than one way of defining such measures exist; apart from passing mention only the 'standard' definitions are used here. This is covered in Section C.1; Section C.2 refers back to earlier chapters with additional detail added in some cases.

C.1 Definitions and notation

Suppose we have n observations $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ on two variables, and wish to measure the strength of the linear relationship between them. For the purposes of this section define

$$X_i = x_i - \bar{x} \qquad Y_i = y_i - \bar{y}$$

where $\bar{x}$ and $\bar{y}$ are the means of the variables, so X_i and Y_i are centered on zero.

The estimated *covariance* between the variables is *defined* as

$$s_{xy} = \frac{\sum_i^n X_i Y_i}{n-1}$$

and is a measure of the linear relationship between the variables. That it is a sensible measure of linearity can be seen from Figure C.1

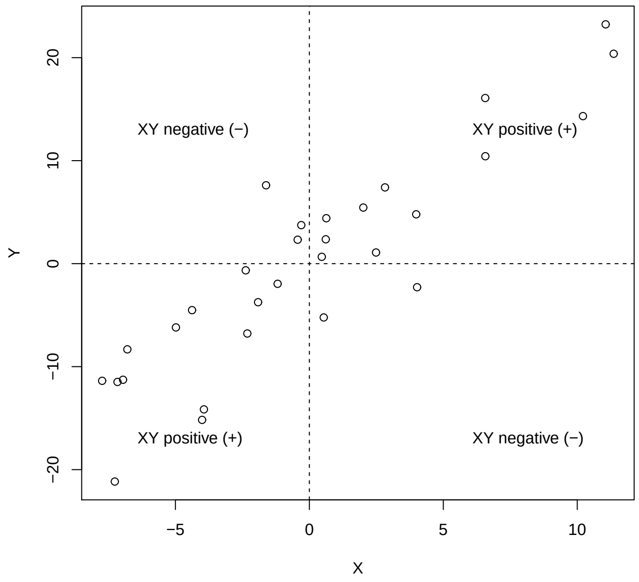

Figure C.1: *Artificial data showing a positive correlation for two variables. The population correlation coefficient from which the data are sampled is* $\rho_{xy} = 0.90$.

The plot is based on artificial data constructed to show a strong positive relationship. In the definition of s_{xy} the numerator is just the sum of terms of the form $X_i Y_i$. It can be seen from the figure that for a strong positive linear relationship these terms lie mostly in the upper-right and lower-left quadrants and are positive (remembering that the product of two negative numbers is positive) so that their sum, and hence s_{xy}, will be positive. If there is a negative relationship points will lie mostly in the upper-left and lower right quadrants, so if X_i is positive Y_i will tend to be negative (and *vice versa*) and their product will be negative. This implies that s_{xy} will be negative; if the plot is random points will be scattered around the quadrants and their effects will tend to cancel, so s_{xy} should be close to zero.

It should be obvious that the value of the numerator will be dependent on the sample size, n, and the divisor of $(n-1)$ in the definition removes this effect[1]. The other feature of a covariance, for descriptive purposes, is that the numerical result depends on the scale of the data. This means that you can't tell, just by looking

[1] The unbiased sample estimate of the population covariance is defined here, hence the divisor of $(n-1)$. Some treatments use n as the divisor which gives the definition of the population covariance. The distinction is not of great importance here except to note that the numerator is being adjusted for sample size.

at the value, whether the covariance is 'large' or 'small', and a comparison of the strength of relationship between two sets of data may not be easy. To get round this problem it is necessary to scale the covariance so it does not depend on the units of measurement.

If the covariance of a variable with itself is measured, write $s_{xx} = s_x^2$, by convention, then this is the estimated variance of x, with s_y^2 similarly defined. This gives s_x and s_y as the estimated standard deviations. The *correlation* between x and y is then defined as[2]

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

with $-1 \leq r_{xy} \leq 1$, and the correlation of a variable with itself is 1. The correlation coefficient is an estimate of the population correlation ρ_{xy}.

It is not necessary to assume that the population from which the data are drawn has a bivariate normal distribution for r (dropping the subscripts) to be useful. The assumption is necessary if formal tests of the null hypothesis $H_o : \rho = 0$ are used, but these are often not useful. If n is large then small correlations of little interest will be statistically significant. For small samples formal tests can guard against reading too much into apparently 'large' observed correlations, but usually graphical inspection is more than adequate to identify any problems.

The chapters on PCA through to that on LDA involve the analysis of multivariate data, where $p > 2$ variables are involved. All possible pairwise covariances can be calculated and summarized in the form of a $p \times p$ *covariance matrix*, $\mathbf{S}$, given by

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1(p-1)} & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2(p-1)} & s_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{(p-1)1} & s_{(p-1)2} & \cdots & s_{(p-1)(p-1)} & s_{(p-1)p} \\ s_{p1} & s_{p2} & \cdots & s_{p(p-1)} & s_{pp} \end{bmatrix}$$

which is symmetric since $s_{ij} = s_{ji}$. The *correlation matrix*, $\mathbf{R}$, is similarly defined to be the $p \times p$ matrix with typical element r_{jk}. The diagonal elements of $\mathbf{R}$ are all equal to 1; otherwise the r_{ij} lie between -1 and +1.

The total variance in a data set can be defined as the sum of the individual variances $s_1^2 + s_2^2 + \ldots + s_p^2$ which is just the sum of the diagonal elements of $\mathbf{S}$. We can write this as $tr(\mathbf{S})$, where $tr(.)$, the *trace operator*, is just the sum

[2] 'Correlation' as defined here is shorthand for the Pearson product-moment correlation coefficient, to give it its full name. This is sometimes needed to distinguish the coefficient from other definitions of correlation such as the non-parametric Spearman's or Kendall's rank order correlations. Usually, though, the shorthand suffices. It may be noted that if the data are ranked and the definition of r_{xy} applied to the ranks Spearman's rank-correlation coefficient is obtained.

of the diagonal elements of a square matrix. From this definition it follows that $tr(\mathbf{R}) = p$, since each of the p diagonal elements of $\mathbf{R}$ is equal to 1.

Suppose $\mathbf{X}$ is an $n \times 2$ data matrix, so that its correlation matrix, $\mathbf{R}$, is 2×2. The data set is said to be two-dimensional. If, however, X_1 and X_2 are perfectly correlated then the true dimensionality is really 1 since, given X_1, we know what X_2 is. If X_1 and X_2 are highly correlated then, in a sense, the data are 'approximately' one-dimensional. More generally, if $\mathbf{X}$ is a $p \times p$ matrix, but the variables are highly correlated, then the true dimensionality of the data will be somewhat less than p and it can be expected that low-dimensional representations of the data (which is what PCA and correspondence analysis attempt) will be quite successful.

C.2 Applications

C.2.1 Linear regression analysis

Linear regression analysis is covered in Chapter 5. The simplest practical linear regression model is

$$y = \alpha + \beta x + \varepsilon$$

where α and β are unknown parameters and ε is the error term. This is equation (5.1); interest commonly centers on obtaining $\hat{\beta}$, an estimate of β. The model can be given wider application by allowing simple transformations, such as logarithmic, of y and x as in equations (5.6) and (5.7).

Chapter 5 eschewed mathematical details in favor of proceeding by example. Introductory quantitative methods for archaeology texts intended for teaching purposes usually deal with simple linear regression (e.g., Shennan, 1997; Drennan, 2009). The texts cited do not derive the formula for $\hat{\beta}$ (which can, however, be obtained using basic calculus) but do give formulae, including computationally efficient versions. Shennan (1997: 137) gives

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

which in our notation is

$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{s_{xy}}{s_x^2}$$

and, if nothing else, is a neater way of expressing the result and makes explicit the role that covariance plays.

The correlation coefficient has been presented in various ways, among them

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}} = \frac{s_{ij}}{s_i s_j}.$$

Following the first'=' the two expressions are the formula for r_{xy} usually presented and a computationally efficient equivalent version if you must do calculations 'by hand' (e.g., Shennan, 1997: 140). The final version, apart from being more compact, makes it clear that the correlation is the covariance rescaled to allow for the different 'spread' of the variables.

Software output for regression analyses typically report an R^2 value – the *coefficient of determination*. For the special case of simple linear regression R^2 is just the square of the correlation coefficient, r_{xy}^2, and can be interpreted as the amount of variation in y 'explained' by variation in x.

C.2.2 Principal component analysis

Chapter 7 dealt with PCA largely by illustration. The mathematics is dealt with in more detail in Appendix D; here only brief notice is provided of the role played by the covariance matrix.

If $\mathbf{S}$ is the covariance matrix of $\mathbf{Y}$, the data matrix used for analysis,

$$\mathbf{S} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$$

where $\mathbf{V}$and $\mathbf{D}^2 = \mathbf{\Lambda}$ are $p \times p$ matrices, and the latter is diagonal with the diagonal elements, assuming they are ordered, corresponding to the variance of the PCs. The elements of the latter, λ_i, are *eigenvalues* – a term that features in some software output. The columns of $\mathbf{V}$ are *eigenvectors* and contain the coefficients of the PCs. Thus, all the ingredients for the practical analysis and interpretation of a PCA can be obtained from the covariance matrix $\mathbf{S}$ (or $\mathbf{R}$ if the data are standardized). Principal component scores are also required and these are obtained from the $n \times p$ matrix $\mathbf{YV}$.

C.2.3 Mahalanobis distance and LDA

More on MD and notation

Define $\mathbf{p}$ to be a vector with p terms, with $\mathbf{q}$ similarly defined. Mahalanobis distance (MD) can be defined as

$$\tilde{d}^2 = (\mathbf{p} - \mathbf{q})'\tilde{\mathbf{S}}^{-1}(\mathbf{p} - \mathbf{q})$$

where definition of the terms depends on the particular version of MD used. If there are G groups denote the estimated covariance matrix of group g as $\mathbf{S}_g$, using $\mathbf{S}$ for the situation where there is just one group. Assuming common population covariance matrices a weighted average can be defined as

$$\mathbf{S}_w = \sum_{g=1}^{G} \frac{(n_g - 1)\mathbf{S}_g}{(N - G)}$$

where $N = (n_1 + n_2 + \ldots + n_G)$ is the sum of the sample sizes, n_g, for each group. The term $\mathbf{S}_w$ is the *within-group* covariance matrix. It can be thought of as a measure of the 'compactness' of the groups. A particular case is when there are two groups, where

$$\mathbf{S}_w = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{(n_1 + n_2 - 2)}.$$

If $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the means for the two groups, MD as defined earlier becomes

$$d_{12}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

The other case of interest here is that of the MD of a single case, $\mathbf{w}_i$, from a group with mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{S}$; this can be written as

$$d_i^2 = (\mathbf{w}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{w}_i - \bar{\mathbf{x}}).$$

Two situations can be distinguished; the first is that $\mathbf{w}_i$ is not a member of the reference group, the second is that $\mathbf{w}_i = \mathbf{x}_i$ is a member of the group. The latter situation introduces the complication that $\mathbf{x}_i$ influences the calculation of the group statistics $\bar{\mathbf{x}}$ and $\mathbf{S}$. This has the effect of reducing the size of the MD compared to calculations that omit it. The obvious way to remedy this is to use leave-one-out (LOO) calculations (Section 11.2.2) where $d_{(i)}^2$, $\bar{\mathbf{x}}_{(i)}$) and $\mathbf{S}_{(i)}$ replace corresponding terms in the above formula, the (i) subscript indicating that case i has been omitted from calculations[3]. The implication would seem to be that to use LOO calculations n separate analyses are needed as the mean and covariance calculations change for each case. This is not necessary because it is possible to convert d_i^2 to $d_{(i)}^2$ without the need for such calculations (e.g., Section 6.4.3 of Baxter, 2003).

MD in LDA

So far MD has simply been defined; some insight into, and 'justification' for, it can be gained by considering LDA for the two-group case assuming equal population covariance matrices. With $G = 2$ there is one discriminant function that, for case i, gives rise to scores of the form $\mathbf{a}'\mathbf{x}_i$ and the task is to determine the p elements of $\mathbf{a}$ that maximize group separation on the transformed scale.

The within-group sample covariance matrix, $\mathbf{S}_w$, has been defined above for two groups. It can be thought of as an averaged 'measure' of the 'compactness' of the groups. It is also possible to define a between-groups sample covariance matrix which, in weighted form and for two groups can be written as

$$\mathbf{S}_b = n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})' + n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})'$$

[3] The term *jackknife* is also sometimes used for LOO calculations.

where $\bar{\mathbf{x}}$ is the mean of all the data. This can be thought of as a 'measure' of the 'separation' of the two groups.

After performing a discriminant analysis, which requires $\mathbf{a}$ to be determined, the centroids of the two groups of transformed data can be denoted as $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$. We want the distance between these to be as large as possible, and this depends on $\mathbf{S}_b$, but this needs to be balanced against the wish to keep the groups as compact as possible which depend on $\mathbf{S}_w$. Fisher's (1936) idea, which does not depend on distributional assumptions, was to determine $\mathbf{a}$ to maximize the ratio

$$\mathbf{a}'\mathbf{S}_b\mathbf{a}/\mathbf{a}'\mathbf{S}_w\mathbf{a}.$$

Thus the initial idea of LDA, with the desire to transform the data so that the predefined groups are as distinct has possible, has been converted to a mathematical problem the solution of which is to obtain $\mathbf{a}$ from the eigenvectors of

$$\mathbf{S}_w^{-1}\mathbf{S}_b.$$

It further transpires that the Euclidean distance betwwen the transformed group means, $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$, is given by

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

which is just the MD as defined, for the comparison of group means, but without explanation, at the start of this section. Here we have seen how MD arises 'naturally' as a measure of distance in the context of LDA.

Constructing confidence ellipsoids

For a single group, when LOO calculations are appropriate, the set of values for which $d_i^2 = c$, where c is a constant, define ellipsoidal contours. To assign a confidence level to the contours it is necessary to assume bivariate normality for two-dimensional plots. Theory exists that shows that a suitable transformation of d_i^2 exists that is approximately distributed as an F-statistic with $(p, n-p-1)$ degrees of freedom, or a chi-squared statistic with p degrees of freedom if n is large in relation to p The choice of a value for F or chi-squared determine the confidence level, The formulae are given in Baxter (2003: 71) with references to the original theoretical derivations.

It can be noted that such calculation allow probabilities of group membership to be assigned to individual cases, so it can be judged if they are plausible group members or not. This differs from the usage in Table 11.1 where results are presented in terms of relative probabilities that assume cases must belong to one or other of the groups included in an analysis.

Appendix D

More on PCA and factor analysis

D.1 Introduction

This chapter provides an account of some of the mathematics that underlies the derivation of principal components, and factors in factor analysis. It providea a source of reference for Chapter 7 and 8 without burdening those chapters with too much algebraic detail. There are fundamental differences between factor analysis and PCA that can be presented in starker relief than is possible using purely verbal descriptions, among them the dependence of factor analysis on a model for the data, and the role that rotation plays in applications of the methods. Rotation is central to factor analysis and an option in PCA, not exercised that much except, perhaps, in papers that confuse the two methodologies (Section 8.4).

A reservation often expressed about the use of factor analysis is what might be called the 'unavoidable indeterminacy' of factor analysis solutions. The model that forms the basis of factor analysis is expressed in terms of unknown parameters that must be estimated to obtain a 'solution'. A variety of choices need to be made to obtain a specific solution, among them distributional assumptions about the nature of random variation in the model, the method of factor extraction adopted, the number of factors selected for subsequent rotation, and the choice of method of rotation itself. What has been published in archaeological applications is often a matter of convention and convenience (e.g., what has gone before, often dictated by defaults in software packages). Earlier applications were dominated by PCA with varimax rotation, wrongly regarded as 'factor analysis' (Section 8.4). It is not an accident that these were the default options in widely-used commercial statistical software packages.

Any model-based method of statistical analysis involves similar kinds of choices (e.g., distributional assumptions, method of estimation) but the outcome of a factor analysis is perhaps more subject to the choices made than other methods. By contrast, PCA depends on the method of data pre-treatment chosen (Section 7.2) but, given this and as usually employed, the results from a PCA are obtained by mathematical means, leading to a unique 'solution' that does not involve assumptions about random variation, the need for estimation and so on. As emphasized in comparative reviews such as the statistical texts of Krzanowski (1985), Everitt and Dunn (2001) and Jolliffe (2002), the methods have different aims and should not be viewed as 'competitors'.

The issue of the indeterminacy of factor analysis solutions has generated a considerable literature. The choices made may sometimes not matter much, though this an empirical matter best resolved on a case-by-case basis. Some scholars are not troubled by such indeterminacy, but sceptics express concern that it allows unwarranted latitude in selecting outcomes that conform with 'theoretical' preconceptions about the phenomenon studied.

Whatever the view taken, it is a fact that any data set can be subjected to a large number of different specific factor analyses. The `fa` function from the `psych` package, used in the examples of Section 8.3.2, has the option of six different methods of factor extraction and 15 methods of rotation – eight orthogonal and seven oblique (Section D.3.2) – so 90 in all. Some of these can be expected to produce similar results, but scope for variation exists. The illustrative examples of Section 8.3.2 show some of the variation that can occur. Other (non-archaeological) illustrations are provided by Jolliffe (2002: 161–65). For convenience of reference Section D.4 summarizes some of the possible sources of indeterminacy in factor analysis applications; these are referred to in Section 8.3.2 without much additional discussion.

D.2 The singular value decomposition

For p variables, $(Y_1, Y_2, \ldots, Y_p)$, define p principal components, $(Z_1, Z_2, \ldots, Z_p)$, where component j is

$$Z_j = \mathbf{a}'\mathbf{y} = a_{j1}Y_1 + a_{j2}Y_2 + \ldots + a_{jp}Y_p$$

and $\mathbf{a} = (a_{j1}\ a_{j2}\ \ldots\ a_{jp})$ is a $(p \times p)$ column vector with transpose $\mathbf{a}'$, a $(p \times 1)$ column vector, and $\mathbf{y} = (Y_1\ Y_2\ \ldots\ Y_p)$ a $p \times 1$ column vector. Define $\mathbf{Y}$ and $\mathbf{Z}$ as $n \times p$ matrices of the data and component score with typical elements y_{ij} and z_{ij} with $\mathbf{A}$ the $p \times p$ matrix of coefficients with typical element a_{ij}; then

$$\mathbf{Z} = \mathbf{Y}\mathbf{A}'. \tag{D.1}$$

The data matrix $\mathbf{Y}$ can be factorized, using the *singular value decomposition* (SVD), as

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}' \tag{D.2}$$

where $\mathbf{U}$ is $n \times p$, $\mathbf{V}$ is $p \times p$ and $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$ the $p \times p$ identity matrix. The matrix $\mathbf{D}$ is diagonal; the diagonal elements are the *singular values*, σ_i, and their squares, $\lambda_i = \sigma_i^2$ are *eigenvalues*. The matrix with diagonal elements given by the eigenvalues is $\mathbf{\Lambda}$.

Let $y_{ij} = (x_{ij} - \bar{x}_j)/(n-1)^{-1/2}$ (i.e. it is centered, but not standardized, and rescaled for convenience). Then

$$\mathbf{Y}'\mathbf{Y} = \mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' \tag{D.3}$$

where $\mathbf{S}$ is the covariance matrix of the data (Appendix C)[1]. It follows from equation (D.3), on post-multiplication by $\mathbf{V}$, that

$$\mathbf{S}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}.$$

By definition the columns of $\mathbf{V}$ are the *eigenvectors* of $\mathbf{S}$, with the diagonal elements of $\mathbf{\Lambda}$ the associated *eigenvalues.*

From equations (D.1) and (D.2) $\mathbf{Y} = \mathbf{Z}\mathbf{A}' = \mathbf{U}\mathbf{D}\mathbf{V}'$ define $\mathbf{Z} = \mathbf{U}\mathbf{D}$ and $\mathbf{V} = \mathbf{A}$. Thus column j of $\mathbf{V} = \mathbf{A}$ is the jth eigenvector of the estimated covariance matrix $\mathbf{S}$ and λ_j is the jth eigenvalue. Furthermore, if $\mathbf{S}_Z$ is the covariance matrix of $\mathbf{Z}$,

$$\mathbf{S}_Z = \mathbf{Z}'\mathbf{Z} = \mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{D} = \mathbf{D}^2 = \mathbf{\Lambda}.$$

The sum of the diagonal elements of $\mathbf{\Lambda}$ (the eigenvalues) is the total variance of the variables defined by $\mathbf{Z}$. It can be shown that this is the same as the sum of the variances of $\mathbf{Y}$ – that is, the sum of the diagonal elements of $\mathbf{S}$ (Baxter, 2003: 68). This shows that the linear combinations, Z_j, are uncorrelated (because $\mathbf{\Lambda}$ is diagonal); that the variances of the Z_j are the eigenvalues of $\mathbf{S}$; and that (by arrangement) the Z_j are ordered in terms of importance as measured by the variances. The variances of the Z_j 'redistribute' the variances of the original data.

These are the properties required of principal components[2]. Numerical values for the a_{ij}, component variances and so on can be obtained by extracting eigenvectors and eigenvalues via the SVD. The necessary computations are applied in `prcomp` and other functions in `R`.

D.3 The factor analysis model

D.3.1 The model

The fundamental difference between factor analysis and PCA in that a statistical model needs to be formulated for the data in factor analysis whereas PCA as

[1] Note: $\mathbf{Y}'\mathbf{Y} = \mathbf{V}\mathbf{D}'\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{V}\mathbf{D}^2\mathbf{V}' = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$ from previous results/definitions.

[2] This development uses the covariance matrix (i.e. unstandardized data). For standardized data the covariance matrix is the correlation matrix $\mathbf{R}$, and defining $y_{ij} = (x_{ij} - \bar{x}_j)/s_j(n-1)^{-1/2}$ so that $\mathbf{S} = \mathbf{R}$ does not affect the development.

usually applied implies no such model (Jolliffe, 2002: 151)). From equation (D.1), $\mathbf{Z} = \mathbf{YA}'$; from the SVD of equation (D.2) $\mathbf{V}'\mathbf{V} = \mathbf{I}$; and following from these $\mathbf{V} = \mathbf{A}$ so $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Thus, on post-multiplying both sides of the expression for $\mathbf{Z}$ by $\mathbf{A}$

$$\mathbf{Y} = \mathbf{ZA} \tag{D.4}$$

or

$$Y_j = a_{1j}Z_1 + a_{2j}Z_2 + \ldots + a_{pj}Z_p.$$

This might be thought of as a 'model' for the data, but in fact is simply a mathematical consequence of the way that components are defined as linear function of the variables, showing that the latter can be expressed as linear functions of the components. This does not involve distributional assumptions of the kind typically associated with models.

In contrast, factor analysis involves a model

$$Y_j = b_{1j}F_1 + b_{2j}F_2 + \ldots + b_{qj}F_q + \varepsilon_j$$

where ε_j is an error term with variance ψ_j. The matrix formulation for this is

$$\mathbf{Y} = \mathbf{FB} + \boldsymbol{\varepsilon}. \tag{D.5}$$

There are important differences between this and the PCA formulation of equation (D.4).

1. Unlike PCA, factor analysis involves a statistical model for the data, expressed as the sum of a systematic and random component.

2. In equation (D.4) $\mathbf{F}$ is an $n \times q$ matrix of unobserved factor scores where $q < p$. If p is large q will typically be a lot smaller. The F_j are *common factors*. In PCA the number of components, p, is known; in factor analysis q is not, and determining or confirming a suitable value is an aspect of the analysis.

3. In equation (D.4) $\mathbf{B}$ is a $q \times p$ matrix of parameters that must be estimated. Determining coefficients in the PCs is a mathematical exercise.

4. Assumptions are needed for estimation; minimally, that errors are uncorrelated and common factors are uncorrelated with the errors and each other. This last assumption can be relaxed, giving rise to *oblique* factors.

D.3.2 Rotation

There is another additional and important difference between the two methods. As seen from equation (D.3) the PCs and their variances can be determined by finding the eigenvectors and eigenvalues of the covariance matrix of the data, $\mathbf{Y}'\mathbf{Y} = \mathbf{S}$. Subject to the constraint imposed on the eigenvectors and the properties required of the PCs this leads to a unique solution.

Given the assumptions of factor analysis the covariance matrix can be written as

$$\mathbf{S} = \mathbf{B}'\mathbf{B} + \mathbf{\Psi} \tag{D.6}$$

where $\mathbf{\Psi}$ is a diagonal matrix with diagonal elements ψ_i. If $\mathbf{T}$ is a $p \times p$ orthogonal matrix (i.e. $\mathbf{T}'\mathbf{T} = \mathbf{I}$) then, defining $\tilde{\mathbf{B}} = \mathbf{TB}$ it follows that

$$\tilde{\mathbf{B}}'\mathbf{T}'\mathbf{T}\tilde{\mathbf{B}} = \mathbf{B}'\mathbf{B}$$

and $\mathbf{TB}$ is as valid a solution to equation (D.6) as $\mathbf{B}$. This solution is called an *orthogonal rotation* and it can be obtained in innumerable ways.

To obtain specific solutions for the factor loadings some choice of $\mathbf{T}$ is needed which requires the imposition of constraints on the factor loadings. Almost invariably 'simple structure' is aimed for, ideally resulting in factor loadings that are either 'high' or close to zero. Ideally a variable will have a high loading with respect to only one factor. Factor analysis, and ideas of rotation are driven by the desire for 'interpretability'; the assumption is that the covariances/correlations between observable variables reflect their relationship with a smaller number of common factors (or latent variables) or constructs that can be assigned some sort of (theoretical) 'meaning' within the domain of study involved.

Matters are complicated by the view that if factors do represent some aspect of an unobservable 'reality' there is no particular reason to expect them to be uncorrelated (Cattell, 1978, 128; cited in Jolliffe, 2002: 152). That is, orthogonal rotation does not lead to the identification 'of correct factor structure'. To allow for this, methods of *oblique rotation* have been developed where the constraint that $\mathbf{T}'\mathbf{T} = \mathbf{I}$ is dropped, leading to the identification of correlated factors[3].

An obvious question is 'how should the choice of rotation method be made?'. There is no prescriptive answer to this question

D.3.3 Factor extraction

No attempt is made here to provide a comprehensive mathematical account of methods of factor extraction, which necessarily precede rotation. The references

[3] The notion of 'correct' factor structure, explicit here, leads to the thought there is no obvious logical reason why 'correct' structure should also be 'simple' – that is, just because a factor is 'interpretable', and why simple structure is sought, doesn't make it 'real' or 'correct'.

given at the end of the chapter may be pursued for technical details. Binford and Binford's (1966) seminal paper that popularized the use of 'factor analysis' in archaeology was, in fact, PCA with rotation, as was the majority of 'factor analysis' applications to the mid-1980s when usage began to decline. It is a commonplace observation in the statistical literature that treating unrotated PCA as factor analysis is wrong. The same is true of PCA with rotation since it takes no account of the error structure which is one of the distinguishing features of the factor analysis model (Section 8.4).

It is, however, possible to use principal component ideas applied to the 'reduced covariance matrix' found from equation (D.6).

$$\mathbf{B}'\mathbf{B} = \mathbf{S} - \boldsymbol{\Psi}$$

as a starting point in iterative methods of estimation. This requires an estimate of $\boldsymbol{\Psi}$, leading to many specific varieties of factor analysis, none of which have any great claim to 'absolute validity' (Jolliffe, 2002: 159). This method of *principal axis* factor analysis has been one of the most commonly used methods of factor extraction when not confused with PCA.

Extraction methods do not necessarily require distributional assumptions to be made about the error terms, other than that they are independent with zero mean. The statistically more 'thoroughgoing' method of maximum-likelihood does, however, require the strong assumption that the errors have a multivariate normal distribution. The method has the advantage, unlike PCA and unlike other methods of factor analysis, that results do not depend on data standardization; it also facilitates the use of inferential methods to assess the quality of results obtained. The 'downside' is that the multivariate normality assumption might often be considered to be unrealistic, though the estimates have some reasonable properties even when normality does not hold[4].

Jolliffe (2002: 156–57) draws an analogy with least squares regression which can be applied regardless of distributional assumptions, but produces maximum-likelihood estimates and inherits their optimality properties if normality (of the errors) can be assumed. A variety of least squares methods exist for parameter

[4] One quite often reads the assertion that factor analysis and PCA require normality assumptions. Other than MLE this is not a requirement in most applications. There is sometimes confusion between normality of the error terms and 'normality' of the variables. As far as the former is concerned the usual PCA formulation does not even involve an error term. As with the commoner applications of regression analysis, where the same erroneous statement is sometimes found, there is no requirement of a normal distribution for the variables. What is the case is that if the observed distribution of variables is 'badly-behaved', containing clear outliers for example, this may unduly affect results, and the problem(s) need to be identified and remedial action taken. This is a matter of practical data analysis rather than the imposition of unnecessary distributional requirements. Variables can be perfectly 'well-behaved' for the purposes of data analysis without approaching a normal distribution.

estimation in factor analysis, several of which are available in the `fa` function in the `psych` package for `R`. The default method of factor extraction in `fa`, `minres` – 'minimum residual' using Ordinary Least Squares estimation – is noted in the documentation to produce 'solutions very similar to maximum-likelihood even for badly behaved matrices', and weighted and generalized least squares (`wls` and `gls`) methods are also available. The documentation also states that maximum-likelihood 'is probably preferred' provided the data are well-behaved.

The default method of rotation in `fa`, `oblimin`, is oblique, replacing the orthogonal method, `varimax`, that was the default in earlier versions. Varimax is the most widely used method in published applications, probably for historical reasons (Jolliffe, 2002: 153–54; Section 8.4).

D.4 Discussion

Here and in Chapter 8 it has been emphasized that PCA and factor analysis are different methods of analysis that should be clearly distinguished. The distinction has not always troubled practitioners and comment to the effect that the methods will often 'produce similar results' is not uncommon (see Section 8.4).

The choices leading to indeterminacy in factor analysis are mostly not relevant in PCA and may produce substantively different interpretations. What are the important choices?

1. *Factor extraction* can make a difference. It might be expected that those methods that require prior estimation of $\mathbf{\Psi}$ will, if the factor model is reasonable and a sensible method of estimation used, produce fairly similar results (Jolliffe, 2002: 159). If, as the `fa` documentation claims, least-squares methods approximate maximum-likelihood it may not be worth troubling too much about which is used unless the latter is wanted for inferential purposes. The example in Section 8.3.2 contrasts methods towards either end of the 'spectrum' in terms of the assumptions needed – principal factor analysis and maximum-likelihood – to see if the choice makes much difference.

2. *Rotation* of factors can be undertaken in many different ways. Jolliffe (2002: 271) suggests, within the class of orthogonal rotations, the choice 'often makes little difference to the results'. The choice between orthogonal and oblique rotation presumably matters; there would otherwise be no reason for developing the latter. Section 8.3.2 provides an example comparing varimax and oblimin rotations.

3. The *number of factors to rotate* is a choice that Jolliffe suggests may be more important than the factor extraction and rotation methods used. Kaiser's

rule and variants of it is a commonly used factor selection method. Where the choice of factors to rotate is not obvious experimenting with different choices, to see if interpretation is much affected, seems sensible.

Interrogating data using different methods and different variants of a method can lay one open to the accusation of being 'unscientific' and 'fishing' for palatable interpretations compatible with initial preconceptions. This *is* a danger. An implication – this is something of a caricature, but not entirely so – is that the proper approach is to make a principled choice of method and live with the consequences. This is a counsel of perfection open to the possible counter-accusation that such a 'principled choice' may itself conceal a variety of strongly held preconceptions and methodological/philosophical views.

The view adopted in these notes is that the exploration of different methods of data analysis is legitimate and sensible. It is in the selective reporting only of those results palatable to the investigator that the danger lies. The dictates of publication doubtless produce pressure to concentrate only on 'significant' results, but it would be helpful for potential consumers of a method to be made aware of circumstances when they don't 'work' and why.

Early developments of factor analysis occurred in the psychometric literature (e.g., Spearman, 1900). Statistician's came late the subject; the date and title of the first edition of Lawley and Maxwell's (1963) text, *Factor Analysis as a Statistical Method*, with the emphasis on *as a Statistical Method*, testifies to this. The development of efficient computational software during and either side of the 1960s promoted the wider use of factor analysis, archaeology not excluded. The popular `SPSS` package first surfaced in 1968; from some perspectives this popularity, and its longevity, has not necessarily been beneficial. The treatment of PCA as a particular case of factor analysis has engendered confusion and is to be regretted.

General statistical texts on multivariate analysis begin to appear in any profusion round about this period. Anderson's (1958) early text was a theoretical treatment and other texts such as Morrison (1967) appear in the 1960s. A 'flurry' of books appeared in the late-1970s to mid-1980s, among them Mardia *et al.* (1979), Chatfield and Collins (1980), Seber (1984) and Krzanowski (1988). The first edition of Jolliffe (2002), with its more specialized focus on PCA, appeared in 1986. Other texts noted elsewhere in these notes that include comparisons of PCA and factor analysis are Krzanowski and Marriott (1995) and Everitt and Dunn (2001). None of these can be described as 'recent' but the underlying mathematics and ideas don't change. What has changed is the ease and flexibility of implementation in packages such as `R`. Statistical developments, of more complex and computer intensive methods with essentially similar aims, have taken place but not yet influenced archaeological practice much. The framework that supports most archaeological applications of PCA, factor analysis and other multivariate methods was in place well before the turn of the century and remains valid.

Chatfield and Collins' (1980) take on factor analysis is, as already noted, a negative one (Section 8.4). Other statistical treatments have been more even-handed. Jolliffe's (2002) emphasis that factor analysis and PCA are different rather than competing methods, and that factor analysis, if appropriate to an analysis and properly executed, has its place as part of the analyst's 'toolkit', is echoed in similar statistical texts.

The questions for archaeologists are whether factor analysis is an appropriate tool for much of what they want to do and, if so, has its value in application been convincingly demonstrated? The first question is an interesting one, and for archaeologists rather than statisticians to address. My own view on the second question, nearly 50 years on from the publication of Binford and Binford (1966), it that is difficult to answer in the affirmative.

Index