

Hatching seeds before they're counted

Graphical methods for interpreting archaeobotanical data

Walton A. Green

Received: 9 October 2008 / Accepted: 20 January 2009 / Published online: 17 February 2009
© Springer-Verlag 2009

Abstract Like other analytic aspects of archaeology, archaeobotany has been growing progressively more quantitative in the past few decades. This may be a sign of the proliferation of increasingly mature and sophisticated methodologies for analyzing botanical data, but associated with the sophistication of quantitative methods is their inherent opacity: the value and applicability of anthropological conclusions drawn from quantitative archaeobotanical data are not only limited by the amount of information that can be extracted from data by sophisticated statistical tools, but also by our ability to draw reasonable anthropological—as opposed to merely statistical—conclusions. Even the words “classification” and “significance” have different meanings in statistics and in anthropology. In this paper, I propose the use of *graphical analysis* for archaeobotanical data in addition to, or instead of, typical statistical tools like significance tests, variable reduction, and clustering. Applied to data from charred seed assemblages from the ancient Near East, the visual representation of quantitative data has the advantage

of handling semiquantitative data better and being interpretable without reliance on the paradigm of a formal statistical test.

Keywords Archaeobotany · Ancient Near East · Tell Brak · Data analysis · Quantification in archaeology · Graphical analysis

Introduction

With the increasing hegemony of quantification in archaeobotany (and, indeed, in the social and natural sciences in general), results that depend on complex statistical arguments are often viewed with scepticism. Much of this suspicion is borne by a set of statistical tools commonly called “multivariate statistics,” which I will call “traditional” or “standard” multivariate statistics to distinguish them from graphical analysis, another method of dealing with multivariate data. Books like *How to Lie with Statistics* (Huff 1954) and aphorisms like the tricolon crescendo “Lies, damn lies, and statistics” indicate the skepticism with which statistical results are sometimes viewed. Some archaeobotanists have pointed out specific limitations of multivariate methods (e.g., Hastorf 1993; Hubbard and Clapham 1992) and more archaeologists simply ignore bodies of work that seem to depend too heavily on statistical arguments.

One reason for this sort of skepticism is that mathematical statistics has become a complicated technical field with many procedures (developed by statisticians) that are either inherently complex or explained in complex technical language. Therefore, the average

W. A. Green
Pitt-Rivers Laboratory,
McDonald Institute of Archaeological Research,
Cambridge University, Cambridge, UK

Present Address:
W. A. Green (✉)
National Museum of Natural History,
Smithsonian Institution,
P.O. Box 37012, MRC 121,
Washington, D.C. 20013-7012, USA
e-mail: wagreen@bricol.net

archaeologist frequently has no way to evaluate the legitimacy of a statistical conclusion.

Another problem is that the assumptions behind statistical tools limit the questions to which they can usefully be applied. This can be stated in technical language (e.g., “every exact probability is conditional upon a finite set of possible outcomes” or “every test of a hypothesis is a test of multiple joint hypotheses”) (Quine 1953; Lakatos 1978). This amounts to the argument that the world is a complicated place and making simplifying assumptions so that the mathematics work out nicely may lead to missing something important.

By the time they are used in archaeology, standard multivariate statistical tools are seldom or never wrong per se, but they are frequently misapplied. Even more frequently, they are correctly applied but contribute nothing of anthropological interest to the question being addressed. Differences in terminology and confusion over concepts allow reporting of results that are statistically correct but irrelevant to the anthropological or historical questions of interest.

In archaeobotany, standard multivariate methods like principal components analysis are increasingly becoming routine tools for data analysis. Therefore, it now seems important to examine their use critically, and make to the suggestion that “graphical analysis,” or the “visual display of graphical information” (Tufte 1983), is a viable alternative or accessory for the routine analysis of archaeobotanical data. This is a solution that has been suggested by statisticians many times in the past few decades (Anscombe 1973; Chernoff 1973; Tukey 1977; Wang 1978; Kleiner and Hartigan 1981; Tufte 1983; Cleveland 1985) but is still underemployed in archaeobotany.

In this paper, I begin by examining some theoretical limitations that arise in archaeobotanical practice, and then apply a form of graphical analysis to a small data set from the fourth millennium BC in what is now eastern Syria in order to show how it copes with some of these theoretical limitations better than standard multivariate statistics. There is a vast and disparate literature on graphical tools for data analysis, which cannot be summarized or reviewed here. Instead, I hope to open a window into an area of statistical thought that is not covered in introductory statistics courses or textbooks for social scientists, and provide one example of how graphical analysis can be applied in archaeobotany.

Semiquantitative data and excess precision

Whatever methods are used to accumulate and analyze archaeobotanical data, at some point in the process

of collection and analysis, the data are represented as a matrix of numbers. Typically, for instance, taxa or anatomical parts form the row labels and the sample names form the column labels of such a matrix. However, the “numbers” that appear in such a matrix are seldom real, fully quantitative “numbers” in the same way as the “five” in the statement “I have five fingers” is a number. Instead, they are numbers like the “five” in “I’ll be there in 5 min.” In the former case, the “five” means that the speaker has five fingers, as opposed to four-and-a-half or three; exactly corresponding to the integral number five; in the latter case, however, “five” could well mean four-and-a-half or (one frequently finds) as many as 25, but is unlikely to mean 5 s or 5 h. In other words, the phrase “five minutes” conveys more information than “pretty soon,” but not as much information as “a number of minutes well approximated by a gaussian random variable with mean 5 and standard deviation 2.” Furthermore, the meaning of “five minutes” varies depending on the reliability of the person using the phrase, the tone of voice in which it was uttered, and the context in which it is said. “Your boiled egg will be ready in 5 min” differs from “just 5 min while I print it out.”

In the same way, a count of 213,987 barley grains in a typical archaeological sample conveys little more information than “about 200,000” or even “thousands of barley grains.” Providing numbers with too many significant figures is sometimes referred to as “excess precision,” but the distinction between “precision” and “accuracy” is not the same as the difference between the quantitative “fingers’ five” and the semiquantitative “minutes’ five”. Unlike accuracy, the “minutes’ five” is fundamentally descriptive, not quantitative.

This kind of unquantifiable error or uncertainty has been pointed out in several different fields. In instrumental analytic chemistry, the term “semiquantitative” is used: numbers produced by an uncalibrated instrument run are referred to as semiquantitative data—useful for internal comparison or order-of-magnitude estimates, but not to be confused with elemental concentrations. In economics, Knight (1921) distinguished true *uncertainty* (now sometimes called Knightian or radical uncertainty) from quantifiable probabilities or *risks*, and in discussion of the history of probability theory, Hacking (1985) drew a similar distinction between “epistemological” and “aleatory” probability. Here, the most useful way of formulating this distinction is to say that some random variables have distributions that can be well-approximated by quantitative parameters and some do not. Statistical models and tests apply only to random variables that meet or approximate the assumptions for which the procedures are designed.

In archaeobotanical data, the precise numerical counts are often the byproducts of many, complex, confounding factors, and frequently, the interesting samples are the statistical outliers. For instance, Hubbard and Clapham (1992) make the point that depending on the context from which an archaeobotanical sample is recovered, taphonomy may be a more significant determinant of the numbers counted than the original distribution of the material on the site. They point out that the contexts in which (charred) archaeobotanical material is found range from a perfectly preserved, closed storage container in its original anthropogenic context, which they call a “Type A” context, through mixed or potentially contaminated but identifiable contexts (“Type B”), to material obtained by flotation of sediment from an indeterminate context (“Type C”). They then argue that quantification of Type C samples, which are the most frequently encountered, is largely pointless because the complex processes that produce such samples can never be sufficiently disentangled that statements can be made about agricultural or dietary practices.

As with all areas influenced by New Archaeology, this debate over the value of quantification may continue. In practice, however, most contemporary archaeobotanists do attempt to quantify their samples. If archaeobotanical methods are ultimately a tool to understand past human societies better, then it seems reasonable to see taphonomy and site formation processes as producing large and potentially unquantifiable errors in the seed counts of flotation samples from unidentified contexts.

Other sources of numerical errors in archaeobotanical data are obvious and easily corrected. For instance, the number of embryo ends (a way of quantifying the minimum number of individual seeds present in a sample) are not directly comparable with counts of every identifiable element. Other sources of error may be ultimately unquantifiable, like the subjectivity of identification and the possibility of contamination or mislabelling of samples. It is often possible with care to reduce or eliminate some of these potential sources of error, but it is never possible to eliminate, or even identify, all possible errors. Thus, we are led to the conclusion that the numbers in a matrix of seed counts are semiquantitative—usually more like the “minutes’ five” than the “fingers’ five.”

Moreover, the anthropological or historical significance of equivalent numbers of seeds can be radically different. A single well preserved seed of *Zea mays* (maize) from a sealed context well dated to the third century AD of the Near East would threaten to revise our entire historical framework of world civilization—

in fact, such a report would be so radical that it would almost certainly be treated as a mistake or a hoax. A single *Hordeum vulgare* (barley) caryopsis from the same context might not even be interesting enough to be worth recording.

Significance, pattern, and meaning

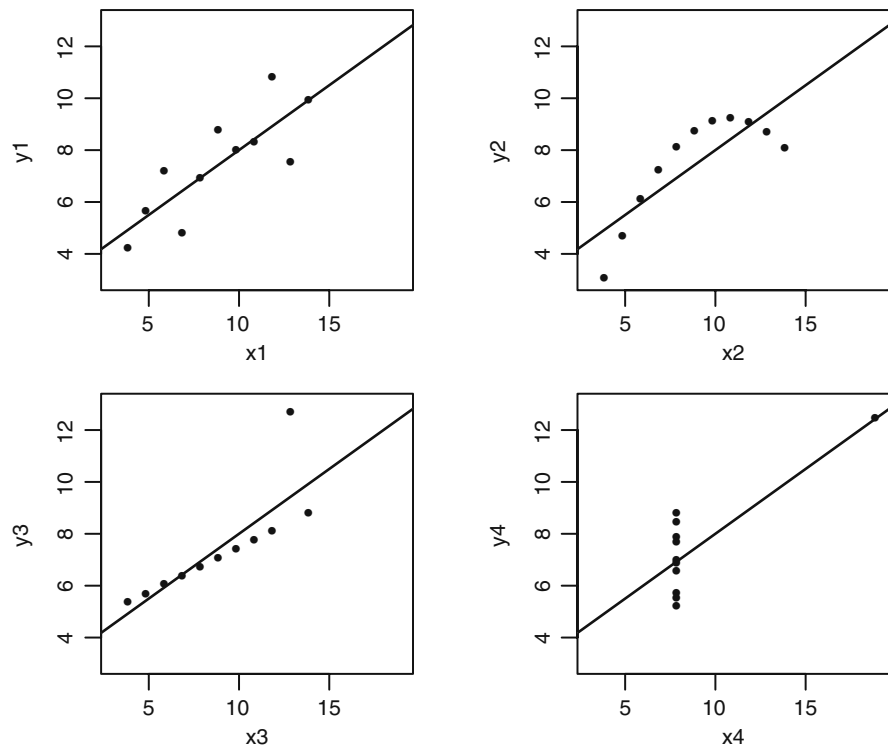
Appropriate statistical tools are available for dealing with some types of semiquantitative data. For instance, nonparametric statistics allow some analyses of quantities whose exact sizes are not known, so long as they can be ranked or ordered from small to large. These are useful, but there remains a second fundamental limitation on statistical analysis of archaeobotanical data: identifying a pattern or statistically significant correlation is not sufficient to show a scientifically interesting effect. This point is essentially the same as is made by the cliché that correlation is not causation, but in practice, there are many patterns that are not correlations.

A simple univariate example of this is the well-known demonstration by Anscombe (1973) that a number of different distributions of points in two-dimensional space can produce the same regression line (see Fig. 1). No one who looks at the plots is fooled, but without a graphical examination, even a statistical procedure as well understood as univariate, least-squares regression can hide important differences between data sets. Multivariate data can prevent the sort of simple visualization that identifies the differences between Anscombe’s data sets. In the case of archaeobotanical data, which is usually highly multidimensional, it is easy to be led astray.

The tools that constitute traditional multivariate statistics are designed for dealing with data in three or more dimensions. Most of the archaeobotanical applications of multivariate statistics have employed what is called Q-mode analysis, which is the practice of classifying or grouping objects based on a number of variables (Legendre and Legendre 1998). With archaeobotanical data, the taxa counted in each sample are generally treated as variables and the samples are treated as objects; in other words, the intent is to classify samples.

Q-mode analysis can be separated into two general sets of tools, both of which have been used by archaeobotanists: clustering algorithms and eigenvalue methods (which including factorial analysis, principle components, canonical correspondence analysis, and multidimensional scaling). The general intent behind both types of Q-mode analysis is to classify objects based on a large number of measured variables. Most

Fig. 1 Figure plotted from data in Anscombe (1973). Four least-squared regressions are shown, which all produce the same regression line: $y = 3 + 0.5x$. Though this line is a mathematically accurate statistical model in all cases, it fails to capture essential aspects of the data in all cases except in the first ($x1$ vs $y1$)



	x1	x2	x3	x4	y1	y2	y3	y4		x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58	7	6	6	6	8	7.24	6.13	6.08	5.25
2	8	8	8	8	6.95	8.14	6.77	5.76	8	4	4	4	19	4.26	3.10	5.39	12.50
3	13	13	13	8	7.58	8.74	12.74	7.71	9	12	12	12	8	10.84	9.13	8.15	5.56
4	9	9	9	8	8.81	8.77	7.11	8.84	10	7	7	7	8	4.82	7.26	6.42	7.91
5	11	11	11	8	8.33	9.26	7.81	8.47	11	5	5	5	8	5.68	4.74	5.73	6.89
6	14	14	14	8	9.96	8.10	8.84	7.04									

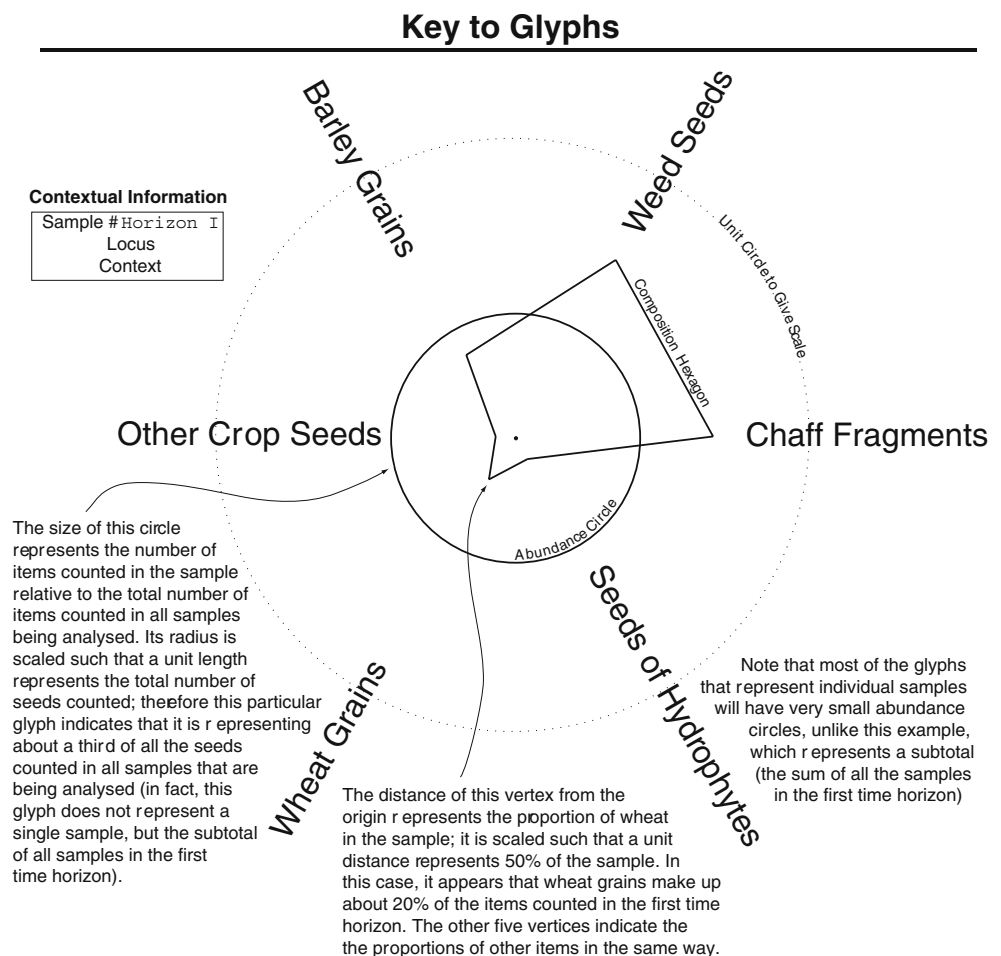
clustering algorithms treat all measured variables as equally important, while eigenvalue methods focus on identifying a few axes along which the objects are spread out (thereby reducing the dimensionality of the data) and classifying objects by reference to these axes.

The major advantage that these forms of mathematical classification have over classifications made by eye is usually referred to as “objectivity,” by which it is meant that they are presumed to reflect “natural” or “inherent” relationships among the objects being classified. In linguistics and anthropology, there has been debate over the role of the classifier (Pike 1967). Happily, for our purposes, there is no difference between “finding natural groups in data” and “classifying data well.” In both cases, the object is to obtain classes or groups that reflect important or interesting aspects of the data and are stable when applied to more data of the same type, under examination by different people at different times and in different places. Therefore, in order to side-step the epistemological issue, we can describe the advantage that mathematical forms of classification

enjoy as “bias of a type radically different from most human biases.”

Such forms of a classification are useful because they provide foils to the biases of trained archaeobotanists, but they are not necessarily less biased than classification by eyeball. They merely substitute arbitrary mathematical biases for rational human ones. To use a simplistic analogy: imagine a room full of people with a number of characteristics that can be measured, like age, height, weight, and hair color, and a number of things one might like to know like life expectancy or cultural affiliation. Actuarial analysis is not needed to realize that age will reveal a lot and hair color virtually nothing about life expectancy, while the significance of the variables is reversed if one is interested in cultural affiliation. Which of these two things one is interested in is a question that no statistical method can answer. By and large, statisticians accept that “There is no obligation to slavishly [sic] accept a numerical solution...every research scientist knows the value of a courteous and able assistant who will make constructive and unbiased suggestions, and this

Fig. 2 Key to star plot; encoding was done with a macro (“M-file”) in Matlab 5.3. Text in the figure explains the details of coding and scaling, but it is only necessary to remember that the size of the circle indicates the abundance of a sample and the hexagon gives its composition

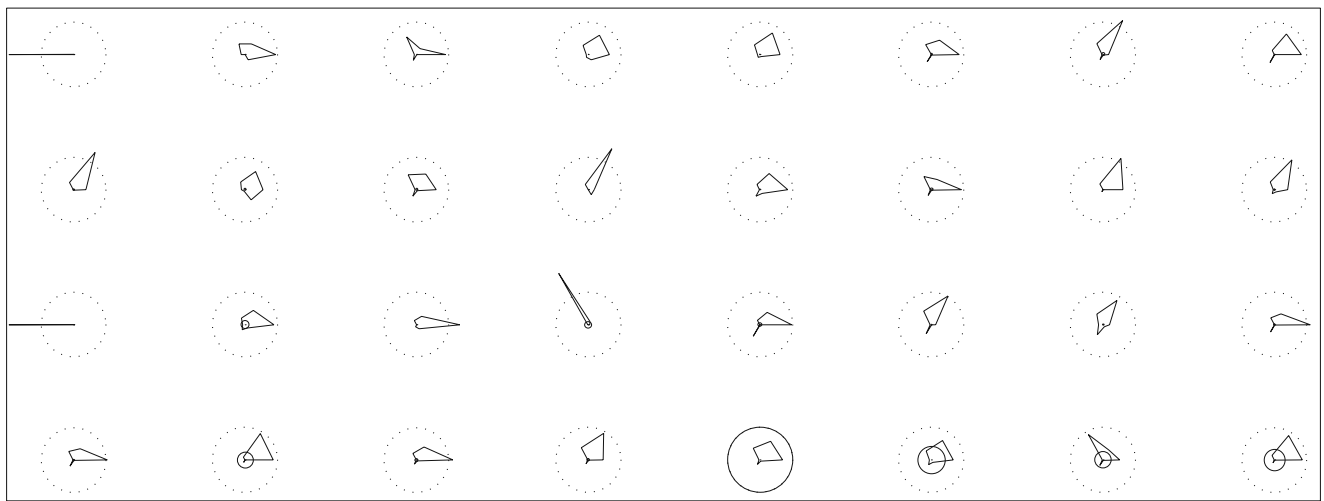


function a computer equipped with a suitable battery of clustering programs is able to fulfil. And like a good assistant, it knows its place: it issues neither judgements nor commands.” (Williams 1971, p. 324). However, the translation between disciplines is confusing. The archaeological consumer needs to distinguish between a classification that is objective merely by virtue of being arbitrary and a classification that explains interesting variation in the archaeobotanical record. As a recent review of multivariate statistics in ecology and systematics puts it: “Much of the misuse of statistical tools is attributed to miscommunication between statisticians and biologists... Statistical usage of terms like ‘effect’ or ‘explanatory variable’ is not meant to imply causation... The objective of the present review is to help the researcher navigate between the Scylla of oversimplification... and the Charybdis of assuming that patterns in data necessarily reflect factors in nature, that they have a common cause, or, worse, that statistical methods alone have sorted out multiple causes.” (James and McCulloch 1990, p. 131f).

Hastorf (1993) has already made the point that “Analysis must transform raw counts quantitatively or qualitatively so that they are interpretable... Because of the biases inherent in the data, most paleoethnobotanists choose to use some form of relative presentation, maintaining an internal comparison, to control for preservation differences and other post-deposition effects... I have found that no one technique gives a complete picture of what the data have to offer archaeologists.”

Exploratory data analysis

As pointed out by Tukey (1977) (among others), the object of exploratory data analysis is not to simplify or summarize data but to represent it in ways that allow us to appreciate it better. This goal is much older than traditional multivariate statistics or its critiques. Graphical display of quantitative information is actually the primitive form of data analysis, and multivariate statistics is



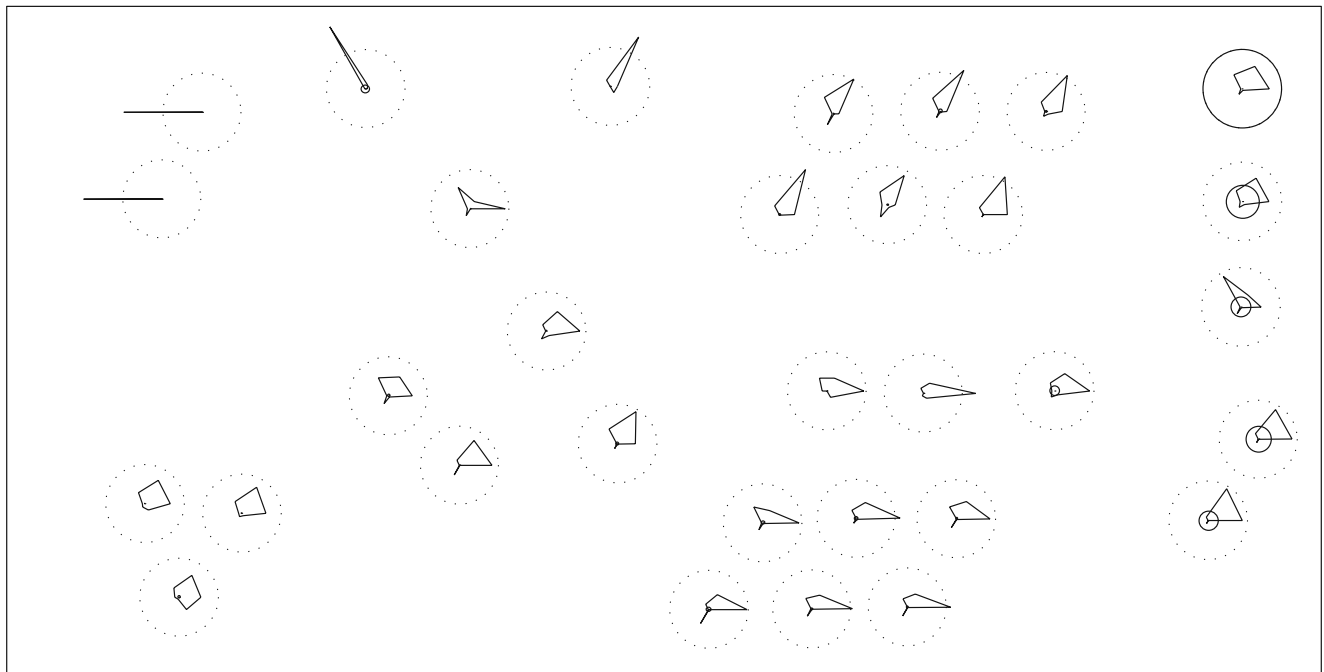
Glyphs arranged in arbitrary order

Fig. 3 Star plots with all contextual information removed, arranged in an arbitrary order

a comparative innovation, so “traditional multivariate statistics” is a misnomer. Tufte, who coined the phrase “visual display of quantitative information” (Tufte 1983) provides a brief discussion of the earlier history of the practice beginning with William Playfair in the eighteenth century.

Data collection and analysis, however, were totally changed by the introduction of the electronic computer

about 40 years ago. The idea of analyzing data by graphical representation was effectively reintroduced and stated in modern statistical terms by Chernoff (1973). Essentially, he argued that it is easier for us to appreciate complex patterns in familiar objects (like faces) than in pages of numbers. Therefore, in order to enhance the power of human pattern recognition, he coded data onto small face-like icons or glyphs.



Glyphs grouped by eye

Fig. 4 Star plots with all contextual information removed, clustered by eye

An elaboration of this strategy was discussed in the *Journal of the American Statistical Association* (see Kleiner and Hartigan 1981: article, comments, and reply), and several examples of its application are given in Wang (1978). The logic behind this approach is that traditional multivariate tools focus on reducing the dimensionality or complexity of data, summarizing, simplifying, or eliminating all but the axes of maximal variation.

Frequently, archaeobotanists already understand the general characteristics of their data and are interested in addressing complicated multifaceted relationships or in finding patterns that were not anticipated. Summarizing or reducing dimensionality can obscure complexity, which is why exploratory data analysis relies heavily on visual representation (Tukey 1977). Note that the term “exploratory data analysis” is most frequently understood to include hierarchical cluster analysis, principle components analysis, and other techniques of data summary that do not involve the testing of explicit hypotheses. All of these techniques have graphical aspects and can provide effective methods of plotting data. I use the term “graphical analysis” to refer to visual representations of data that do not rely on mathematical summary or simplification.

To illustrate how this strategy can be applied to archaeobotanical data, I will use a preliminary matrix of seed counts from the site of Tell Brak in northeastern Syria (Green 1999). (Please note that the seed counts shown in Table 1 in Appendix A and the stratigraphic relationships in Fig. 5 are preliminary and should not be relied upon for archaeological interpretations except in general terms. The archaeological and archaeobotanical conclusions drawn from this material will be presented elsewhere (Charles et al. 2009), as this paper is exclusively about the methods of data analysis.)

Figure 2 shows a star plot (also known as radar plots or rose diagrams) representing an archaeobotanical sample. In this example, six categories of archaeobotanical material are assigned to axes around a center. The number of specimens in each category is plotted along the axes and their apices joined to form an irregular polygon; in this case, a hexagon whose axes correspond to the numbers of wheat seeds, nongrain crop seeds, barley grains, weed seeds, chaff fragments, and seeds of hydrophytic plants. This is labeled the “composition hexagon.” An accompanying circle (the “abundance circle”) gives the total number of specimens in all categories in the sample and another dotted circle gives the total number of specimens counted in all samples for scale (“unit circle”). This is a very general method for exploring complex multivariate data, which, so far as I know, has only been employed once before (Hastorf

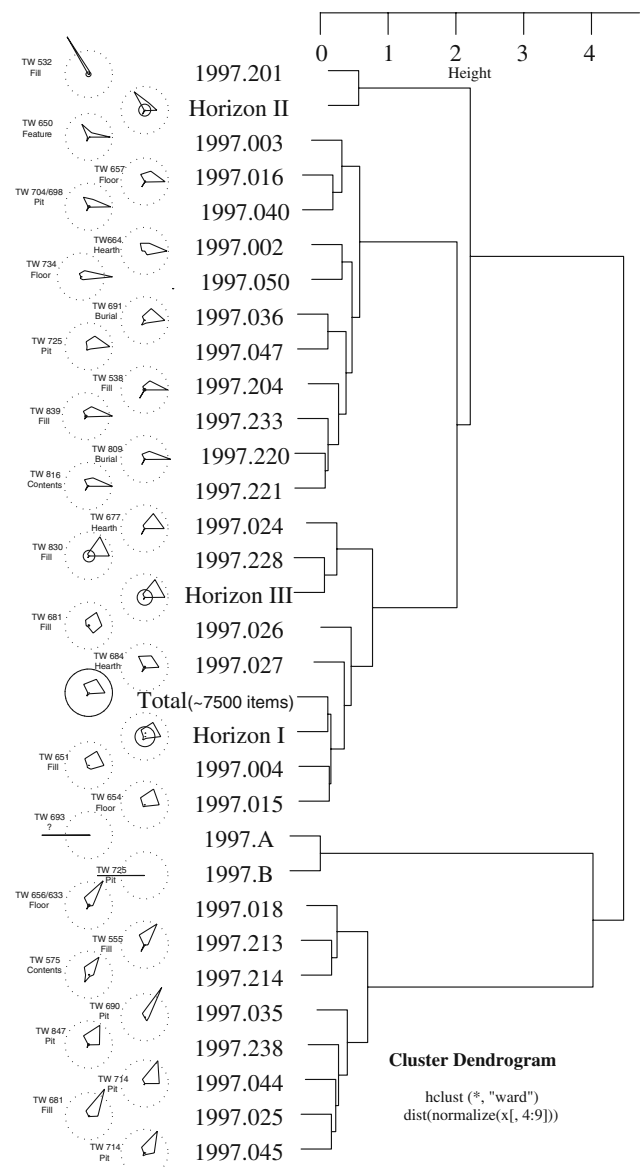


Fig. 5 A hierarchical cluster map (dendrogram) showing the relationships between samples as produced by an algorithmic procedure. The procedure used is one of many possible options using Ward’s method for clustering and a euclidean distance metric. Note that it basically replicates the groups apparent in Fig. 4

1993) for archaeobotanical samples. This particular version was produced by a macro in the program Matlab (Appendix B), though a number of other software packages [in particular, the R statistics and graphics language, R Development Core Team (2004), which, unlike Matlab, is free and open-source] also provide the ability to produce arbitrary graphical representations of numbers.

In Fig. 3, the counts in Appendix A are represented as a series of these star plots, one for each sample.

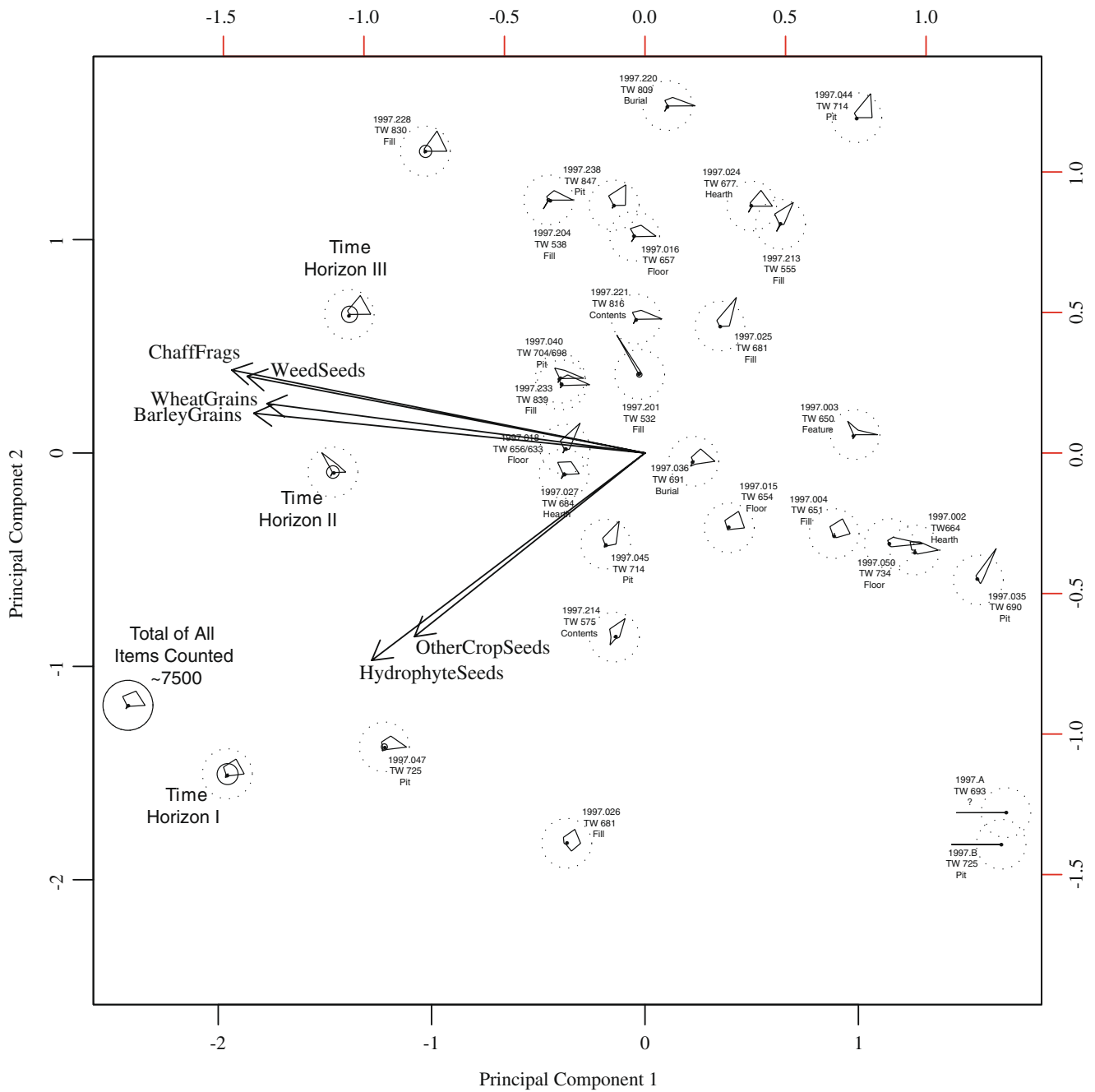


Fig. 6 Here, the star plots are themselves plotted in the bivariate space defined by the first and second principle components. Geometrically, this is the projection of data points in 6-dimensional

space onto the plane whose axes are defined by the linear combination of the six variables that maximize the spread of the scatter of points

The size of the circle represents the abundance of the sample and the vertices of the hexagon give its composition, as described above. In Fig. 3, the glyphs are printed in an arbitrary order, while in Fig. 4, they have been clustered by eye, without reference to contextual data.

The groups displayed in Fig. 4 are based only on pattern recognition and, thus, should not be biased by archaeological preconceptions. If additional objectivity is needed, resampling values analogous to bootstrap or jackknife numbers could be obtained for given groups by showing the top box in Fig. 4 to a large number of

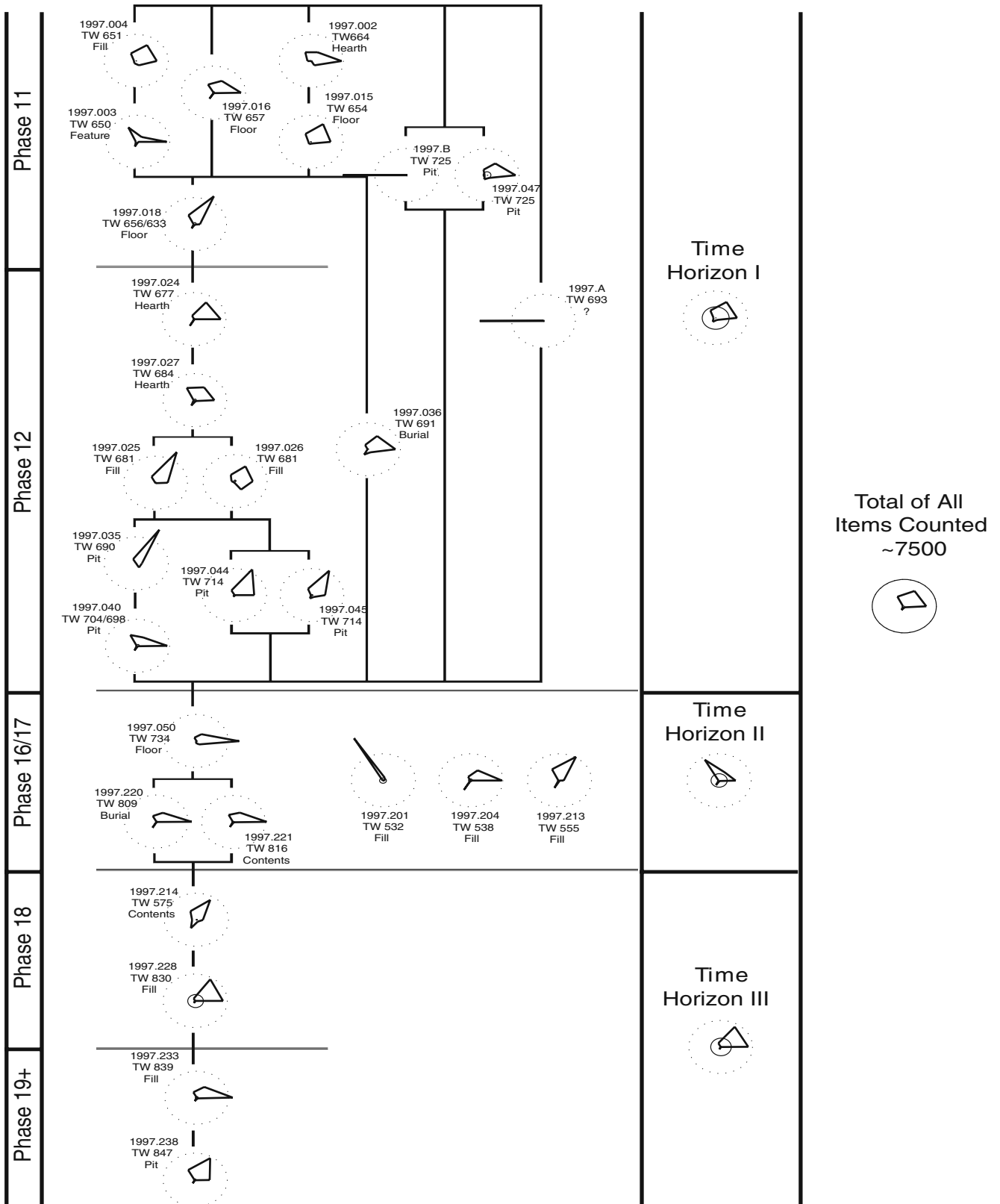


Fig. 7 Harris matrix showing stratigraphic relationships between samples, as well as sample number, name of the stratigraphic unit in which the sample was found, and an abbreviated description of the archaeological context

people and respecting only the groups that are found reliably. The chances are, however, that archaeologically interesting groups will be obvious and revealed by any reasonable method of data presentation. If clustering is so weak as to disappear or change radically when a different method of analysis is applied, it probably contains little of archaeological interest.

In order to compare graphical analysis with standard multivariate tools, Fig. 5 combines a principle components analysis with graphical analysis by plotting the stars from Fig. 3 in the locations where they fall in a hierarchical cluster dendrogram. As can be seen, the groups revealed are very similar to those produced by eye in Fig. 4.

In Fig. 6, the same star plots are ordinated in a bivariate plot of the first two principal components. Again, the groupings that are produced are roughly similar. An additional advantage of this presentation is that the covariation of weed seeds grain and chaff is shown by the cluster of four arrows pointing to the left in a similar direction, which suggests that the first principal component is related to crop processing.

Finally, in Fig. 7, the glyphs representing each sample are inserted into a Harris matrix to show how the flexibility of graphical presentation allows specialized data like stratigraphic relationships and a certain amount of contextual information to be incorporated.

Discussion

Graphical analysis has some limits not shared by some other multivariate tools: on the 2-dimensional page, it is sometimes possible to fit a dozen or more variables (for instance, Fig. 7 gives six quantitative, two categorical, and one relational variable for 31 objects). When there are more than 50 or 100 variables of interest, however,

some form of summary or simplification is needed. Original graphical displays tailored to a particular purpose frequently take more effort to produce and may require either artistic skill or facility with computer programming.

These defects, however, are offset by the advantages visual tools have over traditional multivariate statistics: they are better at coping with the semiquantitative nature of data because we naturally view shapes and sizes by reference to other shapes and sizes (Cleveland 1985), whereas numbers can imply a precision that the data do not support. Graphical displays are also more easily interpreted by readers without reference to the statistical literature.

The best strategy usually consists of the application of several different tools to the same data (Hastorf 1993), but almost all anthropological or historical points can be made by reference to a graphical display instead of a statistical summary or the numerical output of a mathematical manipulation. This is not to imply that graphs should be published alone—it is still important to publish numerical data in as raw a form as practicable. It is also important to recognize that no analytical choices or statistical techniques can compensate for data that are carelessly collected or can add relevance or precision to data that are not pertinent to anthropological or historical questions of interest. Archaeobotanists without a background or interest in statistics or quantitative analysis should be reassured that elaborate statistic manipulations are seldom or never necessary. A picture is worth a thousand words, and often even more numbers.

Acknowledgements Thanks to K. E. Black, J. P. Downs, D. Q. Fuller, D. S. Grewal, W. G. Joyce, K. B. Milek, and two anonymous reviewers for comments on earlier versions of this paper.

Appendix A: Data matrix of seed counts used in analyses**Table 1** Because these are preliminary counts, no details are given of what exactly was counted in each of the six categories (e.g. ‘Chaff Fragments’ is the sum of the counts of *Hordeum* internodes and *Triticum* spikelet forks)

Sample no.	Locus	Phase	Context	Chaff fragments	Weed seeds	Barley grains	Other crop seeds	Wheat grains	Hydro-phyte seeds
1997.A	TW 693	?	?	0	0	0	7	0	0
1997.002	TW 664	11	Hearth	7.5	3	3	1	0	1.5
1997.003	TW 650	11	Feature	13	3	9	1	3	0
1997.004	TW 651	11	Fill	10.5	11	5.2	1	1.7	3
1997.015	TW 654	11	Floor	23.5	29	13	3	4	3
1997.016	TW 657	11	Floor	85.5	50.3	35.2	0	26.3	2
1997.018	TW 656	11	Fill	29.5	211.8	67	6	32.5	4
1997.024	TW 677	12	Hearth	35.5	31.5	6.5	0.5	12.5	0
1997.025	TW 681	12	Fill	34	119.3	22.3	1	1.3	2
1997.026	TW 681	12	Fill	68.5	80.5	34	14	7	46
1997.027	TW 684	12	Hearth	78	71.8	69	6	30.5	5
1997.035	TW 684	12	Hearth	1	8	1	0	0	1
1997.036	TW 691	12	Burial	45	30	9.5	1	12.5	7
1997.040	TW 704	12	Pit	151	58.5	75.5	5	33.5	3
1997.044	TW 714	12	Pit	15.5	27.9	5	0	2	0.1
1997.045	TW 714	12	Pit	41.5	109.5	28	4	15	9
1997.B	TW 725	12	Pit	0	0	0	5	0	0
1997.047	TW 725	12	Pit	418	239	109.5	37.3	86	55
1997.050	TW 734	16/17	Floor	18	4	2	0	1	2
1997.201	TW 532	16	Fill	17.5	42	719	5	4	1
1997.204	TW 538	16/17	Fill	224	100.5	35.3	2	93	2
1997.213	TW 555	16/17	Fill	5	38.2	18	1.3	12	0.3
1997.214	TW 575	17	Vessel contents	16	70.5	31	11	29	6
1997.220	TW 809	17	Burial	108	38	24	0.5	25.5	1
1997.221	TW 816	17	Vessel contents	99.5	38.5	29	2.1	21	2
1997.228	TW 830	18	Fill	788.5	853.8	91	6.1	84	2
1997.233	TW 839	EU	Fill	198.5	81.5	41.5	2.6	27	6
1997.238	TW 847	EU	Fill	62.5	128.5	59	2.1	19.5	1
Horizon I	NA	NA	NA	1,057.5	1,084.2	492.6	92.8	267.8	141.5
Horizon II	NA	NA	NA	488	331.7	858.3	21.8	185.5	14.3
Horizon III	NA	NA	NA	1,049.5	1,063.8	191.5	10.5	130.5	9
Total	NA	NA	NA	2,595	2,479.6	1,542.5	125.2	583.8	164.8

For further details, see Green (1999) and Charles et al. (2009)

Appendix B: Macro used to plot glyphs

```

function englyph (M)
  \%The function englyph takes a matrix with six rows and any number of columns
  \%and plots each column vector in the matrix as a glyph composed of an
  \%irregular hexagon and a circle, both centered at the origin, in which the
  \%six vertices of each hexagon are distant from the origin in proportion to
  \%the six vector entries that the glyph represents (and sum to unity), while
  \%the diameter of the circle is proportional to the sum of the entries of the
  \%vector being represented and scaled so that the vector of all the column
  \%vectors in the matrix whose entries have the greatest sum has unit length
  \%
  \%The input matrix has columns representing archaeobotanical samples and rows
  \%corresponding respectively to counts in each sample of Chaff (Wheat and
  \%Barley), Weed Seeds, Barley Grains, Other Crop Seeds, Wheat Grains, and
  \%Seeds of Hydrophytes.

  totals = sum (M)
  max_sample_abundance = max (totals)
  columns = size (M)
  j= columns (2)

  for i=1:j
    S=M(:,i)
    tot=sum(S) \%calculates total number of items counted in sample
    d=tot/max_sample_abundance \%scales sample vector to unit circle

    figure (1)
    subplot (5,8,i)
    hold on \%allows all values to appear on the same axes

    rectangle ('position', [-1/2,-1/2,1,1], 'curvature', [1,1], 'LineStyle', ':', 'LineWidth', .25)
    \%draws dotted circle with unit diameter around origin

    rectangle ('position', [-d/2,-d/2,d,d], 'curvature', [1,1])
    \%draws circle at origin with diameter d, representing abundance of sample

    [a1,a2] = pol2cart (0, S(1)/tot)
    [b1,b2] = pol2cart (pi/3, S(2)/tot)
    [c1,c2] = pol2cart (2*pi/3, S(3)/tot)
    [d1,d2] = pol2cart (pi, S(4)/tot)
    [e1,e2] = pol2cart (4*pi/3, S(5)/tot)
    [f1,f2] = pol2cart (5*pi/3, S(6)/tot)
    \%calculate cartesian coordinates of glyph vertices

    line ('XData', [a1,b1], 'Ydata', [a2,b2])
    line ('XData', [b1,c1], 'Ydata', [b2,c2])
    line ('XData', [c1,d1], 'Ydata', [c2,d2])
    line ('XData', [d1,e1], 'Ydata', [d2,e2])
    line ('XData', [e1,f1], 'Ydata', [e2,f2])
    line ('XData', [f1,a1], 'Ydata', [f2,a2])
    \%draw glyphs

    line (0,0) \%draws point at origin

    axis ([-1 1 -1 1]) \%sets xmin, xmax, ymin, ymax
    axis square \%equalizes x and y scales
    axis off \%makes the axes invisible
  end

```

References

- Anscombe FJ (1973) Graphs in statistical analysis. *Am Stat* 27(1):17–21
- Charles M et al (2009) Report on the botanical remains. In: Oates D, Oates J, McDonald H (eds) *Excavations at Tell Brak, vol III: the Uruk period*. Cambridge University Press, Cambridge (in preparation)
- Chernoff H (1973) The use of faces to represent points in K-dimensional space graphically. *J Am Stat Assoc* 68(342): 361–368
- Cleveland WS (1985) *The elements of graphing data*. Chapman and Hall, New York
- Green WA (1999) *Agriculture and colonialism: tell brak in the uruk period*. M Phil dissertation, University of Cambridge
- Hacking I (1985) *The emergence of probability*. Cambridge University Press, Cambridge
- Hastorf CA (1993) *Agriculture and the onset of political inequality before the Inka*. Cambridge University Press, Cambridge
- Hubbard RNLB, Clapham A (1992) Quantifying macroscopic plant remains. *Rev Palaeobot Palynol* 73:117–132
- Huff D (1954) *How to lie with statistics*. Norton, New York
- James FC, McCulloch CE (1990) Multivariate analysis in ecology and systematics. *Ann Rev Ecol Syst* 21:129–166
- Kleiner B, Hartigan JA (1981) Representing points in many dimensions by trees and castles. *J Am Stat Assoc* 76(374): 260–269
- Knight FH (1921) *Risk, uncertainty, and profit*. Houghton Mifflin, Boston
- Lakatos I (1978) *The methodology of scientific research programmes*. Cambridge University Press, Cambridge
- Legendre P, Legendre L (1998) *Numerical ecology*, 2nd ed. Elsevier, Amsterdam
- Pike K (1967) *Language in relation to a unified theory of the structure of human behavior*. Mouton, The Hague
- Quine WVO (1953) *From a logical point of view*. Harvard University Press, Cambridge
- R Development Core Team (2004) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>
- Tufte ER (1983) *The visual display of quantitative information*. Graphics, Cheshire
- Tukey JW (1977) *Exploratory data analysis*. Addison Wesley, Reading
- Wang PCC (ed) (1978) *Graphical representation of multivariate data*. Academic, New York
- Williams WT (1971) Principles of clustering. *Ann Rev Ecol Syst* 2:303–326