10
The incorporation of cluster analysis into multidimensional matrix analysis

John Wilcock
School of Computing, Staffordshire University, Stafford, UK
Email: j.d.wilcock@soc.staffs.ac.uk

10.1 Introduction

A previous paper (Wilcock 1993) showed that seriation could be applied simultaneously to more than two dimensions. This paper extends the idea to incorporate cluster analysis, and proposes a number of new types of diagram for the portrayal of multidimensional data, in particular new forms of dendrogram and skyline plot.

10.2 The portrayal of outputs from cluster analysis

It is beyond the scope of this paper to discuss the methodology of cluster analysis. It is, however, relevant to discuss the types of diagram which have been used for the portrayal of outputs from cluster analysis, as a basis for the additional diagrams proposed in the paper.

The *dendrogram* is a tree of relationships, first given this name by Mayr, Linsley and Usinger (1953). Sneath and Sokal (1973) have shown that the 2-dimensional dendrogram cannot be used successfully to portray multidimensional data, and that early attempts to force multidimensional data into hierarchical relationships by means of dendrograms have not resulted in satisfactory classifications. A natural non-overlapping taxonomic hierarchy can be represented in the form of a dendrogram, providing that the pair-function is monotonically increasing or ultrametric. However, if the relationships are not ultrametric they cannot be mapped into a dendrogram (Hartigan 1967; Jardine 1969; Farris, 1969), and if they are not even metric then serious distortion of the taxonomic relationships and reversals frequently occur when the results are represented in dendrogram form. Essentially, the dendrogram branches do not represent a genealogy, but rather affinities between the clustered units. A continuous hierarchy is created by gradually combining units into larger groups, using an agglomerative strategy. Alternatively a dendrogram can be used to portray the results of dividing the total sample into progressively smaller groups using a divisive strategy. Polythetic groups or classes have members which have a large number of properties in common, but no one property is essential, and these classes are regarded as natural; monothetic classes must possess one or more properties, and these classes are regarded as artificial (Sokal and Sneath 1963).

Hodson, Sneath and Doran (1966) used the dendrogram to portray clusters from a sample of brooches from the Iron Age cemetery at Münsingen-Rain, and they concluded that the average-link clustering method (Sokal and Sneath 1963) had produced groups of archaeological significance, while the single-link method had not (Hodson 1969). It is not proposed to cover the well-known controversy of the 1960s of average-link versus single-link, but we may note that Jardine, Jardine and Sibson (1967) in a series of papers, with answers from Hodson *et al.*, pointed out the important theoretical difficulty of *discontinuity*, where a small change in one of the similarity coefficients could affect the dendrogram adversely, not only in the region of the two objects, but right across the assemblage. A good comparison of typical dendrograms resulting from single-link, total-link and average-link cluster analyses is given by Aldenderfer and Blashfield (1984).

The paper of Jardine *et al.* (1967) marks a turning point away from average-link methods to other methods such as 'k-means' (Hodson, 1970; 1971) which were more acceptable to the mathematicians, and still had usefulness for the archaeologists. The k-means method differs from both the average-link and single-link methods in that it splits up the initial assemblage into a specified number of clusters, k. Objects are allowed to migrate between clusters; this migration is not allowed in average-link, with the disadvantage that an object may become trapped in a cluster which was the best in an early stage of the clustering, but is not the best at the end. The end product of k-means is a series of divisions of the objects into the best two clusters, the best three, etc. up to k clusters. Since these clusterings are independent, there is no reason why a given pair of objects should always belong to the same cluster in the different distributions. Thus a dendrogram is not appropriate for portraying the k-means output; instead a graph is used which shows the 'percentage error of fit', the average squared distance of the objects from the centre of their own clusters expressed as a percentage of their average squared distance from the centre of the whole distribution, for the different number of clusters. The percentage error of fit has a value of 100% for one cluster, and decreases as the number of clusters is increased. If there is an obvious elbow in this curve, it may be an indication of the 'correct' number of clusters.

Another method for representing cluster analysis results is the *skyline plot*. This was developed by several workers, notably Ward (1963) and Wirth, Estabrook and Rogers (1966). However, this type of diagram suffers from the same drawbacks as the dendrogram, in that it can portray only two-dimensional data satisfactorily. A comparison of diagrams for the portrayal of archaeological classifications is given by Wilcock (1975).

10.3 The portrayal of multidimensional data

Dendrograms as used in the literature have been confined to the portrayal of two-dimensional data, and archaeologists have taken the *multidimensional scaling* and *principal components analysis* routes for the discovery of further dimensions.

Multidimensional scaling starts with an n-coordinate positioning of points representing objects in n-dimensional space, and then attempts to reduce the number of dimensions, preserving as far as possible the rank order of dissimilarities between pairs of objects in terms of a rank order of distances. This is not possible as the number of dimensions is reduced, and the awkwardness of the fit is expressed as 'strain'. If the strain becomes too big, the reduction stops, but many archaeological distributions do reduce to three or even two dimensions without too much strain, and if so the distribution may be portrayed as a *scalogram.* Multidimensional scaling was first used in psychology. Examples of its use in archaeology are given by Hodson, Sneath and Doran (1966), Bonsall and Leach (1974) and Shennan and Wilcock (1975).

Principal components analysis (e.g. described in Morrison 1967) derives a set of new orthogonal components, each being a synthetic property based on different loadings of the original properties. The early principal components often account for variances equivalent to many properties-worth of the original properties. The usual output diagram for principal components is a *pc plot*, for example *pc1* versus *pc2* (e.g. as shown for the Münsingen-Rain data by Hodson (1969, 94; 1970, 314)). An extension of this would be to plot three pcs on a 3D surface diagram, which may identify clusters as peaks on the surface.

Cluster boundaries may be added to either scalograms or pc plots, as may minimum spanning trees, and the resulting diagrams are referred to as Wroclaw diagrams. Although multiple dimensions are being portrayed on these diagrams, the dimensions are synthetic, made up from different loadings of the original properties, and the original dimensions may be difficult to visualise.

10.3.1 Multidimensional matrices

The use of multidimensional matrices for seriation has been discussed in an earlier paper (Wilcock 1993). The main points made by this former paper were:

- matrices of three or more dimensions have rarely been studied
- all the algorithms described should be applicable to any number of dimensions greater than or equal to two
- For a 2-dimensional matrix, rows or columns may be treated as sub-matrices
- For a 3-dimensional matrix, the sub-matrices are 2-dimensional planes, and so on
- For D dimensions, each sub-matrix will have $(D - 1)$ dimensions
- Problems with four or more dimensions may still be handled, since the algorithms are completely generic in nature. However, the configuration may then no longer be represented as a geometrical model, since it is in hyperspace.

In fact, problems with more than two dimensions are commonplace in archaeology. A hierarchy of typical dimensions applicable in archaeology is:

1. Time
2. Culture
3. Site
4. Phase
5. Assemblage
6. Artefact
7. Property

i.e. at least seven dimensions could be considered. For example, a problem concerning the distribution of Roman coins in Britain has already been implicitly restricted to the Roman culture, to coin artefacts, and geographically to Britain, but time, site, phase, assemblage, and properties of coins remain as dimensions to be considered, i.e. the data has five inherent dimensions. Again, in the study of Medieval bells, the Medieval culture and bell artefacts are implicit, but the development of bells in time, the places where bells are found, the places of manufacture, the makers, and the properties of bells remain to be studied, again five dimensions. Archaeologists are often not sufficiently plain it stating what assumptions they are making about the data, and what dimensions are being implicitly excluded from a study. Sometimes the body of data is submitted blindly to multidimensional scaling, principal components or factor analysis routines in the hope that the variance will emerge as relevant principal components or factors.

10.4 Dimensional Framework for a Study

The dimension under study requires a framework which typically includes the dimension on each side of it in the hierarchy, e.g. a culture is meaningless without being defined in terms of some time period and some geographical location (sites or areas).

As an illustration, the Münsingen-Rain data (Hodson 1968) have been employed in this study described in this paper. This data set has three dimensions (although previously published analyses of the data have confined themselves to two dimensions only). The three dimensions are Hodson's 'horizons' (phases), the graves and the artefacts. The reprocessing of this data in a three-dimensional matrix, with a horizon in each plane, gave results which were found to be comparable with Hodson's original sequence, but which introduce a better conception of the true dimensionality of the data.

1	1 111	2 122	2 113	4 125	5 226	
1 111	2	2 122	4 233	4 135	5 226	
2 122	2 122	4 233	4 233	4 144	6 145	
2 131	4 233	4 233	4 233	6 256	6 256	
4 152	4 253	4 253	4 253	6 256	6 256	Plane 1

1 111	3 222	3 222	3 313	5 316	5 316	
3 222	3	3 222	4 233	5 334	5 316	
3 222	3 222	4	4 233	5 334	6 345	
3 331	4 233	4 233	4 233	6 345	6 345	
4 342	4 342	4 342	6 345	6 345	6 345	Plane 2

3 222	3 222	3 222	5 425	5 425	5 425	
3 222	3 222	3 222	5 334	5 334	5 425	
3 222	3 222	4 233	5	5 334	6 345	
3 331	4 233	4 233	5 334	6	6 345	
4 342	4 342	4 342	6 345	6 345	8 456	Plane 3

3 422	3 422	3 422	5 425	5 425	5 425	
3 331	3 331	5 334	5 334	5 334	7 436	
3 331	3 331	5 334	5 334	7 445	7 445	
3 331	4 342	5 334	7 445	7	7 445	
4 342	4 342	7 454	7 445	7 445	8	Plane 4

Figure 10.1: A typical 3-dimensional matrix. The central diagonal is outlined heavily in black. The remaining cells have two numbers, the first giving the central diagonal cell to which that cell is nearest, and the second giving the coordinates of the cell to which the cell is connected, in the tree structure described in Wilcock (1993).

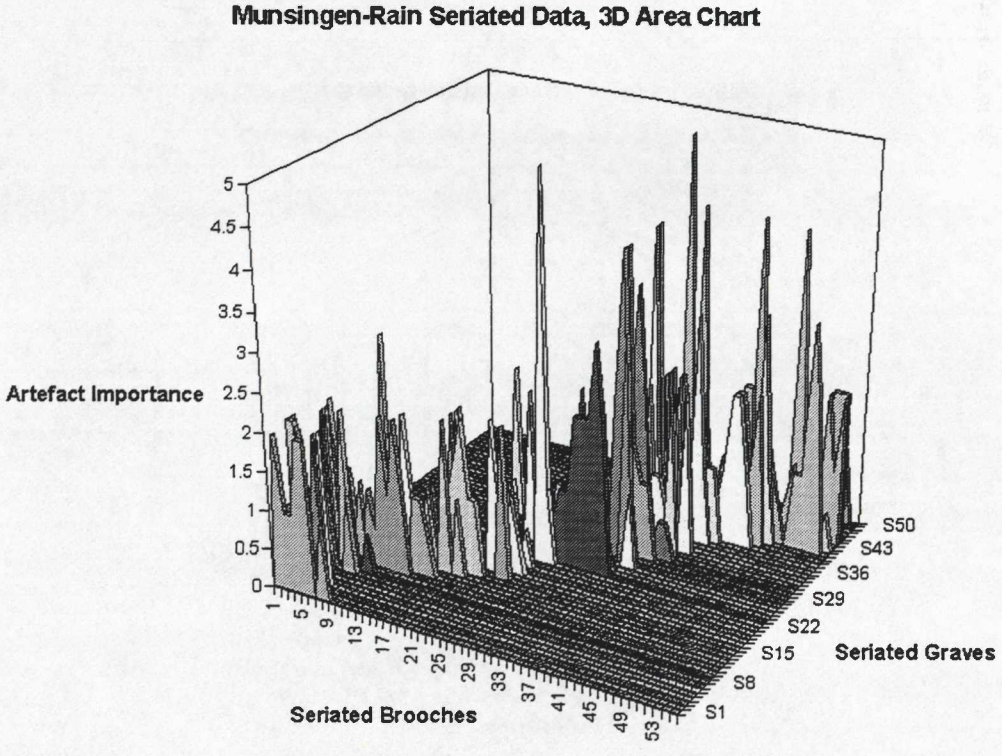

Figure 10.2: 3D Area Chart, Graves versus Brooches, for the Münsingen-Rain seriated data

10.5 A typical 3-dimensional matrix

Figure 10.1 shows a typical 3-dimensional matrix with 4 planes, 5 rows and 6 columns. The four separate planes are shown, with the central diagonal cells outlined heavily in black:

10.6 The graphical portrayal of multidimensional data

10.6.1 3D Area Chart, Graves versus Brooches

It is appropriate to exploit the potential of charting packages to display multidimensional data. Figure 10.2 shows a 3D area chart showing the seriation of graves versus brooches for the Münsingen-Rain data. An area chart shows how values change in proportion to the total (of values for each property) over a period of time. It is similar to a series of line graphs, but emphasises the magnitude of values with respect to time, rather than the flow of time and rate of change of properties.

10.6.2 3D Surface Chart, Graves versus Brooches

The same seriated data may be shown as a 3D Surface Chart with contours for artefact importance (see Figure 10.3).

10.6.3 Stacked Bar Chart, Seriated Brooches

Figure 10.4 uses the *Stacked Bar Chart* form of diagram to show seriated brooch types on the horizontal axis, the height of a bar gives the importance of a brooch type, and the different shadings in a bar give the corresponding graves and the relative importance of those graves for a brooch type.

10.6.4 Stacked Bar Chart, Seriated Graves

Conversely, the Stacked Bar Chart in Figure 10.5 shows seriated graves on the horizontal axis, the overall height of a bar gives the importance of a grave, and the different shadings in a bar give the corresponding types of brooches and the relative importance of those types in a grave.

10.6.5 Stacked Bar Chart, Clustered Graves

Figure 10.6 is similar to Figure 10.5, except that the graves have now been clustered. As with the seriation, the overall height of a bar gives the importance of a grave, and the different shadings in a bar give the corresponding types of brooches and the relative importance of those types in a grave.

10.6.6 Skyline Plot, Clustered Graves

The skyline plot form of display has been mentioned above. In Figure 10.7 the skyline plot for the clustered graves is shown as a 3D Surface Chart. The clusters can be seen, separated by troughs (compare this diagram with Figures 10.8 and 10.9).

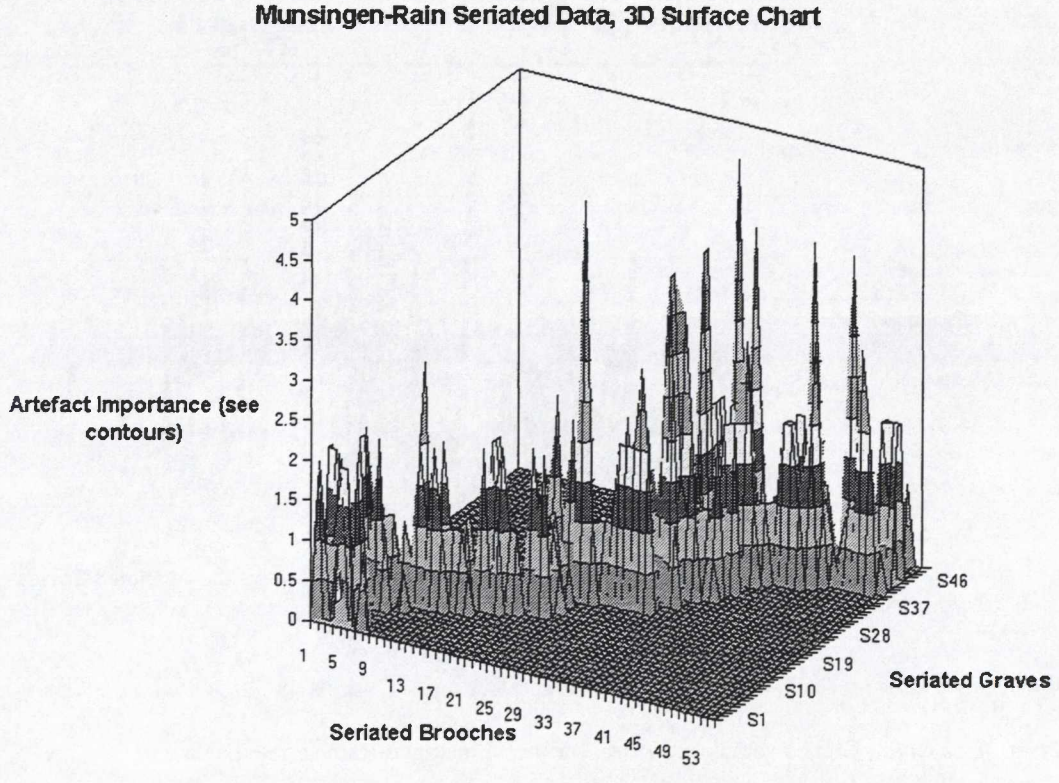

Figure 10.3: 3D Surface Chart, Graves versus Brooches, for the Münsingen-Rain seriated data

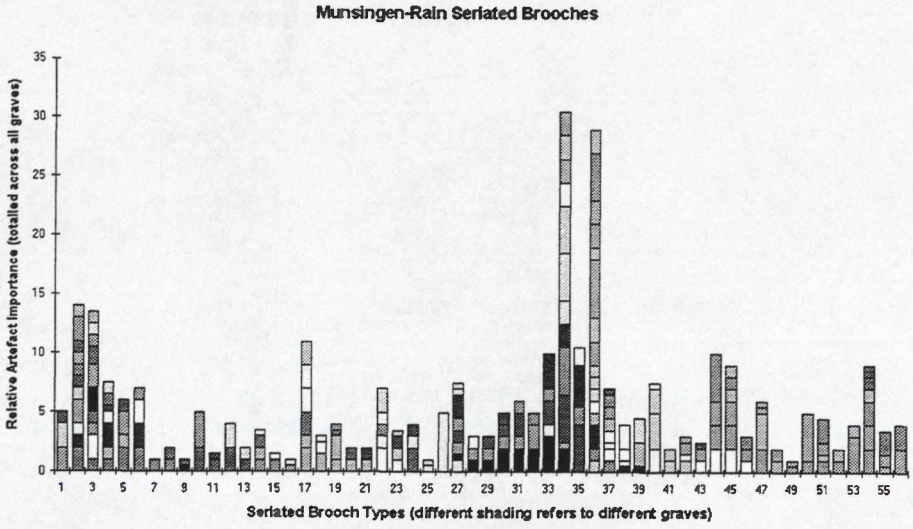

Figure 10.4: Stacked Bar Chart for seriated brooches, Münsingen-Rain data

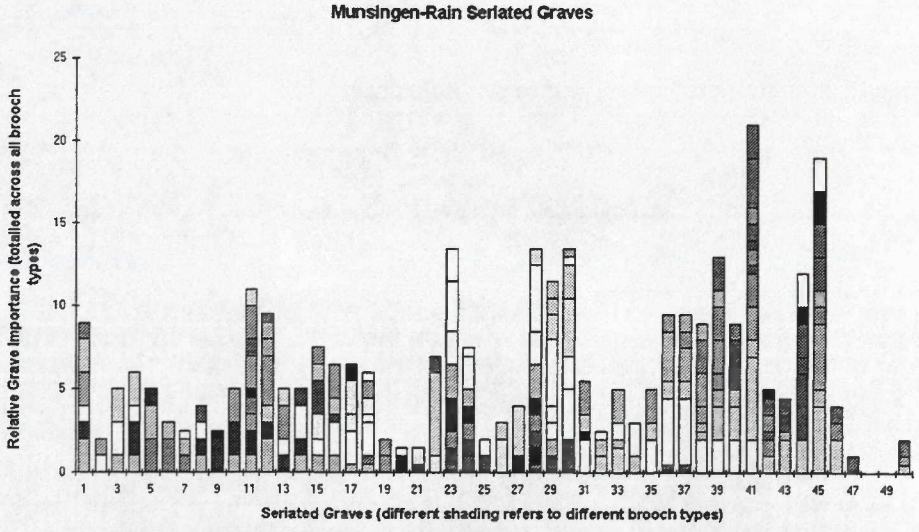

Figure 10.5: Stacked Bar Chart for seriated graves, Münsingen-Rain data

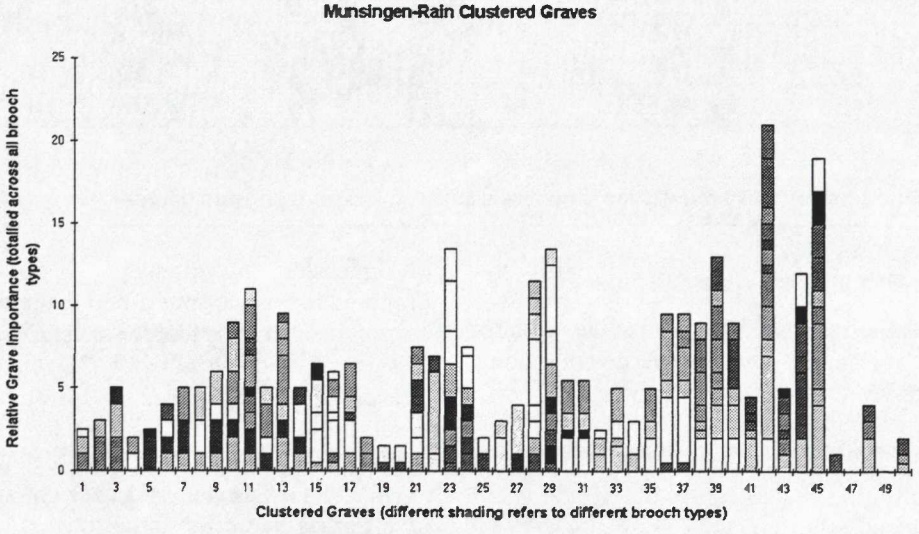

Figure 10.6: Stacked Bar Chart for clustered graves, Münsingen-Rain data

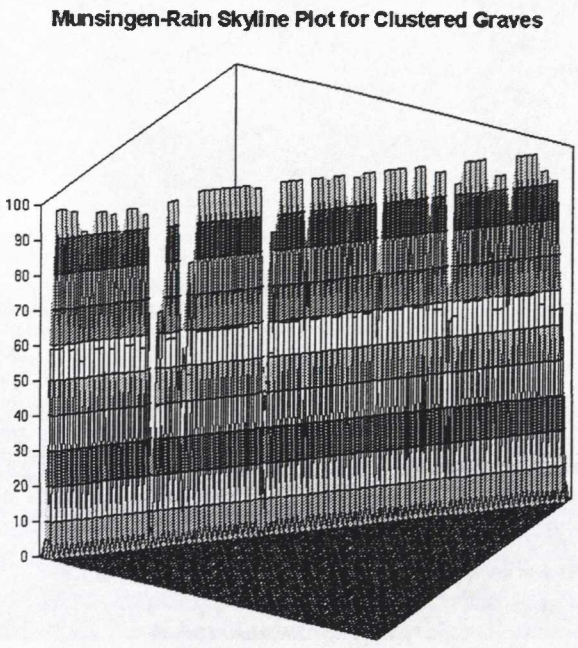

Figure 10.7: Skyline Plot for clustered graves, Münsingen-Rain data

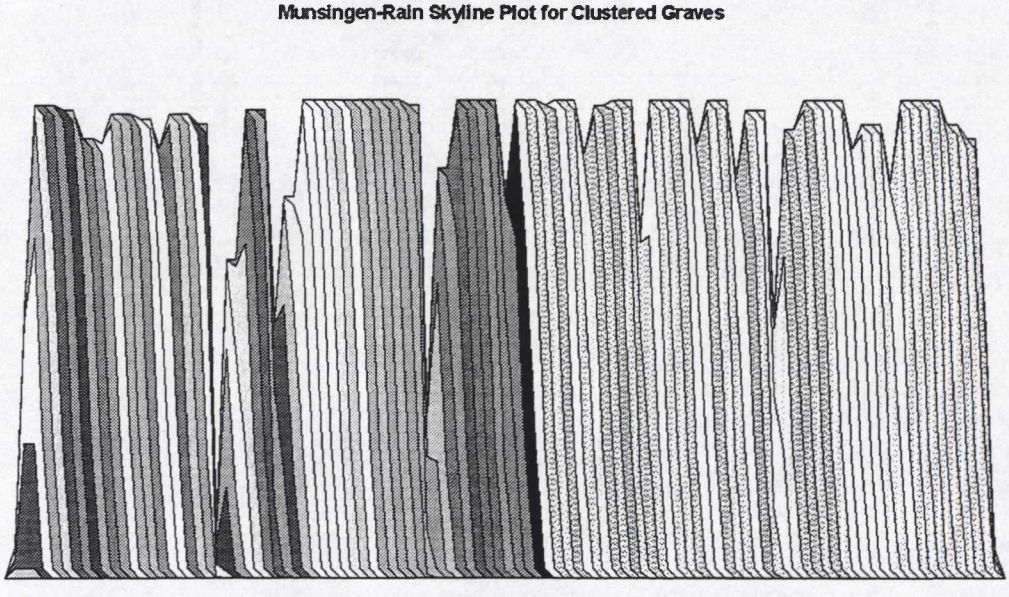

Figure 10.8: Skyline Plot for clustered graves shown as a surface, Münsingen-Rain data

10.6.7 Skyline Plot shown as a Surface

Figure 10.8 shows the same skyline plot for the clustered graves as a more 'sculptured' type of surface with shading for the different graves.

10.6.8 3D Skyline Plot showing Phases related to Graves

However, none of the above diagrams have really shown 3 or more dimensions in a satisfactory manner — the representation has chiefly been the display of a selected pair of dimensions, albeit selected from the larger number of dimensions under study. Figure 10.9 is, however, proposed in this paper as a step in the right direction. The conventional skyline plot for the clustered graves has at its head ONE 'side view' of a 3D matrix (Phases versus Graves versus Brooches), in this case showing Phases vertically, and Graves horizontally. The vertical lines relate each grave to its corresponding phase. Another skyline plot exists at the side, at right-angles, which would give Phases vertically and Brooches horizontally. Thus the third dimension has been introduced into the essentially two-dimensional skyline plot.

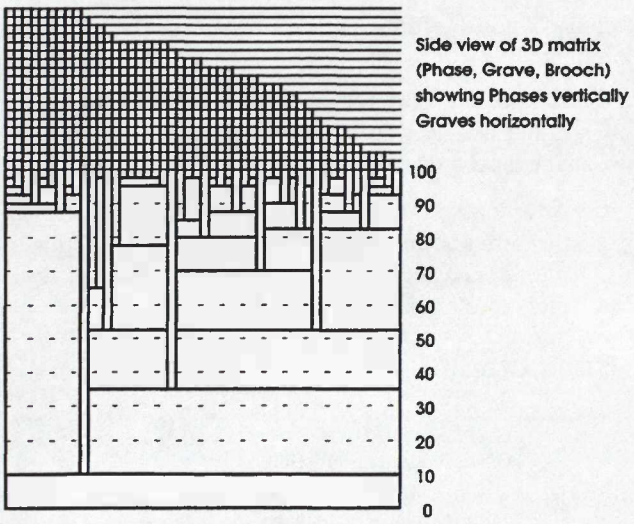

Figure 10.9: 3D Skyline Plot showing Phases and Graves. The block at the head of the conventional Skyline Plot is the side view of a 3D matrix Phases versus Graves versus Brooches.

10.6.9 3D Dendrogram showing Phases related to Graves

Figure 10.10 repeats the procedure, this time for the Dendrogram. The conventional dendrogram for the clustered graves has at its head ONE 'side view' of a 3D matrix (Phases versus Graves versus Brooches), in this case showing Phases vertically, and Graves horizontally. Another dendrogram exists at the side, at right-angles, which would give Phases vertically and Brooches horizontally. Thus the third dimension has been introduced into the essentially two-dimensional dendrogram.

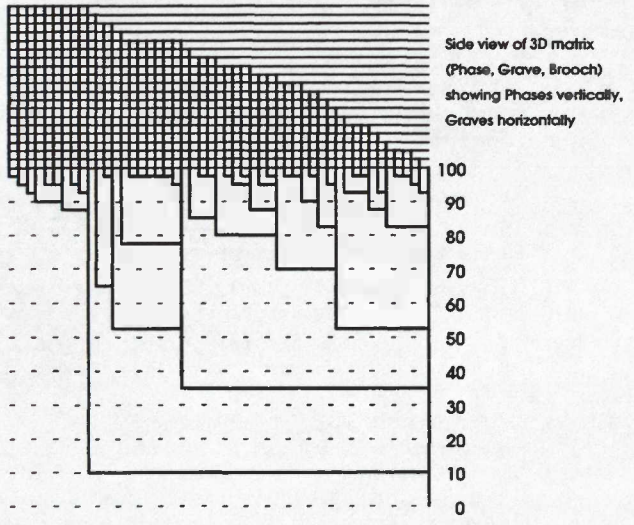

Figure 10.10: 3D Dendrogram showing Phases and Graves. The block at the head of the conventional Dendrogram is the side view of a 3D matrix Phases versus Graves versus Brooches.

10.7 The Triad (Three Dimensions)

It is appropriate to discuss the number of dimensions in a theoretical manner. Figure 10.11 shows a *triad* of three dimensions. The 3 dimensions A, B and C may be combined in three different *pairs*, *[AB]*, *[AC]* and *[BC]*. In our case we may allocate Phases to A, Graves to B and Brooches to C.

Thus we could have a dendrogram for Phases v Graves, another for Phases v Brooches, and a third for Graves v Brooches in our application. The super-dimension appears at the head in each case, and the sub-dimension forms the dendrogram or skyline plot. The diagram above shows a unique allocation of a sub-entity to a super-entity, but in principle sub-entities could occur across several super-entities, and be indicated by some form of bar (as in a *battleship plot*). A bar could also itself have more than one dimension, the 'battleship' sections then becoming more like globular Christmas-tree ornaments in shape. Such a system could show parallel time-lines for different cultures, related to the areas of the world inhabited/controlled by the respective cultures.

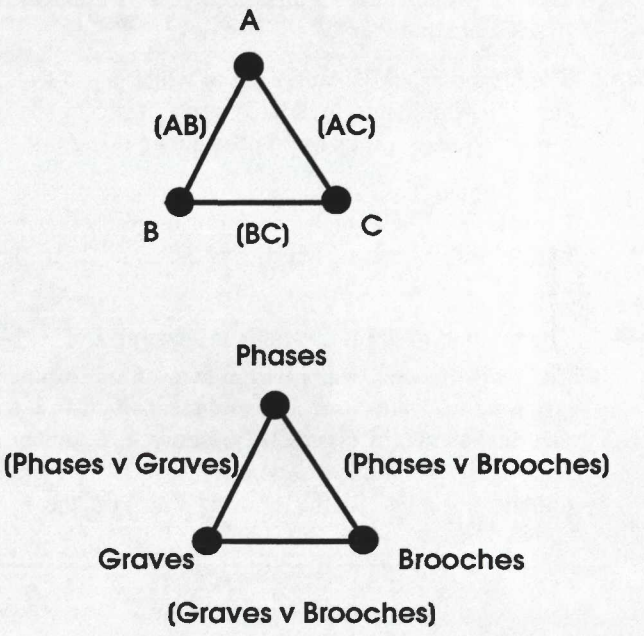

Figure 10.11: The Triad

10.8 The Quad (4 Dimensions)

Figure 10.12 extends this procedure to 4 dimensions, and shows the *quad*.

- Thus 1 Quad has 4 Triads, each of which has 3 Pairs.
- In general for D dimensions, there will be D sub-matrices of $(D-1)$ dimensions, each of which has $(D-1)$ sub-sub-matrices of $(D-2)$ dimensions, and so on.
- For 1 Quad there are: 4 unique Triads, each of which has 3 Pairs, and there are 6 unique Pairs

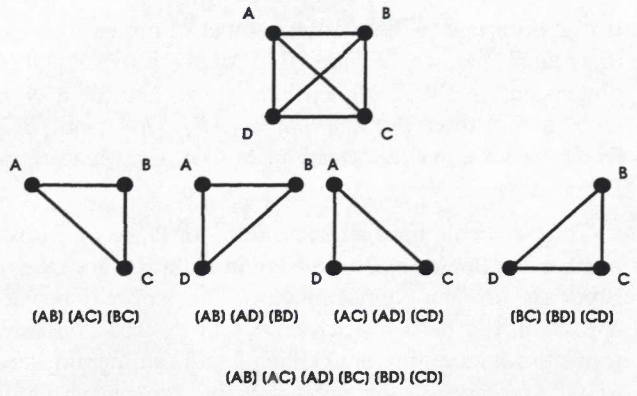

Figure 10.12: The Quad

10.9 Larger numbers of dimensions

The procedure may be extended to ever larger number of dimensions. Figure 10.13 shows the *pentad.*

- The Pentad has 5 Quads, each of which has 4 Triads, each of which has 3 Pairs. There is a total of 10 possible unique Pairs.
- The Hexad has 6 Pentads, each of which has 5 Quads, each of which has 4 Triads, each of which has 3 Pairs. There is a total of 15 possible unique Pairs.
- The Heptad has 7 Hexads, each of which has 6 Pentads, each of which has 5 Quads, each of which has 4 Triads, each of which has 3 Pairs. There is a total of 21 possible unique Pairs.
- The number of unique Pairs is in general $D(D-1)/2$ for D dimensions, which is the sum of an arithmetic progression. This is shown in the half-matrix less diagonal shown in Figure 10.13 above, the number of Pairs being the sum of the newly introduced Pairs of dimensions, plus all the existing Pairs on the rows above.

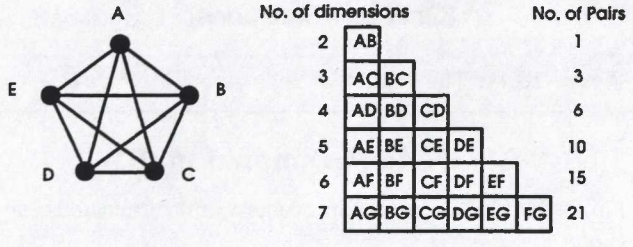

Figure 10.13: The Pentad, and the half-matrix less diagonal which shows all Pairs for dimensions between 2 and 7

10.10 Conclusion

It has been proposed that archaeologists should consider the true dimensionality of their data, and explore ways of portraying results in more than two dimensions. It is apparent that as many as seven dimensions are possible in archaeological data, and five dimensions are commonplace in a study.

- A method of looking at up to 7 archaeological dimensions has been proposed, with the incorporation of cluster analysis.
- New diagrams have been proposed for the portrayal of the multidimensional data, in particular a multidimensional form of the dendrogram and skyline plot types of diagram.

Bibliography

ALDENDERFER, M. . AND R. . BLASHFIELD 1984. *Cluster analysis, Quantitative Applications in the Social Sciences Series*, Sage Publications Inc., Beverly Hills.

BONSALL, J. C. AND C. LEACH 1974. 'Multidimensional scaling analysis of British microlithic assemblages'. In Laflin, S. (ed.), *Computer applications in archaeology 1974*, Univ. of Birmingham, 16.

FARRIS, J. S. 1969. 'On the cophenetic correlation coefficient', *Systematic Zool.* 18, 279–285.

HARTIGAN, J. A. 1967. 'Representation of similarity matrices by trees', *J. Amer. Statist. Ass.* 62, 1140–1158.

HODSON, F. R. 1968. *The La Tène cemetery at Münsingen-Rain, Acta Bernensia* V, Stämpfli, Bern.

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 (3), 299–320.

HODSON, F. R. 1971. 'Numerical typology and prehistoric archaeology'. In Hodson, F.R., D.G. Kendall AND P. Tautu (eds), *Mathematics in the archaeological and historical sciences*, Edinburgh University Press, Edinburgh, 30–45.

HODSON, F. R., P. H. A. SNEATH AND J. E. DORAN 1966. 'Some experiments in the numerical analysis of archaeological data', *Biometrika* 53, 311–324.

JARDINE, N. 1969. 'A logical basis for biological classification', *Systematic Zool.* 18, 37–52.

JARDINE, C. J., N. JARDINE AND R. SIBSON 1967. 'The structure and construction of taxonomic hierarchies', *Math. Biosc.* 1, 173–179.

MAYR, E., E. G. LINSLEY AND R. L. USINGER 1953. *Methods and principles of systematic zoology*, McGraw-Hill, New York.

MORRISON, D. F. 1967. *Multivariate statistical methods*, McGraw-Hill, New York.

SHENNAN, S. J. AND J. D. WILCOCK 1975. 'Shape and style variation in Central German Bell Beakers: a computer-assisted study', *Science and Archaeology* 15, 17–31.

SNEATH, P. H. A. AND R. R. SOKAL 1973. *Numerical taxonomy*, W.H. Freeman and Company, San Francisco, 59.

SOKAL, R. R. AND P. H. A. SNEATH 1963. *Numerical taxonomy*, W.H. Freeman and Company, San Francisco and London.

WARD, J. H. JR 1963. 'Hierarchical grouping to optimize an objective function', *J. Amer. Statist. Ass.* 58, 236–244.

WILCOCK, J. D. 1975. 'Presentation of computer classification results: A comparison of graphical methods', Science and Archaeology 15, 32–37.

WILCOCK, J. D. 1993. 'Analysis of multidimensional matrices for archaeological data'. In Wilcock, J. D. and Lockyear, K. (eds.), *Computer applications and quantitative methods in archaeology 1993*, BAR International Series S, British Archaeological Reports, Oxford.

WIRTH, M., G. F. ESTABROOK AND D. J. ROGERS 1966. 'A graph theory model for systematic biology, with an example for the Oncidiinae (Orchidaceae), *Systematic Zool.* 15, 59–69.