THE STORAGE, RETRIEVAL, AND CLASSIFICATION OF ARTEFACT SHAPES

Peter L. Main

Research Centre for Computer Archaeology
North Staffordshire Polytechnic

This paper describes briefly a package of Algol routines which provides facilities for the storage and manipulation by computer of outline drawings of all types of archaeological artefact. Particular attention is paid to the classification of groups of data into similar types by complete profile comparison, *a new technique based on the* complete *outline shape of each artefact. Classifications of a group of railway-line sections and a group of medieval cooking pots are discussed as examples of the method.*

The routines to be described have been written in ALGOL 60 and are implemented on an ICL 4130 computer at the University of Keele. They have been developed by the author over a period of two years as part of a research programme towards a PhD. Research on many of the techniques used is still in progress.

The paper falls into two parts:

A) The Storage and Retrieval of Artefact Shapes. B) The Classification of Artefact Shapes.

A) *Storage and Retrieval*

Let us imagine for a moment that an archaeologist finds himself with an example of an artefact of a particularly unusual and interesting type. He wishes to know if objects of comparable shape have been found elsewhere. As things stand at present, probably the best he could do would be to start the tedious process of leafing through hundreds of 'likely' excavation reports glancing at illustrations and hoping to come across something similar. Failure in such a search could well mean that he had simply not consulted an obscure enough publication (or had blinked at the wrong moment!)

Imagine now that we have at our disposal

1. A means of storing every drawing appearing on an excavation report on a *centralised* computer data bank, together with information about its source.
2. Means of automatically comparing any two such stored drawings, that is to say some sort of measure of similarity (or dissimilarity) between them.

We then have the basis of a system for searching through past records of other artefacts similar in shape to some given artefact. If we assume that *all* published material (throughout Britain, say) was added to this data bank as a matter of course when it was published, and that the data bank could be accessed by remote terminals, we have a very powerful archaeological information retrieval system, based on outline shapes, which could be implemented alongside, or even incorporated into, a more traditional text-based information retrieval system.

The routines which have been written for storage and retrieval are used in a man-machine interactive mode using a Visual Display Unit and lightpen. Drawings and text can be displayed on the V.D.U. and the sequence of operations is controlled by the user pointing the lightpen at an 'alphabet menu' in order

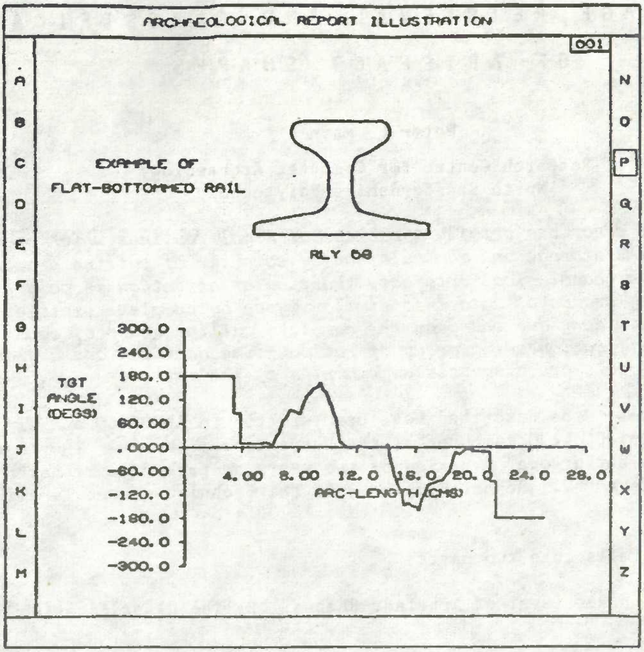


Figure 1.

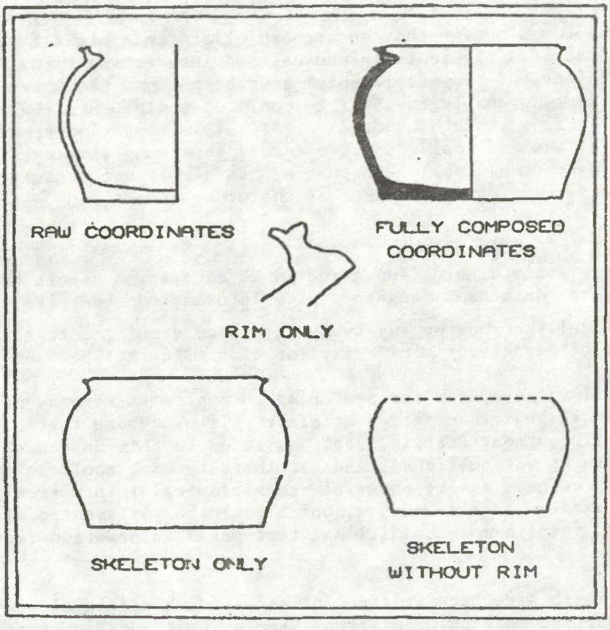


Figure 2.

to select one of the available options. This mode of operation is particularly useful for 'pieceing together' illustrations for an archaeological report, but can also be used; for example, to visually search a file of drawings stored on disc or magnetic tape.

Options currently available include:

1. Input of information in the form of digitised coordinates from an archaeological drawing. Input can be from paper tape, random access disc files or serial access magnetic tape files.
2. Output of such information to disc or magnetic tape.
3. Generation of *tangent profiles* from the digitised coordinates. Generating a tangent profile involves connecting up the digitised points, either by straight lines or by passing smooth curves through them, and mathematically transforming the resulting curve to a graph of tangent angle against arclength. All curves are stored in this form when in main store, this representation being important for the calculation of similarity measures. The original drawing can be easily recovered from its tangent profile for display purposes. Figure 1 shows a flat-bottomed rail section, together with its tangent profile. The frame in this illustration contains the alphabet menu mentioned above.
4. Display of drawings on the V.D.U. screen together with facilities for rotating them, reflecting, expanding, moving them about the screen, adding text, adding lines, framing etc. Particular parts of a drawing can be selected for special attention, and different parts of the same drawing can be displayed with solid or dashed lines, or can be 'blanked out'. Figure 2 shows the basic digitised information of a medieval pot together with various different ways of displaying parts of it.

These routines allow selected drawings to be composed into an illustration which can then be plotted on the digital plotter, in a form suitable for publication in an archaeological report.

B) *Classification*

The similarity measure mentioned earlier can also be used to classify (or 'type') selected groups of objects on the basis of their shapes. This represents a new approach to the classification of shapes. The statistical/archaeological literature is full of experiments, in the classification of (mainly) pottery and handaxes, by employing variations of the following procedure:

1. Select (subjectively) 'features' of the objects and make various measurements related to these features (e.g. on pottery - width of neck, height of neck, width of base, etc. etc.)
2. Use these measurements, or simple ratios of them, as input to a multivariate statistical technique which provides 'clusters' of similar objects.

The weaknesses of this sort of approach are fairly obvious

1. By restricting attention to selected features, the greatest part of the outline profile is ignored.
2. It is not necessarily clear which features should be selected.
3. Choice of features mav radically effect the resulting groupings.
4. Features present on one object may be absent on another.
5. There may be no apparent features to select for measurement in any case.

Of course the weaknesses of this approach are obvious to those who employ it - it is simply that no alternative has been available.

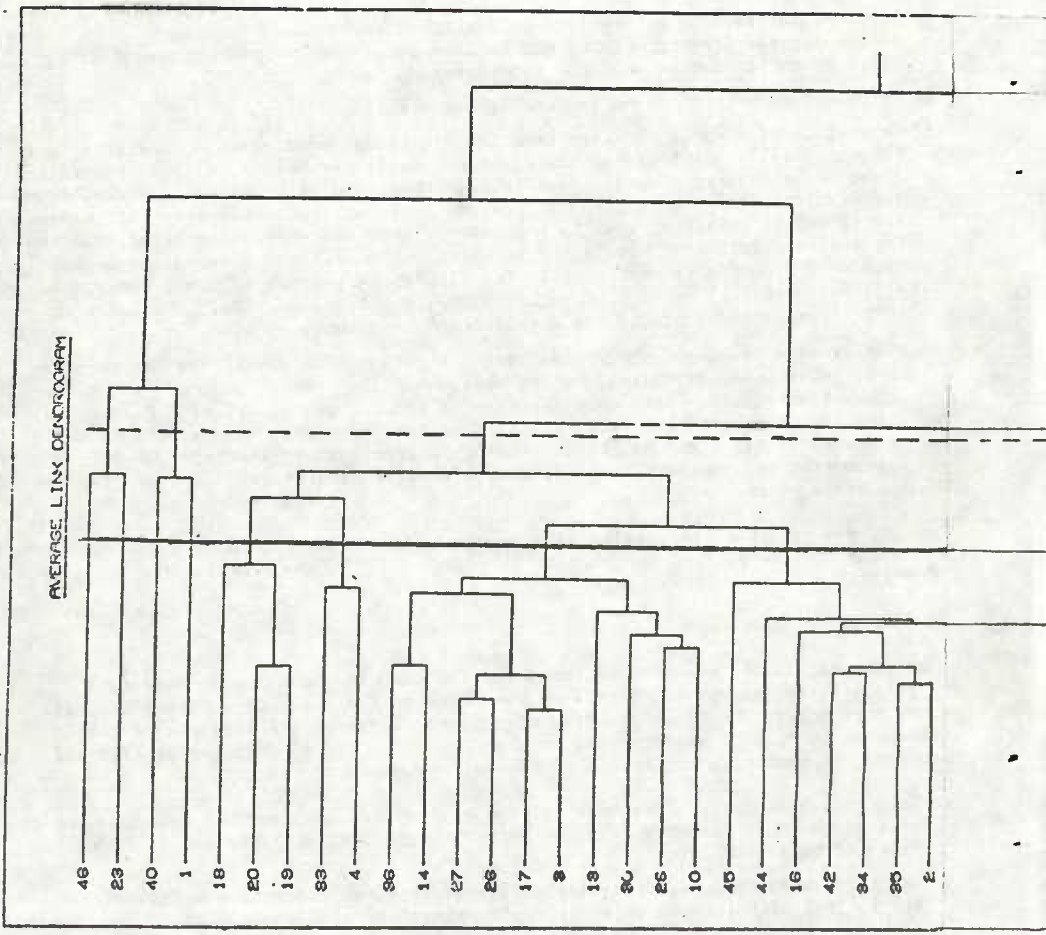
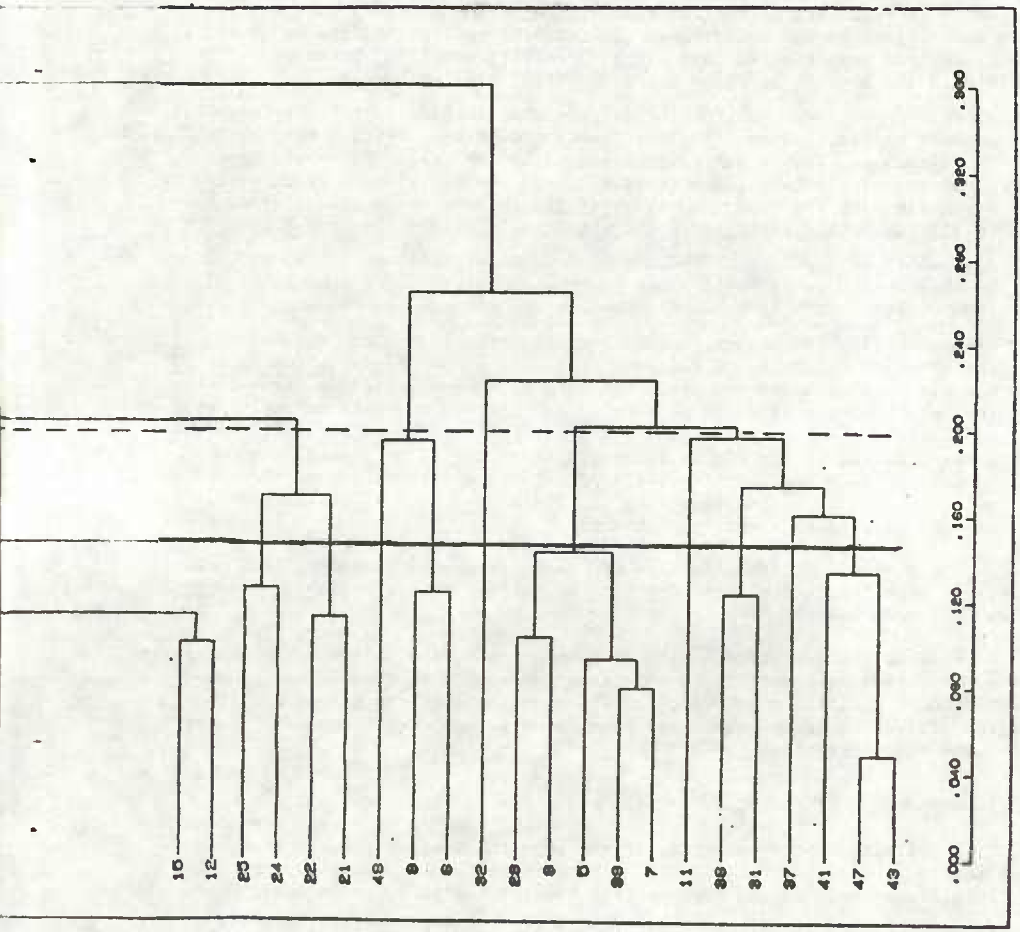


Figure 3.



The routines described here classify objects on the basis of their *entire outline shapes*. Widely differing shapes can be compared. As an extreme example, a measure of similarity between a neolithic axe and a clay pipe could be calculated, though this is not recommended as a particularly enlightening exercise! We will describe such comparison methods as *complete profile comparison* (CPC). In the remainder of this paper, two experiments in the classification of archaeological data by CPC will be described.

It is important to realise that although such a method does not require the identification and measurement of 'features' of the objects under study, this does not mean that no room for subjectivity remains. There are three points in the process at which decisions still have to be made.

1. Since the routines for CPC provide only similarity/dissimilarity measures between shapes, the actual classification must be carried out by established techniques of cluster analysis, using these measures. There are many cluster analysis techniques available, and the decision on which method to use rests with the user. However, if the objects under study 'cluster well', the method employed should not greatly affect the groupings.
2. Any number of types of CPC measures are theoretically possible, and it may be that some measures will prove to be more suitable for some types of data than others. This is presently the subject of research by the author.
3. When comparing two shapes, facilities are built in to the routines allowing the relative weighting of different portions of the curves. For example, the rim area of a pot may be up-weighted or down-weighted relative to the rest of the pot when a CPC measure is calculated, and this weighting will be reflected in the resultant classification or retrieval search. In the extreme cases, if the rim is zero-weighted, it will be ignored for comparison purposes, or if everything *but* the rim is zero-weighted then we will end up with a classification of rims alone, or will retrieve only on rims if we perform a retrieval search.

The author's attitude is that CPC measures should be unweighted, *unless* there is some indication that weighting is desirable (see medieval cooking-pot classification below).

Whereas the decisions having to be made in 1. and 2. above may be regarded as a reflection of a deficiency of current knowledge in these areas, the provision of facilities for *weighted* CPC measures seems to give a desirable degree of freedom to the user, and provides a very powerful tool for the study of shape classification.

A Classification of Railway-line Sections

A collection of 93 drawings of railway-line sections (taken from two publications: see references 2 and 3) were classified by using a combination of a CPC distance measure and Average Link Cluster Analysis, as follows:

1) An upper triangular distance matrix $S = [s_{ij}]$ was calculated, where s_{ij} was a CPC measure of dissimilarity between sections i and j. (The measure was based on usigned area between tangent profiles of i and j, suitably standardised).
2) This matrix was used as input to an average link cluster analysis. The method used was that called 'group-average' by Lance and Williams (reference 1).

This first analysis separated the 93 sections into three very well separated groups. The first group contained all those types of rails known as

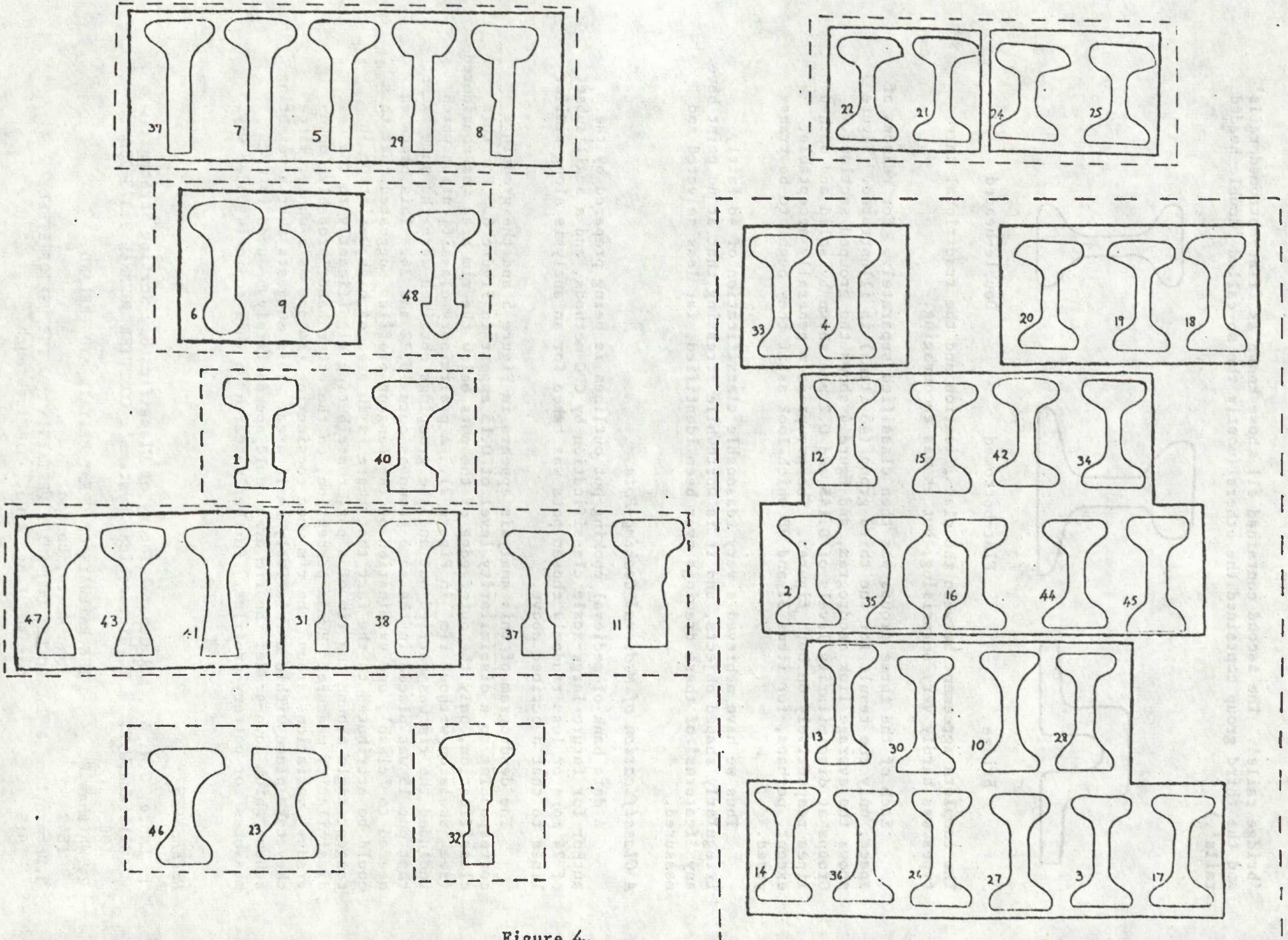
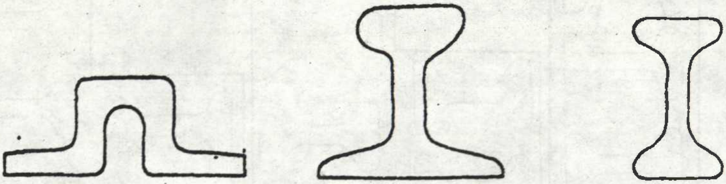

37
7
5
29
8
6
9
48
1
40
47
43
41
31
38
37
11
46
23
32
22
21
24
25
33
4
20
19
18
12
15
42
34
2
35
16
44
45
13
30
10
28
14
36
26
27
3
17

Figure 4.

'bridge rails'. The second contained all those known as 'flat-bottomed rails' and the third group contained the others (mostly the so-called 'double-headed rails'.)



Bridge Flat-bottomed Double-headed

The complete agreement between the classification and the traditional three types was hardly very surprising, but it was encouraging.

Each of the three groups was then classified separately. For reasons of space, only the result for the third group (48 items) is illustrated. Figure 3 shows the average link dendrogram, and Figure 4 shows the grouped sections. Groups at dissimilarity levels of 0.145 and 0.2 are shown by solid and dashed lines respectively on both figures. The groups seem generally acceptable, except, perhaps, for items 37 and 38 which look as if they ought to be transposed.

Thus we have achieved a very reasonable classification of 48 fairly irregularly shaped objects, and it is worthwhile stressing that at no point have any 'features' of these drawings ever been identified, far less selected and measured.

A Classification of Medieval Cooking-Pots

A data bank of medieval cooking-pot outlines is being prepared by the author for future large-scale classification by CPC methods, and a small subset of 24 more or less randomly chosen pots was taken for an analysis along similar lines to that described above.

The dendrogram of this analysis appears in Figure 5 and the groupings corresponding to a dissimilarity level of 0.15 appear in Figure 6. For this classification, only the outer edge of the pots below the rim area was considered (see 'skeleton without rim' in Figure 2). A preliminary classification which included the rim was carried out and the groupings obtained were the same except that pot 15 was placed in the same group as pots 18,22 and 24. This placement seemed to 'clash', and examination of the tangent profiles suggested that this could be attributed to the fact that there is an area of high variation in tangent angle around the rim of a pot, contributing significantly to the dissimilarity measure. Human perception, on the other hand, probably 'down-grades' variations around the rim when considering the pot as a whole, since they take place within a relatively small area. This suggests that a weighting scheme which down-grades the rim area could appropriately be used in CPC measures for pottery profiles. Further research on this topic is under way.

REFERENCES

1. Lance, G.N. & Williams, W.T. 1967 — 'A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems'. COMPUTER JOURNAL 9:373-380
2. Whishaw, F. 1842 — THE RAILWAYS OF GREAT BRITAIN AND IRELAND John Weale, London.
3. Dorman, Long & Co. 1915 — POCKET COMPANION PERTAINING TO THE USE OF STEEL Dorman, Long, & Co., Middlesbrough.

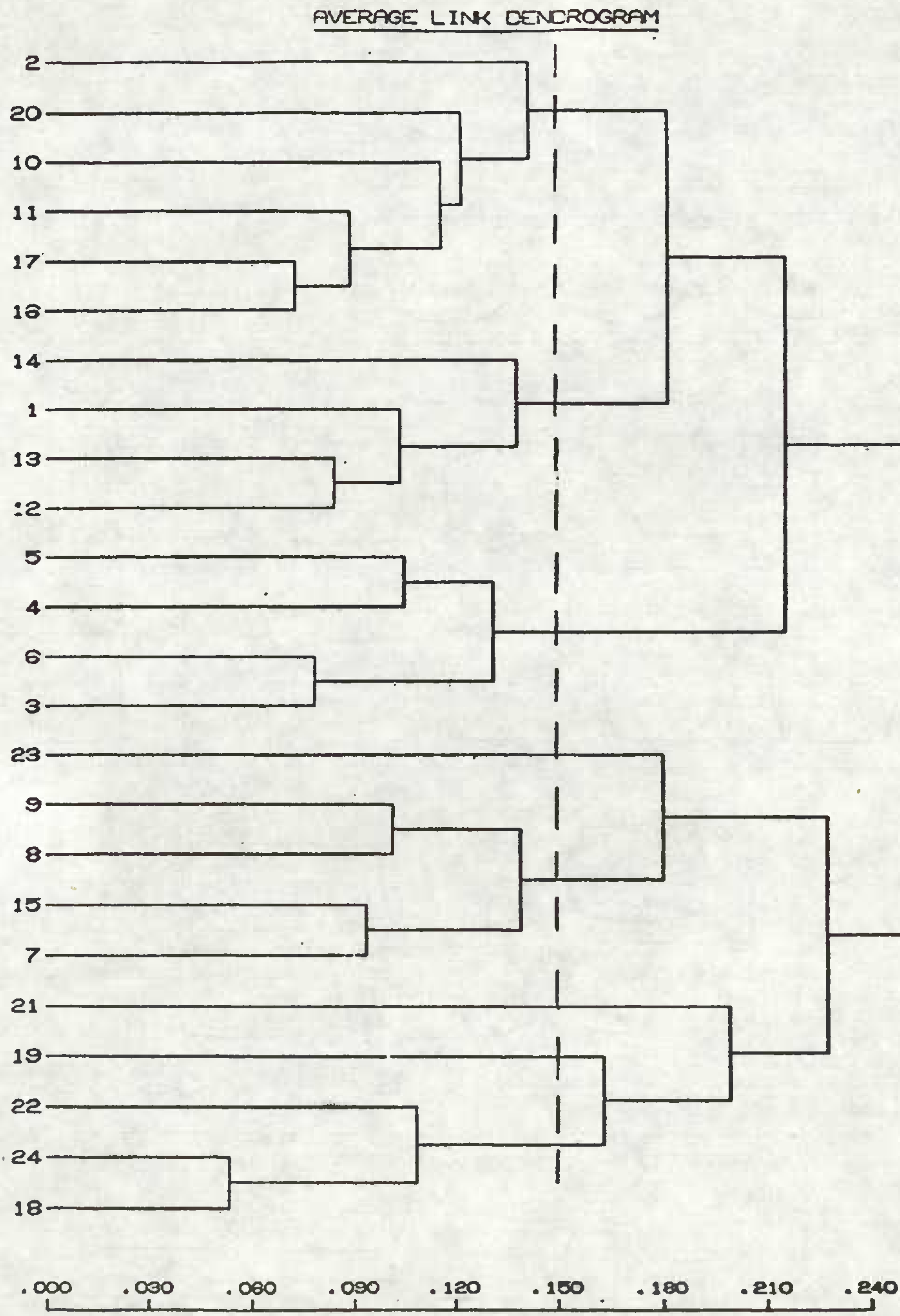

AVERAGE LINK DENDROGRAM
2
20
10
11
17
16
14
1
13
12
5
4
6
3
23
9
8
15
7
21
19
22
24
18
.000
.030
.060
.090
.120
.150
.180
.210
.240

Figure 5.

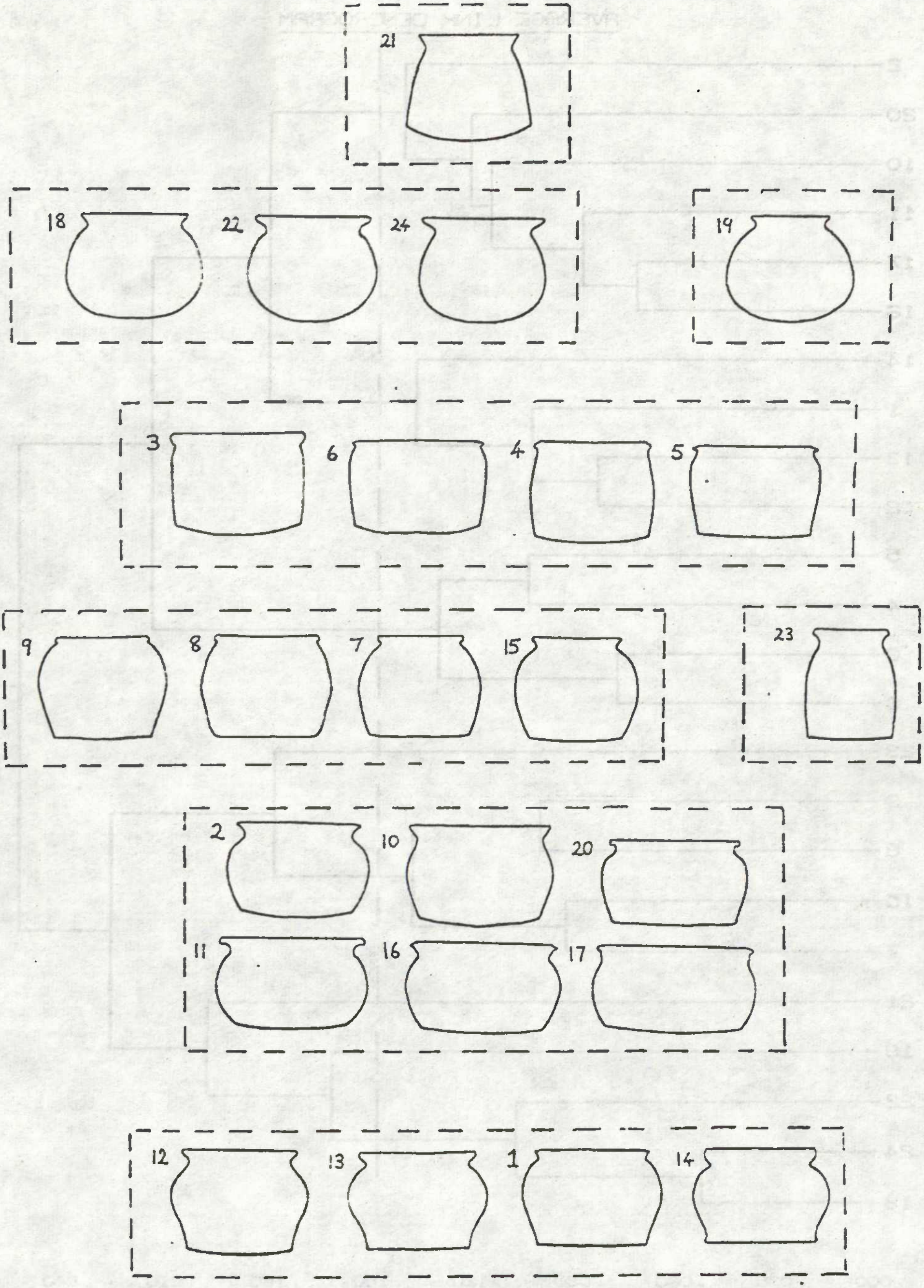


Figure 6.