Cluster Analysis using Fragmentary Data

Patricia Galloway

A student of archaeology who concentrated his study only upon printed reports might be excused for assuming that complete examples of artifacts were frequently found, but as soon as he took to the field he would encounter the all too seldom discussed problem of the fragmentary survival of artifacts. This is a problem which thousands of drawings of reconstructed artifacts will not alleviate, and it is one which must be encountered at some time by those of us who are interested in quantifying large classificatory problems.

I am working at present on a corpus of medieval antler combs from the urban centre of Trondheim, Norway --a corpus which will eventually include about five hundred combs--and fragmentariness was an immediate difficulty. Antler combs, by virtue of their structure and function, are far more likely to be found in a fragmentary state than not; a sample composed entirely of whole combs cannot hope to be representative of the population as a whole. Accordingly, I have been obliged to approach automatic classification with the requirement that it enable me to deal adequately with a body of material of which the majority of members is fragmentary. In addition, I need a method which will allow me to study those examples of which only small fragments remain and to assign them to one or more classes with some measure of acceptability. As a first step in a long-range programme of study, I have tested parts of the CLUSTAN suite of programs on this data. This paper is a report of my efforts and of the progress I was able to make.

The whole question of fragmentary attributes has received very little attention; automatic classification procedures are usually used on objects like metalwork which are most frequently complete. It was not possible to approach even the formation of the attribute list without evaluating in turn each step in the classification procedure, to determine exactly where the difficulties presented by fragmentariness lay and to determine, if possible, how the best could be made not only of the design of the attribute list but also of each decision required for choosing a classification procedure.

It is first necessary to say something of the combs themselves. Much is known of the comb-maker's craft. Evidence drawn from workshop finds has made almost all stages in the construction process clear (Hilczerowna, 1961; Thomas, 1960). For example, we know from finds of shaped connecting plate blanks that the shape of this piece was decided from the outset, and that this shape governed a number of steps which followed. In fact, it can be said with some assurance that every observed feature is an intentional one; finds of spoilt unfinished combs confirm this assertion. On a small,scale, over a number of classes, it can be said that randomness of attributes is essentially unimportant.

Another thing which should be mentioned is the way in which fragmentariness manifests itself with these artifacts. The connecting plate nearly always survives to a great enough extent that its shape can be identified. Conversely, the teeth and end segments of the comb are far less likely to survive, but quite often the shape of one end segment can be ascertained. These two shapes are essential to the conventional definition of types; it is frequent practice in the formation of comb typologies to use just these shapes as the initial determining factors in the division of groups of combs (compare the headings in Roes, 1963, and the typologies in Hilczerowna, 1961, and Kellmer, in press), so it follows that fragmentary combs can often be related to initial large groupings. The real problems ariđe when subclasses within large classes are to be defined and attributes referring to decoration are brought into play. In an hierarchic sense, the fragmentary combs are able to ascend the hierarchic tree easily through the first few branchings, but get 'stuck' at some further branching which is based upon an attribute that they lack.

Finally, some notion of the proportionate numbers of combs in intuitive groups should be outlined. It is fairly clear that some classes were far larger than others; it has even been suggested that a given workshop was likely to produce at any one time many 'production model' combs and a few very handsome and relatively more expensive combs (Hilczerowna, 1963). This means that some classes of relatively simple combs with only a few attributes will be very large, while some classes of more decorated combs will be rather small. For this reason it is necessary that we allow for a large range of class sizes.

Little effective work on the classification of combs and the evolution in their typology has been done because efforts towards a classification system have been plagued by the failure to recognise the fact that this artifact has four clearly defined <u>form</u> categories, each of which tends to be chronologically ubiquitous, within each of which there are very definite <u>typological</u> changes. It is therefore nearly useless to classify by form categories, which are perfectly obvious, and thus to ask of a classification method that it perform four <u>sets</u> of classification simultaneously. To begin with, I judged that each form category should be taken separately. Now, for the period (11th-13th centuries) and the place (Norway), the actual number of combs is dominated by the forms which are composite (made of several pieces of horn), while the single-sided and double-sided forms occur in about equal numbers. I decided to take the form class of single-sided composite combs, which is capable of the most complex variation, for the initial study. This class presents a complex problem whose solution, if found, will answer nearly all questions of comb classification.

With the project confined to this form class, we are free to turn to the choice of attributes. When a corpus contains many fragmentary examples, the definition of the attribute list becomes an even more serious matter than it is ordinarily. Now the attributes which I think are relevant to the classification of this group of combs are for the most part binary. But if many attributes will be missing and therefore unscorable in binary terms, some adjustments must be made. Sokal and Sneath (1963) offer little helpful advice on this question. Discussing the situation in which it is necessary to compare fragmentary with whole specimens, they suggest the introduction of three-state attributes whose values would be 'present', 'absent', and 'no comparison'; in the analysis, they suggest that those comparisons which contain at least one 'no comparison' score be eliminated, but they urge that such instances not be allowed to be frequent because of the error that they introduce into the coefficients of comparison thus calculated(pp. 162-165). With a sample whose members are more likely to be fragmentary than not, such advice is not of much use. Similarly, Doran and Hodson (1975, p. 104) have suggested the use of a three-state attribute where missing attributes must be scored, but give no further specific guidance with the problem beyond the helpful mention of techniques which will be unsuitable if material is fragmentary. It is clear that we must seek some other means of dealing with the attributes, means which will be appropriate to the data and which will prevent the distortion of classes by the dominance of fragmentary members.

I have used and modified an attribute list of my own devising to record comb data for a period of two years. It has provided an adequate system for the description of combs, but it will have to be adapted somewhat for use in classification. The list contains a series of nine measurements, shape specifications for the connecting plates and end segments, decorative motifs used in both of these locations, details regarding the number and placement of the rivets, and finally a series of more contingent attributes, such as whether the comb possesses a certain functional perforation. From the start, some of these attributes can be set aside, some need to be given binary equivalents, and still others must be adjusted to compensate for fragmentariness.

It is not impossible to mix continuous with discrete attributes in one analysis, but it is not conveniently done. In the case at hand, I eliminated many of the measurements on the grounds that they were included in a more general way in some of the shape attributes. It was really necessary to include only one measurement, that of the length of the comb as indicated by the length of the connecting plate. This part of the comb is not only most likely to survive complete, but since it is related to a constant width for the tooth segments, it is also a good indicator of the proportions of the comb. A graph of number of combs against length of connecting plate showed two peaks divided by a major break, so the continuous length attribute was divided into two discrete attributes.

As I have said previously, some attributes are very little affected by fragmentariness, and thus present no problem to binary scoring. If a certain feature is quite likely to suffer damage, however, some alteration must be made if binary attributes are to be used. This has been done by adjusting the attributes most prone to damage so that a negative or absence score actually indicates that the attribute cannot be scored. For example, a piece may have no decoration either because it is actually undecorated or because the portion which would ordinarily bear the decoration is missing. To allow for this, we keep the attribute for decoration and add one for plainness, so that the absence of both indicates that the part is fragmentary.

Once the attribute list is adjusted in these ways, we have a set of presence/absence attributes which are felt to approximate fairly cbosely an objective tabulation of the distinguishing features of the artifact. The next step is to establish just what it is that we expect of an automatic classification procedure in the case of combs.

I decided to begin the process with an examination of the possibilities offered by the methods of cluster analysis. My idea of the way that a clustering procedure for this corpus should proceed is as follows:

1)As a training set, cluster complete combs, permitting the clusters to be polythetic and of variable size, to allow for the distributional distortion introduced by using complete examples. Some hierarchic method would be desirable to show relations between clusters.

2)Assign to these clusters those fragmentary examples which have enough attributes to meet a set similarity threshold.

3)Recalculate cluster centres, including the fragmentary data added in step two.

4)Derive a key for each cluster.

5)Assign remaining fragmentary combs to clusters at whatever hierarchic level they meet similarity requirements.

To a certain extent these steps may be followed by using programs in the CLUSTAN package, or at least so it seems; in the section which follows I will show how the programs were tested against this idealised sequence.

Clustering whole combs

The first step was to find which of the set of programs would provide the 'best' clustering of the complete combs. Having the fragmentary examples in mind, however, we must apply certain constraints from the beginning. First, because of the clarity of the notion of a tree structure in which fragmentary examples would begin at the most general level and ascend branchings until 'caught' at a branch due to lack of information, it is desirable to have a method which will relate clusters in a tree structure. Accordingly, I decided to test both the monothetic divisive DIVIDE and the agglomerative HIERARCHY on the set of complete combs. Decisions regarding coefficients of similarity and correlation as well as clustering methods are also affected by fragmentary data, but these will be discussed below.

To use the program DIVIDE, the user chooses an association coefficient which the program will use to select the attribute for division which defines by its presence and absence the two clusters whose difference is maximum. It should be noted that the CLUSTAN user's manual leaves the issue of coefficients rather confused, for though five 'standard methods' are defined, the table of forty coefficients from which one may also choose leaves nearly half of them unlabelled or inaccurately labelled. I chose two of the standard methods, the so-called empirical coefficient which represents the sum of the absolute values of AD-BC and the chi squared measure for comparison. It was hoped that the first coefficient would allow even for seemingly random correlations and for a rather high degree of inequality in size between clusters. It was felt that neither of these would be allowed for with the chi squared technique.

Because DIVIDE is a monothetic division method, it singles out an attribute and divides a group into two on the basis of its presence or absence. Because it chooses the attribute with the highest positive or negative correlation with other attributes, it runs the risk of creating anomalous clusters. When the program was run with each of the two coefficients on the 57 complete combs, results were quite dissimilar in terms of the relations between clusters, though six of the clusters at the sixteen-cluster level were identical in both cases. My judgment of the 'goodness' of an application depends upon whether I think the groupings into types are correct. By this criterion the 'empirical coefficient' was less effective than the chi-squared method because it tended in two important cases to divide into quite distant groups combs which, in one of the cases, were correlated on every attribute but one. Both had trouble with a large class of related combs which share relatively few common attributes and are prone to several variations in combination with other attributes, but with one exception tha chi squared method managed to group them within a single large cluster while the 'empirical coefficient' split them up at the first division. It seems to me that this, an unfortunate major division, is the great risk in the monothetic division method, which I was reluctant to use for anything but comparison from the start. Its great virtue is that by tracing the process of division and noting which attributes govern successive divisions, it is possible to obtain a very good idea of the relations between attributes in the corpus. And although the material is at least partly polythetic and therefore essentially at odds with the method, it is interesting to see what monothetic structures are to be found in the data.

In contrast, I had an affinity to the methodology of the HIERARCHY program from the beginning. On the basis of a similarity coefficient chosen by the user (I had to discover by experiment that the Jaccard coefficient which I wanted to use was actually coefficient five), it fuses pairs of objects and clusters so that the result is a complete hierarchy from leaves to root. I chose the Jaccard coefficient because it does not take negative matches into account, and if we are to cluster fragmentary combs in the same way, it is the obvious choice. In addition, a clustering method had to be chosen, and I decided upon average-link because of the possible ill effects of complete-link with polythetic data and similar incompatibility between single-link and the anticipated fragmentary data.

The results obtained were certainly better than both DIVIDE results, and a relative degree of coherence was even introduced into the large class which caused trouble before. I found that the complete dendrogram was especially illuminating, and I now feel that much useful information might be obtained by allowing the DIVIDE program to make a complete analysis. But the real strength of the HIERARCHY program is that it allows for polythetic clusters, and I think this explains its success on the comb material.

Clustering all combs

It is not really possible to proceed to step two in the idealised list above, adding fragmentary combs to established clusters, without doing a large amount of sorting not provided for in the CLUSTAN package, so I decided at this stage to cluster the full data set with all the combs exactly as I had with the whole combs alone. Such an analysis would at least provide an idea of the necessary threshold of similarity which would permit only meaningful additions of fragmentary combs to established clusters. It would also permit comparisons showing changes in clustering criteria with the addition of a mass (76 additional combs) of further material. I should remark on a change which is made with the addition of more examples to the analysis, and that is that if the same number of clusters is chosen as before, the resulting dendrogram will then represent a higher level of analysis where even more clusters are included under the leaves.

Expecting results to be unspectacular, I was not surprised to find that both DIVIDE and HIERARCHY tended to make clusters composed entirely of fragmentary combs. At first glance, neither DIVIDE run was too badly condemned on this score, producing three fragmentary clusters with chi squared and four with the 'empirical coefficient', while the HIERARCHY run produced ten of them, all out of sixteen clusters. Looking more closely, however, I found that the inoffensive three and four clusters accounted for seventeen and twenty-four objects respectively, while the ten clusters averaged two objects each. In fact the HIERARCHY program had succeeded much better than the other two runs at integrating the fragmentary combs by the sixteen cluster level, and the fragmentary clusters are in fact only 'late joiners' into established large clusters, for the very good reason that they have very few attributes and have got 'stuck' at a branching. The program proves itself to be extremely robust, for it has actually drawn the same dendrogram, only inserting more leaves.

Such is not the case with the DIVIDE runs, since the added data has altered the correlation of attributes. It is quite instructive, however, to notice how certain patterns of the monothetic divisions are repeated. The addition of fragmentary combs has not allowed one method to show improvement over the other, either; the chi squared division still does the better grouping of clusters.

By untying the CLUSTAN package a bit, one could go somewhat further towards the implementation of a classification procedure; fragmentary combs could be added by hand to a classification array obtained from the HIERARCHY program, which could then be run through the RELOCATION program using a threshold value so that only relatively complete combs would be added and cluster centres recalculated accordingly. The derivation of a key would, however, remain to be done by hand, and the assignment of remaining fragmentary combs could not be done automatically. For these last two steps programs remain to be written; possibly the RELOCATION program could be adapted to do the assignment task. I began this part of the project with the intention of determining which of the two methods would work best on this data, and I have found that one of them works satisfactorily because I like the results. This outcome, however gratifying, does not change the fact that the actual suitability of various methods and options is still unproven in a definitive way. What is needed in general and certainly for this project in particular is a model study which would assess this suitability. I hope that I have at least shown that the problem of fragmentary data should be included in such a study.

References

Doran, J.E., and Hodson, F.R., Mathematics and Computers in Archaeology, Edinburgh, 1975.

Hilczerowna, Zofia, 'Rogonictwo Gdanskie w X-XIV Wieku', Gdanskie Towarzystwo Naukowe, Gdansk, 1961.

Kellmer, Inger, (an as yet untitled study of the combs from excavations in Bergen), Oslo, in press.

Roes, Anna, Bone and Antler Objects from the Frisian Terp-Mounds, Haarlem, 1963.

Sokal, R.R., and Sneath, P.H.A., Principles of Numerical Taxonomy, London, 1963.

Thomas, Sigrid, Studien zu den germanischen Kämmen der römischen Kaiserzeit, Leipzig, 1960.