# Classification and Identification with Incomplete Data

P.H.A. Sneath
University of Leicester.

It is surprising that the problems of classification and identification with incomplete data have received so little serious study. Although these problems are particularly acute in certain fields, such as archaeology, they are in fact ubiquitous. It is not possible to work in any broad field without finding that full comparative information is impossible, or impracticable, to obtain. The botanist compares plants with other plants, the entomologist compares insects with one another, but only rarely does anyone compare plants with insects, and even then only selected instances; it is impracticable to compare all plants with all insects.

Yet the fact that classification and identification are such general (and successful) activities shows that incompleteness of information is not per se a serious drawback. It is particularly important that the data should be as complete as possible within certain sections of the material. When order has been obtained within these sections one can compare a selection with other material that is not too dissimilar; in this way a complex picture can gradually be built up, even though the total proportion of information is small. The more easily one can recognize well-defined types, and choose typical exemplars for further studies, the easier it is to make a satisfactory picture of the whole. It may be thought that if the gaps in the data are entirely haphazard the difficulties would be greatly increased. This may not necessarily be so, however, and we have little information on this point.

The logical form of the classification has a major influence. Two important alternatives are polythetic and monothetic classifications. These may be illustrated by the small data tables below, where the occurrence of various attributes (indicated by letters) is shown in some objects (indicated by numerals).

| Polythetic class | | | | | Monothetic class | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Objects | Attributes | | | | Objects | Attributes | | | |
| 1 | A | B | C | | 5 | E | F | G | |
| 2 | A | B | | D | 6 | E | F | G | |
| 3 | A | | C | D | 7 | E | F | | H |
| 4 | | B | C | D | 8 | E | F | | H |

The difference between these types of classification is that in a polythetic class the members are defined by sharing a high proportion of attributes, but no particular attribute must always be present. The unity of objects 1-4 is evident, even though no attribute is found in every one.

In contrast the class of objects 5-8 is defined by the presence of attributes E and F in every one: no exceptions are permitted. In this instance a monothetic subdivision can be made on the basis of attributes G and H. There are analogous

principles in schemes for identification.

The example above illustrates two main patterns of missing information, scattered gaps on the left, and gaps affecting substantial parts of columns of data, on the right. In both classification and identification the pattern of gaps will affect polythetic and monothetic systems differently. Polythetic schemes are in general resistant to disturbance by scattered, randomly-placed gaps, because their effects are similar to those of the exceptions for which polythetic systems make provision. In contrast, monothetic methods may show serious failure if a gap occurs in an essential attribute. However, we do not know how these logical methods compare under a variety of patterns of missing information, and this is an area of some interest for future study.

The useful term relevance has been suggested for the proportion of available data. Each of objects 1-4 exhibits relevance of 75% with respect to the attributes A-D. The relevance of attributes G and H for objects 5-8 is 50%. A fuller exposition of relevance is given in Sneath & Sokal (1973, p.181). Work relating to this general problem is described by Sokal & Rohlf (1970) and Moss (1971).

It is useful to consider missing information in the same framework as the effect of errors in attributes. One convenient representation of classification and identification is to treat the objects as having a position in a multidimensional space, which has one dimension for each attribute or variable. The effect of errors and of gaps can then be visualized as introducing uncertainty of position of objects (in the sense that the apparent position is not the same as that one would obtain from complete, correct, data). The effect of errors has been studied in classification and identification (Sneath & Johnson, 1972; Sneath, 1974), and they will be later compared with the effect of gaps.

One particular difficulty in handling the effects of gaps is that these are equivalent to reducing the number of dimensions of the multivariate space. There are conceptional problems in visualizing this. It can be shown that under certain simplifying assumptions the amount of uncertainty is related to the chi-square distribution, if the uncertainty is measured as the risk of unrecognized overlap in the reduced dimensions (Sneath, 1980). There are also further effects. The introduction of gaps in the vectors used to calculate similarity coefficients between objects can lead to these coefficients losing certain desirable properties. Thus, the distances between the objects may become non-metric, and suggestions for replacing the gaps by "estimated" values (Jičin & Vašiček, 1969) have received little study. Another effect is that the degrees of freedom of the similarity values are reduced, and the reduction can be erratic, thus leading to difficulties with significance tests.

If we turn to what is known about the effects of missing data on numerically-produced classifications, it may be first noted that a small percentage of randomly-spaced gaps has little effect. Some preliminary experiments with J.W. Carmichael (noted in Sneath & Sokal, 1973, p.181) investigated the correlation between a similarity matrix based on complete data and that based on data with varying proportions of

randomly-placed gaps, using the Simple Matching Coefficent. As the proportion of gaps was increased from 0% to 10%, 20% and 33.3%, the correlation coefficient r fell from 1.0 to 0.98, 0.95 and 0.89 respectively. The fall-off is not linear, though this small study provides little evidence of the exact shape of the curve. The data table was, however, fairly large (85 objects and 49 attributes), and typical of taxonomic work with bacteria. It seems probable that the curve of fall-off is sigmoid: one expects that a few gaps will have little effects, and then as the proportion is increased the correlation will fall rapidly and remain low. This is a parallel to the effects of increasing errors, where a sigmoid behaviour is found (Sneath, 1974). In the latter instance it is explainable if the effect of perturbation is to expand a multivariate normal swarm past a critical radius.

What is of particular interest is the resistance of the usual numerical taxonomic methods to quite large proportions of missing data if it is distributed in a haphazard fashion. Thus Crovello (1968) found that r between a similarity matrix based on almost complete data and one with about 50% of gaps was about 0.89.

Another pattern of missing information is one where the only similarities available are those to a restricted set of reference objects. Thus, a typical specimen of some class may be chosen, and comparisons made between all specimens and that one (but not between all possible pairs of specimens). This gives a similarity matrix containing only certain rows and columns of similarity values. Such incomplete matrices can occur in material where similarities are readily determined, but the underlying attributes are not accessible (for example tables of serological resemblances between organisms). They may also occur where the comparison of all specimens becomes too laborious, and a restricted set of reference specimens is used for practical reasons, and this is likely to be a major cause of this situation.

I have recently made a preliminary study of this (Sneath, 1980) in which a complete similarity matrix is calculated from the incomplete one by treating the initial resemblances to the reference specimens as variables, and computing taxonomic distances. The results of this operation produce very marked distortions of the original relationships. Peripheral objects are pulled in toward the reference specimens. There is strong compression of the objects except the reference specimens. If the material contains clusters, these clusters may not be recognizable after the operation, unless each cluster is represented by a reference specimen. This therefore places the taxonomist in a difficulty: if he has already correctly recognized the clusters he will obtain results that show the clusters fairly well, but if he has not, he may fail to find the clusters at all.

Turning to identification, it may first be noted that the usual diagnostic keys (a common identification system in most branches of science) are closely analogous to monothetic classifications. If the information required by the initial couplet is missing the key is unworkable. Keys are therefore sensitive to absence of particular items of information, and this dependence on critical data is still seen, (though to a

diminishing extent) when keys are elaborated to contain several items per couplet. A good discussion of a botanical example is given by Pankhurst (1975).

In contrast, those identification systems that are analogous to polythetic classifications are not sensitive to the lack of particular items of information, though they are obviously sensitive to the total quantity of information that is provided. The polyclave "Peek-a-boo" systems make provision for exceptional attributes, and are thus partly protected against information loss. Also protected are the identification schemes that utilise a similarity measure based on numerous attributes. These are commonly constructed so that the identification of an unknown specimen is achieved by comparing it in turn with each class to which it might belong. The comparison is often made to the centroid of the class, and the class may also be surrounded by an envelope (commonly spherical) that represents the limit of acceptable class-membership. In such schemes, therefore, a positive identification is obtained when an unknown specimen, scored for a number of attributes, is represented by a geometric position that is well within a single "sphere".

We have some knowledge of the amount of information required to construct a successful identification system (Barnett & Pankhurst, 1974; Sneath & Chater, 1978). The number of attributes needed is close to the number of classes in the system, or perhaps a little less. It is far greater than the theoretical minimum, where m presence-absence attributes can distinguish $2^m$ classes. This is true for monothetic systems like keys, and for polythetic systems. This broad rule seems true for most types of naturally-occurring variation.

Computer-assisted identifications in bacteriology have been recently developed (Lapage et al., 1973), based on polythetic systems, with taxonomic distances or their analogues. These have afforded a good deal of experience on the proportion of errors and gaps that can be tolerated in such systems. It has been noted that one can frequently achieve quite good results when the number of attributes scored is as low as half the number of classes. The proportion of errors that can be tolerated is naturally lower.

In order to get some idea of the relative sensitivity to errors and to gaps, I have made some preliminary experiments with the bacterial matrix of fermenters published by Lapage et al. This matrix contains 56 bacterial species and 56 attributes (scored as percentage positives), and a typical strain of Actinobacillus lignieresi was used as an unknown. To the vector of attributes of this strain successively more gaps, and more errors, were introduced, and also combinations of the two. These were introduced in a random manner, and the probability index used by Lapage et al. was employed as the test statistic.

The behaviour was not entirely smooth, but the probability assigned to the identity Actinobacillus lignieresi fell slowly from over 90% to less than 10% when the number of gaps increased from about 20 to about 50. The curve of successful identification is thus roughly sigmoid (as for classification), but not very steep. The corresponding fall in probability on adding random errors occurred when the number of errors (i.e.

attributes that were re-coded in the opposite fashion, + for -, and - for +) increased from about 10 to about 20, also in sigmoid fashion. For mixtures of gaps and errors the results were roughly intermediate.

Though small, this experiment shows features similar to that for classification. Serious degradation of the correct solution does not occur until the proportion of gaps is about 50%, or the proportion of errors is about 20%. These conclusions are of course based on very restricted material. A theoretical study to relate the effect of the proportion of gaps to the proportion of errors would be of great value, as the results could then be extended to a much wider range of situations.

REFERENCES


Barnett, J.A. & Pankhurst, R.J. 1974 — A NEW KEY TO THE YEASTS. North Holland Publ. Co., Amsterdam.

Crovello, T.J. 1968 — 'The effect of missing data and of two sources of character values on a phenetic study of the willows of California'. MADRONO 19:301-315.

Jičin, R. & Vašiček, A. 1969 — 'The problem of the similarity of objects in numerical taxonomy'. JOURNAL OF GENERAL MICROBIOLOGY 58:135-139.

Lapage, S.P., Bascomb, S., Willcox, W.R. & Curtis, M.A. 1973 — 'Identification by computer: general aspects and perspectives'. JOURNAL OF GENERAL MICROBIOLOGY 77:273-290.

Moss, W.W. 1971 — 'Taxonomic repeatability: an experimental approach'. SYSTEMATIC ZOOLOGY 20:309-330.

Pankhurst, R.J. 1975 — 'Identification by matching'. BIOLOGICAL IDENTIFICATION WITH COMPUTERS, ed. Pankhurst, R.J., Academic Press, London. pp.79-91.

Sneath, P.H.A. 1974 — 'Test reproducibility in relation to identification'. INTERNATIONAL JOURNAL OF SYSTEMATIC BACTERIOLOGY 24:508-523.

Sneath, P.H.A. 1980 — 'The probability that distinct clusters will be unrecognized in low dimensional ordinations'. CLASSIFICATION SOCIETY BULLETIN 4(4):22-43.

Sneath, P.H.A. & Chater, A.O. 1978 — 'Information content of keys for identification'. ESSAYS IN PLANT TAXONOMY, ed. Street, H.E., Academic Press, London. pp. 79-95.

Sneath, P.H.A. & Johnson, R. 1972 'The influence on numerical taxonomic similarities of errors in microbiological tests'. JOURNAL OF GENERAL MICROBIOLOGY 72:377-392.

Sneath, P.H.A. & Sokal, R.R. 1973 NUMERICAL TAXONOMY. W.H. Freeman & Co., San Francisco.

Sokal, R.R. & Rohlf, F.J. 1970 'The intelligent ignoramus, an experiment in numerical taxonomy'. TAXON 19:305-319.