(4) With this approximation and the Chalkley estimate, V', both inserted in equations 1 and 2, the following equations are obtained:

$$V' = \frac{\pi}{4} \Sigma p_i L d_i^2$$

$$\text{and } A' = \pi \Sigma p_i L d_i$$

V', p_i, and d_i values are available from the data. Solving the first equation for L and substituting it in the second we get:

$$L' \text{ (an estimate of } L) = \frac{4\ V'}{\pi\ \Sigma\ p_i\ d_i^2} \quad (4)$$

$$\text{and } A' = 4\ V' \left(\frac{\Sigma\ p_i\ d_i}{\Sigma\ p_i\ d_i^2} \right) \quad (5)$$

The latter provides the estimate of total surface area sought. The fraction on the right is the ratio of the average diameter to the average squared diameter (not average diameter squared). An improper simplification is $A' = 4\ V' / \Sigma p_i\ d_i$. It is possible to obtain estimates of the surface area in each diameter class also, but these are less reliable.

JOHN D. HAYNES

Lederle Laboratories,
Pearl River,
New York.

[1] Vogel, A. W., *Proc. Amer. Assoc. for Cancer Res.*, **4**, 70 (1963).
[2] Chalkley, H. W., *J. Nat. Cancer Inst.*, **4**, 47 (1943).

An Objective Method of Weighting in Similarity Analysis

THE use of 'similarity' methods of hierarchical classification in numerical taxonomy is now well-established, and is particularly associated with the successful studies of Sneath *et al.*[1-3]. In these methods individuals or groups are successively united in larger groups to form a hierarchy or 'family tree', the aim being to unite the most 'similar' individuals or groups first. Various coefficients of similarity have been proposed, some of which are discussed by Dagnelie[4], but they are usually closely related to the Euclidean distance between individuals (or between centroids of groups) plotted in an N-dimensional space where the jth co-ordinate for an individual is 1 if he possesses the jth of the N attributes considered, and 0 if he lacks it.

It has always been known that these methods may fail if very few attributes are available or if many of the attributes are lacked (or possessed) by nearly all the individuals. This disadvantage is not so marked in Tanimoto's[5] method, where the individuals are clustered about a number of apices, or 'most typical members' of their clusters, the apices being determined by the use of a similarity measure closely related to Sneath's. However, this method becomes cumbersome when the number of individuals greatly exceeds the number of attributes. Thus both Sneath's and Tanimoto's methods may be inappropriate in, for example, ecology, where few species may be present and some of these may be rare; a similar difficulty may arise in such human sciences as psychology, sociology or criminology. Sneath has always stressed that his methods are not, and are not intended to be, applicable to such data. The difficulty arises as a result of the intrinsically low information content per individual of the sample to be classified; there may well be a large number of equally good 'best' fusions available at any stage, and the one actually chosen may set the course of the subsequent analysis on an unprofitable path. Further information must, therefore, be imported into the system.

One obvious step is to assert that some attributes are more important than others in determining similarity. This may be decided from outside the data[6], using prior knowledge of the field; but this usually destroys the impersonality that is the single most desirable feature of these systems. Alternatively, we may determine the scale along each axis of our N-space (that is, estimate the relative importance of each attribute) internally from the data themselves.

In another (monothetic) method of hierarchical classification, known as association analysis[7], the 'importance' of the attributes (in a rather different sense) is measured in the following way. χ^2_{jk} is calculated between every pair of attributes j and k (in terms of the number of individuals possessing or lacking them singly or jointly) and the sum $\sum_{k \neq j} \chi^2_{jk}$ is formed of all the χ^2 which involve a particular attribute j. Although association analysis and similarity methods are not truly comparable, we decided empirically to investigate the use of $\sum_{k \neq j} \chi^2_{jk}$ mentioned here as a weighting coefficient for the jth attribute in a similarity analysis. Accordingly, writing the co-ordinates of two individuals (or of the centroids of two groups) as $(x_{11}, x_{12}, \ldots x_{1j} \ldots x_{1N})$ and $(x_{21}, x_{22} \ldots x_{2j} \ldots x_{2N})$ we define the square of the distance between them as $\sum_{j=1}^{N} \left\{ (x_{1j} - x_{2j})^2 \sum_{k \neq j} \chi^2_{jk} \right\}$ and we successively pool the two nearest individuals or groups with nearest centroids (not the groups with nearest neighbouring points as originally proposed by Sneath[1]).

Although we make use of χ^2 calculations, the analysis is not probabilistic. Probabilistic similarity methods could be defined and developed (we are working on this), but here we have simply devised a convenient grouping of points plotted in an arbitrarily-scaled Euclidean space; the purpose of this is to give rise to testable hypotheses about the groups so formed and it is only in the testing of these that probability enters. We shall discuss this more fully elsewhere.

The process has been tested by hand computation on the Beaulieu Road Community, already analysed by association analysis[7]. This contains 615 individuals specified by only 6 attributes, two of which are rare, and generates so many ambiguities by the standard similarity method that the sorting process cannot begin. The new method produces an elegant, unambiguous and informative classification. This suggests that the method may prove to be valuable, both in ecology and in taxonomy *sensu stricto*, for some cases where the existing standard methods have proved unsuitable. A programme is being prepared for the Ferranti *Pegasus* computer.

This work was supported by a grant from the Home Office; we thank Mr. L. Mockett for assistance with the computation.

W. T. WILLIAMS
M. B. DALE

Department of Botany,
University of Southampton.

P. MACNAUGHTON-SMITH

Home Office Research Unit,
Thames House South,
Millbank,
London, S.W.1.

[1] Sneath, P. H. A., *J. Gen. Microbiol.*, **17**, 201 (1957).
[2] Sneath, P. H. A., and Cowan, S. T., *J. Gen. Microbiol.*, **19**, 551 (1958).
[3] Sneath, P. H. A., and Sokal, R. R., *Nature*, **193**, 855 (1962).
[4] Dagnelie, P., *Bull. Serv. Carte Phytogeog.*, B, **5**, 7 (1960).
[5] Tanimoto, T. T., I.B.M. Corp., New York (1958).
[6] Proctor, J. R., and Kendrick, W. B., *Nature*, **197**, 716 (1963).
[7] Williams, W. T., and Lambert, J. M., *J. Ecol.*, **48**, 689 (1960).