SYMPOSIUM
ON
NUMERICAL TAXONOMY

(Friday, 21st. July, 1967).

Chairman : D. A. GRIFFITHS (U.K.).

A SURVEY OF NUMERICAL METHODS USEFUL IN TAXONOMY

BY

J. C. GOWER.

Rothamsted Experimental Station.

1. *Introduction.*

Statistical and other numerical methods have been used in taxonomy for well over fifty years but their value is only now becoming recognized by many taxonomists. The main reasons for this are that biologists find mathematical and numerical work difficult, that early numerical work in taxonomy was in specialised subjects (mainly in anthropometry) and that most of the calculations are impracticable without a computer.

Many taxonomists are still suspicious of numerical methods, particularly when computers are involved. Some of them feel that in future they will merely collect data and feed them into a large computer, which will produce irrevocable taxonomic classifications. This is not so, as the taxonomist's judgement is needed at nearly every stage in the methods described below. These methods may suggest new groups or the artificiality of some previously assigned groups, but it remains for the taxonomist to consider this in the light of all the information he has. Numerical methods should be considered as a further aid rather than superseding more traditional skills.

Because computers are now widely available, and can perform lengthy and repetitive calculations there is increasing use of known numerical methods in many branches of taxonomy, and many new methods, usually for classifying individuals into groups, have been developed. The diversity of these methods suggests that taxonomists should consider more precisely the criteria used for constructing groups.

This paper outlines the more important existing methods. Details of the

numerical techniques are omitted except where these help in understanding the principles, as these are embodied in computer programmes that are fairly readily accessible. To enable taxonomists to decide which programme to use and to evaluate the results, the important thing is to understand the principles behind the numerical methods.

When a taxonomist examines a specimen [1] he observes that it has certain characters [2] some of which other specimens lack. A table listing for every specimen the characters it does and does not have is the starting point of many numerical methods used in taxonomy. This table is often amplified by making quantitative or qualitative observations of existing characters. Thus the length of a character may be measured or its colour noted. If a character does not occur in a particular specimen then no supplementary information of this kind can be recorded. Provision must be made for dealing with missing or non-recorded observations. Observations may be missing for many reasons but no logical distinction is usually made between them ; thus characters may be non-observable (e.g. because the species is extinct and we only have fossils) or the observation may be lost, unreliable, not made, etc.

Most classical taxonomic statistical methods are designed for quantitative data only and do not easily cope with missing values. Thus these methods are useful only with closely related groups of animals, where all the characters being studied occur in some form in all the individuals of the sample. In contrast, the techniques developed by taxonomists themselves were originally designed for presence/absence characters but quantitative and multi-level qualitative data can now be used (see GOWER, 1968 ; GOODALL, 1966).

Most of the methods described here are designed to assist with one of two taxonomic problems ; either to suggest sensible groupings of individuals which the taxonomist might label as species, genera, families, etc. or given such groupings, how to assign correctly a new individual supposedly belonging to one of the groups. These will be referred to as the *grouping* problem and the *identification* problem respectively.

The identification problem is the one on which most statistical work has been done (but very much more work is needed) and the logical principles are best understood, but most taxonomists think of the grouping problem when " numerical taxonomy " is mentioned. Consequently, the grouping problem will be outlined first.

1. The words *specimen* or *individual* are used interchangeably. Depending on context, an individual may represent a whole group of individuals such as a genus. Other words used in a similar connotation are *unit*, *OTU* (original/operational taxonomic unit), *strain*, *species* and *quadrat* (in ecology).

2. The word character will be used throughout, because this is commonly used by taxonomists ; other words with a similar connotation are *feature*, *test*, *attribute*, *property*, *characteristic* and *variate*.

2. *The Grouping Problem.*

We shall start with the simplest case, in which a table of the presence and absence of characters for a sample of individuals is given. Table 1 gives an example of observations that might be made on types of fruit. The presence of the character, named at the top of each column, is denoted by + and its absence by —. The characters chosen are not particularly good because of their variability; for example not all apples are hard, but we shall assume that the fruits examined had the properties listed.

Table 1. A table giving the presence (+) or absence (—) of eight characters for seven types of fruit.

	Hard	Round	Stone	Skin or Peel	Smooth	Sweet	Stalk	Segmentable
Apple	+	+	—	+	+	+	+	—
Pear.....	+	—	—	+	+	+	+	—
Lemon.....	+	—	—	—	—	—	—	+
Orange	+	+	—	—	—	+	—	+
Grapefruit	+	+	—	—	—	—	—	+
Plum	—	—	+	+	+	+	+	—
Greengage.....	—	+	+	+	+	+	+	—

Two distinct steps are necessary to find groups; first calculate a coefficient of similarity between each pair of individuals and then put like individuals into sets and like sets into larger groups of sets and so on; this second step is known as a cluster analysis. There are many ways of doing both operations but we shall illustrate the techniques by simple methods.

One of the simplest coefficients measuring the similarity between two individuals (the simple matching coefficient) is obtained by counting the number of matches (either positive or negative) between two individuals and expressing this as the proportion of all possible matches. Thus comparing the apple and pear of Table 1, there are seven matches (5 positive and 2 negative) out of eight possible matches : a similarity of $\frac{7}{8}$. Whatever the measure of similarity the similarity between an individual and itself must always be 1 and the similarity between two individuals can never be greater than one. When they have no character state in common the similarity must be 0. The full table of similarities derived from Table 1 is given in Table 2.

Table 2. Similarities (simple matching coefficient) between each pair of fruits given in Table 1.

Name	Abbreviated name	Similarity						
Apple.....	A	1						
Pear.....	PR	$\frac{7}{8}$	1					
Lemon.....	L	$\frac{2}{8}$	$\frac{3}{8}$	1				
Orange	O	$\frac{4}{8}$	$\frac{3}{8}$	$\frac{6}{8}$	1			
Grapefruit.....	GF	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{7}{8}$	$\frac{7}{8}$	1		
Plum	PM	$\frac{5}{8}$	$\frac{6}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{0}{8}$	1	
Greengage.....	GG	$\frac{6}{8}$	$\frac{5}{8}$	$\frac{0}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{7}{8}$	1
		A	PR	L	O	GF	PM	GG

This table can be expressed in triangular form because the similarity between, for example, an apple and a grapefruit must be the same as that between a grapefruit and an apple. The table can be considered as a symmetric square matrix, with the same numbers appearing in the upper triangle as in the lower.

To illustrate the cluster analysis we shall use the single linkage method. At a similarity of $\frac{7}{8}$ apples and pears go together as do greengages and plums and also grapefruits with lemons and oranges. At this level of similarity we say that grapefruits, lemons and oranges form a group even though the similarity between lemons and oranges is only $\frac{6}{8}$; this is a characteristic of the single linkage method, where all that is required is that there is at least one set of links joining the three fruits at the prescribed level of similarity $\left(\text{here } \frac{7}{8}\right)$. The situation after sorting at the $\frac{7}{8}$ level may be written down as (A, PR) (L, O, GF) (PM, GG), the brackets enclosing fruits which have been combined into groups at this level of similarity. If we now sort at a level of similarity of $\frac{6}{8}$ some of the groups found at level $\frac{7}{8}$ may merge, because of a link of $\frac{6}{8}$ joining two or more of the groups. There are two such links between (A, PR) and (PM, GG), one joining A to GG and the

other PR to PM and in virtue of either of these links (A, PR, PM, GG) forms a group of the $\frac{6}{8}$ level. Thus, after sorting at the level $\frac{6}{8}$, we have

(A, PR, PM, GG) (L, O, GF).

The closest similarity between any two members of these two groups occurs for oranges and apples which have a similarity of $\frac{4}{8}$; consequently the two groups merge at the $\frac{4}{8}$ level.

The results of this calculation can be exhibited as a dendrogram or family tree, with the levels of sorting written in the left-hand margin (Fig. 1).

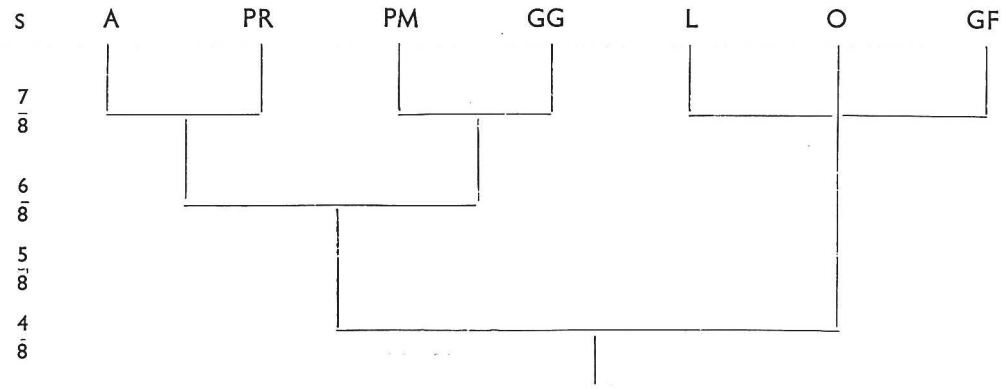


FIG. 1. — A dendrogram giving the results of a single linkage cluster analysis of Table 1.

Sensible groupings have been found, with the citrus fruits grouped separately from the rosaceous and with the rosaceous subdivided into soft fruits with a stone and hard fruits with pips.

Many points arising from this type of analysis can only be mentioned but some references for further reading are given.

It has been suggested that the similarities written down the left-hand side of dendrograms, as in Fig. 1 can be used to define families, genera, species, etc. Thus individuals occurring on different branches which join at a level of similarity below .5 might be regarded as belonging to a different genus, those joining between .5 and .8 as being different species and those joining above .8 as being subspecies or varieties. This is fallacious because the similarity levels themselves can easily be raised by including many further characters common to all individuals but irrelevant to the study ; e.g., in Table 1 we might add characters " does it grow on a tree, " " is it sold in shops, " " can it sneeze ". The effect of adding constant characters is to make all the similarities bigger ; the general form of the dendrogram will be unaltered but greater similarity levels must be written down the side. Another reason for rejecting any attempt to give meaning to absolute values of the similarities is that there are many different ways of calculating similarity

(see SOKAL and SNEATH (1963) for a discussion of many different coefficients and GOWER (1968) for some more). Different coefficients will not only give different values but two different types of coefficient need not be monotonically related; that is to say if A, B and C are three individuals and S_{AB} is the similarity between A and B as measured by one method and S'_{AB} by another, then if S_{AB} is greater than S_{AC} it is not necessarily true that S'_{AB} is greater than S'_{AC}. Clearly different coefficients may sort the individuals out differently even when the same type of cluster analysis is used. There are also very many different types of cluster analysis which will in general give different results. Some of these are discussed in Section 3.

The single linkage method has been explained in terms of grouping individuals with high similarities, it may also be explained in terms of distance. Pairs of individuals with a high similarity may be regarded as closer together than pairs with low similarities. It is not necessary in the single linkage method to know the actual distances so long as we know the ranking of these distances. We can then imagine a diagram as in Fig. 2.

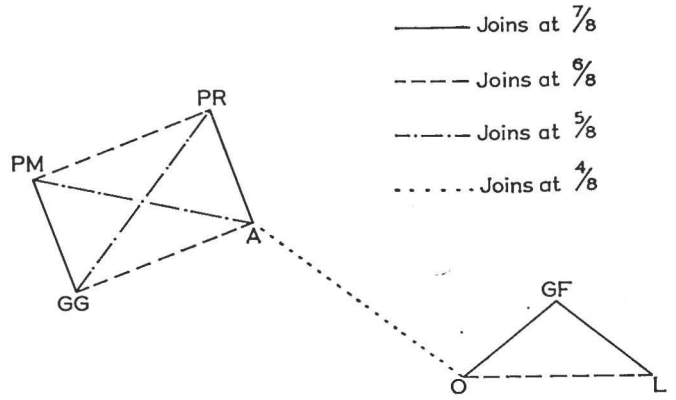


FIG. 2. — An approximate geometrical representation of the similarities of Table 2 in terms of distance.

Thus the three clusters formed at a level of $\frac{7}{8}$ similarity (PM, GG) (PR, A) (O, L, GF) are represented by the solid joining lines, and the first two of these groups merge at $\frac{6}{8}$ similarity because of either one of the two lines PM, PR or GG, A. At $\frac{5}{8}$ no further joining occurs but the group (A, PR, PM, GG) becomes more tightly bound as in fact all possible joins have now occurred between members of this group. The final join is represented by the line A, O at $\frac{4}{8}$ similarity.

The actual coordinate values for the individuals as represented in diagrams like Fig. 2 can be found but it is first necessary to have a table of all the distances, rather like the table of similarities. Again there are many possibilities that pre-

serve the ranking (e.g. each similarity S can be transformed to a distance $1 - S$, $\sqrt{1 - S}$, or $-\log S$).

Another problem in numerical classification is how to select characters to include. If in a classification for a special purpose the characters relevant to that purpose can be identified, there is no problem. If the methods can lay any claim to generality, similar classifications should be obtained when different sets of characters are used, provided the character sets have not been specially selected to bias the classification in a particular direction. This requires that adding new characters will not upset previous classifications. Clearly, if a stable classification exists, information on only one character cannot be expected to find it, so the question then arises as to how many characters must be observed before a stable classification can be achieved. In terms of distance we require that, although the distances between individuals may differ with different choices of characters or when additional characters are included, the relative distances must not be unduly changed. SOKAL and SNEATH (1963) discuss these problems from the taxonomic point of view but a statistical treatment is still required.

In practice the taxonomist's judgement is accepted and it is usual to include in the analysis all the characters he has examined giving them all equal wieght when calculating any similarity coefficient or distance. The decision to weight or not to weight character scores has been a controversial problem for taxonomists ; in general those in favour of using numerical methods prefer not to weight, but the traditional taxonomist holds that taxonomies have always been constructed by recognizing that certain characters are more important than others. At least part of the difficulty seems to come from the fact that with a new set of organisms completely unrelated to any known group, no *a priori* weighting would be acceptable, but once this set has been classified, it becomes clear that certain characters are more suitable than others for constructing an identification key. In any subsequent reclassification, these characters might be regarded as more important and might therefore reasonably be assigned greater weight.

There is no mathematical difficulty in weighting characters unequally but it is difficult to assign differential weights on *a priori* grounds or to do so sensibly from an examination of the data. Several different forms of weighting are possible, e.g., simple weighting of a character or weighting the result of comparing the same character in two individuals — if it agrees it gets one weight, otherwise it gets another. GOWER (1968) discusses this aspect at greater length.

Taxonomists are often advised to include as many characters as possible in their analysis. The rationale here is the notion that there is some " true " value of similarity between pairs of individuals that can be measured only when all the characters are observed but is accurately estimated when enough characters are observed. However, some characters may not contain any useful classificatory information (e.g. fur length in distinguishing cats from dogs). The values of irrelevant characters vary from individual to individual but are not correlated with the values of other characters ; the converse is not true because characters uncorre-

lated in the sample as a whole may be correlated within sub groups determined by a cluster analysis. If there are not too many irrelevant characters, they will contribute little to any similarity coefficient but otherwise they may be the major factor involved. When they are, sensible classification is likely to be difficult unless the irrelevant characters can be eliminated. How this should be done is not known, but it is probably unwise for a taxonomist to include characters he suspects of being irrelevant, just to increase the number of characters in the analysis. The type of subjective judgment involved here is the same as that required to decide what characters should be included in the study and amounts to an *a priori* character weighting (weight 1 = include ; weight 0 = exclude). This subject needs much closer examination than it has so far received.

3. *Other Methods of Cluster Analysis.*

SOKAL and SNEATH (1963), BALL (1965) and ANDERSON (1966) discuss many forms of cluster analysis in detail. Cluster analyses may be agglomerative or divisive, polythetic or monothetic. Single linkage cluster analysis is agglomerative and polythetic. It is agglomerative because the groups determined at any level of sorting are obtained by combining individuals or previously determined groups and polythetic, because groups need not, and usually will not, have constant characters.

WILLIAMS and LAMBERT's (1959) association analysis and the scheme outlined by EDWARDS and CAVALLI-SFORZA (1965) are both examples of divisive methods because groups are repeatedly subdivided. The first is monothetic because groups are divided into two on a single character, one group consisting of all the individuals possessing the chosen character and the other all individuals without it ; the second is polythetic. GOWER (1967 a) discussed these superficially dissimilar methods in terms of distance and showed that they can all be interpreted as grouping individuals, so that the distance between groups (suitably defined) is maximised.

Agglomerative polythetic methods are often preferred when trying to determine groups and divisive monothetic ones for constructing keys (see Section 5 below).

When the individuals fall into well defined groups all methods can be expected to give similar results, but when boundaries between groups are less distinct, different methods are likely to assign borderline individuals differently.

It is in precisely this situation that most taxonomic wrangles occur. The numerical approach merely reflects that intermediate forms are difficult to classify and unless there are overwhelming reasons for accepting one numerical method in preference to all others, no definite method for assigning borderline cases can be given. Thus numerical techniques are unlikely to give a final " correct " answer to the classification of closely related individuals, although diagrams like Fig. 2 do give a general view of the problem and reveal the futility of many arguments between taxonomists.

There is a danger of misinterpreting the results of a cluster analysis because there is no theoretical and little empirical knowledge concerning how the analysis behaves with random data. All clustering methods will produce some sort of hierarchical classification along the lines of Fig. 1 even when the data do not warrant it ; examination of diagrams like Fig. 2 help to expose this type of situation. Some experimental work has been done using dummy data. Cluster analysis of data, with presence and absence of characters being assigned at random with varying frequencies, has given results comparable to the analysis of some real data.

A property of single linkage cluster analysis that some taxonomists find objectionable is that a cluster of individuals may occur on long chains and not in compact groups. The first and last members of such a chain may be very unalike although a chain of individuals links them at some prescribed level of similarity. The situation is shown in Fig. 3.

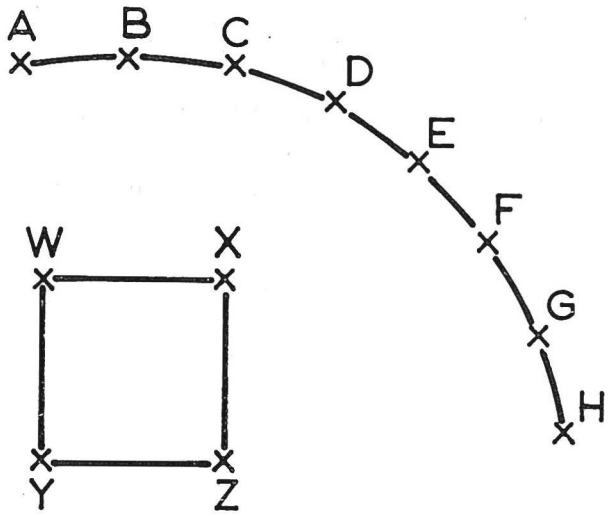


FIG. 3. — (A, B, C, D, E, F, G, H) form a group on single linkage analysis whilst W, X, Y, Z do not join until a lower level of similarity.

SOKAL and MICHENER (1958) suggested a clustering method without this property and which will be used below to illustrate another method of agglomerative polythetic cluster analysis. The method exists in two forms, the unweighted mean pair group (UMPG) and the weighted mean pair group (WMPG), and they are most simply explained geometrically, in terms of distance. The method explained here is not exactly SOKAL and MICHENER's original method but is very similar (GOWER, 1967 a). In UMPG, replace the two nearest neighbours by a single combined member at their centroid (centre of gravity). Keep on repeating this process combining individuals or previously combined groups of individuals. At any stage a combined group of individuals is represented by the centroid of the original positions of the individuals. An objection to this method is that, if the original sample

comprised several members each of several species, when groups representing two species are combined the position of the resulting centroid depends on the number of members of each species ; i.e., on their relative abundance. Taxonomists do not usually want their classifications to depend on abundance, so the alternative WMPG method was devised. This differs from the UMPG method in that when two points (representing individuals or previous combinations of individuals) are joined they are replaced by a single point at the centre of the joining line, irrespective of the number of individuals represented by each point.

A minor disadvantage of these methods is that, when a point A is joined to a point B, the point replacing them may be closer to some other point than A was to B. This property of the methods is very unlikely to give trouble but it may mean that it is not always possible to write consistant similarity levels against each branch of the dendrogram as was possible in Fig. 1.

Both these forms of analysis used on the data of Fig. 2 give substantially the same results as a single linkage cluster analysis. The only difference is that (O, GF) or (L, GF) join arbitrarily at a higher level of similarity than the remaining citrus fruit.

The possible dangers discussed above, arising from the relative abundance of different species are also relevant to some coefficients of similarity. In Table 1, the proportion of times each character occurs can be calculated, thus hardness occurs $\frac{5}{7}$ times and roundness $\frac{4}{8}$ times, etc. These can be interpreted as estimates of probabilities of the occurrence of each character. Coefficients of similarity using these probabilities have been proposed (see e.g. GOODALL (1966) and MCNAUGHTON-SMITH (1965)). The danger of using such probabilities can be seen when we consider a sample consisting of only two species and we consider a character that is always + for one species and always — for the other. The estimated probability for this character will then be an estimate of the relative abundance of the first species ; any value between 0 and 1 can be obtained by adjusting the sample sizes. There may be situations when this type of coefficient is useful but the problems arising from abundance need more attention ; GOWER (1967 a) showed that there was implicit weighting from abundance in WILLIAMS and LAMBERT'S association analysis and also in EDWARDS and CAVALLI-SFORZA'S (1965) cluster analysis method.

4. *Distance and Similarity.*

In the discussion so far, the complementary concepts of distance and similarity were used. In statistical writing the distance concept has been most used, whereas most taxonomists favour similarities. As was shown above it is easy to convert a similarity into a distance but because similarities are restricted to the range 0, 1, derived distances are sometimes also restricted (e.g. 1 — Similarity is restricted to the range between 0 and 1 but — log (Similarity) is not).

The simplest distance to use with quantitative characters is obtained by regarding the character values as coordinate values referred to a set of rectangular axes, and using an extension of Pythagoras's theorem. Table 3 gives values for four quantitative characters for each of two imaginary individuals.

Table 3. The values of four quantitative characters for two individuals.

	Length (mm's)	Breadth (mm's)	Height (mm's)	Weight (Grm's)
Individual 1	10	1.5	3	21
Individual 2	12	1.3	4	24

The distance $/d_{12}$ between the two individuals is calculated as

$$d^2_{12} = (10\text{-}12)^2 + (1.5\text{-}1.3)^2 + (3\text{-}4)^2 + (21\text{-}24)^2$$
$$= 14.04$$

Thus $d_{12} = 3.75$. An obvious objection is that this distance depends on the units of measurements of each character. Thus if length were measured in centimeters and not millimeters the value of d^2_{12} would be 10.08. To avoid this it is usual to express every measurement in some standardized non-dimensional unit. This is usually achieved by dividing each measurement by its standard deviation calculated from all the values observed in the sample. This standardized measure of taxonomic distance, discussed in detail by SOKAL (1961), has been used for over 50 years in statistical writings. Clearly if the two individuals had exactly the same measurements, $d_{12} = 0$ and the more they differ in their measurements the greater the distance becomes. This measure of distance is just as arbitrary as the choice of similarity coefficient ; there is no special reason, apart from convenience, for choosing rectangular axes as a coordinate framework for the character values and the method of normalizing is quite arbitrary. GOWER (1966, 1967 b) discusses these matters further.

When there are many individuals in the sample, a triangular table, similar to that given in Table 2 can be calculated, whose elements are the distances between the individuals. The distances can be evaluated as above or by transforming a table of similarities S or obtained by some direct experimental means. The coordinates of points giving rise to these distances can be found, so that diagrams similar to Fig. 2 can be drawn. These diagrams are helpful in examining interrelationships between closely related individuals, which do not admit hierarchical classification. It will only very rarely be possible to represent the given distances exactly in two or three dimensions. In the particular case of the distance defined above, the coordinates of the individuals can be taken as the actual observed values of the characters and this needs as many dimensions as there are different characters

To economize in the number of dimensions, mathematical techniques have been devised for finding the best representation of the distances in a few dimensions. Details of these methods can be found in GOWER (1966) and examples of this type of analysis using data on mites were given by SHEALS (1964) and by SHEALS (1969). If the distance defined by Pythagoras's theorem is used, the analysis is called a Principal Components Analysis and has some simplifying features ; if any other distance is used I have termed it a Principal Coordinates Analysis. In either, a set of coordinates is found with the desired distances and the coordinate axes are ordered so that the coordinates referred to the first r axes give the best representation of these distances (in a least squares sense) using r dimensions. In principal components analysis these new axes can be related to those originally used to represent the sample, and this property can occasionally be used to give biological meaning to the new axes.

The interpretation of principal axes, is again complicated by the relative proportions of two or more biological populations that may be in a sample. When two characters are distributed statistically in multivariate normal form, then a scatter diagram of these two characters will have approximately elliptical shape, and similarly for more than two characters. Thus either of the ellipses in Fig. 4 could have arisen in this way. The principal axis of either of these two populations is the major axis of the corresponding ellipse. With more samples, the ellipses can be expected to retain the same orientation so that the principal axes will remain fixed (within sampling variation) and are thus a reproducible property of the biological populations that may also have biological meaning. When the combined data are now analyzed, the principal component depends on the distance between the two centroids and the relative sizes of the two populations. If this distance is great compared with the within population scatter, the principal component will be approximately in the direction of the line joining the centroids ; if small it will be approximately the resultant of the two separate principal components weighted by the population sizes, and is unlikely to have any useful meaning.

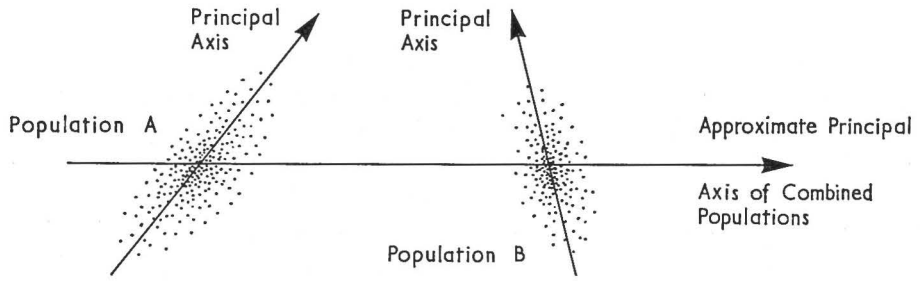


FIG. 4. — A scatter diagram arising from a sample of two characters observed for two biological populations A and B (representing say different species). The areas enclosed will be elliptical when the two characters are normally distributed and the directions of the principal axes of each population are given by the major axes of the ellipses.

When all the characters are measured in identical units it is not necessary to standardize them in defining a distance. An interesting special case of this is when all characters are of the presence/absence type with 1 coded for presence and 0 for absence. If S_{12} is the simple matching coefficient between individuals one and two then $d^2_{12} = n(1 - S_{12})$ when there are n characters in all. A principal components analysis of this type of data is therefore equivalent to choosing a distance $\sqrt{1 - S}$ between pairs of individuals.

In proximity analysis (developed by R. N. SHEPARD) a distance matrix is again the starting point and a low dimensional representation of these distances is sought. However, it sometimes happens that the distances are only very roughly known and, although individuals close together are probably more similar than distant pairs of individuals, the absolute values of these distances cannot be relied on strongly. It is then reasonable to seek a low dimensional representation that preserves, so far as possible, the monotonic relationships among the original distances. SHEPARD (1962a, b) and KRUSKAL (1964a, b) show how this can be done.

Factor analysis is sometimes used by taxonomists, but I do not think it relevant to taxonomic purposes. GOWER (1966) discussed the reasons further and showed that it is likely to give very similar results to the simpler principal components analysis.

5. *The Identification Problem.*

In the grouping problem individuals are grouped into species *because* they have many characters in common ; distances used to solve such problems should take this into account. The more characters there are in common, the stranger the conviction that a " natural " grouping has been found. These common characters may of course be highly correlated in the sample but this is not necessarily so when, for example, the characters have low correlation for some species and high correlation for others. It is sometimes said that correlations should be eliminated when defining distances suitable for the grouping problem, but this seems to be because of confusion with the needs of the identification problem discussed below. An example illustrates the point. Suppose we have a sample of different species of cats and dogs, although we do not as yet know it. Two characters might be :

(1) Does the animal possess claws it can extend ?
(2) Does the animal have a third eyelid which moves transversely across its eye ?

These two characters will be completely correlated in the sample and are two of the characters that lead us to distinguish cats from dogs. If, because of the perfect correlation, we eliminate one of the two characters, we are obscuring this difference. In general when all such characters, except one, are eliminated we are left mainly with those characters that tell us nothing about the differences

between cats and dogs (e.g., length of fur) and the sample would seem homogeneous.

The distances discussed in Section 4 are suitable for the grouping problem but the identification problem is quite different because the different groups are already defined. This may be done by describing single type specimens or by reference to samples of individuals that reflect the natural variation within the group. Species groupings may have been suggested by a cluster analysis of the type discussed earlier. Correlations between characters must now be eliminated, because if two characters are completely correlated within a group, a knowledge of the value of both of them does not add any more information for identification than is available from either one.

Statisticians tend to think of the variable characters of species, particularly quantitative measurements, and have provided techniques to identify sample individuals assumed to belong to one of two or more known populations ; this is known as *discriminant analysis* ; taxonomists have been more interested in constructing keys based on qualitative characters.

The basis of discriminant analysis is to imagine the frequency distributions of each population and to assign individuals so that the probability of their misclassification is minimum. Clearly the more overlap there is between populations the more likelihood of misclassification. Discriminant analysis is therefore most useful in those situations where there are no clear cut distinctions between populations and is mainly relevant to the identification of closely related species or subspecies ; for example, it proved useful in the description of island races of mice and shrews, (DELANY & HEALY, 1964, 1966). The techniques of discriminant analysis have been little examined except when within species variation is multivariate normal (possibly after transformation), preferably with the same dispersion matrix for every species. Under these conditions the boundaries used to discriminate between species are linear combinations of the variate values. The probability of misclassification can be interpreted in terms of a distance D which, when squared, is known as Mahalanobis's D^2-statistic ; D is the distance between the two population means after eliminating the effects of correlations. When there are several populations, a table like Table 2 can be drawn up in which an entry is the value of D^2 between the two populations represented by the row and column in which the entry occurs. The principle coordinates technique can be used to draw diagrams, like that of Fig. 2 in which each point represents a population mean. The variation of each species about its mean can be indicated by drawing circles with the mean as centre. The wider the radius then the greater proportion of the population can be expected to lie within the circle ; it is usual to choose the radius so that the circle contains 90-95 % of the population. Although D^2 is associated with the identification problem, tables of D^2 values may be useful in the grouping problem when it is desired to group the species into higher aggregates such as genera.

The value of D^2 can also be evaluated between a single sample individual and

a population. This gives a simple interpretation to discriminant analysis, because if there is a single sample to be assigned to one of several populations the values of D giving the distance of the sample from each population can be evaluated. The nearest population is the best one to assign the sample to. The probability of getting a value of D as large as the one observed (or equivalently whether the sample point lies within the circles discussed above) can be calculated to see if it is reasonable to regard the sample as belonging to any of the populations. This use of D^2 is valid even when the populations have different dispersion matrices and requires only that characters are distributed symmetrically (possibly after transformation). See COOPER (1963, 1965) for more information on this topic.

When the distribution of the character values do not fall into any of the classes discussed above, practical techniques for discriminant analysis do not exist; although the mathematical theory is known (RAO, 1952) the numerical calculations become unwieldy and difficult to specify. A more serious practical difficulty is that often too little is known about the qualitative and quantitative characters available to set up any acceptable multivariate distribution that they can be supposed to follow.

Quantitative characters may be altered by growth and the sensitivity of discrimination reduced. DELANY and HEALY (1964) showed how the growth of young animals into larger adults can be allowed for using standard statistical regression techniques and BURNABY (1966) discussed how to combine the necessary adjustments into the standard discriminant analysis. There are further difficulties when some characters grow at different rates from others; HOPKINS (1966) discussed this problem and gives further references.

The uses of discriminant analysis were surveyed by BARTLETT (1965), who gives examples of discrimination between populations with different dispersion matrices and possibly with the same mean. BARTLETT also discusses Penrose's method for separating size and shape components and Fisher's method for assessing optimum scores for the different levels of a qualitative character. This latter is useful in many taxonomic situations, but so far as I am aware, has never been used by taxonomists.

When qualitative characters are used there is no question of overlap between the species and identification, strictly speaking, is possible only when complete agreement between the sample and type specimen exists, although even when agreement is not exact, it is still possible to ask what species the sample most resembles.

The question usually asked by taxonomists is how can an identification key be devised so that individuals can be identified with the fewest steps. I know of no general algorithm to answer this question, but many monothetic grouping methods are clearly influenced by the design of keys. GOWER (1967 a) suggested that WILLIAMS and LAMBERT's (1959) monothetic " association analysis " may be better for determining a key than for assigning groups. At any stage the value of one particular character is required and this diminishes the number of possible

species to which the sample can belong ; further steps of this process eventually narrow the identification down to a single species.

The following points need to be kept in mind when trying to devise a key, whether automatically or by trial and error :

(1) If a mistake is made or the individual being identified is aberrant in some way, the wrong branch of the hierarchical key will be followed and identification will be impossible. Cross referencing between one branch and another can be used to overcome the danger.

(2) It is difficult to include quantitative characters in a key, especially when there is overlap between species.

(3) Quicker identification might be obtained by considering the characters in conjunction with one another and not just in isolation.

(4) It might sometimes be better not to use the most direct way of identifying an individual, but to recognize that some characters are more easily or more cheaply observed than others and are therefore more economic to use.

Although taxonomists argue about the advisability of using numerical methods to determine groups, there seems less reason to disapprove of any attempt to mechanize the identification process. This involves both the determination of the best set of rules to follow when identifying an individual and the process of using these rules for identification. The closely related problem of (medical) diagnosis has already been studied intensively.

6. *Phylogenetic Background.*

Classifications found by the methods discussed earlier are based on general resemblance between pairs of individuals and therefore cannot claim any evolutionary basis. Because of the biological phenomena of mimicry and evolutionary convergence, a morphological classification may be quite different from a phylogenetic one. Nevertheless, when enough characters are taken, some morphological, some biochemical, some histological, etc., those characters that are similar from convergence or mimicry, might be expected to be far outweighed by the others and so the phenetic classification might be accepted as being at least an approximation to a phylogenetic one.

Most taxonomists would rightly view such an approach with suspicion and would want to base their phylogenetic classifications on any fossil record and collate their results with geological information. If fossils are pheneticly classified there is no guarantee that accepted evolutionary principles will not be broken. For example, it may be found that a classification is suggested in which C is supposed to have evolved from B and B from A ; this might involve one or more characters in changing from a primitive state to a specialised one and back again. To avoid this sort of difficulty, Camin and Sokal (1965) have suggested that a

phylogenetic classification might be constructed directly from the table of character states. Each character state is given an integer number representing its supposed stage of development from the most primitive (0). To do this, a fair amount must be known about the development of the group. The next step is to fit an evolutionary tree to the character states without violating a set of rules governing the type of step allowed in an evolutionary change of character. Usually, many possible trees can be fitted and the criterion taken is to select the tree with the fewest evolutionary steps. CAMIN and SOKAL (1965) were not able to give an exact logarithm for finding this but suggested several approximate methods. They found their methods useful and have for example successfully constructed phylogenies for the fossil horses. They claim also that their methods reveal characters that have been wrongly coded and have shown empirically (for the fossil horses, for example) that presumed ancestral forms can be reconstructed within reasonable bounds when doubtful characters are omitted from the analysis.

This type of analysis seems promising and is more in agreement with the classical taxonomists' approach than other forms of cluster analysis. The rules governing the evolutionary steps allowed may be a source of disagreement, but one advantage of this general approach is that the effects of choosing different sets of rules can be studied. For example, CAMIN and SOKAL (1965) appear to allow a character to take the same step on different branches of phylogeny ; this permits the same evolutionary change to occur independently on more than one occasion and might not be allowed by other taxonomists.

A branch of mathematics that may be useful in the construction of trees with the minimum number of steps is that part of " mathematical programming " that deals with finding the shortest path through a set of points. A typical problem of this kind is to find an optimum route from a point A_1 to a point A_n given a map of a set of points $A_1,A_2,...,A_n$ which has some (but not necessarily all) of the joins A_iA_j and their lengths. In its simplest form the shortest route from A_1 to A_n is required but it might also be stipulated that all the points A_i be visited at least once (this is the travelling salesman problem) and further we might require that A_1 is the same as A_n (i.e., we must return to the starting point). An example of an algorithm for a problem of this sort can be found in NICHOLSON (1966), which gives further references. BOOTHROYD (1967) has published Algol programs for Nicholson's and related methods.

In the present context a map might be drawn where each point A_i represents an individual. It will not in general be possible to go from A_i to A_j by a set of forward evolutionary steps but there is always a hypothetical common ancestor A. Thus we consider a map with the points A_i and a hierarchy of common ancestors from which we select the minimum spanning tree containing all the original individuals.

Conclusion.

This brief survey has outlined the main types of numerical methods useful to taxonomists. It is clear that much remains to be done and that much that is already done is far from perfect. This is no reason to reject the whole structure out of hand, rather it demonstrates a need for discussion and constructive criticism, so that any deficiencies are clearly understood and so far as possible remedied. I suspect that one of the most important results of the increased use of numerical methods in taxonomy has been the re-examination of basic principles.

Summary.

The principles underlying some numerical methods useful in taxonomy are described. The mathematical treatment is elementary and is intended to give an introduction to the subject that can be supplemented by further reading. Classification, identification and the construction of phylogenies are discussed. As a first step towards improving the methods, attention is drawn to some of their deficiencies. Only by understanding the principles can taxonomists judge the usefulness, or otherwise, of numerical methods.

Acknowledgment.

I thank Dr. J. H. RAYNER for reading this paper during my absence.

REFERENCES

ANDERSON (A. J. B.), 1966. — A review of some recent developments in numerical taxonomy. — M.Sc. Diss. ; U. of Aberdeen.

BALL (G. H.), 1965. — Data analysis in social sciences. What about details ? — Proc. Fall Joint Computer Conference 1965, Stanford Research Institute, Manlo Park, pp. 533-559.

BURNABY (T. P.), 1966. — Growth-invariant discriminant functions and generalized distances. — Biometrics, **22** : 96-110.

BARTLETT (M. S.), 1965. — Multivariate statistics, being Chapter 8 of Theoretical and Mathematical Biology, edited by Waterman & Morowitz. Blaisdell Publishing Co.

BOOTHROYD (J.), 1967. — Algorithms 22, 23 and 24. — Comp. J., **10** : 306-308.

CAMIN (J. M.) & SOKAL (R. S.), 1965. — A method for deducing branching sequences in phylogeny. — Evolution **19** : 311-326.

COOPER (P. W.), 1963. — Statistical classification with quadratic forms. — Biometrika, **50** : 439-448.

COOPER (P. W.), 1965. — Quadratic discriminant functions in pattern recognition. — IEEE Trans. on Information IT **11** : 313-315.

DELANY (M. J.) & HEALY (M. J. R.), 1964. — Variation in the long-tailed field mouse (*Apodemus Sylvaticus* (L.), in North-West Scotland. II. Simultaneous examination of all characters. — Proc. roy. Soc., B, **161** : 200-207.

DELANY (M. J.) & HEALY (M. J. R.), 1966. — Variation in the white-toothed shrews (*Crocidura* spp.) in the British Isles. — Proc. roy. Soc., B, **164** : 63-74.

EDWARDS (A. W. F.) & CAVALLI-SFORZA (L. L.), 1965. — A method for cluster analysis. — Biometrics, **21** : 362-375.

GOODALL (D. W.), 1966. — A new similarity index based on probability. — Biometrics, **22** : 882-907.

GOWER (J. C.), 1966. — Some distance properties of latent root and vector methods used in multivariate analysis. — Biometrika, **53** : 325-338.

GOWER (J. C.), 1967a. — A comparison between some methods of cluster analysis. — Biometrics, **23** : 623-637.

GOWER (J. C.), 1967b. — Multivariate analysis and multidimensional geometry. — The Statistician, **17** : 13-28.

GOWER (J. C.), 1968. — A general coefficient of similarity and some of its properties. (Submitted to Biometrics).

HOPKINS (J. W.), 1966. — Some co-reductions in multivariate allometry. — Biometrics, **22** : 747-760.

KRUSKAL (J. B.), 1964a. — Multidimensional scaling by optimising goodness of fit to a non-metric hypothesis. — Psychometrika, **29** : 1-27.

KRUSKAL (J. B.), 1964b. — Non-metric multidimensional scaling : a numerical method. — Psychometrika, **29** : 115-129.

MCNAUGHTON-SMITH (P.), 1965. — Some statistical and other numerical techniques for classifying individuals. — A Home Office Research Unit Report, No. 6. H.M.S.O. London.

NICHOLSON (T. A. J.), 1966. — Finding the shortest route between two points in a network. — Comp. J., **9** : 275-280.

RAO (C. R.), 1952. — Advanced statistical methods in biometrical research. — J. Wiley — New York.

SHEALS (J. G.), 1964. — The application of computer techniques to acarine taxonomy : a preliminary examination with species of the *Hypoaspis-Androlaelaps* complex (acarina). — Proc. Linn. Soc., Lond., **175** : 11-21.

SHEALS (J. G.), 1969. — Computers in acarine taxonomy. — Acarologia, **11** (3) : 376-394.

SHEPARD (R. N.), 1962a. — The analysis of proximities : multidimensional scaling with an unknown distance function. I. — Psychometrika, **27** : 125-139.

SHEPARD (R. N.), 1962b. — The analysis of proximities. Multidimensional scaling with an unknown distance function. II. — Psychometrika, **27** : 219-246.

SOKAL (R. R.), 1961. — Distance as a measure of taxonomic similarity. — Syst. Zool., **10** : 70-79.

SOKAL (R. R.) & MICHENER (C. D.), 1958. — A statistical method for evaluating systematic relationships. — Univ. Kansas Sci. Bull., **38** : 1409-1438.

SOKAL (R. R.) & SNEATH (P. H.), 1963. — The principles of numerical taxonomy. — W. H. Freeman, San Francisco and London.

WILLIAMS (W. T.) & LAMBERT (J. M.), 1959. — Multivariate methods in plant ecology. I. — J. Ecol., **47** : 83.