Multivariate Methods in Plant Ecology: V. Similarity Analyses and Information-Analysis

W. T. Williams; J. M. Lambert

The Journal of Ecology, Vol. 54, No. 2. (Jul., 1966), pp. 427-445.

MULTIVARIATE METHODS IN PLANT ECOLOGY

V. SIMILARITY ANALYSES AND INFORMATION-ANALYSIS

BY W. T. WILLIAMS* AND J. M. LAMBERT

Botany Department, University of Southampton

G. N. LANCE

C.S.I.R.O. Computing Research Section, Canberra, Australia

CONTENTS

INTRODUCTION

In the current controversy over the relative merits of classification and ordination in vegetational analysis, it has been argued elsewhere (Lambert & Dale 1964) that the initial choice rests more on the convenience of the user than on preconceptions as to the continuity or discontinuity of the vegetation: if the prime requirement is to produce vegetational units which can be used for mapping or description, then classificatory methods are more applicable. In circumstances where classification is desired, there is a further user choice in the overall type of classification to be adopted, namely between hierarchical and non-hierarchical (i.e. reticulate) systems. It has been pointed out (Lance & Williams 1966) that hierarchical methods seek to subdivide the population progressively by the most efficient steps, while non-hierarchical methods—such as the many variants of cluster analysis—are aimed at the erection of efficient groupings irrespective of the route by which they are obtained; and, since no method is yet available which simultaneously maximizes hierarchical efficiency and group homogeneity, the user must decide whether to optimize the groupings or the route. In general, hierarchical methods are better known, less cumbersome, and more widely used in ecological work than direct clustering techniques, and—without prejudice as to the possible value of the latter under certain requirements—we have so far concentrated on the hierarchical

* Now at the C.S.I.R.O. Computing Research Section, Canberra, Australia.

approach. The first four papers of this series* (Williams & Lambert 1959, 1960, 1961; Lambert & Williams 1962) were concerned with a particular hierarchical method known as association-analysis, with the ecological results assessed over a number of test-communities. The present paper and subsequent communication will examine critically an alternative set of of hierarchical methods known collectively as 'similarity analyses', and compare the strategy and results of the most effective of these methods with those of association-analysis.

II. GENERAL CONSIDERATIONS

Hierarchical methods of classifying elements into sets are subject to two independent choices. First, the strategy may be *divisive*, in that the population is progressively subdivided into groups of diminishing size, or *agglomerative*, in that individuals are progressively fused into groups of increasing size until the entire population is synthesized. Secondly, the strategy may be *monothetic*, every group at every stage (except the entire population) being definable by the presence or lack of specified attributes, or *polythetic*, the groups being defined by their general overall similarity of attribute structure. Of the four systems so generated, agglomerative monothetic methods cannot exist, except in a trivial sense. Furthermore, of the two existing divisive polythetic methods, one (Edwards & Cavalli-Sforza 1965) is computationally out of reach for all except very small populations, and the other (Macnaughton-Smith *et al.* 1964) is not yet sufficiently developed for application to ecological problems. In practice, therefore, the choice at present lies between divisive-monothetic and agglomerative-polythetic.

Agglomerative-polythetic methods (i.e. similarity methods) are historically the older, deriving at least from the work of Kulczynski (1927); in their less-developed forms they are also simpler, and even amenable to hand-calculation. It is not, therefore, surprising that a number of variants, many of them frankly inefficient, already exist in the literature; a good general review is presented by Sokal & Sneath (1963). We can conveniently think of the overall strategy as determined by two choices. The first of these concerns the measure of similarity to be employed. The number of coefficients which at one time or another have been suggested is legion; it would be completely impracticable, and almost certainly unprofitable, to attempt to compare them all. Criteria have been suggested (Williams & Dale 1965) which may be used as a guide in the selection of coefficients for further study, and our present selection has been based on these recommendations. The second choice concerns the precise strategy to be used in making the successive fusions. Here again the alternatives have been discussed elsewhere (Williams & Dale 1965) and two have been selected for comparison.

A difficulty in the past has been that, owing to the restricted computer facilities hitherto available, any one worker has commonly had access to a single sorting method and often a single coefficient; there has therefore been virtually no opportunity for the comparative assessment of different coefficients and different fusion strategies on the basis of the results obtained from a single set of data. With improved computing facilities it is now easier, and more urgently necessary, to undertake comparative studies. Our results have been obtained by the use of the 'flexible similarity programmes' QUALNEAR and CENTROID on the Control Data 3600 computer in the C.S.I.R.O. Computing Research Section at Canberra; these programmes have already been briefly announced in a communication (Williams & Lance 1965) primarily concerned with the inferential problems presented by all intrinsic numerical classifications.

* For brevity, we shall subsequently cite these papers as Papers I–IV.

III. COMPARISON OF SIMILARITY METHODS

A. *Theoretical*

1. *The fusion strategies*

It will be convenient to consider the fusion strategies—usually termed 'sorting methods'—first. A number of alternatives have been suggested in different contexts, and the two selected for detailed study are described below. However, preliminary consideration must be given to two problems which may arise in all agglomerative systems, and which are related to the same cause—namely, that a hierarchical strategy involves irrevocable fusions; a 'bad' fusion early in the analysis is thus in principle capable of directing the subsequent fusions along an unprofitable path.

The first problem is that the data may be subject to appreciable error or bias in the sampling. Undesirable results can then be minimized by using a coefficient incorporating information from the population as a whole, such as the 'objectively weighted squared Euclidean distance' of Williams, Dale & McNaughton-Smith (1964); but in a complex situation the resulting dependence on overall population structure may destroy local concentrations of interest. A better solution would be to carry out a duplicated analysis and compare the results. In practice, it must be admitted that the usual procedure is simply to ignore this problem, but it is important that its existence be realized.

The second problem is simply that, unless the number of attributes is very large, ambiguities may be encountered. This can only be resolved by appeal to a different coefficient—which may not improve the situation—or, again, by importing information from the population as a whole or from outside the data. In our present study this problem, too, has been ignored; to avoid computational difficulties in the pilot programmes, the first of a set of ambiguities encountered has been used for action.

We proceed to outline our two chosen strategies:

(a) '*Nearest-neighbour*' *or* '*single-link*' *sorting*. This is the simplest of all agglomerative procedures. The process begins with the calculation of a similarity coefficient between all pairs of individuals; these coefficients, tagged with the numbers of the individuals concerned, are then sorted into a linear order with the most-similar pair at the beginning. All coefficients are then examined in turn and subjected to the following strategy:

(i) If neither of the members of the next most similar pair is already in a group (this is always true at the start of an analysis) designate them as forming a new group.

(ii) If one is in an existing group, add the other to the group.

(iii) If both are in different groups, add the groups.

(iv) If both are in the same group, discard.

The process continues until all individuals are fused into a single group. If there are n individuals, $\frac{1}{2}n(n-1)$ coefficients are calculated at the start, but there is no further calculation. The method is in theory lacking in power, since the structure of the groups as they form is not itself used in calculation; the analysis never rises above the information-level of a single individual.

(b) '*Centroid*' *sorting*. This process also begins with the calculation of all $2n(n-1)$ similarity coefficients, which are tagged but not in this case sorted into order. The strategy is then as follows:

(i) The most similar pair of individuals are *added together*, attribute by attribute, to form a new synthetic individual which is allotted the next available serial number.

(ii) The records of the individual members of the pair concerned are deleted, together with all coefficients involving either of them.

(iii) Coefficients are calculated between the new individual and all other remaining individuals; the process then returns to operation (i).

When all individuals have been fused into a single group, $(n-1)^2$ coefficients will have been calculated. The theoretical advantage of this method is that the groups grow in information content as the analysis proceeds, and become progressively less sensitive to errors and accidents in the data. More coefficients have to be calculated than for 'nearest-neighbour' sorting, so that the process is slower. More important than this is the fact that it requires more computer storage space, and is unsuitable for small computers. It is doubtless for this reason that the method, though long known in principle, has in the past been little used in classificatory studies.

2. *The coefficients*

We have used four coefficients, with two versions of one of them, making five in all; they are as follows:

(i) *Correlation coefficient.* This needs no definition. For qualitative data and 'nearest-neighbour' sorting it can be calculated from a 2 × 2 table as the Pearson ø-coefficient; for numerical data, and therefore for centroid sorting, the usual product-moment coefficient has been calculated. Special provision must be made for individuals lacking or possessing all attributes, since for such cases the coefficient is not defined. In the Canberra programme relationships with such individuals are allocated the impossible coefficient of $-2{\cdot}0$, which enables them to be segregated from the rest of the analysis.

(ii) *Squared Euclidean distance.* In a spatial model, let the co-ordinates of two individuals, or of the centroids of two groups, be $(x_{11}, x_{12}, \ldots, x_{1j}, \ldots, x_{1p})$ and $(x_{21}, x_{22}, \ldots, x_{2j}, \ldots, x_{2p})$; then the square of the distance between them is given by

$$\sum_{j=1}^{p} (x_{1j}-x_{2j})^2.$$

Qualitative data are accommodated by taking the jth co-ordinate for an individual as 1 if it possesses the attribute considered, and 0 if it lacks it; in the usual (a, b, c, d) symt olism of a 2 × 2 table, the squared distance between two qualitatively specified individuals then reduces to $(b+c)$. In centroid sorting, each attribute-entry is divided by the number of individuals in the group before calculating the distance; the values so obtained are the co-ordinates of the centroid of the group, and it is from this particular case that centroid methods derive their name.

(iii) *Standardized squared Euclidean distance.* This is merely a variant of (ii). It might plausibly be suggested that, in qualitative data, the joint presence of two rare attributes (or joint absence of two common ones) is more meaningful than the joint presence of two common attributes (or joint absence of two rare ones). To weight such joint occurrences appropriately, the attributes are standardized to zero mean and unit variance before the analysis begins.

(iv) *Non-metric coefficient.* We imply by this the coefficient, for qualitative data, $(b+c)/(2a+b+c)$. It is the complement of the familiar 'coefficient of floral community' which seems first to have been used by Czekanowski (1913), and which is monotonic with the coefficient $a/(a+b+c)$, probably first used by Jaccard (1908), and subsequently by Sneath in his early work (Sneath 1957) to avoid counting double-negative matches. Its quantitative form, in the symbols used in (ii) above, is $(\Sigma|x_{1j}-x_{2j}|)/\Sigma(x_{1j}+x_{2j})$, and as such (though in different symbols) is the familiar coefficient used by Curtis (1959) for ordination. For algebraic reasons which need not concern us here, the synthetic individuals of centroid sorting are again reduced by division to centroid co-ordinates before the calculation. The coefficient is undefined if both individuals being compared are everywhere zero. Since it is desirable that such individuals should be grouped together as identicals, the coefficient is put equal to zero if $(2a+b+c)$ or $\Sigma(x_{1j}+x_{2j})$ is zero.

Attention has been drawn elsewhere (Williams & Dale 1965) to some mathematical shortcomings of this coefficient; nevertheless, it has been used so often in early ecological work that we have felt it necessary to include it if only for its historical importance.

(v) *Information statistic.* The suggestion that statistics of this type should be used in classificatory problems is not novel—*vide*, e.g. Rescigno & Maccaccaro (1960). We deal with the derivation and relationships of the form we use in a parallel paper concerning computer problems (Lance & Williams 1966), and shall here content ourselves with a definition. Let a group of n individuals be specified by the presence or absence of p attributes, and let there be a_j individuals possessing the jth attribute. Then we define a statistic I, such that

$$I = pn \log n - \sum_{j=1}^{p} \left[a_j \log a_j + (n-a_j) \log (n-a_j) \right]$$

The statistic arises from the concept of entropy, and may be regarded as a measure of the disorder of the group; it becomes zero if all members of the group are identical. The most efficient route through the hierarchy is obtained by fusing those two individuals or groups which, on fusion, produce the smallest *increase* in I (I-gain or ΔI); but the absolute *value* of I for the resulting group is a consequent property of the group that may be of interest. For classificatory purposes the base of the logarithms is at arbitrary choice; we have utilized the tables of $n \log n$ to base e given in Kullback (1959).

The statistic is not defined for truly quantitative (i.e. continuously varying) data. In the qualitative case, if only two individuals are being fused, it reduces, in the symbolism of a 2×2 table, to $2(b+c)\log 2$; with 'nearest-neighbour' sorting, where all fusions are of this type, it therefore reduces to a constant multiple of squared Euclidean distance and the classification is identical in form with that produced by (ii).

3. *The problem of hierarchical levels*

The immediate outcome of any similarity analysis is a string of instructions for successive fusions of individuals, each fusion being associated with the value of the coefficient which brought it about. For assessment purposes, however, it is usually desirable not only to know the sequence of the fusions, but also to define in some way a set of levels which can be associated with these fusions.

To appreciate fully the issues involved, we must first consider the properties of the system defined by the ordered fusions, without regard to the coefficient values. If there are n dissimilar individuals, we have a dichotomously-branched system in which the n individuals are progressively fused into the entire population, passing through $(n-2)$ intermediate populations *en route*. This system, which we have called a hierarchy, and which Sokal & Sneath (1963) call a dendrogram, is topologically a *tree*, consisting of *nodes* joined by *line-segments*, with the individuals and the intermediate and final populations occupying the nodes. Any tree has certain important properties, viz.: (i) there are one fewer line-segments than there are nodes; (ii) the system is *connected*, in that there is a continuous route from any node (i.e. individual or population) to any other node; and (iii) if no line-segment is to be traversed more than once this route is unique and passes through a fixed string of nodes. These properties are invariant; the line-segments may be of any length, and a map of the system may be crumpled, twisted or stretched without losing these properties. A tree can always be represented in two dimensions. Since a hierarchical system is intrinsically directional, the only internodal routes of interest or meaning are those joining the population node to any of the ultimate nodes occupied by individuals; we shall refer to these as major routes.

We shall adopt the convention of disposing the individuals horizontally along a baseline with the population node above them and the intermediate nodes in the space between. It is now natural to regard the baseline as an abscissa, and a line perpendicular to this and passing through the population node as an ordinate. However, the height of any intermediate node above the baseline is meaningless without further definition; for the line-segments can be so adjusted that any intermediate node lies above any other.

Since each of the intermediate nodes represents a sub-population which may itself possess features of intrinsic interest, it is again natural to wish to import additional information so that the vertical distance of any such node above the baseline shall be associated with some meaningful property of the sub-population which occupies that node. This is the concept of *hierarchical levels.* A simple example would be to associate each node with the number of individual elements in the sub-population at that node; every node would now occupy a definite position along the ordinate (i.e. a definite distance above the baseline), and every major route would define a monotonic string of integers.

Universally, however, there has been an instinctive feeling that, since the tree itself is generated by a series of similarity coefficients, and since the formation of every population-node is associated with such a coefficient, it would be desirable to use these coefficients for the additional purpose of defining hierarchical levels. Our next problem, therefore, is to investigate the extent to which this is possible.

We consider two sub-populations (j) and (k) which fuse to give a third sub-population (i). Of the five coefficients here under consideration, the first four coefficients have this in common: they are (j, k) coefficients, in the sense that they provide a measure of the difference between (j) and (k), but they provide no measure of the heterogeneity of (i). However, it is reasonable to suppose that, the more dissimilar are (j) and (k), the more heterogeneous is (i), and it is thus reasonable to regard the (j, k) coefficient as a measure of the heterogeneity of (i) so far as a single division of fusion is concerned. If these measures could be accumulated over the hierarchy, a genuine measure of the overall heterogeneity of (i) could be obtained. Unfortunately, none of these four coefficients is additive in this sense; the convention in the past has therefore been to take the (j, k) coefficient, technically only the measure of a single fusion, as the best available measure of the heterogeneity of (i) taken over the whole of the hierarchy up to that point.

The situation is entirely different for the fifth coefficient, the information statistic; for this is an (i, jk) coefficient, defining the difference between (i) on the one hand and (j) and (k) jointly on the other, leaving the (j, k) measure undefined. Moreover, the coefficient is completely additive; if we write $I(i)$ for the total information content of (i), we have by definition:

$$I(j)+I(k)+\triangle I(i, jk) = I(i).$$

Individuals (or groups of identicals) have zero information content; so that if the I values are accumulated successively, a value of the information content (i.e. heterogeneity) is obtained which genuinely applies to the node attained. It is now possible to place the (*i*), (*j*) and (*k*) values on their appropriate point along the ordinate.

However, it is clearly, also possible, to use the convention which the other four coefficients impose on us for setting hierarchical levels; that is, to use the (*i*, *jk*) value for a single fusion—ΔI—as if it itself were a measure of the heterogeneity of (*i*). Only by so doing can we genuinely compare the hierarchies generated by all five coefficients, since the (*i*) values are at least then all subject to the same type of restriction. In our comparative assessment of results from the different methods, therefore, we shall use this convention throughout; the vertical scale will be that of the actual similarity coefficients (ΔI in the case of the information statistic), treated as properties of the group produced *after* the fusion they define.

4. *General criteria for comparison*

The use of a classificatory programme implies that the user has already decided that classification, and not ordination, is the aim; and the use of basically hierarchical system, rather than a clustering technique, likewise implies an interest in the actual path of fusion as well as in the groups. Given that the general requirement is to maximize the information specifically needed at the expense of other less relevant properties of the data, we can erect two basic criteria by which to compare the general effectiveness of the various methods under examination. These are as follows:

(a) The classification should be as clear-cut as possible, i.e. the hierarchy should consist of well-marked groups at well-separated levels as far as the data permit. First, this implies that the coefficient used should be somewhat sensitive to group size, so that the fusion of large groups is delayed as long as possible; the information statistic necessarily possesses this property, the correlation coefficient and Euclidean distance do not, and the properties of the non-metric coefficient are obscure. Secondly, the hierarchy should rise continuously to successively higher levels, i.e. the values used in constructing the hierarchy should be monotonic throughout; hierarchies from 'nearest neighbour' sorting are monotonic by definition, those from centroid sorting not necessarily so unless the coefficient is itself monotonic.

(b) The results must be profitable, i.e. they must suggest groupings which are ecologically meaningful when tested by appeal to other relevant information from outside.

Although on theoretical grounds alone, some of the methods under discussion appear to have certain intrinsic advantages over the others, the extent to which these operate in practice still requires empirical test. In the following section we compare the results obtained from parallel sets of analyses of two test-communities, using each of the two sorting strategies combined in turn with each of the five coefficients.* For ease of reference, we shall subsequently refer to analyses using 'nearest neighbour' sorting as the 'A' series, and centroid sorting as the 'B' series; the coefficients are numbered 1–5 in the order in which they have been discussed.

B. *Community analyses*

The test-communities chosen were two which have been used for earlier work in this series, namely 'Tumulus Heath' (20 sites/76 species) and 'Hoveton Great Broad'(56 sites/73

* Although for 'nearest neighbour' sorting the results from the information statistic are necessarily coincident with those from squared Euclidean distance (see p. 431), the two analyses are both included in all the comparisons for the sake of completeness.

species); their general ecological characteristics have already been described (Papers II, III and IV), and need not be repeated here. For all methods, each community was subjected to both a 'normal' and an 'inverse' analysis, i.e. to a classification of the sites in terms of the species present, and to a classification of the species in terms of the sites in which they occur. In every case the hierarchy was plotted exactly as it emerged from the computer, with no subjective shuffling of the elements.

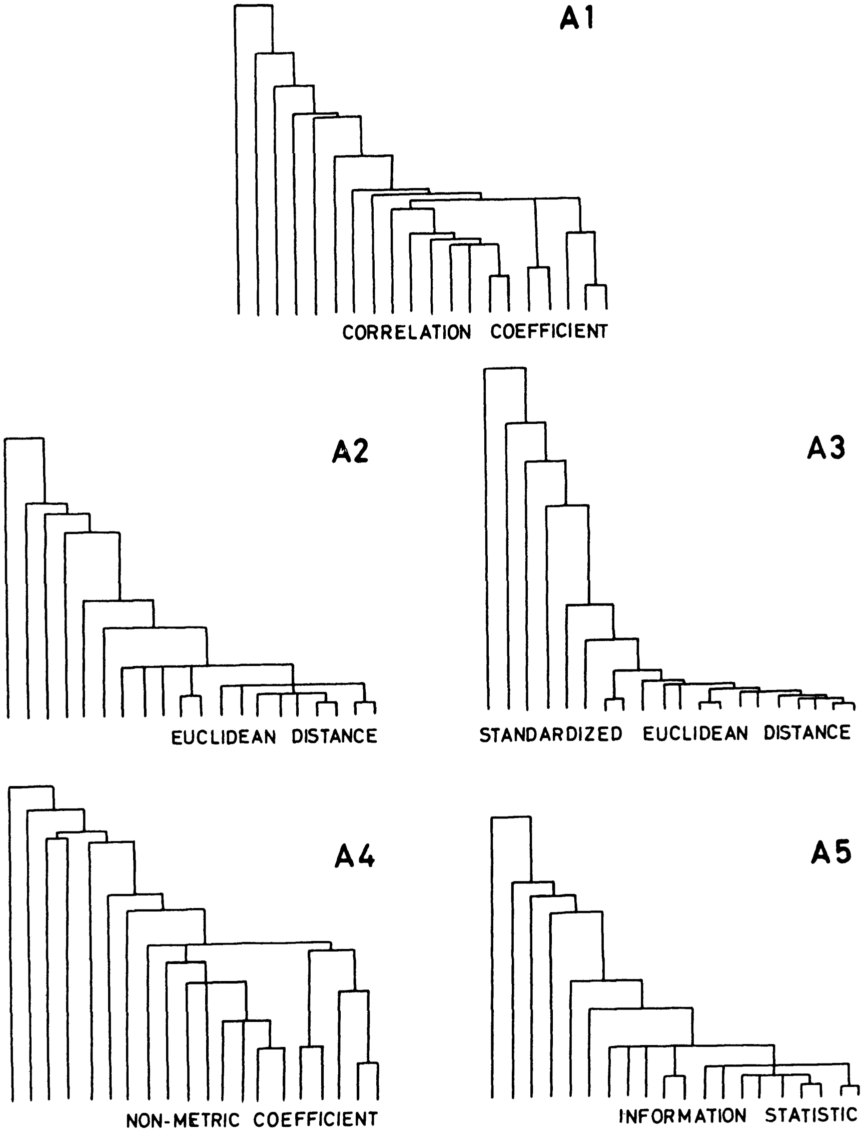

FIG. 1. Tumulus Heath: 'Similarity' analyses. Hierarchies from 'nearest neighbour' sorting methods.

The hierarchies from the normal analysis of Tumulus Heath, reduced roughly to the same scale in terms of range of coefficient, are shown in Figs. 1 and 2. The other three sets (Tumulus Heath inverse; Hoveton normal and inverse), with larger and more complex hierarchies, are not depicted here for reasons of space; the results, however, are included in our general assessment.

It is obvious from the figures that the different analyses give very different results; we may now apply our basic criteria to them.

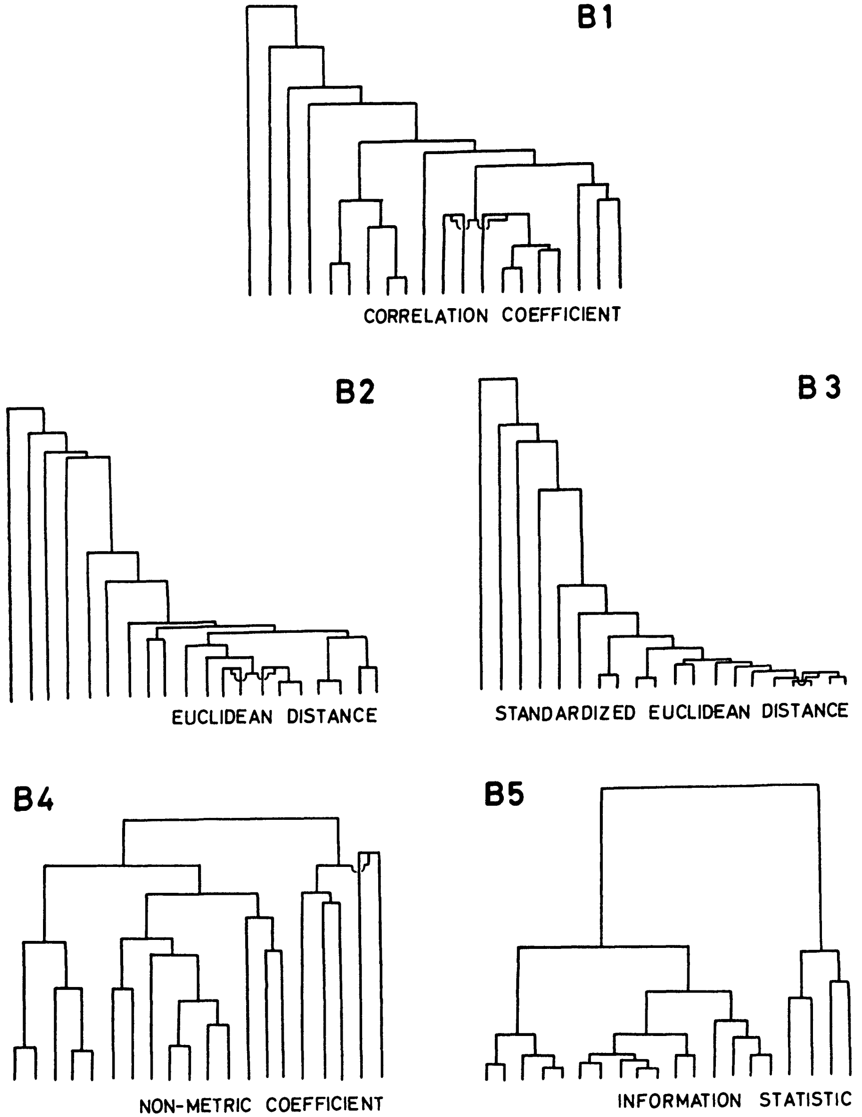

Fig. 2. Tumulus Heath: 'Similarity' analyses. Hierarchies from centroid sorting methods.

1. *General form of the hierarchy*

Here there are two separate features to be considered: (a) the degree of grouping, and (b) the distinctness of the groups depending on degree of what we shall loosely refer to as 'stratification'.

(a) *Grouping*. The essence of a useful classification is that the bulk of the individuals should be absorbed as quickly as possible into groups of higher order: we already know that the individuals (except identicals) are different from one another in some respect, and our main concern is to discover distinctive sets of individuals with properties in common. Furthermore, the groupings at a given level should preferably be of roughly comparable size as far as the data permit; otherwise extrinsic information needed for interpretation will be unequally distributed. In general, therefore, on *a priori* grounds a roughly symmetrical hierarchy is to be preferred to one with a high degree of 'chaining'.

By chaining, we mean the tendency for a given group to grow in size by the addition of single individuals or groups much smaller than itself, rather than by fusion with other groups of comparable size. Since chaining is due to inequality in the numbers in the two sub-populations concerned at each fusion in a hierarchy, a simple assessment of this tendency is possible. A dichotomous hierarchy defining n dissimilar individuals has $(n-1)$ junctions; let these be numbered in any order from 1 to $(n-1)$. Observe, at each junction, the *difference* between the numbers in the two sub-populations which fuse at that point, and let this difference at the ith node be denoted by δ_i. Then we define a coefficient of chaining, C, such that

$$C = \frac{2\sum_{i=1}^{n-1}\delta_i}{(n-1)(n-2)}$$

It is easily shown that this coefficient varies between zero for even divisions throughout (only attainable, of course, if n is a binary power) and unity for complete chaining.

Table 1. '*Chaining coefficient*' *values*

	'Nearest neighbour' sorting					Centroid sorting				
	A1	A2	A3	A4	(A5)	B1	B2	B3	B4	B5
TUMULUS HEATH										
Normal	0·77	0·88	0·86	0·67	(0·88)	0·55	0·75	0·79	0·18	0·18
Inverse	0·69	0·74	0·67	0·85	(0·74)	0·24	0·51	0·55	0·25	0·14
HOVETON										
Normal	0·37	0·35	0·68	0·38	(0·34)	0·26	0·37	0·62	0·21	0·07
Inverse	0·48	0·55	0·63	0·45	(0·55)	0·15	0·50	0·55	0·29	0·04
Mean	0·58	0·63	0·71	0·59	(0·63)	0·30	0·53	0·63	0·23	0·11

Table 1 gives the value of C, over both sorting strategies and all similarity coefficients, for the normal and inverse analyses of both test-communities. To summarize the major differences, the sums of squares of deviations from the grand mean were calculated as in preparation for an analysis of variance. First, this showed that the highest percentage of variation (31%) lies in the difference between the two sorting strategies, the 'A' analyses being in general more highly chained than the 'B'. This is perhaps only to be expected, since, with rather continuous data (as is frequent in ecology), the ability of an individual to link with *any member* of an existing group is likely to prejudice the formation of new groups. It is interesting in this connection that in Tumulus Heath, where the samples were all variants of heathland, not only are the chaining values generally higher than those for the more discontinuous Hoveton vegetation, but the distinction between the two sorting methods is somewhat more marked; in fact, although 18% of the total variation is caused by differences between the communities themselves, community–strategy interaction is still responsible for a further 4%.

Secondly, we must turn to the effect of the coefficients. Although the overall difference between them is appreciably less than that for the sorting strategies, this difference is nevertheless responsible for some 22% of the variation. However, first-order interactions between sorting strategy and coefficient account for a further 12%, due largely to the fact that the chaining values for the 'B' series are more variable than those for 'A'. Indeed, reference to Table 1 shows that the values for B2 and B3 approach the universally high 'A' values, and even surpass them in specific instances. The means for the first-order interactions between strategy and coefficients have been included in the table, from which it is clear that the overall difference between the methods lies mainly in the difference between the low values for B1, B4 and B5 and the high values for the remainder. These three thus fulfil our requirement for good grouping; and, of the three, B5 is clearly the best, though B4 and B1 are sufficiently close to merit further attention.

(b) *Stratification.* Though a tendency towards a symmetrical grouping is a valuable characteristic of a hierarchical method, this is not sufficient in itself. The groups must further be distinguished as sharply as possible from one another, and the levels at which they arise must be unequivocal: the picture is only confused if the criterion for ordering the groups runs counter in a given instance to that used in assessing their relative importance, i.e. the value of the coefficient. Moreover, since the essence of hierarchical classification is that the groupings towards the top have more intrinsic value than the rest, these need to be particularly discrete.

Table 2. *Number of 'reversals' in the value of the similarity coefficient*

	'Nearest neighbour' sorting					Centroid sorting				
	A1	A2	A3	A4	(A5)	B1	B2	B3	B4	B5
Tumulus Heath										
Normal	–	–	–	–	–	1	1	1	1	0
Inverse	–	–	–	–	–	2	3	4	2	0
Hoveton										
Normal	–	–	–	–	–	6	7	5	1	0
Inverse	–	–	–	–	–	10	14	9	5	0
Total	–	–	–	–	–	19	25	19	9	0

The 'nearest neighbour' methods cannot by definition give rise to reversals in the value of the coefficient for successive groups, but this is not true of the others. Reference to the figures for Tumulus Heath shows that, even in this simple community, B1–4 have one reversal each; in B4, the reversal is particularly serious in that it occurs in the upper part of the hierarchy.

The overall situation for the two test-communities is shown in Table 2, which gives the actual numbers of reversals produced by the different methods over all the analyses. B2, with a total of twenty-five, is particularly bad in this respect; B1 and B3, with nineteen each, are not much better; B4, with nine, still has too many for convenience; and it is only B5, with none, which meets this particular requirement completely.

Though the presence of reversals may confuse the stratification at certain points, the general picture is also affected by the proportion of the total range of the coefficient value occupied by successive fusions. Ideally, the most conspicuous changes should be towards the top, so that the major groupings stand out above the rest. Admittedly this can be adjusted by appropriate algebraic means; but since the solution is unique for a given population and cannot be used directly for another, such tricks are not to be recommended.

A rough guide to the general distribution of the groupings can be obtained by calculating the proportion of the total range of coefficient occupied by a given percentage of the fusions. In a complex population, it is usually possible to assess only the upper few, so that the bulk of the sub-populations can be ignored. For this reason, we have calculated, in Table 3, the proportion of the range occupied by the top 15% of fusions in all our analyses. This again shows B5 to be far superior to the rest, and confirms us in our preference for this method on grounds of the evidence so far.

Table 3. *Proportion of total range of similarity coefficient value occupied by top 15% of fusions*

	'Nearest neighbour' sorting					Centroid sorting				
	A1	A2	A3	A4	(A5)	B1	B2	B3	B4	B5
TUMULUS HEATH										
Normal	0·39	0·36	0·42	0·19	(0·36)	0·36	0·18	0·37	0·18	0·69
Inverse	0·50	0·50	0·52	0·48	(0·50)	0·60	0·74	0·73	0·39	0·86
HOVETON										
Normal	0·44	0·18	0·29	0·69	(0·18)	0·69	0·17	0·41	0·56	0·91
Inverse	0·16	0·54	0·77	0·19	(0·54)	0·64	0·41	0·85	0·17	0·92
Mean	0·37	0·39	0·50	0·39	(0·39)	0·57	0·37	0·59	0·32	0·83

Finally, before we leave this section, we must look briefly at the B5 hierarchy in its genuine form, i.e. with its levels plotted according to *total* information content instead of mere information gain (see p. 433); we shall call this other version B5′. It is clear from Fig. 3, which shows the two forms of B5 side by side, that the desirable feature of marked stratification has been accentuated by the new levels; moreover, there are slight changes in the relative levels of certain fusions which could be ecologically important.

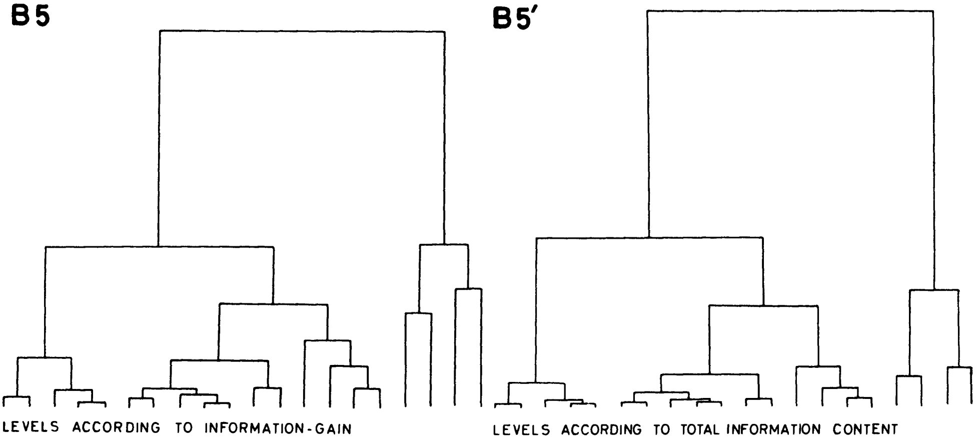
B5
B5′
LEVELS ACCORDING TO INFORMATION-GAIN
LEVELS ACCORDING TO TOTAL INFORMATION CONTENT

FIG. 3. Tumulus Heath: Comparison of hierarchial levels in the B5 method.

2. *Ecological assessment*

Though a method may appear to have unassailable theoretical advantages, the acid test is whether the information it imparts is acceptable to the user—in this case, the ecologist. Frequently, a method developed *in vacuo* fails miserably when actually applied, either because of other unexpected and hitherto unexamined properties of the method itself, or because of special features of the data which make it inapplicable. We shall therefore complete this study by comparing in general terms the ecological information extracted from the two-test communities by the various methods.

(a) *Method of assessment.* The difficulty with a comparison involving the concept of ecological acceptability is to find objective criteria in an essentially subjective situation by which to differentiate between the analyses. Two courses are open to us in this respect: either we can examine the results from each analysis directly and assess the extent to which the groupings produced are readily interpretable in the light of other ecological experience; or we can erect a set of groupings *ab initio* from this experience and assess the extent to which these groupings are reflected by the analyses. Since ecological experience has many facets, some are likely to be more relevant to the one situation and some to the other; the two approaches will therefore not necessarily give exactly the same results. However, if only one is to be employed, the choice between the two is largely a matter of convenience. In present circumstances, we find it easier to adopt the second system.

With the given test-communities, the threshold for our acceptance of any of the hierarchical methods under study will be that the major groupings which arise shall not be fewer than, or markedly different from, those recognized intuitively as distinct ecological entities at the time the data were collected. The stability of these entities has already been partly confirmed by the fact that they reappeared (together with additional information) when these communities were subjected to association-analysis in our earlier work (Paper III, p. 723; Paper IV, pp. 792, 799); we therefore feel justified in setting a minimum standard based on them. For Tumulus Heath, we shall require a differentiation between wet heath, dry heath and grass-heath at least; and for Hoveton, a similar separation will be needed between open water, reedswamp, primary fen, mowing marsh, swamp carr and fen carr groupings. These we shall call our *standard ecological categories.*

As a second criterion, we may reasonably ask that the groupings in the different categories for either community shall all be differentiated at roughly the same hierarchial level; the very fact that these groupings are required to reflect the lowest level of intuitive recognition of distinctive vegetation types in the original subjective survey suggests that in some sense they should be equivalent in degree of ecological homogeneity.

Since it is clearly impracticable to give detailed consideration here to the separate results of forty different analyses, a crude allocation system for overall comparison was devised:

(i) Each quadrat and species was first subjectively allocated to one or other of the predetermined ecological categories on the basis of previous experience of the vegetation; individuals of uncertain ecological affinities, such as anomalous or ecotonal quadrats, or widely tolerant species, were left uncategorized. The categorized individuals were known as *standard units.*

(ii) Each grouping in each analysis was then examined for ecological homogeneity in terms of the standard units it contained. To qualify for recognition as a *standard grouping*, a group must have accumulated at least 50% of the total number of standard units in a given category without the addition of any unit or grouping of units of different type. To give as much latitude as possible, the uncategorized individuals were regarded as ‘floaters’, i.e. they could be attached to any group without destroying or adding to its ecological integrity.

(iii) The hierarchical level at which a standard grouping achieved 50% of its total possible membership was designated its *minimum level of organization*; the ultimate level which it reached before becoming contaminated by the addition of one or more extraneous units, or fused with another standard group, was designated its *level of maximum differentiation*. Where the picture was confused at a critical point by reversals in level (see p. 437), the sequence of fusions was allowed to override the relative values of the levels.

(b) *Results*. Table 4 gives the tally of standard groupings for all the hierarchies under examination, including the two versions of B5; a positive entry for a particular ecological category indicates the presence of a standard grouping of that type, while a negative entry means that the requisite degree of differentiation was not achieved. As a rough indication of the relative levels at which the various standard groupings became organized in any one analysis, the single lowest level over all the groupings at which a grouping was maximally differentiated was taken as a dividing line; those groupings which existed at this level are shown in bold-face type, while the others, whose minimum level or organization lay higher than this line, are shown in normal type.

It is clear from the table that even the very crude allocation system employed is sufficient to discriminate cleanly between the analyses. We shall first consider the emergence of the groupings themselves, and then the relative levels at which they arise.

The first point of interest lies in the general pattern of the table as a whole. For both Tumulus Heath and Hoveton, there are certain standard groupings which appear with all the methods. In Tumulus Heath, for instance, wet and dry heath are differentiated throughout in the normal analyses, while in Hoveton the open water and mowing marsh groupings appear with all methods on both normal and inverse sides; these entities are clearly sufficiently distinct for every method to extract them, whatever its sorting strategy or coefficient. Conversely, there are other groupings which appear infrequently, such as the 'normal' grass-heath grouping in Tumulus Heath, and the 'normal' swamp carr and 'inverse' fen carr in Hoveton; and it is the less well-circumscribed groupings like these which serve best to test the relative sensitivity of the different methods.

Secondly, there is some interest in comparing the relative efficiency of the extraction on the normal and inverse sides. For most of the analyses, the *number* of groupings produced is similar on both sides, though the *categories* may differ. However, for the two methods using the non-metric coefficient (A4 and B4), there is an overall tendency for the normal groupings to be better than the inverses. This is clearly a function of this particular coefficient, which is sensitive only to positive matches: since the species, being abstractions, are inherently more likely to be more variable than the quadrats, they will tend to have fewer occurrences in common to bind them together into discrete groups.

With regard to the overall number of groupings extracted by the individual methods, B5 is clearly best, with all eighteen possible groupings represented; B4, with fourteen groupings despite its failure on the inverse side, is next, followed by A1 and B1 with thirteen each. In general, the methods using the centroid strategy show a slightly better performance than their counterparts with 'nearest neighbour' sorting; but the fact that A3 and B3 show the lowest returns in the 'A' and 'B' series respectively, while those for A1 and B1 are relatively high, is a pointer to the importance also of the effect of the different coefficients.

We may now turn to our second criterion concerning the relative levels of differentiation of the standard groups extracted from a single analysis. Even with the extremely crude device we have used to identify the groupings which become organized only at high

Table 4. *Extraction of ecological groupings by the different methods*

	Tumulus Heath						Hoveton												
	Normal			Inverse			Normal						Inverse						
	Wet heath	Dry heath	Grass-heath	Wet heath	Dry heath	Grass-heath	Open water	Reedswamp	Primary fen	Mowing marsh	Swamp carr	Fen carr	Open water	Reedswamp	Primary fen	Mowing marsh	Swamp carr	Fen carr	Total standard groups (max. = 18)
A1	**+**	**+**	−	+	−	**+**	**+**	+	+	+	−	**+**	**+**	−	+	**+**	**+**	−	13
A2	**+**	**+**	−	+	−	**+**	**+**	−	+	+	−	+	+	−	−	**+**	+	−	11
A3	**+**	**+**	−	+	−	**+**	**+**	−	−	+	−	−	+	−	−	**+**	+	−	9
A4	**+**	**+**	−	−	−	−	**+**	**+**	+	**+**	−	**+**	**+**	−	−	**+**	**+**	−	10
(A5)	(**+**)	(**+**)	(−)	(+)	(−)	(**+**)	(**+**)	(−)	(+)	(+)	(−)	(+)	(+)	(−)	(−)	(**+**)	(+)	(−)	(11)
B1	**+**	**+**	−	**+**	−	−	**+**	+	+	**+**	−	**+**	**+**	+	+	**+**	**+**	−	13
B2	**+**	**+**	−	**+**	+	**+**	**+**	−	+	+	−	+	+	−	−	**+**	+	−	12
B3	**+**	**+**	−	+	+	**+**	**+**	−	−	+	−	−	+	−	−	**+**	+	−	10
B4	**+**	**+**	**+**	**+**	**+**	−	+	**+**	**+**	**+**	**+**	**+**	**+**	−	−	**+**	**+**	−	14
B5	**+**	**+**	+	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	+	**+**	**+**	**+**	18
B5′	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	18

levels, the general picture is clear. For the majority of methods, at least some of the groupings fail to form until at least one another grouping has already been fully differentiated and has lost its identity; in such a case, only a proportion of the total number of standard groupings extracted can exist together at any one hierarchical level, and the value of these groupings as ecological units for mapping or descriptive purposes (as in the original survey) is correspondingly reduced.

On this particular criterion, A4 and B4 are good, with only one failure each; but, because of its other advantages, the interest naturally centres on B5. Here, since we are now specifically concerned with hierarchical levels, we may legitimately extend our comparison to include both versions of the hierarchy which can be constructed from the B5 type of analysis (see p. 438); and we find that, whereas the form which is strictly comparable with the other hierarchies has two failures, the other—B5′—has none. The B5′ hierarchy is thus the only one which fulfils our initial requirements completely, both in the range of standard groupings extracted and the levels at which they are formed.

(c) *Discussion.* The immediate results from the crude tests applied for ecological assessment appear at first sight to eliminate all except B5 from further serious consideration. However, it could be argued that to concentrate exclusively on the ability of a method to extricate predetermined groupings at roughly predetermined levels is placing too much reliance on preconceived ideas as to the nature of the situation in the given test-communities. Ideally, a parallel appraisal should be made using the alternative method of assessment (see p. 439), in which the groupings themselves are examined directly in their own right, using an external measure of some sort to compare their ecological significance. Apart from the difficulty of erecting such a measure, the extra expenditure of time could hardly be justified here: the real difference between the methods lay not in the ecological *nature* of the groups, but whether substantial groupings were formed *at all*, i.e. those methods which failed to provide the requisite number of 'standard' groupings did not in general provide alternative groupings to be assessed. Some of the failures were certainly due to intermingling of individuals from allied ecological categories, such as reedswamp and open water, or the two types of carr; but these were failures due to lack of differentiation, rather than to the erection of genuinely new and unexpected groupings. In short, there was little evidence of migration of tagged individuals into new composite groups, such as would justify an independent assessment from other external evidence.

It is particularly interesting in this context that, in the few cases where substantial new-groupings did in fact emerge, these usually concerned the unallocated individuals; these groups were thus *additional* to the standard groups instead of replacing them. Since we have been primarily concerned with threshold criteria for the acceptance of a method we have so far ignored these supernumerary groupings in our general assessment. They were almost entirely restricted to B4 and B5 and, as we shall be examining the B5 hierarchies in more detail subsequently, we shall content ourselves for the moment with the comment that most of them seemed ecologically meaningful.

Before a final decision is made in favour of B5, B4 deserves a last consideration; though failing on some counts, such failures were usually rather trivial. For instance, detailed examination of the B4 hierarchies showed that failure to extract inverse standard groupings was due largely to the incorporation of an alien element at a critical stage, rather that to a genuine lack of grouping. It is, in fact, its emphasis on rather trivial features which operates particularly against B4 for ecological work: the inability of the coefficient to use information from negative matches means that chance records—such as entries for casual species—are given undue importance. Nevertheless, the overall performance of B4 is such that with more symmetrical data—as in certain types of taxonomic work—its particular disadvantages could easily be outweighed (see e.g. Watson, Williams & Lance 1966). For present purposes, however, B5 seems superior on all counts. We have therefore selected B5 alone for further consideration in the following paper; and we shall henceforth refer to B5—the method using centroid sorting and the information statistic as its coefficient—as *information-analysis.*

IV. GENERAL DISCUSSION

With interest increasing in the use of numerical methods, it is inevitable that techniques developed in one field of study should find their way into others. The history of taxonometrics is no exception to this: strategies and coefficients developed for vegetational work have been adopted for the classification of individual organisms, and *vice versa.* However, without some regard for differences in the nature of the material to be manipulated, such practices can easily lead to failure. For instance, a rather crude classificatory method can appear strikingly efficient if used on material already partly classified subjectively before analysis, as in much taxonomic work; but if the same method is then required to extra groupings from more continuous vegetational data, it may prove insufficiently sensitive for the purpose. The essence of this difference in material to be handled is concisely stated by Webb (1954): 'The majority of species are guaranteed some measure of objectivity, stability and discriminability by the genetic pattern .... No comparable factor is available to stabilize plant communities .... The fact is that the pattern of variation shown by the distribution of species among quadrats over the earth's surface chosen at random hovers in a tantalizing manner between the continuous and the discontinuous'.

With such material, it is only natural that plant ecologists have wavered between the relative merits of classification and ordination as a means of reflecting vegetational relationships. At first sight, therefore, a method which is sensitive only to fairly discrete groups and then 'chains' the other individuals might be regarded as a useful compromise between the two. However, although the chained part of a hierarchy may appear superficially like an ordination, the order in which the chained individuals appear is related only to their degree of similarity with the groupings previously formed; there is, in fact, no necessary immediate affinity between two juxtaposed chained individuals, and hierarchies showing excessive chaining have therefore little useful function to perform.

The very fact that different analyses of exactly the same set of data show every gradation between excessive chaining (as in A3) and fairly symmetrical grouping (as in B5), itself is an indication of the futility of arguments as to the 'real' nature of vegetation. All that such arguments mean, is that different observers are instinctively using different values to assess similarities and differences between one individual and another, or between an individual and a pre-erected group. Some ecologists may be most impressed in the field with likenesses between neighbouring vegetation samples, and are intuitively using 'nearest neighbour' sorting; others, with perhaps more power of integration, are using a mental process akin to centroid sorting; others, again, may be particularly struck by the presence of rare species and the absence of common ones, and are mentally using a measure like the standardized Euclidean coefficient; while, yet again, others may be especially susceptible to the size of the area covered by a given type of vegetation, and tend to make mental adjustments allied to the group-size sensitivity of the information statistic.

With such diverse conceptions as to what is 'important' in vegetational analysis, the only objective criterion we can use in selecting a 'best' method is to choose that method which will most efficiently perform the particular function which is asked of it. Our attitude here has been that, in a complex ecological situation, the clarity of exposition of the results is all-important. The great advantage of information-analysis in this respect is that, not only does it produce a clear-cut hierarchy, but the method itself is internally consistent so that different mathematical models are not confused: by minimizing the variables in the method of analysis itself, the variables in the situation under study are thus more clearly exposed.

ACKNOWLEDGMENTS

We are indebted to Mr P. Macnaughton-Smith, of the Home Office Research Unit, for his original suggestion of the possible application of the information statistic in the present context. The diagrams for this paper were drawn by Mr B. A. Lockyer.

SUMMARY

Agglomerative-polythetic methods (commonly known as 'similarity methods') of hierarchically classifying elements into sets can take a large number of different forms, according to: (a) the type of fusion strategy ('sorting method') employed; and (b) the coefficient used to measure similarity. Ten selected versions, using two different sorting methods combined in turn with five different coefficients, are tested empirically for their relative efficiency, using both theoretical and ecological criteria. The results from the comparative analyses of two test-communities show that, whereas 'centroid' sorting in general gives better results than 'nearest neighbour' sorting, there is also an interaction between sorting strategy and coefficient. The method combining centroid sorting with an information-statistic coefficient is shown to be greatly superior to the others in producing clear-cut and ecologically acceptable hierarchies; and this method, called *information analysis*, is selected for further test.

REFERENCES

Curtis, J. T. (1959). *The Vegetation of Wisconsin.* Wisconsin.

Czekanowski, J. (1913). *Zarys metod statystycznych* (*Die Grundzuge der statischen Metoden*). Warsaw. (*ex* Curtis, 1959).

Edwards, A. W. F. & Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics*, **21**, 362–75.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. vaud. Sci. nat.* **44,** 223–70. (*ex* Sokal & Sneath 1963).

Kulczynski, S. (1927). Zespoly roslin w Pieninach. (Die Pflanzen-associationen der Pieninen). *Bull. intern. Acad. pol. Sci. Lett., Cl. Sci. Math., et Nat.* Ser. B. Suppl. II, **57,** 203. (*ex* Curtis, 1959).

Kullback, S. (1959). *Information Theory and Statistics.* New York.

Lambert, J. M. & Dale, M. B. (1964). The use of statistics in phytosociology. *Adv. ecol. Res.* **2,** 59–99.

Lambert, J. M. & Williams, W. T. (1962). Multivariate methods in plant ecology. IV. Nodal analysis. *J. Ecol.* **50,** 775–802.

Lance, G. N. & Williams, W. T. (1966). Computer programs for hierarchical polythetic classification ('Similarity analyses'). *Brit. Comp. J.* (In press).

Macnaughton-Smith, P., Williams, W. T., Dale, N. B. & Mockett, L. G. (1964). Dissimilarity analysis: a new technique of hierarchical sub-division. *Nature, Lond.* **202,** 1034–5.

Rescigno, A. & Maccaccaro, W. B. (1960). The information content of biological classifications. In: *Symposium on Information Theory.* London.

Sneath, P. H. A. (1957). The application of computers to taxonomy. *J. gen. Microbiol.* **17,** 201–26.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy.* San Francisco.

Watson, L., Williams, W. T. & Lance, G. N. (1966). Angiosperm taxonomy: a comparative study of some novel numerical techniques. *J. Linn. Soc.* (In press).

Webb, D. A. (1954). Is the classification of plant communities either possible or desirable? *Bot. Tidsskr.* **51**, 362–70.

Williams, W. T. & Dale, M. B. (1965). Fundamental problems in numerical taxonomy. *Adv. bot. Res.* **2**, 35–68.

Williams, W. T., Dale, M. B. & Macnaughton-Smith, P. (1964). An objective method of weighting in similarity analysis. *Nature, Lond.* **201**, 426.

Williams, W. T. & Lambert, J. M. (1959). Multivariate methods in plant ecology. I. Association-analysis in plant communities. *J. Ecol.* **47**, 83–101.

Williams, W. T. & Lambert, J. M. (1960). Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis. *J. Ecol.* **48**, 689–710.

Williams, W. T. & Lambert, J. M. (1961). Multivariate methods in plant ecology. III. Inverse association-analysis. *J. Ecol.* **49**, 717–29.

Williams, W. T. & Lance, G. N. (1965). Logic of computer-based intrinsic classifications. *Nature, Lond.* **201**, 159–61.

(*Received* 3 *November* 1965)