EXPERIMENTAL APPRAISAL OF CERTAIN PROCEDURES FOR THE CLASSIFICATION OF DATA

By R. T. LANGE,* N. S. STENHOUSE,† and CHRISTINA E. OFFLER*

[*Manuscript received August* 10, 1965]

Summary

An experimental appraisal of two of the commoner techniques for empirical classification, namely Sneath Q-study and Williams and Lambert normal analysis, is presented. It is shown that the methods have unresolved problems associated with them and that results obtained by their use are not necessarily liable to effective interpretation. A number of objections are raised and it is suggested that biologists using the methods should reconcile their conclusions with these objections.

Data are presented for the first time on a range of relevant features including matched results, comparisons of split data, the influence of random variation, the relative performance of χ^2 versus percentage similarity, the nature and stability of percentage similarity, and on situations where the techniques as they stand are very suitable.

I. INTRODUCTION

Classification of data, particularly in biology, has led recently to the development of "numerical taxonomies". Examples are provided by Sneath (1957*a*, 1957*b*), Sneath and Cowan (1958), Cheeseman and Berridge (1959), Hill (1959), Lysenko and Sneath (1959), Liston (1960), Colwell and Liston (1961*a*, 1961*b*), Cerbon and Bojalil (1961), Beers *et al.* (1962), Sokal and Rohlf (1962), Pohja and Gyllenberg (1962), Floodgate and Hayes (1963), Proctor and Kendrick (1963), and Graham (1964). They all start by considering data which describe numerous things according to some sets of characteristics. Such data are properly described as an "entities" × "attributes" matrix.

Procedures of numerical taxonomy have been developed concurrently in various places and are now assembled into an interdisciplinary field which is the subject of books (Sokal and Sneath 1963) and parts of books (Davis and Heywood 1963; Grieg-Smith 1963), of reviews (Sneath 1962, 1964*a*; Lambert and Dale 1964), a newsletter (*Taxometrics*), and a large expanding journal literature ranging, for example, from microbiology (Hill *et al.* 1961), entomology (Sokal and Michener 1958), higher plant taxonomy (Rogers and Tanimoto 1960; Rogers and Fleming 1964), plant sociology (Williams and Lambert 1959, 1960, 1961), and paper chromatography (Cheeseman and Berridge 1959) to even (for instructional purposes) heraldic beasts (Sneath 1964*b*).

Many biologists will wish to use these methods in the same manner as they use analytical statistics, that is, to achieve an objective by methods largely taken on faith; it is to these biologists that this paper is addressed.

* Department of Botany, University of Adelaide.

† Division of Mathematical Statistics, CSIRO, Adelaide.

The situation in numerical taxonomy contrasts with that in statistics. Our accumulated experiences provide enough instances of deficiencies within numerical

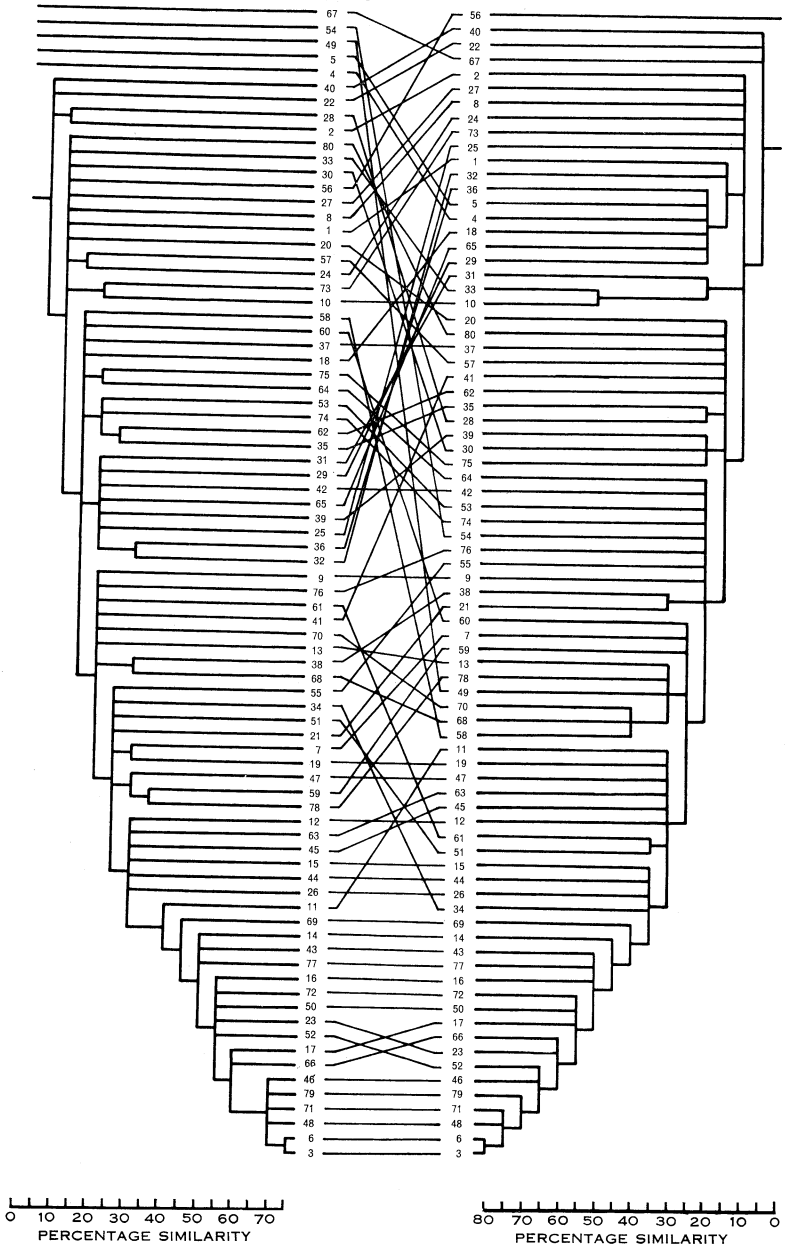

Fig. 1.—Dendrograms expressing firstly the interrelationships between 80 plant species on the basis of their incidence patterns in 165 quadrats, and secondly the difference between the results of two successive determinations. For further explanation see Sections II(*a*) and III(*a*)(vi).

taxonomy to dispose of any suggestion that its methods can be taken on faith. Some deficiencies are fundamental and extremely serious, permitting the biologist to

arrive at spurious inferences, and many precedent papers involve methods in which these deficiencies must have occurred (see literature already cited).

It is clear that the biologist must check any method he wishes to use, and decide for himself if it provides a sound basis for inference. This experimental appraisal is such a check on the two methods of numerical taxonomy which we wished to use, those on the lines described by Sneath (1957*a*, 1957*b*) and by Williams and Lambert (1959, 1960, 1961), and the reader who wishes to follow it is assumed to have read these five papers.

The appraisal is presented in three sections. The first concerns comparisons, which previously have not been developed very far, between these two very similar methods. It is shown that the rational basis for their application is different, and that they assess different properties of data, but produce the same groupings. It is also shown that the measure "percentage similarity" basic to one of the methods takes into its significant values different sorts of information in an uncontrolled fashion. Further, some steps are shown to be unnecessarily arbitrary.

These features are pointed out and discussed in relation to interpretation and inference.

The second section concerns the influence of chance in results obtained by these methods, a feature which has received insufficient attention previously. Chance is shown to be capable of introducing much instability to groupings, impairing the reliability of inferences drawn from them.

While we have been concerned mainly with objections, we have also found situations where the techniques as they stand are suitable. The palaeobotanical example presented in the third section illustrates how one of the methods achieves unqualified success.

II. Experimental

Except where otherwise stated, all calculations were done on a CDC 3200 computer at the CSIRO Computing Research Section at Adelaide University, and were programmed by N. S. Stenhouse:

(1) In a field study, data were obtained on the incidence of over 100 plant species within 330 quadrats located in vegetation by restricted randomization. Half of the quadrats were selected at random from each stratum of the original sampling, and a table compiled listing incidence within these quadrats of the 80 most frequent species. A second table was then compiled listing the incidence of the same species in the other 165 quadrats. The contents of each table were then analysed by the method of Sneath (1957*a*—Q-study), and the results expressed in the dendrograms of Figure 1.

(2) From data in one of the two preceding tables, percentage similarities of species incidence patterns were calculated by the method of Sneath (1957*a*), and the nature and extent of species associations by calculating χ. χ is a standard normal deviate which can take positive and negative values and is the square root of the more familiar χ^2. A negative value of χ indicates inverse association. These values are plotted in Figure 2.

(3) Species associations and percentage similarities according to the data in the two preceding tables of (1) were calculated and the results of each interspecific comparison according to each of the two data tables are plotted in Figures 3 and 4.

(4) Random data simulating the incidence of plant species in 100 quadrats in vegetation were obtained. These data were subjected to normal analysis according to the method advanced by Williams and Lambert (1959). Results are presented in Figure 5. This analysis was done on a CDC 3600 computer at the CSIRO Computing Research Centre, Canberra, programmed by G. N. Lance.

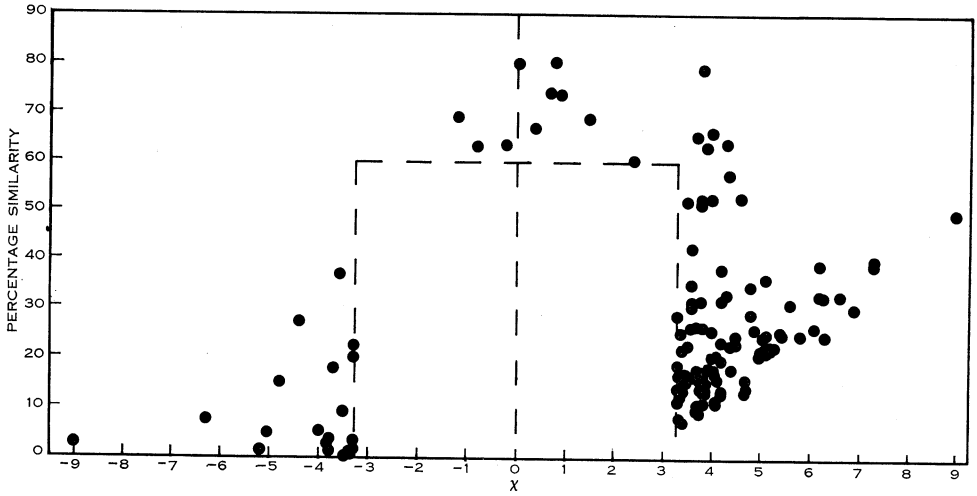

Fig. 2.—Relationships between percentage similarities and interspecific associations determined for 80 plant species on the basis of their incidence in 165 quadrats in vegetation.

(5) Vegetative shoot morphology of Australian coniferous plants was examined in detail from authentic herbarium sources. Concurrently, collections from Arcoona Plateau fossil floras (Oligocene ?) were searched for vegetative shoots closely resembling any known type of Australian coniferous shoot architecture. Twenty-one fossil types were detected and similarly were examined in detail for features of their vegetative shoot morphology. Descriptive common ground for comparisons between shoots in overall extant-fossil comparisons extended to 62 features. An entities $\times$ attributes table was compiled describing both the extant and the 21 fossil species on the basis of the 62 common-ground features of their vegetative morphology. These data were subjected to Sneath Q-study, and the results are expressed in the right-hand dendrogram of Figure 6.

The data were also subjected to Williams and Lambert's normal analysis (by G. N. Lance on the CDC 3600 computer at Canberra), reading vegetative shoots in place of quadrats, and morphological features in place of species representation. These results are expressed in the left-hand dendrogram of Figure 6.

III. Results and Discussion

(a) *Comparisons between Sneath Q-Study and Williams and Lambert Normal Analysis*

(i) *Rationales Compared*

Sneath Q-study was intended to classify into groups bacteria according to their overall similarities in respect to common-ground attributes. The reasoned argument basic to this has been presented in full by Sneath (1957*a*) and elsewhere since (Sneath 1962, 1964*a*). It is an argument readily transferable to the general case in biology.

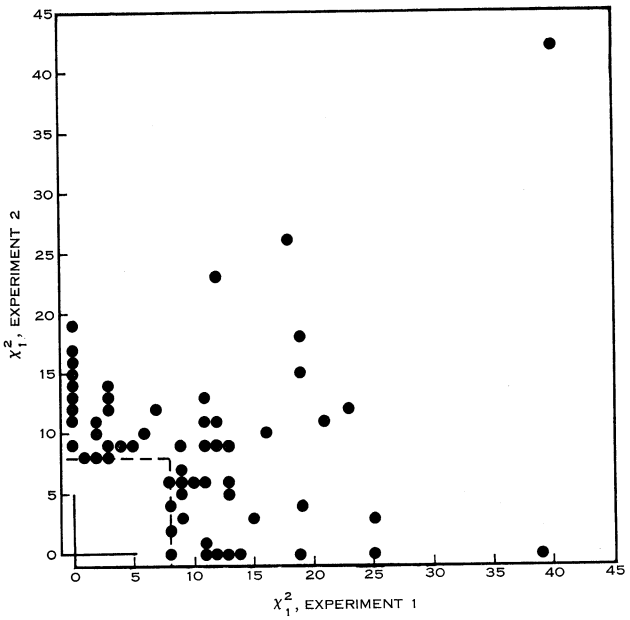

Fig. 3.—χ^2 was determined for interspecific comparisons between 80 plant species in 165 random quadrats, and this was repeated for the same species in a second 165 random quadrats from the same area. Figure 3 illustrates the relationships between the corresponding χ^2's in the two trials. Values of χ^2 less than 8·0 are not shown.

The "similarity" of any two things is taken as the ratio of the number of attributes for which both are positive, to the number of attributes for which at least one is positive, and a matrix of percentage similarity values is computed and assembled. This is then sorted in an attempt to display entities in a schema where "similar" ones are placed together and "dissimilar" ones are separated. Attempts involve much compromise because relationships involve many more dimensions than can be illustrated easily by diagrams.

Williams and Lambert normal analysis concerns the grouping of vegetation samples into aggregates representing plant communities. Sneath Q-study is also applicable to the problem of classifying vegetation samples, reasoning that these group into communities directly according to the overall similarities of their floristic content. However, normal analysis reasoning is quite different. Williams and

Lambert subdivided samples so that attribute association demonstrable by significance tests in the pooled samples was absent from subdivisions. The logical basis of this is that association of the attributes in the population is due to the variation between the subdivisions. This idea was developed by Goodall (1953), after Tuomikoski (1942).

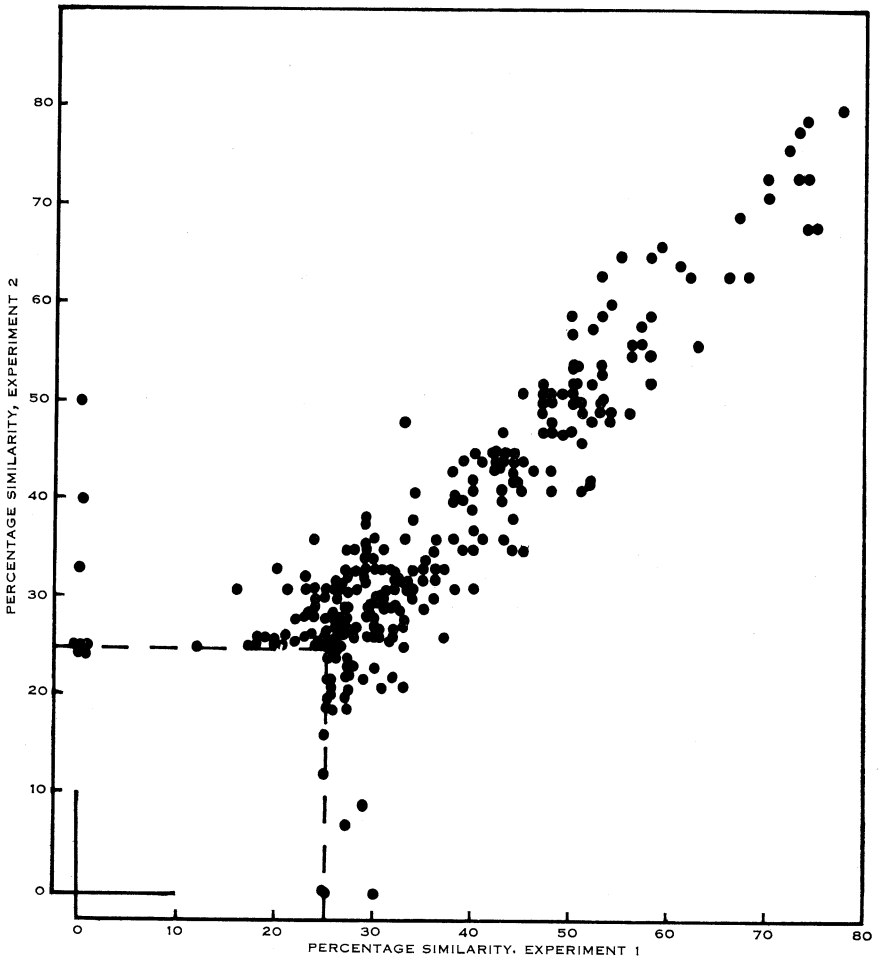

Fig. 4.—Percentage similarity was determined for comparisons between 80 plant species in 165 random quadrats, and this was repeated for the same species in a second 165 random quadrats from the same area. Figure 4 illustrates the relationship between corresponding percentage similarities in the two trials, for each of the comparisons attaining $\geqslant 25\%$. The few erratic points about zero on one axis are the result of comparisons based on species of exceedingly low frequencies from 1 to 6%.

Given that over some defined area exactly n plant species are represented, then if association exists between any m species in that area, the vegetation is heterogeneous in that some components (the m species) differ from others in their incidence pattern in the area. The botanist infers this to be due to differential reaction by the m associated species to something environmental, and concludes that more than one community exists in the area, the fundamental community being envisaged as that

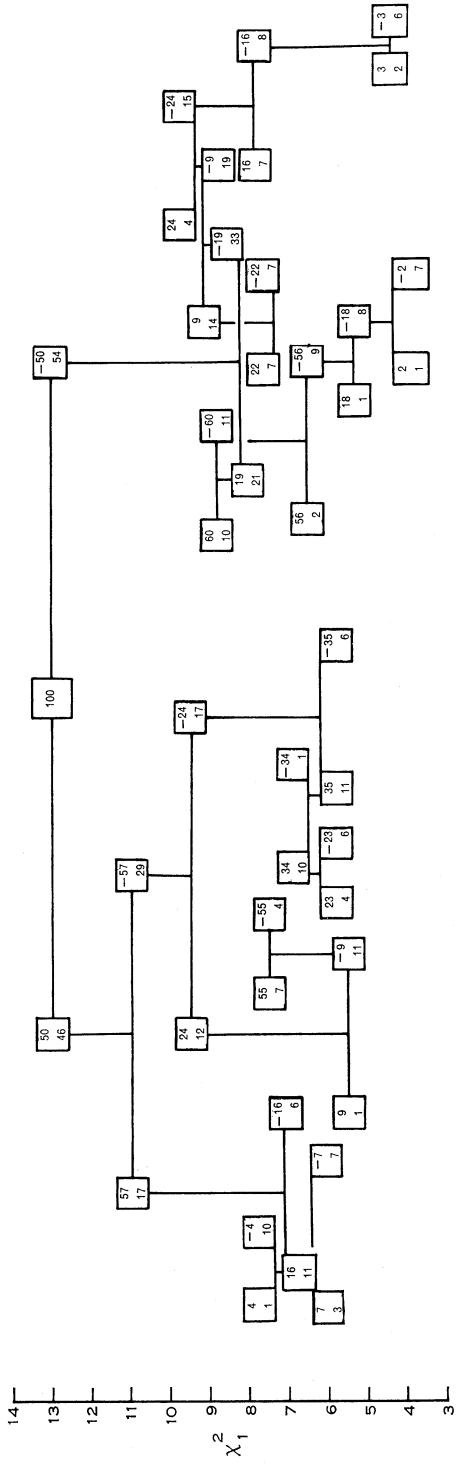

Fig. 5.—Normal analysis of random data simulating plant incidence in 100 quadrats from vegetation.

part of the total area defined by the incidence of a certain set of species but without their exhibiting the above basis for further subdivision.

Compared with Q-study, normal analysis at first appears to be at a disadvantage since its underlying reasoning is not phrased for cases other than the vegetation analysis for which it was developed. That is, biologists do not as a rule think of their purpose when classifying things as extracting groups of them wherein attributes show no mutual statistical associations. Certainly such an approach would not serve at all well in many biological contexts. It would be irrational to apply Williams and Lambert's normal analysis to the classification of plants or animals, for example, unless one is first convinced that what one really seeks are groups wherein attributes show no statistical associations (influences on the probabilities of each other's incidence) above some arbitrary level. In contrast, the reasoning backing a Sneath Q-study may appear at first sight to be attractively straightforward. Percentage similarity seems very versatile and widely applicable. It is, however, not straightforward, but on the contrary may be misleading, since it takes into its significant values such different sorts of information as statistical association, and ubiquity, as will be shown [see Section III(*a*)(v)].

(ii) *Implementations Compared*

Figure 7 illustrates that the initial manoeuvres by these methods are very similar indeed. It is only in their final stages that they differ much.

Things to be classified (here called entities) are each described by a series of positive/negative scores for a series of attributes, which make up the common ground available for developing comparisons and are of equal value. In Q-study the entities were bacteria and attributes were their cultural and biochemical properties, etc. In normal analysis the entities are quadrats in vegetation and the attributes are incident species. In this latter case the attribute range is never at issue, since there is usually no doubt about which plant species occur in vegetation under study. In the former case, however, the attribute range is not naturally circumscribed, and attributes must be selected from an indeterminate class. Numerous subtleties complicate this and render it controversial (see Sneath 1957*a*). The investigating microbiologist is responsible for the judicious selection not only of the entities, but also of the attributes by which to describe them. The big difference between these methods is in their terminal phases. Q-study compares entities directly on the basis of attributes, but normal analysis first compares attributes on the basis of entities (see Fig. 7). While Q-study then proceeds to a "cluster analysis" by direct sorting of entities, normal analysis sorts to eliminate attribute associations from entity subgroups, and hence sorts entities only incidently.

(iii) *The Indices "Percentage Similarity" and "χ^2" Compared*

The marked difference between these two indices is illustrated by Figure 2, where percentage similarity is plotted against χ, rather than χ^2, in order to distinguish positive from negative associations. Certain features are outstanding. Percentage similarity elevates to prominence species pairs which more often than not fail to exhibit any outstanding significant association. These owe their high similarity of

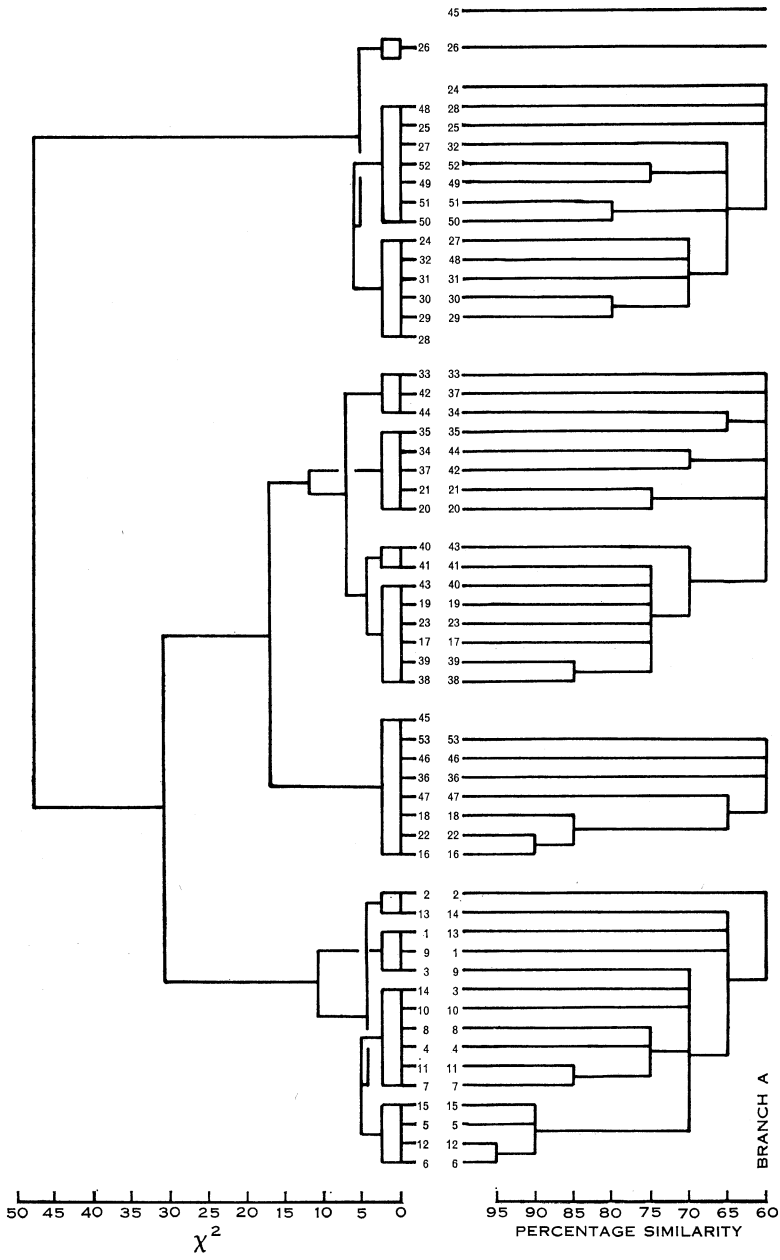

Fig. 6.—Dendrograms expressing the results of experimental Section II(5). The right-hand dendrogram exhibits the results of a Sneath Q-study on some palaeobotanical entities × attributes data, and the left-hand dendrogram shows the results of normal analysis on the same data. The entities are shown to be classified similarly by the two methods.

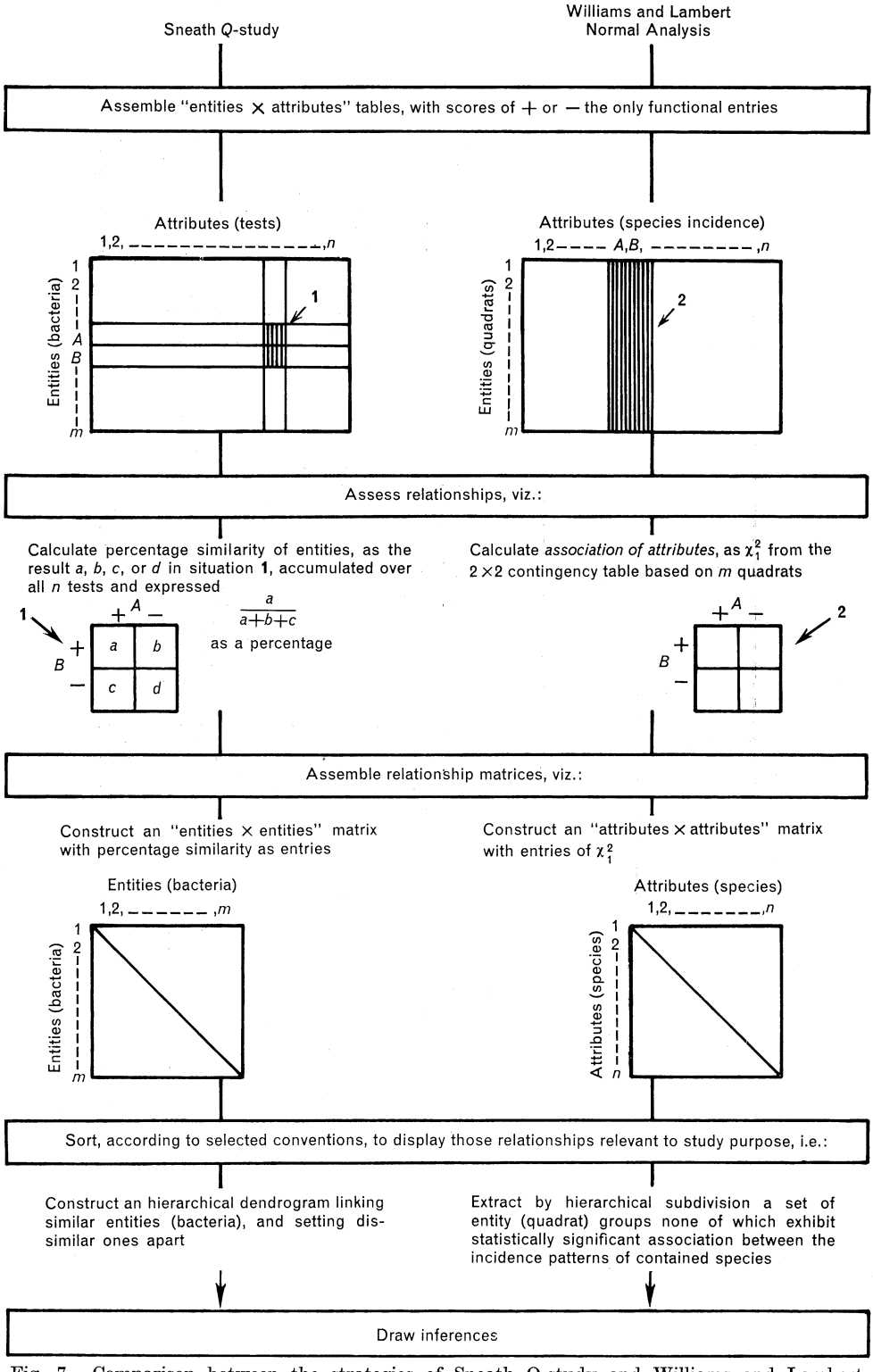

Fig. 7.—Comparison between the strategies of Sneath Q-study and Williams and Lambert normal analysis.

range incidence to their ubiquitous nature, which constrains them to high percentage similarity irrespective of any interaction in the statistical sense. Another feature is that most species pairs attaining very highly significant χ values of positive sign do not have outstanding percentage similarity. Also, Sneath's percentage similarity, by excluding negative associations from consideration, neglects the kind of information provided by the χ values of negative sign, some of which equal in magnitude the most significant of the positive associations. Altogether, Figure 2 provides ample basis for requiring a clear statement of purpose when electing to use either percentage similarity or χ^2, since these index very different attributes of data.

(iv) *Results Compared*

Considering the difference between these techniques both in basic reasoning and in the performances of their operative indices, one would hardly expect them to yield nearly identical results when applied to the same data. Nonetheless, this is shown empirically to be the case (see, for example, Fig. 6). The agreement between the significant aspects of groupings by the two methods is remarkably good. Only in one case in Figure 6 is there much difference. Entity 45, segregated as an independent type by Sneath's polythetic procedure, is located in a group by the monothetic process of normal analysis. Sokal and Sneath (1963, p. 281) reported that they had experience of matching the two methods on the same data, but advanced no further information except that substantial agreement was observed.

Since in general the aims of the two techniques differ, it must be pointed out that while normal analysis corroborates Q-study, there are no clear reasons to expect that it should, so far as rational argument about the biological situation is concerned.

(v) *Nature of Percentage Similarity*

In the matrix of percentage similarity values calculated for 80 plant species on the basis of their incidence patterns in 330 quadrats, attention was directly to features of the top ranking combinations attaining 20% similarity or more. These are the pairs which dictate the significant features of a Sneath dendrogram by virtue of being top rank, and hence determine the nature of group nuclei and the successive high-level cross-links. In this case they constituted the top 7% of the total 3160 scores. Table 1 lists some of these, deliberately selected according to the frequencies with which the species involved were represented in the 330 quadrats.

High percentage similarities in category C are due to the mutual high density of positive scores which pairs of entities possess and are otherwise meaningless unless supported by a significant value of χ, as are the first two pairs. In category A, in which the frequencies of occurrence are extremely low, it is possible to find highly significant associations which may be botanically important for which the percentage similarities are relatively small. It is clear that a percentage similarity should never be considered in isolation but always in relation to its statistical significance and also the numbers from which it was derived.

(vi) *Cluster Analyses Compared*

The course by which data are subdivided in normal analysis is rigorously specified, and is based on ideas of efficient subdivision according to information theory (see Williams and Lambert 1959). The courses by which data are subdivided in

Goodall's methods (1953) are also specified closely, and are different from those of normal analysis. All these methods lead to valid but different classifications of the data, and the important decision for the biologist is which of these classifications best serves his purpose.

While subdivision of data by Q-study can be similarly rigorous, reliance on the shaded similarity triangle for the presentation of results permits interpretation so arbitrary as to render the whole analysis most unattractive. In such presentation

TABLE 1

CHARACTERISTICS OF ENTITY-PAIRS DICTATING THE SIGNIFICANT FEATURES OF A SNEATH DENDROGRAM BY VIRTUE OF ATTAINING THE TOP-RANKING PERCENTAGE SIMILARITIES AND HENCE DETERMINING THE NATURE OF GROUP NUCLEI AND THE MAIN SUCCESSIVE HIGH-LEVEL CROSS-LINKAGES

Entities Compared	Numbers of the 330 Attributes for which they are Positive (frequencies)	Percentage Similarity	χ (to nearest whole number)	Comment
4, 5	2, 3	25	+7	Category A: lowest of the low frequencies observed
70, 58	5, 7	20	+5	
58, 68	7, 9	22	+6	
64, 75	26, 32	22	+5	Category B: some moderate frequencies
7, 63	41, 95	22	+4	
7, 19	41, 72	28	+6	
3, 46	292, 245	72	+3	Category C: highest of the high frequencies observed, approaching those of ubiquitous species
3, 48	292, 265	76	+3	
6, 71	287, 277	74	−1	
3, 71	292, 277	76	0	
3, 6	292, 289	78	0	
	Trend for increase in frequency down this column	Trend for increase in magnitude down this column within the top 7% of cases	Trend for decrease in χ down this column, to non-significant levels in category C	

the upper right-hand part of a matrix is ignored (being a mirror image of the lower left-hand part), values for percentage similarity in the entities × entities matrix are classified into a few class-intervals, and the cells are then shaded in graded intensities corresponding to class-intervals, darkest for highest similarities and lightest for lowest ones. This effectively transposes the contents of the matrix for rapid visual appraisal. An attempt is then made to order the entities so that very similar ones are adjacent, on the dictum (Sneath 1962, p. 307) that the ordered matrix will contain areas of high similarity showing as dark triangles of shaded cells. It is on the appearance of such dark triangles that interpretation is attempted.

Such shaded similarity triangles are unsatisfactory in that they do not reveal the compromises adopted in arriving at them, and unnecessary in that they provide no further information than does the cluster analysis upon which they should be based. At their worst, they may present only the outcome of empirical "juggling", which still occurs in biology (Curtis and McIntosh 1951; Bray 1956; and Anderson 1963). An assumption underlying the whole analysis is that there is some particular pattern of similarities to reveal, which has to be sought and recognized among many possible patterns. Any two-dimensional depiction necessarily must involve distortion and compromise (except in cases trivially simple) and, to provide for rational interpretation, it is necessary to specify closely both the compromises adopted and the manner of introducing them. Sneath (1962) and Sokal and Sneath (1963) illustrate some of these *specified compromises*, termed "cluster analyses" or "nodal analyses". For reasoned interpretation of a cluster analysis, the compromises must be kept sight of and understood. A shaded similarity figure is not of itself a cluster analysis; it may be a summary of the results of one, but is not necessarily so, and of itself does not indicate the compromises adopted in its construction.

A commonly employed convention for interpretable cluster analysis (used by Sneath 1957) is illustrated in Figure 6, right-hand dendrogram. Here a dendrogram is constructed relative to an ordinate scale of percentage similarity. Entity pairs with the highest similarities are first cross-linked at the level of their similarity (or more usually at the lower limit of a convenient class-interval into which they fall) to give a series of groups. Pairs with the next highest similarities (or first appearing in the next lowest class interval) are then cross-linked into further groups, and linked also with pre-existing groups if pre-existing and new groups possess at that level of similarity a common entity. (Linking of groups thus involves an all-or-nothing compromise.) Much of the resulting dendrogram's apparent pattern depends on the conventions adopted for its organization. Thus in Figure 6 that branch of the dendrogram which possesses the highest-level amalgamations is placed such that it is succeeded by groups successively lower in their highest amalgamations. The same order of listing applies within groups. Dendrograms must be interpreted carefully. Members linked into one group may yet have zero similarity. Other biases incorporated into this form of cluster-analysis are discussed below in Section III(*c*).

To qualify for inclusion in the entities × attributes table, an attribute must differentiate at least two entities. Descriptive attributes possessed by all the entities, or not possessed by any of them, are not considered in computations, irrespective of their relevance to wider subsequent comparisons. Hence, while magnitudes ascribed to relationships in any Q-study are correct in a relative sense, their *scale* has no intrinsic significance to the biologist.

(*b*) *Influence of Random Variation*

(i) *Random Data*

The usual features of a Sneath Q-study dendrogram or a Williams and Lambert normal analysis may be obtained by starting with entity × attribute scores generated entirely at random. That is, an hierarchical schema results with a formidable set of divisions and cross-linkages (Fig. 5), which, despite their complexity and the exhibition of very highly significant associations, are quite fortuitous. When therefore

random variation is known to influence appreciably the entries in an entities × attributes table (as it will in biological sampling situations), then the validity of inferences drawn from the dendrogram becomes, to some unspecified extent, questionable. As they stand, the procedures make no provision for delimiting chance effects.

(ii) *Split Data*

Williams (personal communication) suggested that if the original data be split randomly in halves, and each half analysed separately, the extent of agreement between the two sets of results could usefully indicate the stability of the classification. Figure 1 illustrates a comparison made in this manner using data split and then analysed by Sneath's method. These data and their analysis are referred to in Section II(*a*) above. The lack of correspondence between the two classifications at similarity thresholds less than about 40% clearly indicates that the sorting process is without value at lower levels in this instance. Above 40% the agreement of the two halves, although not perfect, is sufficiently good for the acceptance of the classification in regard to the first 17 species sorted.

(iii) *Stability of* χ^2

To examine this instability of classification, interspecific associations between the 80 plant species in the 330 quadrats were each calculated as χ^2, according to each of two random halves of the original data, and these values are plotted in Figure 3. There is obviously a very marked absence of stability in high values of χ^2, on the basis of these data. In fact, the plot indicates absurdities such as interactions exhibiting in one experiment probabilities ranging from 10^{-5} to 10^{-9}, and in a replicate experiment probabilities of more than 0·05. It is clear that this relates to the density of positive scores, for on inspection, interactions exhibiting such variable results are found to be based on species from category A, Table 1. To eliminate erratic values such as these one would first have to cull the data to exclude those entities which are positive for very few of the attributes, or alternatively for nearly all of them. In the biological context this may mean abandoning sparse data, however interesting, to comply with requirements for analysing the bulk. The biologist may not wish to do this, but must appreciate that it is not always possible to reconcile both aspects.

(iv) *Stability of Percentage Similarity*

Percentage similarities were calculated for the same species as in Section III(*b*)(iii) above, using each half of the randomly-split data, and the results plotted in Figure 4. In marked contrast to the high-level χ^2 values, the top-ranking percentage similarities exhibit a very close correlation indeed with the obvious exception of some erratic values associated with low-frequency data. In the light of Figure 4 the difference between the two dendrograms of Figure 1 is not to be attributed to undue instability of percentage similarity, because in the main the agreement between replicates is very good, while the species involved in unstable values are nevertheless placed consistently in the two dendrograms. Instead, the dendrogram convention for cluster analysis is revealed as unduly sensitive, amplifying the effects of relatively small change in a percentage similarity in such a way as to impose large changes in

the emerging pattern of cross-linkages. This undesirable property is unfortunately a consequence of the same all-or-nothing rule for cross-linking which makes the method feasible for cluster analysis in the first place.

(*c*) *Inference where Random Variation does not Influence the Data*

In the study on extant and fossil vegetative shoots, entries in the table compiled for analysis were not in any way influenced by random processes since each of the entities was an abstract type (a botanical species or a palaeobotanical form species), the scored attributes of which were invariable by definition. The validity of recognizing such entities is not in question here, but the nature of overall relationships between them is. The raw table of entities × attributes is mentally indigestible, and it is to digest such a large block of scores that recourse is reasonably taken to Sneath's method. The dendrogram summarizing the results (Fig. 6, right-hand part) is of special interest for a number of reasons. Because of the convention whereby groups are linked at any particular level on the basis of sharing a single common entity, they will often contain paired entities possessing mutual similarity much lower than the similarity level of the group which contains them. This detracts considerably from group significance since the bias towards linking groups on the least available evidence inevitably coalesces some nodes of relationship which are essentially separate. However, there is no such qualification on the interpretation of the schisms between groups; these signify that at the similarity levels through which they extend there is absolutely no similarity between the groups they separate.

With this in mind, the right-hand dendrogram of Figure 6 has considerable biological significance. Down to the 60% level, above which there is obviously ample scope for the similarity patterns to emerge, branch A is absolutely unrelated to any of the other branches; branch A is the only group not containing a fossil type, and branch A contains exclusively all of the *Callitris* and *Actinostrobus* species examined. That is, all of the architectural types of extant Australian conifer shoots are represented in the fossil floras of Arcoona Plateau except the callitro–actinostroboid architecture. This permits the surprising inference that on the basis of this evidence plants were present in these assemblages resembling, in shoot architecture, all of the extant Australian conifer groups *except* those conifers for which contemporary Australia is particularly noted, viz: endemic species of *Callitris* and *Actinostrobus*.

(*d*) *Conclusion*

Sneath Q-study and Williams and Lambert normal analysis embody two of the best contemporary approaches to empirical classification. Nevertheless they have various deficiencies which the rational biologist can ignore only at the risk of making faulty inferences. *Caveat emptor*!

IV. Acknowledgment

The authors are extremely grateful to Professor W. T. Williams for his suggestion [see Sections III(*a*)(i) and III(*b*)(ii)], which has led to an irrefutable case for terminating sorting procedures before or upon reaching "noise level". Criteria for terminating sorting in a single set of data are currently under investigation by the authors.

V. References

ANDERSON, D. J. (1963).—The structure of some upland plant communities in Caernarvonshire. III. The continuum analysis. *J. Ecol.* **51**: 403–14.

BEERS, R. J., FISHER, J., MEGRAW, S., and LOCKHART, W. R. (1962).—A comparison of methods for computer taxonomy. *J. Gen. Microbiol.* **28**: 641–52.

BRAY, J. R. (1956).—A study of the mutual occurrence of plant species. *Ecology* **37**: 21–8.

CERBON, J., and BOJALIL, L. F. (1961).—Physiological relationships of rapidly growing mycobacteria. Adansonian classification. *J. Gen. Microbiol.* **25**: 7–15.

CHEESEMAN, G. C., and BERRIDGE, N. J. (1959).—The differentiation of bacterial species by paper chromatography. VII. The use of electronic computation for the objective assessment of chromatographic results. *J. Appl. Bact.* **22**: 307–16.

COLWELL, R. R., and LISTON, J. (1961*a*).—Taxonomic relationships among the pseudomonads. *J. Bact.* **82**: 1–14.

COLWELL, R. R., and LISTON, J. (1961*b*).—Taxonomy of *Xanthomonas* and *Pseudomonas*. *Nature, Lond.* **191**: 617–19.

CURTIS, J. T., and MCINTOSH, R. P. (1951).—An upland continuum in the prairie–forest border of Wisconsin. *Ecology* **32**: 476–96.

DAVIS, P. H., and HEYWOOD, V. H. (1963).—"Principles of Angiosperm Taxonomy." (Oliver and Boyd: Edinburgh and London.)

FLOODGATE, G. D., and HAYES, P. R. (1963).—The Adansonian taxonomy of some yellow-pigmented marine bacteria. *J. Gen. Microbiol.* **30**: 237–44.

GOODALL, D. W. (1953).—Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. *Aust. J. Bot.* **1**: 39–63.

GRAHAM, P. H. (1964).—The application of computer techniques to the taxonomy of the root-nodule bacteria of legumes. *J. Gen. Microbiol.* **35**: 511–17.

GRIEG-SMITH, P. (1963).—"Quantitative Plant Ecology." (Butterworths: London.)

HILL, L. R. (1959).—The Adansonian classification of staphylococci. *J. Gen. Microbiol.* **20**: 277–83.

HILL, L. R., TURRI, M., GILARDI, E., and SILVESTRI, L. G. (1961).—Quantitative methods in the systematics of actinomycetales. II. *Giorn. Microbiol.* **9**: 56–72.

LAMBERT, J. M., and DALE, M. B. (1964).—The use of statistics in phytosociology. *Adv. Ecol. Res.* **2**: 59–99.

LISTON, J. (1960).—Some results of a computer analysis of strains of *Pseudomonas* and *Achromobacter*, and other organisms. *J. Appl. Bact.* **23**: 391–4.

LYSENKO, O., and SNEATH, P. H. A. (1959).—The use of models in bacterial classification. *J. Gen. Microbiol.* **20**: 284–90.

POHJA, M. S., and GYLLENBERG, H. G. (1962).—Numerical taxonomy of micrococci of fermented meat origin. *J. Appl. Bact.* **25**: 341–51.

PROCTOR, J. R., and KENDRICK, W. B. (1963).—Unequal weighting in numerical taxonomy. *Nature, Lond.* **197**: 716–17.

ROGERS, D. J., and FLEMING, H. (1964).—A computer programme for classifying plants. II. A numerical handling of non-numerical data. *BioScience* **14**: 15–28.

ROGERS, D. J., and TANIMOTO, T. T. (1960).—A computer programme for classifying plants. *Science* **132**: 1115–18.

SNEATH, P. H. A. (1957*a*).—Some thoughts on bacterial classification. *J. Gen. Microbiol.* **17**: 184–200.

SNEATH, P. H. A. (1957*b*).—The application of computers to taxonomy. *J. Gen. Microbiol.* **17**: 201–26.

SNEATH, P. H. A. (1962).—The construction of taxonomic groups. 12th Symp. Soc. Gen. Microbiol. pp. 289–332.

SNEATH, P. H. A. (1964*a*).—New approaches to bacterial taxonomy: use of computers. *A. R. Microbiol.* **18**: 335–346.

SNEATH, P. H. A. (1964*b*).—Computers in bacterial classification. *Advmt. Sci., Lond.* **20**: 572–82.

SNEATH, P. H. A., and COWAN, S. T. (1958).—An electro-taxonomic survey of bacteria. *J. Gen. Microbiol.* **19**: 551–65.

SOKAL, R. R., and MICHENER, A. A. (1958).—A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **38**: 1409–38.

SOKAL, R. R., and ROHLF, F. J. (1962).—The comparison of dendrograms by objective methods. *Taxonomy* **11**: 33–40.

SOKAL, R. R., and SNEATH, P. H. A. (1963).—"Principles of Numerical Taxonomy." (W. H. Freeman: San Francisco and London.)

TUOMIKOSKI, R. (1942).—Untersuchungen über die Untervegetation der Bruchmoore in Ostfinnland. I. Zur Methodik der pflanzensoziologischen Systematik. *Ann. Bot. Vanamo* **17**: 1–203.

WILLIAMS, W. T., and LAMBERT, J. M. (1959).—Multivariate methods in plant ecology. I. Association-analysis in plant communities. *J. Ecol.* **47**: 83–101.

WILLIAMS, W. T., and 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., and LAMBERT, J. M. (1961).—Multivariate methods in plant ecology. III. Inverse association analysis. *J. Ecol.* **49**: 717–29.