Fundamental Problems in Numerical Taxonomy

W. T. WILLIAMS and M. B. DALE

Department of Botany, The University, Southampton, England

I. Introduction

In any field of endeavour which transgresses the boundary between fundamental and applied disciplines there tend to be two alternative approaches: the user's approach, "What do I wish to do, and how can it best be done?" and the more fundamental "What can most efficiently be done, and what can it be used for?" The approaches are more different than is commonly realized, and both are necessary. These reflections are prompted by the appearance of the first major text-book devoted to numerical taxonomy, that due to Sokal and Sneath (1964). This will provide an admirable introduction for those botanists wishing to enter this rapidly developing field; and it is no denigration of this important work to suggest that the authors are less rigorous in their examination of the methods than they are in their use and interpreta-

tion, for it is with these latter aspects that they are primarily concerned. The user's interests in plant ecology are similarly met by Greig-Smith (1964) and, in a more limited context, by an article which to some extent complements our own (Lambert and Dale, 1964). Excellent bibliographies have been provided for taxonomy by Sokal and Sneath (1964) and for ecology by Goodall (1962) and Greig-Smith (1964).

Our intention is different. The newcomer to this field is faced with a formidable diversity of methods, all apparently fulfilling closely similar functions. It is nevertheless our contention that the number of fundamentally distinct methods is very small, and that criteria can be erected which will clarify the distinctions between them, and between their numerous variants. This is the aim of this communication. We shall not be concerned with the problem of allocation to an existing classification, which is the province of discriminant analysis.

Although all the methods we shall discuss are in principle applicable to botanical problems, few have yet been so applied; our references will therefore of necessity be drawn from a wide variety of disciplines. Symbols used will be conventional; but in the 2×2 contingency table arising from the possession (J,K) or lack (j,k) of two attributes J and K, two conventions now exist for the number of individuals in each class: the alternatives are set out below:

	$+J$	$-J$	
$+K$	a	b	(K)
$-K$	c	d	(k)
	(J)	(j)	N

(*i*)

	$+J$	$-J$	
$+K$	n_{JK}	n_{jK}	n_K
$-K$	n_{Jk}	n_{jk}	n_k
	n_J	n_j	n

(*ii*)

Scheme (*i*) is older, and has long been used in elementary statistical texts; scheme (*ii*) is used by Sokal and Sneath. The latter is more informative, but is clumsy in algebraic expressions and in our experience is easily misread. When such a table is at issue, we shall therefore adhere to the (a,b,c,d) convention.

II. The Nature and Properties of Classifications

A. THE BASIC AXIOMS

Most general discussions on classification are concerned to define, and to distinguish between, existing types of classification; such, for example, are the discussions in Lawrence (1951), Beckner (1959), Gilmour (1951) and Sokal and Sneath (1964). It is important for our purposes, however, to establish the minimum requirements which all classifications must meet, and we restate the problem as follows. A population consists of elements, each of which can be individually described by reference to a predetermined list of "relevant characteristics". This population is subdivided into sets of elements. what requirements must be fulfilled by these sets for the sub-division to rank as a classification? We submit that the following axioms will suffice.

(1) Within every many-membered set there must be, for every member of the set, at least one other member with which it shares at least one relevant characteristic.
(2) Membership of the set may not itself be a relevant characteristic.
(3) Every member of any one set must differ in at least one relevant characteristic from every member of every other set.

Axiom (1) introduces a concept of "likeness" and ensures that an element cannot be classified if nothing is known about it. Axiom (2) has two important consequences. First, division into groups defined solely as possessing a stated number of members (such as dividing a population into groups of ten, or dividing it equally into eight parts) is excluded; secondly, all classifications must be open-ended—there may be no *known* members to add to a set, but it must not be impossible by definition to add more. Axiom (3) not only ensures that identicals cannot be distributed between different sets, but makes provision for the single-membered set.

Although these axioms will suffice to define a classification, they are not in general sufficient to define one which is useful. We therefore need to discover what additional constraints must be imposed to enable our classification to meet specific external requirements, and it is from this point of view that we now proceed to examine some of the basic problems in numerical taxonomy.

B. MONOTHETIC AND POLYTHETIC CLASSIFICATIONS

These terms were introduced by Sneath (1962) to replace Beckner's (1959) terms "monotypic" and "polytypic" (without changing Beckner's definitions), since these terms have other meanings. The sets in a mono-

thetic classification are completely defined by the presence or absence of specific characteristics. Since such classifications are always generated in practice by successive sub-division, it follows that there must always be at least one set *all* of whose members share at least one relevant characteristic. It is quite possible to construct a population, classifiable by reference to the axioms, from which no such set can be extracted; in such a case monothetic classification is impossible. Monothetic classifications may nevertheless be useful. They have proved valuable in ecology, where the concept of "indicator species" has long been familiar; they may well be needed in criminology, in which a decision may have to be taken quickly and based on as few attributes as possible. They are normally unacceptable in taxonomy; in medical taxonomy, for instance, one does not wish a man to be treated for the wrong disease because he has one aberrant symptom. The occasional criticism that monothetic systems produce misclassification is, however, invalid, since the criticism automatically assumes that a polythetic system is desired, and the argument is circular. The real objection to monothetic classifications is that they assume a property of the population which it may not in fact possess. Polythetic classifications imply no properties beyond those involved in the basic axioms, and are therefore always possible.

C. MAXIMIZATION

1. Principles of maximization

The basic axioms will serve to define a large number of alternative classifications, and a further constraint is needed to select from among these. The constraint universally required by users is that, in a sense yet to be defined, the members of any one set are to be as alike as possible and as unlike the members of other sets as possible. Differences within sets are to be minimized, differences between sets are to be maximized. Formal work in this field, usually loosely known as "maximization", has been largely confined to discriminant situations, particularly in the field of pattern recognition (*vide*, e.g. Sebestyen, 1962); but the diverse methods of numerical taxonomy are simply variant methods of maximization.

The methods fall into two fundamentally distinct groups.

i. Self-structuring methods

(*a*) A function of the relevant characteristics is defined between pairs of elements.

(*b*) An element may be either a member of a population or an entire set; if a set, then the set may be defined by one of its members, by all

of its members, or by an element constructed from all of its members.
(*c*) Sets are to be constructed so that the function is minimum (or maximum) within them, maximum (or minimum) between them, or both.

ii. Derived-structuring methods

(*a*) A function is defined between pairs of relevant characteristics over a given set of members.
(*b*) A characteristic, or a group of characteristics, is found for which the function, or a derivative of the function, is maximal.
(*c*) Sets of members are defined in relation to the characteristic(s) so selected.

For certain purposes it is desirable that the analysis can be "inverted", in the sense that the elements and characteristics change places. For this to be possible the data must fulfil certain conditions which we explore later (Section IV A). The apparent diversity of methods in the literature largely concerns self-structuring methods, and in these the diversity is largely one of the function selected. Monothetic methods necessarily employ derived-structuring.

2. Internal and external classifications

It is assumed in the foregoing paragraph that the members as defined by their relevant characteristics form a self-sufficient set within which maximization is to be effected; such systems, which comprise almost the whole of existing literature in numerical taxonomy, we shall call "internal" classifications. It may nevertheless be desired to impose a restraint in the form of an external element or set of elements (self-structuring) or an external characteristic or set of characteristics (derived-structuring). In such cases the maximization is entirely between the reference unit on the one hand and the internal sets on the other, the internal sets needing only to satisfy the basic classificatory axioms. The process of maximization is, however, itself different from the all-internal case. The primary maximization is of the *range* of the selected function, in that the internal sets are to be as like *or* unlike as possible to the reference set.

The main use of these "external" classifications is likely to be predictive; if the population is heterogeneous in the sense we shall define in Section IV C, they will be more powerful than the classical regressions taken over the whole population. Their possible application to problems in plant ecology is also under investigation. The only example known to us in the literature is the derived-structure "predictive attribute analysis" of Macnaughton-Smith (1963), with whom we are currently collaborating in the development of more general systems.

3. Simultaneous alternative classifications: clumps

Suppose the system be restricted by the requirements (*i*) that maximization shall extract only one sub-set from the population, and (*ii*) that this sub-set shall be subject to a specified constraint; the constraint normally imposed is that the subset must contain a specified element or group of elements which will act as its nucleus. Such a sub-set is normally termed a "clump" and the remainder of the population is without interest. Let this process be successivly repeated on the entire population by specifying a new constraint on each occasion; the ultimate result is a set of clumps. This set is sometimes loosely termed an "overlapping classification", but such an extension of the term "classification" is not to be recommended; the clumps need not exhaust the population, and any one element can, and usually does, occur in more than one clump. Systems of this type are particularly associated with the work of Needham (*vide*, e.g. Needham, 1962; Needham and Jones, 1964) on linguistic data arising from problems in documentation and information retrieval; but they have also found some application in anthropology and medicine (Bonner, 1964). They have been developed to meet circumstances in which simplicity and speed of computation are more important than power, and they may well require re-examination before they can satisfy the more rigorous demands of plant taxonomy and ecology.

A system of clumps can similarly be generated by the use of a changing external criterion as constraint. The groups delimited by the "deme" terminology (Gilmour and Heslop-Harrison, 1954) of plant taxonomy together form a system of precisely this nature, but it seems never to have been the subject of numerical study. We shall not be further concerned with clump systems in this article.

4. Weighting

Sokal and Sneath (1964) accept the Adansonian postulate that "every character is of equal weight". We need not so restrict ourselves, and we shall first distinguish between *a priori* and *a posteriori* importance.

i. Importance a priori

Classifications in, for example, medical or criminological contexts may be used as guides to action; in such cases particular characteristics may be of overriding importance. It might be regarded as undesirable to send epileptics to prison, no matter what their other characteristics suggested. Such cases do not disturb the systems we are considering, since they do not alter the classifications, but only the use that is made of them. It has, however, frequently been suggested (*vide*, e.g.

Proctor and Kendrick, 1963) that characteristics should be assigned a differential importance from prior knowledge of the field; as we have already pointed out (Williams *et al.*, 1964) this destroys the objectivity which is the single most valuable feature of numerical taxonomy, and we cannot recommend it.

ii. Importance a posteriori

Derived-structuring methods maximize some function of the characteristics. After maximization, therefore, each characteristic will be associated with a numerical value which reflects its contribution to the overall maximization, and which may therefore be regarded as a measure of its importance. The application of this concept presents different problems in different systems, and the situation may best be explored by consideration of firstly, monothetic derived-structure, and secondly, polythetic self-structure.

(*a*) *Monothetic derived-structure.* If a population is such that it contains many shared characteristics and so can be defined as a set of final classes, a very large number of alternative monothetic classifications is possible. The characteristics used may be selected solely for external convenience, or even indiscriminately, and there is no internal maximization. Such are the "special classifications" (into, e.g., food- or fibre-plants) and the dichotomous keys in floras. These, which are in fact perfectly good external classifications, are commonly termed "artificial". It is therefore tempting to equate "artificial" with "absence of internal maximization"; but we defer to the views of Sneath (*in litt.*) to the effect that the terms "natural" and "artificial" have been so variously used that to provide them with new statistical definitions would confuse rather than clarify the situation.

In contrast to these classifications, the method of Association Analysis, whose properties are discussed in Section V D *2*, is a monothetic method whose defining characteristics have been obtained by a process of internal maximization. The characteristics now differ in *a posteriori* importance, and this has by some workers been regarded as "weighting".

(*b*) *Polythetic self-structure.* Here again it is theoretically possible to effect classification without maximization, but since most real-life populations already themselves satisfy our Axiom (1) for a classification, the solution is usually trivial. A single maximization is therefore necessary in practice. All the "similarity" methods discussed in Sokal and Sneath (1964) are of this type: they use the least maximization which is in practice essential. However, the first step in such an analysis might be a derived-structure maximization, so that the characteristics were as a first step provided with "importance" measures; a second

maximization, using these weighted characteristics, would be necessary to complete the classification. The only such doubly-maximized methods known to us are those in whose development we have ourselves collaborated (Macnaughton-Smith *et al.*, 1964; Williams *et al.*, 1964).

5. Unintentional weighting: "nuisance correlations"

The selection of attributes is not itself the concern of numerical taxonomy. Nevertheless, methods which employ derived-structure functions are prone to difficulties arising from so-called "nuisance correlations"—groups of attributes linked for reasons unconnected with the purpose of the analysis. This problem does not arise in those ecological studies in which the attributes are plant species, since these are necessarily different things. The questionnaires of sociological studies, however, normally contain much redundant information; this is deliberate, since a question which may be avoided in one form may be answered readily in another. Some of the attributes are therefore logically linked, and these linked groups may dominate the subsequent analysis. It must be remembered that questionnaires have not normally been designed with modern numerical methods in mind, and the increasing use of these methods will doubtless in time influence the design of questionnaires; but meanwhile the problem exists. The nature of the problem, however, has not always been clearly understood. The objection to these links is simply that they can be known to be links without recourse to analysis; if they could not be so known they would be of interest. It does not follow that they are in every case easily recognized, and a preliminary numerical analysis may serve to establish them. This is possible if the system is such that elements and characteristics can change places, so that the characteristics can be grouped into sets; if such a set inescapably suggests the hypothesis that the members are linked for reasons—such as intrinsic redundancy in a questionnaire —in which the investigator is not interested, the group can be replaced by one or more of its members or by a new attribute constructed from all of them. Despite statements to the contrary in the literature, we submit that the objection to nuisance correlations does not lie in their logically necessary links; the sole criterion is the interest or otherwise of the user.

D. HIERARCHICAL AND NON-HIERARCHICAL CLASSIFICATIONS

Hierarchical classifications are of very real advantage to the taxonomist, since they enable him to compare taxa at any desired level. This has probably contributed to the fact that the vast majority of existing numerical methods are hierarchical in nature. However, it

may also generate a requirement that each level in division is associated with some measure which shall fall as the hierarchy descends. It is not always realized that this places an additional constraint on the choice of maximizing function; some functions (notably Euclidean distances and information statistics) possess this property, whereas others (most of the derived-structure coefficients and the "statistical distance" coefficients) do not.

The term "reticulate classification" seems to include two quite different concepts. The first is the unmaximized external classification with an embarrassing choice of alternatives, such as arises in classifying books; this need not concern us. Truly reticulate classifications arise out of an interest in inter-set relationships after division into sets has been completed. If only the inter-set functions are required, a completely non-hierarchical method could be used; but, as shall later point out, the choice of such methods is extremely restricted. In most cases, therefore, both the hierarchy and the inter-set functions are of interest, and the problem is to generate either from the other.

We shall later demonstrate that maximization may, or may not, be uniform over the entire mathematical model in use. If it is uniform, as with unweighted Euclidean distances or information statistics, no difficulty arises: inter-set functions and hierarchical divisions are everywhere compatible. In those methods with which we ourselves have been associated, the maximization is deliberately non-uniform over the model; in these cases, which are hierarchical, no compatible inter-set function has yet been defined (*vide* Sections III D (*i*) (*ii*)). It is not permissible to define a completely new function, since the original hierarchical maximization may then fail; this is the cause of the "recombination of sets" difficulty which Goodall (1953a) experienced in his pioneer studies in divisive methods.

E. PROBABILISTIC AND NON-PROBABILISTIC CLASSIFICATIONS

This particular dichotomy has generated more confusion—and probably more rancour—than any other. It underlines the commonly-expressed doubts as to whether these methods can, or cannot, be classed as statistics, and so has caused Greig-Smith (1964) to use the term "quantitative" and Sokal and Sneath (1964) and ourselves to fall back on "numerical". It underlies, too, the misgivings that authors frequently express concerning the "significance" of their results. The difficulty has been exacerbated by the fact that modern statistics is almost entirely concerned with estimates of probability, so that if well-known statistical parameters—χ^2 or the correlation coefficient, for example—are used for maximizing, it is assumed that these are

estimates which should be associated with measures of probability. In fact, the methods of numerical taxonomy are not, or need not be, probabilistic systems at all, but hypothesis-generating systems. We shall outline the two alternative approaches.

1. The non-probabilistic approach

From this point of view, the methods of numerical taxonomy may be regarded as stemming from a branch of statistics of respectable antiquity—that concerned with finding mathematical formulations which will serve as a concise and economical description of an otherwise intractably cumbersome mass of data. Though superficially so dissimilar, their logical relatives are to be found among such projects as the fitting of Pearson curves to actuarial data (Elderton, 1938); the search for a flexible growth-curve (Richards, 1959; Nelder, 1961); and the application of contagious Poisson distributions to distributions of plants in the field (Archibald, 1948). The fitting of regression lines is itself a member of the same family, extended by the probabilistic concept of the significance of the parameters which the fitting requires.

Now, these concise mathematical descriptions can with perfect validity be used to generate hypotheses concerning the nature of the data, but only if two conditions are rigidly satisfied. First, as always, the hypotheses must be capable of being tested; secondly, any test must depend on new observations, and cannot again use the data from which the hypothesis was generated. Generation of the hypothesis may not be used as its own evidence; we forbear to cite examples of this practice, contenting ourselves by remarking that they can be found in biological literature.

The precise statistical context of these methods can most clearly be demonstrated by comparing a vegetation survey in ecology with an agronomic experiment in, say, mineral nutrition. In the agronomic context the hypothesis is set from previous experience, and this determines the details of an experiment, which issues in a quantity of data; statistical methods are applied to these data in order to test the hypothesis—usually in the form of the probability of obtaining a given deviation from a null hypothesis by chance alone. In the ecological context, although experience may have informed its collection, the data is the starting-point; functions are selected and appropriately maximized in order to reduce the data to simpler form; this simpler form is used to generate a hypothesis—often in the form of "there is a change of some sort in this region"; and the hypothesis is tested by new, direct observations in the field. Examples of this type of hypothesis-validation may be found in the work on Association Analysis (Williams and Lambert, 1960).

Nor is validation difficult in applied taxonomy. In medicine, for example, the individuals classified may be disease-producing organisms, or symptoms; in criminology they are normally delinquents. In these cases the hypothesis takes the form of a suggestion for treatment. It may be remarked in passing that the power of the mathematical methods used is all-important in these fields, for although falsification of a hypothesis might gratify a dispassionate experimenter, it is apt to be disastrous if a human individual is concerned.

The problem is more difficult in "classical" taxonomy. Here it is tempting to enunciate a phylogenetic hypothesis, normally based on inter-set functions, but fossil records are such that hypotheses of this type are rarely testable (*vide*, e.g. Sneath and Sokal, 1962). The basic requirement of taxonomy *sensu stricto* is stability, both of the membership of sets and of the pattern of characteristics that their members display within them. In the first case (the membership of sets), addition of new characteristics followed by re-maximization should not change the membership of the sets. In the second (the pattern of characteristics), let a new element be discovered whose characteristics are imperfectly known; if from the known characteristics it can be allocated unequivocally to an existing set, the pattern of its remaining characteristics, when these are examined, should conform to the pattern for the set.

On this approach, therefore, the methods of numerical taxonomy are hypothesis-generating systems; and a hypothesis-generating system is neither valid nor invalid. Probability enters only, if indeed it enters at all, in the testing of the hypotheses that are generated. This approach exposes a possible danger, which we do not believe taxonometric writing has always avoided. This is that computer classifications might be regarded as in some sense absolute—as "objective" and therefore "better". They are not objective, since they depend on the user's personal choice of maximizing function; and they are only better if they can be shown to fulfil a stated requirement more efficiently.

2. *The probabilistic approach*

It is, as we shall show, easy to conceive of probabilistic classifications in theory; but we are here concerned to defend the thesis that such classifications are usually both impracticable and unprofitable. First, it should be noted that a probabilistic classification requires a null hypothesis; this will normally take the form of stating that the pair-functions available for maximization in a given population or set could have been generated by a random process. The null hypothesis cannot, in fact, be independent of the function selected for maximization.

i. Difficulties inherent in null hypotheses

Since the null hypothesis depends on the maximizing function, it will be convenient to select two well-known cases for consideration.

(*a*) *Multivariate normal populations.* In this case the null hypothesis would state that the observed variation in characteristics could have been generated by a set of independent normal variates, usually the characteristics themselves. The function available for test would probably be the correlation matrix. Now, Bartlett's (1950) test for the roots is not available if the matrix is singular, and experience suggests that it is sensitive to departures from normality. To demand that all the coefficients be individually significant is normally regarded as too stringent; and Goodall (1953a) has in effect suggested that the coefficients be themselves treated as normal deviates, so that the proportion of them which exceeds the individual significance level be regarded as a test of significance of the whole matrix. There is, in fact, no simple, unequivocal and robust test available.

(*b*) *Qualitative populations.* The function used (though others are available) is often related to the Euclidean distance between elements (or set centroids) plotted in an n-dimensional space where the jth co-ordinate for an element is 1 if it possesses the jth attribute and 0 if it lacks it. The problem now is to state a null hypothesis at all. Use of the binomial expansion would imply that possession of all characteristics was equally likely; and the solution obtained by Rohlf (1962) for even n makes assumptions as to the distribution of the frequencies. If we assume, however, that the hypothesis should not involve the frequencies, an obvious solution would be to retain the frequency totals and to construct from them the entirely dissociated class-frequencies; that is, the numbers of individuals that would be required in all possible sub-classes if, without change in the total numbers possessing each attribute, all pairs of attributes were to have zero association. It is straightforward, though tedious, so to calculate the probabilities (for 0, 1, $\sqrt{2}$) in the two-characteristic case; but the resulting algebraic expressions are extremely cumbersome, and lend little hope of extension. In any case, construction of the general null population may present formidable difficulties. If we write (A) for the number of individuals possessing attribute A, (AB) for the number possessing both A and B, and so on, then in the completely dissociated population

$$\frac{(ABC\ldots.)}{N} = \frac{(A)}{N}\cdot\frac{(B)}{N}\cdot\frac{(C)}{N}\cdot\ldots.$$

Unfortunately, for more than two characteristics, this relationship is necessary but not sufficient (*vide*, e.g. Yule and Kendall, 1950), and cannot therefore be used as a generating function.

It has been suggested to us (Macnaughton-Smith, *in litt.*) that information statistics might provide a solution of the qualitative problem, in view of their remarkable additive properties and their relationship to χ^2. Let a group of n individuals be specified by the possession or lack of p attributes, and let the number possessing the jth attribute be a_j; we have made preliminary observations, using ecological data, on the behaviour of the statistic:

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

Ecological data not uncommonly contain groups of identical or near-identical individuals, and these groups may vary greatly in size; the data will have the properties of a stratified, rather than of a random, sample. Unfortunately, we find that the statistic above is sensitive to this particular form of non-randomness, and is therefore unduly sensitive to set size—sets tend to be fused if they contain comparable numbers of members. This is incompatible with our second classificatory axiom, since it implies that a function of the set may determine the allocation of one of its members. This difficulty may be removed by normalizing for group size, though, in some forms of analysis, at the expense of replacing it by the generation of an "ambiguity" problem related to that arising from unweighted Euclidean distances (Section V D *3* (*i*)). Nevertheless, these statistics have many desirable properties and would repay further investigation.

(*c*) *Goodall's coefficient.* Very recently Goodall (1964) has proposed a probabilistic similarity index. For every pair of individuals, the probability that the two are as similar as in fact they are is calculated for each attribute separately, and the attribute-probabilities then combined. The method is cumbersome for qualitative data, but it is the only method known to us which is in principle applicable to mixed data i.e. data in which the attributes are so unlike that any common scaling would be unrealistic. No example of its use has yet been published.

ii. Application of probabilistic classifications

Suppose an appropriate criterion of significance, and therefore an appropriate null hypothesis, to be available; and suppose a population to have been divided by maximization into two sets whose distinction fails to reach significance. It still does not follow that the division should not be effected. For the population may be so intractably large that the best possible sub-division, though non-significant, may be more useful than none at all. However, although the overall characteristic-pattern may not define a significant difference, sub-sets of characteris-

tics may exhibit stability (this phenomenon may occur if the population exhibits *noda*, which are briefly discussed in Section IV C 4). In either case, it is the usefulness of the division which will be of importance; the division will therefore in any case be subjected by the user to a second, pragmatic, test which will override the first, probabilistic, test. We are therefore not convinced that any useful purpose is served by the probabilistic test, quite apart from its inherent difficulties.

III. The Choice of Mathematical Model

A. Introduction: metrics

The ultimate test of a numerical method is whether the user finds it useful. However, all methods are of *some* use to the user; and if he is to bear the sole responsibility of deciding between them, he will be faced with an immense amount of empirical work, still with no assurance that the method may not fail under extreme conditions—as, we believe, some existing "similarity" methods have already failed. The literature contains many despondent remarks on the paucity of available information relating to comparison of methods. This is particularly true of the pair-functions themselves, often loosely classed as "similarity coefficients". The best-known have been reviewed by Goodman and Kruskal (1954, 1959), Dagnelie (1960) and Sokal and Sneath (1964); but it is doubtful whether even these extensive collections are complete. The problems would be relatively unimportant if all such functions were jointly monotonic, in the sense that, if element-pairs are so ordered that one function forms a monotonic series (i.e. a series which either increases or decreases over the whole of its length), the remainder will also be monotonic. To take only three well-known functions, $2a/(2a+b+c)$, $(a+d)/(a+b+c+d)$, and the correlation coefficient, it is easily shown that no one of these is jointly monotonic with either of the others. A choice is therefore necessary; and the testing difficulty can be overcome, at least in part, if the methods and functions are required to fulfil appropriate mathematical conditions.

We consider it essential that any measure used for maximization should define a model, and, if possible, a model in Euclidean space. The advantages of such systems are threefold. First, many simple, robust and powerful methods are available in Euclidean systems that are not available outside them. Secondly, as mentioned in Section II D, they have hierarchical advantages. Thirdly, and perhaps most important, our daily experience gives us an intuitive perception of Euclidean systems, and thereby enables us to grasp their properties and to predict these properties in extreme cases. If the function is such that it is not known

to be associated with any particular probabilistic or spatial model (models which are neither probabilistic nor spatial are possible, but we know of no published work on them) we propose that it must, as a minimum requirement, be a metric; it will then necessarily define a space whose properties can be explored. We deal in this section with the general problem of metrics, and it will be convenient first to state the conventional definitions. The subject is fully discussed in geometrical texts; the formulation we use is substantially that of Kelley (1955).

Definition. A numerical function $d(x,y)$ of pairs of points of a set E is said to be a metric for E if it satisfies these conditions:

(1) $d(x,y) = d(y,x) \geqslant 0$ (symmetry)
(2) $d(x,z) \leqslant d(x,y)+d(y,z)$ (triangle inequality)
(3) if $d(x,y) = 0$ then $x = y$ (distinguishability of non-identicals)
(4) $d(x,x) = 0$ (indistinguishability of identicals).

A system in which (3) is not everywhere true we shall call semi-metric ("pseudometric" of some writers); a system in which (2) is not everywhere true we shall call quasi-metric.

In the context of numerical taxonomy, metric properties may fail for two reasons: the characteristics may be intrinsically metric but the pair-function selected is not; or only a pair-function exists, and this is not metric. We deal with these cases in order.

B. METRIC PROPERTIES OF PAIR-FUNCTIONS

It would be unprofitable to examine the properties of all the many functions in the literature, and we shall largely confine our observations to the three most familiar: $(a+d)/(a+b+c+d)$; $2a/(2a+b+c)$; and the correlation coefficient.

1. $(a+d)/(a+b+c+d)$

This is the coefficient normally used by Sneath. Using the model mentioned in Section II E 2(*i*)b, and following Sokal and Sneath in writing Δ for the Euclidean distance between the two points, the coefficient is equal to $(1-\Delta^2/N)$. It is based on a Euclidean metric and satisfies the requirements we have suggested.

2. $2a/(2a+b+c)$

This coefficient is probably among the oldest in the literature; it specifies the ratio between the number of characteristics common to two elements and the arithmetic mean of the numbers possessed by each. It is monotonic with the coefficient $a/(a+b+c)$, used by Sneath for the purpose of excluding double-negative matches; the intention

here, in the context of our model, is to prevent points being grouped solely because they are near the origin. Enumeration of the 3-characteristic case for either coefficient will immediately demonstrate that the coefficients are quasi-metric; they are also necessarily semi-metric, and thus do not satisfy our requirements on either count. It is of interest to note that Sokal and Sneath (1964) no longer recommend the exclusion of double negatives.

The earlier coefficient can be used to generate a different model. If, in the ordinary 2×2 table, we have $(2a+b+c) \neq N$, there must be $|a-d|$ double-positive or double-negative matches; there may be more but there cannot be less. If the coefficient is written in the form:

$$1 - \frac{\Delta^2}{2a+b+c}$$

it may be regarded as derived from a distance whose dimensions are scaled to remove all logically-necessary matches. The dimensions of the model now change from place to place, so that the model is non-Euclidean; it could be topologically embedded in a Euclidean space of not more than $2N$ dimensions, but we are not ourselves competent to explore the utility of this approach.

3. The correlation coefficient

Several alternative models have been suggested for this function, only one of which fulfils our requirements. First, it should be noted that of the four requirements for a metric, simple unsigned derivatives of this coefficient (such as $(1-r)$) fail to satisfy requirements (2) and (3), and cannot therefore be handled in this way (its semi-metric properties are of value if "shape" coefficients are required—*vide* Rohlf and Sokal, 1963). A model commonly used in factor analysis, however (*vide*, e.g. Cattell, 1952), supposes the points to be rigidly attached to their co-ordinate axes by extensible perpendiculars. If the axes are now rotated about the origin until all correlations are zero, the final angles between them will be the inverse cosines of the original coefficients. These angles between pairs of lines now serve to define a Euclidean space with oblique axes; providing this model is in use, our requirements are therefore satisfied.

4. Asymmetric functions

Goodall (1953b), in the course of an examination into the phyto-sociological concept of "fidelity", has suggested that asymmetric functions would be of value in this context. Several such functions are in fact on record in the literature, although only Goodall appears to have appreciated their nature and possible application. So far as we are

aware, no practicable strategy for the maximization of asymmetric functions has ever been suggested or even sought; until this is done, further investigation of their properties will remain unprofitable.

C. INTRINSICALLY NON-METRIC SYSTEMS

In the previous section it has been assumed that all characteristics are, or can be regarded as being, measurable; failure of metricity is due only to the calculation of a measure which is not a metric. We are here concerned with two problems of greater fundamental difficulty; first, the case in which individual characteristics, though they exist, cannot be provided with a simple measure; and secondly, the case in which characteristics do not exist, though a pair-function between elements does.

1. Individual characteristics

This situation arises when all that can be measured in respect of a given characteristic is some comparison between two elements—commonly in the form of a difference or a ratio. It is then necessary to operate on these comparisons in such a way as to generate a metric which will uniquely order the elements along that characteristic considered as a dimension. This problem, usually known as "scaling", is of great importance in psychometric work, and has given rise to an extensive literature; the recent communication by Phillips (1963) will serve as an introduction to the field. We know of no botanical work of this type; but since comparative measures are not unknown in taxonomic descriptions, the methods may yet prove applicable, and botanists should be aware of their existence.

2. Isolated pair-functions

Consider a sociological study in which has been recorded, for the members of each pair of individuals in a group, the number of times they met each other in a given period; all that is available for analysis is a pair-function. Functions of this type are often semi-metric (some pairs of individuals never meet) and are almost always quasi-metric. The problem is to generate a Euclidean system of hypothetical characteristics such that the distances between elements shall be related to the original pair-functions. A solution is provided by the "proximity analysis" of Shepard (1962a,b). This generates a system of co-ordinates, of the lowest order which will permit a unique solution, such that the Euclidean distances between the elements are monotonic with the original pair-functions. The solution is iterative, and a computer

program exists. Again, we know of no published botanical applications, but the method might conceivably be of interest in competition studies where the records took the form of the number of times pairs of species were in contact.

D. NON-EUCLIDEAN SYSTEMS

1. Introduction

A Euclidean space is necessarily metric, but the converse is not true. It will be convenient to begin with a conventional definition:

A Euclidean space of order n is the set of all n-tuples $(x_1 \ldots x_i \ldots x_n)$ where all x_i are real numbers and where the distance between two points is given by $[d(x,y)]^2 = \sum_{i=1}^{n}(x_i - y_i)^2$. It can be shown that such distances are metrics.

There are three obvious ways in which Euclidean properties may be lost. Firstly, n itself may not be constant, so that the dimensions vary locally; this is the situation for the coefficient discussed in Section III B *2* above. Secondly, the n-tuples may be constrained in some way, e.g. to the surface of a sphere; we know of no application in numerical taxonomy. Thirdly, the distance function may fail; in the cases we shall consider, the distance-function holds within sets, but fails between some or all of them. The space defined is thus *locally* Euclidean, and hence (if varying continuously) Riemannian; and no difficulty arises unless inter-set functions are required. We shall discuss briefly three methods in which this type of problem arises.

2. Examples

(*i*) *Attempts to use "statistical distance"*

The group of statistics of which the Mahalanobis D^2 is the best-known is essentially probabilistic in concept; it relates inter-set distance to a common within-set function, and involves the postulation of a common dispersion matrix for the two sets. If the sets are manifestly unlike, this is an unrealistic assumption; and if such a matrix is artificially constructed, it may well be singular. It is not uncommon (cf. Harberd, 1962) to postulate that the common dispersion matrix is an identity matrix, and to regard the Euclidean distance so calculated as a derivative of the Mahalanobis statistic; we ourselves believe this to represent an unrealistic model, and consider that, as suggested by Kendall (1957), a Riemannian metric is needed.

ii. Weighted polythetic subdivision

This method (Macnaughton-Smith *et al.*, 1964) employs a Euclidean model with axes scaled by *a posteriori* importance measures. As in the previous case, the scaling depends on the dispersion (or correlation) matrix, but it is the individual axes which are affected. New scales are calculated before each sub-division; as a result, any two sets derived by sub-division of a single set share the same metric, but this is not true of set-pairs in general. The final model resembles a Riemannian system in being locally Euclidean; but the space is now divided into blocks with the local metric changing abruptly at the boundaries, and may best be described as a "disjoint metric space". Although models of this general type have received some attention from topologists, we have been unable to trace any work on the metrization of such a space. The difficulty is exacerbated by the fact that the space remains undefined between sets.

iii. Association Analysis

Since this is a pure derived-structure method, no measure of inter-set distance arises naturally from the maximization. Again, however, the axes vary in *a posteriori* importance from set to set, and a Euclidean metric would be unrealistic.

E. CONCLUSIONS

We may now conveniently classify the acceptable coefficients under three headings:

1. Information statistics

These can be maximized over the whole model; as a result, they automatically provide inter-set functions and progressively-falling hierarchy measures. Their relationship with χ^2 permits them to be used in a probabilistic context. If our misgivings (Section II E *2(i)b*) as to their dependence on set-size prove to be unfounded, or can be overcome, they will be very attractive; but more work is needed.

2. Euclidean distances

These, too, can be maximized over the whole model, and provide inter-set functions and hierarchical measures with the desired properties. They seem likely to be probabilistically intractable, but we have given reasons (Section II E *2(ii)*) for regarding this as relatively unimportant. Compared with the doubly-maximized coefficients they appear to lack power, especially in populations defined by small numbers of characteristics.

3. Riemannian and disjoint-space functions

It is our opinion that these provide the most realistic models and the most powerful methods for classification; but work on inter-set functions is badly needed. For the biologist (including the present authors) the mathematics required for such work is out of reach; but the difficulties are not entirely mathematical. In the probabilistic case, since the dispersion matrices are known to be different, what is the null hypothesis which is to be tested? In methods using sub-division with changing weights, what properties is an inter-set function required to possess? Before the geometers can be expected to collaborate, the users must be prepared to consider these questions.

IV. The Basic Euclidean Model

A. Duality: the R/Q problem

We have already given reasons for our preference of a Euclidean model; this is not incompatible with essentially Riemannian or disjoint models, since the region of space undergoing maximization is always locally Euclidean. We therefore now consider the problems that arise in setting up such a model. We begin with n elements (which we shall henceforth call *individuals*) specified by p characteristics (which we shall henceforth call *attributes*, a term we use in an extended sense to include variables and variates). Provided all attributes can be given values, either inherently or by the methods of generation outlined in Section III C, the system is symmetrical; the data-matrix can be transposed so that the individuals and attributes exchange status. Two models immediately present themselves: a set of n points in a p-space, or a set of p points in an n-space.

This duality has given rise to the symbols R and Q. Unfortunately, two mutally incompatible traditions as to the definition of these symbols exist side-by-side in the literature; and this confusion—to which we have ourselves contributed—must now be resolved. The early workers in factor analysis commonly refer to a model in which the individuals are points imagined as in a space specified by co-ordinate axes representing attributes: this is our n points in a p-space that we shall for the moment call an attribute-space. The arithmetical operations were carried out on a matrix of correlations (or occasionally covariances) between attributes. Such a method was called an R-method. Later, the entire process was transposed for certain purposes; the model is now p points in an n-space that we shall for the moment call an individual-space, and the arithmetical operations were carried out on a matrix of correlations between individuals. This was a Q-method.

This now is the problem: do *R* and *Q* refer to the model or the matrix? In the early work there was no obvious ambiguity; but consider a matrix of Euclidean distances between individuals. Since it is a matrix between individuals, it is *Q*; but since it is based on a model in the attribute-space, it is *R*.

Sokal and Sneath (1964) define the symbols unequivocally by the matrix, and the individual-distance matrix is, for them, *Q*; but in all publications from our laboratory we have defined the symbols by the model, and the individual-distance matrix has been, for us, *R*. We have now decided that the Sokal-Sneath definition should prevail, for two reasons. First, the widespread circulation that their book will deservedly attain will ensure that many taxonomists previously unfamiliar with the symbols will first meet them in the Sokal-Sneath sense; and to attempt to assert a rival definition would cause unjustifiable confusion. Secondly, there is some historical precedent. Most early work obtained approximate solutions for principal axes by the "centroid" method. Although this operates arithmetically on an attribute-correlation matrix, it is based on a model in the individual-space; but it has always been known as *R*, though by the Southampton definition it would be *Q*. We suggest, then, that *R* and *Q* refer to the matrix; but it will still be convenient to have symbols for the model, and we suggest the symbols *A* (for a model in the attribute-space) and *I* (for a model in the individual-space).

It seems likely that the indecision frequently expressed concerning the relative merits of *R* and *Q*-methods stems partly from inadequate understanding of the models implied, and we believe that the introduction of the new symbols will clarify the situation. A matrix of inter-individual distances and an inter-individual correlation matrix are both *Q*; but the former implies relationships between points in an *A*-space, the latter between angles in an *I*-space. An attribute-correlation matrix is *R*, and an individual-distance matrix is *Q*; but both are *A*-space models, the first concerned with angles and the second with points. In fact, two *Q*-methods may require models which differ from each other more fundamentally than do some *R*/*Q* pairs.

B. ADJUSTMENTS TO THE MODEL

Virtually all numerical methods involve difficulties concerned with the dimensions of physical quantities. Only in truly qualitative data do these difficulties not arise; whatever the nature of the quantities which have been dichotomized, addition of either rows or columns of the data-matrix is interpretable in terms either of the number of individuals possessing an attribute or of the number of attributes possessed by an

individual. If the data is quantitative and not all in the same units, addition of attributes across a single individual is not technically possible. This difficulty arises immediately in Euclidean distances or principal components in the A-space, and in correlation coefficients ("between persons") in the I-space. It is discernible, too, in the tendency to regard principal components as "taking out" a proportion of a variance constructed by the illegitimate addition of separate variances. It is the invariable, and inevitable, convention in numerical taxonomy to regard all attributes as dimensionless, and hence available for arithmetical manipulation; but the highly autocratic nature of this convention must be clearly realized.

Methods which involve the addition of different attributes are not, in general, invariant under changes in scale of the co-ordinate axes. The initial scaling of axes is thus irrevocable, and will in a sense determine the results of the analysis. Euclidean distances are very sensitive to the scales of the axes, but independent of the position of the origin; principal components are very sensitive to both. In the case of Euclidean distances we have, we believe (Macnaughton-Smith *et al.*, 1964; Williams *et al.*, 1964), turned this sensitivity to scale to advantage, though at the expense of ending the analysis with a disjoint-space model. There tend to be two schools of thought concerning principal components, those who leave the variances unchanged and those who standardize them all to unity; one advantage of such standardization is that it renders the variates genuinely dimensionless. It would be equally permissible to standardize the variates by "importance" measures as in the case of our scaled distances; this might conceivably increase the power of the method, but it has never been tried. In factor analysis it is usual to rescale the factors to unit variance after their extraction. The position of the origin is a far more intractable problem, since in highly heterogeneous data there is no obvious "best" place for its location. It is traditional to take the decision appropriate to the multivariate normal distribution and locate the origin at the common mean; we ourselves have no better solution to offer.

C. HETEROGENEITY

1. Introduction

If material is presented for classification, it must be suspected of being heterogeneous in some way. In the context of our model, this heterogeneity may take two forms. In the first, all attributes may be meaningful for, and measurable on, all individuals; but attributes, either singly or in linked groups, may be markedly polymodal. In the

A-space model, the points form discrete galaxies (we discuss the possibility of non-galactic heterogeneity below). In the second, the attributes may, again either singly or in groups, become zero. If only the zero or non-zero nature is at stake (qualitative data) there is no difficulty; the difficulty arises when measurable attributes are sometimes zero. For the concept of "zero" embraces two quite distinct concepts—that which happens to be zero and that which must be zero: the number of hairs on the third pair of legs *may* be zero in an insect, but it *must* be zero in man. In many cases the pattern of zeros and non-zeros is itself of primary importance.

2. The data-classes

Since data may be homogeneous, or heterogeneous in either or both of two ways, we find it convenient to distinguish four classes of data.

Class 1. Co-ordinates measurable on all axes; no sub-populations everywhere zero on sub-sets of axes; distributions substantially unimodal on all axes.

Class 2. Co-ordinates measurable on all axes; no sub-populations everywhere zero on sub-sets of axes; polymodal on at least some axes, the points forming galaxies in the A-space.

Class 3. Qualitative data, the co-ordinates taking only the values 0 or 1; sub-populations exist which are everywhere zero on sub-sets of axes.

Class 4. Co-ordinates measurable on all axes; sub-populations exist which are everywhere zero on sub-sets of axes; distributions on the non-zero axes may be polymodal.

Class 1, of course, approaches the multivariate normal distribution, and is of no interest in classificatory problems. Class 2 data is the raw material of taxonomy, so long as the individuals are known to be closely similar. Since random sampling of individuals of widely disparate nature would be pointless for taxonomic purposes, this requirement is normally fulfilled. Class 3 is the familiar "presence-or-absence" data of the ecologist; perhaps because of the modesty of its mathematical demands, it has been extensively studied by "biological" biometricians. Moreover, owing to the relative ease with which such data can be analysed, it is frequently generated from Class 4 data by dichotomizing the variates. It has, however, been pointed out to us by Macnaughton-Smith (*in litt.*) that this raises a new difficulty. In ecology the 1/0 situation is truly asymmetrical, in that only the presences carry useful information, but this is not true in sociology; if 1 is taken to represent drunkenness, 0 represents sobriety, and both are meaningful. It

remains only to say that Class 4 data are normal in sociology, and in ecology if a measured attribute (such as percentage cover) is used.

3. Transposition of data-classes

The data-classes are not necessarily invariant under transposition. Classes 1 and 2 may become inter-converted under A/I transposition; so may Classes 3 and 4, with certain limitations. The nature of the data may thus appear to change markedly when transposed. We believe that this is one cause of the prevailing uncertainty regarding R/Q differences: an R/Q difference is inherently likely to be greater if it also involves an A/I difference.

4. Noda

The term "nodum" was apparently introduced by Poore (1955) in the context of phytosociology; its numerical implications have so far been examined only in the case of monothetic classifications of Class 3 data (Williams and Lambert, 1961a; Lambert and Williams, 1962). This concept is, however, most easily illustrated in Class 2. Consider points in a 3-space, disposed within two elliptical cylinders whose long axes are parallel to the Z-axis. The projection on the (X, Y)-plane will show two sharply-defined galaxies; projections on the other two planes will show no strikingly galactic structure, and may not even separate the two cylinders. A nodum, in our definition, is an enumeration both of a set of points and of the set of axes in which the points constitute a galaxy or "cluster". In Class 3 data, a nodum consists of an enumeration of a set of individuals and of a set of attributes for which they are substantially all non-zero.

Noda may be regarded as foci around which the population is varying; they are potentially of great value as a basis for shedding peripheral information. Unfortunately, no general method of extracting them is yet known. We now incline to the view that the solution provided by Williams and Lambert (1961a) is open to objections; it is in any case applicable only to monothetic situations. One of us (Dale, 1964) has carried out a preliminary investigation into the characterization and combinatorial properties of noda, but the problem is as yet far from solution.

The difficulty is more fundamental than may appear at first sight. The existing method involves setting up the two models (n points in p-space and p points in n-space) and collating the results; but the concept of a nodum as "central" information intrinsically requires that the individuals and attributes be manipulated simultaneously. This is impossible so long as either is regarded as a set of points in a space defined by co-ordinate axes of the other. Despite our advocacy

of a Euclidean model, and despite its incontestable power, the search for nodal techniques may yet force us to abandon spatial models and seek methods of maximizing some function of a data-matrix which shall be symmetrical as regards rows and columns.

5. Non-galactic heterogeneity

In Class 2 data, the points need not cluster into galaxies; they might, for example, be dispersed along intertwined filaments, or on the surfaces of concentric spheres. We are not aware that any such data have been reported (except, of course, in ecological situations where pattern on the ground is at issue); but if it were to be suspected a particular strategy of analysis is indicated (*vide* Section V D *1*(*iii*) below).

V. Strategy of Analysis

A. Simplification methods

By "simplification" we intend some means of reducing the dimensions of the original Euclidean model, so that the data can be displayed in a small number of dimensions with the minimum loss of information. The process may fulfil any of three quite distinct functions, though these are seldom clearly distinguished in the literature.

1. Subjective classification of complex data

A taxonomist may legitimately not wish to invoke objective numerical methods, preferring for some specific purpose to utilize his own knowledge and experience to delimit taxonomically intractable material. The data may nevertheless be specified by too many attributes for the taxonomist to handle confidently; the requirement is to find transformations of the original attributes which can be graphed in two or three dimensions. Principal component analysis is commonly used for this purpose, but is intrinsically liable to produce a dilemma. If the dispersion matrix is used, the data is in no way distorted; but if any of the attributes have appreciably higher variance than the remainder, these attributes will dominate the analysis, so providing information which could have been more simply obtained by univariate inspection. If, on the other hand, the correlation matrix is used—as it normally is—the data being classified is not the original data. The Hotelling solution commonly given in text-books involves two successive standardizations to unit variance—first of the attributes, then of the components—and so still further distorts the original data.

Factor analysis has occasionally been used for the same purpose, but the elements being so classified are further removed again from those

specified by the original data. However, the case reported by Pettett (1960), of a population of *Viola* spp. which showed marked discontinuity on the first factor but not on the first component, is potentially of great interest, and would merit further investigation.

Relatively crude approximations to such methods exist in the literature; the method of Curtis (1959), for example, can be regarded as an approximation to a component analysis with the origin outside the population, at the point of intersection of the tangent-planes perpendicular to the axes of the hyper-ellipsoid. The coefficient used is the quantitative counterpart of $2a/(2a+b+c)$ and is thus non-metric. Such methods were unavoidable while computing facilities were limited; now that fast programmes exist for the calculation of large dispersion or correlation matrices, and for the extraction of their roots and vectors, there is little point even in such approximations as the centroid solution.

2. Preliminary evaluation of data

A non-probablistic approach in practice necessarily assumes that there is heterogeneity to be found; but it may well be desired to explore the general configuration of this heterogeneity, whether or no it is everywhere sharply-defined, and whether or no it is galactic. Unless the data is exceptionally complex, the first two or three principal components will normally provide the information required.

3. Generation of "underlying factor" hypotheses

It may be desired to erect hypotheses more far-reaching than those [Section II E *1*] which are purely classificatory; such hypotheses normally take the form of postulating the existence of a small number of underlying "factors" which would be sufficient to generate the interrelationships within the entire numerical system under study. It is natural to explore the simplest possible hypotheses—i.e. those that can be contained in the smallest number of postulated factors—with due regard to Kendall's (1957) warning that "this seems to assume on Nature's part a much more indulgent behaviour than we have any right to expect". If classical methods are in use—as distinct from those maximum likelihood methods (Lawley and Maxwell, 1963) which do not require rotation—the appropriate solution is iterated communalities, double standardization and rotation to simple structure. Rotation normally aims at providing the simplest possible relationship between old and new axes; but it may instead be required to seek the simplest possible relationship between individuals and new axes—the solutions are not necessarily identical. In cases of extreme heterogeneity, Dale (1964) has shown that there may be no factor-analytic solution: there may be no real values of the communalities which will substantially

reduce the order of the matrix, and the centroid iteration of communalities may fail to converge.

These methods are currently out of favour, probably as a result of the incautious claims which have in the past been made for them. They do not "reveal" or "demonstrate" any structure in the data, and we deprecate the tendency to "identify" the factors which are extracted. If they are regarded purely as hypothesis-generating systems their use is unexceptionable, and they are potentially of great power.

B. PARTITION

Attributes may be such that not only are they present or absent (and the pattern of presences or absences important) but they may also, if present, be measurable. Such situations are more common than is usually realized. In particular, the data of plant ecology are of this type if a measure such as percentage cover is in use. They arise in pure taxonomy if, for given types of specimens (such as herbarium specimens), some attributes cannot be observed; and they arise in sociology if parts of a questionnaire are not answered. Essentially, the data is of Class 4, and the primary need is to ascertain whether the major heterogeneity is qualitative or quantitative. We have suggested elsewhere (Williams and Dale, 1962) a method by which this may be effected, but the computation is heavy and no computer programme exists at present. The method consists essentially of a partition into qualitative and quantitative elements; it can be extended without difficulty to the threefold system (known/unknown): (if known, present/absent): (if known and present, then measured). We incline to the opinion that Class 4 data should normally be partitioned—i.e. separated into Class 2 and Class 3 elements—before numerical analysis; but the methods of subsequent analysis will require modification from their normal forms, and no investigation of this kind has yet been attempted.

C. NON-HIERARCHICAL METHODS

The most familiar non-hierarchical method is that which uses canonical variates (Rao, 1952) for the comparison of groups of individuals. Recent examples are mainly zoological, though the method has been used successfully on *Populus*, *Betula* and *Ulmus* spp. by J.M.R. Jeffers (personal communication) and his collaborators. Like all methods related to the Mahalanobis D^2 statistic, it is not applicable to individuals or to groups not known *a priori* to be sufficiently similar to share a common within-group dispersion matrix, and its detailed consideration is therefore outside the scope of this article.

Component analysis and factor analysis are non-hierarchical, but are normally made the basis of subjective classification: completely objective non-hierarchical methods, such as the "multi-dimensional group analysis" under development by R. Jancey (personal communication), seem to be extremely rare. We have already pointed out (Section II D) that hierarchical classifications are commonly regarded as desirable by the users, and it is presumably for this reason that they dominate the literature. Despite their intrinsic theoretical interest, we incline to the view that non-hierarchical methods are of limited value in numerical taxonomy. An exception should perhaps be made for the method associated with Tanimoto: but this, though non-hierarchical, is closely related to certain hierarchical systems, and it will be more convenient to defer its consideration to the section which follows.

D. HIERARCHICAL METHODS

1. General considerations

i. Pairs

Hierarchical methods are completely dominated by the concept of all possible pairs of points or of axes. There is no difficulty in conceiving methods based on all possible triangles or tetrahedra of points, or all possible solid angles. We know of no work of this type. It would involve far more computation than do the pair-systems, and until it is certain that all possible power has been extracted from such systems, it is doubtful whether more complex methods are worth pursuing.

ii. Direction

The analysis may either begin with the entire population and progressively break it down (divisive methods), or begin with the individuals and progressively fuse them (agglomerative methods). The relative advantages and disadvantages are best discussed in connexion with specific systems; it is only necessary here to point out that, if a hierarchical classification is required, monothetic methods cannot (except in a trivial sense) be agglomerative; after the first groups have been formed there may be no attribute by which they can be fused.

iii. Sorting

The problem here may be reduced to that of defining a distance between a point and a set of points, and is particularly relevant to the agglomerative methods. Three methods are in use. In the first, the distance is defined as that between the point and the nearest member of the set ("nearest-neighbour" sorting). Since this uses only a small

part of the available information concerning the set, the method is normally regarded as lacking in power. It is, however, computationally very economical, requiring the calculation of $\frac{1}{2}n(n-1)$ distances, and it is the only form of sorting which will elucidate non-galactic heterogeneity. In the second method, the distance is defined as the average of all the distances between the point and the individual members of the set. It requires the calculation of $(n-1)^2$ distances or averages of distances, but demands complex sorting procedures to use the calculations economically. It is nevertheless the only method which has regard to set density. Lastly, the set may be represented by the coordinates of its centroid. This also requires $(n-1)^2$ calculations, but the computational strategy is very simple; it is our opinion that the simplicity and elegance of strategy that this method allows conclusively justifies its use.

2. Monothetic sub-divisive: "association analysis"

For detailed accounts of the use of the method, see Williams and Lance (1958), Williams and Lambert (1959, 1960, 1961b). It uses derived-structure maximization; χ^2_{jk} is calculated between every pair of attributes j and k (in terms of the number of individuals possessing or lacking them singly or jointly) and the sum $\sum_{k \neq j} \chi^2_{jk}$ is formed of all the χ^2 which involve a particular attribute j. Sub-division is on the attribute for which $\sum_{k \neq j} \chi^2_{jk}$ is maximum. Since for the 2×2 case $\chi^2 = Nr^2$, the parameter may be regarded as $\sum_{k \neq j} r^2_{jk}$; in this form the method is possibly applicable to quantitatively-specified data (*vide* Dale, 1964, for a method of sub-division on a quantitative variable) but no work of this type has yet been undertaken. The original investigations in fact used $\sum_{k \neq j} |r_{jk}|$ as sub-division-parameter; but private communications from H. Stein (using a multiple-regression model) and from P. Macnaughton-Smith (using an information-theory model) have independently demonstrated that Σr^2 is the efficient parameter if the greatest reduction of residual variance is required. Lawley (*in litt.*) has pointed out that, as originally suggested, $\Sigma|r|$ may be regarded as a crude approximation to a factor analysis (using averoid communalities), and thus may perhaps be treated as a monothetic approximation to an essentially polythetic system. Empirical trials on ecological data have suggested that $\Sigma|r|$ has in fact certain advantages: in particular it is less sensitive to the presence of "outlying" individuals, whose innate similarities it may preserve. The more efficient Σr^2 tends to split off outlying indivi-

duals as single-membered sets, thus fragmenting the analysis. Further comparative tests on different types of data are desirable. The method has now been used in a variety of contexts and appears robust in that it is not unduly sensitive to occasional errors in transcription of data. It is, however, extremely sensitive to "nuisance correlations" as defined in Section II C *4*; because of the large contributions that such correlations can make to Σr^2, they are intrinsically liable to dominate the analysis.

The largest individual χ^2_{jk} has been used as a measure of "rank" for each successive sub-division, and Williams and Lambert (1960) give reasons for not using more sophisticated parameters. The measure is nevertheless unsatisfactory, since it does not necessarily fall with the hierarchy; this is particularly troublesome at the lower levels of sub-division. We have some reason to believe that $\sum_{k \neq j} \chi^2_{jk}$ would be a better measure, and we propose to subject this possibility to empirical test.

3. Polythetic agglomerative: "similarity" analyses

Most of the published accounts of such methods use parameters which we consider unsatisfactory for reasons given in Section III B, often combined with inevitable but relatively inefficient hand-sorting; these methods need no critical examination. We shall also exclude information statistics and the Goodall probablistic coefficient, since no fully developed methods are yet in use. With these provisos, there are currently only three genuinely distinct methods, associated respectively with Sneath, with Tanimoto and with ourselves. We consider these in turn.

i. Sneath: unweighted methods

References: Sneath (1957); Sneath and Cowan (1958); Sokal and Sneath (1964). The earlier work used the quasimetric coefficient $a/(a+b+c)$, though in his more recent writing Sneath, like ourselves, inclines towards the fully metric $(a+d)/(a+b+c+d)$; the earlier work also used nearest-neighbour sorting as a computational convenience, though here also Sneath concedes the greater power of group-sorting techniques when computational facilities are available. The important feature of his methods is the strict adherence to the Adansonian postulate that, unless there is some special reason for so doing, all attributes are equal and should not be weighted. A difficulty immediately arises if only few attributes are available or if many of the attributes are lacked or possessed by nearly all the individuals: the intrinsic information content per individual is so low that it is impossible to specify the "best" fusion at any stage. Sneath has always made it

clear that his methods are not applicable to such data, and stresses that the number of attributes used should not be less than about 40. Despite the undoubted successes that his methods have achieved with suitable data, we believe that this limitation is a severe and undesirable restriction on the wide application of the method.

It is clear that this restriction can be overcome if further information can be imported into the system at the individual level, which will necessarily involve some form of weighting. On the assumption that *a priori* importance measures are undesirable, the only remaining source of information is contained in such *a posteriori* measures as can be obtained from the population as a whole. The remaining two methods offer different solutions to this problem.

ii. Tanimoto: weighted individuals

Tanimoto (1958); Rogers and Tanimoto (1960). This method in fact uses a quasimetric coefficient, but this is not important in the present context. The coefficients are summed for all individuals, thus providing an *a posteriori* importance measure for each individual; the individuals with the highest values are used as "apices" for beginning the aggomerative process. Unfortunately, the existing sorting process is non-hierarchical and involves decisions on the part of the operator, and, as Sokal and Sneath (1964) point out, the increasing tendency to separate operator from computer renders "steered" programmes undesirable. Despite the early successes of the method, its sorting strategy requires revision; if its concept of information-importing can be combined with the use of a fully-metric coefficient and a mechanical (and preferably hierarchical) sorting system, the method, already of great intrinsic interest, might be a widely applicable strategy of considerable power.

iii. Williams et al.: *weighted attributes*

Williams, Dale and Macnaughton-Smith (1964). This method uses Euclidean distances in an A-space with axes permanently scaled by the *a posteriori* importance measure of Association Analysis, i.e. $\sum_{k \neq j} \chi^2_{jk}$. Its successful classification of a 6-attribute population demonstrates that it is free from the attribute limitations of Sneath's unweighted method. Its chief demerit is the use of invariant weights: the analysis is necessarily dominated by what may loosely be regarded as "first-factor" relationships.

4. Polythetic divisive

Several authors (*vide*, e.g., Roscigno and Maccaccaro, 1960; Cochran

and Hopkins, 1961; Macnaughton-Smith *et al.*, 1964) have considered the general problem of finding what is in some sense the best of all possible alternative sub-divisions; but the only practical method known to us is that of Edwards (1963). This calculates between/within partitions of Euclidean distances for all possible sub-divisions into two groups. Since there are $2^{n-1}-1$ such sub-divisions, the method is necessarily limited to a small number of individuals; and since the distances are unweighted, some difficulties due to ambiguity may be expected at low levels of division.

If the number of individuals is to be increased to realistic proportions, some form of "directed search" is inevitable. The Macnaughton-Smith *et al.* "dissimilarity analysis" finds, in accordance with a stated criterion, the individual least representative of the population as a whole; this individual is then used as the basis for a sub-population, and the remaining individuals allocated sequentially to this sub-population or to the remainder of the population. The "objectively-weighted Euclidean distance" of Section V D *3(iii)* above has been used as criterion in the preliminary trials. The method has two advantages. First, the computation required, though still considerable, is considerably less than is required for an "all possible sub-divisions" method. Secondly, the weights for the axes can be recalculated for each successive sub-division, thus removing the "first-factor" dependence of the corresponding agglomerative method. The present criterion has the disadvantage of defining a disjoint-space model.

5. General conclusions

If a monothetic classification is desired, association analysis is clearly indicated; if a monothetic classification is acceptable, and if $p \ll n$ (as is often the case), the computation required is less than for other methods, and association analysis is again indicated. If a polythetic classification is essential, a sub-divisive method which will provide the major discontinuities at the beginning of analysis is obviously preferable; we can only say that "dissimilarity analysis" shows considerable promise, though further development and experience is necessary before it can be unreservedly recommended.

It is in the agglomerative field that we find ourselves most at variance with current practice. We believe that the completely unweighted methods lack power, and are only suitable where very sharply defined heterogeneities exist; we suspect that the "cloudy clusters" stigmatized in a recent Aslib discussion (1962, p. 258) as "a criticism not of the method, but of the material" may yet be found to be due to using a method of insufficient power. Furthermore, we consider that weights calculated internally from the data contravene only the letter, and not

the spirit, of the Adansonian postulates: Adanson could hardly have foreseen the possibility of internal weighting.

It will not have escaped notice that we incline towards the use of methods in whose developments we have ourselves been concerned. This is inevitable, since had we not been dissatisfied with existing methods we should not have been led to devise our own. It does not follow that we are right: when all the programmes are freely available, the users will decide.

Acknowledgments

We are indebted to many mathematicians and statisticians for their patient assistance, among whom we particularly wish to acknowledge the help of Mr. P. Macnaughton-Smith, Dr. F. Rhodes and Mrs. N. Wilson; but these must not be held responsible for any heretical views we may have expressed. We are also greatly indebted to Dr. P. H. A. Sneath for enabling the senior author to read Sokal and Sneath's book in proof. The article incorporates material arising from work carried out by one of us (M.B.D.) during the tenure of a D.S.I.R. Studentship.

References

Archibald, E. E. A. (1948). *Ann. Bot. N.S.* **12**, 221.
ASLIB Conference on Classification (1962). *Aslib Proc.* **14**, 215.
Bartlett, M. S. (1950). *B.J.P. Statist.* **3**, 77.
Beckner, M. (1959). "The Biological Way of Thought". Columbia University Press, New York.
Bonner, R E. (1964). *I.B.M. J.* **8**, 22.
Cattell, R. B. (1952). "Factor Analysis". Harper & Bros., New York.
Cochran, G. and Hopkins, C. E. (1961). *Biometrics* **17**, 10.
Curtis, J. T. (1959). "The Vegetation of Wisconsin". University of Wisconsin Press.
Dagnelie, P. (1960). *Bull. Serv. Carte phytogeogr. B* **5**, 7.
Dale, M. B. (1964). Ph.D. Thesis, University of Southampton.
Edwards, A. W. F. (1963). *5th Int. Conf. Biometrics* (Cambridge, 1963).
Elderton, W. P. (1938). "Frequency Curves and Correlation", 3rd ed. Cambridge.
Gilmour, J. S. L. (1951). *Nature, Lond.* **168**, 400.
Gilmour, J. S. L. and Heslop-Harrison, J. (1954). *Genetica* **27**, 147.
Goodall, D. W. (1953a). *Aust. J. Bot.* **1**, 39.
Goodall, D. W. (1953b). *Aust. J. Bot.* **1**, 434.
Goodall, D. W. (1962). *Excerpta bot.* B **4**, 16.
Goodall, D. W. (1964). *Nature, Lond.* 203. 1098.
Goodman, L. A. and Kruskal, W. H. (1954). *J. Amer. statist. Ass.* **49**, 732.
Goodman, L. A. and Kruskal, W. H. (1959). *J. Amer. statist. Ass.* **54**, 123.
Greig-Smith, P. (1964). "Quantitative Plant Ecology". Butterworths, London.
Harberd, D. J. (1962). *J. Ecol.* **50**, 1.

Kelley, J. L. (1955). "General Topology". Van Nostrand, New York.
Kendall, M. G. (1957). "A Course in Multivariate Analysis". Griffin, London.
Lambert, J. M. and Dale, M. B. (1964). "Advances in Ecological Research," Vol. 2. Academic Press, London.
Lambert, J. M. and Williams, W. T. (1962). *J. Ecol.* **50**, 775.
Lawley, D. N. and Maxwell, E. A. (1963). "Factor Analysis as a Statistical Method". Butterworths, London.
Lawrence, G. H. M. (1951). "Taxonomy of Vascular Plants". Macmillan, New York.
Macnaughton-Smith, P. (1963). *Biometrics* **19**, 364.
Macnaughton-Smith, P., Williams, W. T., Dale, M. B. and Mockett, L. G. (1964). *Nature, Lond.*
Needham, R. M. (1962). *In* "Information Processing 1962", p. 284. North Holland Publishing Co. Amsterdam.
Needham, R. M. and Jones, K. S. (1964). *J. Document.* **20**, 5.
Nelder, J. A. (1961). *Biometrics* **17**, 89.
Pettett, A. (1960). Ph.D. Thesis, University of Southampton.
Phillips, J. P. N. (1963). *Nature, Lond.* **200**, 1347.
Poore, M. E. D. (1955). *J. Ecol.* **43**, 226, 245, 606.
Proctor, J. R. and Kendrick, W. B. (1963). *Nature, Lond.* **197**, 716.
Rao, C. R. (1952). "Advanced Statistical Methods in Biometric Research." Wiley, New York.
Rescigno, A. and Maccaccaro, G. A. (1960). The Information Content of Biological Classifications. *In* "Symposium on Information Theory". Butterworths, London.
Richards, F. J. (1959). *J. exp. Bot.* **10**, 290.
Rogers, D. J. and Tanimoto, T. T. (1960). *Science* **132**, 1115.
Rohlf, F. J. (1962). Ph.D. Thesis, University of Kansas.
Rohlf, F. J. and Sokal, R. R. (1963). *Kans. Univ. Sci. Bull.*
Sebestyen, G. S. (1962). "Desision-making Processes in Pattern Recognition". Macmillan, New York.
Shepard, R. N. (1962a). *Psychometrika* **27**, 125.
Shepard, R. N. (1962b). *Psychometrika* **27**, 219.
Sneath, P. H. A. (1957). *J. gen. Microbiol.* **17**, 201.
Sneath, P. H. A. (1962). *Symp. Soc. gen. Microbiol.* **12**, 283.
Sneath, P. H. A. and Cowan, S. T. (1959). *J. gen. Microbiol.* **19**, 551.
Sneath, P. H. A. and Sokal, R. R. (1962). *Nature, Lond.* **193**, 855.
Sokal, R. R. and Sneath, P. H. A. (1964). "Numerical Taxonomy". W. H. Freeman, San Francisco and London.
Tanimoto, T. T. (1958). "An Elementary Mathematical Theory of Classification and Prediction". I.B.M. Corp., New York.
Williams, W. T. and Dale, M. B. (1962). *Nature, Lond.* **196**, 602.
Williams, W. T. and Lambert, J. M. (1959). *J. Ecol.* **47**, 83.
Williams, W. T. and Lambert, J. M. (1960). *J. Ecol.* **48**, 689.
Williams, W. T. and Lambert, J. M. (1961a). *Nature, Lond.* **191**, 202.
Williams, W. T. and Lambert, J. M. (1961b). *J. Ecol.* **49**, 717.
Williams, W. T. and Lance, G. N. (1958). *Nature, Lond.* **182**, 1755.
Williams, W. T., Dale, M. B. and Macnaughton-Smith, P. (1964). *Nature, Lond.* **201**, 426.
Yule, G. U. and Kendall, M. G. (1950). "An Introduction to the Theory of Statistics". Griffin, London.