R. Hamilton

The Use of Computers in Classifying Marine Bottom Communities

W. STEPHENSON

Department of Zoology, University of Queensland, St. Lucia, Brisbane

Abstract

Different schools of terrestrial botanists have recognised biocoenotic communities in terms of dominance, constancy and fidelity; others have stressed the existence of ecological continua. There are parallels in studies of marine benthos.

Different forms of data (binary to quantitative) are listed and also the possibilities of data reduction and transformation. Measures of similarity/dissimilarity are outlined, and the use of visual matrices. Classificatory techniques include divisive and agglomerative, monothetic and polythetic, non-hierarchial and hierarchial. The main hierarchial strategies are compared with emphasis on group-average and flexible sorting and their combined (non-hierarchial) use.

The following are stressed: (i) the wide range of methods available, (ii) some of the criteria for selecting the "better" methods, and (iii) the general value of selected methods for use with "Petersen-type" data. Reference is made to the use of numerical methods in marine ecology, benthic, and otherwise.

INTRODUCTION

There seem to be only three reasons why anyone might wish to use computers in the present context and these are:

(1) to appear "up-to-date"

(2) to try out methods for application in another context, and

(3) to attempt to analyse data too complex for adequate consideration by "commonsense" techniques.

Disregarding the first reason, the author has been involved in computer exercises involving the second and third. Beginning with the third (Stephenson, Williams, and Lance 1970), we found it impossible in analysing dredge catches from Moreton Bay to pick out any clear Petersen-type communities, in which dominant animals characterise a number of the sampled sites, and were forced because of the complexity of the data to use computer methods. Our second paper (Stephenson and Williams 1971) concerned a limited number of grab catches from a New Guinea situation in which Petersen-type communities could be recognised intuitively, and in which we compared a number of numerical analyses. The most satisfactory of these enabled us to confirm and extend our intuitive conclusions.

Recently we have re-analysed the results of Petersen (1914) upon which the original concepts of bottom communities were established, and have shown that with the methodologies now available it is possible to make considerably more use of his data and obtain somewhat different conclusions (Stephenson, Williams, and Cook, in press).

It is evident that we have found numerical classification to be helpful, and it is equally evident from the literature that there is increasing application by other workers. Recent benthic literature includes Cassie and Michael (1968), G. F. Jones (1969), Lie and Kelley (1970), Bayer, Voss, and Robins (1970), Martin, Duke, Bloom, and McGinnis (1970), Boudouresque (1970, 1971), Hughes and Thomas (1971a, b), Popham and Ellis (1971), Day, Field, and Montgomery (1971) and Boesch (1971). There have been similar analyses of intertidal biotas by Field and McFarlane (1968), Field and Robb (1970), and others are envisaged (see Moore 1971 for example). Different methods have been used by these various workers, and by others concerned with plankton, for example Cassie (1961), Williamson (1961, 1963), Fager and McGowan (1963), Colebrook (1964), Colebrook and Robinson (1964), McConnaughey (1964), Sheard (1965), Brown (1969) and Thorrington-Smith (1971). Examination of the above literature suggests that some of the methods used are open to serious criticisms, and that a brief review would be of value. This is the object of the present paper.

Oceanography of the South Pacific 1972, comp. R. Fraser. New Zealand National Commission for UNESCO, Wellington: 1973.

RECOGNITION OF COMMUNITIES

COMMUNITIES AND ASSOCIATIONS

To appreciate more fully the problems which are involved, it is desirable to outline briefly the history of the concepts of communities or associations which have been developed by terrestrial botanists. Here we shall eliminate the normal extensive discussions of community concepts by discarding the biotope concept originated by Dahl (1908), and concentrating on biocoenosis concepts originated by Möbius (1877). We shall also avoid the debates upon whether or not communities should be regarded as organisms, quasi-organisms, or super-organisms (Clements 1905, 1916, 1920; Tansley 1920; and Emerson 1939) because in marine bottom communities the supposedly intimate cross-linking of species is largely unproven. In another paper we have synonymised communities with the "associations" of the botanists (Stephenson, Williams, and Cook, in press) and followed Clark (1946), the main definition of Fager (1963, p. 418), and Mills (1969) in only being concerned with co-occurrence of species.

BOTANICAL WORK UPON RECOGNITION OF COMMUNITIES

Studies in terrestrial botany have preceded those in marine biology and there have been three main schools, a dominance-consistency school which originated in northern Europe at about the turn of the century, the Braun-Blanquet school (see Braun-Blanquet 1951) originating in southern Europe a little later, and the school of individualistic dissenters originating in America, and associated with the names of Gleason (1926), Whittaker (1962, 1970), and McIntosh (1967). The dominance-constancy school recognised associations by one or two dominant organisms (e.g., trees) which were constantly present and dominated the community. Objections to this approach are that it is difficult to apply in complex floristic situations where many species vie for dominance, and that it neglects subdominant species. The Braun-Blanquet school's main contributions in the present context are its recognition that three concepts merit consideration in choosing species which characterise communities—these are dominance, constancy, and fidelity. In this form they have been virtually ignored by marine biologists until recently (see Fager 1963; Stephenson, Williams, and Lance 1970) but the basic ideas have been used. As an example, plankton indicators are faithful species which like most others tend to be infrequent and are often inconstant.

The bulk of modern work in terrestrial plant ecology appears to indicate that many species have essentially individualistic distributions, not closely related to any other species, and if so recognition of associations by any method will present problems. Computer methods have been developed by phytosociologists to analyse these difficult situations, but until recently have found little application in marine studies in general and benthic studies in particular.

RELATIONSHIP BETWEEN BOTANICAL AND BENTHIC COMMUNITY STUDIES

Petersen-type communities have a close affinity to those described by the dominance-constancy school of botanists and are open to the same general objections: First the concepts are doubtfully applicable to complex situations such as appertain in warmer waters. While several workers have recognised Petersen-type communities in these waters, most of the later workers failed to find communities with a few predominating species. Examples of the former are Hartman and Barnard (1958, 1960) and Barnard and Ziesenhenne (1961) in Southern California, Longhurst (1957, 1958) and Buchanan (1958) in West Africa, Seshappa (1953) in India, Plante (1967) in Madagascar, and Stephenson and Williams (1971) in New Guinea—but in a possibly special situation. Examples of workers in warmer waters who were unable to recognise Petersen-type communities include Hartman (1955) and G. F. Jones (1969) in California, Parker (1956, 1960) in the Gulf of Mexico, Parker (1964) in the Gulf of California, Day (1963) in South Africa, Thorson (1966) off Thailand, Gallardo (briefly quoted in Thorson 1966) off South Vietnam, and Stephenson, Williams, and Lance (1970) in Moreton Bay, South Queensland. The late Dr Thorson wrote to me in these terms of his Thailand work: "We never found any community pattern. . . . On the contrary while the parallel communities always were built up by fairly few species found in large quantities we here found a huge number of species represented by only a few individuals each." Our own work in Moreton Bay led to a similar state of "disorder"; we said (Stephenson, Williams, and Lance 1970, p. 484): "No communities based upon these (or other) dominants were revealed either by 'commonsense' consideration or the computer analyses which were undertaken. Examination of distributional patterns of the 36 most frequently occurring species did not reveal satisfactory groupings apart from ill-defined relationships to substratum, and it was evident that neither a simple dominance concept nor a co-dominance concept was applicable to the present situation. In other words, Petersen-type communities could not be recognised."

Other criticisms have been levelled at the Petersen school, particularly in the selection of characterising species. Following Petersen these have been recognised on an intuitive basis of number *and* weights; we have shown that classifications based on numbers and on weights give different results (Stephenson, Williams, and Cook, in press). Although bottom sampling data using a Petersen grab or one of the multiplicity of later designs is approximately quantitative, selection of characterising species has been partly subjective and in the words of Thorson (1957, p. 476): "Commonsense, combined with experience and intuition play an essential part in selecting characterising species." Somewhat cynically one might suppose that part of the experience was knowledge of the literature, and that there may have been unconscious selection of characterising species which would show resemblances with communities already described.

NB Application of the ideas of dominance, constancy, and fidelity to marine benthic situations has particular relevance to the use of numerical methods of analysis, and the last two merit discussion. A species is highly constant if it appears in all the samples* or quadrats* within an association, but it need not be restricted to a single association. Conversely a species is highly faithful if it occurs only in a single association, but it need not occur within all the samples within the association. Constant species are often dominants, but faithful species which have a narrow spectrum of ecological requirements are often far from obvious in the communities they characterise.

The interrelationships between constancy and fidelity are important in classifying sites* by their species composition: there are no firm rules how far the classification ought to proceed. At one extreme we make a single site-group and have not classified at all, whereas at the other extreme each site is separate and again we have gained nothing. Some arbitrary intermediate level must be sought which gives the greatest overall "sense" to the data, and which avoids the extremes of oversplitting and overlumping. With more site-groups, the moderately common species fit into "neater parcels" and the constancy of the system increases. Simultaneously the less common species, which include the faithful ones, lose their fidelity by being spread between several groups. In the analyses we have so far undertaken, the numbers of groups we wished to consider have stressed constancy with sacrifice of fidelity. In the end we obtain, but by a route very different from Petersen's, communities which are determined by the dominance and constancy of the species.

The concept of individualistic distribution and its corollary of an ecological continuum has been stated or inherent in many benthic studies. Examples include Stephen (1933), Lindroth (1935), Macginitie (1935, 1939), Tischler (1950), Sanders (1960), Wieser (1960), Margalef (1962), Eisma (1966), Udvardy (1964), Glemarec (1969), and Mills (1969). Additional references are given in Pearson (1970). A typical approach has been to plot species populations along a tranverse, and to note that boundaries do not coincide (e.g., Sanders 1960, Glemarec 1969, Mills 1969). Eisma (1966) used correlation coefficients and showed that the benthic molluscs he studied tended to have their own distributional patterns.

The concept of a complete ecological continuum is one which the human brain can scarcely comprehend. A similar difficulty arises in taxonomy when the change in specific features through geological time is considered. Haldane (1956, p. 96) has stated: "Thus in a complete palaeontology all taxonomic distinctions would be as arbitrary as the division of a road by milestones. . . . To sum up, the concept of a species is a concession to our linguistic habits and neurological mechanisms . . .". Ager (1956, p. 105) continues: "The problem therefore lies in defining the indefinable, and can only be clarified by agreed arbitrary solutions."

Several workers have pointed out that by using numerical methods and in ecology one may arrive at a situation somewhere between a continuum and discrete communities. In bottom community studies, both Jones (1969) and Stephenson, Williams, and Lance (1970) came to similar conclusions. In order to "sharpen" the boundaries in a near-continuum certain techniques are preferable.

---

*These, and "stations" are treated as being synonymous.

INTRODUCTION TO CLASSIFICATORY TECHNIQUES

Here ordination is excluded although noting its use by ourselves (Stephenson, Williams, and Lance 1970; Stephenson and Williams 1971) and by Boudouresque (1970, 1971) in benthic studies.

"Dissection" is included (see Kendall 1966, Kendall and Stuart 1966, Cormack 1971); this is the splitting of a continuous into a discontinuous series, and in a strictly continuous ecological classification we dissect rather than classify. Because presumably there is merging between complete continua, near-continua, and discontinuity in ecological situations, it seems best to merge dissection into classification and use this term in its wider sense.

There are many kinds of classification, and there have been several discussions of and classifications of classifications, for example by Macnaughton-Smith (1965), Kendall (1966), Good (1965), Lance and Williams (1967), Jardine and Sibson (1968), Jardine and Sibson (1971a, b), Cormack (1971), and Williams (1971). One may note (with disfavour) an increasingly mathematical flavour in this field of literature. It is fortunate that many of the alternatives are not relevant to the present situation and that the remainder are sufficiently understandable in commonsense terms to permit comprehension by a practising marine biologist.

Only one of the various alternative types of classification merits comment at this stage, this is the difference between normal and inverse analyses. In ecological surveys we note the occurrences of certain species in certain sites (recognising that occurrences may be in presence/absence form or may be in quantitative form), and we can analyse our data in two ways. First we can classify our sites, grouping like with like, using the specific data as attributes. Second we can classify our species, grouping like with like, using the site data as attributes. These two types of classification have had a moderately long history (for example in psychological and educational research) and the terms "Q" and "R" have been applied. Unfortunately their meaning have been reversed on occasion, and following Ivimey-Cook, Proctor and Wigston (1969), and also Hughes and Thomas (1971a) it is clear this nomenclature should be discontinued. Following Williams and Lambert (1959, 1961) the terms "normal" and "inverse" should be used, the former for site-grouping and the latter for species-grouping.

Later we shall return to the different methods of classification but meanwhile it is necessary to list the types of data available for classification, and the preliminary treatments which they may receive.

THE DATA

General Forms of Data

Five forms exist:

(1) *Binary*. In ecology this means "present" and "not present", and whenever possible in marine studies more informative data should be sought.

(2) *Multistate*. For example "red", "white", and "blue"; this is not applicable in the present case.

(3) *Ordinal or graded multistate*. For example "abundant", "common", "present". With improved methods of data analysis, it is preferable if possible to have full data.

(4) *Ranked*. This is essentially graded within a collection, and again should be avoided if possible.

(5) *Quantitative*. These data may be meristic (in whole numbers) or in continuous form. If possible these data should be obtained and most of the remainder of this account will assume it.

Data Reduction, and the Problem of "Double Negatives"

Classificatory programmes are commonly used by biologists for either taxonomic or ecological purposes, and it is not always appreciated that those suitable for taxonomy might not be so for ecology. An initial difference is in the form of the data, in taxonomy most are usually in binary or multistate form, in marine ecology the bulk are usually in one of the other forms.

Another difference can best be appreciated by reference to binary data in which there is a subtle difference between binary data in taxonomy and in ecology. In ecology the alternatives are present/not present, whereas in taxonomy it is often the choice of two different types of "presence". In taxonomy, moreover, the absence of a character is usually very meaningful, while the absence of a species from a site may mean only it is relatively rare anyway. In ecology if there are many infrequent species the value of "double negative" recordings becomes doubtful and needs consideration.

One way to reduce them is to exclude the less common species and numerous benthic workers have done this. It has the added advantage of saving in the time required for coding and processing, and in fact certain programmes are only possible with a reduced data matrix. Further methods of data reduction are available—we have successfully employed a divisive monothetic classification (see later).

Data Transformation

It appears that many if not all sampling techniques for marine benthos, plankton, and nekton produce large numbers of a few species and small numbers of many. If raw numbers are used in many analytical techniques, results are strongly biassed by the few abundant species. It has become common practice to use transformed numbers, aiming at an approximately normal distribution after transformation. There is a considerable literature on these transformations (see Macfadyen 1963) and it is growing (see Ebeling *et al.* 1970, Edden 1971). The problem is complicated by the fact that different transformations may be needed for different species and that data upon numbers and weights may require different treatments. We have used a square root transformation for numbers data and a logarithmic one for weights data (Stephenson, Williams and Cook, in press). In this case to avoid negative logs for values less than unity it has become customary to use log (n + 1).

√N

MEASURES OF SIMILARITY AND DISSIMILARITY

Data in their final form are used, in most normal analyses (the polythetic ones—see later) to obtain measures of similarity (or dissimilarity) between all possible pairs of stations. Usually similarity and dissimilarity measures are closely related, often one is unity minus the other.

The literature contains a great variety of measures—see, for example, Goodman and Kruskal (1954, 1959), Dagnelie (1960), Sokal and Sneath (1963), Macfadyen (1963), and Cormack (1971). These measures can be divided into those appropriate to binary data only, and those for other data (typically quantitative).

Using Binary Data

Although binary data should generally be avoided, many marine workers (including ourselves) have used them and various of the measures, often termed coefficients, which are derived therefrom. To understand these it is best to refer to the standard 2 × 2 table, which in the present context, using + for presence and − for absence is:

	+ Species 1	− Species 1
+ Species 2	a	b
− Species 2	c	d

a signifies co-occurrences, *b* presence of species 2 but not species 1, *c* presence of species 1 but not species 2, *d* signifies conjoint absence.

The two coefficients which appear particularly suited for ecological work are the Jaccard (1908) and the Czekanowski (1913), often attributed to Dice (1945).

The Jaccard coefficient is: $\dfrac{a}{a + b + c}$, while the

Czekanowski coefficient is $\frac{2a}{2a + b + c}$, and gives double weighting to joint occurrences. The Jaccard coefficient has been used on intertidal data by Field and McFarlane (1968) and by Popham and Ellis (1971). We have used the Czekanowski coefficient, unfortunately naming it the Jaccard coefficient, in preliminary consideration of data on the Moreton Bay benthos (Stephenson, Williams, and Lance 1970) and it has also been used in marine studies by Field and McFarlane (1968) and by Day, Field and Montgomery (1971). It should be noted that both the above similarity coefficients (and many of the others) range between zero for no association to unity for complete association. $0 \rightarrow 1$

Other coefficients used in marine work include Fager's (1957) index of affinity, modified by Fager and McGowan (1963) This is:

$$\frac{a}{\sqrt{(a + b)(a + c)}} - \frac{1}{2\sqrt{a + b}}$$

It has only been used in inverse analyses to determine the "recurrent species groups", and has had extensive use mostly by American marine workers. References include Fager and McGowan (1963), Sheard (1965), Jones (1969), Longhurst (1969), Lie and Kelley (1970), Bayer, Voss, and Robins (1970), Martin, Duke, Bloom, and McGinnis (1970) and Boesch (1971).

More complex coefficients have been used. For example Lie and Kelley (1970) for normal analysis use a coefficent which they call Kendall's rank correlation coefficient, but as quoted this is Pearson's (1926) product moment coefficient, which for binary date is:

$$\frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$$

It will be noted that this involves the double negative d. Apart from this there are reasons for avoiding the Pearson coefficient, particularly in normal classification (see later).

An even more complex coefficient derived from Preston's (1962) resemblance equation has been used in marine phytoplankton studies by Thorrington-Smith (1971)- the equation has no general solution.

To appreciate matters raised later, it is important to understand the concept of *ecological distance*, and this can be illustrated by reference to binary data. Suppose there are two samples differing by the presence of a species in one sample only. This can be represented by two points (the two samples) on a line which is one unit long. For a presence/absence difference of a second species we can either have a line two units long or, and preferably, generate a second line at right angles to the first. The two points (samples) are now separated by the hypotenuse of a right-angled triangle and the ecological distance apart is $\sqrt{2}$. With a third presence/absence of a species we have a third axis, and a right-angled triangle with one side $\sqrt{2}$ (as established) and another with unity. The ecological distance becomes $\sqrt{3}$. More generally the ecological distance is $\sqrt{n}$, where there are n presence/absence differences between the stations. Referring to the standard 2 × 2 table this is $\sqrt{b + c}$.

For binary data it is also possible to use information gain on fusion of stations (a good introduction to information theory is given by Edwards, 1964). In the above notation it is $(b + c)2 \log 2$, in other words it is directly proportional to the square of the ecological distance.

USING NUMERICAL DATA

A variety of measures which look promising have proved disappointing in practice. The Pearson (1926) correlation coefficient can be applied satisfactorily in the inverse (species/species) analyses and has been used by a variety of plankton workers beginning with Cassie (1961) and in benthic studies (see Eisma 1966, Jones 1969), sometimes in a preliminary way (Stephenson, Williams, and Lance 1970). One disadvantage is that double negatives are included. When used for the normal (site/site) classification it involves standardisation by sites and hence operates on relative abundance.

Measures based on information theory appear attractive, following their successful application to taxonomy, and these measures imply their own classificatory strategies. We have used and discarded two methods--the fusion of sites is dominated by those with the sparsest biota (Stephenson and Williams 1971).

Ecological distance, also called Euclidean distance, has been variously used. If x_{1j} and x_{2j} are the numbers of the jth species at two sites to be compared, this distance is $\sqrt{\Sigma_j(x_{1j} - x_{2j})^2}$. In its "raw" form it will be dominated by the abundant species but this can be avoided by standardising all species to unit variance (Orloci 1967a, b). This standardisation was used in recent benthic work by Hughes and Thomas (1971a, b) as a useful method of combining data upon numbers of animals with weights of marine plants. It has also been considered and rejected by Thorrington-Smith (1971) in phytoplankton work, where it is described as McIntosh's (1967) dissimilarity measure.

Two coefficients have recently been employed as dissimilarity measures. The first is widely known as the "Bray-Curtis" measure and is the complement of that used by Bray and Curtis (1957). It was actually used by previous workers (Motyka *et al*, 1950) and is a quantitative extension of Czekanowski's (1913) coefficient. Using the previous notation the dissimilarity measure between two sites is:

$$\frac{\Sigma_j | x_{1j} - x_{2j} |}{\Sigma_j(x_{1j} + x_{2j})}$$

(In this the $| x_{1j} - x_{2j} |$ indicates that the value of the difference is always positive.) It should be noted that in this coefficient the denominator is the total of all individuals of all species at the two sites. Hence it tends to be greatly influenced by occasional outstanding values. This coefficient has been used by Field and Macfarlane (1968), Field (1969), Day, Field, and Montgomery (1971), and by Stephenson and Williams (1971).

Another dissimilarity measure, apparently first used for rainforest ecology by Dr W. T. Williams, proved valuable in our analyses of Petersen's data (Stephenson, Williams, and Cook, in press). This is the Canberra metric, which is:

$$\frac{1}{j} \Sigma j \frac{| x_{1j} - x_{2j} |}{(x_{1j} + x_{2j})}.$$

It differs from the Bray-Curtis complement in being the sum of a series of fractions. An outstandingly abundant species can only contribute to one of the fractions and the same applies to an outstanding difference. We have found this measure preferable to the Bray-Curtis. It has one disadvantage—in the above expression, so long as x_2 is zero, the fraction $\frac{x_1 - x_2}{x_1 + x_2}$ contributed by a particular species is unity. This we have circumvented in cases where one of the values is zero by replacing it with a small number, its value being lower than any of the recorded values, and judged by trial and error. The criterion for choice was the constancy and numerical dominance which was revealed in final analyses.

MATRICES AND THEIR VISUAL ANALYSIS

The similarity or dissimilarity measures between all pairs of sites (or between all pairs of species) form a matrix, and all classification consists of matrix analysis.

Matrices can be displayed. Taking a simple arbitrary example, we might have four sites I, II, III, IV, and similiarity measures as follows: I–II 0.5, I–III 0.3, I IV 0.1, II III 0.5, II IV 0.3, III IV 0.7. These can be expressed as a two dimensional matrix or trellis diagram in the following form:

	I	II	III	IV
I		0.5	0.3	0.1
II			0.5	0.3
III				0.7
IV				

Only the upper triangle of the trellis is given, clearly the lower half is a repetition.

Matrix analysis can begin at a very simple and visual level. Typically the continuous data are graded into a small number of categories and then in the words of Macfadyen (1963, pp. 194–6): "The indices are arranged in the squares of the trellis diagram . . ., the linear order of the samples being the same in the rows and the columns. This linear order is then shuffled round by trial and error so as to bring the highest numbers to the centre diagonal and the lowest numbers farther from it. A moment's thought will show that this results in samples with high affinity, as measured by the index, coming together. When arrangement has been carried out as far as possible it is usually found that some of the samples fall into distinct groups."

Visual analyses of "trellis" matrices are being extensively used in recent publications on marine ecology including the benthos. Examples are Sanders (1960), McNulty, Work, and Moore (1962), Parker (1964), Eisma (1966), Boudouresque (1970), Lie and Kelley (1971), Pearson (1971), Popham and Ellis (1971), and by Boesch (1971). Thorrington-Smith (1971) in phytoplankton work has recently published complete matrices without grading or rearrangement.

With the techniques now available to analyse matrices only by visual methods seems insufficient.

ALTERNATIVE TYPES OF CLASSIFICATION

All apply to both normal and inverse analyses, and are given below as a series of alternatives, usually with preferences indicated. They are illustrated by reference to normal (site) classification.

Divisive and Agglomerative

The former divides the sites into groups from the "top" (considering all the data simultaneously), while the latter groups the data into smaller and then larger entities starting from the "bottom" (starting from individual sites). Both in the end give similar results, typically shown by dendrograms—a convenient term first coined by Mayr, Linsley, and Usinger (1953). Because in both cases the lower portions of the dendrograms are scarcely ever used, the divisive method is more economical of computer time than the agglomerative.

The classificatory programmes involving divisive processes have until recently all been *monothetic*, i.e., they operate by considering only one attribute at a time. In the present context sites are divided into two groups, typically by the presence and absence of one species. This can result in misclassification, a site with very considerable resemblences to others can be "misdirected" on the basis of a single species. For this reason divisive monothetic methods have limited application in benthic classification, although we have used them (mostly for data reduction) on two occasions (Stephenson, Williams, and Lance, 1970; Stephenson, Williams, and Cook, in press).

For benthic work the effective choices lie amongst the agglomerative *polythetic* methods. These consider simultaneously all the attributes (species) of all the sites, and because they involve handling more data are computationally more elaborate, and also tend to produce "better" results.

Non-hierarchial and Hierarchial

Williams (1971) has defined these terms, noting that in hierarchial strategy optimises the route followed in classification but at possible expense in the homogeneity of the groups. A non-hierarchial strategy optimises the structure of the individual groups, and as such has advantages in marine ecological studies Unfortunately in Williams' words ". . . their curren state of development lags far behind that of thei hierarchial counterparts, which at their best are mor flexible, provide a wider range of facilities, ar numerically better understood, and are computationally faster."

Here we will only consider hierarchial strategies up to the point of combining them to obtain a non-hierarchial result (see later).

Hierarchial Strategies Available

Listed from Lance and Williams (1966) and Lance and Williams (1967) these are: nearest neighbour, furthest neighbour, group average, centroid, median, and flexible sorting.

(1) *General remarks*

Before considering them, several of their properties as determined by use must be discussed. These are "chaining", "reversals", and "group-size dependence".

"Chaining" involves consideration of further terms: "space-conserving", "space-dilating" and "space-contracting". Reference should be made to the critical paper in this field, Lance and Williams (1967). If the original inter-site measures are regarded as occurring in a given space, then sometimes the properties of this space remain unaltered as groups form, but in other strategies the groups alter the properties of the space near them. The authors state (p. 374): "In a space-contracting system a group will appear, on formation, to move nearer to some or all the remaining elements; the chance that an individual element will add to a pre-existing group rather than act as the nucleus of a new group is increased, and the system is said to 'chain' (for a measure of chaining, see Williams, Lambert and Lance, 1966). In a space-dilating system groups appear to recede on formation and growth; individual elements not yet in groups are now more likely to form nuclei of new groups."

"Reversals" occur when a descending branch of a dendrogram gives rise to groupings at higher levels than the original. Examples are given in Lance and Williams (1967). To avoid reversals there is a "monotonicity requirement" and of the methods listed, this is conspicuously lacking in centroid sorting.

"Group-size dependence" has been discussed by Williams, Clifford, and Lance (1971) and applies to space-dilating strategies. When groups are small they admit new members readily, but as they grow entry becomes more difficult. This disadvantage applies to flexible sorting as normally operated, but more so to most information statistic strategies.

(2) *Nearest neighbour*

Lance and Williams (1967, p. 374) have commented: "This is the oldest of the conventional strategies. The distance between two groups is defined as the distance . . . between their closest elements, one in each group. . . . As a group grows it must appear to move closer to some elements and further from none; it is thus a space-contracting strategy, and its consequential chaining tendencies are notorious." Because of its chaining properties this method is not particularly suitable in practice. However Jardine and Sibson (1968) have erected a set of mathematical criteria which they believe should be imposed on classificatory strategies, and this virtually confines one to using nearest-neighbour fusion. It appears that if the nearest-neighbour strategy has the greatest mathematical justification it gives the least artificial sharpening of boundaries within a near-continuum and hence least assists the mental processes of the ecologist. A recent example of use of nearest-neighbour strategy in a marine ecological study is given in Thorrington-Smith (1971). As many as 12 species are linked in chains in her species/species analysis and the overall conceptual picture is considerably weakened thereby.

(3) *Furthest neighbour*

This, like nearest neighbour, is a "single linkage" method, operating between, for example, a site to be admitted to a group and a single site already in the group. It was suggested by Macnaughton-Smith (1965), and we tried it recently (Stephenson, Williams, and Cook, in press).

Of furthest neighbour sorting to quote Lance and Williams (1967, p. 374): "It is the exact antithesis of the foregoing, in that the distance between the two groups is now defined as that between the most remote pair of elements, one in each group. . . . Since on growth a group will recede from some elements and move nearer to none, it is markedly space-dilating." In other words it gives much more intense clustering than nearest-neighbour sorting.

(4) *Centroid*

This appears to date from Sokal and Michener (1958) who termed it "the weighted-group method". Here fusion of a site into a group, or fusion of groups, depends upon the co-ordinates of the centroids when the group is considered in multi-dimensional Euclidean space. The principle of centroid sorting is easy to understand, and it is space-conserving. Because of its susceptibility to reversals it should be avoided. Sokal and Sneath's (1963) weighted pair method and weighted group method are centroid sortings. This has been applied to marine benthos by Popham and Ellis (1971).

(5) *Median*

This was suggested by Gower (1966), and avoids another disadvantage of the centroid system, as shown below in fig. 1 in diagrammatic form.

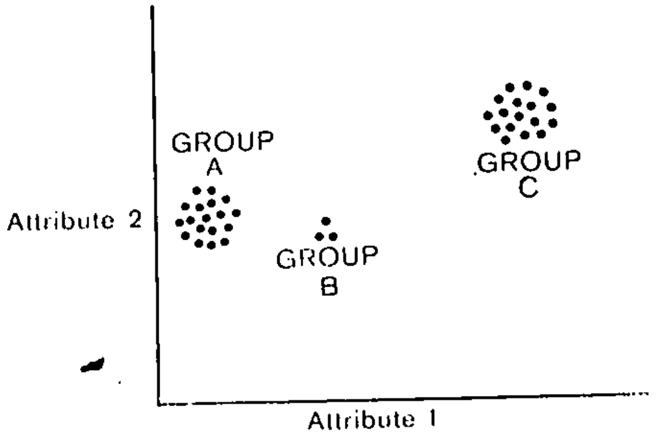

FIG. 1. Groups of entities in two-dimensional space. In centroid grouping B loses its individuality; in median grouping its individuality is retained.

In centroid grouping, as group B fuses with group A the new centroid comes to lie within the confines of group A. To emphasise the individuality of group B the properties of the combined group should be intermediate. This can be effected by regarding the groups as being of unit size (unweighted) and the new position of the fused group (A+B) must lie between them. This grouping has been little used and because it is quite liable to give reversals it should be avoided.

(6) *Group average*

In its form of element/group fusions this was originated by Sokal and Michener (1958) and termed the "unweighted group mean method"; it has also been called the "unweighted pair-group method". In the element/group situation the mean distance of a site to each site within a group is derived, and the site is fused to that group with the smallest mean distance. Lance and Williams (1967, p. 375) state: "The system is less rigorously space-conserving than is centroid but, since it has no marked tendencies to contraction or dilation, it may be regarded as a conserving strategy." In other words it gives only moderately sharp clustering. It has the advantage of being monotonic and is very little prone to misclassification. Group average sorting has generally satisfactory properties and has been used in benthic marine work by G. F. Jones (1969), Day, Field, and Macfarlane (1971) and by Stephenson, Williams, and Cook (in press). It has also been used in studies of animals netted in deep water by Ebeling *et al.* (1970).

(7) *Flexible*

This system was introduced by Lance and Williams (1966, 1967) who derived a general equation which permits the five foregoing strategies to be expressed in similar form, by changing the values of certain constants. This avoids the necessity for writing five separate computer programmes. They proposed a new flexible strategy in which the values of constants were constrained to avoid reversals, and in which by altering the constants within these constraints it is possible (see Williams, Clifford, and Lance 1971) to move from a space-contracting to a space-dilating model. The critical constant β has been termed the cluster-intensity coefficient (Williams 1971). It has become conventional to set its value at -0.25, when the strategy becomes space-dilating. This results in the sharpening of clusters which is generally desirable in marine benthic surveys and we have used this strategy successfully on three occasions (Stephenson, Williams, and Lance 1970; Stephenson and Williams 1971; Stephenson, Williams, and Cook, in press).

To summarise the above the three strategies which appear most satisfactory are furthest neighbour, group average, and flexible ($\beta = -0.25$). However as might be expected they give different results (see Stephenson, Williams, and Cook, in press), so further considerations are involved as outlined below.

TWO-WAY COINCIDENCE TABLES AND COMBINATION OF HIERARCHIAL STRATEGIES

As indicated earlier, one can use the same data to obtain a normal classification into site-groups and an inverse classification into species-groups. It is desirable to combine these in a two-way coincidence table which indicates which species (and with quantitative data what quantities of them) characterise which site-groups.

When one compares the two-way tables produced by alternative methods (for example Bray-Curtis versus Canberra metric coefficients, or group average versus flexible sorting) it is evident that there are differences in both site-grouping and species-grouping. Questions arise as to which is "best".

The objective of classification is to promote understanding of a complex system and this is facilitated by the production of discrete groups. These are best indicated by the degrees of constancy and fidelity of the various species. For example with higher constancy of more species there will be "tighter" groupings and hence a "better" classification. In testing three sorting strategies—furthest neighbour, group average, and flexible, it became clear that the choice lay between the second two, but beyond this the choice was uncertain (Stephenson, Williams, and Cook, in press).

Careful scanning of two-way tables usually reveals, at least for flexible sorting, a small number of apparent misclassifications which appear to be inevitable in agglomerative hierarchial strategies. By no means all the differences between strategies are due to these misclassifications. It is possible by "handwork" to determine the corresponding portions of the different end-results of classifications, and it is equally possible to re-allocate the "non-conformist" elements, which we did in the above work. By so doing we were converting two hierarchial strategies into a single non-hierarchial method. We think the method was successful with the data we used; if others have a similar experience it may be desirable to programme the method and avoid tedious handwork.

CONCLUSION

It will be evident that, in handling marine benthic data there are many alternative methods. Choices are needed in the amounts of data to discard, the possible transformations of data, the measures of similarity or dissimilarity, and the strategies of classification. Most of these choices may be made by the ecologist for the very good reasons of common sense, but it seems necessary at the present stage to check the final results of a number of alternatives to determine which is "best". Preference is not in terms of which precise groupings emerge but whether the groupings are precise or nebulous.

Beyond this further "extrinsic" criteria are available particularly for the site-groupings. Other things being equal we would expect that sites would be grouped in proximity and form some sort of topographical patterns. Also other things being equal we would not expect amongst mixed environments there would be close groupings between vastly dissimilar substrates. This did happen between clean coral grit and black stinking mud environments using certain techniques with our New Guinea work (Stephenson and Williams 1971) and led to rejection of the analytical methods.

It will be clear from this paper and from the vast literature in other fields that computer techniques have come to stay, are developing rapidly, and that they are likely to undergo further evolution. In ecology in general and in bottom communities studies in particular they are likely to be of great value, and subject to critical scrutiny. It is hoped that from such use and scrutiny the evolution may be facilitated. If this review assists in making itself obsolescent it will have performed its function.

ACKNOWLEDGMENTS

I am deeply grateful to Messrs R. Bradbury and A. R. Jones who read the manuscript and made valuable suggestions. I am particularly grateful to Dr W. T. Williams who initiated me to computer methods, has guided my steps, and who also made valuable suggestions on the present paper at a manuscript stage.

ADDENDUM

Important information received between submission of this paper and its acceptance is outlined below.

(a) A paper by J. G. Field has appeared entitled "A numerical analysis of changes in the soft-bottom fauna along a transit across False Bay, South Africa (*J. exp. mar. Biol. Ecol.*, 7, 215–253, 1971). Field excluded from analysis difficult taxonomic groups (e.g., Nemertea, Nematoda) and all species with less than ten individuals in 55 samples. Data were log-transformed, and the "Bray-Curtis" (called Czekanowski) coefficient of association was employed. The sorting strategies used in normal classification were nearest-neighbour, centroid (= average-member) and group average-criticisms of the first two have been outlined earlier in this paper. Field obtained broadly similar results by his three methods.

(b) Dr J. B. Jillett presented a paper at the present conference (cf. Jillett and Mitchell, 1973). Jillett excluded all but eight of his planktonic taxa, and hence in essence eliminated all but dominant species. Species-groups and site-groups were deduced from visual matrices, using the Spearmann rank-coefficient as a measure of association.

(c) Dr W. T. Williams (personal communication) is developing an information-statistic classificatory strategy for ecological use which avoids the objections to previous methods we have tried (see Stephenson and Williams, 1971).

REFERENCES

Ager, D. V. 1956: Geographical factors in the definition of fossil species. pp. 105–109 *in* "The species concept in paleontology". (Sylvester-Bradley, P. C. ed.) The Systematics Association, London.

Barnard, J. L.; Ziesenhenne, F. C. 1961: Ophiuroid communities of southern Californian coastal bottoms. *Pacif. Nat. 2 (5)*: 131–152.

Bayer, F. M.; Voss, G. L.; Robins, C. R. 1970: Report on the marine fauna and benthic shelf-slope communities of the isthmian region. Battelle Memorial Institute, Ohio.

Boesch, D. F. 1971: Distribution and structure of benthic communities in a gradient estuary. Ph.D. thesis, Virginia Institute of Marine Science, Virginia.

Boudouresque, C. F. 1970: Recherches sur les concepts de biocoenose et de continuum au niveau de peuplements benthiques sciaphiles. *Vie et Milieu, B., 21:* 103–36.

—— 1971: Le concept de nodum en bionomie et sa generalisation. *C. R. Acad. Sci. Paris, 272:* 1260-3.

Braun-Blanquet, J. 1951: "Pflanzensoziologie". (2nd Ed.): Springer, Wien.

Bray, J. R.; Curtis, J. T. 1957: An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr. 27:* 325–49.

Brown, Sharon D. 1969: Grouping plankton samples by numerical analysis. *Hydrobiologia. 33:* 289–301.

Buchanan, J. B. 1958: Bottom fauna communities across the continental shelf off Accra, Ghana. *Proc. Zool. Soc. London. 130:* 1–56.

Cassie, R. M. 1961: The correlation coefficient as an index of ecological affinities in plankton populations. *Mem. Inst. Ital. Idrobiol. 13:* 151–77.

Cassie, R. M.; Michael, A. D. 1968: Fauna and sediments of an intertidal mud flat: a multivariate analysis. *J. exp. Mar. Biol. Ecol. 2:* 1–23.

Clark, E. T. 1946: Studies in the ecology of British grasshoppers. *Trans. R. Ent. Soc. London. 99:* 173–222.

Clements, F. E. 1905: "Research methods in ecology". Univ. Publ. Co., Lincoln, Nebraska.

Clements, F. E. 1916: Plant succession: an analysis of the development of vegetation. *Publ. Carneg. Instn.*, 242.

—— 1920: Plant indicators. The relation of plant communities to process and practice. *Publ. Carneg. Instn.*, 290.

Colebrook, J. M. 1964: Continuous plankton records: a principal component analysis of the geographical distribution of plankton 1948-68. *Bull. Mar. Ecol. 6:* 78–100.

Colebrook, J. M.; Robinson, G. A. 1964: Continuous plankton records: annual variation of abundance of plankton 1948-60. *Bull. Mar. Ecol. 6:* 52-69.

Cormack, R. M. 1971: A review of classification. *J. R. statist. Soc. A. 134:* 321–367.

Czekanowski, J. 1913: Zarys metod statystycznyck: Warsaw.

Dagnelie, P. 1960: Contribution à l'étude des communautés végétalés par l'analyse factorielle. *Bull. Serv. Carte Phytogéogr. (B), 5:* 7–71, 93–195.

Dahl, F. 1908: Die Lycosiden oder Wolfspinnen Deutschlands und ihre stellung im Haushalte der Natur. *Nova Acta Leop. Carol. Dtsch. Akad. Naturforsch. 88:* 174–678.

Day, J. H. 1963: The complexity of the biotic environment. pp. 31–49. *in* "Speciation in the sea" (J. P. Harding and Norman Tebble Ed.) *Publs. Syst. Ass.* 5.