EPA-600/3-77-033
March 1977

Ecological Research Series

APPLICATION OF NUMERICAL CLASSIFICATION IN ECOLOGICAL INVESTIGATIONS OF WATER POLLUTION

Environmental Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Corvallis, Oregon 97330

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into five series. These five broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the ECOLOGICAL RESEARCH series. This series describes research on the effects of pollution on humans, plant and animal species, and materials. Problems are assessed for their long- and short-term influences. Investigations include formation, transport, and pathway studies to determine the fate of pollutants and their effects. This work provides the technical basis for setting standards to minimize undesirable changes in living organisms in the aquatic, terrestrial, and atmospheric environments.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

EPA-600/3-77-033
March 1977

APPLICATION OF NUMERICAL CLASSIFICATION IN ECOLOGICAL INVESTIGATIONS OF WATER POLLUTION[1]

by

Donald F. Boesch
Virginia Institute of Marine Science
Gloucester Point, Virginia 23062

Grant No. R803599-01-1
ROAP/TASK NO. 21 BEI

Project Officer

Richard C. Swartz
Marine and Freshwater Ecology Branch
Newport Field Station
Corvallis Environmental Research Laboratory
Newport, Oregon 97365

CORVALLIS ENVIRONMENTAL RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
CORVALLIS, OREGON 97330

[1] Special Scientific Report No. 77, Virginia Institute of Marine Science.

DISCLAIMER

This report has been reviewed by the Corvallis Environmental Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

FOREWORD

Effective regulatory and enforcement actions by the Environmental Protection Agency would be virtually impossible without sound scientific data on pollutants and their impact on environmental stability and human health. Responsibility for building this data base has been assigned to EPA's Office of Research and Development and its 15 major field installations, one of which is the Corvallis Environmental Research Laboratory.

The primary mission of the Corvallis laboratory is research on the effects of environmental pollutants on terrestrial, freshwater, and marine ecosystems; the behavior, effects, and control of pollutants in lake systems; and the development of predictive models on the movement of pollutants in the biosphere.

This report describes classificatory techniques for demonstrating similarities in the distribution of species or in the composition of biological communities. Numerical classification offers a promising quantitative method for analyzing the impact of pollution on aquatic community structure.

A. F. Bartsch, Director
Corvallis Environmental
Research Laboratory

ABSTRACT

Numerical classification encompasses a variety of techniques for the grouping of entities based on the resemblance of their attributes according to mathematically stated criteria. In ecology this usually involves classification of collections, representing sites or sampling periods, or classification of species. Classification can thus simplify patterns of collection resemblance or species distribution patterns in an instructive and efficient manner.

Procedures of numerical classification are thoroughly reviewed, including data manipulations, computation of resemblance measures and clustering methods. The importance and effects of transformations and standardizations are discussed. It is particularly critical to choose an appropriate resemblance measure which best corresponds with the investigator's concept of ecological resemblance. Clustering methods form groups on the basis of patterns of inter-entity similarity. Various types of clustering methods exist but currently the most useful and best developed are those which are exclusive, intrinsic, hierarchical and agglomerative. Agglomerative clustering methods which distort spatial relationships and intensely cluster are often most useful with ecological data.

The value of post-clustering analyses in the interpretation of the results of numerical classifications is stressed. These include reallocation of misclassified entities, comparison of classifications of collections with those of species (nodal analysis), comparing alternate classifications, testing differences among groups, relating classification to extrinsic environmental factors and interfacing classification with other multivariate analyses.

The usefulness of numerical classification is demonstrated for objective analysis of the data sets resulting from field surveys and monitoring studies conducted for the assessment of effects of pollution. However, to date few pollution biologists have applied the more powerful classificatory techniques and post-clustering analyses.

CONTENTS

		Page
Foreword		iii
Abstract		iv
List of Figures		vii
List of Tables		viii
Acknowledgments		ix
Sections		
I	Conclusions	1
II	Recommendations	2
III	Introduction	3
	Numerical Classification	4
	Procedures of Numerical Classification	5
IV	Data	10
	Forms of Data	10
	Data Reduction	12
	Transformations	14
V	Resemblance Measures	18
	Normal vs. Inverse Analyses	18
	General	18
	Similarity Coefficients	20
	Euclidean Distance	30
	Correlation Coefficients	32
	Information Content Measures	34
	Probablistic Measures	36
VI	Clustering Methods	38
	General	38

Sections		Page
VI	Clustering Methods (cont.)	
	Classification of Clustering Methods	38
	Non-hierarchical Methods	42
	Agglomerative Hierarchical Methods	45
	Divisive Hierarchical Methods	56
VII	Interpretation of Numerical Classifications	59
	Stopping Rules	59
	Reallocation	61
	Nodal Analyses	62
	Comparing Classifications	68
	Testing Differences Among Groups	68
	Relating Classifications to Extrinsic Factors	70
	Interfacing Classification and Ordination	71
VIII	Applications of Ecological Classification	73
	Designing a Classificatory Analysis	73
	Application of Classification to Water Pollution Problems	79
IX	References	98
X	List of Publications	114

FIGURES

No.		Page
1	Sequence of Procedures in Numerical Classification	7
2	Trellis Diagram	8
3	2 x 2 Contingency Table	21
4	Computation of Euclidean Distance	31
5	Classification of Clustering Methods	39
6	Recurrent Groups of Demersal Fishes	44
7	Combinatorial Computation of Inter-Group Distance	46
8	Extensive Chaining in a Single Linkage Clustering	50
9	Effect of Varying Cluster Intensity Coefficient	54
10	Nodal Constancy in a Two-way Table	65
11	Nodal Fidelity in a Two-way Table	67
12	Classification of Intertidal Assemblages around a Sewer Outfall	84
13	Classification of Intertidal Assemblages Impacted by Oil Spill	86
14	Classification of Benthic Assemblages in Hampton Roads	88
15	Classification of Los Angeles - Long Beach Harbor Stations, November 1954	92
16	Nodal Constancy For Species and Site Groups, Los Angeles - Long Beach Harbors	97

TABLES

No.		Page
1	Binary Similarity Coefficients	23
2	Parameters for Combinatorial Clustering Methods	48
3	Classification of Reish's Los Angeles - Beach Harbor Sites	90
4	Classification of Species from Los Angeles - Long Beach Harbor	93

ACKNOWLEDGEMENTS

Dr. Richard C. Swartz of the Environmental Protection Agency had the vision to plan comprehensive reviews of the use of mathematical methods of community analysis in water pollution investigations. He provided me with encouragement, support and guidance during the preparation of this review.

My understanding and approach to numerical classification have been strongly influenced by W. Stephenson and W. T. Williams, two delightful Englishmen whose coincident option for the Australian noonday sun has resulted in major impetus to the application of numerical classification in aquatic ecology.

Kenneth A. Dierks assisted in analysis of data and in literature search. Finally, I thank my associates Robert J. Diaz and David J. Hartzband for reviewing drafts of the report and providing constructive criticism.

SECTION I

CONCLUSIONS

The wide variety of numerical classificatory techniques available is bewildering but affords the investigator a choice of the methods which most appropriately simulate ecologically meaningful criteria. Guidelines for the choice of classificatory strategies based on the efficacy of the various techniques are given but the most appropriate design depends on circumstances and the questions posed.

Currently the most useful and easiest to apply are combinatorial agglomerative clustering methods applied to similarity, distance or correlation resemblance matrices. Polythetic divisive methods are theoretically attractive but are at this point poorly developed and are not widely available.

Data manipulations, including reduction, transformations, and standardizations can profoundly affect the results of a numerical classification. Their use should be justified and only manipulations appropriate to the ecological questions posed should be applied.

Analyses performed on the results of the numerical classification greatly enhance interpretation of the classification and thus ecological insight. In particular, relating normal classifications (of collections) to inverse classifications (of species) in two-way tables, referred to here as nodal analysis, is simple and effective.

Although numerical classification has been used effectively in water pollution investigations, its use is not widespread and most studies have not employed particularly effective techniques. Appropriate classificatory techniques applied with properly designed sampling approaches should prove very useful in future impact assessments.

SECTION II

RECOMMENDATIONS

Ecologists employing numerical classification should become familiar with the wide range of methods available, should understand the strengths and weaknesses of these methods in the analysis of ecological data and should design appropriate sampling and analytical approaches to assess the environmental problem at hand. This task should be made easier by the recent appearance of several texts on the subject of numerical classification as well as by this report.

Computer programs for a wide variety of numerical classification methods should be more available to and useable by the practicing ecologist.

Methodological advances are needed in several areas, notably in polythetic divisive clustering methods, in objective procedures for reallocation of entities after initial classification, and for statistically testing differences among classificatory groups.

The use of numerical classification in water pollution investigations should be encouraged, in particularly where biotic assemblages are diverse and patterns of occurrence complex. However, choice of techniques depends on the investigator's ecological criteria and the circumstances of the study. The approach should be rational rather than routine.

SECTION III

INTRODUCTION

The use of multivariate analytical techniques in community ecology has expanded tremendously in recent years. These techniques have the appeal of objective analysis and simplification of the complex arrays of data generated in field studies. These data arrays typically take the form of measures of abundance of the various species represented in a series of collections. Mental capacity to perceive patterns in such data arrays quickly diminishes with the size of the array, i.e. with the number of collections and the number of species. Thus, except in very limited studies or in cases of extremely low species richness, reproducible procedures for the detection and description of patterns are indeed desirable.

The wide availability of computers have spurred a surge in development of multivariate techniques for the analysis of complex data sets and their wide application in ecology, taxonomy, other biological sciences and in such disparate fields as medicine, criminology, anthropology, geology, remote sensing, engineering and the humanities (Anderberg 1973, Sneath and Sokal 1973, Sokal 1974). As a consequence of the broad-based and rapid development of multivariate analyses, the relevant literature on techniques and applications is diffuse and often obscure. A review of the applications of multivariate analyses in aquatic ecology shows that most practitioners were unaware of or lacked facility with the broad range of techniques now in existence, but instead have been restricted to familiar or readily available techniques. Also, because of the mathematical nature of the techniques and the excessive amount of unstandardized jargon common in the discipline, application of multivariate techniques is often more obfuscating than illuminating to the non-specialist.

The use of multivariate analyses in field ecological

research on man's impacts and in environmental baseline studies is rapidly increasing. Need for a compilation, review and evaluation of the various techniques available was recognized in the development of the Environmental Protection Agency's research program on Biological Indices for Marine Ecosystems. Thus, this critical review was commissioned to assist the Agency in evaluating and conducting research on environmental impacts and to serve as a reference for practicing aquatic ecologists. This report constitutes a general review of numerical classification or cluster analysis. Subsequent reports resulting from research on a new grant (R 804127-01-0) will focus on new developments in numerical classification, on ordination and related techniques, and on applications of techniques.

Several important texts and reviews treating numerical classification have recently been published, some while this report was in preparation. The reader is especially directed to books by Clifford and Stephenson (1975) and Orloci (1975) on ecological applications of multivariate analyses, Sneath and Sokal (1973) on biological (chiefly taxonomic) applications of numerical classification, and Anderberg (1973) and Hartigan (1975) on general applications of cluster analysis. Because of these existing references, no attempt is made here to describe many of the techniques in detail nor to be exhaustive in coverage. Rather an overview summarization of techniques will be provided together with an evaluation of their application in aquatic ecology. Since bothersome terminological differences exist in the diverse literature, frequent cross-referencing of terms will be made.

NUMERICAL CLASSIFICATION

In simplest terms, classification is the ordering of entities into groups or sets on the basis of the relationships of their attributes. Classification is an important biological process which must predate man, but the science of classification has had a fairly recent and parallel development in several disciplines (Sokal 1974).

In ecology the entities most often classified are biological collections or observations. The classification of collections or observations, either conscious or subconscious, is central to the ecologist's conception of communities. Ecologists also classify species on the basis of their ecological attributes. Thus, we think of tropical,

intertidal or demersal species and carnivores or deposit-feeders on the basis of where they occur or what they do.

Numerical classification or cluster analysis encompasses a wide variety of techniques for ordering entities into groups on the basis of certain formal pre-established criteria rather than on subjective and undefined conceptions. Numerical classifications have certain advantages over subjective classifications, notably: (1) they can be based on a much larger number of attributes than is allowed by human mental capacity; and (2) once the classificatory criteria are set, their results are repeatable by any investigator studying the same data set.

It is important to distinguish classification from several other processes and analyses. First, the process of "identification", involving the allocation of additional unidentified entities to the most appropriate class, once such classes have been established (Dagnelie 1971, Sneath and Sokal 1973, Sokal 1974), is here excluded from classification. The use of techniques of numerical identification (e.g. discriminant analysis) both in reallocating members of classes to improve classifications and in assigning new members to classes will be considered in a future report. On the other hand, "dissection", or the optimal splitting of a continuous into a discontinuous series (Clifford and Stephenson 1975), is here considered a case of classification.

Secondly, various multivariate analyses other than numerical classification may be applied to ecological data. These include, in addition to various regression and correlation approaches, a broad group of techniques referred to by biologists as ordination. In ordination the relationships among entities are expressed in a simplified spatial model of a few dimensions, with no attempt to group or draw boundaries between classes (Pielou 1969, Whittaker 1967, Whittaker and Gauch 1973, Sneath and Sokal 1973, Orloci 1975). Ordination includes such techniques as principal components analysis, factor analysis, principal coordinates analysis, correspondence analysis, and multidimensional scaling.

PROCEDURES OF NUMERICAL CLASSIFICATION

To orient the reader to the following sections, a brief description of the chain of procedures in numerical

classifications is in order. Numerical classifications are generally directed by a set of algebraically expressed criteria (an algorithm). This chain of operations begins with the original data, in one or more forms which may be further transformed to conform to certain preconditions. In ecological applications the original data are generally in the form of a matrix of some measure of abundance of each species in a series of collections (Fig. 1). Section IV considers the different forms data may take and reductions or transformations which may be performed before proceeding with a clustering algorithm.

From the original or transformed data matrix most numerical classifications then require the computation of a resemblance measure between all pairs of entities being classified. This is a numerical expression of the degree of similarity, or, conversely, dissimilarity, between the entities on the basis of their attributes. In ecology, the entities being classified may be collections (representing sites, stations, or temporal intervals) with species content as the attributes. This may be referred to as a normal classification as opposed to an inverse classification of species as entities with their presence or abundance in the collections as attributes (Williams and Lambert 1961a). "Normal" and "inverse" are synonymous with the widely used terms "Q analysis" and "R analysis," respectively. However the Q/R distinction has been confused in the past (Ivimey-Cook, Proctor and Wigston 1969) and the normal/inverse terminology is fast becoming standard in ecology. The wide variety of resemblance measures used or proposed are reviewed in Section V.

Matrices of inter-entity resemblance measures are usually required to perform normal or inverse analyses (Fig. 1). These matrices are symmetric in that one corner is the mirror image of the other across the "self-match" diagonal and thus it is necessary to display only half the matrix, as in Fig. 1, as the excluded portion is repetitious. A familiar type of resemblance half-matrix is an inter-city distance finder commonly found on road maps. Resemblance matrices are often presented, sometimes as familiar shade-coded "trellis diagrams" (Fig. 2), in the ecological literature (Macfadyen 1963). From the resemblance matrix one can go further and seek to group entities into groups on the basis of their patterns of resemblance (Fig. 1). This is the essence of clustering. The great variety of clustering methods available are summarized in Section VI.

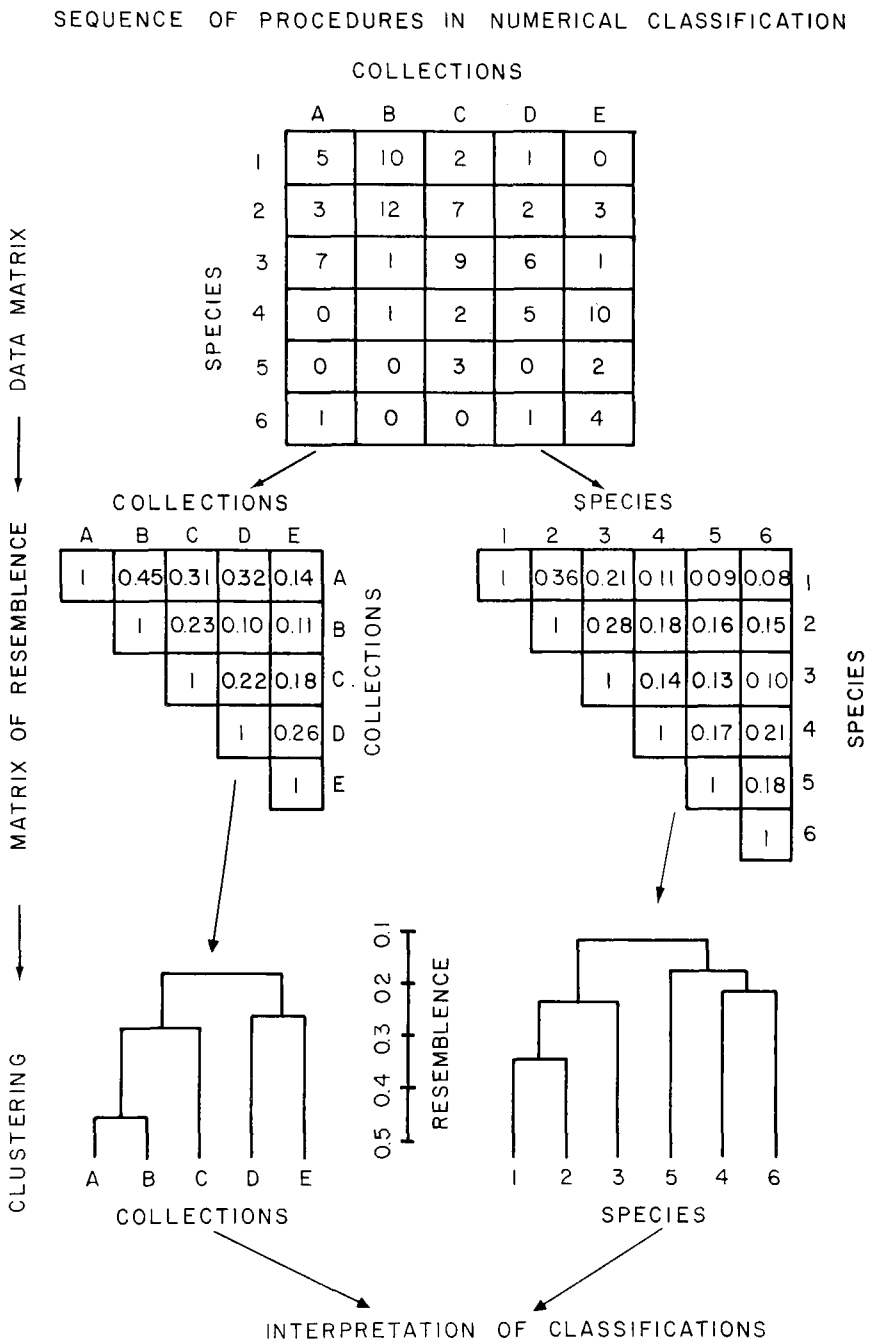


Fig. 1. Illustration of the sequence of procedures in numerical classification.

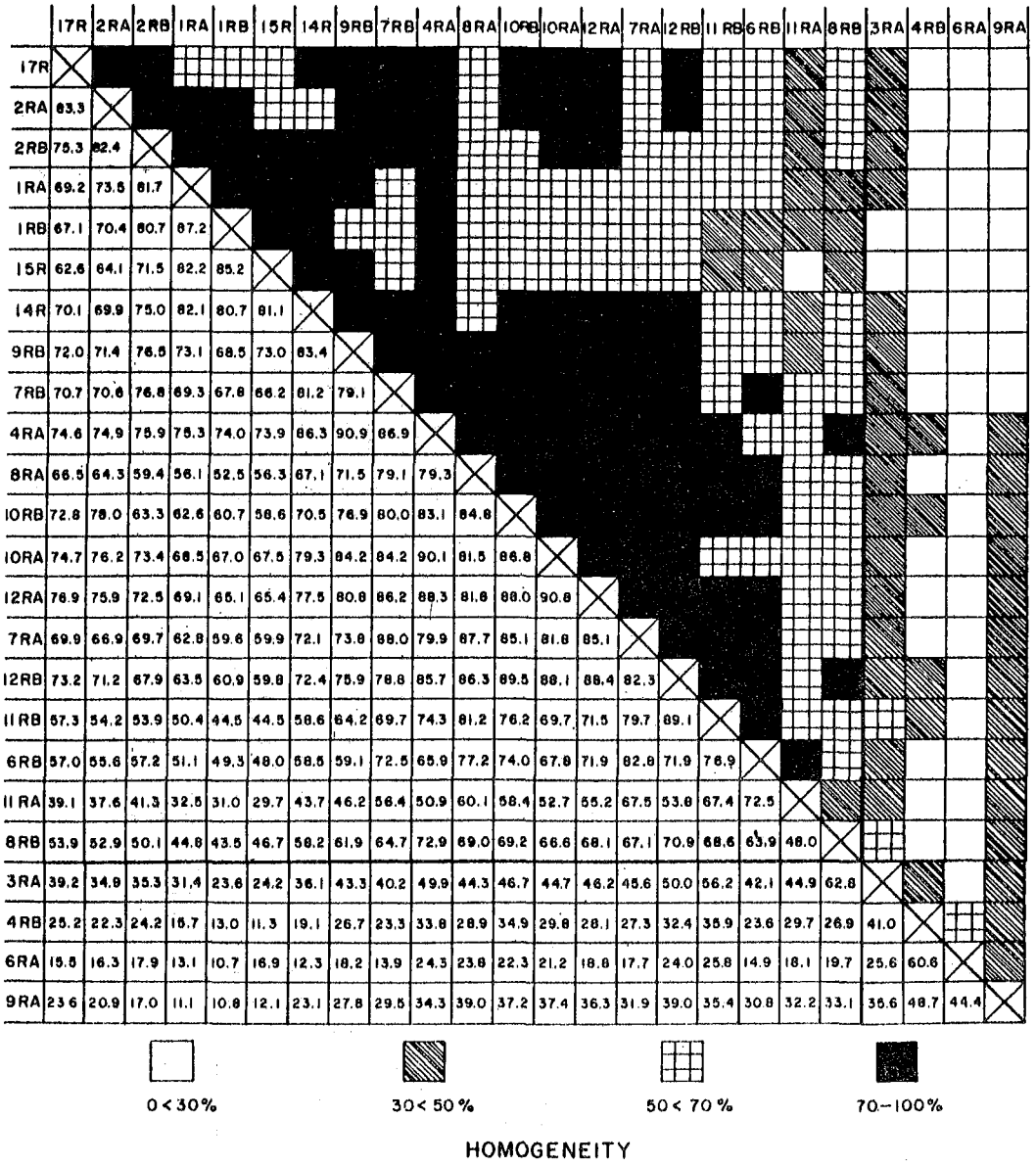


Fig. 2. Example of a "trellis diagram" or a rearranged resemblance matrix with degree of resemblance shade coded (from Sanders, 1960).

All too frequently, the results of numerical classification are presented with painfully little interpretation. Recognizing that classificatory techniques attempt only to simplify complex data sets and not to provide ecological interpretations, post-clustering analyses and interpretive techniques are emphasized in Section VII.

SECTION IV

DATA

Despite the recent proliferation of texts on mathematical ecology (Pielou 1969, 1974, 1975, Sokal and Rohlf 1969, Poole 1974) there is a paucity of comprehensive treatments of the problems of ecological data and data manipulation appropriate to applications of numerical classification. Clifford and Stephenson (1975, Chapters 5 and 7) present a thorough discussion of the types of data and data manipulations ecologists are likely to use with numerical classification. The following discussion of data problems is intended to be complementary to their treatment.

FORMS OF DATA

The usual form of ecological data to which numerical classificatory techniques are applied is the presence or some quantitative measure of importance (numerical abundance, biomass, productivity, rank, etc.) of taxa in collections. However, entities may be classified on the basis of other ecological attributes, for example classifying sites on the basis of abiotic environmental variables.

In general terms, data may be considered to be of one of five basic types (Clifford and Stephenson 1975):
(1) Binary - possessed of two character states, in ecology generally species present or absent.
(2) Disordered multistate - possessing three or more contrasting forms each ranking equal, e.g. red, white, blue.
(3) Ordered multistate - possessing a hierarchy of contrasting forms, which encompasses the total variation in the range of entities under study, e.g. abundant, common, rare.
(4) Ranked - graded within a collection, e.g. most abundant, second most abundant, etc.

(5) Quantitative - quantitative data may be meristic (counts) or continuous (size).

Binary Data

In most ecological applications data will be binary or quantitative. The use of binary data was generally the rule in early ecological applications of multivariate analyses, but the use of quantitative data is growing fast. However, use of binary data is still quite common and in certain applications, e.g. in biogeography, may be the only practical approach. Many ecologists have generally disdained the use of binary data in situations where quantitative data may be collected (Grieg-Smith 1964, p. 160, Clifford and Stephenson 1975, p. 39, Stephenson 1973). Others have noted that, especially if there are many zeros (species absences) in the data matrix, use of binary rather than quantitative data involves loss of relatively little information (Lance and Williams 1967b, Williams et al. 1973).

Actually, the choice between the use of binary or quantitative data involves a decision as to the ecological question asked by the analyst. In a normal analysis resemblance measures based on binary data ask "How similar are the species lists of two collections?" In an inverse analysis the question is "What is the degree of co-occurrence of two species?" Collections may have identical species lists, but vast differences in the relative abundances or dominance of the species, and species may be continuously sympatric, but have distinct habitat preferences.

Quantitative Data

Various types of quantitative data may be used, although the most common are counts or densities (meristic) and biomass (continuous). Other continuous data forms such as productivity, respiration, or cover may also be used. If many replicate samples are taken, frequency of species occurrence in the replicates may be used as an importance measure. Again, the choice of data type is an ecological rather than an analytical question and often the data form is dictated by circumstances. The use of different types of quantitative data, e.g. numerical density versus biomass, may yield vastly different classifications (Clifford and Stephenson 1975, p. 44).

In some ecological situations where analyses of samples is quantitative but sampling effort is inconsistent or unquantified, e.g. with dredge hauls, resulting data may not be quantitatively comparable between collections. In such cases, the investigator not content with basing a classification solely on binary data may express the data as ordered multistate or ranked by using a numerical scoring system. Alternately the data may be collection-standardized (see below) by expressing species importance as percent of the total in the collection.

The non-random form and typically great inequalities of quantitative data bare frequent problems. Thus, one often has to compare very large, and sometimes aberrant, quantities with small quantities to determine resemblance. Data transformations of various types are often used to alleviate this problem. Transformations are increasingly routine in ecological classification, but their application is frequently unthinking or arbitrary and their effects on classifications poorly understood.

DATA REDUCTION

Ecological surveys often generate very large data matrices, due in part to the great abundance of relatively rare species in many communities. Large data matrices are commonly reduced before performing numerical classifications. This is done by the elimination or amalgamation of certain collections or by the elimination of certain species.

Clifford and Stephenson (1975) list three reasons why data reduction may be desirable: (1) to reduce the number of computations, and therefore the resultant expense; (2) to permit the use of certain classificatory strategies which would not otherwise be available because of the mass of data; and (3) to exclude data which have little or no biological meaning.

Most commonly, data matrices are reduced by elimination of species. The simplest and most widely used criterion for elimination is frequency in the collections. Thus, one may eliminate species occurring only once, twice, etc. The rationale is that since the probability of occurrence of very rare species in any given collection is small, co-occurrence relationships of these species may be due more to chance than to similar habitat requirements. The

occurrence of very rare species is often patternless, at least within the limits of reasonable sampling effort. However, before excluding species occurring in less than some arbitrary frequency, the data should be studied for rare species which seem to be habitat-restricted. These should be retained if possible. Alternately, some investigators have excluded species whose overall or maximum abundance fell below a given level (e.g. Day, Field and Montgomery 1971).

Other criteria for exclusion are also possible. Boesch (1976) excluded species on the basis of habitat-constancy. Only species which exceeded a minimum level of overall constancy in the seasonally replicated samples at a site were included in the analysis. Stephenson and his associates have used several more complicated techniques to decide on the elimination of species. These have included: (1) establishing a minimum inter-species resemblance level (i.e. a species must at least have a certain resemblance to another species to be included) (Stephenson, Williams and Lance 1970); (2) sorting out of species which do not contribute much to the overall "pattern" by a divisive monothetic clustering method (see Section VI) (Stephenson, Williams and Lance 1970, Stephenson, Williams and Cook 1972); (3) assessing the contribution of a species to the variance of the data matrix (Williams and Stephenson 1973, Stephenson, Williams and Cook 1974); and (4) testing the conformity of species to predetermined collection groups (Williams and Stephenson 1973; Stephenson, Williams and Cook 1974). Each of Stephenson's techniques tend to accentuate habitat-specificity at the expense of ubiquity. There is a danger of excluding moderately common, ubiguitous species from the analysis, thus yielding an exaggerated "sharpness" in the classification.

In the past, ecologists applying classification to collection data have often been far too cavalier and arbitrary in the elimination of species from data sets. Exclusion criteria ultimately depend on the ecological question one is attempting to pose in the analysis. The intuitive criteria in most cases are themselves multivariate, thus it is reasonable to impose several criteria in making decisions on exclusion. An elaborate attempt to incorporate a variety of criteria was made by Grigal and Ohmann (1975) who ranked species according to six different criteria, including overall frequency, mean abundance, deviation of the standard deviation of their abundance from that predicted from the mean, information content of binary occurrence, contribution

to inter-collection differences, and sums of loadings in an inverse principal components analysis.

Reduction of the data matrix by elimination of collections may be made on more straightforward bases. Collections of doubtful quality, i.e. "bad hauls," may be eliminated on practical grounds. Adjacent samples may be combined if suitable homogeneity exists. Temporal samples from a station may be combined if the primary aim of the analysis is to elucidate spatial patterns while ignoring temporal interactions and, conversely, contemporary samples may be combined over a series of stations to examine overall temporal patterns (Stephenson, Williams and Cook 1974).

TRANSFORMATIONS

Transformations of original data may be suggested because of one or several of the following reasons: (1) ecological collections usually produce large numbers (or biomass) of a few species and small numbers of many; (2) the distribution of species abundance tends to be non-normal; and (3) sampling effort may be inconsistent. It is important to distinguish between two basic types of "transformations": transformations (sensu stricto) and standardizations. Transformations are alterations to the attribute scores (species abundance) of entities without reference to the range of scores within the population as a whole. Common transformations are square root, logarithmic and arcsine (Sokal and Rohlf 1969). Standardizations are alterations which depend on some property of the array of scores under consideration. A common standardization is the conversion of values to percentages, e.g. percent of the total number of individuals in a sample by each species.

Transformations

Perhaps the most common transformation is conversion of species scores into logarithms. Usually, because of the presence of zero scores, the transformation takes the form log (x+1). This transformation may be applied when the mean population estimates are positively correlated with their variance to normalize the distribution of sample estimates (Sokal and Rohlf 1969). Logarithmic transformation has the other very important effect in numerical classification of reducing the discrepancy between large and small values in the computation of resemblance measures. In ecological

terms this reduces the relative contribution of very abundant species to inter-collection resemblance and reduces the relative contribution of high density occurrences to inter-species resemblance. Clifford and Stephenson (1975) present a detailed discussion of the effects of transformations on commonly used resemblance measures.

Other types of transformations are exponential (e.g. $x^{1/2}$, $(x+c)^{1/2}$, where c is a small number, $x^{1/3}$, etc.) and arcsine or angular (especially appropriate to percentages or proportions) transformations. Another type of transformation which has been used with the Canberra metric resemblance measure (see Section V) involves the addition of a small number to all species scores (Stephenson et al. 1972, Boesch 1973) to decrease the relative contribution of 1/0 matches to resemblance.

Standardizations

The most common standardization is by collection total, $(x_i/\Sigma\ x_i$, where x_i is the importance of the i-th species in a collection such that the original species scores become proportions or percentages of the total. Collection total standardization is implicit in the widely used "percentage similarity" (also known to marine ecologists as index of affinity or dominance affinity) as a measure of resemblance (Sanders 1960, Goodall 1973). Collection total standardization is most appropriate when unequal sampling effort disallows direct comparison of absolute abundance data. Alternately, values may be standardized by species total, i.e. species abundance values are divided by the total abundance of the species in all collections $x_j/\Sigma\ x_j$, where x_j is the importance of the species in question in the j-th collection. Clifford and Stephenson (1975) discuss in detail the reasons for applying collection and species standardization and the effects of standardizations on resemblance measures.

Other standardizations which have also been used include (Noy-Meir 1971, Burr 1968): centering by expression of species scores as deviates from the mean quantity of the species in all collections; division by species norm $(\Sigma\ x_j^2)^{1/2}$ or collection norm $(\Sigma\ x_i^2)^{1/2}$; division by collection or species maximum, range, mean or standard deviation; and double standardization by totals or norms of both species and collections. All this may seem less

confusing when one realizes that the familiar product-moment correlation coefficient (Sokal and Rohlf 1969) is entity-mean centered and entity-norm standardized.

Double standardizations have intriguing properties in that they may alleviate scale problems (i.e. comparing large numbers with small) in both normal and inverse analyses using the same transformed data set (Boesch 1973). Bray and Curtis (1957) used a two-step (successive) standardization involving division of scores by the maximum for that species followed by division of the new scores by the total for the collection. Simultaneous double standardizations have been applied by Benzecri (1969), Austin and Noy-Meir (1971), and Boesch (1973). The double standardization used by Austin and Noy-Meir and Boesch produced transformed elements,

$$Y_{ij} = \frac{x_{ij}}{(\sum_j x_{ij} \sum_i x_{ij})^{1/2}},$$

where x_{ij} is the unstandardized value of the i-th species in the j-th collection.

Classifications of the same data with different standardizations can yield strikingly different results (Austin and Grieg-Smith 1968). Standardization involves weighting information from different species or collections in the overall multivariate analysis. The choice of standardization in any particular study is therefore critical and should be based on consideration of the purposes of the classification and the nature of the data, rather than a "cookbook formula" (Noy-Meir 1971).

A case in point is the frequent use of collection total standardizations in "percentage similarity" comparisons. It is common in ecological data sets for abundant species to vary widely in abundance and to be periodically collected in unusually high numbers. The effect of such variations is to cause artificial inter-collection differences in the standardized values of species whose absolute abundances are fairly evenly distributed. Thus standardization by collection total only seems appropriate where sampling effort is variable or unquantified (e.g. with dredge and trawl hauls or an unmetered plankton tow), where there are considerable concordant differences in the abundances of

most species within the collection set, or where monopolization of a habitat is an important ecological criterion (e.g. space cover on rocky shores or fouling plates).

SECTION V

RESEMBLANCE MEASURES

NORMAL VS. INVERSE ANALYSES

The ecological questions posed by normal and inverse analyses are substantially different. A normal resemblance measure expresses the degree of overall "likeness" between assemblages of organisms, and an inverse resemblance measure reflects the similarity in the distribution patterns (spatial or temporal) between species. However, classificatory algorithms can proceed identically for both normal and inverse analyses. Thus, while some authors consider normal resemblance measures separate from inverse measures (Goodall 1973, Anderberg 1973), I will describe the two simultaneously and identically by referring to entities and attributes, with the implicit understanding that in normal analyses collections are the entities and species are the attributes and in inverse analyses species are entities and collections are attributes.

GENERAL

Large numbers of resemblance measures have been proposed in the literature and many have been more or less restricted to certain disciplines, e.g. numerical taxonomy, social sciences, etc. It far exceeds the scope of this report to list even all of those that have been used in ecology. Instead, only those measures which have been used in aquatic ecological investigations or show promise for application are treated and reference is made to their application in the literature. This section should serve only as a starting point for the reader interested in application of one or more resemblance measures. For more exhaustive discussions one should consult Sneath and Sokal (1973, Chap. 4), Anderberg (1973, Chaps. 4 and 5), Goodall (1973), Clifford and Stephenson (1975, Chap. 6), and Orloci (1975, Chap. II).

As orientation to the notation used in this summarization of resemblance measures, consider the following m x n data matrix, whose n columns represent the n entities to be grouped on the basis of resemblances and whose m rows are m unit attributes. Each entry x_{ij} in such a matrix is the score of entity j for attribute i.

ATTRIBUTES (Normal:species; Inverse:collections)	ENTITIES (Normal:collections; Inverse:species)			
	1	2	. . .	n
1	$x_{1,1}$	$x_{1,2}$	. . .	$x_{1,n}$
2	$x_{2,1}$	$x_{2,2}$	. . .	$x_{2,n}$
.				
.				
.				
m	$x_{m,1}$	$x_{m,2}$	. . .	x_{mn}

Other authors use different symbolism and terminology, thus the expression of similarity measures contained herein may appear different in other sources. Note that the entity-attribute terminology is consistent with that of Clifford and Stephenson (1975) except that they refer to inverse classifications as clustering of attributes, whereas I prefer to switch the entity-attribute distinction depending on the type of analysis. "Entity" and "attribute" may be considered equivalent to Sneath and Sokal's (1973) "OTU" and "character," respectively, and Anderberg's (1973) "data unit" and "variable," respectively.

Various taxonomies of resemblance measures are also used in the texts listed above. In most, divisions among some of the types of measures is rather arbitrary and some of the authors apply identical terms to different types of measures. The terminology used here is modified from Clifford and Stephenson (1975) by referring to their "coefficients of association" as "correlation coefficients" to remove the ambiguity with Sneath and Sokal's (1973) use of "association coefficients." Thus, I refer to (1) similarity coefficients as those measures constrained between 0 and 1, (2) correlation coefficients as those constrained between -1 and 1, (3) Euclidean distance, (4) information content measures, and (5) probabilistic measures.

SIMILARITY COEFFICIENTS

As used here, similarity coefficients are those resemblance measures which are 1 (or very close to it) when entities are identical and 0 (or very close to it) when entities have no attributes in common. Many authors have expressed similarity in percentages, in which case the value range from 100 to 0. The complement of similarity (1-S, if S is a similarity coefficient) is dissimilarity (D). Some investigators use the concept of dissimilarity rather than similarity and compute dissimilarity coefficients for the sake of operational ease. Also, dissimilarity can be considered analagous to inter-entity distance, allowing the use of dissimilarity measures in certain clustering and ordination techniques based on Euclidean distances (see below). I use similarity rather than dissimilarity here because it seems intuitively clearer to most ecologists, but it is a very simple matter to convert one to the other, e.g. if S = 0.6, D = 0.4 and vice versa.

Qualitative Similarity Coefficients

Coefficients of comparison of entities based on binary data (i.e., species presence or absence) can be conveniently explained using the symbolism of a 2 x 2 contingency table which lists the frequencies of agreement and disagreement of their binary attributes. The general form of the 2 x 2 contingency and the meanings of its elements in ecological terms in both normal or inverse analysis are given in Fig. 3. Note that the sum a+c is the total number of positive attributes (occurrences) for entity 2, the sum a+b is the total number of positive attributes for entity 1, and the sum a+b+c+d is the total number of attributes for which entities have been compared.

Table 1 lists the commonly used similarity coefficients for binary data and some of their properties and constraints. The first coefficient, the simple matching coefficient, differs from the others in the inclusion in the expression of d, the number of joint absences or "double zero matches." As Clifford and Stephenson (1975) point out, in many circumstances it would seem ridiculous to regard two entities as similar largely on the basis of them both lacking something. With most ecological data sets joint absences of species has relatively little meaning, given the rarity and contagious distribution of some species, and for this reason similarity coefficients involving conjoint absences are

2 x 2 CONTINGENCY TABLE

GENERAL

ENTITY 2 \ ENTITY 1	1	0
1	a	b
0	c	d

NORMAL ANALYSIS

COLLECTION B \ COLLECTION A	1	0
1	NO. SPECIES IN COMMON	NO. SPECIES IN B BUT NOT A
0	NO. SPECIES IN A BUT NOT B	NO. SPECIES NOT REPRE-SENTED IN A OR B

INVERSE ANALYSIS

SPECIES 2 \ SPECIES 1	1	0
1	NO. OF CO-OCCURRENCES	NO OF OCCURRENCES OF 2 WITH-OUT 1
0	NO. OF OCCURRENCES OF 1 WITH-OUT 2	NO. OF TIMES NEITHER 1 OR 2 OC-CURRED

Fig. 3. 2 x 2 contingency tables showing elements a, b, c and d used in computation of binary similarity coefficients.

generally not used in ecology (Green 1971, Field 1971). However, when most species are common or where there is a high degree of species fidelity to particular collection types, the simple matching coefficient may be useful.

Only three of the coefficients listed have been frequently used in aquatic ecology--Jaccard, Dice and Fager coefficients. The Jaccard and Dice coefficients are simple and similar, with the difference that the Dice measure doubly weights shared positive attributes (joint presences), and thus will always be greater than or equal to the Jaccard measure. Column 4 in Table 1 suggests that in the case of disparate number of positive attributes, the Dice coefficient yields more intuitively accurate values. Furthermore, Clifford and Stephenson (1975) offer that in cases where there are relatively few conjoint presences in the data set the Dice coefficient is more attractive, and with relatively many conjoint presences the Jaccard coefficient is more attractive because it will give a wider spread of values in the upper end of the range. Goodall (1973) shows that the sampling distribution of the Dice coefficient is slightly more biased than that for the Jaccard coefficient.

The Fager coefficient has been widely used in marine ecology primarily because its author was active in that field. The Fager coefficient is the Ochiai coefficient (Table 1) modified by subtraction of a "correction factor" which means that the measure is not constrained between 0 and 1; rather with no shared positive attributes the measure is slightly less than 0 and with identical entities the measure is slightly less than 1. Because of these and other undesirable properties Field (1971) and Clifford and Stephenson (1975) raise objections to the use of the Fager coefficient. The incorporation of a geometric mean term in the denominator of the Ochiai coefficient does make this "uncorrected" form of the Fager coefficient more attractive than the Jaccard coefficient when the entities have a disparate number of positive attributes (Sepkoski and Rex 1974).

In summary, the most attractive similarity measures for binary ecological data appear to be the Jaccard, Dice and Ochiai coefficients. The selection of the most appropriate coefficient depends on the nature of the data. If the task is to discriminate relationships among closely similar entities one might choose the Jaccard coefficient. If, on the other hand, the entities vary widely in their number of positive attributes (e.g. rich and poor collections in a normal analysis or common and rare species in an inverse

Table 1. COMMONLY USED BINARY SIMILARITY COEFFICIENTS AND THEIR PROPERTIES. VARIABLES AS IN FIG. 3. EXPRESSIONS IN COLUMN 1 RESULT WHEN TWO ENTITIES HAVE THE SAME NUMBER OF POSITIVE ATTRIBUTES; THOSE IN COLUMN 2 WHEN THEY SHARE NO POSITIVE ATTRIBUTES; THOSE IN COLUMN 3 WHEN THEY ARE IDENTICAL; AND THOSE IN COLUMN 4 WHEN ONE SAMPLE HAS TWICE AS MANY POSITIVE ATTRIBUTES AS THE OTHER AND THE NUMBER OF ATTRIBUTES IN COMMON IS ONE-HALF THE NUMBER IN THE ENTITY WITH THE FEWER POSITIVE ATTRIBUTES. ASSUME $b \geq c$. MODIFIED FROM VALENTINE (1973) AND CLIFFORD AND STEPHENSON (1975).

Coefficient	1 $a+c = a+b$	2 $a = 0$	3 $a = a+b+c$	4 If $\frac{a+c}{a+b}=1/2$ and $\frac{a}{a+c}=1/2$
(1) Simple matching $\frac{a+d}{a+b+c+d}$	$\frac{a+d}{a+2c+d}$	$\frac{d}{b+c+d}$	1	$\frac{a+d}{5a+d}$ [if a=d, then=1/3]
(2) Jaccard (=Iverson) $\frac{a}{a+b+c}$	$\frac{a}{a+2c}$	0	1	1/5
(3) Dice (=Sørensen, Czekanowski) $\frac{2a}{2a+b+c}$	$\frac{a}{a+c}$	0	1	1/3
(4) Kulczynski first $\frac{a}{b+c}$	$\frac{a}{2c}$	0	∞	1/4

Table 1 (continued). COMMONLY USED BINARY SIMILARITY COEFFICIENTS AND THEIR PROPERTIES.

Coefficient	1 $a+c = a+b$	2 $a = 0$	3 $a = a+b+c$	4 $\frac{a+c}{a+b}$ and $\frac{a}{a+c}=1/2$
(5) Kulczynski second $\frac{a}{2}\left(\frac{1}{a+b}\ \frac{1}{a+c}\right)$	$\frac{a}{a+c}$	0	1	3/8
(6) Simpson $\frac{a}{c}$	$\frac{a}{a+c}$	0	1	1/2
(7) Braun-Blanquet $\frac{a}{b}$	$\frac{a}{a+c}$	0	1	1/4
(8) Ochiai (=Otsuka) $\frac{a}{\sqrt{(a+b)\ (a+c)}}$	$\frac{a}{a+c}$	0	1	$1/\sqrt{8}$
(9) Fager $\frac{a}{\sqrt{(a+b)\ (a+c)}}-\frac{1}{2\sqrt{a+b}}$	$\frac{a}{a+c}-\frac{1}{2\sqrt{c}}$	$-\frac{1}{2\sqrt{a+b}}$	$1-\frac{1}{2\sqrt{a+b}}$	$(1/\sqrt{8})-(1/4\sqrt{c})$

analysis) one should choose the Dice or Ochiai. Another advantage of the Dice coefficient is that it is the binary equivalent of the most commonly used quantitative similarity measure, the Bray-Curtis or Czekanowski coefficient (see below).

Quantitative Similarity Coefficients

As in the case of binary similarity measures, many quantitative similarity coefficients have been proposed or employed, although only a handful have been applied in aquatic ecology. An important class of quantitative similarity coefficients are derivatives of metric distance functions (Minkowski metrics) whose general form can be stated as

$$D_{jk} = (\sum_i |x_{ij}-x_{ik}|^p)^{1/p}. \tag{10}$$

In particular, coefficients are derived from the Manhattan metric in which p = 1, thus

$$D_{jk} = \sum_i |x_{ij}-x_{ik}| \tag{11}$$

Metrics in their basic forms are unconstrained (they range from zero to infinity) and are distance rather than similarity measures. The metric derivatives discussed here are expressed as constrained similarity/dissimilarity coefficients.

Bray-Curtis Coefficient -

The Bray-Curtis similarity coefficient (Clifford and Stephenson 1975) is perhaps the most widely employed quantitative measure in ecology. It can be expressed as a similarity or dissimilarity measure:

$$S_{jk} = \frac{2 \sum_i \min (x_{ij}, x_{ik})}{\sum_i (x_{ij}+x_{ik})} = 1 - D_{jk} \tag{12}$$

$$D_{jk} = \frac{\sum_i |x_{ij}-x_{ik}|}{\sum_i (x_{ij}+x_{ik})} = 1 - S_{jk} \tag{13}$$

This measure has been often referred to as the Czekanowski coefficient (Field 1971, Day, Field & Montgomery 1971) as it is a quantitative extension of the binary similarity coefficient used by Czekanowski (1909) and referred to above as the Dice coefficient. If the scores are standardized by entity-total (i.e. expressed as proportion or percent), Bray Curtis similarity becomes "percentage similarity" widely used in American plant ecology and made popular in marine ecology by Sanders' (1960) application in the study of a marine benthic community. (Sanders refers to the coefficient as "dominance affinity"). If the scores are expressed as proportion of the total-for the entity ($p_{ij} = x_{ij} / \Sigma x_{ij}$), then

$$S_{jk} = \sum_i \min (p_{ij}, p_{ik}) \tag{14}$$

or the sum of the minimum proportions (or percentages) of each attribute.

The Bray-Curtis coefficient both in its unstandardized and "percent standardized" forms has been extensively used in marine ecology. Some examples are Barnard (1970), Bloom, Simon and Hunter (1972), Day et al. (1971), Eagle (1973, 1975), Field (1970, 1971), Field and MacFarlane (1968), Gage (1974), Hartzband and Hummon (1974), Kay and Knights (1975), Markle and Musick (1974), Mauchline (1972), Nichols (1970), Sanders (1960), Sanders and Hessler (1969), Santos and Simon (1974), Stephenson and Williams (1971), Stephenson, Williams and Cook (1972), Wade (1972), Ward (1973), and Warwick and Gage (1975).

Ruzicka Coefficient -
A variant of the Bray-Curtis coefficient was proposed by Ruzicka (1958) and is expressed as

$$S_{jk} = \frac{\sum_i \min (x_{ij}, x_{ik})}{\sum_i (x_{ij}+x_{ik}) - \sum_i \min (x_{ij}, x_{ik})} = \frac{\sum_i \min (x_{ij}, x_{ik})}{\sum_i \max (x_{ij}, x_{ik})} . \tag{15}$$

The dissimilarity measure thus becomes

$$D_{jk} = \frac{\sum_i |x_{ij}-x_{ik}|}{\sum_i \max\ (x_{ij}+x_{ik})} = 1 - S_{jk} . \qquad (16)$$

The difference between this and the Bray-Curtis coefficient is that the Ruzicka measure divides the sum of the minimum shared attributes by the sum of the maximum attribute scores whereas in the Bray-Curtis measure the sum of the minimums is divided by the sum of the average (between the two entities) attribute scores. Because of this the Ruzicka coefficient is more affected by large differences, thus high attribute scores, which makes it less sensitive in the middle range of resemblance than the Bray Curtis coefficient. Despite the drawback the coefficient has recently been used by Dutch marine phytoecologists (Colijn and Koeman 1975, Van den Hoek, Cortel-Breeman and Wanders 1975).

Canberra metric Coefficient -

A principal difference between the Bray-Curtis similarity coefficient and the aforementioned binary similarity coefficients is the effect of size of the score on the measure. In the Bray-Curtis coefficient and many other quantitative resemblance measures, attributes with high scores largely determine the value of the measure whereas attributes with low scores are relatively unimportant. In ecological terms this means that abundant species largely determine inter-collection (normal) resemblance and dense occurrences largely determine inter-species (inverse) resemblance. Indeed in many ecological circumstances this might be an intuitively appealing characteristic, but in others it may be tantamount to basing inter-collection resemblance on only one or two species. To overcome this characteristic of quantitative metric and correlation measures Lance and Williams (1966, 1967b) proposed the Canberra metric coefficient which is usually expressed in its dissimilarity form

$$D_{jk} = \frac{1}{m} \sum_i \frac{|x_{ij}-x_{ik}|}{(x_{ij}+x_{ik})} \qquad (17)$$

The similarity form of the coefficient is

$$S_{jk} = \frac{1}{m} \sum_i \frac{2 \min (x_{ij}, x_{ik})}{(x_{ij}+x_{ik})} = 1 - D_{jk} \qquad (18)$$

Thus, the Canberra metric is the average of a series of fractions representing the inter-entity agreement of each attribute and, as such, has a built-in attribute standardization. An outstandingly large attribute score can contribute to only one of the summed fractions and so does not dominate the coefficient. In this regard the Canberra metric coefficient can be considered intermediate between other quantitative similarity, distance and correlation measures and binary resemblance measures.

The incorporation of zero scores in the Canberra metric is subject to certain conventions (Clifford and Stephenson 1975). Double zero matches (i.e. when attribute scores of both entities being compared are zero) are usually ignored for the same reasons that binary coefficients incorporating the joint absence contingency are disfavored. Thus the appropriate divisor is not m, the total number of attributes, but m-r where r is the number of double zero comparisons. Secondly, since when one of the attribute scores is zero the fraction contributed to the sum is one, small numbers may be substituted for zero in the case of single zero matches to ensure a greater contribution to dissimilarity of an attribute difference of 1000 to 0 than of a difference of 1 to 0, for example.

If applied to binary data, with the supression of double-zero matches, the Canberra metric reduces to S=a/a+b+c, i.e. the Jaccard coefficient.

Use of the Canberra metric in aquatic ecology to date has been confined to associates of the Canberra (Australia) school of numerical classification, e.g. Boesch (1973) and Stephenson et al. (1972).

Morisita Coefficient -
The Morisita coefficient (Morisita 1959) is not derived from metric distance functions, rather it is related to both correlation and information content resemblance measures. However, since it ranges from 0 (no resemblance)

to $\pm$ 1 (identity) it is here considered a similarity coefficient. The coefficient, often referred to as C_λ or C_δ is given by

$$S_{jk} = \frac{2 \sum_i x_{ij} x_{ik}}{(\lambda_j + \lambda_k) \sum_i x_{ij} \sum_i x_{ik}} \tag{19}$$

where

$$\lambda_j = \frac{\sum_i x_{ij} (x_{ij} - 1)}{[\sum_i x_{ij} ((\sum_i x_{ij}) - 1)]} \tag{20}$$

and

$$\lambda_k = \frac{\sum_i x_{ik} (x_{ik} - 1)}{[\sum_i x_{ik} ((\sum_i x_{ik}) - 1)]} \tag{21}$$

The terms λ_j and λ_k are the Simpson (1949) diversity measures for the attributes of entities j and k, respectively.

The basic term of this coefficient, as in other correlation coefficients, is the product of the two attribute scores being compared rather than the differences in the two scores as in coefficients derived from metrics. This leads to a heavier weighting of the importance of attributes with high schores than, say, the Bray-Curtis coefficient. Thus, the Morisita coefficient can be expected to reflect primarily the resemblance of scores of the most abundant species in a normal analysis and the resemblance of outstanding abundances of species in an inverse analysis. On the other hand, correlations, in general, are less influenced by scale differences between entities than are the "metric" expressions, which are based on differences in attribute scores. In ecological terms this means that in an inverse comparison a usually abundant species will have low resemblance to a species which is usually not very abundant,

even if their abundances are correlated, when a "metric" derived measure is used but may have high resemblance on the basis of the Morisita coefficient.

The Morisita coefficient has been used in marine ecology by Barnard (1970), Bloom et al. (1972), Livingston (1975) and Ono (1961).

EUCLIDEAN DISTANCE

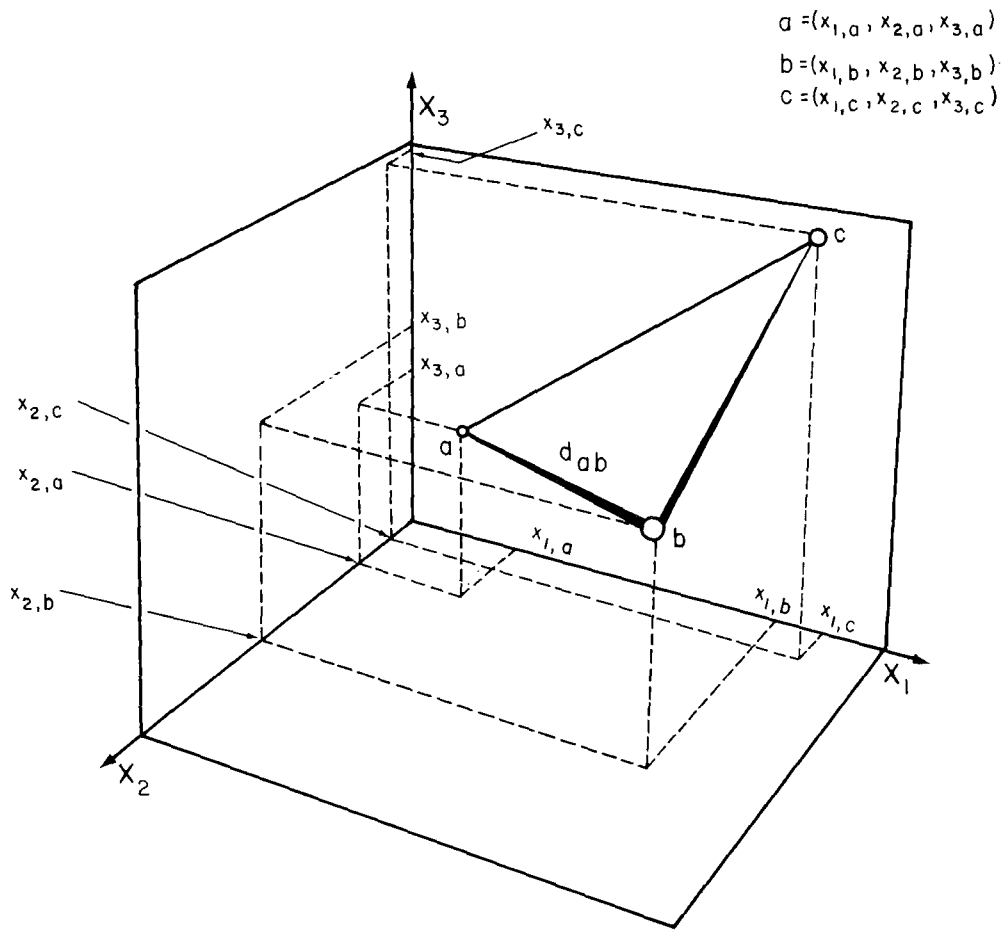
If an entity is construed to be represented by a point in an m dimensional space each dimension of which corresponds to an attribute and is orthagonal (at a right angle) to the other dimensions (or axes), the Euclidean distance is the linear distance between any two points (entities) in that hyperspace. The coordinates of the m axes are the scores of the attributes represented by the axes and the distance between two points can be computed as the square root of the sums of the squared differences between attribute scores,

$$D_{jk} = [\sum_i (x_{ij} - x_{ik})^2]^{1/2} \qquad (22)$$

You may recognize this as a Minkowski metric (Equation 10) where p = 2. Euclidean distance may, of course, range from 0 (when entities are identical) to infinity. Either Euclidean distance itself or its square may be used as the distance measure.

The concept and computation of Euclidean distance may be made clearer by consideration of Fig. 4 which depicts the spatial relationship of three points (entities) in three dimensions. The distance between any two points can be computed by squaring the difference of their coordinates on each axis, summing those squared values and taking the square root of the sum. This can be expanded to additional dimensions with the addition of attributes. The squared differences between the scores of each additional attribute can simply be added on to the squared distance.

Because differences between attribute scores are squared, Euclidean distance heavily weights attributes with high scores and worsens the scale problem between high scoring and low scoring entities compared to the Bray-Curtis



$$d_{ab}^2 \quad (x_{1,a} \quad x_{1,b})^2 + (x_{2,a} \quad x_{2,b})^2 + (x_{3,a} \quad x_{3,b})^2$$

Fig. 4. Illustrations of the computation of Euclidean distance between entities defined by coordinates representing three attributes.

coefficient (Orloci 1967, 1973, 1975, Clifford and Stephenson 1975). Thus, in ecological applications Euclidean distance may place overemphasis on dominance or outstanding abundances and may result in artificially high resemblance between entities which do not have many attributes in common but whose attribute scores are low. To overcome these weaknesses transformations or standardizations are usually applied to the data. Williams and Stephenson (1973) and Stephenson, Williams and Cook (1974) used a cube-root transformation. Standardizations by entity-norm (Orloci 1967, Pielou 1969, Noy Mier 1971) and by species variance (Hughes and Thomas 1971a, b) are frequently used.

With binary data the Euclidean distance reduces to $D = \sqrt{b+c}$, using the notation of the 2 x 2 contingency table (Clifford and Stephenson 1975).

Marine ecological applications of Euclidean distance as a resemblance measure include Holland and Dean (1976), Hughes and Thomas (1971a, b), Polgar (1975), Stephenson et al. (1974) and Williams and Stephenson (1973).

CORRELATION COEFFICIENTS

Most correlation coefficients range from -1 (completely dissimilar) to +1 (completely similar). Many, but not all, are based on a probabilistic model, offering the potential advantage of testing the significance of resemblance. However, it is only appropriate to apply tests of significant correlation between species and not collections, because of the assumptions of independence in the tests. Even then assumptions of the parametric significance tests (normality, randomness, etc.) are seldom met and one should show caution in interpreting the results of tests of significance of interspecies correlations.

Binary Correlation Coefficients

Two different binary correlation coefficients have been used in aquatic ecological investigations. The point correlation coefficient (also referred to as Kendall's coefficient of association) (Looman and Campbell 1960, Goodall 1973) as given in the standard terms of the 2 x 2 contingency is

$$r = \frac{ad - bc}{[(a+b)\ (c+d)\ (a+c)\ (b+d)]^{1/2}} \tag{23}$$

The significance of r can be tested by a Chi-square comparison (Looman and Campbell 1960) but as discussed earlier the meaning of this significance test is dubious, particularly since the test assumes that a+b, a+c and d are known, when these variables are in practice subject to sampling error. The point correlation coefficient was employed by Lie and Kelley (1970) and Nichols (1970) in studies of marine benthic communities.

The second binary correlation coefficient was proposed by McConnaughy (1964) in an analysis of planktonic communities:

$$r = \frac{a^2 - bc}{(a+b)\ (a+c)} \tag{24}$$

Quantitative Correlation Coefficients

A commonly used resemblance measure, particularly in inverse analyses, is the product-moment correlation coefficient (Sneath and Sokal 1973, Goodall 1973, Clifford and Stephenson 1975)

$$r_{jk} = \frac{\sum_i (x_{ij} - \bar{x}_j)\ (x_{ik} - \bar{x}_k)}{[\sum_i (x_{ij} - \bar{x}_j)^2\ \sum_i (x_{ik} - \bar{x}_k)^2]^{1/2}} \tag{25}$$

where $\bar{x}_j$ and $\bar{x}_k$ are the mean values of all m attributes of entities j and k, respectively. This is the entity-mean centered and entity-norm standardized form of the general correlation expression and other forms of quantitative correlation are possible (Noy-Meir 1973a).

In summary, although the product-moment correlation coefficient is useful in expressing the relationship of the shape of species distribution patterns over a series of collections in an inverse analysis, the correlation coefficient suffers from several undesirable characteristics (Clifford

and Stephenson 1975, Field 1970, Orloci 1967, 1973, Sneath and Sokal 1973). It tends to exaggerate the contribution of outstandingly large scores to resemblance. It can suggest spurious patterns of resemblance if there are many zero values in the data matrix (a common condition except when only abundant or ubiquitous species are included in the analysis). Finally, since it is a "shape" measure and not a "size" measure, perfect correlations can occur between nonidentical entities. Furthermore, tests of significance of inter-entity correlation should be applied with caution. In particular, it is inappropriate to apply probablistic tests to inter-collection correlation because the attributes (species) do not represent a "variable" in the statistical sense and they may not be independent (Sneath and Sokal 1973, Clifford and Stephenson 1975). Moreover, tests of significance of species correlation assume normal frequency distributions and linear relationships between species scores--conditions which often do not obtain in ecological data.

The product-moment correlation coefficient has been used extensively in marine ecological applications of numerical classification, e.g. Angel and Fasham (1973), Chardy (1970), Ebeling et al. (1970), Eisma (1966), Jones (1969), and Mauchline (1972) and has received even wider use in applications of principal components and factor analyses (types of ordination).

INFORMATION CONTENT MEASURES

The term information is here used in a strictly technical context and relates more to the degree of uncertainty or surprise than to knowledge (Orloci 1969, 1971). The measures discussed here have the same information theoretical basis as the familiar Shannon diversity measure. We can express the information content of the attribute scores of an entity as

$$I_j = (\sum_i x_{ij}) \log (\sum_i x_{ij}) - \sum_i x_{ij} \log x_{ij} \qquad (26)$$

using the same notation, x_{ij}, for the elements of the data matrix. The information content, I_k, of another entity k can be similarly computed. The information content of the combined pair of entities can be expressed:

$$I_{j+k} = (\sum_i x_{ij}+x_{ik}) \log (\sum_i x_{ij}+x_{ik}) - \sum_i (x_{ij}+x_{ik}) \log (x_{ij}+x_{ik}) \quad (27)$$

The increase in information from that represented by I_j and I_k to that represented by I_{j+k} can be used as a distance measure

$$\Delta I_{j,k} = I_{j+k} - (I_j + I_k) \ . \quad (28)$$

The common or mutual information between the two entities can alternately be expressed:

$$I_{jk} = I_j + I_k - I_{j+k} \ . \quad (29)$$

Similarily, the information content of arrays of binary attributes may also be computed by one of several methods (Lambert and Williams 1966, Dale and Anderson 1972, Clifford and Stephenson 1975) and information gain or, conversely, mutual information can be calculated.

Inter-entity matrices of information content resemblance measures may be passed directly to clustering algorithms which form groups on the basis of the resemblance matrix alone (combinatorial clustering strategies). Alternately, clustering may take place by procedures which require recomputation, following consultation of the data matrix, of the information gain for each clustering iteration (non-combinatorial strategies). These clustering strategies are more fully discussed in Section VI.

Information content resemblance measures have been little used in marine ecology, although they have been widely applied in plant ecology and taxonomy (Sneath and Sokal 1973, Clifford and Stephenson 1975). Stephenson and Williams (1971) in a study of marine benthos attempted the use of agglomerative information-gain clustering using both quantitative and binary data but were dissatisfied with their results. Stephenson et al. (1971, 1972) and Moore (1973) used the divisive information analysis DIVINF (Section VI) in studies of marine benthos.

Several resemblance measures which do not express information content *per se* but which incorporate information terms have also been used in ecology. Horn (1966) proposed a measure of inter-entity "overlap" which is an information analog of the Morisita coefficient

$$S_{jk} = \sum_i [(p_{ij}+p_{ik})\log(p_{ij}+p_{ik}) - \Sigma p_{ij}\log\Sigma p_{ij} - \Sigma p_{ik}\log p_{ik}] / [(\sum_i x_{ij}+\sum_i x_{ik})\log(\sum_i x_{ij}+\sum_i x_{ik}) - (\sum_i x_{ij})\log(\sum_i x_{ij}) - (\sum_i x_{ik})\log(\sum_i x_{ik})] \quad (30)$$

where $p_{ij} = x_{ij}/\sum_i x_{ij}$ and $p_{ik} = x_{ik}/\sum_i x_{ik}$.

The measure is constrained between 0 and 1 and is appropriately classed as a similarity coefficient. Horn's overlap coefficient was used by Kohn (1968) to study ecological relationships among marine snails of the genus *Conus* and Bloom et al. (1972) in a study of intertidal benthos. Other information measures have been similarly used to express "niche overlap" or "habitat overlap" (Colwell and Futuyma 1971, Pielou 1975) between pairs of species. Finally, Hummon (1974) formulated a complex similarity coefficient which is a mixture of components of percentage similarity (Bray-Curtis coefficient), mutual information and the Fager similarity coefficient, and applied it in a study of marine gastrotrich taxocenes.

PROBABILISTIC MEASURES

In addition to the correlation coefficients discussed above, several other measures may be employed to test differences between pairs of entities. As with the correlation coefficients, however, their use as a probabilistic test of significant differences between pairs of collections is questionable and they are most often applied in the inverse case of testing the significance of associations between pairs of species.

Since, except for the correlation coefficient, no other probabilistic measures have been used much in numerical

classification, I will not elaborate on the methods. Several appropriate techniques are thoroughly reviewed by Pielou (1969, 1974). Chi-square tests of binary occurrence data bsed on the 2 x 2 contingency table are the most commonly used methods (Pielou 1969, 1974).

SECTION VI

CLUSTERING METHODS

GENERAL

It was common in earlier years and it remains an occasional practice in ecology to simply present resemblance matrices or trellis diagrams as the end point of a multivariate analysis. Frequently, the elements of the resemblance matrix are rearranged so that the highest resemblance scores are closest to the diagonal of the half-matrix, i.e. to rearrange the order of entities so that they are close to those entities they most resemble. Usually this is done by eye, although Lie and Kelley (1970) presented a procedure for the rearrangement of the resemblance matrix by objective criteria. Some investigators have attempted to draw, more or less by eye, a simple spatial model or "plexus" of the patterns of inter-entity relationships based on the resemblance matrix. Such matrix and plexus techniques (McIntosh 1973) are more appropriately considered forms of ordination rather than classification. Rather, this section treats numerical procedures by which entities can be objectively grouped based on their resemblances.

CLASSIFICATION OF CLUSTERING METHODS

Various classifications of clustering methods have been proposed (Pielou 1969, Williams 1971, Sneath and Sokal 1973) and the dichotomized scheme presented in Fig. 5 encompasses most of their salient features.

Exclusive versus Non-Exclusive

An exclusive classification is one in which an entity may occur in only one group while in non-exclusive

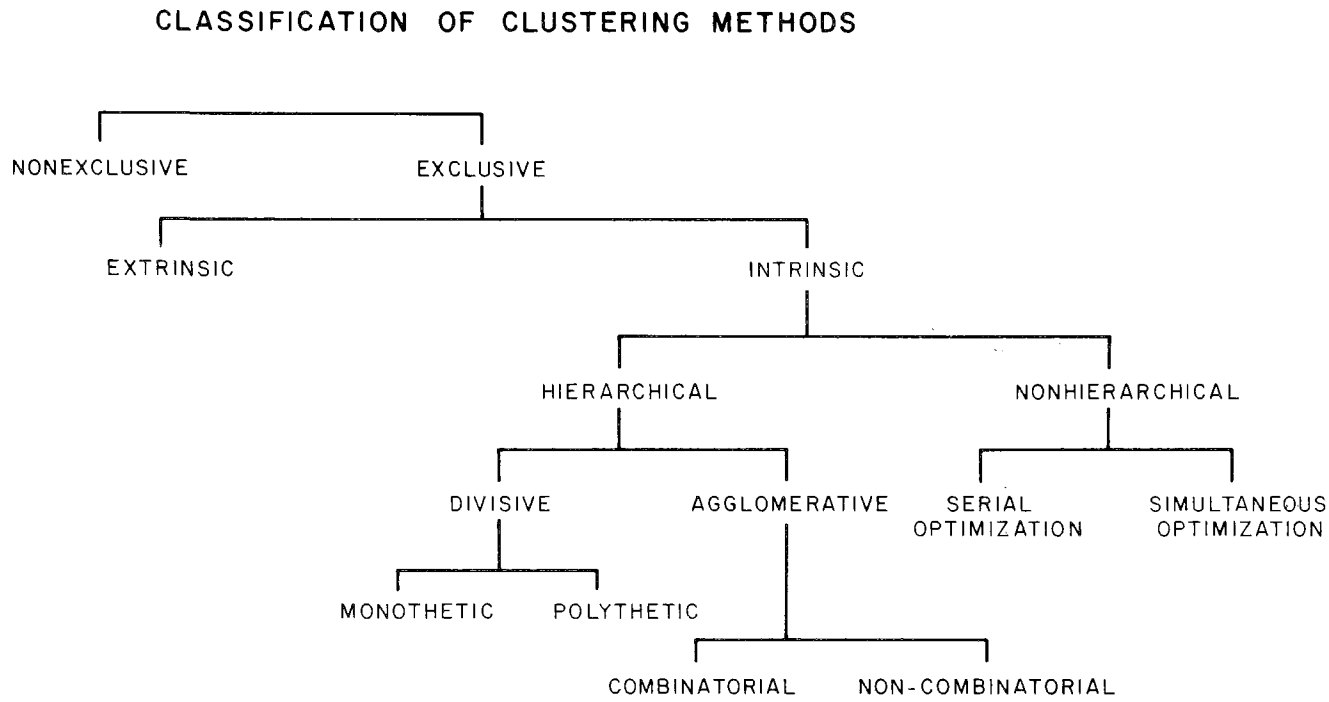


Fig. 5. A classification of clustering methods as discussed in text (from Williams 1971).

classifications entities may be members of more than one group. Sneath and Sokal (1973) use the terms non-overlapping and overlapping as synonymous with exclusive and non-exclusive. Although in certain cases the use of non-exclusive classifications in ecology may make some sense, they have not been used except in a very few cases (e.g. Yarranton et al. 1972) and will not be discussed further.

Extrinsic versus Intrinsic

Intrinsic classifications form groups based solely on their attributes whereas in extrinsic classifications the resulting groups, although based on internal attributes, are required to reflect predetermined external attributes as much as possible. In ecology only intrinsic classifications have been used but the resulting intrinsic groups are often related to extrinsic attributes (e.g. abiotic environmental parameters).

Hierarchical versus Non-Hierarchical

Hierarchical clustering methods optimize a route between the individual entities to the entire set of entities by progressive fusions or fissions. The results of hierarchical classifications are usually expressed as a dendrogram (Fig. 1) or tree-diagram, which depicts the optimal route from the whole to the individual entities. Non-hierarchical clustering methods, on the other hand, optimize the homogeneity of the groups formed, without defining a route between groups and their constituent entities or between groups (Williams 1971). Hierarchical clustering methods are better developed, more versatile and better understood than are non-hierarchical methods. Although most of the ensuing discussion will concern hierarchical methods, non-hierarchical methods will also be briefly discussed because one non-hierarchical technique, Fager's (1957, Fager and McGowan 1963) recurrent group analysis, has been extensively used in aquatic ecology.

Serial versus Simultaneous Optimization

All hierarchical clustering methods are serially optimized, but non-hierarchical methods may be serially or simultaneously optimized. In serially optimized non-hierarchical

clustering, once a group is formed it is removed from the population of entities, a second group is formed from the remainder, and so on until all the entities are accounted for. In simultaneously optimized clustering, the groups are obtained simultaneously, usually by iterative optimization of partitions of the population of entities.

Agglomerative versus Divisive

Agglomerative hierarchical clustering proceeds by progressive fusion beginning with the entities and ending with the complete population. Divisive hierarchical clustering progressively splits the entire set of entities into smaller and smaller groups. Agglomerative clustering strategies are the most widely used in ecology. Williams (1971) pointed out that agglomerative methods suffer from some computational disadvantages and are inherently prone to a small amount of misclassification, because they begin at the inter-entity level, where the possibility of error is greatest. On the other hand, most divisive clustering strategies are monothetic (see below) which severely handicaps their utility.

Monothetic versus Polythetic

Divisive clustering methods may be monothetic, in which case fissions are based on a single attribute (i.e. in the binary case the presence or absence of an attribute), or polythetic, in which case the division is based on resemblance over all attributes. Clearly, monothetic methods, which would, for example, split two collections on the presence or absence of only one "indicator" species, are of limited utility in ecology. However, polythetic divisive strategies which appear to be the ideal hierarchical clustering methods are poorly developed or impractical in terms of computation time. Several new short-cut polythetic divisive methods have recently been proposed and, although none has yet been used in aquatic ecology, they will be reviewed because of the promise they show.

Combinatorial versus Non-combinatorial

Agglomerative hierarchical clustering methods can have combinatorial or non-combinatorial solutions (Lance and

Williams 1967a, Williams 1971). With combinatorial methods group/group and group/entity resemblance measures can be calculated successively from the inter-entity resemblance matrix and thus, once that matrix is computed, it is no longer necessary to retain the original data matrix. Such a procedure has obvious computational advantages over a non-combinatorial method in which the original data matrix must be retained for the calculation of measures required during successive agglomerations.

By far the most widely used clustering methods are combinatorial, agglomerative, and hierarchical. However, non-combinatorial agglomerative, monothetic divisive, and serially optimized non-hierarchical methods have also been used in aquatic ecology. Clustering methods falling in these categories, plus the intuitively attractive polythetic divisive category, are discussed below.

NON-HIERARCHICAL METHODS

Serially Optimized Methods

Recurrent Group Analysis -

The only non-hierarchical method receiving much use in ecology is Fager's (1957) recurrent group method. Fager (1957) gives detailed instructions for the formation of clusters and I will only attempt an abbreviated restatement. Starting with an inter-entity resemblance matrix, it is first necessary to select an arbitrary level of resemblance at which the investigator considers two entities associated. Thus, the resemblance matrix is converted into a matrix of binary attributes, "associated" and "non-associated." One then determines the largest group of associated entities which can be formed. These entities are termed the first group and are removed from further consideration together with any other entities which only have associations with members of the first group. This procedure is repeated with the remaining unclassified entities again and again until all entities with positive associations are placed in a group. The relationships among the groups may be indicated by the proportion of the number of inter-entity associations which are positive (e.g. if there were 3 positive associations between entities in 2 groups, one with 3 members and the other with 5, the connectivity would be $3/(3 \times 5) = 0.20$).

Recurrent group analysis has usually been employed in inverse classifications based on the Fager binary similarity coefficient (Table 1). An example of a recurrent group classification is given in Fig. 6 and reflects the patterns of co-occurrence of demersal fishes off Southern California. The technique has similarly been used widely in marine ecology in the study of plankton (Fager and McGowan 1963, Sheard 1965, Stone 1969, Tash and Armitage 1967, Venrick 1971), benthos (Jones 1969, Lie and Kelley 1970), and fishes (Fager and Longhurst, 1968, Mearns, 1974).

Recurrent group non-hierarchical clustering has some serious disadvantages. The minimum resemblance required for grouping entities must be stated a priori and is constant for all entities and the technique does not recognizes degrees of association. Changing of the arbitrary level of resemblance necessary for association can produce very different results. An entity may be "captured" by a large group early in the clustering and may appear unassociated with entities which have high resemblance to it but not to all other members of its group. The analysis typically produces a few large groups and many small remnant groups, whose entities, together with those entities attached to, but not members of groups, are not informatively classified. With these criticisms in mind and with the present wide availability of superior hierarchical clustering programs, there remains little value in the continued use of the recurrent group analysis and it is best considered obsolete.

Other Methods -
Various other serially optimized non-hierarchical methods have been proposed, some of which operate on a resemblance matrix and some of which do not involve the computation of the entire matrix. Some methods are reviewed by Lance and Williams (1967c) and Anderberg (1973).

Simultaneously Optimized Methods

These are of basically two types: those which operate on an inter-entity resemblance matrix and those which operate on subsets of entities and involve prior declaration of the number of groups sought (Lance and Williams 1967c). Simultaneously optimized methods generally proceed by first partitioning the entities in some way and optimizing groups by an iterative process of reallocation. Methods are

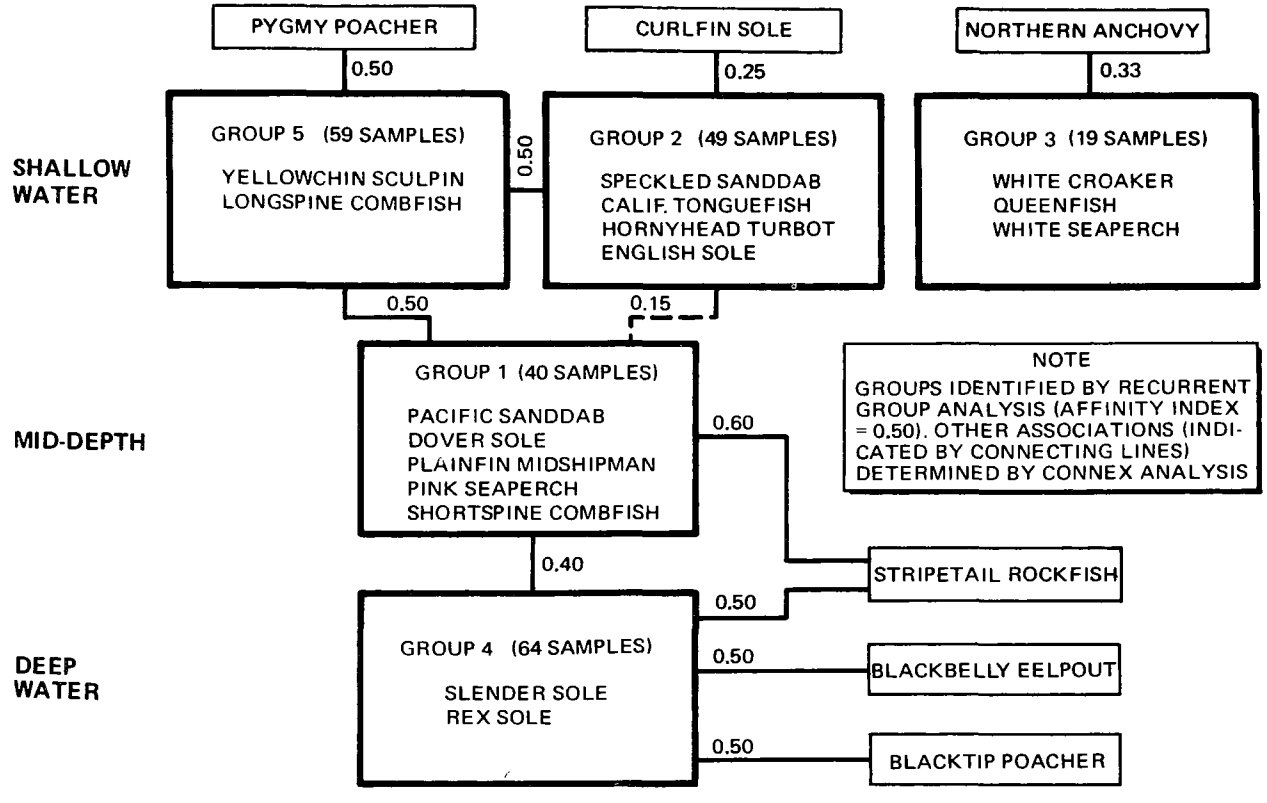


Fig. 6. Species groups of demersal fishes on the Southern California shelf as defined by recurrent group analysis using the Fager binary similarity coefficient (from Mearns 1974).

reviewed by Lance and Williams (1967c) and Anderberg (1973).

Although non-hierarchical clustering methods offer the attractive promise of optimization of within-group homogeneity, in practice the available techniques either have serious drawbacks in performance, are limited in the types of data or resemblance measures with which they can be used, or are computationally difficult. Consequently, it is recommended that the practicing ecologist use hierarchical methods and avoid, at least for the time being, non-hierarchical clustering.

AGGLOMERATIVE HIERARCHICAL METHODS

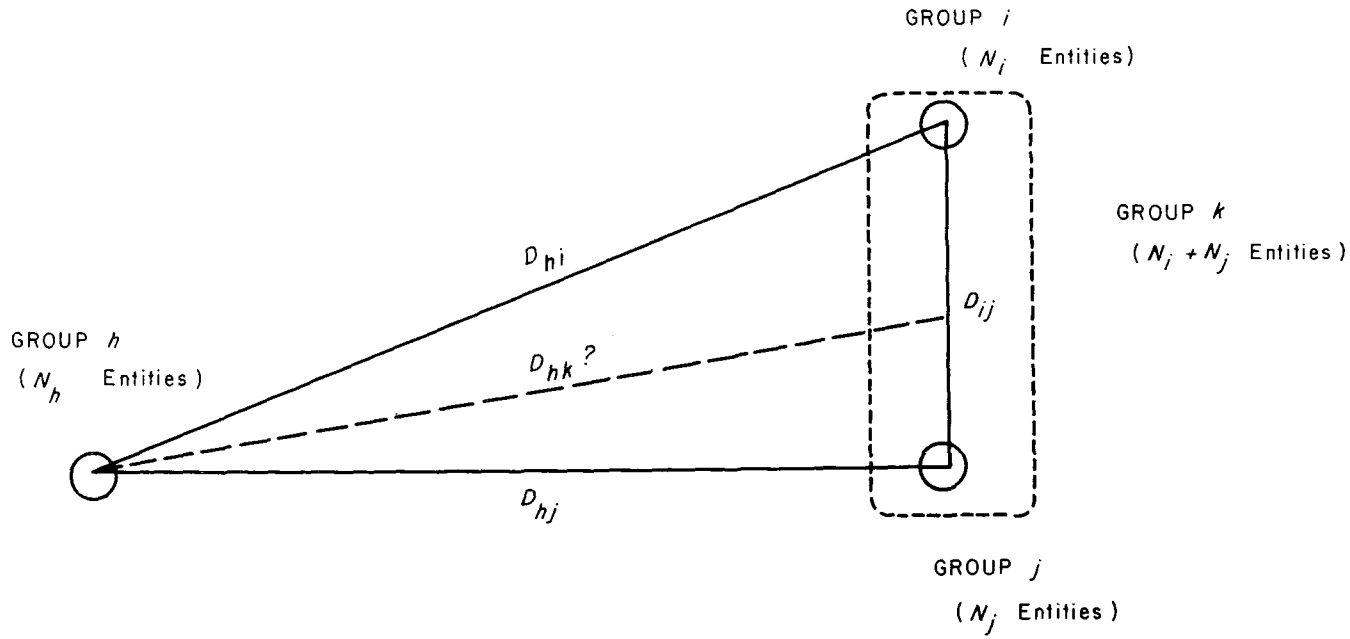
Agglomerative hierarchical clustering strategies all operate by an iterative process of fusing pairs of entities, then pairs of groups of entities until the total population is fused. During each fusion cycle that pair of entities or groups most similar (or least dissimilar or distant) are joined and new resemblances determined between the new group and the all remaining entities and groups. Combinatorial strategies allow the new resemblances to be computed from the preceding resemblance matrix, while with non-combinatorial methods, the original data matrix must be used in the computation of the new resemblances.

Combinatorial Methods

Lance and Williams (1966, 1967a) showed that for a variety of combinatorial strategies group/group or group/individual resemblances can be computed by variants of a single linear equation. The problem of defining these new resemblances when entities or groups are fused is geometrically illustrated in Fig. 7. Two groups i and j are fused to form group k, what then is the resemblance of group k to another group, group h? Given the resemblances (expressed as dissimilarity or distance) D_{hi}, D_{hj} and D_{ij}, what is D_{hk}? The Lance-Williams combinatorial solution is

$$D_{hk} = \alpha_i D_{hi} + \alpha_j D_{hj} + \beta D_{ij} + \gamma |D_{hi} - D_{hj}|, \qquad (31)$$

where the parameters α_i, α_j, β and γ determine the nature of the strategy.



$$D_{hk} = \alpha_i D_{hi} + \alpha_j D_{hj} + \beta D_{ij} + \gamma \left| D_{hj} - D_{hj} \right|$$

Fig. 7. Lance - Williams combinatorial computation of distance between Group h and a new group, Group k, formed by the fusion of Groups i and j.

Values of these parameters for the common combinatorial strategies are listed in Table 2 and the strategies are further discussed below.

Single Linkage -
In this clustering method, also referred to as "nearest neighbor" clustering, the resemblance between two groups is defined as the resemblance of their most similar entities, one in each group. If the resemblances are represented by distances as in Fig. 7 it can be seen that as a group grows it must appear to move closer to some groups or entities, but further from none, thus it is a space contracting strategy. As a result single linkage clustering produces excessive chaining in the hierarchical clustering route, in which entities are fused to a few nuclear groups one at a time rather than forming new groups. This results in classifications in which many entities are not effectively clustered but must be considered as individuals. A good example of an extensively chained, single-linkage agglomeration is given in Fig. 8, which shows a classification of marine phytoplankton collections from the Indian Ocean. Note that the classification of large numbers of collections is indeterminate because of excessive chaining.

Jardine and Sibson (1968, Sibson 1971) defined a set of theoretical conditions which should be met by a hierarchical clustering method which would virtually confine one to single-linkage clustering. However, many authors (Williams et al. 1971, Pritchard and Anderson 1971, Cunningham and Ogilvie 1972) have pointed out the severe shortcomings of single-linkage clustering.

In aquatic ecology single-linkage clustering has been used principally by British marine biologists (Field and MacFarlane 1968, Thorrington-Smith 1971, Angel and Fasham 1973). However, because of restrictions to its utility, the use of single-linkage clustering is not recommended.

Complete Linkage -
This method, also called furthest-neighbor clustering, is the exact opposite of single linkage clustering in that the resemblance between two groups is defined as the resemblance of their least similar entities, one in each group. As a group grows it will recede from some groups or entities but become nearer to none, thus it is a space dilating strategy. Whereas single-linkage agglomeration results in

Table 2. VALUES OF PARAMETERS OF THE LANCE AND WILLIAMS (1967a) LINEAR SOLUTION (EQUATION 31) FOR COMPUTATION OF INTER-GROUP RESEMBLANCE FOR EIGHT COMBINATORIAL CLUSTERING METHODS, WHERE n_h, n_i AND n_j ARE THE NUMBERS OF ENTITIES IN GROUPS h, i AND j, RESPECTIVELY, AND n_k IS THE NUMBER OF ENTITIES IN GROUP k RESULTING FROM THE FUSION OF i AND j (i.e. $n_k = n_i + n_j$).

Method	α_i	α_j	β	γ	Space distortion
Single linkage (Nearest neighbor)	1/2	1/2	0	-1/2	contracting
Complete linkage (Furthest neighbor)	1/2	1/2	0	1/2	dilating
Group average (UPGMA)	n_i/n_k	n_j/n_k	0	0	conserving
Simple average (WPGMA)	1/2	1/2	0	0	conserving
Centroid* (Unweighted centroid)	n_i/n_k	n_j/n_k	$-\alpha_i\alpha_j$	0	conserving
Median (Weighted centroid)	1/2	1/2	-1/4	0	conserving
Flexible	$(1-\beta)/2$	$(1-\beta)/2$	1	0	$B \cong 0$, conserving $B>0$, contracting $B<0$, dilating

Table 2.(continued). VALUES OF PARAMETERS OF THE LANCE AND WILLIAMS (1967a) LINEAR SOLUTION (EQUATION 31).

Method	α_i	α_j	β	γ	Space Distortion
Incremental sums of squares	$(n_h+n_i)/(n_h+n_k)$	$(n_h+n_j)/(n_h+n_k)$	$-n_h/(n_h+n_k)$	0	dilating

* Centroid method combinatorial only for squared Euclidean distance.

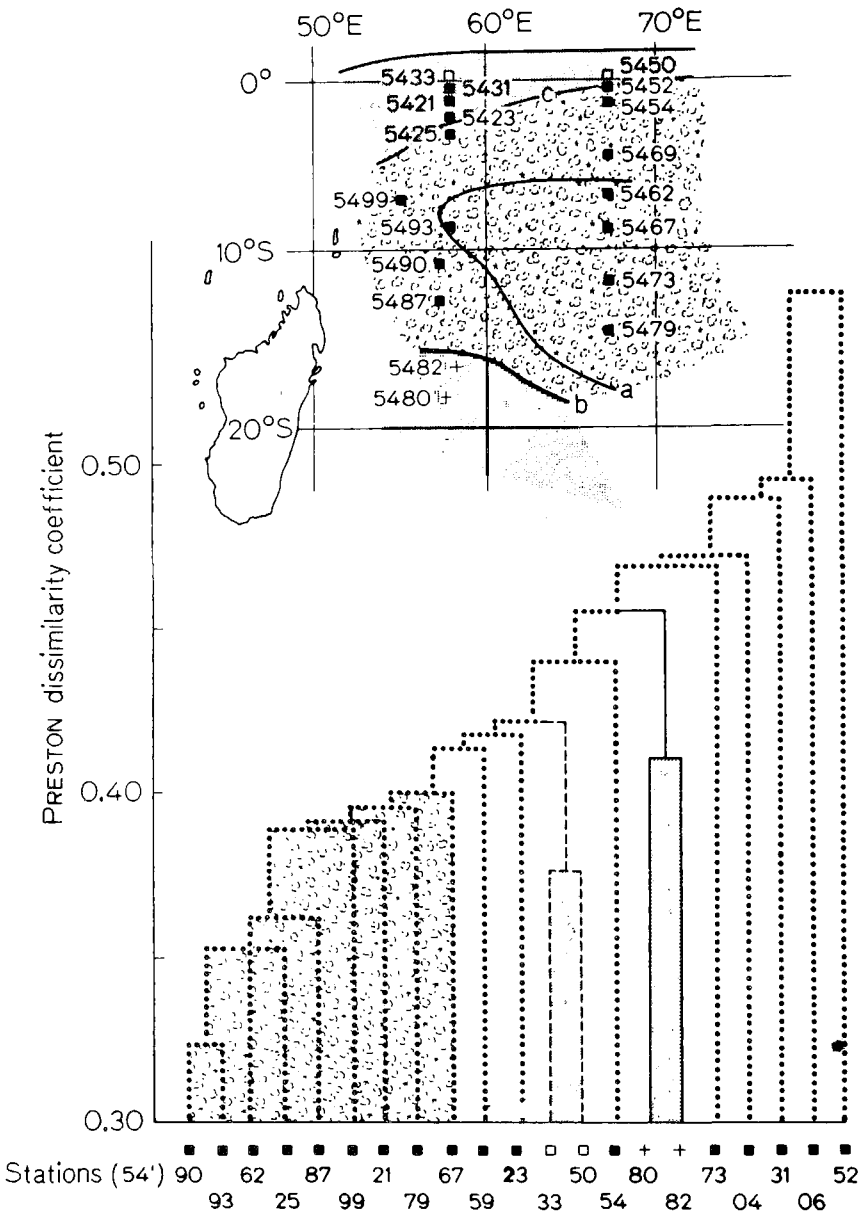


Fig. 8. Dendrogram from a single-linkage clustering showing excessive chaining. Example is from Thorrington-Smith's (1971) study of phytoplankton assemblages in the Indian Ocean off Madagascar.

chaining, complete-linkage agglomeration typically results in intense clustering by forming discrete groups.

Although intense clustering is often a desirable property, one often desires a cluster intensity intermediate between that of single- and complete linkage. Furthermore, it is desirable to base inter-group resemblance on more information than just maximum or minimum resemblance between entities in the two groups. For these reasons, the combinatorial strategies yet to be discussed are generally preferred.

Group Average -
In this method inter-group resemblance is defined as the mean of all resemblances between members of one group to members of another. This solution is widely referred to by numerical taxonomists and American biologists as the "unweighted pair-group method using arithmetic averages" or UPGMA (Sneath and Sokal 1973). Group average clustering has no marked tendencies to space contraction or dilation, and thus may be regarded as space conserving. Hence, it produces only moderately sharp clustering but introduces relatively little distortion to the relationships originally expressed in the inter-entity resemblance matrix (Cunningham and Ogilvie 1972).

Group average agglomeration is now the most widely used clustering method in ecology and it has been extensively employed in aquatic ecology. A non-exhaustive list of applications include: Boesch (1973), Bowman (1971), Cairns and Kaesler (1969, 1971), Cairns, Kaesler and Patrick (1970), Crossman, Kaesler and Cairns (1974), Day et al. (1971), Eagle (1973, 1975), Ebeling et al. (1970), Field (1970, 1971), Gage (1974), Kaesler and Cairns (1972), Kaesler, Cairns and Bates (1971), Kay and Knights (1975), Loya (1972), Jones (1969), Roback, Cairns and Kaesler (1969), Santos and Simon (1974), Stephenson et al. (1972), Ward (1973), and Warwick and Gage (1975).

Simple Average -
This method is equivalent to Sneath and Sokal's (1973) "weighted pair-group method using arithmetic averages" or WPGMA and it differs from group average clustering by weighting the entities most recently admitted to a group equal with all previous members. In practice the results of simple average agglomeration are quite similar to those produced by group average clustering. The method is space-conserving and introduces slightly more distortion to the

actual resemblances than does the group average method.

Centroid -
In this strategy entities are considered defined as points in Euclidean space and when grouped defined by the co-ordinates of the centroid, or geometric center of the points in the group. Centroid clustering is combinatorial only for squared Euclidean distance and variance-covariance used as resemblance measures. For correlation coefficients and similarity measures the original data must be retained for computation of centroids.

Centroid clustering is space conserving but is somewhat prone to distortion (Cunningham and Ogilvie 1972). It suffers from a particular problem in that reversals in the agglomeration may be produced. That is, two groups may fuse at a given level of resemblance and may be subsequently fused with a third group at a higher level of resemblance than the first fusion (for example see Fig. 5-9 in Sneath and Sokal (1973) and Fig. 8.4 in Clifford and Stephenson (1975). Largely for this reason centroid clustering has been recently disfavored. It has been employed in marine ecology by Popham and Ellis (1971), Colijn and Koeman (1975) and Van den Hoek et al. (1975).

Median -
This method, referred to by Sneath and Sokal (1973) as "weighted centroid," weights fused groups as co-equal despite differences in sizes of the groups in a similar fashion as the simple average method. Thus the centroid of the fused group is the centroid of the centroids of the precursor groups. Its properties are more or less similar to that of the centroid method, including the lack of monotonicity of the sequence of fusion levels which results in reversals.

Flexible -
The development of a linear equation for inter-group distance in combinatorial clustering strategies allows the use of continuously variables coefficients in the equation, effectively creating an infinite number of clustering strategies. Lance and Williams (1967a) proposed a flexible strategy based on Equation (31) with the following constraints ($\alpha_i + \alpha_j + \beta = 1$; $\alpha_i = \alpha_j$; $\beta<1$; $\gamma = 0$). By varying β (the cluster intensity coefficient) one can purposefully cause space distortion, as β increases from 0 the strategy is space-contracting and as β decreases from

0 it is space dilating. Fig. 9 shows the effect of varying on the clustering of entities defined by the same resemblance matrix.

Flexible sorting with $\beta = -0.25$ has produced satisfactory results with a wide range of data sets and this value has become more or less conventional (Williams 1971, Clifford and Stephenson 1975). At this level of β, flexible clustering is an intensely clustering, moderately space-dilating strategy. In practical terms, this means that as agglomerations are made, there is a bias against an entity or group joining an already large group and a bias in favor of entities or small groups joining to form separate branches of the hierarchy, i.e. it is group-size dependent. It is important to keep in mind, however, that β can be varied to simulate any level of cluster intensity, although there is little point in using $\beta>0$.

Flexible sorting has been criticized on the grounds that objectivity is lost if that cluster intensity is chosen which most closely fits preconceptions about the data (Sneath and Sokal 1973). However, the use of a variably space-dilating strategy seems sensible in some ecological contexts. For example, a common feature of many ecological data sets is high resemblance among the common or abundant species and much lower resemblance among the rarer species. It seems reasonable to accept a significantly lower resemblance between rare species than between common ones, and in practice intense flexible sorting often compensates for this discrepancy by forming groups of rare species which would be chained on to larger nuclear groups in space-conserving clustering. Intense clustering strategies are often prone to misclassifications and one often has to choose between non-classifications due to weakly clustering strategies or misclassifications due to intensely clustering strategies. The best approach depends on the data set, but with large data sets, especially in inverse analyses, it is often better to use an intensely clustering strategy followed by reallocation of misclassified entities.

Marine ecological applications of flexible clustering include Stephenson et al. (1970, 1972, 1974), Stephenson and Williams (1971), Williams and Stephenson (1973) and Boesch (1973). Another enlightening application of flexible clustering was by Williams et al. (1973) in a study of pattern in rain-forests.

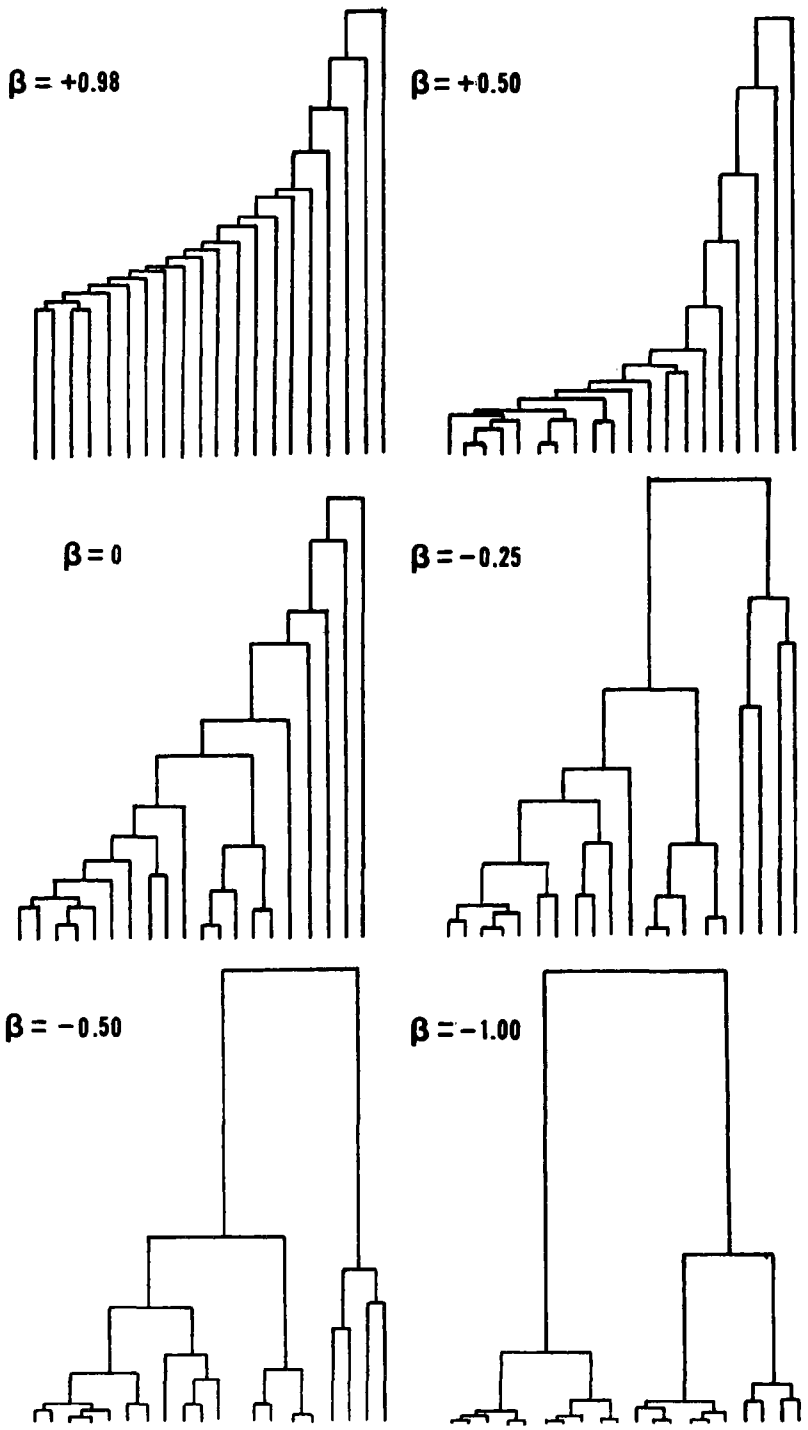


Fig. 9. Effect on agglomerative hierarchy of varying the cluster intensity coefficient in flexible clustering. (after Lance and Williams 1967a).

Incremental Sums of Squares -
It can be seen in Fig. 4 that squared Euclidean distance (D^2) is an additive function, as is variance and total information content. Several authors (Ward 1963, Orloci 1967, Burr 1970) have proposed clustering methods in which entities or groups are successively agglomerated so that fusion will cause the smallest possible increment in the sums of squares of Euclidean distances. Burr (1970) showed that this strategy is combinatorial using the constants as listed in Table 2. The strategy can also be applied to other distance measures, but in practice is usually used with Euclidean distance or standardized Euclidean distance measures.

The incremental sums of squares strategy is an intensely clustering, group-size dependent method. Thus the technique is powerful in imposing structure in relatively patternless data, but, like other space-dilating strategies, is prone to misclassification and may produce clusters of entities which have relatively little in common except their paucity of attributes.

This strategy has been applied by Hughes and Thomas (1971a, b), Hughes et al. (1972), Polgar (1975) and Holland and Dean (1976) in studies of marine and estuarine benthic communities.

Non-combinatorial Methods

In general, non-combinatorial clustering methods have a serious drawback in efficiency because the original data must be retained for computation of resemblance matrices after each fusion cycle. Thus they are likely to be impractical for large data sets.

Centroid -
As noted earlier, centroid clustering is combinatorial only for squared Euclidean distance. For other distance measures, distances between the centroid of a newly formed group and the centroid of another group must be recalculated based on the average scores of all attributes for the two groups. Since centroid distances are mainly useful only for Euclidean metrics (for which a combinatorial solution is available) and because of the drawbacks of centroid clustering mentioned above, centroid methods are seldom used for non-Euclidean distances.

Information Gain -
As mentioned in the discussion of information content measures in Section V, agglomerative clustering can be based on the minimum information gain in fusion of entities or groups. Methods exist for clustering based on binary or multistate data (Williams, Lambert and Lance 1966) and quantitative data (Lance and Williams 1967b, Orloci 1969, 1971, 1975, Dale, Lance and Albrecht 1971). Information gain agglomeration is an intensely clustering strategy and thus suffers attendant disadvantages.

The application of agglomerative information-gain clustering to plant ecological problems has been thoroughly reviewed by Dale (1971), Dale and Anderson (1972) and Williams et al. (1973). Stephenson and Williams (1971) attempted information-gain agglomeration in a study of marine benthos using both quantitative and binary data but were dissatisfied with their results. Similar techniques were applied by Jeffrey and Carpenter (1974) on ranked abundance data in a study of seasonal succession of coastal phytoplankton.

DIVISIVE HIERARCHICAL METHODS

Divisive clustering methods offer the obvious advantage of starting with the whole, when total information available is maximum, and then subdividing along natural breaks in the whole data set. In practice, though, the most developed and practical divisive methods are monothetic, i.e. divisions are based on the presence or absence of one attribute, and are thus of limited usefulness with most ecological data sets.

Monothetic Methods

Association Analysis -
With this method the divisions are made on the basis of chi^2 values summed over the attributes, such that the attribute with the greatest contribution to chi^2 is used as a basis of division of each successive set into two groups, one of the entities possessing the attribute and the other lacking it (Williams and Lambert 1959, 1961a, Lance and Williams 1968).

Moore (1973) used both normal and inverse association analysis in a study of communities associated with kelp

holdfasts and the effects of turbidity and pollution thereon. Although extensively used by plant ecologists through the early 1960's, association analysis has been more or less displaced by monothetic divisive methods based on information content.

Information Content -
Entities are successively divided on the basis of presence or absence of attributes when such divisions result in the maximal reduction in information (Lambert and Williams 1966, Lance and Williams 1968). That is, if I_c is the information content of the total population to be divided, and if I_a and I_b are two monothetically determined subsets, then the value $I_c - (I_a + I_b)$ is maximized during each fusion cycle. A similar divisive method capable of using multistate and continuous data in addition to binary data has been developed by Lance and Williams (1971).

Monothetic divisive information clustering using the Australian CSIRO program DIVINF has been applied to marine ecological problems by Stephenson et al. (1971, 1972), Moore (1973) and Jeffrey and Carpenter (1974).

Polythetic Methods

Polythetic divisive clustering methods are theoretically the optimal hierarchical strategies (Williams 1971). Unfortunately, their development has lagged due to the computational difficulties arising from the very large $(2^{n-1} - 1)$ number of dichotomous splits for each subdivision. For example, Edwards and Cavalli-Sforza (1965) proposed a method by which divisions are made so that the sum of squares of Euclidean distances between subdivisions is maximum. But for hierarchical division of more than 16 entities the computation time is indeed enormous (Gower 1967). Recently, several new polythetic divisive methods have been proposed which are based on short-cut solutions to the problems of examining all possible subdivisions in fission sequence. These are of two main types: those which base subdivisions on an ordination model, and those involving some form of directed search to limit the number of splits which must be examined.

Methods Based on Ordination -
Lambert (1971) and Noy-Meir (1973b) developed methods of optimized division based on principal components scores followed by various iterations to further optimize the divisive structure of the classifications. With the use of principal components analysis the inter-entity resemblances are Euclidean distances. However, the use of other ordination methods allows resemblances to be expressed by other distance measures (e.g. similarity coefficients).

Another divisive classificatory method based on an ordination model has been proposed by Hill, Bunce and Shaw (1975) under the name "indicator species analysis." The method proceeds by ordinating entities using correspondence (reciprocal averaging) ordination, and then successively divides the population of entities based on scores of a few indicator attributes which are most responsible for the ordination structure.

Polythetic divisive classifications based on ordination models show a great deal of promise, but it is too early to say which of the proposed methods is best. However, all such methods may be subject to some of the disadvantages of ordination approaches. Furthermore, all of the extant methods seem to have a bias toward forming subdivisions of approximately equivalent size during each division, when arrays of ecological entities are often not so symmetrical.

Methods Based on Directed Search -
Lambert et al. (1973, Smartt, Meacock and Lambert 1974) developed two methods for polythetic divisive classification which seek to form a preliminary split in the population of entities and then examine the robustness of that split by iterative examinations. In AXOR, the initial strategy is to extract principal component or principal coordinate axes and then investigate all n-1 "ordered" splits on the axis to find the best division. Improvements in the split are then made by relocation of entities one at a time in the second and subsequent axes of the ordination until the consideration of a new axis gives no further improvement. In MONIT, the population is first split monothetically and improvements in the split are made by relocation of individuals one at a time until further iterations produce no improvement. Lambert et al. (1973) and Smartt et al. (1974) report consistently better results with these strategies than with various monothetic divisive and polythetic agglomerative strategies.

SECTION VII

INTERPRETATION OF NUMERICAL CLASSIFICATIONS

Frequently ecological applications of numerical classification end with the definition of clusters or classificatory hierarchies or, at best, with a brief description of the relationship of the classification to spatial or temporal distributions of collections or species. However, numerical classification is best viewed as a method for simplifying complex data sets, allowing ecological analysis to proceed more efficiently, rather than as an end in itself. Furthermore, numerical classifications should be viewed critically. The various clustering algorithms discussed in the preceding sections are only algebraic approximations of ecological criteria for classification and the great variety of methods available produce variable results. Thus, further refinements of the objectively produced classifications are frequently needed.

STOPPING RULES

A common problem in the interpretation of hierarchical classifications is the determination of operational groups within the hierarchy. If we consider the results of a hierarchical classification as a dendrogram, the question is which branches are considered groups with reasonable internal resemblance. Frequently, investigators have drawn a line across the dendrogram at some given level of resemblance and stipulated that each branch crossing that line represents a group. Thus, the "stopping rule" is fixed. The fixed resemblance level may be independently determined based on some assumed level of "significance" or it may be based on the resemblance level at which a given number of branches exist in the dendrogram.

Alternately, other investigators have used a variable

stopping rule for definition of operational groups. Usually this involves studying the dendrogram, often in consultation with the original data matrix or two-way tables, to determine "reasonable" groups. Thus, one may specify two groups which cluster together at a higher level of resemblance than that found within a third group.

Although application of a fixed stopping rule obviously lessens subjectivity in interpretation of the classification, there are two characteristics which suggest variable stopping rules are often more appropriate. The first concerns the group size dependence and space distortion properties of some hierarchical clustering methods. There is little justification for a stopping rule of fixed resemblance when inter-group and entity-group resemblance depends on the size of the group. Thus, there seems little sense in applying a fixed stopping rule with classifications formed by intensely clustering methods such as incremental sums of squares and flexible (with negative β) clustering. The second characteristic concerns the nature of ecological data and is particularly important in inverse analyses. Most data sets include species which are more or less ubiquitous and species which are much more rare. It seems reasonable to require higher intra-group resemblance in groups of ubiquitous species than in groups of rare species for which the probability of cooccurrence is low.

A parallel problem to the definition of stopping rules in hierarchical classification is the definition of required intra-group homogeneity with some non-hierarchical methods. For example, Fager's (1957) recurrent group analysis requires the setting of a minimum resemblance level that an entity must have with all members of a group to be included in that group. Varying this minimum resemblance level can severely affect the classification which is produced (Jones 1969). Fager and McGowan (1963) found that a minimum value of 0.5 using the Fager similarity coefficient for binary data produced satisfactory results, but efficacy is dependent on the resemblance measure used and the nature of the data. Selection of a fixed level of intra-group homogeneity is also subject to the same criticism as fixed stopping-rules for hierarchical clustering, namely that intuitive ecological criteria for grouping are not necessarily fixed.

REALLOCATION

A "misclassification" occurs when an entity is placed in one group by a numerical classification when it would improve within-group homogeneity if it were placed in another. With many non-hierarchical clustering methods misclassification problems are minimal because homogeneity is implicitly optimized. Rather, the difficulties with non-hierarchical methods are often quite the opposite -- entities are not effectively classified rather than misclassified.

However, misclassifications are frequent problems in hierarchical clustering. They can occur in divisive clustering because similar entities may be separated in early divisions or in agglomerative clustering where an entity may resemble only one member of a group in which it is included because of an early fusion. As discussed earlier, misclassifications are most frequent in space-dilating hierarchical methods, which tend otherwise to be particularly useful with complex ecological data.

In the case of misclassifications, reallocation of entities from one group to another is appropriate. Although many ecologists using numerical classification have noted obvious misclassifications, relatively few have attempted reallocation of entities. Subjective reallocation nullifying the objectivity of the analysis has apparently troubled many users. On the other hand, rather casual reallocation of entities to conform to preconceptions extrinsic environmental characteristics, or visual inspection of the data or resemblance matrices has sometimes been practiced (Stephenson et al. 1972, Boesch 1973, Clifford and Stephenson 1975).

One may be able to detect misclassifications by examination of the inter-entity resemblance matrix to uncover the presence of entities which have average resemblance to another group higher than that in which it has been classified.

Another convenient way to detect misclassifications is to rearrange the original data matrix both by collection and species groups as determined by normal and inverse numerical classifications, respectively. This allows the examination of the concentration of species occurrence or abundance within the "cells," or coincidences of collection groups with species groups, in this "two-way coincidence table"

(Clifford and Stephenson 1975). Interpretive analyses based on the coincidence of collection and species groups are discussed below under the heading of "nodal analysis." One can then reallocate collections or species by appropriately adjusting the rows and columns of the two-way coincidence table to sharpen the classifications by increasing the "cell density" of scores. This is equivalent to the elaborate table rearrangements used in the Braun-Blanquet or Zurich-Montpellier approach of European plant ecologists (Westhoff and van der Maarel 1973, Mueller-Dombois and Ellenberg 1974, Popham and Ellis 1971).

Although reallocation based on the two-way coincidence table has usually been done visually, objective criteria for reallocation can also be employed. For example, Boesch (1973) reallocated species if the average constancy of species in a group within collection groups could be increased in dense cells or decreased in sparse cells. He further reallocated some species based on an interpretation of an ordination model. Even then, some discretion is involved. Ceska and Roemer (1971) proposed an objective and automated technique for the rearrangement of two-way tables which shows some promise for future application in reallocation.

The development of better and more objective reallocation methods is sorely needed. Lance and Williams (1967c) attractively suggest that simultaneously-optimized non-hierarchical clustering serve to reallocate entities in groups based on an initial hierarchical clustering. Grigal and Ohmann (1975) used multiple discriminant analysis (referred to as a canonical analysis) in the reallocation of entities into groups in order to resolve differences among four different classifications of the data set. However, most suggested procedures for reallocation do not allow the use of the same resemblance function that was initially used to classify the entities.

NODAL ANALYSES

Most investigators who have applied numerical classification to aquatic ecological problems have classified sites (i.e. normal analysis) only. A few investigators, particularly those applying "recurrent group analysis," have classified species (i.e. inverse analysis) only. Relatively few have conducted both normal and inverse

analysis of the same data set (e.g. Stephenson and Williams 1971, Stephenson et al. 1972, 1974, Hughes and Thomas 1971a, b, Boesch 1973, Moore 1973, Sepkoski and Rex 1974). Relating normal and inverse classifications greatly enhances the interpretation of the results of numerical classification and is recommended as a routine post-clustering analysis.

Normal-inverse coincidences can be conveniently examined in a two-way table (Clifford and Stephenson 1975) which is simply the original data matrix rearranged by collection and species groups. As described above, two-way tables are most helpful in identifying misclassifications and in assisting in reallocation. But beyond that they are extremely useful in assisting ecological interpretation of the classifications. Differences among collection groups can be conveniently described on the basis of frequency or abundance of members of the species groups. Conversely differences in the distribution patterns of species groups can be elucidated by the relative frequency or abundance of the species in the various collection groups.

Williams and Lambert (1961b, Lambert and Williams 1962) termed this approach "nodal analysis" since one attempts to describe and interpret the dense cells or "nodes" of the data matrix in which a group of species and group of collections coincide. This concept of nodal analysis is further expanded by Noy-Mier (1971) who developed procedures for the inter-relationship of normal and inverse ordinations.

Further nodal analysis interpretations can be made in expression of the degree of collection group and species group coincidence by using the classic ecological concepts of dominance, constancy and fidelity (Fager 1963, Westhoff and van der Maarel 1973). Stephenson et al. (1972) and Boesch (1973) expressed the pattern of constancy of species belonging to particular species groups in particular collection groups as relative densities of cells of the two-way table. Constancy was arbitrarily graded as high, medium, low, etc. based on percentages or proportions of the number of occurrences of species in the collection group to the total possible number of such occurrences. Algebraically this constancy index can be expressed as

$$C_{ij} = a_{ij}/(n_i n_j) \tag{32}$$

where a_{ij} is the actual number of occurrences of members of species group i in collection group j and the n_i and n_j are the numbers of entities in the respective groups. The index will take a value of 1 when all species occurred in all collections in the group and 0 when none of the species occurred in the collections.

Fig. 10 gives an example of a nodal constancy diagram which also shows the hierarchical relationships of the collection (site) and species groups. The underlying reasons for the classifications and the relationships of the groups are clearly apparent in terms of the patterns of species group constancy. The analysis was based on a data set representing the abundances of 68 species of macrobenthic animals collected from 47 sites on the shallow continental shelf off Virginia (D. F. Boesch, in prep.). The site classification strongly reflects substrate differences among the sites with groups A, B and C consisting of muddy-sand sites, group D consisting of hard-packed fine sand sites and groups E, F, G and H consisting of the coarser sand sites. The nodal constancy patterns conveniently demonstrate the faunal differences between collection groups. One can see, for example, that both the muddier sites (groups A and B) as well as the sites with coarser sediments (groups F, G and H) are characterized by species which are constant there and not elsewhere, but that the sites with intermediate sediment grain size are characterized by species (e.g., groups 5 and 6) which, while highly constant at those sites, are widely distributed with respect to sediment type.

Similarily, one can examine the fidelity of species groups to collection groups in order to give an indication of the degree to which species "select" or are limited to collection types (habitats, seasons or whatever). A simple index of fidelity is an expression of the constancy of species in a collection group compared to the constancy over all collections. Thus, the fidelity of species group i in collection group j can be defined as

$$F_{ij}=(a_{ij}\sum_j n_j)/(n_j\sum_j a_{ij})$$

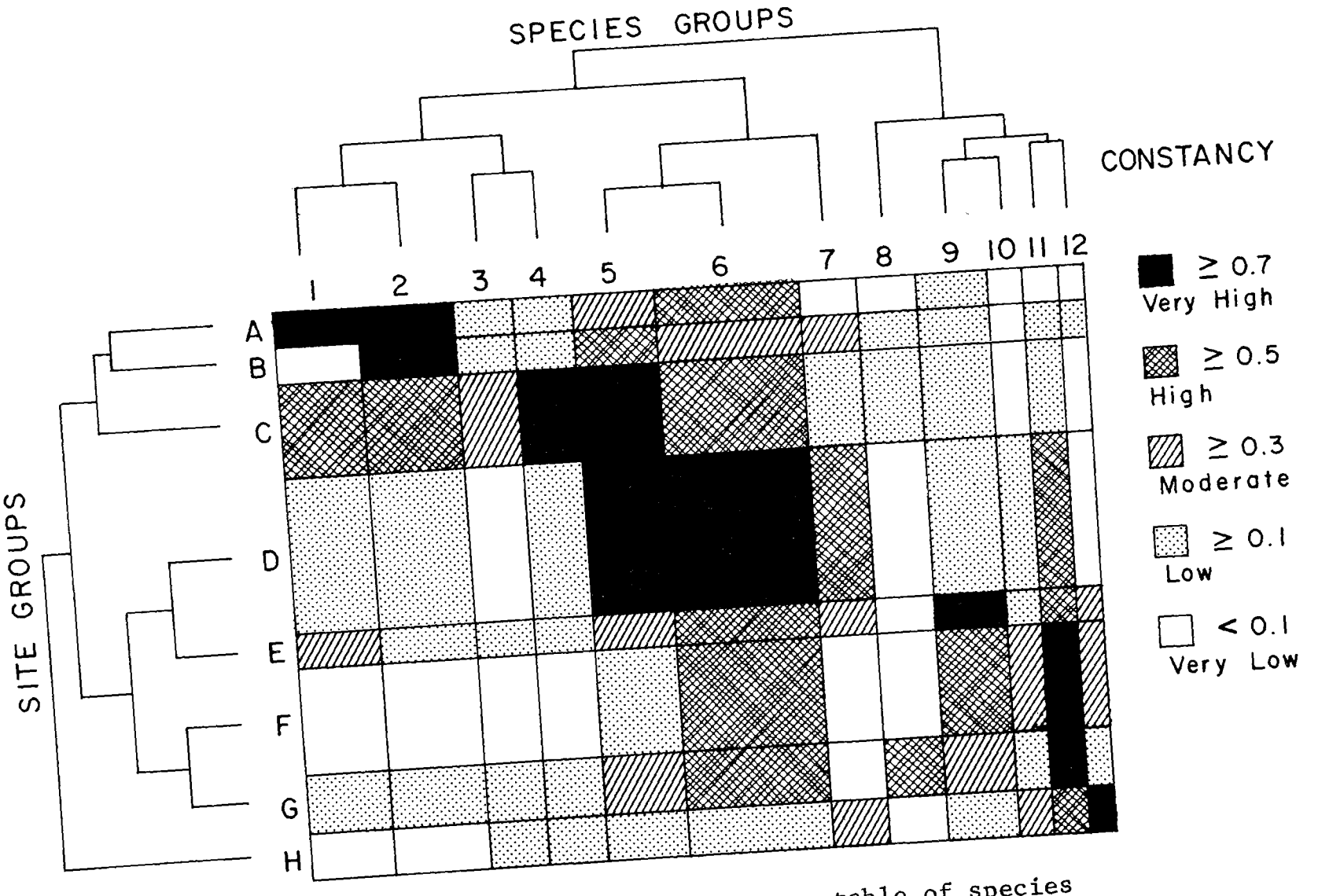


Fig. 10. Nodal constancy in a two-way table of species groups in site groups from an analysis of distribution patterns of macrobenthos from the shallow continental shelf off Virginia (D. F. Boesch, in prep.).

using the same terms as in the constancy index. This index is unity when the constancy of a species group in a site group is equivalent to its overall constancy, greater than 1 when its constancy in that collection group is greater than that overall and less than 1 when its constancy is less than its overall constancy. Values of the index greater than 2 suggest strong "preference" of species in a group for a collection group, i.e. indicating that the average frequency of occurrence of those species in those collections is twice what it is considering all collections. Values of the index much less than 1 suggest "avoidance" of the spatial or temporal habitats represented by the collection group or negative fidelity.

Fig. 11 shows nodal fidelity patterns for the same two-way table as in Fig. 10. Note that some species groups (e.g. 5 and 6), although highly constant in some collection groups, are not very faithful in any. Also some species groups (e.g. 3), although not highly constant in any collection group, are highly faithful to some groups.

Using quantitative data one can also express the concentration of abundance of species in the collection groups. For each species the average abundance in the collection group is divided by its average abundance overall. These ratios can be averaged over all species in the species group to reflect the average concentration of abundance for the node.

Alternate approaches have been taken by Stephenson and his associates (Williams and Stephenson 1973, Stephenson et al. 1974, Clifford and Stephenson 1975) in relating species distribution patterns to collection groups. He has used various tests of "conformity" of individual species to collection groups. In this sense conformity is analagous to fidelity or concentration of abundance, as used above. Species conformity can be tested probabilistically using F-tests or non-parametric tests of significance. The contribution of a species to the collection classification can then be described in terms of its conformity and importance (i.e. relative abundance).

If the nodal constancy diagram is drawn with the width of the rows and columns proportional to the number of entities in the respective collection and species groups as in Fig. 10 the diagram is also useful in explaining gross differences in the species richness of collections. For example, it is clear from Fig. 10 that collections in site groups F,

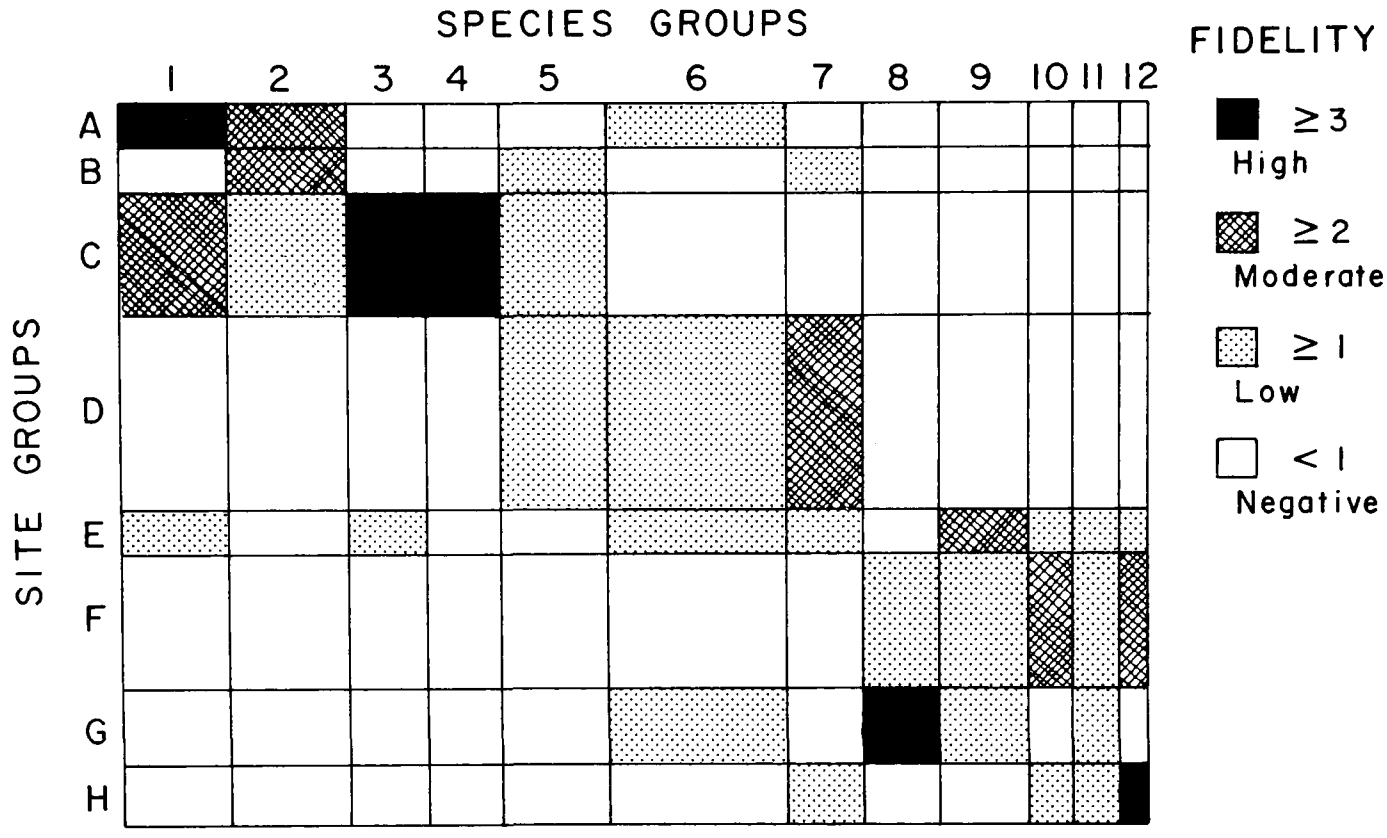


Fig. 11. Nodal fidelity in the same two-way table as in Fig. 10.

G and H generally contain fewer species than collections in group C. Furthermore, it is possible to directly explain these differences in terms of species composition of the collections and patterns of species distribution. Boesch (1973) used such an interpretation of a nodal constancy table to explain patterns of species richness in estuarine benthos in a polluted harbor.

COMPARING CLASSIFICATIONS

It is frequently useful to apply several clustering algorithms to the same data set and compare the results of the alternate classifications. This is helpful not only in selecting the most appropriate classification, but in interpretation of the nature of the patterns exhibited in the data, e.g. qualitative versus quantitative patterns.

In addition to simple comparisons by subjective visual examination, a variety of quantitative methods have been proposed to measure the congruence of two or more classifications. Rohlf (1974) reviewed the methods of comparing classifications so comprehensively that further elaboration here is not necessary. Sufffice it to say, that for the more common hierarchical classifications most methods involve correlating matrices either of the original resemblance measures or of new resemblances based on separation of entities in the classificatory hierarchy.

TESTING DIFFERENCES AMONG GROUPS

For certain purposes it may be desirable to test the reality of the groupings of a classification by application of tests of significant differences among the groups. Statistical techniques for this purpose have not been extensively developed but several different approaches have been used.

As discussed above, Stephenson and his associates have variously tested the conformity of species to normal classifications. However, these are essentially univariate tests and do not constitute tests of differences among either species or collection groups.

Field (1969) proposed a test of differences between clusters based on the information gained by each fusion

in agglomerative clustering. A transform of the information gain, 2 I, is tested for significance with the degrees of freedom based on the number of attributes possessed by the entities or groups being fused. Unfortunately, as proposed, the test is limited to comparisons of binary attributes, considerably reducing the usefulness of the test. The information gain statistic has been used as a test of internal homogeneity of classified groups by Field (1971), Day et al. (1971) and Santos and Simon (1974).

In the previous section on correlation coefficients I outlined several reasons why they were inappropriate for expressing significant relationships among classified groups. Nonetheless, correlation tests have been frequently applied in this manner.

Mountford (1971) developed a probability model describing the joint distribution of resemblance measures which allows a conservative test of significance of clusters defined by internal criteria. It appears not to have been applied subsequently, thus it is difficult to assess its usefulness. Mountford's model, as in the case of others, predicts that the resemblance measure, or a transform of the measure, is normally distributed. However, the sampling distributions of most measures are unknown and Mountford concluded that his test is more readily applicable to indices of similarity constructed according to probability considerations.

The use of multiple discriminant analysis (Cooley and Lohnes 1971) in the test of significance of resemblance among groups of entities shows some promise (Goldstein and Grigal 1972b, Grigal and Ohmann 1975, Polgar 1975). When groups are compared by discriminant analysis, the between-groups sums of squares are maximized with respect to the within-groups sums of squares. The maximization procedure extracts canonical axes onto which each entity can be mapped as a point. The distances among entities can be computed and tested for significance. However applications of the tests do require certain assumptions about the data (e.g. homogeneity of variance, independence of variables and equality of group size) which may not be met by the data.

From this discussion it is clear that further research is needed on the sampling distributions of resemblance measures and on tests of significance among clusters. Numerical classification methods are hypothesis generating

rather than hypothesis testing techniques. They provide hypothetical generalizations on the structure of multivariate data. Testing the constructed hypotheses could be greatly assisted by the availability of non-parametric multivariate tests of significance which impose few assumptions about the nature of the data or the resemblance measures.

RELATING CLASSIFICATIONS TO EXTRINSIC FACTORS

Ways to relate numerical classifications to extrinsic factors (abiotic environmental variables, etc.) are limited only by the imagination of the investigator. Approaches range from plotting the distribution of site (collection) or species groups on maps of the sampling area to statistical comparison of the extrinsic factors corresponding to the groups. With regard to statistical analyses, non-parametric tests may be more appropriate than parametric tests because of problems regarding homogeneity of variance and unequal group size. Extrinsic variables are usually individually related to the classification but multivariate techniques of canonical correlation and multiple regression may be useful (Dagnelie 1971).

One approach which has been only infrequently used is to independently classify or ordinate collections based solely on their associated abiotic factors. The abiotic factor classification of collections can then be compared to the biotic intrinsic classification. Smith (1973), in a study of benthos along a transect in the vicinity of waste discharges, ordinated sites on the basis of water and sediment quality parameters and plotted the distribution of species groups (as determined by a numerical classification) on this ordination.

A frequent problem in the application of numerical classifications in ecology is in the analysis of collections taken over both space and time, e.g. from a series of sites which are sampled seasonally. Several approaches have been used to classify such collections. Some investigators have chosen to classify the collections from each season separately (Field 1971), while others classified the combined temporal collections for each site to elucidate spatial patterns and the combined collections made during each sampling period to elucidate temporal patterns (Jones 1973, Stephenson et al. 1974, Raphael 1974). Boesch (1973)

classified the entire set of collections from sites sampled over time and felt this approach allowed a better understanding of the spatial-temporal interactions which were important in his study.

Williams and Stephenson (1973) developed a technique for the classificatory analysis of the three dimensional data matrix (sites x species x sampling periods). Using Euclidean distance as a resemblance measure, one is able to partition squared distance (as with variance in an analysis of variance) to produce site/species classifications eliminating the effect of temporal changes, and sampling period/species classifications eliminating the effect of spatial differences. Similarly, this method allows the classification of species based solely on either their spatial patterns of occurrence or on their temporal patterns. One is also able to judge the relative importance of spatial patterns, temporal patterns and spatial-temporal interactions. The method is further discussed by Stephenson et al. (1974) and Clifford and Stephenson (1975).

INTERFACING CLASSIFICATION AND ORDINATION

There has been much debate among plant ecologists regarding the most appropriate type of multivariate analysis of ecological data -- classification or ordination (Anderson 1965, Goodall 1970, Whittaker 1973). On one hand, ecologists interested in describing vegetation units or communities have tended to use classification approaches. On the other, those believing that species are distributed more or less independently preferred ordination. Anderson (1965) discusses the controversy and concludes that the problem is non-existent. Numerical classification and ordination are both useful tools although one may be more relevant in certain circumstances than the other. Classification is more useful in simplifying large, complex data sets. Ordination may be more useful in the analysis of smaller, more homogeneous data sets when one is more interested in interpreting the detailed relationships among entities.

Moreover, ordination may be useful in the interpretation of classification and vice versa. Classification and ordination can be interfaced in several ways. The distribution of member entities of classificatory groups in ordination space can be plotted in order to show the

integrity of the groups or, alternately, their overlap (Lie and Kelly 1970, Hughes and Thomas 1971a, b, Hughes et al. 1972, Boesch 1973). In this fashion ordination can also be used in the reallocation of entities to new groups to sharpen the classification (Boesch 1973, Grigal and Ohmann 1975). Alternately, classificatory groups can be ordinated, either as the centroids of the spatial cluster of their constituent entities or by ordination of the groups as defined by their aggregate attributes (Stephenson and Williams 1971). Multidimensional ordinations of groups may provide more accurate depictions of the inter-group relationships than classificatory hierarchies, which are essentially one-dimensional. As discussed above, ordination of groups of entities in multiple discriminant space may allow tests of significance among clusters.

SECTION VIII

APPLICATIONS OF ECOLOGICAL CLASSIFICATION

DESIGNING A CLASSIFICATORY ANALYSIS

Questions Addressed by Classification

The first step in any analysis of ecological data should be a statement of the objectives of the analysis. One should clearly pose the ecological question(s) of interest before selecting an analytical approach. Numerical classification is simply a technique for optimal grouping of entities according to the resemblance of their attributes as expressed by given criteria. It is not a panacea for all data analysis problems nor a procedure by which a computer can do ecology.

Stephenson (1973) lists three reasons why an investigator might wish to apply numerical classification: (1) to appear "up-to-date," (2) to try out methods for application in another context and (3) to attempt to analyze data too complex for adequate consideration by "common sense" techniques. The point of his admonition is that there is an apparent trend to uncritically use numerical classification and other multivariate analyses simply because they are currently popular. For example, there may be little to gain in the application of numerical classification in the analysis of small data sets in which the patterns of entity relationships are clearly apparent.

It must also be remembered that numerical classification is most appropriately a hypothesis generating technique and, with minor exceptions, significance tests which would allow hypothesis testing are not inherent in classificatory techniques. The classificatory algorithm in effect develops a hypothesis, based on predetermined criteria, about the nature of the data. One must then use other techniques to

evaluate hypotheses regarding the homogeneity of the groups formed, the differences among the groups or the relationships of the groups to extrinsic factors.

Alternatives to Classification

Depending on the questions posed, other mathematical techniques may be more appropriate in the analysis of ecological data than classification. If one is primarily interested in relating biotic entities (e.g. species) to abiotic attributes, various correlation or regression techniques, both univariate and multivariate, may be apropos. If one is more interested in the degree of relationships among ecological entities rather than a simplification of a large number of entities into a smaller number of entities (groups) which can be studied more effectively, ordination may be a more appropriate analysis. However, ordination methods are not without their pitfalls and potentials for misuse. As a general rule, ordination become less useful as the data set becomes larger and more complex, whereas numerical classification becomes more attractive under these circumstances. However, there is a wide range of circumstances where both classification and ordination approaches may be useful and often complementary.

There are other ecological problems requiring multivariate pattern seeking for which both classification and ordination may be inappropriate. A good case in point is in the analysis of patterns along an ecological gradient or ecocline. Classification may be appropriately applied if the question posed is "how is the ecocline optimally dissected into zones," but in gradient studies the question often of most interest is "what are the relative rates of biotic change along the gradient?" On first consideration, it would seem that ordination is ideally suited for addressing the latter question, allowing the coenocline (biotic part of the ecocline) to be expressed as a spatial (hopefully linear) model which can be directly compared to the extrinsic gradient vector. However, it has been shown that most, if not all, ordination techniques produce considerably distorted models of coenoclines (Austin and Noy-Meir 1971, Whittaker and Gauch 1973). Terborgh (1971) proposed a simple graphical approach to the analysis of coenocline patterns in which inter-site resemblances are plotted on an abscissa representing the extrinsic environmental gradient. This technique and its modifications have been successfully applied to study the distribution of

benthos along estuarine gradients (Boesch 1976) and the distribution of stream benthos along an elevation gradient (Allan 1975). This technique, which I have termed "coenocline similarity projection" (Boesch 1976), may be useful in analyzing patterns along near source away from source transects often sampled in pollution studies.

Criteria for Selecting Algorithms

Data Manipulations -

There has been a tendency among those using numerical classification not to give proper consideration to the nature of the data and transformations of the data before proceeding with a classificatory analysis. Data reduction, transformation and standardization can profoundly affect the results of the classification. Of course, to a large measure the nature of the data is determined by the particular study or by practical limitations. For example, in some cases truly quantitative data cannot be collected and classifications must be based on binary or ranked data.

Data reductions may be justified for one of three reasons: (1) the data set is too large for computational practicality, (2) aberrant collections exist due to sampling problems and (3) for the exclusion of species which are very rare or inconsistently identified. Justifications are too study-specific to recommend general criteria for data reduction, but appropriate criteria should be consistently applied and clearly stated.

The justifications for data transformations were discussed in Section IV. The application of logarithmic or exponential transformations is appropriate in many ecological cases when large variations in abundance exist. It must be remembered, however, that in addition to "normalizing" species distributions, such transformations may profoundly affect inter-entity resemblance by reducing differences in the size of abundance estimates.

Decisions concerning the use of standardizations are difficult because of the wide range of possible standardizations and uncertainty regarding the effects of their application. The investigator is urged to consider the effects (as discussed on p. 16) of standardizations by collection total (percent standardization), as is frequently done in ecology, and to judge whether these effects are indeed desirable in the case at hand. A second

recommendation is that species-standardizations are often appropriate (depending on resemblance measure used) in inverse classifications in order to reduce the scale problem which exists between abundant and non-abundant species. Standardizations are particularly required with resemblance measures which heavily weight scale differences, e.g. Euclidean distance measures.

Numerical and Ecological Resemblance -
The ecologist interested in applying numerical classification is often confused by the bewildering variety of resemblance measures available. Often selections are made from convenience (e.g. based on program availability) or because of precedent rather than rational criteria. Selection of an appropriate resemblance measure is critical, for it is here more than anywhere else in the clustering algorithm that one attempts to express ecological ideas in algebraic expressions. Therefore, the first step in the selection of resemblance measures should be a verbal statement, in ecological terms, of the criteria for resemblance between entities. Is the investigator more interested in qualitative or quantitative resemblance; how important is dominance (in normal analyses) or outstanding abundance (in inverse analysis) in defining ecological resemblance; and are there underlying spatial or probabilistic conceptualizations in the investigator's perception of ecological resemblance?

Although the selection of qualitative or quantitative resemblance often depends on whether quantitative data are available, ecologists are often interested in patterns of both qualitative and quantitative resemblance among entities. The controversy over whether qualitative comparisons are as informative as quantitative comparisons (Grieg-Smith 1964, Dale and Anderson 1972, Moore•1974, Clifford and Stephenson 1975) is moot because qualitative and quantitative patterns may indeed be quite different. Insight into distributional patterns can often be enhanced by comparing qualitative and quantitative resemblances (Boesch 1976, Boesch, Diaz and Virnstein in press).

If the investigator's concept of inter-collection resemblance is based largely on the similarity of abundance of dominant species a variety of quantitative resemblance measures may be used in normal classification. The Bray-Curtis coefficient is the preferred similarity measure for this purpose. Euclidean distance, correlation and

information content measures weight dominance even more. However, there may be reasons for use of Euclidean distance and information measures because of the requirements of clustering methods or for their additive, spatial or probabilistic properties. One must take care to appropriately standardize Euclidean distances (Orloci 1975) in order to avoid nonsense results which primarily reflect coincidental aberrations.

If, on the other hand, the investigator's concept of inter-collection resemblance is based on more or less equal weighting of all the species in the collection and he wishes to account for quantitative as well as qualitative differences between collection, he can choose to use the Canberra Metric coefficient or one of the aforementioned measures after application of species-standardization.

Ecological criteria for inter-species resemblance may likely be different than those for inter-collection resemblance. Thus, the investigator may choose different data standardizations and resemblance measures for inverse analysis than used for normal analysis.

The scalar differences in species abundances pose a problem in inverse analyses and if the algorithm is not adjusted for them, the ultimate classification may be one which largely separates abundant species from those which are not -- a finding hardly worth the effort. Species-standardizations may help alleviate this effect in the computation of resemblance. As with normal analyses, similarity coefficients derived from the Manhattan metric may be more appropriate than Euclidean metrics or correlation measures when very large discrepancies exist in the scale of attribute scores, because Manhattan metrics are based on absolute differences rather than squared differences or products. If comparison of shape of the distribution patterns between species makes ecological sense (and it often does), correlation coefficients may be useful when there are not large numbers of zero-values in the data matrix.

Classification Structure -

Once a satisfactory method of reflecting ecological resemblance is chosen, the investigator must then choose a strategy for optimum grouping of entities based on these resemblances (Section VI). Given the poor state of development of non-hierarchical clustering methods and the theoretical advantage but impracticality of divisive

hierarchical clustering methods, the clustering methods currently most available and useful are agglomerative and hierarchical. Among these, the computational simplicity of combinatorial methods make these clustering methods the most useful. Armed with three of these clustering methods-group average, flexible and incremental sums of squares-the investigator would have a versatile array of methods to suit most purposes.

Group average clustering has space conserving properties which produce clusters with little distortion of the actual resemblance relationships. Its advantages over centroid clustering (also space conserving) include its combinatorial properties for all distances measures and the fact that it is not susceptible to reversals. However, as discussed in Section VI, group average clustering often does not cluster ecological data intensively enough for effective interpretation. Thus, group average clustering is most useful when entities are relatively few, when space conservation in the classification is required, or as a first look at the unaltered relationships among entities before proceeding to a more intensively clustering method.

Flexible clustering advantageously allows continuous variability of clustering intensity and is more useful than group average clustering when many entities are being classified and their patterns of resemblances are complex. As discussed in (Section VI) flexible clustering is particularly helpful in inverse classifications of large numbers of species of varying abundance. Incremental sums of squares clustering is an intense clustering strategy which may have certain advantages over flexible clustering when Euclidean distance is used as a resemblance measure.

Program Availability

A very important limitation to the custom design of classificatory algorithms has been, and continues to be, the availability of computer programs for their execution. Versatile program systems for numerical classification should include options for the use of various transformations, standardizations, resemblance measures and clustering methods. I know of no set of programs which has facility for application of all the techniques described in this report.

Classificatory programs require extensive computer storage

and are time-consuming in operation, thus programs are usually written to maximize efficiency of operation at a particular computer facility. This means most programs are highly machine dependent and considerable reprogramming is often required in order to make them operable at another facility. The alternative of de novo programming is even more expensive in terms of development time and cost. Several program systems or descriptions of available program systems have been published (Wishart 1969, Rohlf, Kishpaugh and Kirk 1971, Goldstein and Grigal 1972a, Anderberg 1973, Hartigan 1975) and the only alternative for prospective users is to attempt to modify one of these or some other extant program system to suit their needs. A program is currently under development at this laboratory for combinatorial, polythetic, agglomerative hierarchical clustering (COMPAH) which will include options for most of the data manipulations and resemblance measures described in this report and the eight combinatorial agglomerative strategies in Table 2. We are attempting to make the system relatively machine independent and plan to publish a thoroughly documented listing of the programs.

APPLICATION OF CLASSIFICATION TO WATER POLLUTION PROBLEMS

Approaches to Assessing Effects of Pollutants on Community Structure

Field surveys have long been conducted to assess the effects of pollutants or other anthropogenic stresses on aquatic ecosystems. The aim of such sampling approaches is to assess the effect of the stress on community composition and structure. If the effects are catastrophic, there is usually no difficulty in detecting and describing effects based on sampling before and after the stress or surveying impacted and control areas. However, it is more difficult to detect and quantify effects when the level of impact is intermediate. This is because natural communities are usually composed of populations of many species which are often highly variable. Several approaches have been taken to simplify the problem of interpretation of such collection data (Swartz 1972).

Indicator Species -

Perhaps the earliest approach was to concentrate attention on certain "indicator species" which either are particularly sensitive and thus likely to be eliminated by the

stress or, more commonly, are particularly hardy and opportunistic and thus likely to be favored by the stress. For example, it has long been known that organic pollution in rivers and streams may result in the elimination of many insect, molluscan and crustacean taxa and favor the establishment of dense populations of a few species of chirinomid insect larvae and tubificid oligochaetes. The indicator species approach has also been applied in studies of lakes and marine waters, but with less success than in running freshwater habitats. In both lacustrine and marine habitats, species favored by anthropogenic stress are more closely related to the constituents of natural assemblages than in streams. In streams the faunal replacements are often at the order of phylum level, whereas changes may be at subfamily levels in lacustrine, estuarine or marine environments.

Heavy reliance on the indicator species concept has been widely criticized on the grounds that it discards from consideration potentially valuable information on the distribution of the large number of species not considered indicators *a priori*. It has also been noted that indicators presumably favored by pollution are also constituents of natural communities. The so-called pollution indicators are adapted for exploitation of resources following disturbances, or thrive under stress conditions which reduce biotic pressures, and they can frequently be abundant in unpolluted situations (Grassle and Grassle 1974). Nonetheless, the indicator species concept is fundamental to the interpretation of community data collected for the assessment of impacts. Any changes in the composition or structure of communities can only be understood after consideration of the habitat preferences and life history characteristics of species whose abundances are affected. If extensive knowledge of these characteristics is available for a local biota, the indicator species approach can be effectively and meaningfully applied. For example, it has been successful in research on the effects of pollution on benthic communities in various parts of the Baltic Sea (Leppäkoski 1975, Anger 1975). There the extensive collective experience of Scandinavian and German investigators has allowed the classification of large numbers of species into grades of progressiveness, regressiveness or indifference with respect to the response of their populations to pollution stress.

In regions where the responses of the biota are less well known or where community patterns are more complex than those of the Baltic benthos, exclusive use of an indicator species approach to interpretation of data is likely to be less reliable. In such cases numerical classification and other multivariate analyses should prove valuable because they allow analysis of data on all (or at least a larger portion) of the biota. Inverse classifications can produce objective groupings of species corresponding to the response of their populations to pollution, thus generating hypotheses about the relative effects of the pollution on components of the biota which may be experimentally or empirically tested.

Species Diversity -

It has long been known that pollution stress often reduces the species diversity of communities (Jacobs 1975). Species diversity, in the sense of species richness or the number of species in a community, was used as a criterion for assessing the effects of pollution by early workers on stream pollution. The mid 1960's witnessed an explosive increase in interest in species diversity and the quantification of diversity in ecology which has profoundly affected aquatic pollution ecology. A paper by Wilhm and Dorris (1968) suggested that quantitative measures of species diversity be incorporated as water quality criteria and introduced many pollution biologists to information diversity measures. Now the use of diversity indices in investigations of the effects of pollution on aquatic community structure is virtually universal and is often required by contractors and regulatory agencies.

A typical approach in the use of species diversity in pollution studies is the computation of one or more indices of diversity and the correlation (casual or statistical) of these indices with pollution stress and other environmental factors. Often this is only analysis of the multispecies data resulting from the collections. In addition to some theoretical problems concerning the diversity indices used (Hurlburt 1971, Peet 1974), there are severe practical limitations to the usefulness of this approach. Summarizing community structure in one parameter, such as a diversity index, involves a drastic reduction in the information contained in the multivariate entity summarized, i.e. the collection or community. Biotic assemblages with different numbers of taxa and concentration of dominance can, depending on the diversity index, have similar diversity.

Furthermore, assemblages without any species in common can have the same diversity. Numerous cases have been reported in which pollution stress resulted in changes in species populations but actually increased the species diversity. Frequently this is a result of increased evenness of distribution of abundance among the species rather than an increase in species.

Multivariate Analyses -

Numerical classification and other multivariate techniques allow simplification of patterns of multispecies distribution which involve far less loss of the information originally contained in the data than do diversity indices. Furthermore, in classification, comparisons are based on the identity of the species in the collections and the species are not simply treated as strictly numerical entities as in the computation of diversity indices. As in the case of indicator species, numerical classification may be very useful in the interpretation of species diversity analyses of a set of collections (e.g. Boesch 1973). To argue for the exclusive use of one of the three approaches--indicator species, species diversity or multivariate analysis--is foolhardy, for they are in fact complementary. Multivariate techniques serve to provide one level of simplification of the collection data by defining optimal structure of the inter-entity relationships. The conceptualized structure of the data, whether it results from numerical classifications or the investigators subjective appraisal, should then be interpreted in terms of what is known about the biology of the constituent species, which in turn provides the basis of designation of indicator species. Species diversity indices provide quantification of one important aspect of the ecological structure of communities. Multivariate analyses and biological interpretations allow placing species diversity in a larger framework of the total community structure.

Previous Applications of Classification

Despite the explosive increase in the use of numerical classification in aquatic ecology, in particular marine benthic ecology, classificatory techniques have not been widely applied to water pollution problems. However, judging from inquiries and knowledge of ongoing work, applications of multivariate analyses will soon become

widespread in pollution biology. This brief review of published applications of numerical classification is not exhaustive but is intended to illustrate some of the ways classification has been applied in the assessment of water pollution problems.

Crossman et al. (1974) used normal classifications (Jaccard similarity coefficient, group average clustering) of sampling sites based on presence-absence of macrobenthic species in the assessment of effects of spills of hazardous materials in the Clinch River, Virginia. They found the analyses useful in documenting the downstream effects of the spills. The analyses were also found "useful in determining the effects of type of substrate, time of sampling, longitudinal succession, and flooding on the composition of the macrobenthic community." By performing cluster analyses of stations based on various taxonomic groups considered separately, they were able to describe different recovery rates for insects and gastropods.

Mearns (1974) used recurrent group analysis in a study of distribution patterns of demersal fishes off Southern California (Fig. 6). He was particularly interested in determining the effects of pollution from ocean outfalls on fish distribution. Species groups defined by the analysis largely reflected depth distributions, and the effects of the outfalls were not apparent in this analysis except for the absence of the yellowchin sculpin (Fig. 6) in one localized region. However, the limitations of the recurrent group analysis (i.e. binary similarity and sequentially optimized non-hierarchical clustering) restrict the power and reliability of the analysis.

Littler and Murray (1975) used a normal classification (product-moment correlation coefficient, simple average clustering) of quadrat samples of rocky intertidal organisms to assess the effects of a small sewage discharge in California. Quadrat groups were identified on the basis of the species which were cover dominants (Fig. 12). The distribution of quadrat groups showed a modification of the normal zonation patterns around the outfall by the replacement of a stratified, diverse algal cover by a low turf of blue-green, green and red algae (*Gelidium* and *Ulva*) in the mid intertidal and by calcareous tube worms (*Serpulorbis*) and calcareous red algae (*Corallina*) in the lower intertidal (Fig. 12).

Cimberg, Mann and Straughan (1973) applied cluster analysis

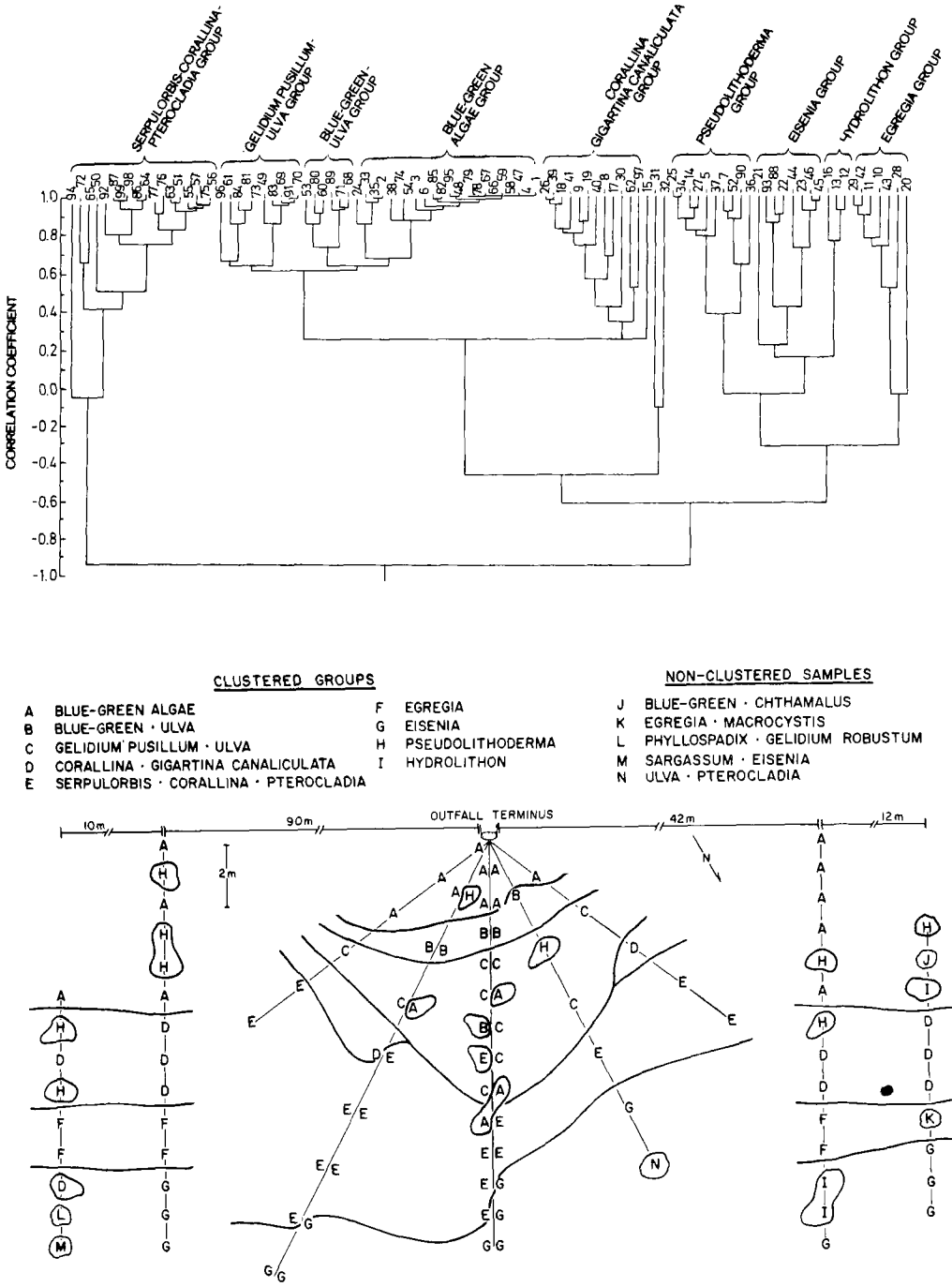


Fig. 12. Classification of quadrat collections from transect surveys of intertidal organisms adjacent to and removed from a small sewage outfall. Distribution of collection groups plotted in lower figure (from Littler and Murray 1975).

(percent standardized Bray-Curtis similarity coefficient, group average clustering) of sites, each representing line transect surveys of rocky intertidal organisms, in a study of the long term effects of the Santa Barbara oil spill along the Southern California coast (Fig. 13). They interpreted the results to indicate that sand coverage and substrate stability were the most important environmental factors influencing species composition, whereas oil apparently had a minor influence. This conclusion was based on the fact that the classification did not separate those beaches oiled from those not. Of course there is a danger in interpreting the classification as indicating that there were no effects (or at most very minor effects) of oiling on intertidal organisms. The effects of seasonal coverage of the rocky surfaces by sand are understandably greater than any effect oil coating might have had. Likewise, the well documented effects of the Santa Barbara oil spill were greatest in the higher intertidal zones. Classifications of collections representing intertidal zones, rather than whole transects, from beaches which were similar in terms of sand burial and substrate stability, but differed in the degree of oiling, would certainly have been more instructive.

Moore (1973) used monothetic divisive analyses in the classification of collections and species from kelp holdfasts in northeast Britain. He found association analysis most useful for normal classification and divisive information analyses (DIVINF) most useful for inverse analysis. Comparisons of the classifications in a nodal analysis showed that turbidity was a primary factor governing the distribution of the holdfast fauna and sites could be characterized by the presence of turbid water species, various groups of clean water species and turbidity indifferent species. Moore (1973, 1974) concluded that the effects of pollution on this fauna, reported by others as important, were not apparent but could not be ruled out because their "definition becomes complicated by the intervention at lower levels of heterogeneity of other correlating factors, e.g. holdfast morphology."

The problem of interpretation of the results of Mearns (1974), Cimberg et al. (1973) and Moore (1973) point out the need for design, if possible, of sampling approaches which establish suitable controls to mitigate the effect of overwhelming natural environmental factors on the assessemnt of effects of pollution. Classification cannot mysteriously decipher pollution effects in complex data

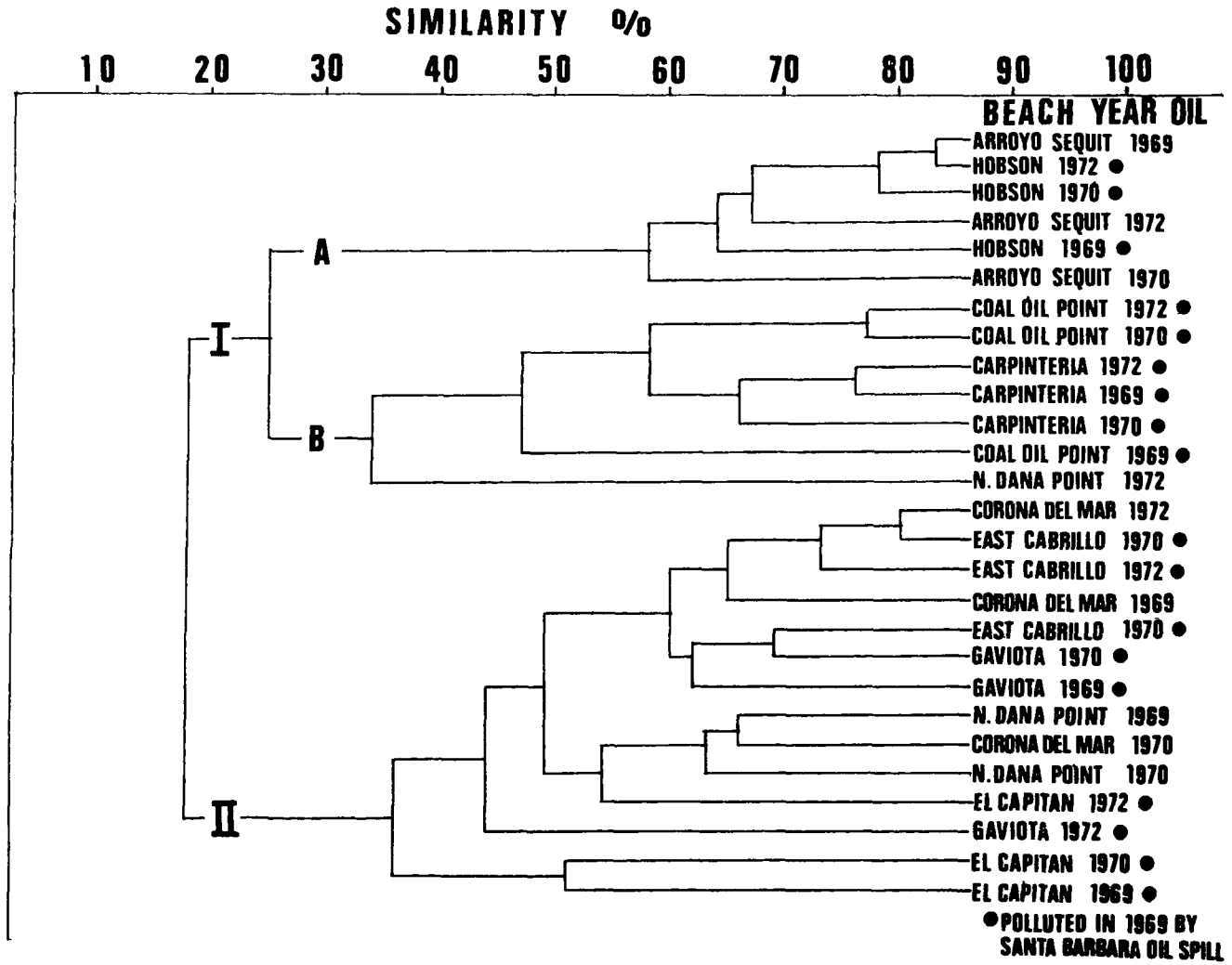


Fig. 13. Classification of transect collections of rocky intertidal organisms from beaches along the Southern California coast (from Cimberg et al. 1973).

sets which are strongly influenced by other environmental factors, and their inability to do so should not necessarily be construed as proving the lack of pollution effects.

Smith (1973) used an inverse classification (Bray-Curtis similarity coefficient, group average clustering) to group macrobenthic species taken from seven stations aligned along a transect from pollution sources in Los Angeles Harbor. He plotted the distribution of species groups on an ordination based on water and sediment quality parameters. He was thus able to relate species distributions to the composite "quality" of the habitat.

Boesch (1973) applied normal and inverse classifications (simultaneous-double standardization, Canberra metric, flexible and group average clustering) in the study of distributional patterns of macrobenthos in a multi-use harbor. The classifications and nodal analysis were useful in interpreting substrate and seasonal patterns. Normal analysis clearly separated the collections from (Elizabeth River) the most heavily polluted part of the harbor, from those from other muddy-sand bottoms (Fig. 14). Inverse analysis interpreted via two-way tables indicated the shifts in species occurrence and abundance which were responsible for these differences.

Smith and Greene (1976) used normal and inverse classifications (cube root transformed, species mean standardized, Bray-Curtis similarity and flexible clustering) in the analysis of macrobenthos on the continental shelf and slope around a Southern California sewage outfall. Sites around the outfall were distinctly grouped and a two-way coincidence table showed that they were characterized by a group of ubiquitous species which were unusually abundant at these sites. The results of the classification were compared with those of ordinations and related to environmental variables.

It is clear from this review that relatively few of the available techniques have yet been used in water pollution applications of numerical classification. Some of the analyses used, e.g. monothetic divisive methods, recurrent group analysis, and binary similarity coefficients, suffer severe limitations. Furthermore, few investigators have attempted both collection and species classifications or have interpreted classifications in two-way tables. It is also clear that some investigators have had naive expectations about what numerical classification can show and have

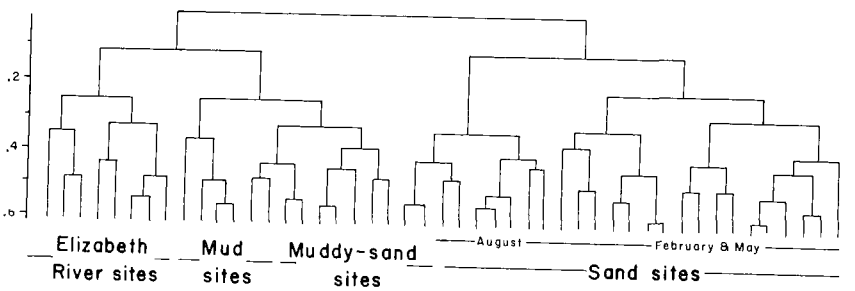


Fig. 14. Classification of assemblages of macrobenthic animals from the Hampton Roads area, Virginia, showing the clear separation of the heavily polluted Elizabeth River sites from the other muddy-sand sites. (From Diaz, et al. 1973 after Boesch 1973).

not designed their sampling or analytical approaches to meaningfully assess the effects of pollution.

A New Example of a Classificatory Approach

To further illustrate the potential usefulness of numerical classification in assessment of the effects of water pollution, data from Reish's (1959) classic study of effects of pollution in Los Angeles - Long Beach Harbors were reanalyzed by normal and inverse classification. Reish sampled macrobenthos at a large number of sites in the inner and outer harbors during January, June and November 1954. He subjectively classified those sites as "healthy," "semi-healthy I," "semi-healthy II", "polluted" or "very polluted" (no macrofaunal life) on the basis of the composition of the collections at the sites, relying heavily on a few indicator species of polychaetes.

Reish identified 141 species in his study. Those species inconsistently identified or present in fewer than five collections were excluded from the classification analysis, leaving 78 species. The Bray-Curtis similarity coefficient based on log-transformed data and group average clustering were used for both normal and inverse classifications. This clustering algorithm is perhaps the most widely used in marine ecology. Normal cluster analysis was run on all collections in which animals were present for each sampling period separately. Inverse cluster analysis was performed on the 78 species with abundance in all collections (i.e. at each of the stations during each of the periods) as attributes.

Site groups selected for each of the three sampling periods are listed in Table 3 and the hierarchical classification for November is given in Fig. 15. The agreement of the objective numerical classification with the subjective classification of Reish for the November collections is remarkable. Only one of Reish's "healthy" sites is not clustered in Site Group 1. The numerical classificatory separation of the "semi-healthy I and II" and "polluted" sites was not completely congruent with Reish's classification, but the general trends are in agreement. Species diversity indices show the expected trend of higher diversity at the "healthy sites" but the ranges of diversity among both the numerical and Reish site groups broadly overlap. This warns against the exclusive use of summary

Table 3. CLASSIFICATION OF COLLECTIONS OF BENTHIC INVERTEBRATES FROM SITES IN LOS ANGELES AND LONG BEACH HARBORS DURING THREE SAMPLING PERIODS FROM DATA OF REISH (1959). SITE CLASSIFICATIONS MADE INDEPENDENTLY FOR EACH SAMPLING PERIOD.

Site Groups	Stations
JANUARY 1954	
0 (no animals collected)	LA 32, 35, 51 LB 5, 15
1	LA 4A, 5A, 6, 7, 8, 13, 22, 26, 48A, 55 LB 1, 2, 2A, 6, 7, 12, 18, 20, 21
2	LA 11, 30C, 41 LB 10, 11, 17, 23
3	LA 30, 30A, 30B, 31, 33, 36, 38, 40, 43A, 49A, 54
4	LA 37
5	LA 29, 43
6	LA 10
7	LA 16, 20, 29A, 34, 39, 49, 50 LB 14
8	LA 45
JUNE 1954	
0 (no animals collected)	LA 11, 32, 33, 34, 35, 36, 39, 49, 49A, 50, 51 LB 10, 14
1	LA 4A, 5A, 6, 7, 10, 22, 26, 55 LB 1, 2, 2A, 5, 6, 7, 12, 17, 18, 20, 21, 23

Table 3 (continued). CLASSIFICATION OF LOS ANGELES-LONG BEACH BENTHOS

Site Groups	Stations
JUNE 1954 (cont.)	
2	LA 29A, 30, 30A, 30B, 30C, 31, 37, 38, 40, 41, 43, 48A
3	LA 13
4	LA 16, 29, X3A, 45 LB 15
5	LB 11
6	LA 20, 54
NOVEMBER 1954	
0 (no animals collected)	LA 11, 16, 20, 32, 34, 35 36, 49, 49A, 50, 51, 54
1	LA 4A, 5A, 6, 7, 8, 22, 26, 55 LB 1, 2, 2A, 5, 6, 7, 10, 12, 17, 18, 20, 21
2	LA 10, 13 LB 11
3	LA 29, 29A, 30. 30A, 30B, 31, 37, 40, 41 LB 23
4	LA 30C, 38, 43, 43A, 45, 48A LB 14, 15
5	LA 33, 39

Fig. 15. Classification of stations in Los Angeles - Long Beach Harbors based on data of Reish (1954) for November, 1954. Reish's designation of degree of pollution effects indicated for each station as is the species diversity according to Shannon's information measure (Pielou 1975).

Table 4. CLASSIFICATION OF SPECIES OF BENTHIC INVERTEBRATES COLLECTED BY REISH (1959) FROM LOS ANGELES AND LONG BEACH HARBORS. DATA FROM ALL THREE SAMPLING PERIODS WERE INCLUDED. ENVIRONMENTAL INDICATOR SPECIES USED BY REISH LABELLED H (HEALTHY) SHI (SEMIHEALTHY I) SHII (SEMIHEALTHY II) AND P (POLLUTED).

Species Group	Indicator	Species
A	H	*Nereis procera*
	H	*Tharyx parvus*
	H	*Cossura candida*
		Nemerteans
		Chaetozone corona
		Lumbrineris minima
		Capitata ambiseta
		Marphysa sp.
		Spiophanes missionensis
		Tellina buttoni
		Tharyx multifilis
		Paraonis gracilis
		Spiochaetopterus sp.
		Armandia bioculata
		Hypoeulalia bilineata
		Petricola californiensis
	SHII	*Cirriformia luxuriosa*
		Chione undatella
		Neanthes caudata
		Saxidomus nutalli
A_2		*Hesperone complanata*
		Anaitides williamsi
		Chone minuta
		Prionospio heterobranchia
		Cirriformia spirabranchia
		Crepidula onyx
A_3		*Glycera americana*
		Pherusa inflata
		Tagelus californianus

Table 4 (continued). CLASSIFICATION OF SPECIES OF BENTHIC INVERTEBRATES

Species Group	Indicator	Species
B_1	SHI	Polydora paucibranchiata
	SHI	Dorvillea articulata
		Macoma nasuta
	P	Capitella capitata
		Podarke pugettensis
		Oligochaetes
B_2		Diopatra splendissima
		Lumbrineris erecta
		Polydora brachycephala
B_3		Corophium acherusicum
		Caprellids
C		Lyonsia californica
		Acteon punctocoelata
		Thyasira barbarensis
		Lumbrineris latreilli
		Prionospio cirrifera
D_1		Melinna cristata
		Epinebalia sp.
		Platynereis bicanaliculata
		Nuculana taphria
		Crenella decussata
D_2		Pinnixa franciscana
		Callianassa californiensis
		Psephidia ovalis
		Polydora cirrosa
		Protothaca staminea
E_1		Amphicteis scaphobranchiata
		Pectinaria californiensis
		Prionospio pinnata
		Holothurians
		Stylatula elongata
		Ostracods
		Terebellides stroemi
		Streblospio crassibranchiata

Table 4 (continued). CLASSIFICATION OF SPECIES OF BENTHIC INVERTEBRATES

Species Group	Indicator	Species
E_2		*Drilonereis nuda* *Lumbrineris* sp. (juv.) *Diopatra* sp. (juv.) *Axiothella rubrocincta* *Lumbrineris* l. *japonica* *Asychis disparidentata* *Laonice cirrata*
F_1		*Chione fluctifragra* *Haploscoloplos elongus* *Chone mollis* *Fusinus kobetti*
F		*Nephtys caecoides* *Eteone californica* *Polydora* sp. *Acteocina magdalensis*

statistics in the description of differences in community structure.

Species groups selected are listed in Table 4. The taxa used are those of Reish. The three polychaetes Reish designated as indicators of "healthy" bottoms, Nereis procera, Tharyx parvus and Cossura candida, are clustered in the large Species Group A and were in fact very closely clustered in the infra-group hierarchy. The single "semi-healthy II" indicator, Cirriformia luxuriosa, was also included in Species Group A. This species was broadly distributed among the sites and occurred in very large densities at some inner harbor stations, Reish's "semi-healthy II" sites. The two species indicative of "semi-healthy I" bottoms, Polydora paucibranchiata and Dorvillea articulata, and the single polychaete indicative of polluted bottoms, Capitella capitata, were clustered in Species Group B_1. Members of this group are pollution tolerant species which largely comprised the fauna of the inner harbor. Often only Capitella was found in the most severely affected zones of the inner harbor, which Reish termed "polluted." However Capitella widely cooccurred with the other members of Group B and is understandably grouped with them.

Normal and inverse analyses are compared in nodal constancy diagrams in Fig. 16. Note the high constancy of Species Group B, in the "semi-healthy" sites, moderate constancy (due to Capitella) in the polluted sites and low to very low constancy in the healthy sites. Also observe the moderate constancy of Species Group A at the "healthy" sites and low constancy at "semi-healthy" and "polluted" sites. Most of the other species groups consist of relatively inconstant species, i.e. the rarer forms, some of which demonstrate complex spatial-temporal patterns of constancy. Most of the species in these groups are largely restricted to "healthy" bottoms.

Further analyses of these data provided even more insight into the interactions of species distribution and community structure. However, this brief description demonstrates the utility of the approach. The basic conclusions one would reach via the numerical classification are similar to those of Reish, based on his extensive experience with the fauna, and thus the technique was efficacious. Furthermore, the numerical classificatory approach allows greater insight into patterns than empowered simply by our limited multivariate mental processes.

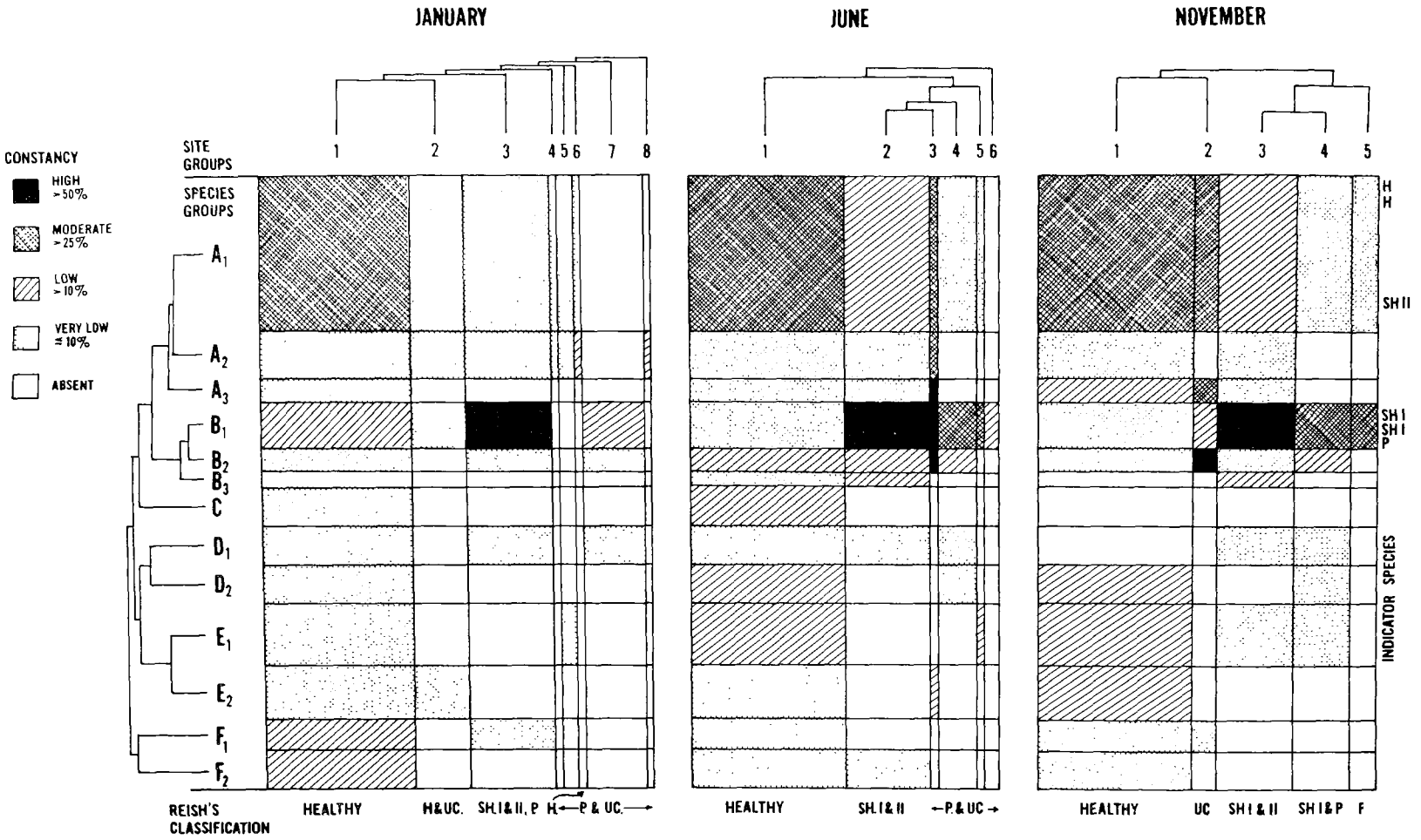


Fig. 16. Nodal constancy in two way tables (site classifications performed separately for each sampling period) for species groups in site groups determined for Reish's (1954) Los Angeles - Long Beach Harbor data.

SECTION IX

REFERENCES

Allan, J. D. 1975. The distributional ecology and diversity of benthic insects in Cement Creek, Colorado. Ecology 56:1040-1053.

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, New York. 359 p.

Anderson, D. J. 1965. Classification and ordination in vegetation science: controversy over a non-existent problem? J. Ecol. 53:521-526.

Angel, M. V. and M. J. R. Fasham. 1973. Sond Cruise: 1965: Factor and cluster analyses of the plankton results, a general summary. J. Mar. Biol. Ass. U.K. 53:185-231.

Anger, K. 1975. On the influence of sewage pollution on inshore benthic communities in the south of Kiel Bay. Part I. Qualitative studies on indicator species and communities. Merentutkimaslait. Julk./Havsforshningsinst. Skr. 239:116-122.

Austin, M. P. and P. Grieg-Smith. 1968. The application of quantitative methods to vegetation survey. II. Some methodological problems of data from rain forest. J. Ecol. 56:827-844.

_____ and I. Noy-Meir. 1971. The problem of non-linearity in ordination: experiments with two-gradient models. J. Ecol. 59:763-773.

Barnard, J. L. 1970. Benthic ecology of Bahia de San Quintin, Baha, California. Smiths. Contr. Zool. 44:1-60.

Benzecri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data, p. 35-74. In S. Watanabe (ed.) Methodologies of pattern recognition. Academic Press, New York & London.

Bloom, S. A., J. L. Simon, and V. D. Hunter. 1972. Animal-sediment relations and community analysis of a Florida estuary. Mar. Biol. 13:43-56.

Boesch, D. F. 1973. Classification and community structure of macrobenthos in the Hampton Roads area, Virginia. Mar. Biol. 21:226-244.

_____. 1976. A new look at zonation of benthos along the estuarine gradient. In B. C. Coull (ed.) Ecology of marine benthos. Univ. South Carolina Press, Columbia.

_____, R. J. Diaz and R. W. Virnstein. in press. Effects of Tropical Storm Agnes on soft-bottom macrobenthic communities of the James and York estuaries and the lower Chesapeake Bay. Chesapeake Sci.

Bowman, T. E. 1971. The distribution of calanoid copepods off the southeastern United States between Cape Hatteras and southern Florida. Smiths. Contr. Zool. 96:1-58.

Bray, J. R. and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 27:325-349.

Burr, E. J. 1968. Cluster sorting with mixed character types. I. Standardization of character values. Austral. Comput. J. 1:97-99.

_____. 1970. Cluster sorting with mixed character types. II. Fusion strategies. Austral. Comput. J. 2:98-103.

Cairns, J., Jr. and R. L. Kaesler. 1969. Cluster analysis of Potomac River survey stations based on protozoan presence-absence data. Hydrobiologia 34:414-432.

_____ and _____. 1971. Cluster analysis of fish in a portion of the upper Potomac River. Trans. Amer. Fish. Soc. 100:750-756.

Cairns, J., Jr., R. L. Kaesler and R. Patrick. 1970. Occurrence and distribution of diatoms and other algae in the upper Potomac River. Not. Nat. Acad. Natur. Sir Philadelphia 436:1-12.

Ceska, A. and H. Roemer. 1971. A computer program for identifying species releve groups in vegetation studies. Vegetatio 23:255-277.

Chardy, P. 1970. Ecologie des crustaces peracarides des fonds rocheux de Banyuls-sur-Mer. Amphipodes, isopodes, tanaidaces, cumacés, infra et circalittoraux. Vie et Milieu 21:657-728.

Cimberg, R., S. Mann and D. Straughan. 1973. A reinvestigation of Southern California rocky intertidal beaches three and one-half years after the 1969 Santa Barbar oil spill: A preliminary report, p. 697-702. In Proc. Joint Conf. Provent. Contr. Oil Spills, March 13-15, 1975. American Petroleum Inst., Washington.

Clifford, H. T. and W. Stephenson. 1975. An introduction to numerical classification. Academic Press, New York. 229 p.

Colijn, F. and R. Koeman. 1975. Das Mikrophytobenthos der Wahen, Strande und Riffe un Hohen Knechtsand in der Wesermundung. Forschungsstelle Norderney, Jahresb. 1974 26:63-83.

Colwell, R. K. and D. J. Futuyma. 1971. On the measurement of niche breadth and overlap. Ecology 52:567-576.

Cooley, W. W. and R. P. Lohnes. 1971. Multivariate data analysis. Wiley, New York.

Crossman, J. S., R. L. Kaesler and J. Cairns, Jr. 1974. The use of cluster analysis in the assessment of spills of hazardous materials. Amer. Midl. Nat. 92:94-114.

Cunningham, K. M. and J. C. Ogilvie. 1972. Evaluation of hierarchical grouping techniques: a preliminary study. Comput. J. 15:209-213.

Czekanowski, J. 1909. Zur differential Diagnose der Neandertalgruppe, Korrespbl. dt. Ges. Anthrop. 40:44-47.

Dagnelie, P. 1971. Some ideas on the use of multivariate statistical methods in ecology, p. 167-180. In G. P. Patil, E. C. Pielou, and W. E. Waters (eds.) Statistical Ecology. Vol. 3, Pennsylvania State Univ. Press, University Park.

Dale, M. B. 1971. Information analysis of quantitative data, p. 133-148. In G. P. Patil, E. C. Pielou,and W. E. Waters (eds.) Statistical Ecology. Vol. 3, Pennsylvania State Univ. Press, University Park.

_____ and D. J. Anderson. 1972. Qualitative and quantitative information analysis. J. Ecol. 60:639-653.

_____, G. N. Lance, and L. Albrecht. 1971. Extensions of information analysis. Austral. Comput. J. 3: 29-34.

Day, J. H., J. G. Field and M. P. Montgomery. 1971. The use of numerical methods to determine the distribution of the benthic fauna across the continental shelf of North Carolina. J. Anim. Ecol. 40:93-125.

Diaz, R. J., M. E. Bender, D. F. Boesch and R. Jordan. 1973. Water quality models and aquatic ecosystems. Status, problems and prospectives, p. 137-153. In R. A. Deminger (ed.) Models for environmental pollution control. Ann Arbor Science, Ann Arbor.

Eagle, R. A. 1973. Benthic studies in the southeast of Liverpool Bay. Estuar. Coast. Mar. Sci. 1:285-299.

_____. 1975. Natural fluctuations in a soft bottom benthic community. J. Mar. Biol. Assn. U.K. 55:865-878.

Ebeling, A. W., R. M. Ibara, R. J. Lavenberg, and F. J. Rohlf. 1970. Ecological groups of deep-sea animals off Southern California. Bull. Los Angeles County Mus. Nat. Hist. Sci. 6:1-43.

Edwards, A. W. F. and L. L. Cavalli-Sforza. 1965. A method for cluster analysis. Biometrics 21:362-375.

Eisma, D. 1966. The distribution of benthic marine molluscs off the main Dutch coast. Neth. J. Sea Res. 3:107-163.

Fager, E. W. 1957. Determination and analysis of recurrent groups. Ecology 38:586-593.

_____. 1963. Communities of organisms, p. 415-433. In M. N. Hill (ed.) The sea. Ideas and observations on progress in the study of the seas. Wiley-Interscience, New York.

_____ and A. R. Longhurst. 1968. Recurrent group analysis of species assemblages of demersal fishes in the Gulf of Guinea. J. Fish. Res. Bd. Canada 25:1405-1421.

_____ and J. McGowan. 1963. Zooplankton species groups in the North Pacific. Science 140:453-460.

Field, J. G. 1969. The use of the information statistic in the numerical classification of heterogeneous systems. J. Ecol. 57:565-569.

_____. 1970. The use of numerical methods to determine benthic distribution patterns from dredging in False Bay. Trans. Roy. Soc. S. Afr. 39:183-200.

_____. 1971. A numerical analysis of changes in the soft-bottom fauna along a transect across False Bay, South Africa. J. Exp. Mar. Biol. Ecol. 7:215-253.

_____ and G. McFarlane. 1968. Numerical methods in marine ecology. 1. A quantitative "similarity" analysis of rocky shore samples in False Bay, South Africa. Zool. Afr. 3:119-137.

Gage, J. 1974. Shallow water zonation of sea-loch benthos and its relation to hydrographic and other physical features. J. Mar. Biol. Assn. U.K. 54:223-249.

Goldstein, R. A. and D. F. Grigal. 1972a. Computer programs for the ordination and classification of ecosystems. Oak Ridge Natl. Lab., Ecol. Soc. Div. Publ. 417.

_____ and _____. 1972b. Definition of vegetation structure by canonical analysis. J. Ecol. 60: 277-284.

Goodall, D. W. 1970. Statistical plant ecology. Ann. Rev. Ecol. System. 1:99-124.

Goodall, D. W. 1973. Sample similarity and species correlation, p. 107-156. In: R. H. Whittaker (ed.) Ordination and classification of communities. Handb. Veg. Sci. 5, Junk, The Hague.

Gower, J. C. 1967. A comparison of some methods of cluster analysis. Biometrics 23:623-637.

Grassle, J. F. and J. P. Grassle. 1974. Opportunistic life histories and genetic systems in marine benthic polychaetes. J. Mar. Res. 32:253-284.

Green, R. H. 1971. A multivariate statistical approach to the Hutchinsonian niche: Bivalve molluscs of central Canada. Ecology 52:543-556.

Grieg-Smith, P. 1964. Quantitative plant ecology. 2nd ed. Butterworths, London. 256 p.

Grigal, D. F. and L. F. Ohmann. 1975. Classification, description, and dynamics of upland plant communities within a Minnesota wilderness area. Ecol. Monogr. 45: 389-407.

Hartigan, J. A. 1975. Clustering algorithms. Wiley-Interscience, New York. 351 p.

Hartzband, D. J. and W. D. Hummon. 1974. Sub-community structure in subtidal meiobenthic Harpacticoida. Oecologia 14:37-51.

Hill, M. O., R. G. H. Bunce and M. W. Shaw. 1975. Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in Scotland. J. Ecol. 63:597-613.

Holland, A. F. and J. M. Dean. 1976. The community biology of intertidal macrofauna inhabiting sand bars in the North Inlet area, South Carolina. In B. C. Coull (ed.) Ecology of marine benthos. Univ. South Carolina Press, Columbia.

Horn, H. S. 1966. Measurements of overlap in comparative ecological studies. Amer. Nat. 100:419-423.

Hughes, R. N. and M. L. H. Thomas. 1971a. The classification and ordination of shallow water benthic samples from Prince Edward Island, Canada. J. Exp. Mar. Biol. Ecol. 7:1-39.

Hughes, R. N. and M. L. H. Thomas. 1971b. Classification and ordination of benthic samples from Bedeque Bay, an estuary in Prince Edward Island, Canada. Mar. Biol. 10: 227-235.

_____, D. L. Peer and K. H. Mann. 1972. Use of multivariate analysis to identify functional components of the benthos in St. Margaret's Bay, Nova Scotia. Limnol. Oceanogr. 17:111-121.

Hummon, W. D. 1974. S_H': A similarity index based on shared species diversity, used to assess temporal and spatial relations among intertidal marine Gastrotricha. Oecologia 17:203-220.

Hurlburt, S. H. 1971. The nonconcept of species diversity: A critique and alternate parameters. Ecology 52:577-586.

Ivimey-Cook, R. B., M. C. F. Proctor and D. L. Wigston. 1969. On the problem of the R/Q' terminology in multivariate analyses of biological data. J. Ecol. 57:673-675.

Jacobs, J. 1975. Diversity, stability and maturity in ecosystems influenced by human activities, p. 187-207. In W. H. van Dobben and R. H. Lowe-McConnell (eds.) Unifying concepts in ecology. Junk, The Hague.

Jardine, N. and R. Sibson. 1968. The construction of hierarchic and non-hierarchic classifications. Comput. J. 11:177-184.

Jeffrey, S. W. and S. M. Carpenter. 1974. Seasonal succession of phytoplankton at a coastal station off Sydney. Aust. J. Mar. Freshwat. Res. 25:361-369.

Jones, A. R. 1973. The structure of an association of nektobenthic invertebrates from Moreton Bay. Unpublished Ph.D. Thesis, University of Queensland.

Jones, G. F. 1969. The benthic macrofauna of the mainland shelf of Southern California. Allan Hancock Monogr. Mar. Biol. 4:1-219.

Kaesler, R. L. and J. Cairns, Jr. 1972. Cluster analysis of data from limnological surveys of the upper Potomac River. Amer. Midl. Nat. 88:56-67.

Kaesler, R. L., J. Cairns, Jr. and J. M. Bates. 1971. Cluster analysis of non-insect macroinvertebrates of the upper Potomac River. Hydrobiologia 37:173-181.

Kay, D. G. and R. D. Knights. 1975. The macroinvertebrate fauna of the intertidal soft sediments of south east England. J. Mar. Biol. Ass. U.K. 55:811-832.

Kohn, A. J. 1968. Microhabitats, abundance and food of Conus on atoll reefs in the Maldine and Chagos Islands. Ecology 49:1046-1062.

Lambert, J. M. 1971. Theoretical models for large-scale vegetation survey, p. 87-109. In J. N. R. Jeffers (ed.) Mathematical models in Ecology. Blackwell, Oxford.

_____ and W. T. Williams. 1962. Multivariate methods in plant ecology. IV. Nodal analysis. J. Ecol. 50:775-802.

_____ and _____. 1966. Multivariate methods in plant ecology. VI. Comparison of information-analysis and association-analysis. J. Ecol. 54:635-664.

_____, S. E. Meacock, J. Barrs and P. F. M. Smartt. 1973. AXOR and MONIT: Two new polythetic-divisive strategies for hierarchical classification. Taxon 22:173-176.

Lance, G. N. and W. T. Williams. 1966. A generalized sorting strategy for computer classifications. Nature 212:218.

_____ and _____. 1967a. A general theory of classificatory sorting strategies. I. Hierarchical systems. Comput. J. 9:373-380.

_____ and _____. 1967b. Mixed-data classificatory programs. I. Agglomerative systems. Aust. Comput. J. 1:15-20.

_____ and _____. 1967c. A general theory of classificatory sorting strategies. II. Clustering systems. Comput. J. 10:271-277.

_____ and _____. 1968. Mixed-data classificatory programs. II. Divisive systems. Austral. Comput. J. 1:82-85.

Lance, G. N. and W. T. Williams. 1971. A note on a new divisive classificatory program for mixed data. Comput. J. 14:154-155.

Leppakoski, E. 1975. Assessment of degree of pollution on the basis of macrozoobenthos in marine and brackish-water environments. Acta Acad. Aboensis, Ser. B 25:1-90.

Lie, U. and J. C. Kelley. 1970. Benthic infauna communities off the coast of Washington and in Puget Sound: identification and distribution of the communities. J. Fish. Res. Bd. Canada 27:621-651.

Littler, M. M. and S. N. Murray. 1975. Impact of sewage on the distribution, abundance and community structure of rocky intertidal macro-organisms. Mar. Biol. 30:277-291.

Livingston, R. J. 1975. Impact of kraft pulp-mill effluents on estuarine coastal fishes in Apalachee Bay, Florida, USA. Mar. Biol. 32:19-48.

Looman, J. and J. B. Campbell. 1960. Adaptation of Sorensen's K (1948) for estimating unit affinities in prairie vegetation. Ecol. 41:409-416.

Loya, Y. 1972. Community structure and species diversity of hermatypic corals at Eilat, Red Sea. Mar. Biol. 13: 100-123.

McConnaughey, B. H. 1964. The determination and analysis of plankton communities. Mar. Res. Indonesia Spec. p. 1-40.

Macfadyen, M. A. 1963. Animal ecology; aims and methods. Pitman, London.

McIntosh, R. P. 1973. Matrix and plexus techniques, p. 157-191. In R. H. Whittaker (ed.). Ordination and classification of communities. Handb. Veg. Sci. 5. Junk, The Hague.

Markle, D. F. and J. A. Musick. 1974. Benthic slope fishes found at 900 m depth along a transect in the western North Atlantic Ocean. Mar. Biol. 26:225-233.

Mauchline, J. 1972. Assessing similarity between samples of plankton. J. Mar. Biol. Ass. India 14:26-41.

Mearns, A. J. 1974. Southern California's inshore demersal fishes: Diversity, distribution, and disease as responses to environmental quality. Calif. Coop. Ocean Fish. Invest. Rep. 18:141-148.

Moore, P. G. 1973. The kelp fauna of northeast Britain. II. Multivariate classification: turbidity as an ecological factor. J. Exp. Mar. Biol. Ecol. 13:127-162.

_____. 1974. The kelp fauna of northeast Britain. III. Qualitative and quantitative ordinations, and the utility of a multivariate approach. J. Exp. Mar. Biol. Ecol. 16:257-300.

Morisita, M. 1959. Measuring of interspecific association and similarity between communities. Mem. Fac. Sci., Kyushu Univ. Ser. E Biol. 3:65-80.

Mountford, M. D. 1971. A test of the difference between clusters, p. 237-257. In G. P. Patil, E. C. Pielou, and W. E. Waters (eds.) Statistical Ecology. Vol. 3. Pennsylvania State Univ. Press, University Park.

Mueller-Dombois, D. and H. Ellenberg. 1974. Aims and methods of vegetation ecology. Wiley, New York. 547 p.

Nichols, F. H. 1970. Benthic polychaete assemblages and their relationship to the sediment in Port Madison, Washington. Mar. Biol. 6:48-57.

Noy-Meir, I. 1971. Multivariate analysis of the semi-arid vegetation in southeastern Australia. I. Nodal ordination by component analysis. Proc. Ecol. Soc. Austral. 6:159-193.

_____. 1973a. Data transformations in ecological ordination. 1. Some advantages of non-centering. J. Ecol. 61:329-341.

_____. 1973b. Divisive polythetic classification of vegetation data by optimized division on ordination components. J. Ecol. 61:753-760.

Ono, Y. 1961. An ecological study of the brachyuran community on Tomoika Bay, Amakusa, Kyushu. Rec. Oceanogr. Wks Japan. Spec. No. 5:199-210.

Orloci, L. 1967. An agglomerative method for classification of plant communities. J. Ecol. 55:193-206.

Orloci, L. 1969. Information analysis of structure in biological collections. Nature 223:483-484.

_____. 1971. Information theory techniques for classifying plant communities, p. 259-270. In G. P. Patil, E. C. Pielou, and W. E. Waters (eds.) Statistical Ecology. Vol. 3. Pennsylvania State Univ. Press, University Park.

_____. 1973. Ordination by resemblance matrices, p. 249-286. In R. H. Whittaker (ed.) Ordination and classification of communities. Handb. Veg. Sci. 5. Junk, The Hague.

_____. 1975. Multivariate analysis in vegetation research. Junk, The Hague. 276 p.

Peet, R. K. 1974. The measurement of species diversity. Ann. Rev. Ecol. System. 5:285-307.

Pielou, E. C. 1969. An introduction to mathematical ecology. Wiley-Interscience, New York. 286 p.

_____. 1974. Population and community ecology. Principles and methods. Gordon and Breach, New York. 424 p.

_____. 1975. Ecological diversity. Wiley-Interscience, New York. 165 p.

Polgar, T. T. 1975. Characterization of benthic community responses to environmental variations by multiple discriminant analysis, p. 267-292. In S. B. Saila (ed.) Fisheries and energy production: A symposium. Lexington Books, Lexington, Mass.

Poole, R. W. 1974. An introduction to quantitative ecology. McGraw-Hill, New York. 532 p.

Popham, J. D. and D. V. Ellis. 1971. A comparison of traditional, cluster, and Zurich-Montpellier analyses of infaunal pelecypod associations from two adjacent sediment beds. Mar. Biol. 8:260-266.

Pritchard, N. M. and A. J. B. Anderson. 1971. Observations on the use of cluster analysis in botany with an ecological example. J. Ecol. 59:727-747.

Raphael, Y. I. 1974. The macrobenthic fauna of Bramble Bay, Moreton Bay, Queensland. Unpublished MSc. Thesis, University of Queensland.

Reish, D. J. 1959. An ecological study of pollution in Los Angeles - Long Beach Harbors, California. Allan Hancock Found. Occ. Pap. No. 22. 119 p.

Roback, S. S., J. Cairns, Jr. and R. L. Kaesler. 1969. Cluster analysis of occurrence and distribution of insect species in a portion of the Potomac River. Hydrobiologia 34:414-432.

Rohlf, F. J. 1974. Methods of comparing classifications. Ann. Rev. Ecol. System. 5:101-113.

_____, J. Kishpaugh and D. Kirk. 1971. NT-SYS. Numerical taxonomy system of multivariate statistical programs. Tech. Rep. State University of New York at Stony Brook, New York.

Ruzicka, M. 1958. Anwendung mathematisch-statistischer Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen). Biologia, Bratisl. 13:647-661.

Sanders, H. L. 1960. Benthic studies in Buzzards Bay. III. The structure of the soft-bottom community. Limnol. Oceanogr. 5:138-153.

_____ and R. R. Hessler. 1969. Ecology of the deep-sea benthos. Science 163:1419-1424.

Santos, S. L. and J. L. Simon. 1974. Distribution and abundance of the polychaetous annelids in a south Florida estuary. Bull. Mar. Sci. 24:669-689.

Sepkoski, J. J., Jr. and M. A. Rex. 1974. Distribution of freshwater mussels: coastal rivers as biogeographic islands. Syst. Zool. 23:165-188.

Sheard, K. 1965. Species groups in the zooplankton of eastern Australian slope waters. Aust. J. Freshw. Res. 16:219-254.

Sibson, R. 1971. Some observations on a paper by Lance and Williams. Comput. J. 14:156-157.

Simpson, E. H. 1949. Measurement of diversity. Nature 163:688.

Smartt, P. F. M., S. E. Meacock and J. M. Lambert. 1974. Investigations into the properties of quantitative vegetational data. J. Ecol. 62:735-759.

Smith, R. W. 1973. Numerical analysis of a benthic transect in the vicinity of waste discharges in outer Los Angeles Harbor. In Marine studies of San Pedro Bay. Part II, Biological investigations. Allan Hancock Foundation, Univ. Southern California.

_____ and C. S. Greene. 1976. Biological communities near submarine outfall. J. Wat. Pollut. Contr. Fed. 48:1894-1912.

Sneath, P. H. A. and R. R. Sokal. 1973. Numerical taxonomy. The principles and practice of numerical classification. Freeman, San Francisco. 573 p.

Sokal, R. R. 1974. Classification: Purposes, principles, progress, prospects. Science, 185:1115-1123.

_____ and F. J. Rohlf. 1969. Biometry, the principles and practice of statistics in biological research. Freeman, San Francisco. 776 p.

Stephenson, W. 1973. The use of computers in classifying marine bottom communities, p. 463-473. In R. Fraser (comp.) Oceanography of the South Pacific. N.Z. Nat. Comm. for UNESCO. Wellington.

_____ and W. T. Williams. 1971. A study of the benthos of soft bottoms, Sek Harbour, New Guinea, using numerical analysis. Aust. J. Mar. Freshwat. Res. 22:11-34.

_____, W. T. Williams and S. D. Cook. 1972. Computer analyses of Petersen's original data on bottom communities. Ecol. Monogr. 42:387-415.

_____, _____ and _____. 1974. The benthic fauna of soft bottoms, Southern Moreton Bay. Mem. Qd. Mus. 17:73-123.

_____, _____ and G. N. Lance. 1970. The macrobenthos of Moreton Bay. Ecol. Monogr. 40:459-494.

Stone, J. H. 1969. The Chaetognatha community of the Agulhas Current. Its structure and related properties. Ecol. Monogr. 39:433-463.

Swartz, R. C. 1972. Biological criteria of environmental change in the Chesapeake Bay. Chesapeake Sci. 13 (Suppl.): S17-S41.

Tash, J. C. and K. B. Armitage. 1967. Ecology of zooplankton of the Cape Thompson area, Alaska. Ecology 48: 129-139.

Terborgh, J. 1971. Distribution on environmental gradients: Theory and a preliminary interpretation of distributional patterns in the avifauna of the cordillera Vilacabamba, Peru. Ecology 52:23-40.

Thorrington-Smith, M. 1971. West Indian Ocean phytoplankton: a numerical investigation of phytohydrographic regions and their characteristic phytoplankton associations. Mar. Biol. 9:115-137.

Valentine, J. W. 1973. Evolutionary paleoecology of the marine biosphere. Prentice-Hall, Englewood Cliffs, N. J. 511 p.

Van den Hoek, C., A. M. Cortel-Breeman and J. B. W. Wanders. 1975. Algal zonation in the fringing coral reef of Curacao, Netherlands Antilles, in relation to zonation of corals and gorgonians. Aquatic Bot. 1:269-308.

Venrick, E. L. 1971. Recurrent groups of diatom species in the North Pacific. Ecology 52:614-625.

Wade, B. A. 1972. A description of a highly diverse soft-bottom community in Kingston Harbor, Jamaica. Mar. Biol. 13:57-69.

Ward, A. R. 1973. Studies on the sublittoral free-living nematodes of Liverpool Bay. I. The structure and distribution of the nematode populations. Mar. Biol. 22:53-66.

Ward, J. H. 1963. Hierarchic grouping to optimise an objective function. J. Amer. Statist. Ass. 58:236-244.

Warwick, R. M. and J. D. Gage. 1975. Nearshore zonation of benthic fauna, especially Nematoda, in Loch Etive. J. Mar. Biol. Ass. U.K. 55:295-311.

Westhoff, V. and E. van der Maarel. 1973. The Braun-Blanquet approach, p. 617-726. In R. H. Whittaker (ed.) Ordination and classification of communities. Handb. Veg. Sci. 5. Junk, The Hague.

Whittaker, R. H. 1967. Gradient analysis of vegetation. Biol. Rev. 42:207-264.

_____. 1973. Approaches to classifying vegetation, p. 325-354. In: R. H. Whittaker (ed.). Ordination and classification of communities. Handb. Veg. Sci. 5. Junk, The Hague.

_____ and H. G. Gauch, Jr. 1973. Evaluation of ordination techniques, p. 289-321. In: R. H. Whittaker (ed.), Ordination and classification of communities. Handb. Veg. Sci. 5. Junk, The Hague.

Wilhm, J. L. and T. C. Dorris. 1968. Biological parameters for water quality criteria. BioScience 18:447-481.

Williams, W. T. 1971. Principles of clustering. Ann. Rev. Ecol. Syst. 2:303-326.

_____ and J. M. Lambert. 1959. Multivariate methods in plant ecology. I. Association-analysis in plant communities. J. Ecol. 47:83-101.

_____ and _____. 1961a. Multivariate methods in plant ecology. III. Inverse association-analysis. J. Ecol. 49:717-729.

_____ and _____. 1961b. Nodal analysis of associated populations. Nature 191:202.

_____, _____ and G. N. Lance. 1966. Multivariate methods in plant ecology. V. Similarity analysis and information analysis. J. Ecol. 54:427-446.

_____, G. N. Lance, M. B. Dale and H. T. Clifford. 1971. Controversy concerning the criteria for taxonometric strategies. Comput. J. 14:162-165.

_____, _____, L. J. Webb and J. G. Tracey. 1973. Studies in the numerical analysis of complex rainforest communities. VI. Models for the classification of quantitative data. J. Ecol. 61:47-70.

Williams, W. T. and W. Stephenson. 1973. The analysis of three-dimensional data (sites x species x times) in marine ecology. J. Exp. Mar. Biol. Ecol. 11:207-227.

Wishart, D. 1969. An algorithm for hierarchical classifications. Biometrics 22:165-170.

Yarranton, G. A., W. J. Beasleigh, R. G. Morrison and M. I. Shafi. 1972. On the classification of phytosociological data into non-exclusive groups with a conjecture about determining the optimum number of groups in a classification. Vegatatio 24:1-12.

SECTION X

LIST OF PUBLICATIONS

The following publications have been produced partially as a result of Grant No. R803599-01-1:

Boesch, D. F. 1977. A new look at the zonation of benthos along the estuarine gradient. In B. C. Coull (ed.) Ecology of marine benthos. Univ. South Carolina Press, Columbia.

Boesch, D. F., M. L. Wass and R. W. Virnstein. 1976. The dynamics of estuarine benthic communities, p. 177-196. In M. L. Wiley (ed.). Estuarine processes, Vol. I. Academic Press, New York.

Boesch, D. F. in press. Classification and ordination: "Objective" interpretation of complex biological data? In J. A. Sherk (ed.). Biological indicators. U.S. Fish and Wildlife Service. Tech. Paper.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/3-77-033	2.	3. RECIPIENT'S ACCESSION NO.
4. TITLE AND SUBTITLE APPLICATION OF NUMERICAL CLASSIFICATION IN ECOLOGICAL INVESTIGATIONS OF WATER POLLUTION		5. REPORT DATE March 1977
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Donald F. Boesch		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Virginia Institute of Marine Science Gloucester Point, Virginia 23062		10. PROGRAM ELEMENT NO. 1BA608
		11. CONTRACT/GRANT NO. R803599-01-1
12. SPONSORING AGENCY NAME AND ADDRESS Corvallis Environmental Research Laboratory U.S. Environmental Protection Agency 200 SW 35th Street Corvallis, Oregon 97330		13. TYPE OF REPORT AND PERIOD COVERED final
		14. SPONSORING AGENCY CODE EPA/600-02
15. SUPPLEMENTARY NOTES		

16. ABSTRACT

Numerical classification encompasses a variety of techniques for the grouping of entities based on the resemblance of their attributes according to mathematically stated criteria. In ecology this usually involves classification of collections representing sites or sampling periods, or classification of species. Classification can thus simplify patterns of collection resemblance or species distribution patterns in an instructive and efficient manner. Procedures of numerical classification are thoroughly reviewed, including data manipulations, computation of resemblance measures and clustering methods. The importance and effects of transformations and standardizations are discussed. It is particularly critical to choose an appropriate resemblance measure which best corresponds with the investigator's concept of ecological resemblance. Clustering methods form groups on the basis of patterns of inter-entity similarity. Various types of clustering methods exist but currently the most useful and best developed are those which are exclusive, intrinsic, hierarchical and agglomerative. Agglomerative clustering methods which distort spatial relationships and intensely cluster are often most useful with ecological data. The usefulness of numerical classification is demonstrated for objective analysis of the data sets resulting from field surveys and monitoring studies conducted for the assessment of effects of pollution. However, to date few pollution biologists have applied the more powerful classificatory techniques and post-clustering analyses.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
species classification species distribution indicator species clustering methods numerical classification	aquatic ecology mathematical methods of community analysis in water pollution	06/C,F

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)* UNCLASSIFIED	21. NO. OF PAGES 125
	20. SECURITY CLASS *(This page)* UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (9-73)

☆ U. S. GOVERNMENT PRINTING OFFICE: 1977—797-582/101 REGION 10