In: R.P. Higgins & H. Thiel (Eds.) Introduction to the study of Meiofauna.
Washington and London: Smithsonian Institution Press, 1988, pp. 197-231.

Communication no. 418 Delta Institute for Hydrobiological Research, Vierstraat 28
Yerseke, The Netherlands.

14. Data Processing, Evaluation, and Analysis

Carlo Heip, Peter M.J. Herman, and Karlien Soetaert

This chapter aims at introducing a number of methods for the post-hoc interpretation of field data. In many ecological studies a vast amount of information is gathered, often in the form of numbers of different species. The processing of these data is possible on different levels of sophistication and requires the use of tools varying from pencil and paper to a mini computer. However, some "number crunching" is nearly always necessary and with the increasing availability and decreasing prices of the personal computer, even in developing countries, there is little point in avoiding the more advanced statistical techniques. There is also a more fundamental reason: ecological data are often of a kind where the application of simple statistics is not only cumbersome, but also misleading and even incorrect. The basic assumptions on which uni- and bivariate statistics are based are often violated in the ecological world: concepts such as homogeneity of variance, independence of observations etc. do not belong to the "real" world of the ecologist.

All this is not to say that ecology may be reduced to statistics or that numbers convey anything that was not already there. However, structure and relationships in the data, and their significance, are in most cases best studied using the appropriate statistics.

The increasing availability of computers and the software required to perform most of the more powerful eco-statistical methods have at least one important drawback: this technology entrains many possible pitfalls for the user who is not fully aware of the assumptions underlying the methods used (e.g., are the data of a form suitable for that particular analysis?). Thus, the user may misinterpret or fail to interpret the results that computers offer so generously (Nie et al., 1975). This chapter was written with the software package user especially in mind.

The chapter is structured as follows: following this introduction is part 1 on the subject of data storage and retrieval. It is intended to introduce the reader to the software that exists and how it can be accessed. The chapter then continues with the methods used to study spatial pattern (part 2), structure (classification and ordination) in part 3, species-abundance distributions in part 4, diversity and evenness indices in part 5, and temporal pattern in part 6. Each of these topics is discussed independently and each topic can be consulted without having to read the others. For this reason, each part will be preceded by its own introduction.

Part 1. Data Storage and Retrieval

The existence of many software packages, useful in ecology, will not be unknown to most ecologists. Many computer centers provide their users with routines which facilitate the input, transformation and appropriate output of the data, and the linking of available software packages into one flexible system. As the bulk of mathematical treatments of ecological data is already available, it is far more efficient to use the existing software from data centers or to implement software packages on a small computer than to try to program the necessary analyses oneself, especially when one envisages that different techniques are to be applied to the same data (Legendre and Legendre, 1979).

Software Packages

SPSS (*Superior Performing Software Systems*).--Includes: data transformations, arithmetic expressions, basis statistics, distribution statistics, regression, correlation, contingency and association analysis, analysis of variance and covariance (principal component analysis), factor analysis, and discriminant analysis. See Nie et al., 1975.

BMD and BMDP.--Includes: encoding of data and subsequent calculations, transposition of data matrices, regression, correlation, contingency and association analysis, dispersion matrix/missing values, statistical tests, cumulative frequencies, analysis of variance and covariance, data transformation, principal component analysis, discriminant analysis, harmonic analysis and periodic regression, and spectral analysis. See Dixon, 1973, and Dixon and Brown, 1977.

CEP (*Cornell Ecology Programs*).--Includes: DECORANA (Hill, 1979a): reciprocal averaging/

detrended correspondence analysis; TWINSPAN (Hill, 1979b): hierarchical classifications by two-way indicator species analysis; ORDIFLEX (Gauch, 1977): weighted averages, polar ordination, principal component analysis, reciprocal averaging; COMPCLUS (Gauch, 1979): non-hierarchical classifications by composite clustering; CONDENSE (Singer and Gauch, 1979): reading data matrices and copying this into a single, standardized format (efficient for computer processing); DATAEDIT (Singer, 1980): data matrix-edit options.

CLUSTAN.--Includes: classification methods, and principal component analysis as an option. See Wishart, 1978.

NT-SYS (Numerical Taxonomy System) and CLASP.--These are classification program packages, mainly featuring agglomerative polythetic techniques. See Rohlf et al., 1976.

SAS.--This is an integrated software system providing data retrieval and management, programming, and reporting capabilities. Statistical routines include descriptive statistics, multivariate and time series analyses. All analytical procedures utilize a consistent general linear model approach. See SAS Institute, 1985.

Data Bases

A *data base* is a computerized record keeping system designed to record and maintain information (Date, 1981). A *field* is the smallest named unit of data stored in a data base. A *record* is a collection of associated fields. In most systems, the record is the unit of access of the database. A *file* is a collection of records.

In multidisciplinary research, or when several ecologists work together, a varied and large amount of information concerning the ecosystem studied becomes available. It can be advantageous to group all these data in one file, store it in a data base, and let every researcher access parts of the file which are useful to him. In this way the validation of common data has to be performed only once, and, as everybody has access to all data contained in the file, the extension of one's own data with data gathered by other workers is made easy (Frontier, 1983).

In many data centers, there exists a data base management system (DBMS); that is, a layer of software between the data base itself (i.e., the data as actually stored) and the user of the system. The DBMS thus shields the data base-user from hardware-level detail and supports user-operations. These user-operations can be performed on a high level (simplified statements) (Date, 1981).

SIR (*Scientific Information Retrieval*).--This is one of several DBMS in common usage in North America. It furnishes a direct interface with SPSS and BMDP and is well documented by a users guide (Robinson et al., 1979) and a pocket guide (Anderson et al., 1978).

Dbase III + and LOTUS 123.--These are powerful packages that interface with each other and with several word processors. With the increasing availability of personal computers, it is to be expected that these and similar data bases and spreadsheets will play a major role in the storage and exchange of data for all but the largest data sets.

Standardization

In order to exchange data between different research institutes or data centers, certain standards should be adhered to in the representation of the data. At the moment, several systems are being developed that involve the standardized coding of different parameters.

The Intergovernmental Oceanographic Commission General Format 3 (IOC GF-3).--This is a system for formatting oceanographic data onto magnetic tape. The ICES (International Council for the Exploration of the Sea) Oceanographic Data Reporting System (hydrochemical and hydrographic data) has been based on this system.

The Biological Data Reporting Format (Helcom System).--This format is being developed by ICES, and is designed to interface with the ICES Oceanographic Data Reporting System. A data file consists of four types of records, structured hierarchically (Figure 14.1): (1) The File Header Record (called Biomaster), is the file header for all the biological data obtained at one station. It contains 80 columns, the first 27 of which are identical to the first 27 columns of the cards used in the ICES Oceanographic Data Reporting System. In this way the relevant hydrographic, hydrochemical, and meteorological data may be obtained easily. The Biomaster further identifies the parameters that have been measured at that station (e.g., phytoplankton, meiobenthos) and how many Series Header Records have been prepared for each parameter. (2) The Series Header Record (called Type Master), defines the method that has been used to obtain the data for the parameter (e.g., cores, box-corers etc.). (3) The Data Cycle Record contains the data obtained for the parameter with the method encoded by the Type Master at the station defined by the Biomaster.

For instance, it may contain information on the number or biomass of each species present at the station. (4) Finally, there is a Plain Language Record that may be used at any level in the format to insert comments which are relevant to the interpretation of the data.

The system is not complete at the time of writing this chapter but has been described rather extensively since other systems are often quite similar and it has been adopted by the ICES Benthos Ecology Working Group for cooperative programs.

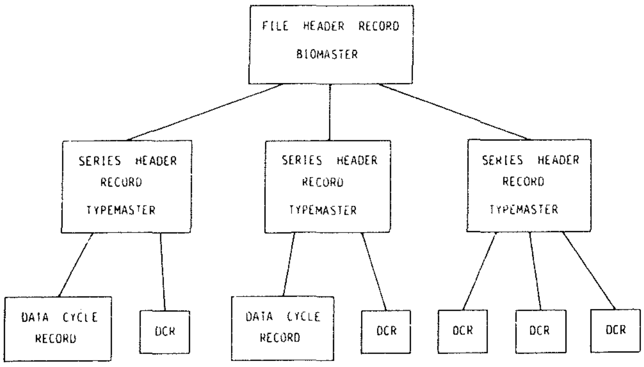

Figure 14.1.--Scheme of the structure of "Helcom system" data base files. See text for an explanation of the terms and a description of the contents of the different records.

NODC (National Oceanographic Data Center, USA.).--This agency has a format for use with benthic organisms. A file contains seven types of 80-character records: (1) The Header record defines the cruise (date, investigator, etc.). (2) There is a Station Header Record for dredge tows that defines the station and describes the tow (type of dredge). (3) There also is a Station Header Record for point sampling that defines the station and describes the pointsampler (e.g., grabs, corers). (4) The Environmental Record describes surface environmental conditions for each station. (5) The Bottom Characteristics Record describes environmental conditions of the bottom and core information. (6) The Taxonomic Data Record gives the number and weight of each taxon collected within a sample. (7) In addition, there is a Text Record.

Taxonomic Codes.--For data exchange, a single system of taxonomic coding would be preferable. However, there are several such systems: (1) For field reporting, the user requires a system that is simple and easy to use, and which minimizes the risk of errors. This implies a mnemonic code based preferably on the scientific name. When using standard abbreviation principles (e.g., Rubin Code: first four letters of genus name, space, first three letters of species name, see Osterdahl and Zetterberg, 1981), the risk of duplicate codes exists. The full latin name, on the other hand, is long and subject to typing and spelling errors, and the software to handle full scientific names is more complicated.

For data exchange in computer compatible form, structured numeric codes are more useful (e.g., the NODC code). These codes can be allocated, taking into account the taxonomic position of the species.

As one should aim at establishing one single coding system, a mechanism to relate the various taxonomic codes to a single common system is needed, such that translation to the common system could be carried out automatically by the computer.

For more information, the reader should contact the Intergovernmental Oceanographic Commission, Place de Fontenoy, Paris, France, or the International Council for the Exploration of the Sea, Palaegade 2-4, DK-1261 Copenhagen K, Denmark.

Part 2. Spatial Pattern

Patterns in the distribution of meiofaunal populations may be studied at length scales that range from that of an individual to the global scale of biogeographic studies. We will restrict our discussion to the small-scale pattern of a deme, the distribution in space of the potentially interbreeding individuals of the population. Examples may be a nematode species on a particular beach or a copepod species in an estuarine creek. Within such large areas, where environmental gradients may or may not exist (but usually will), individual animals occupy a much smaller space during their life. In this smaller space, individual processes and interactions occur: feeding, reproduction, mortality, etc. Small-scale pattern refers to these length scales, which for meiofauna may be between 10^{-2} and 10^{2} m, probably mostly below 10 m. Analysis of this pattern may generate hypotheses on the environmental gradients or factors which influence the species. Also, the information obtained from spatial pattern analysis may be used in the design of an adequate sampling scheme.

In meiofauna ecology the sample is mostly taken by a core of some size varying from drinking straws to 10 cm^{2} surface area cores. These cores may be contiguous or discontiguous. One should always be aware of the importance of core size on the analysis, and cores should not be too small, i.e., they must be much larger than an individual (see below).

Random Patterns

The Poisson Distribution.--At a particular scale the distribution of animals may be random or not; the detection and analysis of randomness or non-randomness is a starting point for further investigation of the factors involved.

A random pattern implies that the probability of finding an individual at a point in the area sampled is the same for all points. When this probability is very low, e.g., when $p < 0.01$, it may be calculated using the Poisson distribution. Alternatively, a Poisson distribution is appropriate when the probability of finding an individual in a sample of the size of an individual is very small. In these cases the Poisson distribution is the standard to which randomness is checked. It should be noted that the mean number of animals per sample is irrelevant in deciding whether or not a Poisson distribution is appropriate: a Poisson distribution applies if the maximum possible number that could occur in a sample is much higher.

The variance of the Poisson distribution is equal to the mean, and the distribution is thus determined by only one parameter. It may be calculated from the arithmetic mean $\bar{x}$ of the observed frequency distribution of the number of individuals per sample. The probability that a sample contains 0, 1, 2, ..., p, ... individuals is given by:

$$e^{-\bar{x}},\ \bar{x}\, e^{-\bar{x}},\ \frac{\bar{x}}{2}\, e^{-\bar{x}}, \ldots,\ \frac{\bar{x}^{p}}{p!}\, e^{-\bar{x}}, \ldots$$

in which each term can be calculated easily from the previous one. The agreement between the observed frequency distribution and the Poisson distribution can then be calculated as X^2 with n−2 degrees of freedom (Pielou, 1969).

The $s^2/\bar{x}$ Ratio.—Since in the Poisson distribution the variance equals the mean, the criterion most frequently used to detect departure from randomness is $s^2/\bar{x} = 1$. Indeed, almost all criteria are in fact based on calculation of the first and second moments of the observed frequency distribution (see Greig-Smith, 1983, for an extensive review). When the ratio is significantly larger than one, the pattern is assumed to be aggregated. When the ratio is significantly smaller than one the pattern is considered to be regular (this has never been observed for meiofauna and will not be considered further). The significance of departure from one can be tested using the fact that $(n-1)s^2/\bar{x}$ is distributed as X^2 with n−1 degrees of freedom.

The Binomial Distribution.--If the number of individuals actually occurring in a sample approaches the maximum possible, the use of the Poisson distribution becomes inappropriate and the frequencies of the density per sample will approximate to a binomial distribution obtained by the expansion of $(p+q)^n$. In this formula p is the probability of finding an individual in the sample, $q = 1-p$ and n is the maximum number possible per sample. Since n is hard to calculate in meiofauna data, the binomial distribution is difficult to apply. A hypothetical meiofauna animal of 1 mm length and 50 µm width has a surface of 0.05 mm^2. With a density of 1000 per 10 cm^2, 5% of the area is covered by the species and that is the chance of finding one if a corer of 0.05 mm^2 surface area is used. At such high densities, the use of the Poisson distribution may thus become doubtful.

Aggregated Patterns

Contagious Frequency Distributions.--When the variance of the observed frequency distribution is significantly larger than the mean, aggregation is inferred. There may be several mechanisms that generate this, and many possible theoretical distributions describe such situations. They are usually called contagious distributions. Examples are the Neyman and Thomas distributions, which are combinations of two Poisson distributions (Neyman, 1939, described clusters of insect eggs and number of eggs per clusters as being both Poisson-distributed; Thomas, 1949, described randomly distributed colonies with individual populations following a Poisson distribution).

Most frequently used is the Negative Binomial Distribution: it may arise in several ways, e.g., when groups are randomly distributed and the number per group follows a logarithmic distribution, or, when groups are randomly distributed with means that vary according to X^2. It may also arise from true contagion, when the presence of an individual increases the chance that another will be there.

The negative binomial distribution is completely defined by two parameters, the arithmetic mean $\bar{x}$ and a positive exponent k. The probability that a sample contains x individuals is given by:

$$p_x = \frac{(k+x-1)!}{(k-1)!\,x!}\ \frac{(\bar{x}/k)^x}{(1-(\bar{x}/k))^{k+x}}$$

The variance of the Negative Binomial Distribution is equal to $s = \bar{x} + \bar{x}^2/k$. From this it follows that:

$$k = \frac{\bar{x}}{(s^2/\bar{x})-1}$$

To obtain an unbiased estimation of k one has to use the maximum likelihood method described by Bliss and Fisher (1953). A computer program for this is given by Davies (1971).

The value of k is a measure of the degree of aggregation (contagion): the smaller it is, the larger the degree of aggregation. Comparisons can only be made for equal sample sizes.

Taylor's Power Law.--Taylor (1961) observed that often a relationship between the variance and the mean exists, such that:

$$s^2 = a\,\bar{x}^b$$

The exponent b is an index of aggregation. If b=0, the pattern is regular ($s^2 = a$); if b=1, the pattern is random (if a=b=1, one obtains the Poisson distribution); if b>1, the pattern is aggregated.

A plot of log s^2 against log $\bar{x}$ yields a straight line:

$$\log s^2 = \log a + b \log \bar{x}$$

Plotting log $(s^2/\bar{x})$ versus log $\bar{x}$ also yields a straight line:

$$\log (s^2/\bar{x}) = \log a + (b-1) \log \bar{x}$$

The advantage of using Taylor's power law lies in the fact that the b-value is independent of sample size. If b=2, $s^2 = a\bar{x}^2$ and $s^2/\bar{x} = a\bar{x}$. A linear relationship between $s^2/\bar{x}$ and $\bar{x}$ has been found by Heip (1975) for harpacticoid copepods. In this case there exists a general relationship between s^2, x, and k, such that $\bar{x}=\sqrt{ks}$, relating density, variability and the degree of aggregation (Heip, 1975).

Description of Patches

If the distribution is contagious, then, whatever the exact mechanism producing it, individuals are found in patches with higher or lower abundance. Information on the size and shape of these patches is clearly desirable. Some attempts to measure patch size in benthic data have been published (Heip and Engels, 1977; Findlay, 1982; Hogue, 1982), but they are not entirely satisfactory. Methods used in plant ecology are described by Greig-Smith (1983). They are based on plotting variance against increasing sample size; a peak in such a plot corresponds to a patch and indicates its size. This method depends on contiguous samples.

Crowding.--The degree of crowding experienced by an individual animal may be estimated by the Index of Mean Crowding (Lloyd, 1967) as:

$$x^* = \bar{x} + (s^2/\bar{x} - 1)$$

The ratio of mean crowding to mean density was called patchiness (Lloyd, 1967).

Location of Patches.--To utilize the information contained in the geographical location of the patches, several methods have been proposed. Krishna Iyer (1949) proposed to classify samples into two categories, one with values higher than the mean, one with values lower than the mean. The expected number of connections between samples with high density values if they are located at random in space can be calculated (see Pielou, 1969 for an extensive description) and compared with the number of connections actually observed. This method has been used by Heip (1976) and Heip and Engels (1977) to describe spatial patterns of the ostracod *Cyprideis torosa* and of six species of harpacticoid copepods.

Another possibility is to use the ratio of the autocovariance to the sample variance weighted as a function of the distance between the samples as a measure of spatial autocorrelation (Jumars et al., 1977). Two statistics proposed by Cliff and Ord (1973) may be used for this purpose:

$$I = \left(\frac{n}{w}\right) \sum_i^n \sum_j^n w_{ij}\, z_i\, z_j \Big/ \sum_i^n z_i^2$$

$$C = \left(\frac{n-1}{w}\right) \sum_i^n \sum_j^n w_{ij}\, (x_i-\bar{x})^2 \Big/ \sum_i^n z_i^2$$

where x_i is the variate value in sample i, n is the number of samples, $z = x_i-\bar{x}$ and w_{ij} are the weights as a function of spatial distance between the samples:

$$w = \sum_i \sum_j w_{ij}$$

The selection of weights is discussed by Jumars et al. (1977). Often used is:

$$w_{ij} = (\text{distance})^{-2}$$

The use of I,c and $s^2/\bar{x}$ is illustrated by Jumars and Eckman, (1983). If the pattern is random, $I \sim 1/(n-1)$ and $C \sim 1$. Significant departures from these values can be calculated according to the formulae found in Cliff and Ord (1973) or Jumars et al. (1977).

Spectral Analysis

The method of spectral analysis, normally applied to temporal series, may also be applied to spatial data. Hogue and Miller (1981) used the autocorrelation of the data to detect periodicity in nematode density data.

Part 3. Classification and Ordination

Classification (clustering) and ordination are two sets of techniques capable of synthesis and ordering of the data collected to describe communities. Classification involves arranging objects (in benthic ecology usually the samples or stations) into groups setting them apart from the members of other groups, and a typical product of classification is a graph called a dendrogram. Ordination attempts to place objects in a space defined by one or more axes in such a way that knowledge of their position relative to the axes conveys the maximum information about the objects. This information may relate to species composition of samples (stations) or to occurrence of species in samples (stations).

Ordination and classification both start from a data matrix, which in benthic ecology takes the form of a n x p matrix with p stations (samples) as columns and n species as rows. More generally, the columns of the data matrix represent the objects (independent variables), the rows represent the descriptors used to describe these objects (dependent variables). The entries in the matrix are a measure of the abundance y_{ij} of species i in station j.

Some measure is then used to compare all the rows, two by two, or all the columns, two by two, of the data matrix. In this way two association matrices may be derived from the data matrix. These association matrices are square matrices with particular properties which make them apt for further analysis. In the Q-mode, the p x p association matrix Q is derived from the comparison between objects (stations), in the R-mode, the n x n association matrix R is derived from the comparison between descriptors. The choice of a suitable association measure depends on whether the analysis is in the Q- or the R-mode. Numerous measures have been proposed; the following account is broadly based on the lucid review of Legendre and Legendre (1979) which should be consulted for more detail.

Association Measures - Q-Mode

Distance Measures.--The association between the p stations (objects) can be conceptualized as resulting from ordering the different stations in a n-dimensional space in which the n axes are formed by the n species (descriptors). The most obvious way to measure association is to measure the distance between the stations (objects). In a two-dimensional space (only two species are used to describe the stations) the distance between the two stations is given by the familiar formula for the hypothenusa of a rectangular triangle derived by Pythagoras (Figure 14.2):

$$D = \sqrt{(y_{22}-y_{21})^2+(y_{12}-y_{11})^2}$$

This is readily generalized to n dimensions as:

$$D_1 = \sqrt{\sum_i (y_{i1}-y_{i2})^2}$$

However, there are several problems in applying the Euclidean distance: it has no upper limit but increases with the number of species, and it depends on the scale of the descriptors. Consequently, numerous other distance measures have been proposed. Their calculation is shown in Table 14.1. First, one may reduce the effect of the number of species by dividing by n, to find an average distance:

$$D_2 = \sqrt{\frac{1}{n} \sum_i (y_{i1}-y_{i2})^2}$$

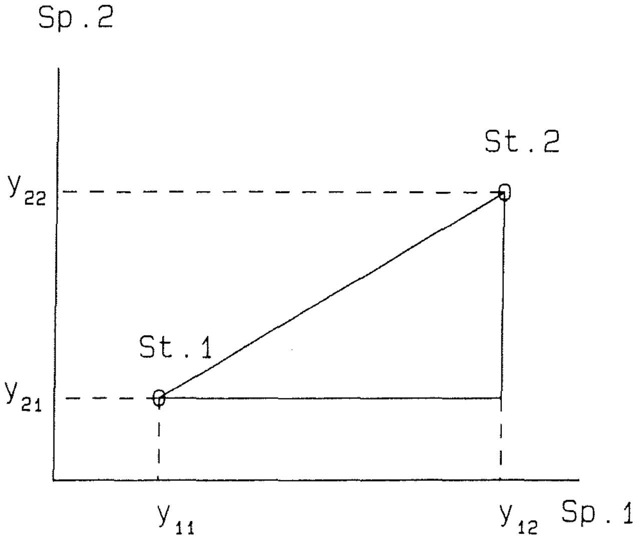

Figure 14.2.--Calculation of the distance between two stations (St. 1 and St. 2) in two-dimensional space (i.e., each station is described by two descriptors, one along each axis). The distance between the two points can easily be calculated by the Pythagorean formula.

Orloci (1967) proposes the chord distance D_3, varying between 0 and $\sqrt{n}$, which is an Euclidean distance after standardization of the descriptors (standardization: mean divided by the standard deviation; the objects will then be situated on a hypersphere with radius one). This may be applied to abundance data:

$$D_3 = \sqrt{2\left(1 - \frac{\sum_i y_{i1}y_{i2}}{\sqrt{\sum_i y_{i1}^2 \sum_i y_{i2}^2}}\right)}$$

Table 14.1. -- A practical example of the calculation of the different distance, similarity, and dependence measures given in the text. The data for this example are given at the start of the table. The steps necessary for the calculation of the distance and similarity between stations 1 and 2 are exemplified next, together with the results for all the measures defined. The third section exemplifies the calculation of the dependence measures for metric and ordered descriptors. Finally the calculation of two dependence measures for binary descriptors are shown.

1. Data Matrix

	St.1	St.2
Spec. 1	13	10
Spec. 2	1	4
Spec. 3	0	2
Spec. 4	1	0
Spec. 5	2	1
Spec. 6	0	2

2. Distance and Similarity Between Stations 1 and 2

	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5	Sp.6	Σ	Σy^2
Stat. 1								
y	13	1	0	1	2	0	17	175
y/Σy	0.76	0.06	0.00	0.06	0.12	0.00	1.00	
pres./abs.	1	1	0	1	1	0	4	
Stat. 2								
y	10	4	2	0	1	2	19	125
y/Σy	0.53	0.21	0.11	0.00	0.05	0.11	1.00	
p/a	1	1	1	0	1	1	5	

Calculations

1.	y1-y2	3	-3	-2	1	1	-2	4
2.	[y1-y2]	3	3	2	1	1	2	12
3.	$(y1-y2)^2$	9	9	4	1	1	4	28
4.	y1+y2	23	5	2	1	3	2	36
5.	$(y1+y2)^2$	529	25	4	1	9	4	572
6.	y1·y2	130	4	0	0	2	0	136
7.	%y1-%y2	0.24	0.15	0.11	0.00	0.07	0.11	0.72
8.	min(y1,y2)	10	1	0	0	1	0	12
9.	min(%y1,%y2)	0.53	0.06	0.00	0.00	0.05	0.00	0.64
10.	[y1-y2]/(y1+y2)	0.13	0.60	1.00	1.00	0.33	1.00	4.06
11.	a (p/p)	1	1			1		3
12.	b (p/a)				1			1
13.	c (a/p)			1			1	2

Table 14.1 (continued)

Distance Coefficients

D1 = $\sqrt{28}$ = 5.29

D2 = $\sqrt{28/6}$ = 2.16

D3 = $\sqrt{2(1 - \frac{136}{\sqrt{175.125}})}$ = 0.401

D4 = arc cos $(1 - \frac{0.401^2}{2})$ = 23.15

D6 = 12

D7 = 12/6 = 2

D8 = (1/2) (0.72) = 0.36

D8' = 2 (1 - 0.64) = 0.72

D9 = 4.06

D10 = 12/36 = 0.33

Similarity Coefficients

S1 = $\frac{3}{3+1+2}$ = 0.50

S2 = $\frac{6}{6+1+2}$ = 0.67

S3 = $\frac{3}{3+2+4}$ = 0.33

S4 = $\frac{2.12}{17+19}$ = 0.67

S5 = $\frac{1}{2}(\frac{12}{17} + \frac{12}{19})$ = 0.67

3. Dependence Between Stations

	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5	Sp.6	Σ	Σy^2	$\bar{y}$
Station 1									
y	13	1	0	1	2	0	17	175	2.8
$y-\bar{y}$	10.2	-1.8	-2.8	-1.8	-1.8	-2.8			
$(y-\bar{y})^2$	103.4	3.4	8.0	3.4	0.7	8.0	126.9		
R	1	3.5	5.5	3.5	2	5.5			
p/a	1	1	0	1	1	0			
Station 2									
y	10	4	2	0	1	2	19	125	3.2
$y-\bar{y}$	6.8	0.8	-1.2	-3.2	-2.2	-1.2			
$(y-\bar{y})^2$	46.7	0.7	1.4	10.0	4.7	1.4	64.9		
R	1	2	3.5	6	5	3.5			
p/a	1	1	1	0	1	1			
Calculations									
$(y1-\bar{y}1)(y2-\bar{y}2)$	69.5	-1.5	3.3	5.8	1.8	3.3	82.2		
d_j	0	1.5	2	2.5	3	2			
d^2_j	0	2.25	4	6.25	9	4	25.5		
a (p/p)	1	1			1		3		
b (p/a)				1			1		
c (a/p)			1			1	2		

Metric descriptors

Covariance s_{ki} = $\frac{82.2}{5}$ = 16.4

Variance y1: s^2_y = $\frac{126.9}{5}$ = 25.4

Variance y2: s^2_y = $\frac{64.9}{5}$ = 13.0

Correlation r_{12} = $\frac{16.4}{25.4 \times 13.0}$ = 0.903

Table 14.1 (continued)

Ordered descriptors

Spearman's rank correlation coefficient:

$$R_s = 1 - \frac{6 \times 25.5}{6^3 - 6} = 0.271$$

Kendall's rank correlation coefficient:

$$\tau = \frac{S}{\sqrt{(p^2\text{-}p\text{-}T_1)(p^2\text{-}p\text{-}T_2)}}$$

a) calculation of S:

yl (ordened):	1	2	3.5	3.5	5.5	5.5
y2	1	5	2	6	3.5	3.5

N higher ranks 5 1 3 0 0 0 N = 9 to the right in y2 (i.e., first value has rank 1, higher ranks are 5, 2, 6, 3.5 and 3.5, total 5. Second value has rank 5, higher value is 6, total 1. Third value has rank 2, higher values are 6, 3.5 and 3.5, etc.)

S = 4N - p(p-1) = 4x9 - 6x5 = 6

b) calculation of correction term for ties:

T = t(t-1), summation over all groups of ties, with t the number of values (ranks) in each tie.

T1 = 2x1 + 2x1 = 4 (two groups of ties with 2 members each in yl)

T2 = 2x1 = 2 (one group of ties with 2 members in y2)

c) final calculation:

$$\tau = \frac{6}{\sqrt{(36\text{-}6\text{-}4)(36\text{-}6\text{-}2)}} = 0.222$$

4. Binary descriptors (presence/absence)

Point correlation coefficient:

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

$$r = \frac{3x0 - 1x2}{\sqrt{(3+1)(3+2)(1+2)(1+0)}} = -0.258$$

Chi square:

$$X^2 = \frac{p((ad\text{-}bc) - (p/2))^2}{(a+b)(a+c)(b+c)(b+d)}$$

$$= \frac{6(-2-3)^2}{(4)(5)(3)(1)} = 2.5$$

It may be transformed into the geodesic metric as shown below. This measures the length of an arch on the surface of a hypersphere with radius one. An example for two species is shown in Figure 14.2.

$$D_4 = \text{arc cos } (1 - \frac{D_3^2}{2})$$

Another general distance measure has been proposed by Minkowski:

$$D_5 = \left(\sum_i |y_{i1}-y_{i2}|^r \right)^{1/r}$$

It reduces to the Euclidean distance for r = 2 and several other variants may be deduced from it. For r = 1 one obtains the Manhattan-metric:

$$D_6 = \sum_i |y_{i1}-y_{i2}|$$

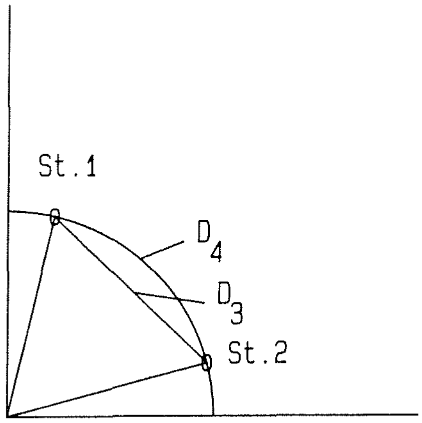

Figure 14.3.--Graphic representation of the difference between the distance measures D_3 and D_4 for two stations in two-dimensional space. In two dimensions, the stations are situated on a circle with radius 1 after standardization. D_3 is then the length of the chord (straight line) between the two points. D_4 is the length of the arch between lines.

The Manhattan-metric has the same disadvantages as the Euclidean distance. One therefore derives:

$$D_7 = \frac{1}{n} \sum_i |y_{i1}-y_{i2}|$$

The index of association of Whittaker is derived as:

$$D_8 = \frac{1}{2} \sum_i \left| \frac{y_{i1}}{\sum_i y_{i1}} - \frac{y_{i2}}{\sum_i y_{i2}} \right|$$

or as:

$$D'_8 = 2 \left[1 - \sum_i \min \left(\frac{y_i}{\sum_i y_i} \right) \right]$$

It is also applied to species abundances.

A variant of the Manhattan-metric is the Canberra-metric (Lance and Williams, 1966), quite popular in benthic ecology:

$$D_9 = \sum_i \frac{|y_{i1}-y_{i2}|}{(y_{i1}+y_{i2})}$$

Also widely used is the index proposed by Bray and Curtis:

$$D_{10} = \frac{\sum_i |y_{i1}-y_{i2}|}{\sum_i (y_{i1}+y_{i2})}$$

Distance and Similarity.--In the Q-mode association measures are of two forms: distance and similarity. Similarity S is maximal for two identical objects, distance D is maximal for two completely dissimilar objects. Between similarity and distance simple relationships exist such as $D = 1-S$, $D = \sqrt{1-S}$ or $D = \sqrt{1- S^2}$. Williams and Dale (1965) urged as a minimum requirement that the distance measure used should be metric (M), i.e., should have the following properties: (a) $D \geq 0$; (b) $D(a,b) = D(b,a)$; (c) $D(a,c) \leq D(a,b)+D(b,c)$ (inequality of the triangle). When this condition is not satisfied the distance measure is called semimetric (SM), when both conditions (a) and (c) are not satisfied it is called non-metric (NM). D1 to D9 are metric distance coefficients, D10 is semimetric.

Similarity Coefficients.--Similarity coefficients are used to measure the association between stations (objects). They are never metric and may not be used to position objects in a metric space as in ordination. However, the complements of several similarity coefficients are metric and may be used in ordination.

Similarity may be measured from simple presence and absence of species, based on the following table:

		Station 1	
		present	absent
Station 2	present	a	b
	absent	c	d

in which a, b, c and d are the numbers of species present in both stations, in station 1 or 2 only, or in none of the stations. In ecological applications, the double absence of a species is normally not taken into account. The following coefficients are among the more frequently used:

$$S_1 = \frac{a}{a+b+c} \quad \text{(Jaccard) (M)}$$

$$S_2 = \frac{a}{2a+b+c} \quad \text{(Sorensen) (SM)}$$

$$S_3 = \frac{a}{a+2b+2c} \quad \text{(Sokal and Sneath) (M)}$$

Similarity coefficients based on abundance but with exclusion of the double zeros may be derived from the binary coefficients shown above by calculating the amount of concordance, which may be based on presence or absence but also on abundance classes.

$$S = \text{concordance}/(\text{n-double zeros})$$

When abundance data are available, other coefficients use this information better. The following of frequent use are based on raw abundance data arranged as in the following example:

	Station 1	Station 2	Minimum
Sp. 1	13	10	10
Sp. 2	1	4	1
Sp. 3	0	2	0
Sp. 4	1	0	0
	A = 15	B = 16	W = 11

$$S_4 = \frac{2W}{A+B} \quad (=22/31 = 0.71) \text{ (Kulczynski) (SM)}$$

$$S_5 = \frac{1}{2}\left(\frac{W}{A} + \frac{W}{B}\right) \quad (=0.71) \text{ (Steinhaus) (SM)}$$

Coefficients of Dependence - R-Mode

In the R-mode (clustering and ordination of species) the primary goal is to determine the relationship between the species, although the association matrix may be used for an ordination of objects as well in certain cases (Principal Component Analysis). The coefficients that measure dependence between species (descriptors) are different according to the nature of the variables (metric, ordered non-metric, non-ordened).

Metric descriptors may be classified or ordinated with parametric dependence coefficients: the covariance s_{ki} and Pearson's correlation coefficient r_{ki} between species k and i.

$$s_{ki} = \frac{1}{p-1} \sum_j (y_{kj}-\bar{y}_k)(y_{ij}-\bar{y}_i)$$

$$r_{ki} = \frac{s_{ki}}{\sqrt{s_k^2 \, s_i^2}}$$

Note that the correlation coefficient is the covariance of the centered standardized variables.

Ordened non-metric descriptors (e.g., ranks, abundance classes) may be classified using the non-parametric correlation coefficients of Spearman R_s and Kendall τ:

$$R_s = 1 - \frac{6 \, \Sigma d_j^2}{p^3-p}$$

in which d_j is the difference in rank of the two species in sample (station) j.

$$\tau = \frac{2S}{p^2-p}$$

To calculate S, the two samples (stations) are ordened according to increasing rank of the species in one of the samples. One then adds one point for all p(p-1)/2 pairs of species of the other sample between which there is also an increase and substracts one point when there is a decrease. When there are ties (equal ranks), the formula has to be corrected (see Table 14.1 for an example of calculation).

Spearman's R_s is founded on the difference in rank occupied by the same object in the two series that have to be compared. Kendall's τ is based on the number of ranks of higher and lower order in the two series.

For non-ordened descriptors, the analysis is based on the use of contingency tables and the basic coefficient is X^2. Alternatively, the analysis may be based on the amount of information common to both descriptors B and the amount of information A and C exclusive to one of them, as calculated by the Shannon information function (see Legendre and Legendre, 1979).

Analysis of Species Abundances.--Since species abundances are normally metric, they may be classified using the parametric dependence measures described above. If the data are normalized to ln (1+y), covariance and correlation coefficients are appropriate. However, a large number of double zeros will affect the analysis. This may be remedied by excluding rare species or double zeros. Another problem is that covariance and correlation measure the correlation between abundance fluctuations, the degree to which species fluctuate together. Since one can define associations on the basis of co-occurrence as well, binary coefficients such as Jaccard and Sorensen, or the point correlation-coefficient, which is based on presence (1) and absence (0) (Daget, 1976), may be used instead.

Classification or Clustering

Classification or clustering aims at grouping objects that are sufficiently similar according to some criterion and the usual product is a graph called a dendrogram.

In benthic ecology, classification is nearly always based on hierarchical methods, i.e., the ensemble of samples (stations) is subdivided or the samples (stations) are grouped successively so that members of groups of inferior rank are also members of the groups with higher rank. Non-hierarchical systems exist as well (see Lance and Williams, 1967, for a review).

Two further choices of strategy are required. Firstly, the ensemble may be progressively subdivided into groups of diminishing size (divisive strategy) or a hierarchy may be constructed by fusing individual stations progressively into groups of increasing size (agglomerative strategy). Secondly, there is a choice to be made between monothetic and polythetic strategies. In the first strategy, groups are formed based on the presence or absence of a single descriptor (species). In the polythetic strategy the dichotomies are based jointly on a number of attributes so that groups are defined in terms of the overall similarity of their members.

Divisive Methods.--Except for agglomerative-monothetic methods, the three other possible combinations are potentially useful. Divisive methods are in principle based on all the available data. The best known divisive-monothetic procedure is that of Association Analysis (Williams and Lambert, 1959). The analysis is based on binary descriptors and one looks at the species that is most associated with the others by calculating X^2_{ik} between all possible pairs of descriptors (species) i and k from 2x2 contingency tables using the familiar formula:

$$\chi^2_{ik} = \frac{p(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

with a,b,c and d the numbers in the four cells of the contingency table.

After this, the sum of X^2_{ik} for all descriptions of each descriptor k is calculated:

$$\sum_i \chi^2_{ik} \qquad i \neq k$$

The largest sum corresponds to the species that is most strongly correlated with the others. The first division of the ensemble follows the descriptions of that descriptor. One forms a first group of stations coded 0 and a second group coded 1. The descriptor is then eliminated from the analysis and calculations are redone for each of the two groups separately, and so on.

Agglomerative-Polythetic Methods.--Benthic data matrices commonly have a large proportion of empty cells so that a large proportion of the data from a particular station are negative, i.e., absence of species. Treatment of benthic data is usually based on agglomerative methods. Agglomerative methods involve the successive fusion of those separate groups which are more similar according to some criterion at each stage. Two decisions have to be made in the selection of the procedure: the similarity measure, already discussed, and the strategy of fusion.

Starting with the matrix of between-group distances, the first operation is always to fuse the two most similar (nearest) stations or, if several pairs have the same smallest distance, to fuse the members of each pair. Two stations fused initially form a group and it is necessary to define the distance of other stations from this group. Lance and Williams (1964) considered five possible definitions:

(1) Nearest-neighbor: the distance between two groups is defined as the shortest distance between each possible pair of stations, one from each group.

(2) Furthest-neighbor: the distance between two groups is defined as the greatest distance between each possible pair of stations, one from each group.

(3) Centroid: a group is replaced on formation by the coordinates of its centroid, i.e., by the same number of stations each having the average composition of the group. The distance between two groups is the distance between the centroids.

(4) Median: a new group is formed as in centroid sorting but is placed midway the positions of the two groups forming it.

(5) Group-average: the distance between two groups is defined as the average distance between all possible pairs of stations, one from each group. A modification, weighted-average sorting exists, in which the group average distances of a third group to each of the two fusing groups are weighted equally.

Lance and Williams (1966) considered three properties to be important in fusing strategies:

(1) Compatible versus incompatible: measures calculated later in the analysis are of exactly the same kind as the initial measures versus not.

(2) Combinatorial strategies: consider two groups i and j forming a new group k, with n_i and n_j stations per group and intergroup distance d_{ij}. Consider a third group h. If the distance d_{hk} of groups h and k can be calculated from d_{hi}, d_{ij}, n_i, and n_j, the strategy is said to be combinatorial. Lance and Williams postulated:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma (d_{hi} - d_{hj})$$

In the following table the values for the coefficients are given:

	α_i	α_j	β	γ
Near. neighbor	1/2	1/2	0	-1/2
Furth. neighbor	1/2	1/2	0	1/2
Centroid	n_i/n_k	n_j/n_k	$-\alpha_i/\alpha_j$	0
Median	1/2	1/2	-1/4	0
Group Average	n_i/n_k	n_j/n_k	0	0
Weight. Average	1/2	1/2	0	0
Flexible Sorting	$1/2(1-\beta)$	$1/2(1-\beta)$	β	0

(3) Space-conserving or space-distorting: the initial distances between the separate stations may be regarded as defining a space with known properties. By forming groups this space may be conserved or not. When not, it may be contracted when there is a tendency for stations to join an already existing group rather than act as a nucleus for a new group, resulting in a chained hierarchy of little ecological interpretation, as in nearest-neighbor. On the other hand, with furthest-neighbor groups appearing to move away from some or all of the remaining stations and space is dilated.

The potential value of a varying degree of space distortion led Lance and Williams to propose a flexible sorting strategy with the constraints:

$$\alpha_i + \alpha_j + \beta = 1 \quad ; \quad \alpha_i = \alpha_j \quad ; \quad \beta < 1 \quad ; \quad \gamma = 0$$

This strategy is combinatorial and compatible for Euclidian distance (and Sorensen's coefficient) but not for the correlation coefficient. As β approaches unity, the system becomes increasingly space-contracting, as β falls to zero and becomes negative it becomes increasingly space-dilating. In most ecological practice, flexible sorting with a coefficient $\beta = -0.25$ appears to be satisfactory.

Which Strategy to Use?.--For the reader not wishing to go into all the details of this overview: in benthic ecology a now often used and satisfactory procedure is group average sorting of stations based on the Bray-Curtis similarity index calculated on double root transformed densities of the species.

Ordination

When dealing with a large number of species (descriptors) or stations (objects) the simple application of uni- or bivariate statistical methods not only is cumbersome but may be misleading as well. Multivariate methods have become indispensable tools in ecology since the general availability of large computers made their use practical. They have two basic roles: (1) to discover structure in the data, and (2) to summarize the data objectively. In contrast with classical statistics, which are concerned with hypotheses testing, ordination tries to elicit some internal structure from which hypotheses can be generated (Williams and Gillard, 1971).

Ordination is simply an operation by which objects are placed along axes that correspond to relationships of order, or on graphs formed by two or more of these axes. The relationships may be metric or not. Ordination tries to reduce the number of dimensions in which the dispersion of stations or species is represented so that the great tendencies of variability in the sample for the ensemble of all descriptors are distinguished. The dispersion of stations (objects) is first represented in a mulidimensional graph with as many axes as species (descriptors). One then looks at the projections in planes of these multi-dimensional graphs which are of most interest. These planes are defined by new axes that permit representation of the variability in the data in an optimal way in a space with reduced dimensions. Therefore, the end product of an ordination is a graph, usually two-dimensional, in which similar species or samples (or both) are near each other and dissimilar ones are far apart.

Many software packages exist in which the most important ordination techniques are included, among which BMD, SPSS, CLUSTAN and NTSYS are widely available. Other programs have been published by Blackith and Reyment (1971), Davies (1971), Orloci (1978), and Hill (1974, 1979a).

Principal Component Analysis (PCA).--The most powerful of ordination techniques, Principal Component Analysis (PCA), is most readily visualized when the stations are described by only two species (Figure 14.4). The stations may then be represented in a two-dimensional graph with the two species as axes. The relationships between the stations can be represented as well by any two other axes in the same plane. PCA in its original form works on centered data, meaning that first the origin of the axes is moved to the centroid of the data (the point representing the mean densities of the species over all stations). The analysis then proceeds by projecting the stations onto a line through the centroid and so oriented that the sum of the squared distances of the stations to the line is minimized, or, what amounts to the same, the sum of the squared projections of the stations on the line is maximized. This line, obtained as a rotation of the original axes around the centroid, is the first principal component or axis. In the two species case, the second principal component is perpendicular to the first in the original plane. If there are more than two species, the second axis could be placed in any direction perpendicular to the first. It is so orientated that a maximum amount of the variability not explained by the first axis is explained by the second.

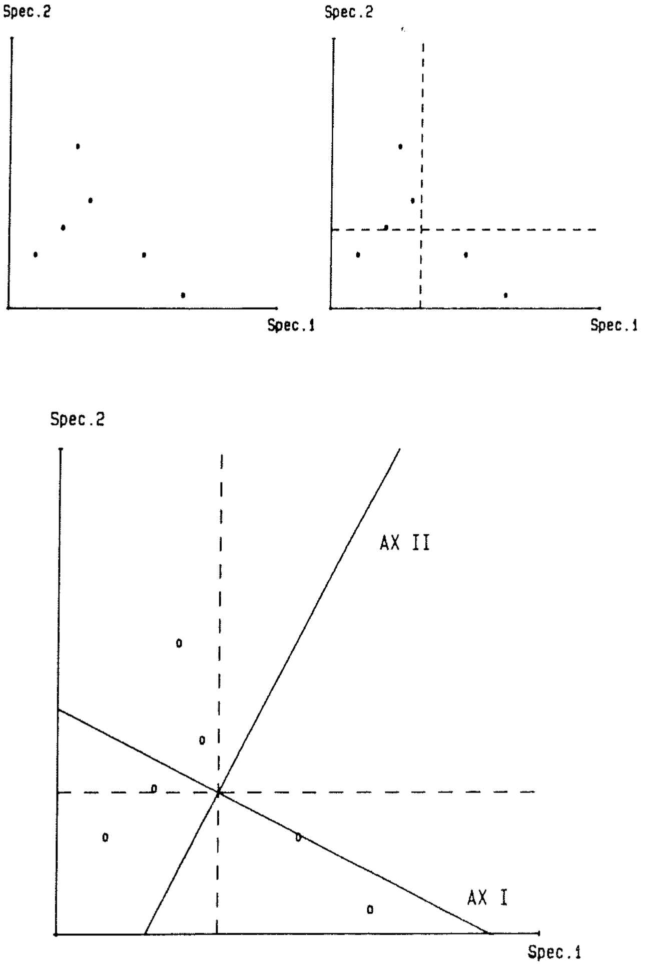

Figure 14.4.--The basic operations of Principal Component Analysis, exemplified for a two-dimensional space. Stations are plotted in this space (top left), axes are centered (top right), and rotated (bottom). The figure is fully explained in the text.

PCA in itself does not produce a reduction in dimensionality, since the number of components is equal to the original number of species. However, by changing from the original axes to a new orthogonal set of axes, it concentrates the variability in the data in the successively derived axes. If the first few axes extracted are accepted as adequate to display the information in the original data, then a reduction in dimensionality has been achieved.

The new axes (z) derived from PCA may be represented as algebraic functions of the old original axes (y):

$$z_i = a_{i1}y_1 + a_{i2}y_2 + \ldots + a_{in}y_n$$

for the i-th new axis in a system containing n species. The a's are constants calculated in such a way that the station's distances from the z-axes are minimized: they are called the components loadings. These loadings represent the characteristic of a particular species along a particular axis and species with similar distributions will thus have similar loadings.

In general, one has to resort to matrix algebra and a computer in order to solve for the new axes and the position of the stations. The starting point is an association matrix A of similarities between stations based on correlation coefficients or covariances/variances. Such a matrix has N eigenvalues which are the values on the diagonal of the matrix when all other elements have been made zero. Associated with each eigenvalue λ is an eigenvector V. The eigenvectors give the loadings of the n species on the p axes and thus define the principal axes. When normalized to unity they permit one to find the principal components. The value of λ for a component is proportional to the total variability accounted for by that component. When covariances are used, $\Sigma\lambda$ is equal to the total variance, when correlation coefficients are used, $\Sigma\lambda$ is equal to the number of species.

Transformation.--PCA has been developed in relation to analyses of psychological tests where centering and standardization of the data are appropriate and sometimes necessary. The use of covariances and correlations implies such transformation. However, it is not always clear in an ecological context whether these explicit or implied transformations are valid. From an extensive discussion by Noy-Meir (1973) and Grieg-Smith (1983) we retain that among several possibilities the following may be noted:

(1) Raw data: y_{ij} : product-moment correlation.

(2) Centered data $y_{ij} - \bar{y}$: covariance.

(3) Standardized by total y_{ij}/ny_i.

(4) Standardized and centered: $y_{ij}-\bar{y}_i/s_i$:correlation.

When untransformed data are used, stations will be weighted by species richness for presence/absence data or total density for quantitative data. Centering will alter the weighings.

Other Ordination Methods

PCA.--PCA can only validly be used on cross-product similarity coefficients and relationships between stations may be more validly expressed by distance measurements. Principal coordinate analysis (Gower, 1966) permits the analysis of every association matrix Q based on a metric distance coefficient.

The calculations are as follows:

(1) The starting point is a matrix of metric distances D_{ij}.

(2) This matrix is transformed into a new matrix A by defining $a_{ij} = -\frac{1}{2}D^2_{ij}$.

(3) The new matrix A is centered to form a matrix α by calculating $\alpha_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}$.

(4) The eigenvalues and eigenvectors of α are calculated; the eigenvectors are standardized to the square root of their eigenvalue.

(5) The standardized eigenvectors are placed in columns: the rows in such a table are the Principal Coordinates of the stations (objects).

If an environmental gradient exists there often is a non-linearity between similarity measures and the between-station distance along the underlying gradient. This phenomenon is at the basis of most of the unsatisfactory features of PCA. Ordination based on ranking of similarity seems attractive and has been developed by Kruskal (1964) in a technique called nonmetric multidimensional scaling.

The calculations are as follows:

(1) The starting point may be every matrix of similarity or distance or even non-metric ordered scores. From this a matrix of distances D_{ij} (pxp) is calculated.

(2) A number of axes, t, thought to be sufficient is chosen a priori (often t=2).

(3) The configuration of the p objects in the reduced t-dimensional space is determined: starting from a random order, a matrix Δ with distances δ_{ij} is calculated in reduced space using an appropriate distance measure.

(4) The dispersion of distances D_{ij} in relation to distances δ_{ij} is graphed and a regression between them is calculated.

(5) The regression coefficients aid in the calculation of an approximation $\hat{D}_{ij}$ of the value necessary to conserve the distance relationship monotonously. The values form a new matrix $\hat{D}$.

(6) One then calculates the "stress" S between the matrices $\hat{D}$ and Δ and repeats steps (4) to (6) until a minimum value of this stress is found.

$$S = \sqrt{\sum_i \sum_j (\delta_{ij}-\hat{D}_{ij})^2 / \sum_i \sum_j (\delta_{ij}-\bar{\delta})^2}$$

Reciprocal Averaging and Detrended Correspondance Analysis.--Reciprocal Averaging (Hill, 1974) is a form of PCA in which the species ordination scores are averages of the station ordination scores and vice versa. It amounts to a double standardization and has considerable advantages in that being non-centered it is efficient with heterogeneous data but has two faults. It shows the arch or horseshoe effect: a linear gradient of composition is expressed as an arch in two dimensions of the ordination. The second fault is that equivalent differences in composition are not represented by the same differences in first axis position (see Gauch, 1984, for an extensive discussion).

The modified version of reciprocal averaging, detrended correspondance analysis (DCA) (Hill, 1979a) is now one of the most widely used ordination methods in benthic ecology. The arch effect is removed by adjusting the values on the second axis in successive segments by centering them to zero mean. The variable scaling on the axes is corrected by adjusting the variance of species cores within stations to a constant value. The software for this ordination technique is available in the package DECORANA.

The calculation is as follows:

(1) Each abundance value y_{ij} of the matrix is standardized according to rows and columns:

$$\frac{y_{ij}}{\sqrt{r_i c_j}}$$

in which

$$r_i = \sum_i y_{ij} \text{ and } c_j = \sum_j y_{ij}$$

(2) The matrix of covariances S is then calculated.

(3) The eigenvalues and eigenvectors of S are extracted.

The direct iteration method (Hill, 1974) is an efficient algorithm for obtaining one to several axes. One starts by assigning arbitrary species ordination scores. The weighted averages are used to obtain sample scores from these species scores. The second iteration produces new species scores by weighted averages of the sample scores, and so on. Iterations are continued until the scores stabilize. The scores converge to a unique solution, not affected by the initial, arbitrary scores.

Ordination - Space Partitioning

A set of very powerful methods has been developed in which the data are first subjected to an ordination and then to a divisive classification. They are explained fully in Pielou (1984) on which this account is based.

The Minimum Spanning Tree.--When the p stations are plotted in a n-dimensional species space, the points of the swarm may be linked by a minimum spanning tree, a set of line segments linking all the n points in the swarm in such a way that every pair of points is linked by one and only one path. None of the paths form closed loops. The length of the tree is the sum of the lengths of its constituent line segments. The minimum spanning tree of the swarm is the tree with the minimum length. When this tree is formed, it is divided by cutting in succession first the longest link in the tree, then the second longest link etc.

This is done starting from a pxp matrix of distances. The first segment in the tree is the shortest distance in the matrix. Next, the shortest distance linking a third point to either one of the first two points is found. Next, the shortest distance linking a fourth point to either one of the first three points is found, etc. (this procedure amounts to nearest neighbor clustering).

Lefkovitch's Partitioning Method.--This method consists in first ordering the data in n-space by means of a centered PCA. Then the first division is made by breaking the first axis at the centroid, the second by breaking the second axis at the centroid, etc.

TWINSPAN.--This method consists of carrying out a one-dimensional Reciprocal Averaging ordination and breaking the axis at the centroid so as to divide the data points into two classes. Each of the two classes is then split in the same way, after a RA ordination, etc. (Hill, 1979b).

Part 4. Species-Abundance Distributions

If one records the abundances of different species in a community, invariably one finds that some species are rare, whereas others are more abundant. This feature of ecological communities is independent of the taxonomic group(s) or the area(s) investigated. An important goal of ecology is to be able to describe consistent patterns in different communities, and explain them in terms of biotic and abiotic interactions.

Although "community" is defined as the total set of organisms in an ecological unit (biotope), it must be specified as to the actual situation. No entities exist within the biosphere with absolutely closed boundaries, i.e., without interactions with other parts. Some kind of arbitrary boundaries should always be drawn. Pielou (1975) recommends to specify explicitly the following features: (1) The spatial boundaries of the area or volume containing the community and the sampling methods, (2) the time limits between which observations were made, (3) the taxocene (i.e., the set of species belonging to the same taxon) treated as constituting the community.

The results of a sampling program of the community take the form of species lists, indicating for each species a measure of its abundance (usually number of individuals per unit surface, although other measures, such as biomass, are possible). Many methods are used to plot these data. The method chosen often depends on the kind of model one wishes to fit to the data. Different plots of the same (hypothetical) data set are shown in Figure 14.5.

It can readily be seen that a bewildering variety of plots is used. They yield quite different visual pictures, although they all represent the same data set. Figure 14.5a-d are variants of the Ranked Species Abundance (*RSA*) curves. The *S* species are ranked from 1 (most abundant) to *S* (least abundant). Density (often transformed to percentage of the total number of individuals *N*) is plotted against species rank. Both axes may be on logarithmic scales. It is especially interesting to use a log-scale for the Y-axis, since then the same units on the Y-axis may be used to plot percentages and absolute numbers (there is only a vertical translation of the plot).

In so-called "*k*-dominance" curves (Lambshead et al., 1983) (Figure 14.5e-f) the cumulative percentage (i.e., the percentage of total abundance made up by the *k*th-most dominant plus all more dominant species) is plotted against rank *k*, or log rank *k*. To facilitate comparison between communities with different numbers of species *S*, a "Lorenzen curve" may be plotted. Here the species rank *k* is transformed to *(k/S) * 100*. Thus the X-axis always ranges between 0 and 100 (Figure 14.5g).

The "collector's curve" (Figure 14.5h) addresses a different problem. When one increases the sampling effort, and thus the number of animals *N* caught, new species will appear in the collection. A collector's curve expresses the number of species as a function of the number of specimens caught. Collector's curves tend to flatten out as more

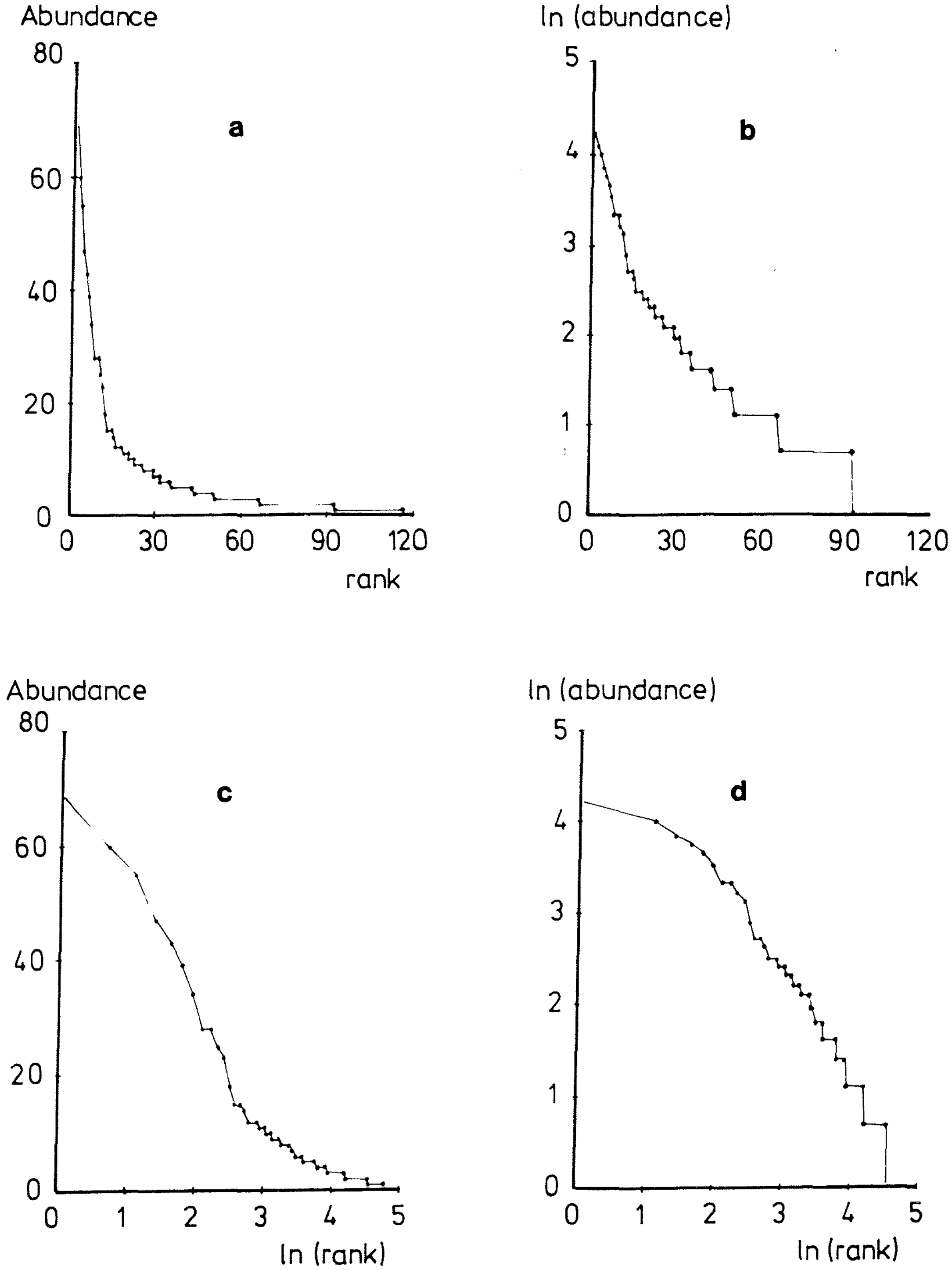

Figure 14.5.—"Fifty ways to lose your reader" in representing species-abundance data. The different figures represent the same species-abundance data. The curves are explained in detail in the text. **a-d**, Ranked Species Abundance curves with none, one, or both axes on a log scale.

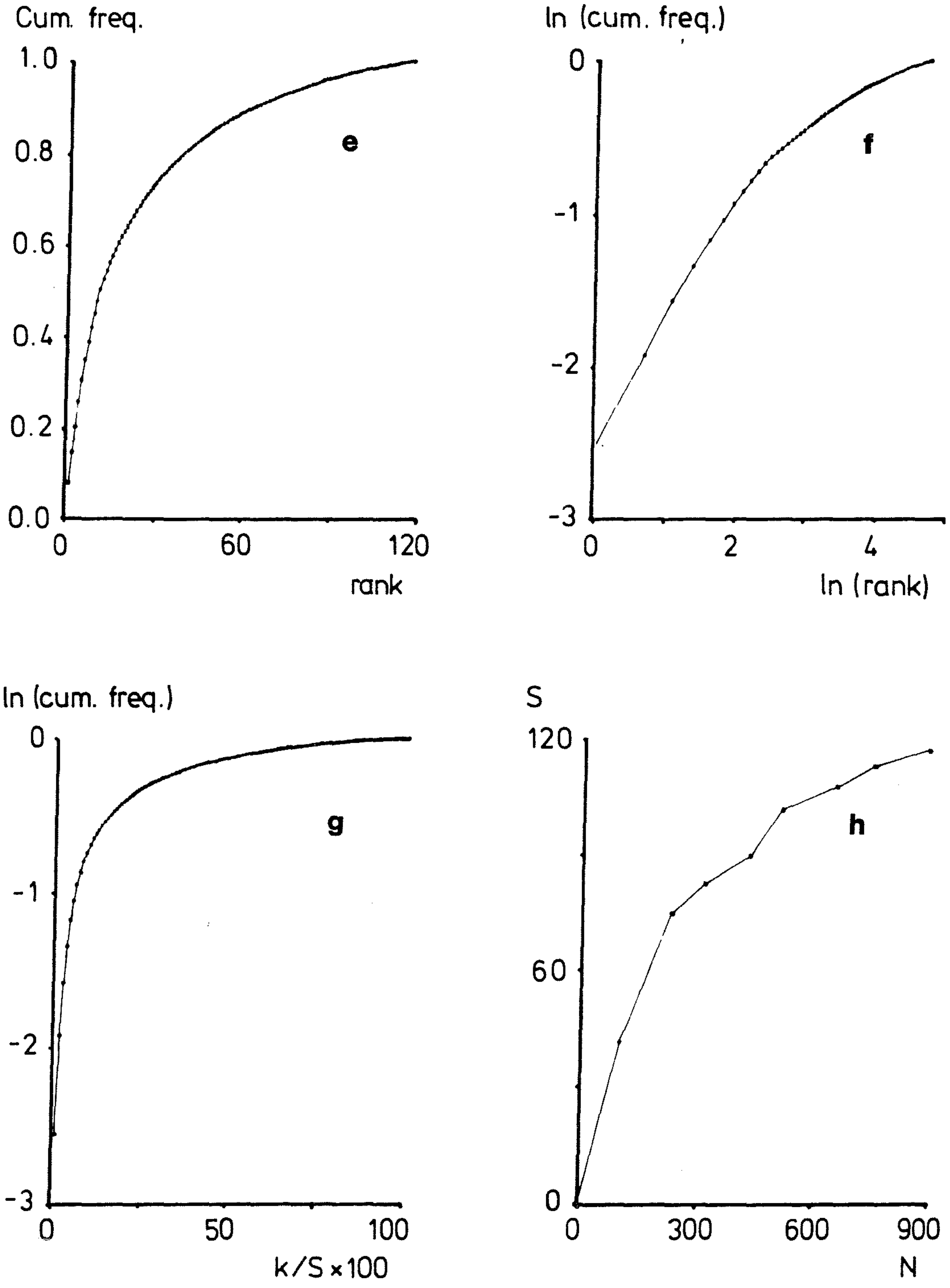

Figure 14.5 (continued).– e-f, k-dominance curves; g, "Lorenzen curve"; h, collector's curve.

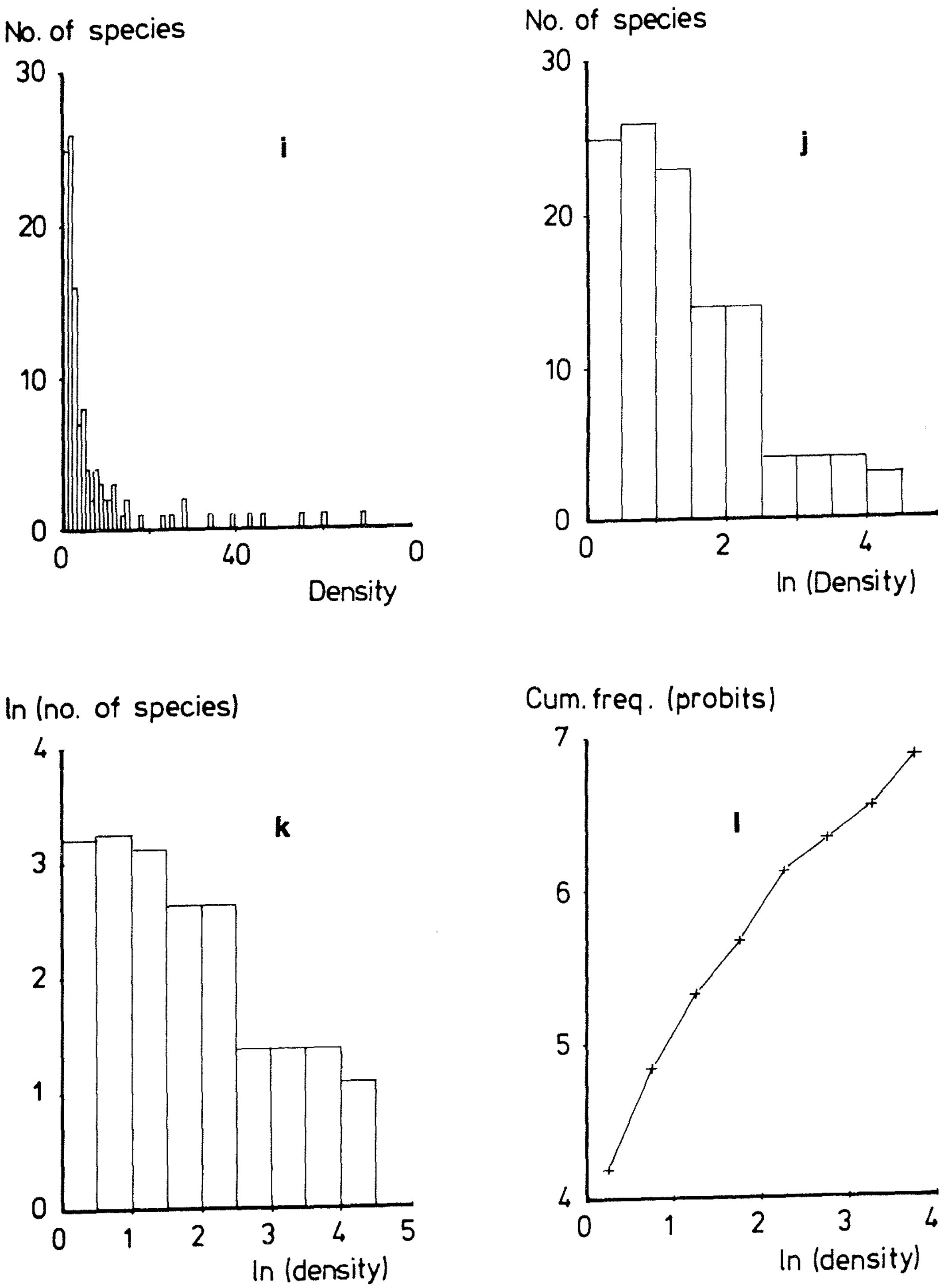

Figure 14.5 (continued).– i-k, species abundance distributions; l, cumulative species abundance distribution on probit scale.

specimens are caught. However, due to the vague boundaries of ecological communities they often do not reach an asymptotic value: as sampling effort (and area) is increased, so is the number of slightly differing patches.

The plots in Figure 14.5i–k are species abundance distributions. They can only be drawn if the collection is large, and contains many species (a practical limit is approximately S>30). Basically, a species–abundance distribution (Figure 14.5i) plots the number of species that are represented by r = 0, 1, 2, ... individuals against the abundance r. Thus, in Figure 14.5i there were 25 species with 1 individual, 26 species with 2 individuals, etc. More often than not, the species are grouped in logarithmic density classes. Thus one records the number of species with density e.g., between 1 and exp(0.5), between exp(0.5) and exp(1), etc. (Figure 14.5j). A practice, dating back to Preston (1948), is to use logarithms to the base 2. Thus one has the abundance boundaries 1, 2, 4, 8, 16, etc. Although these so-called "octaves" are still used, they have two disadvantages: the class boundaries are integers, which necessitates decisions as to which class a species with an abundance equal to a class mark belongs; the theoretical formulation of models is "cluttered" (May, 1975) by factors ln (2), which would vanish if natural logs were used. The ordinate of species–abundance distributions may be linear or logarithmic. Often one plots the cumulative number of species in a density group and all less abundant density groups on a probit scale (Figure 14.5l).

Species–Abundance Models

Two kinds of models have been devised to describe the relative abundances of species. "Resource apportioning models" make assumptions about the division of some limiting resource among species. From these assumptions a ranked abundance list or a species–abundance distribution is derived. The resource apportioning models have mainly historical interest. In fact, observed species abundance patterns cannot be used to validate or discard a particular model, as has been extensively argued by Pielou (1975, 1981). One should consult these important publications before trying to validate or refute fortuitously a certain model!

"Statistical models" make assumptions about the probability distributions of the numbers in the several species within the community, and derive species–abundance distributions from these.

The Niche Preemption Model (Geometric Series Ranked Abundance List).--This resource apportioning model was originally proposed by Motomura (1932). It assumes that a species preempts a fraction k of a limiting resource, a second species the same fraction k of the remainder, and so on. If the abundances of the species are proportional to their share of the resource, the ranked–abundance list is given by a geometric series:

$$k,\ k(1-k), \ldots,\ k(1-k)^{(s-2)},\ k(1-k)^{(s-1)}$$

where S is the number of species in the community. May (1975) derives the species abundance distribution from this ranked abundance list (see also Pielou, 1975).

The geometric series yields a straight line on a plot of log abundance against rank. The communities described by it are very uneven, with high dominance of the most abundant species. It is not very often found in nature. Whittaker (1972) found it in plant communities in harsh environments or early successional stages.

The Negative Exponential Distribution (Broken Stick Model).--A negative exponential species abundance distribution is given by the probability density function:

$$\Psi(y) = S\ e^{-Sy}$$

Stated as such, it is a statistical model, an assumption about the probability distribution of the numbers in each species. However, it can be shown (Webb, 1974) that this probability density function can be arrived at via the "broken–stick model" (MacArthur, 1957).

A limiting resource is compared with a stick, broken in S parts at $S-1$ randomly located points. The length of the parts is taken as representative for the size of the S species subdividing the limiting resource. If the S species are ranked according to size, the expected size of species i, y_i, is given by:

$$E(y_i) = \frac{1}{S} \sum_{x=1}^{S} \frac{1}{x}$$

The negative exponential distribution is not often found in nature. It describes a too even distribution of individuals over species to be a good representation of natural communities. According to Frontier (1985) it is mainly appropriate to describe the right–hand side of the rank frequency curve, i.e., the distribution of the rare species. As these are the most poorly sampled, their frequencies depend more on the random elements of the sampling, than on an intrinsic distribution of the frequencies.

Pielou (1975, 1981) showed that a fit of the negative exponential distribution to a field sample does not prove that the mechanism modeled by the broken–stick model governs the species–abundance pattern in the community. Moreover, the broken-stick model is not the only mechanism leading to this distribution. The same prediction of relative abundance can be derived by at least three other models besides the niche partitioning one originally used (Cohen, 1968; Webb, 1974).

The observation of this distribution does indicate (May, 1975) that some major factor is being somewhat evenly apportioned among the community's constituent species (in contrast to the lognormal distribution, which suggests the interplay of many independent factors).

The Log-series Distribution.--The log-series was originally proposed by Fisher et al. (1943) to describe species abundance distributions in large moth collections. The expected number of species with r individuals, E_r, is given as:

$$E_r = \alpha \frac{X^r}{r}$$

(r = 1,2,3,...). α ($\alpha > 0$) is a parameter independent of the sample size (provided a representative sample is taken), for which X ($0 < X < 1$) is the representative parameter. The parameters α and X can be estimated by maximum likelihood (Kempton and Taylor, 1974), but are conveniently estimated as the solutions of:

$$S = -\alpha \ln (1-X)$$

and

$$N = \frac{\alpha X}{1-X}$$

The parameter α, being independent of sample size, has the attractive property that it may be used as a diversity statistic (see further). An estimator of the variance of $\hat{\alpha}$ is given as:

$$\hat{\text{var}}\ (\hat{\alpha}) \cong \frac{\hat{\alpha}}{-\ln X(1-X)} \quad \text{(Anscombe, 1950).}$$

Kempton and Taylor (1974) give a detailed derivation of the log-series distribution. It was fitted to data from a large variety of communities (e.g., Williams, 1964; Kempton and Taylor, 1974). However, it seems to be generally less flexible than the log-normal distribution. In particular, it cannot account for a mode in the species-abundance distribution, a feature often found in a collection. According to the log-series model, there are always more species represented by 1 individual than there are with 2. The truncated log-normal distribution can be fitted to samples with or without a mode in the distribution.

Caswell (1976) derived the log-series distribution as the result of a neutral model, i.e., a model in which the species abundances are governed entirely by stochastic immigration, emigration, birth and death processes, and not by competition, predation or other specific biotic interactions. He proposes to use this distribution as a "yardstick," with which to measure the occurrence and importance of interspecific interactions in an actual community. Other models have been proposed to generate the log-series distribution (Boswell and Patil, 1971) but they all contain the essentially neutral element as to the biological interactions. However, the proof that any form of biological interaction will yield deviation from the log-series is not given. Neither is it proven that "neutral" communities cannot deviate from the log-series. Therefore, we think that the fit of this distribution cannot be considered as a waterproof test for species interactions.

The Log-normal Distribution.--Preston (1948) first suggested the use of a log-normal distribution for the description of species-abundance distributions. It was shown by May (1975) that a log-normal distribution may be expected, when a large number of independent environmental factors act multiplicatively on the abundances of the species (see also Pielou, 1975).

When the species-abundance distribution is log-normal, the probability density function of y, the abundance of the species, is given by:

$$\Psi(y) = \frac{1}{y\sqrt{2\pi V_z}} \exp \left(\frac{-(\ln y - \mu_z)^2}{2 V_z} \right)$$

The mean and variance of y are:

$$\mu_y = \exp \left(\mu_z + \frac{V_z}{2}\right)$$

and

$$V_y = (\exp(V_z)-1)\exp(2\mu_z+V_z)$$

where μ_y and V_2 are the mean and variance of $z = \ln(y)$.

If the species abundances are lognormally distributed, and if the community is so exhaustively sampled that all the species in the community (denoted S^*) are represented in the sample, the graph of the cumulative number of species on a probit scale (Figure 14.51) against log abundance will be a straight line. This is not normally the case.

In a limited sampling a certain number of species S^*-S will be unrepresented in the sample (S being the number of species in the sample). The log-normal distribution is said to be truncated. In the terminology of Preston (1948) certain species are hidden behind a "veil line" (see insets in Figure 14.6), it follows that it is not a good practice to. estimate the parameters of the lognormal distribution from a cumulative plot on probit scale. In fact if one does not estimate the number of unsampled species, it is impossible to estimate the proportion of the total number of species in a particular log density class. Species abundances that are lognormally distributed will not yield straight lines if one takes into account only the species sampled (see Figure 14.6). Note also that the normal regression analysis is not applicable to highly correlated values such as cumulative frequencies. (If the frequencies are replaced by evenly distributed random numbers, their cumulative values on probit scale still yield very "significant" correlations with log abundance). In order to fit a lognormal distribution it is absolutely necessary to estimate the number of unrepresented species, S^*-S.

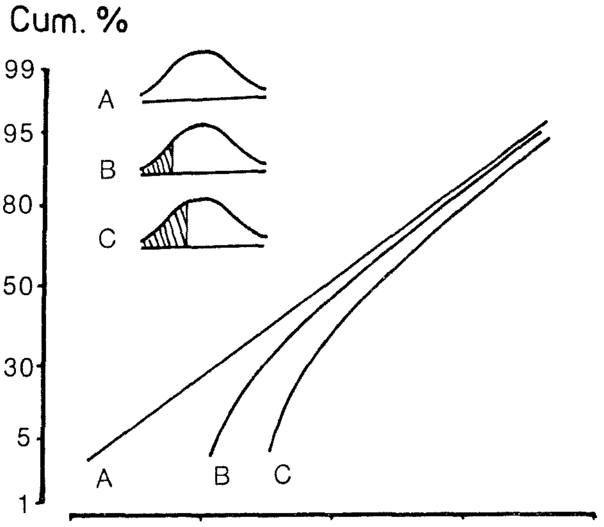

Figure 14.6.--Effect on the probability plot of cumulative % of species vs. abundance of truncating the log-normal distribution (A), by 15% (B), and 30% (C). The hatched areas in the insets represent the unsampled portion of the species (i.e., the species hidden behind the "veil line"). Scale of x-axis arbitrary. After Shaw et al. (1983).

In fitting the log-normal two procedures are used (apart from the wrong one already discussed). The conceptually most sound method is to regard the observed abundances of species j as a Poisson variate with mean λ_j, where the λ_j's are lognormally distributed. The probability, P_r, that a species contains r individuals is then given by the Poisson log-normal distribution (see Bulmer, 1974). P_r can be solved approximately for $r > 10$, but must be integrated numerically for smaller values of r. Bulmer (1974) discusses the fitting to the data by maximum likelihood. Pielou (1975) argues that the fitting of the Poisson lognormal, though computationally troublesome, is not materially better than the alternative procedure, consisting in the direct fitting of the continuous lognormal. The complete procedure in recipe-form is given in Pielou (1975).

The Mandelbrot Model.--This relatively flexible model was derived in information science to model rank-frequency curves of messages in complicated systems (e.g., words in a natural language). It describes the frequency of a species with rank r as:

$$f_r = f_o(r+\beta)^{-\gamma}$$

in which $-\gamma$ is the slope of the asymptote towards which the curve approximates; β is related to the deviation at the left-hand side of the curve (Figure 14.7). This model is extensively discussed in Frontier (1985), where useful references may be found. The model is particularly useful to describe the rank-frequencies of the dominant species in a community. However, for the rare species large deviations may be found.

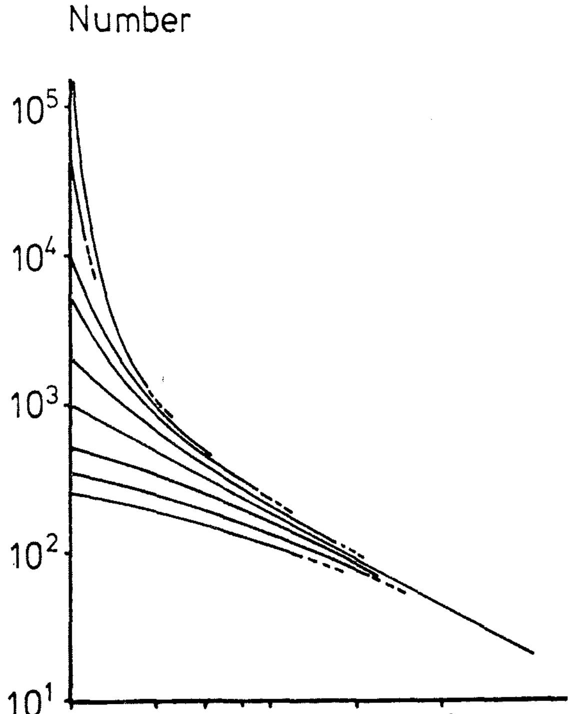

Figure 14.7.--Curves for the Mandelbrot model on a log-log scale, with N_0 = 1000, γ = 1, and the following values for β (from top down): -1, -0.98, -0.8, -0.5, 0 (straight line), 1, 2, 3. After Frontier (1985).

On Fitting Species-abundance Distributions.--Ever since Fisher et al. (1943) used the log-series, and Preston (1948) proposed the log-normal to describe species-abundance patterns, ecologists have been debating about which model is the most appropriate. Especially the log-normal and the log-series have (had) their fan-clubs, recently also among benthologists (e.g., Shaw et al., 1983; Gray, 1983, and other papers). In our opinion, these debates are spurious. As Pielou (1975) remarked, the fact that, e.g., the log-normal fits well in many instances, tells us more about the versatility of the log-normal than about the ecology of these communities. Although most of the distributions have a kind of biological rationale (to make them more appealing to a biological audience?) the fact that they fit does not prove that the "biological" model behind them is valid in the community.

The fitting of a model to field data is meaningful if the parameter estimates are to be used in further analysis. This is analogous to the use of the normal distribution in ANOVA: in order to perform an ANOVA, the data should be normally distributed. Of course this must be checked, but only as a preliminary condition. No one draws conclusions from the fit or non-fit of the normal distribution to experimental data, but from the test performed afterwards. Similarly, if a particular model fits reasonably well to a set of field data, the parameter estimates can be used, e.g., in respect to the diversity of the communities.

Indices of Diversity and Evenness

It is common practice among ecologists to complete the description of a community by one or two numbers expressing the "diversity" or "evenness" of the community. For this purpose a bewildering *diversity* of diversity indices have been used or proposed.

Two different aspects are generally accepted to contribute to the intuitive concept of diversity of a community. These are (following the terminology of Peet, 1974) *species richness*, a measure somehow related to the total number of species in the community (note that the actual number of species in the community is usually unmeasurable), and *equitability*, which expresses how evenly the individuals are distributed among the different species. Some indices, called *heterogeneity* indices by Peet (1974), incorporate both aspects.

It has been clearly demonstrated (May, 1975) that no single diversity index can summarize the species-abundance distribution in a community. However, this is seldom a goal in itself in ecology. Usually, one tries to show how some characteristic feature(s) of the ecosystem may change in relation to evironmental variables. Depending on the situation, several indices may give a good indication of these relations. Anyway, it is useful to keep in mind that a complete specification of the species-abundance relationship contains more information than a single index.

Indices Derived from Species-abundance Distributions

Historically, the first diversity measure was derived by Fisher, Corbet, and Williams (1943) as a result of the derivation of the log-series distribution. As mentioned earlier, the parameter α of the log-series distribution is independent of sample size. From the first equation under the topic of log-series distribution, it is easily seen that α is the only parameter describing the relationship between number of individuals and number of species in the sample. Thus, this parameter describes the way in which the individuals are divided among the species, which is a measure of diversity. Note that, in adopting the log-series model for the species-abundance distribution, the equitability is already specified, so that α only measures the relative species richness of the community. α, as determined by the fitting of the log-series model to the sample, is only valid as a diversity index when the log-series fits well to the data. The same reasoning can be extended to the log-normal distribution. Preston (1948) expressed the diversity as the (calculated) total number of species in the community, S^*.

The use of the log-series α was taken up again, and extended by Kempton and Taylor (1974). Taylor et al. (1976) showed that, when the log-series fits the data reasonably well, α has a number of attractive properties. The most important of these was that (compared to the information statistic H' and Simpson's index; see below) it provided a better discrimination between sites, it remained more constant within each site (all sites were sampled in several consecutive years), it was less sensitive to density fluctuations in the commonest species, and it was normally distributed. On the other hand, when the data deviate from the log-series, α is more dependent on sample size than the other indices.

Kempton and Taylor (1976), and Kempton and Wedderburn (1978) extended this approach, by noting that for the log-series distribution the parameter α is the asymptotic expectation of Q, a "mid-range statistic," defined as:

$$Q = (S^*/2) \;/\; \log(P_{S^*/4} \;/\; P_{3S^*/4})$$

where S^* is the total number of species in the community and the proportional abundances P_i(i = 1, 2, ... S^*) are arranged in descending order of size.

For discrete data this index is estimated from the sample statistic:

$$a_Q = \frac{\frac{1}{2}n_{R_1} + \sum_{r=R_1+1}^{R_2-1} n_r + \frac{1}{2} n_{R_2}}{\log (R_2/R_1)}$$

where the sample quartiles, R1 and R2 are chosen such that:

$$\sum_{r=1}^{R_1-1} n_r < \frac{S^*}{4} \leq \sum_{1}^{R_1} n_r$$

$$\sum_{r=1}^{R_2-1} n_r \leq \frac{3S^*}{4} < \sum_{1}^{R_2} n_r$$

and n_r is the number of species with abundance *r*. It can be seen from Figure 14.8 that *Q* depends mostly on the abundance of the moderately abundant species. *Q* may either be estimated directly from the data or alternatively from the parameter estimates of a fitted species-abundance distribution. For the log-series, $\hat{Q}=\alpha$; for the log-normal, $\hat{Q}=0.371\ S^*/\hat{\sigma}$ provides a reasonable approximation if more than 50 % of the species are represented in the sample. If *Q* is estimated in a series of samples by fitting the same model to each of the data sets, it is an index of species richness (or, if the fit is too bad or the sample too small, of nothing anymore). However, if *Q* is estimated directly from the data, it incorporates both species richness and equitability.

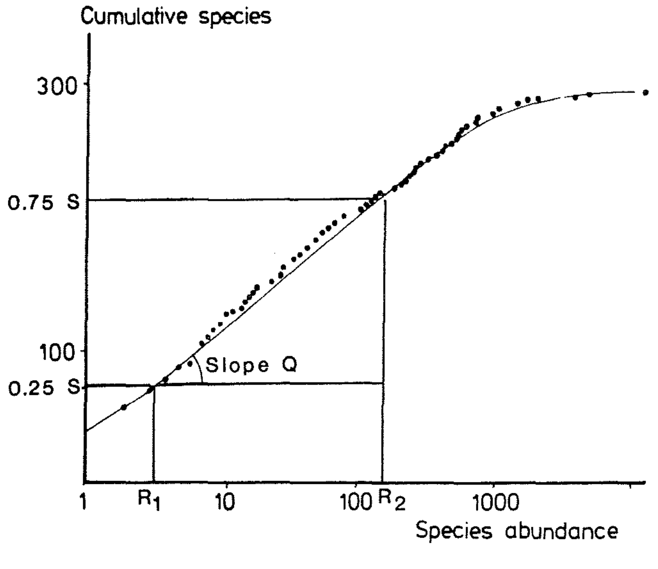

Figure 14.8.--Graphic determination of the "mid-range statistic" Q from a plot of cumulative number of species vs. log (species abundance). Q can be seen to approximately equal the slope of this curve for the moderately abundant species. After Kempton and Wedderburn (1978).

Rarefaction

An obvious index of species richness is the number of species in the sample. However, it is clear that this measure is highly correlated with sample size, an undesirable property. Sanders (1968) proposed a method to reduce samples of different sizes to a standard size, so as to make them comparable in terms of the number of species. The formula used by Sanders (1968) was corrected by Hurlbert (1971), who showed that the expected number of species in a sample of size n is given by:

$$E(S) = \sum_{i=1}^{S} \left[1 - \binom{N-N_i}{n} / \binom{N}{n} \right]$$

where N_i is the number of individuals in the *i*-th species in the full sample, which had sample size *N* and contained *S* species. Alternatively, random samples can be drawn by computer from the original sample (Simberloff, 1972).

Hill's (1973) Diversity Numbers

Hill (1973) provided a generalized notation that includes, as a special case, two often used heterogeneity indices. Hill defined a set of "diversity numbers" of different order. The diversity number of order *a* is defined as:

$$N_a = (\sum_i p_i^a)^{1/(1-1)}$$

where p_i is the proportional abundance of species *i* in the sample.

N_0 can be seen to equal *S*, the number of species in the sample.

N_1 is undefined by the first of these equations. However, defining

$$N_1 = \lim_{a \to 1} (N_a)$$

it can be shown that

$$N_1 = \exp (- \sum_i p_i (\ln p_i))$$

$$= \exp (H')$$

where *H'* is the well known Shannon-Wiener diversity index:

$$H' = -\sum_i p_i\ (\log p_i)$$

Note that in the usual definition of the Shannon-Wiener diversity index, logarithms to the base 2 are used. Diversity then has the peculiar units "bits ind^{-1}". The diversity number N_1 is expressed in much more natural units. It gives an equivalent number of species, i.e., the number of species S' that yields N_1 if all species contain the same number of individuals, and thus if all $p_i=1/S'$. This can be seen in the last equation which in this case reverts to:

$$N_1 = \exp\ (-\ln(1/S')) = S'$$

An additional advantage of N_1 over H' is that it is approximately normally distributed. It has been argued (see e.g., Pielou, 1975) that for small, fully censused communities the Brillouin index should be used. This index is given by:

$$H = \frac{1}{N} \log \frac{N!}{\prod_i N_i!}$$

We do not recommend this index for meiofaunal assemblages. The information theoretical argument for its use should be regarded as allegoric: it has no real bearing to ecological theory. Furthermore, finite collections that are non-destructively sampled do not occur in meiobenthic research. The most compelling argument, however, is given by Peet (1974), who shows with an example that the Brillouin index has counterintuitive properties: depending on sample size, it can yield higher values for less evenly distributed communities.

The next diversity number, N_2, is the reciprocal of Simpson's dominance index, which is given by:

$$\lambda = \sum_i p_i^2$$

for large, sampled, communities. If one samples at random and without replacement, 2 individuals from the community, Simpson's index expresses the probability that they belong to the same species. Obviously, the less diverse the community is, the higher is this probability.

In small, fully censused communities, the correct expression for Simpson's index is:

$$\lambda = \sum_i \frac{N_i(N_i-1)}{N(N-1)}$$

where N_i = number of individuals in species *i*, and N is the total number of individuals in the community. Pielou (1969) shows that for large communities which are sampled, the first equation is a biased estimator of λ. An unbiased estimator is given by the last equation, where λ and N are then sample values, not parametric values as in the case of fully sampled communities.

In order to convert Simpson's dominance index to a diversity statistic it is better to take reciprocals, as is done in Hill's N_i, than to take $1-\lambda$. In that way the diversity number is again expressed as an equivalent number of species.

The diversity number of order + infinity (N_{+oo}) is equal to the reciprocal of the proportional abundance of the commonest species. It is the "dominance index." May (1975) showed that it characterizes the species-abundance distribution "as good as any, and better than most" single diversity indices. Recently it has received some attention in the context of pollution monitoring (Shaw et al., 1983.

Hill (1973) showed that the diversity numbers of different orders probe different aspects of the community. The number of order $+oo$ infinity only takes into account the commonest species. At the other extreme, N_{-oo} is the reciprocal of the proportional abundance of the rarest species, ignoring the more common ones. The numbers N_0, N_1, and N_2, are in between in this spectrum. N_2 gives more weight to the abundance of common species (and is, thus, less influenced by the addition or deletion of some rare species) than N_1. This, in turn, gives less weight to the rare species than N_0, which, in fact, weighs all species equally, independent of their abundance. It is good practice to give diversity numbers of different order when characterizing a community. Moreover, these numbers are useful in calculating equitability (see below).

The Subdivision of Diversity

Hierarchical Subdivision.--In the calculation of diversity indices, all species are considered as different, but equivalent: one is not concerned with the relative differences between species. However, in nature some species are much more closely related to some other species than to the rest of the community. This relation may be considered according to different criteria, e.g., taxonomic relationships, general morphological types, trophic types, etc. Therefore, it may be desirable to subdivide the total diversity in a community in a hierarchical way. Pielou (1969) shows how the Shannon-Wiener diversity H' can be subdivided in a hierarchical way. The species are grouped in genera, and the total diversity H'_t equals:

$$H'_T = H'_g + H'_{wg}$$

where

$$H'_g = - \sum_i q_i \log q_i$$

is the between-genera diversity, and

$$H'_{wg} = \sum_i q_i \left(- \sum_j r_{ij} \log r_{ij}\right)$$

is the average within-genus diversity.

The same procedure may be repeated to partition the between-genera diversity into between-families and average within-family diversity.

This approach was generalized by Routledge (1979) who showed that the only diversity indices that can be consistently subdivided are the diversity numbers of Hill (1973) (of which H' can be considered a member, taking into account the exponential transformation). The decomposition formulae are:

$$\left(\sum_i \sum_j t_{ij}^a\right)^{1/(1-a)} = \left(\sum_i q_i^a\right)^{1/(1-a)} *$$

$$*\left\{\left(\sum_i q_i^a \sum_j r_{ij}^a\right)/\sum_i q_i^a\right\}^{1/(1-a)}$$

for $a \neq 1$

and

$$\prod_i \prod_j t_{ij}^{-t_{ij}} = \prod_i q_i^{-q_i} \prod_i \left(\prod_j r_{ij}^{-r_{ij}}\right)^{q_i}$$

for $a=1$

In these equations, q = the proportional abundance of the group (e.g., genus) *i*, r_{ij} = the proportional abundance of species *j* in group *i*, and t_{ij} = the proportional abundance of species *j* (belonging to group *i*) relative to the whole community:

$$t_{ij} = q_i r_{ij} \quad \text{and} \quad q_i = \sum_k t_{ik}$$

It can be seen that the community diversity is calculated as the *product* of the group diversity and the average diversity within groups, weighted by the proportional abundance of the groups. Note that this is consistent with Pielou's formulae noted in the previous equations since $N_1 = \exp(H')$.

The hierarchical subdivision of diversity may be useful to study the differences in diversity between two assemblages. Is a higher diversity in one assemblage mainly attributable to the addition of some higher taxa (suggestive of the addition of new types of niches), or rather to a diversifying of the same higher taxa that are present in the low-diversity assemblage?

It may also be useful to study groups other than taxonomic ones. Natural ecological groupings, such as the feeding types of nematodes (Wieser, 1953) or the body types of harpacticoids (Hicks and Coull, 1983) may be particularly interesting. Heip et al. (1984, 1985) used as a "trophic diversity index" to describe the diversity in feeding types of nematodes, the index:

$$\Theta = \sum_{i=1}^{4} \vartheta_i^2$$

where q_i is the proportion of feeding type *i* in the assemblage. This index can be seen to be the reciprocal of N_2 (for estimation purposes, it would be better to use the equation for Simpson's index although the bias is small when 200 nematodes are sampled). This approach can be naturally extended, using last equation to (1) diversity indices of other orders, and (2) achieve a more complete description of the assemblage, by a subdivision of the total diversity.

Spatio-Temporal Diversity Components.--All ecological communities are variable at a range of spatio-temporal scales. Thus if one examines a set of samples (necessarily) taken at different points in space, and possibly also in time, and calculates an overall diversity index, it is unclear what is actually measured. Whereas diversity may be small in small patches at a particular instant, additional diversity may be added by the inclusion in the samples of diversity components due to spatial or temporal patterns.

Following Whittaker (1972) one often discerns for the spatial component: (1) *Alpha*-diversity - the diversity within a uniform habitat (patch); (2) *Beta*-diversity - the rate and extent of change in species composition form one habitat to another (e.g., along a gradient); and (3) *Gamma*-diversity - the diversity in a geographical area (e.g., the intertidal range of a salt marsh).

The subdivision of total diversity H' in ecological components is discussed by Allen (1975). He treats a sampling scheme where S species are sampled in q sites, each consisting of r microhabitats. The problem is different from a hierarchical subdivision, since the same species may occur in different microhabitats and sites (it can, of course, only belong to one genus, one family, etc., in hierarchical subdivision). Allen (1975) presents two solutions. One can treat the populations of the same species in different microhabitats as the fundamental entities. Total diversity is then calculated on the basis of the proportional abundance (in relation to the total abundance in the study) of these populations. This total diversity can then be subdivided hierarchically.

Alternatively, one can subdivide the species diversity in the total study in average within microhabitat diversity, average between microhabitat (within site) diversity, and average between site diversity components. The latter computations are generalized for Hill's (1973) diversity numbers by Routledge (1979).

Equitability

Several equations have been proposed to calculate equitability (evenness) from heterogeneity measures. The most frequently used measures, which converge for large samples (Peet, 1974) are:

$$E = (D-D_{min}) \cdot / \ (D_{max}-D_{min})$$

and

$$E = D \ / \ D_{max}$$

where D is a heterogeneity index, and D_{min} and D_{max} are the lowest and highest values of this index for the given species number and the sample size. To this class belongs Pielou's $J = H' / H'_{max} = H' / \log S$.

As discussed by Peet (1974) these measures depend on a correct estimation of S^*, the number of species in the community. It is quasi impossible to estimate this parameter. Substituting S, the number of species in the sample, makes the equitability index highly dependent on' sample size. It also becomes very sensitive to the chance inclusion-exclusion of rare species in the sample.

Hill (1973) proposed to use ratios of the form:

$$E_{a:b} = N_a \ / \ N_b$$

as equitability indices (where N_a and N_b are diversity numbers of order a and b respectively). Note that $H' - H'_{max} = 1n \ (N_1 N_0)$ belongs to this class, but that $J' = H' / H'_{max}$ does not. These ratios are shown to possess superior characteristics, compared with J'. Hill (1973) also showed that in an idealized community, where the hypothesized number of species is infinite and the sampling is perfectly random, $E_{1:0}$ is always dependent on sample size. $E_{2:1}$ stabilizes, with increasing sample size, to a true community value. However, in practice all measures depend on sample size.

Heip (1974) proposed to change the index $E_{1:0} = e^{H'}/S$ to $(e^{H'}-1)/(S-1)$. In this way the index tends to 0 as the equitability decreases in species-poor communities. Due to a generally observed correlation between equitability and number of species in a sample, $E_{1:0}$ tends to $1/S$ as both $e^{H'} \longrightarrow 1$ and $S \longrightarrow 1$.

In general, one cannot attach too much importance to equitability indices. Species-abundance distributions show more information about the equitability than any single index.

The Choice of an Index

Lambshead et al. (1983) have noted that, whenever two k-dominance curves do not intersect, all diversity indices will yield a higher diversity for the sample represented by the lower curve. Equivocal results arise as soon as the k-dominance curves intersect. Different measures of diversity are more sensitive to either the commonest or the rarest species (see Hill's diversity numbers). An elegant approach to the analysis of this sensitivity is provided by the response curves of Peet (1974).

In order to summarize the diversity characteristics of a sampled community, it is advisable to provide the diversity numbers N_0, N_1, N_2, possibly also N_{+00}, the dominance index. If permitted by the sampling scheme, one can use these indices in a study of hierarchical and/or spatio-temporal components of diversity. Equitability indices should be regarded with caution. Hill's (1973) ratios $E_{a:b}$ seem preferable if an index is desired. However, one should rather use species-abundance plots to study equitability.

In any case, it should be remembered that the indices depend on sample size, sample strategy (e.g., location of the samples in space and time), spatio-temporal structure of the community, and sampling error.

Although formulae for the estimation of the variance of H' have been proposed, these do not include all these sources of error. Frontier (1983) estimated the background noise in a time series of H' in plankton samples as roughly one unit (bit/ind.). However, in a species-poor community of meiobenthic copepods, Heip and Engels (1974) found that the variance of H' was conservatively estimated by the formula of Hutcheson (1970). Typical coefficients of variation (s/x) were about 0.3. No index was normally distributed.

The diversity statistics derived from species-abundance patterns are less sensitive to the inclusion/exclusion of rarities, and to the density variations of the most abundant species. They may be more useful in the description of inherent diversity characteristics of communities, and in the discrimination between sites (see Kempton and

Wedderburn, 1978). On the other hand, large samples are required for their estimation. Thus, a relatively large number of environmental patches may be pooled, without the possibility of separation.

Finally, we should stress the possibilities and limitations of diversity indices. A diversity index must be regarded as a summary of a structural aspect of the assemblage. As has been stressed throughout this chapter, different indices summarize slightly different aspects. In comparing different assemblages, it is useful to compare several indices: this will indicate specific structural differences.

The variability of an index within one assemblage cannot be estimated very well. As has been indicated, the estimate of the variance of an index (Hutcheson, 1970) should be used with caution. Spatio-temporal variability of the assemblage should preferably be included in the study, in order to evaluate the variability of the index. Indices may then be transformed (if necessary) so that they are approximately normally distributed, and compared using standard statistical techniques.

A diversity index summarizes the structure, not the functioning of a community. Thus, it is very well possible that two assemblages have a similar diversity, whereas the mechanisms leading to their structures are completely different (e.g., Coull and Fleeger, 1977). Often these functional aspects cannot readily be studied by observing resultant structures, and may require an experimental approach.

Part 6. Time Series Analysis

The analysis of ecological time series is a topic of both practical and theoretical interest. In pollution monitoring studies, pollution effects can only be detected as a divergence between the observed state of some ecological variable, and the range predicted for undisturbed conditions. In the absence of knowledge on the dynamics of a system, prediction is possible when it is based on a long enough time series (Poole, 1978).

From a theoretical point of view, a wide variety of temporal patterns in populations and ecosystems is predicted by many models. Population dynamics models show that populations can exhibit types of behavior in time which range from extreme stability to apparently chaotic behavior (May and Oster, 1976), including periodic and pseudoperiodic cycles. In the field of systems theory, the dynamics of non-linear open systems far from equilibrium are obviously appealing for theoretical ecology. The essence of these systems' dynamics is an unstable but coherent behavior consisting of periodicities and cycles. Finally, some methods of time series analysis may be applicable to studies of spatial structure, in particular where some environmental characteristic varies in an oscillatory way along a gradient (e.g., ripple marks: Hogue and Miller, 1981).

Time series can either be studied in the time domain or in the frequency domain. In the time-domain a probabilistic model is fitted to the data. This model can be used for prediction. The core of the frequency domain studies is spectral analysis. It is analogous to an analysis of variance: the total variance of the series is attributed to oscillations of different periodicities. Whereas the time domain studies are usually superior for prediction purposes, the study in the frequency domain is rather a probe into the structure of the time series. It may lead to an identification of processes acting on the variable. The methods for time series analysis have mainly been developed in the physical and economic sciences. They have a number of features which may represent serious limitations for their application in ecology, and which should be kept in mind when analyses of this kind are planned.

The main body of theory has been developed for the analysis of single series, although extension to a limited number of simultaneous series is possible. For the analysis of the evolution of community structure in time, where a relatively large number of species densities may be recorded at any moment, the ecologist must recur to methods based on ordination or classification.

Many methods in time series analysis are only applicable to relatively long series (more than 100 data points is often considered as constituting a "short series"), which may involve a lot of work if e.g., species densities must be recorded on all dates. Some of the methods discussed below, however, may be used for shorter series. The data should be equispaced for almost all analyses. This is a very important point to consider before starting the sampling. Although in some cases it may be possible to interpolate missing values, one should try to keep the sampling interval as constant as possible from the start onwards. Again this restriction may be weakened for some simple, robust analyses (see e.g., Herman and Heip, 1986). Also in another way the sampling strategy determines the possible outcomes of the analyses. When looking for systematic oscillations in the data (as e.g., in spectral analysis) it is impossible to detect oscillations with a period longer than the time span sampled. Usually a more restrictive factor is given, and anyhow oscillations with a period longer than 1/4 of the time span sampled should be treated with caution. On the other hand, the shortest period that can be resolved is 2 Δt (where Δt is the sampling interval). If important oscillations are missed because the sampling interval is too long, these may still show up in the results of the analyses and yield erroneous conclusions, a phenomenon known as aliasing.

Diagnostic Tools

An essential first step in time series analysis is to plot the data as a function of time. Several features may be apparent: the presence of trend, seasonality, outliers, turning points (where a rising trend changes into a declining one or vice versa), etc. It is often apparent from inspection of the data if a transformation is desirable.

A time series is called stationary if its mean, variance, and higher moments are time-independent. Most analyses require the series to be stationary; this should be checked early in the analysis, usually at the plotting stage. In practice one is usually concerned with non-stationarity of the mean and variance. The mean is non-stationary when, e.g., an increasing or decreasing trend, or an obvious periodic ("seasonal") fluctuation is present.

Biological data, especially population densities, often need transformation. Transformation may stabilize the variance when, in the original series, it varies with the mean. Moreover, it can make seasonal variation additive instead of multiplicative. Several transformations are used in ecology. Legendre and Legendre (1986) give an instructive overview of their properties.

Trend and seasonal effects may be described and/or removed by several techniques (see below). A relatively simple but effective test for the presence of trend makes use of Kendall's correlation coefficient (Kendall, 1976). Consider a series $x_1,...,x_n$, and count the number of times $x > x_i$ for $j > i$. Call this number P. Its expected value in a random series is $n(n-1)/4$, i.e., half the number of comparisons made. P is related to Kendall's τ by the formula:

$$\tau = \frac{4P}{n(n-1)} - 1$$

The significance of τ can be tested by comparison to tabulated values. A positive significant τ points to a rising trend, a negative value to a falling trend.

An important diagnostic tool in time series analysis is the correlogram. This is a plot of the autocorrelation function with lag *k* against *k*. The autocorrelation function with lag *k* expresses the correlation between observations k units apart in time. It is estimated (for reasonably large *N*, and *k* not greater than about *N*/4) by:

$$r_k = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{N} (x_t - \bar{x})}$$

Defining the (sample) autocovariance function C_k as:

$$c_k = \frac{1}{N} \sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})$$

it can be seen that $r_k = c_k/c_o$. This is also the computational formula.

An approximate 5% confidence interval for r_k is $\pm 2 \sqrt{N}$. In a purely random series 5% of the values are expected to lie outside this range.

The correlogram reveals much of the structure of a time series: trend, periodic oscillations, alternation of points (zig-zag type of time series). Typical patterns in the correlogram point to suitable models to fit to the series. In analogy to the autocovariance function, a cross-covariance function of two simultaneous series is defined as:

$$c_{xy}(k) = \begin{cases} \sum_{t=1}^{N-k} (x_t - \bar{x})(y_{t+k} - \bar{y})/N \\ \quad [k=0,1,..,(N-1)] \\ \sum_{t=1-k}^{N} (x_t - \bar{x})(y_{t+k} - \bar{y})/N \\ \quad [k=-1,..,-(N-1)] \end{cases}$$

The cross-covariance plot reveals at which time-lag(s) the two series show a maximum correspondence.

Description and Removal of Trend

The presence of trend is most often the reason for the non-stationarity of the mean in time series. Trend must be removed from the series for most further analyses, but it can be of interest in itself. For example, Heip and Herman (1985) conclude from the near absence of trend in copepod respiration (as opposed to trend in the densities of the separate species, and in "structural" characteristics of the copepod assemblage) that the assemblage is continually changing in composition and structure, but is functionally stable.

Trend may be described (and removed) by different methods. A function (e.g., linear, quadratic, exponential, Gompertz, hyperbolic,...) may be fitted by least squares to the data. This is analogous to a regression calculation and yields entirely valid estimates of the parameters. However, when applying a regression program for these calculations, one should not use the tests of significance, nor the estimates of the standard error of the parameters which are usually provided in the program output. Due to the serial correlation usually present in time series, the basic assumption of independence of the Y's in regression is violated.

An advantage of curve-fitting is that the trend may be summarized in a few parameters. Disadvantages are that updating is computationally cumbersome (for every data point added to the series, the whole fitting must be redone), and that the estimate of trend in, e.g., the first year of the time series is continually changing with each data point added to the series, even 10 years later.

The fitting of moving averages avoids these problems. In essence, one fits a polynomial of order p (to the choice of the investigator) to the first 2m+1 data points in order to estimate the trend term in the (m+1)th point. Then a polynomial of the same order is fitted to the 2nd through (2m+2)th points to estimate the trend in the point m+2 etc.. In practice this procedure reverts to the calculating of linear combinations of the data points with tabulated coefficients. These differ with the order of the polynomial and the length 2m+1 of the data segment. The coefficients, and guidelines for a proper choice of the order p of the polynomial, can be found in Kendall (1976) and Kendall and Stuart (1968).

Inevitably, the fitting and distraction of a moving average distorts the cyclical and random elements of the remainder of the series. Cyclical terms with periods longer than the length of the segment taken into account for the computation of the moving average, will tend to be incorporated into the trend. Terms with shorter periods will remain in the residuals. Moreover, the distraction of a moving average from a purely random series will induce non-zero autocorrelations in the remainders, and thus give the impression that a systematic pattern is present in the data. This phenomenon is known as the Slutzky-Yule effect. It is well studied, and the period of the induced oscillation can be well approximated (Kendall, 1976; Legendre and Legendre, 1986).

Finally, an effective method to remove non-stationarity due to trend from the data is to difference the series. For a linear trend, define the difference operator ∇, such that the differenced series:

$$y_t = \nabla x_{t+1} = x_{t+1} - x_t$$

Occasionally second-order differencing is used, where the differenced series is differenced again. This method is most often used in the fitting of ARIMA models to a series (see below).

Seasonal Effects

Seasonal effects are often prominently present in ecological data. Their causes are usually well known, and are often not the incentive for the long-term study of the system. They may therefore be hindering the interpretation of other, more interesting features.

Seasonal effects may be removed in several ways. A simple, robust method is described by Kendall (1976). A 1-year moving average is calculated, and subtracted from the data. The residuals for all Januaries, all Februaries, etc. (in the case of monthly data) are averaged over all the observations for that month. These are the raw estimates of the seasonal effects. The final estimates are calculated by scaling the seasonal effects such that they sum to zero. The seasonal effects are then subtracted from the data.

Alternatively, a seasonal differencing may be used. This is done with the seasonal difference operator ∇^d, where $\nabla^d x_t = x_t - x_{t-d}$.

Time Domain Studies

Time domain studies use a class of models called ARIMA models (Autoregressive Integrated Moving Average). This class of models was developed for an integrated modeling strategy by Box and Jenkins (1976). Basically, ARIMA models contain several "building blocks":

--An autoregressive process of order *p* is defined as:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + a_t$$

where the error terms a_t are independently, identically distributed normal random variables with zero mean and variance $\sigma^2{}_a$, and the series is corrected for a (stationary) mean. This equation has a structure similar to a multiple regression, but the "regression" is not on independent variables, but on the series itself. This explains the term "autoregression."

--A moving average process of order *q* is:

$$x_t = a_t + \theta_1 a_{t-1} + \ldots + \theta_q a_{t-q}$$

A MA is a model of processes where random events affect not only the present state of the variable, but also have repercussions on its future state.

AR and MA processes are related to each other: AR processes of finite order can be written as infinite-order MA processes and vice versa. For parameter parsimony it is often useful to combine both processes in a "ARMA (p,q)" process:

$$x_t = \phi_1 x_{t-1} + .. + \phi_p x_{t-p} + a_t + \theta_1 a_{t-1} + .. + \theta_q a_{t-q}$$

ARMA processes are stationary, and applicable only to stationary series.

An ARMA model, fitted to a differenced series, is called an ARIMA model of the original series. It is called "integrated" because it has to be summed or integrated to provide a model of the non-stationary (undifferenced) series.

The fitting of an ARIMA model is performed in steps. The data are transformed and differenced to obtain a normally distributed, stationary series. The autocorrelation function is plotted and compared with theoretical autocorrelation functions of ARIMA processes. A suitable model is selected and fitted. The residuals are checked for systematic deviations from the model. If necessary the model is reformulated, fitted, etc., until a model has been selected that provides the best fit with the least possible number of parameters. In practice a suitable computer program package is necessary, and the model fitting requires a good deal of experience. The reader is referred to Chatfield (1976) for an introductory text and useful references. Detailed descriptions of the method are provided by Box and Jenkins (1976).

Examples of the use of ARIMA models in an ecological context are Poole (1978) and Keller (1987).

Spectral Analysis

The power spectral density function, or spectrum of a discrete stationary time series is a function $f(\omega)$ of frequency ω, with the following physical interpretation: $f(\omega)d\omega$ is the contribution to the variance of the series by components with frequencies in the range $(\omega, \omega + d\omega)$. When $f(\omega)$ is plotted against ω, the surface under the curve equals the total variance in the series. A peak in the spectrum indicates a frequency with a particularly important contribution to the explanation of the variance.

Thus the spectrum of a time series with a clear seasonal oscillation (e.g., temperature in a temperate climate) will have a high peak on a frequency of 1 yr^{-1}.

The spectrum is a parametric function (in the same way as μ is usually defined as the parametric mean of a set of data). Its estimation from the data is the aim of spectral analysis.

The interpretation of the spectrum becomes clear if we look at the derivation of the periodogram, which is closely related to it.

According to a fundamental result of Fourier analysis a finite time-series $\{x_t\}(t = 1,2,...,N)$ can be decomposed in a sum of sine and cosine functions of the form:

$$x_t = a_o + a_{N/2} \cos \pi t + \sum_{p=1}^{N/2-1} \{a_p \cos(2\pi pt/N) + b_p \sin(2\pi pt/N)\}$$

where $a_o = \bar{x}$

$$a_{N/2} = \Sigma\ (-1)^t\ x_t/N$$

$$a_p = 2\ \{\Sigma\ x_t\ \cos(2\pi pt/N)\}/N$$

$$b_p = 2\ \{\Sigma\ x_t\ \sin(2\pi pt/N)\}/N$$

$$[p=1,..,N/2-1]$$

Note that the first equation has *N* parameters to describe *N* observations. It fits the data exactly. The component with frequency:

$$\omega_p = 2\pi p/N$$

is called the p-th harmonic. the amplitude of the p-th harmonic is given as:

$$R_p = \sqrt{(a_p^2 + b_p^2)}$$

It can be shown that:

$$\Sigma\ (x_t - \bar{x})^2/N = \sum_{p=1}^{N/2-1} R_p^2/2 + a_{N/2}^2$$

This equation expresses how the total variance of the series is divided over the harmonic components. If we assume that the series has a continuous spectrum, we can regard $R^2{}_p/2$ as the contribution to the explanation of the variance of all components with frequencies in the range $\omega_p \pm \pi/N$.

The periodogram is a plot of $I(\omega)$ against ω, where

$$I(\omega) = N\ R_p^2/4\pi$$

$$\text{for } \omega_p - \pi/N < \omega \leq \omega_p + \pi/N$$

and

$$I(\omega) = N\ a_{N/2}^2/\pi$$

$$\text{for } \pi(N-1) < \omega \leq \pi$$

It can be seen that the surface under the plot of $I(\omega)$ in the range $\omega_p \pm \pi/N$ equals

$$\frac{N\,R_p^2}{4\,\pi}\,\frac{2\,\pi}{N} = \frac{R_p^2}{2}$$

The periodogram is an unbiased estimator of the spectrum, but it is not consistent: its variance does not decrease as N becomes infinitely large, and neighboring values of $I(\omega)$ are asymptotically uncorrelated. Several methods are devised to make the estimate of the spectrum more consistent. In essence, these methods revert to some form of smoothing of the periodogram.

Smoothing may be performed directly on the periodogram. This possibility has become popular since the development of the Fast Fourier Transform algorithm. Calculation of the periodogram with the "classical" Fourier transform takes about N^2 operations. This is drastically reduced with FFT, to 2 $\log_2(N)$. Therefore, calculation of the periodogram has become a reasonable possibility, and for long series ($N > 10^4$) it is the only feasible method.

In fact, all methods for the calculation of the spectrum are mathematically equivalent to a smoothing of the periodogram. In this smoothing operation a compromise should always be reached: smoothing over a broad range reduces the variance of the spectrum estimates, but increases the "bandwidth": a peak in the spectrum is spread over a rather broad range of frequencies. Inversely, reducing the bandwidth increases the variance, and in general: bandwidth x variance = constant.

Spectral analysis can be extended to the study of bivariate processes, where two simultaneous time series are studied.

The cross-spectrum is the finite Fourier transform of the cross-covariance function. However, in contrast to the case of a single time series, this cross-spectrum is complex. In order to study the cross-spectrum, several functions are defined from it. These express the real and imaginary parts, and various combinations thereof. The subject will not be treated in this text. For an introduction and useful references, see Chatfield (1976).

MESA

Maximum Entropy Spectral Analysis is a recently developed method for spectral analysis, which has been especially designed for the analysis of short series. An excellent account of the method and an annotated program listing are provided by Kirk et al. (1979). Burg (1967) has proposed the method on the basis of the following reasoning. In using windows to estimate the spectrum, one makes implicit assumptions on the unavailable data, i.e., the data of the time series before the observations have started, and after they have ended. Given a set of autocorrelation values, with the condition that the spectrum be non-negative definite, there usually exist infinitely many power spectra which are consistent with the given data. MESA selects among these the most random spectrum, i.e., the spectrum with the maximum entropy.

Equivalently, one can say that the autocorrelation values are extrapolated beyond the length M ($<= N$) in the most random way. The method thus corresponds to making the least stringent assumptions possible on the unavailable data.

It can be shown (see Kirk et al., 1979, for a summary) that this method is equivalent to the estimation of the spectrum from the least squares fitting of autoregressive (AR) model of order M to the data. The spectrum is then directly estimated from the coefficients of the AR model, as are the extrapolated autocorrelation values.

MESA is especially suited for short time series. It provides a better resolution in the low frequency range, does not produce sidebands in the spectrum, and can predict periods in the same order as the length of the time series (Kirk et al., 1979). Major drawbacks of the method are the computational effort required, and especially the problems in choosing an appropriate order M of the AR filter. This problem is as yet unsolved. MESA is therefore not very well suited for long series. However, for short series, as are almost all ecological time series, it is a powerful method. The method has been applied in meiobenthic research by Herman and Heip (1984) and Heip and Herman (1985) for the spectral analysis of 7-year series in a copepod community.

Which Method to Choose?

It is difficult to guide the choice of an analytical method, as it depends on a large number of factors. For the analysis of short, irregularly spaced series a simple analysis of trend and seasonal components may suffice (see Herman and Heip, 1986). When longer, more or less equispaced data series are available, the choice becomes more complicated. For the aim of prediction, statistical (ARIMA) models usually are superior to spectral analysis. Ecological series often show pseudoperiodicity: swings in the data are more or less regular, but contain phase shifts and changes in amplitude. Extending "harmonic components" into the future for the purpose of prediction, cannot take such features into account.

On the other hand, spectral analysis does reveal on which time scales the most important variability is to be found in the series. This in itself is an important structural feature of the series, which can yield scientific insight into the structure and functioning of the system.

The specific method chosen for spectral analysis will depend on the length of the time series (MESA may be superior for shorter series, FFT is the only feasible method for very long series). Often it will suffice to plot a periodogram without pursuing the analysis further, e.g., if it is the purpose to show the influence of tides, seasonal influences, etc.

The computer programs available are also an important factor to take into account. It is impossible to perform any of these analyses without a computer. It is also advisable to use standard software packages.

Acknowledgments

The authors acknowledge the relevant criticism of two anonymous reviewers. The three authors were supported through grants from the Belgian Fund for Scientific Research (NFWO) an through grant FKFO 32.9007.82 from the Belgian Fund for Collective Fundamental Research.

References

Allen, J.D.
1975. Components of Diversity. *Oecologia,* 18:359-367.

Anderson, G.D., E. Cohen, W. Gazdzic, and B. Robinson
1978. *User's Pocket Guide to SIR with Sections on SPSS and BMDP.* 108 pages. Evanston, Illinois: S.I.R., Incorporated.

Anscombe, F.J.
1950. Sampling Theory of the Negative Binomial and Logarithmic Series Distributions. *Biometrika,* 37:358-382.

Blackith, R.E., and R.A. Reyment
1971. *Multivariate Morphometrics.* 412 pages. London: Academic Press.

Boswell, M.T., and G.P. Patil
1971. Chance Mechanisms Generating the Logarithmic Series Distribution Used in the Analysis of Number of Species and Individuals. Pages 100-130 in G.P. Patil, E.C. Pielou, and W.E. Waters, editors. *Statistical Ecology.* Volume I. University Park: Pennsylvania State University Press.

Box, G.E.P., and G.M. Jenkins
1976. *Time-series Analysis, Forecasting and Control.* 553 pages. San Francisco: Holden-Day.

Bulmer, M.G.
1974. On Fitting the Poisson Lognormal Distribution to Species-abundance Data. *Biometrics,* 30:101-110.

Burg, J.P.
1967. *Maximum Entropy Spectral Analysis.* Paper presented at the 37th Annual International Meeting, Society of Exploratory Geophysicists, Oklahoma.

Caswell, H.
1976. Community Structure: A Neutral Model Analysis. *Ecological Monographs,* 46:327-354.

Chatfield, C.
1976. *The Analysis of Time Series: Theory and Practice.* 263 pages. London: Chapman and Hall.

Cliff, A.D., and J.K. Ord
1973. *Spatial Autocorrelation.* 178 pages. London: Pion.

Cohen, J.E.
1968. Alternate Derivations of a Species-abundance Relation. *American Naturalist,* 102:165-172.

Coull, B.C., and J.W. Fleeger
1977. Long-term Temporal Variation and Community Dynamics of Meiobenthic Copepods. *Ecology,* 58:1136-1143.

Daget, J.
1976. *Les Modeles Mathematiques en Ecologie.* 172 pages. Paris: Masson.

Date, C.J.
1981. *An Introduction to Database Systems.* Third edition. 574 pages. Reading, Massachusetts: Addison-Wesley Publishing Company.

Davies, R.G.
1971. *Computer Programming in Quantitative Biology.* 492 pages. London: Academic Press.

Dixon, W.J., editor
1973. *BMD Biomedical Computer Programs.* 773 pages. Los Angeles: University of Calififornia Press.

Dixon, W.J., and M.B. Brown, editors
1977. *BMDP-77 Biomedical Computer Programs. P-series.* 880 pages. Berkeley: University California Press.

Findlay, S.E.G.
1982. Influence of Sampling Scale on the Apparent Distribution of Meiofauna on a Sand-flat. *Estuaries,* 5:322-324.

Fisher, R.A., A.S. Corbet, and C.B. Williams
1943. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology,* 12:42-58.

Frontier, S., editor
1983. *Strategies d'Echantillonnage en Ecologie.* 494 pages. Paris: Masson.
1985. Diversity and Structure of Aquatic Ecosystems. *Oceanography and Marine Biology Annual Review,* 23:253-312.

Gauch, H.G., Jr.
1977. *ORDIFLEX - A Flexible Computer Program for Four Ordination Techniques: Weighted Averages, Polar Ordination, Principal Components Analysis, and Reciprocal Averaging.* 185 pages. New York: Ecology and Systematics, Cornell University.
1979. *COMPCLUS - A FORTRAN Program for Rapid Initial Clustering of Large Data Sets.* 59 pages. New York: Ecology and Systematics, Cornell University.
1984. *Multivariate Analysis in Community Ecology.* 298 pages. Cambridge: Cambridge University Press.

Gower, J.C.
1966. Some Distance Properties of Latent Roots and Vector Methods Used in Multivariate Analysis. *Biometrika,* 53:325-338.

Gray, J.S.
1983. Use and Misuse of the Log-normal Plotting Method for Detection of Effects of Pollution - A Reply to Shaw et al. (1983). *Marine Ecology Progress Series,* 11:203-204.

Greig-Smith, P.
1983. *Quantitative Plant Ecology.* 359 pages. Oxford: Blackwell Scientific Publications.

Heip, C.
1974. A New Index Measuring Evenness. *Journal of the Marine Biological Association of the United Kingdom,* 54:555-557.
1975. On the Significance of Aggregation in Benthic Marine Invertebrates. Pages 527-538 in H. Barnes, editor, *Proceedings of the 9th European Marine Biology Symposium.* Aberdeen: Aberdeen University Press.
1976. The Spatial Pattern of *Cyprideis torosa* (Jones, 1850) (Crustacea, Ostracoda). *Journal of the Marine Biological Association of the United Kingdom,* 56:179-189.

Heip, C., and P. Engels
1974. Comparing Species Diversity and Evenness Indices. *Journal of the Marine Biological Association of the United Kingdom,* 54:559-563.
1977. Spatial Segregation in Copepod Species from a Brackish Water Habitat. *Journal of Experimental Marine Biology and Ecology,* 26:77-96.

Heip, C., and P.M.J. Herman
1985. The Stability of A Benthic Copepod Community. Pages 255-263 in P.E. Gibbs, editor, *Proceedings of the 19th European Marine Biology Symposium.* Cambridge: Cambridge University Press.

Heip, C., R. Herman, and M. Vincx

1984. Variability and Productivity of Meiobenthos in the Southern Bight of the North Sea. *Rapports et Proces-verbaux des Reunions. Conseil Permanent International pour l'Exploration de la Mer,* 183:51-56.

Heip, C., M. Vincx, and G. Vranken

1985. The Ecology of Marine Nematodes. *Oceanography and Marine Biology Annual Review,* 23:399-490.

Herman, P.M.J., and C. Heip

1984. Long-term Dynamics of Meiobenthic Populations. *Oceanologica Acta. Proceedings of the 17th European Marine Biology Symposium:* pages 109-112.

1986. The Predictability of Biological Populations and Communities: An Example from the Meiobenthos. *Hydrobiologia,* 142:281-290.

Hicks, G.R.F., and B.C. Coull

1983. The Ecology of Marine Meiobenthic Harpacticoid Copepods. *Oceanography and Marine Biology Annual Review,* 21:67-175.

Hill, M.O.

1973. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology,* 54:427-432.

1974. Correspondence Analysis: A Neglected Multivariate Method. *Applied Statistics,* 23:340-354.

1979a. *DECORANA - A FORTRAN Program for Detrended Correspondence Analysis and Reciprocal Averaging.* 52 pages. New York: Ecology and Systematics, Cornell University.

1979b. *TWINSPAN - A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-Way Table by Classification of the Individuals and Attributes.* 90 pages. New York: Ecology and Systematics, Cornell University.

Hogue, E.W.

1982. Sediment Disturbance and the Spatial Distributions of Shallow Water Meiobenthic Nematodes on the Open Oregon Coast. *Journal of Marine Research,* 40:551-573.

Hogue, E.W., and C.B. Miller

1981. Effects of Sediment Microtopography on Small-scale Spatial Distribution of Meiobenthic Nematodes. *Journal of Experimental Marine Biology and Ecology,* 53:181-191.

Hurlbert, S.H.

1971. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology,* 52:577-586.

Hutcheson, K.

1970. A Test for Comparing Diversities Based on the Shannon Formula. *Journal of Theoretical Biology,* 29:151-154.

Jumars, P.A., D. Thistle, and M.L. Jones

1977. Detecting Two-Dimensional Spatial Structure in Biological Data. *Oecologia (Berlin),* 28:109-123.

Jumars, P., and J.E. Eckman

1983. Spatial Structure within Deep-sea Benthic Communities. Pages 399-451 in G.T. Rowe, editor, *Deep-Sea Biology.* New York: John Wiley and Sons.

Keller, A.

1987. Modeling and Forecasting Primary Production Rates Using Box-Jenkins Transfer Function Models. *Canadian Journal of Fisheries and Aquatic Sciences,* 44:1045-1052.

Kempton, R.A., and L.R. Taylor

1974. Log-series and Log-normal Parameters as Diversity Discriminants for the Lepidoptera. *Journal of Animal Ecology,* 43:381-399.

1976. Models and Statistics for Species Diversity. *Nature* 262:818-820.

Kempton, R.A., and R.W.M. Wedderburn

1978. A Comparison of Three Measures of Species Diversity. *Biometrics,* 34:25-37.

Kendall, M.G.

1976. *Time-series.* 2nd edition. 197 pages. London: Charles Griffin and Company.

Kendall, M.G., and A. Stuart

1968. *Advanced Theory of Statistics.* Volume 3, 2nd edition. 552 pages. London: Charles Griffin and Company.

Kirk, B.L., B.W. Rust, and W. Van Winkle

1979. *Time Series Analysis by The Maximum Entropy Method.* ORNL-5332. 218 pages. Oak Ridge, Tennessee: Oak Ridge National Laboratory.

Krishna Iyer, P.V.

1949. The First and Second Moments of Some Probability Distributions Arising from Points on a Lattice and Their Application. *Biometrika,* 36:135-141.

Kruskal, J.B.

1964. Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika,* 29:115-129.

Lambshead, P.J.D., H.M. Platt, and K.M. Shaw

1983. The Detection of Differences Among Assemblages of Marine Benthic Species Based on an Assessment of Dominance and Diversity. *Journal of Natural History,* 17:859-874.

Lance, G.N., and W.T. Williams

1966. A Generalized Sorting Strategy for Computer Classifications. *Nature,* 212:218.

1967. A General Theory of Classificatory Sorting Strategies. II. Clustering Systems. *Computer Journal,* 10:271-277.

Legendre, L., and P. Legendre

1986. *Ecologie Numerique.* Tome 1, 192 pages; Tome 2, 247 pages. Quebec: Masson, Les presses de l'Universite du Quebec, 2eme edition.

Lloyd, M.

1967. Mean Crowding. *Journal of Animal Ecology,* 36:1-30.

MacArthur, R.H.

1957. On the Relative Abundance of Bird Species. *Proceedings of the National Academy of Science of the United States of America, Washington,* 43:293-295.

May, R.M.

1975. Patterns of Species Abundance and Diversity. Pages 81-120 in M.L. Cody and J.M. Diamond, editors, *Ecology and Evolution of Communities.* Cambridge, Massachusetts: Belknap Press.

May, R.M., and G.F. Oster

1976. Bifurcations and Dynamical Complexity in Simple Ecological Models. *American Naturalist,* 110:573-599.

Motomura, I.

1932. Statistical Study of Population Ecology. [In Japanese.] *Doobutugaku Zassi,* 44:379-383.

Neyman, J.

1939. On a New Class of Contagious Distributions, Applicable in Entomology and Bacteriology. *Annals of Mathematical Statistics,* 10:35-57.

Nie, N.H., C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Bent

1975. *-SPSS- Statistical Package for The Social Sciences.* Second edition. 675 pages. New York: McGraw-Hill Book Company.

Noy-Meir, I.

1973. Data Transformations in Ecological Ordination. I. Some Advantages of Non-centering. *Journal of Ecology,* 61:329-341.

Orloci, L.

1967. An Agglomerative Method for Classification of Plant Communities. *Journal of Ecology,* 55:193-205.

1978. *Multivariate Analysis in Vegetation Research.* 451 pages. The Hague: Junk.

Osterdahl, L., and G. Zetterberg

1981. *Rubin-Species Codes and Species Numbers.* Stockholm: National Swedish Environment Protection Board Report 1427.

Peet, R.K.

1974. The Measurement of Species Diversity. *Annual Review of Ecology and Systematics,* 5:285-307.

Pielou, E.C.

1969. *An Introduction to Mathematical Ecology.* 286 pages. New York: John Wiley and Sons.

1975. *Ecological Diversity.* 165 pages. New York: John Wiley and Sons.

1981. The Broken-stick Model: A Common Misunderstanding. *American Naturalist,* 117:609-610.

1984. *The Interpretation of Ecological Data.* 263 pages. New York: John Wiley and Sons.

Poole, R.W.
1978. The Statistical Prediction of Population Fluctuations. *Annual Review of Ecology and Systematics,* 9:427-448.

Preston, F.W.
1948. The Commonness and Rarity of Species. *Ecology,* 29:254-283.

Robinson, B., G.D. Anderson, E. Cohen, and W. Gazdzic
1979. *S.I.R., Scientific Information Retrieval.* 474 pages. Evanston, Illinois: SIR, Incorporated.

Rohlf, F.J., J. Kispaugh, and D. Kirk
1972. *NT-SYS Numerical Taxonomy System of Multivariate Statistical Programs.* Stony Brook, N.Y.: State University of New York.

Ross, G.J.S., F.B. Lauckner, and D. Hawkings
1976. *CLASP Classification Program.* Harpenden, England: Rothampsted Experimental Station.

Routledge, R.D.
1979. Diversity Indices: Which Ones Are Admissible? *Journal of Theoretical Biology,* 76:503-515.

Sanders, H.L.
1968. Marine Benthic Diversity: A Comparative Study. *American Naturalist,* 102:243-282.

SAS Institute, Incorporated
1985. *SAS User's Guide: Statistics.* 584 pages. Cary, North Carolina: SAS Institute, Incorporated.

Shaw, K.M., P.J.D. Lambshead, and H.M. Platt
1983. Detection of Pollution - Induced Disturbance in Marine Benthic Assemblages with Special Reference to Nematodes. *Marine Ecology Progress Series,* 11:195-202.

Simberloff, D.
1972. Properties of the Rarefaction Diversity Measurement. *American Naturalist,* 106:414-418.

Singer, S.B.
1980. *DATAEDIT - A FORTRAN Program for Editing Data Matrices.* 42 pages. New York: Ecology and Systematics, Cornell University.

Singer, S.B., and H.G. Gauch, Jr.
1979. *CONDENSE - Convert Data Matrix from Any Ordiflex Format into A Condensed Format by Samples.* 7 pages. New York: Ecology and Systematics, Cornell University.

Taylor, L.R.
1961. Aggregation, Variance and the Mean. *Nature,* 189:732-735.

Taylor, L.R., R.A. Kempton, and I.P. Woiwod
1976. Diversity Statistics and the Log-series Model. *Journal of Animal Ecology,* 45:255-272.

Thomas, M.
1949. A Generalization of Poisson's Binomial Limit for Use in Ecology. *Biometrika,* 36:18-25.

Webb, D.J.
1974. The Statistics of Relative Abundance and Diversity. *Journal of Theoretical Biology,* 43:277-291.

Whittaker, R.H.
1972. Evolution and Measurement of Species Diversity. *Taxon,* 21:213-251.

Wiser, W.
1953. Die Beziehung zwischen Mundhoehlengestalt, Ernaehrungsweise und Vorkommen bei freilebenden marinen Nematoden. *Arkiv för Zoologie,* 4:439-484.

Williams, C.B.
1964. *Patterns in the Balance of Nature.* 324 pages. New York: Academic Press.

Williams, W.T., and M.B. Dale
1965. Fundamental Problems in Numerical Taxonomy. *Advances in Botanical Research,* 2:35-68.

Williams, W.T., and P. Gillard
1971. Pattern Analysis of a Grazing Experiment. Australian Journal of Agricultural Research, 22:245-260.

Williams, W.T., and J.M. Lambert
1959. Multivariate Methods in Plant Ecology. I. Association-analysis in Plant Communities. *Journal of Ecology,* 47:83-101.

Wishart, D.
1978. *CLUSTAN Users Manual.* 3rd edition. Edinburgh: Edinburgh University.