# UTRECHT MICROPALEONTOLOGICAL BULLETINS

M. M. DROOGER


Project no. 1

QUANTITATIVE RANGE CHART ANALYSES

26

# UTRECHT MICROPALEONTOLOGICAL BULLETINS

Editor C. W. Drooger
Department of Stratigraphy and Paleontology
State University of Utrecht
Budapestlaan 4, Postbus 80.021
3508 TA Utrecht, Netherlands

In the series have been published:

Bull. 1. T. FREUDENTHAL – Stratigraphy of Neogene deposits in the Khania Province, Crete, with special reference to foraminifera of the family Planorbulinidae and the genus *Heterostegina*. 208 p., 15 pl., 33 figs. (1969) *f* 32,–

Bull. 2. J. E. MEULENKAMP – Stratigraphy of Neogene deposits in the Rethymnon Province, Crete, with special reference to the phylogeny of uniserial *Uvigerina* from the Mediterranean region. 172 p., 6 pl., 53 figs. (1969) *f* 29,–

Bull. 3. J. G. VERDENIUS – Neogene stratigraphy of the Western Guadalquivir basin, S. Spain. 109 p., 9 pl., 12 figs. (1970) *f* 28,–

Bull. 4. R. C. TJALSMA – Stratigraphy and foraminifera of the Neogene of the Eastern Guadalquivir basin, S. Spain. 161 p., 16 pl., 28 figs. (1971) *f* 44,–

Bull. 5. C. W. DROOGER, P. MARKS, A. PAPP et al. – Smaller radiate *Nummulites* of northwestern Europe. 137 p., 5 pl., 50 figs. (1971) *f* 37,–

Bull. 6. W. SISSINGH – Late Cenozoic Ostracoda of the South Aegean Island arc. 187 p., 12 pl., 44 figs. (1972) *f* 57,–

Bull. 7. author's edition. F. M. GRADSTEIN – Mediterranean Pliocene *Globorotalia*, a biometrical approach. 128 p., 8 pl., 44 figs. (1974) *f* 39,–

Bull. 8. J. A. BROEKMAN – Sedimentation and paleoecology of Pliocene lagoonal-shallow marine deposits on the island of Rhodos (Greece). 148 p., 7 pl., 9 figs. (1974) *f* 47,–

Bull. 9. D. S. N. RAJU – Study of Indian Miogypsinidae. 148 p., 8 pl., 39 figs. (1974) *f* 38,–

Bull. 10. W. A. VAN WAMEL – Conodont biostratigraphy of the Upper Cambrian and Lower Ordovician of north-western Öland, south-eastern Sweden. 128 p., 8 pl., 25 figs. (1974) *f* 40,–

Bull. 11. W. J. ZACHARIASSE – Planktonic foraminiferal biostratigraphy of the Late Neogene of Crete (Greece). 171 p., 17 pl., 23 figs. (1975) *f* 52,–

Bull. 12. J. T. VAN GORSEL – Evolutionary trends and stratigraphic significance of the Late Cretaceous *Helicorbitoides-Lepidorbitoides* lineage. 100 p., 15 pl., 14 figs. (1975) *f* 37,–

Bull. 13. E. F. J. DE MULDER – Microfauna and sedimentary-tectonic history of the Oligo-Miocene of the Ionian Islands and western Epirus (Greece). 140 p., 4 pl., 47 figs. (1975) *f* 45,–

Bull. 14. R. T. E. SCHÜTTENHELM – History and modes of Miocene carbonate deposition in the interior of the Piedmont Basin, NW Italy. 208 p., 5 pl., 54 figs. (1976) *f* 56,–

(continued on back cover)

# QUANTITATIVE RANGE CHART ANALYSES

I.G.C.P. Project no. 1

M. M. DROOGER

Printed in the Netherlands by Loonzetterij Abé, Hoogeveen
15 juni 1982

CONTENTS

3 plates, 32 figures

de taal der wiskunde is alleen maar een gebrekkig hulpmiddel voor de menschen om wiskunde aan elkaar mee te delen . . . het projecteeren van wiskundige systemen op de ervaring is een vrije daad.

L. E. J. Brouwer, 1907, pp. 169, 179

## PREFACE

Max Drooger died suddenly on November 27, 1981, at the age of 31, as a consequence of pneumonia and medical negligence.

Chapters I – X of the present bulletin are based on a manuscript which the author had not finally checked for possible incorrect statements or duplications. We have decided to send the manuscript to the printer's as it was.

At some places we have left in the text subchapters pertaining to the computer programs. These programs are available on request from the Department of Stratigraphy and Paleontology of the Utrecht State University, *not* from the sales office or the editor.

We have added two more chapters which formed part of an earlier version of the manuscript. The original position of chapter XI was between I and II, that of chapter XII between VII and VIII. The text of chapter XIII on diversity estimates existed as a separate manuscript.

G. J. Leppink and G. J. van der Zwaan have assisted in finalizing parts of the text, Miss S. M. McNab has made linguistic improvements. A. van Doorn made the drawings and W. A. den Hartog composed the three plates.

We gratefully acknowledge the financial support given by the Netherlands Organization for the Advancement of Pure Science, Z.W.O. (The Hague) within the framework of the "Accuracy in time" project 75–94.

# Chapter I

# INTRODUCTION

## I.1. Distribution charts and range charts

Samples from sediments frequently yield a wealth of microfossils. Because of their relative abundance these fossils have become the primary tool for biostratigraphic and paleoecological analyses carried out by micropaleontologists, both in industry and in academic institutions. Micropaleontologists usually consider only a single group of microfossils, although there are teams of specialists who study different groups from the same samples. These groups occupy widely different positions in the animal and plant kingdoms and include benthonic foraminifera, planktonic foraminifera, radiolarians, ostracodes, conodonts, calcareous nannofossils, dinoflagellates and diatoms (Haq & Boersma, 1978), and spores and pollen. This list is not exhaustive but it accounts for roughly 90 per cent of all micropaleontologists' activities.

The systematic treatment applied to all these groups is the same, or at least it can be the same. Although preparation techniques differ, they all yield per sample a large collection of fossils, which are grouped and labelled by means of some sets of basic rules. Groupings are usually made on the basis of morphological similarity and the units are equated with the taxa distinguished for Recent organisms, whether species or genera, or units of lower or higher rank. Labelling commonly follows the rules of Linnean nomenclature. However, there are other systems of grouping and labelling as well.

For our purpose it is important that a sample contains a collection of taxa, the names of which are understood (at least approximately) by other specialists in the same field of research. The collection of taxa (or names) is thought to give information about either the relative geological age or the environment (or both) of the sediment from which the sample was derived.

In practice the collections of taxa from two or more samples are compared for a specific purpose. For instance if all samples are from Recent sediments of a certain area, i.e. from a single time plane, the aim of the analysis is usually to draw inferences about environment.

Samples and taxa may be represented in a kind of matrix, called a distribution chart. By establishing degrees of similarity in sample contents the specialist is able to draw conclusions. It should be emphasized that there is

no predetermined order of the taxa for such a matrix, and this is not necessary for the samples either. One might postulate that some geographic grid ought to influence the sample order, but in practice such an order is not imposed because the two-dimensional grid cannot be transformed logically into the linear scale of the available axis. As the analysis is expected to yield amongst other things a meaningful geographic pattern of the sample groupings, generally no order is imposed beforehand.

The matrix cannot be freely arranged for the samples in the case of stratigraphic sections. One cannot escape from the logical order of the samples along the stratigraphical column. The taxa arrangement along one axis remains free, along the other axis the samples will be arranged according to the succession of the strata from bottom to top (or top to bottom), i.e. according to relative age. This type of distribution chart is usually called a range chart.

The information drawn from such range charts is usually for stratigraphic purposes; the specialist looks for entries, exits and ranges of taxa from which he constructs datum levels and zones in the biostratigraphic sequence.

If certain precautions are taken a range chart can be used as if it were a distribution chart, i.e. the time sequence is "cut off" and the data in the matrix can be analyzed for paleoecological purposes. Usually the proportions of the taxa will be checked for correlation with other data from the sediment, whether recurrent along the column or not.

There are surprisingly few generally accepted rules relating to the methods by which charts should be constructed. Consequently very few charts are of good quality from the point of view of mathematical statistics. The range charts given in the literature commonly contain no more than presence-absence data of the taxa, and frequently the ranges are simplified by solid lines between the entry and exit levels, suggesting that the taxon was found in all samples in between. Biostratigraphers frequently seem to be concerned only about the few items they are interested in and carelessly throw away all other information they have gathered during their observations.

At least as far as foraminifera are concerned, the micropaleontologists analyzing Recent environments tend to give better numerical information, but their methods differ considerably. Unit weight or unit volume of sediment give markedly different totals of individuals, which reveal little more than the degree of "dilution" by sediment. Since relative frequencies of the taxa are thought to give the most relevant data we frequently find percentages either without a total number or with a calculated total that has no relation with the number of specimens from which the percentages were calculated (Phleger & Parker, 1951).

Instead of percentages, the range charts often show only the degree of abundance of taxa in samples by means of symbols in the cells of the charts. For instance, a taxon with a numerical proportion of less than one per cent in a sample is marked as · in the corresponding cell. A proportion between 1 and 10 per cent may be recorded as ○, and a proportion greater than 10 per cent as ●. In the case of absence the cell is left blank.

If these "semi-quantitative" data come from quick inspections of the trays or slides, an enormous amount of subjectivity is introduced, especially if the observer expects to see something in advance and adjusts his observations to what he expects.

In our opinion the only correct way to make reliable numerical estimates of taxa frequencies in assemblages is to scan and determine "randomly" not too small a number of specimens from the tray or slide, classify them one by one and count them per taxon. The procedure that has been followed in Utrecht for many years is to count a fixed total number (in general 200) of specimens per sample and to record the score for each taxon in each count in the range chart. Such fixed number counts were first used by C. W. Drooger and Kaasschieter (1958) for foraminifera on the Orinoco shelf.

Assuming that the multinomial model is valid for such counts, one can easily calculate the standard error of the proportion of every taxon in some sample in order to establish the reliability of that proportion estimate (Zachariasse et al., 1978). These fixed number counts are however unsatisfactory for estimating numerical proportions of rare taxa because the number of specimens to be considered is too small. In Zachariasse et al. (1978) the present author outlined a procedure of logarithm-transformed estimates of proportions, which can be used for rare taxa, and he has given formulae for calculating standard errors of such estimates (1978, p. 38–45). These logarithmic estimates will be described with a more ample discussion in chapter XI of the present bulletin.

## I.2. Scope of the investigation

The aim of I.G.C.P. project no. 1, "Accuracy in time" is to give us a better understanding of the tools of the stratigrapher. One of the problems in this context is the (in)accuracy range of datum levels (e.g. corresponding to an entry) and a special part of this problem is the numerical expression or expectation for such an entry. This problem was soon found to be part of a much wider subject: the most correct interpretation of numerical data in range charts. The statistical basis of such numerical analyses was found to be rather poor, and various multivariate techniques applied in the recent

literature seem to lead to dubious results. As a consequence checking and improving statistical range chart analysis has become an almost continuous task during the last few years. This paper concentrates on the problems of range chart analysis, and may be applicable primarily in paleoecology and ecostratigraphy. Other problems tackled within the scope of the Accuracy project have been published (or will be published) elsewhere.

### I.3. Basic assumptions

Before starting our discussion of mathematical procedures and techniques, we need to review the axioms and specific characteristics pertaining to the items we are dealing with. These may give an impression of the problems we were facing, and – we hope – make it easier to understand why we chose particular procedures to find our "best" solutions.

We always compare the numerical data of faunal or floral associations of a single group of microfossils (e.g. benthonic foraminifera) at one time. Every taxon of such a group is considered to contain equal individuals. This means that differences in morphology due to evolution or to ecology are neglected, unless these differences have led to the classification of the individuals in a separate taxon, i.e. when they are given another name.

Each taxon is assumed to have a set of environmental requirements under which it functions optimally. In each spot or area with such optimum conditions the taxon is expected to have its highest "absolute" frequencies; these frequencies will diminish in a horizontal sense (i.e. in the same time plane) towards less favourable environments. The biotope of a taxon thus consists of an area with a frequency maximum and numerical gradients from this maximum to close to zero, situated along the margin of its distribution. The conditions under which all other taxa will flourish will be more or less different and therefore the frequency patterns will differ too.

We can postulate that the biotopes of different taxa show different degrees of overlap or remoteness. Overlap may be visualised in a geographical sense, but remoteness does not necessarily mean geographical distance. Some environmental factors like depth and energy have horizontal components which are expected to cause geographical gradients, but others such as oxygen content, nutrient availability or presence or absence of vegetation are less predictable in a horizontal sense. Their occurrence may be patchy, superimposed on those of the factors which have more distinct horizontal gradients, and thus cause local frequency peaks.

If we knew the map of absolute frequencies of all taxa of the group under

consideration for a certain time plane, it would give a good reflection of the sums of all environmental parameters. Even if we had only a restricted knowledge of the environmental factors and their gradients, we would accept the frequency map as a correct image of environment. For each spot on the map the sum of the total numbers of the taxa gives the parameters of what is called the biofacies.

With regard to the Recent oceans our knowledge of environmental parameters at the bottom is rather scanty; about some parameters we know next to nothing. And there are very few data on the frequencies of benthonic taxa available. As a consequence the frequency combination of taxa in the Recent oceans and seas is a poor basis for the interpretation of fossil faunas in terms of paleoenvironments. For the Recent forms we have only some vague ideas about a few environmental requirements for a restricted number of the taxa; and these ideas are hardly ever based on quantitative observations.

As soon as we start to deal with fossils we lose even more control. Environmental factors cannot be measured any more. Inorganic parameters of the sediment may lead to a number of fairly subjective interpretations. And the "absolute" numbers of individual taxa and of entire "populations" become rather elusive.

With regard to these total numbers it should be emphasized that there are several restrictions. The geological samples have an average thickness of five cm (and a width of some 10 cm). If an average sedimentation rate of five cm per thousand years is supposed and if the life time of one generation is taken to be of the order of one year, then the absolute number of the taxon in the sample will reflect the sum of at least one thousand generations, buried in an area of some 50–100 $cm^2$. The paleontologist's "population" evidently is an artefact relative to the biological population; it is too narrow in a horizontal sense, much too large in a vertical sense.

For statistics this population is acceptable but its value is restricted. We know that sampling in the same layer at different spots may yield associations of distinctly different composition (Zachariasse et al., 1978) firstly because of the horizontal shift in environment and secondly because it is impossible to select exactly the same time interval at different places of a layer. It can be said that the association from each geological sample is unique. It cannot be copied from another sample; various degrees of similarity is all that we can expect (M. M. Drooger & C. W. Drooger, 1979).

Furthermore, in the comparison of different geological samples – which is our ultimate aim – absolute numbers appear to fluctuate considerably.

Differences may amount to several orders of magnitude because of factors which have nothing to do with the optimum conditions of life, such as the amount of dilution by inorganic sediment.

Caution is called for if one tries to use the large differences in total numbers in a statistical treatment. This means that the approximation of the faunal composition by means of relative numbers in a random count is of greater interest for our interpretation than the ill-defined absolute numbers.

The relative composition of the taxa in a statistical sample taken from a geological sample from some sediment layer in itself gives little information about environment. We may be able to make no more than a general paleoecological assignment, because our compositional data on Recent faunas are too poor, and because taxa need not have had the same environmental requirements throughout their geological life time (Van der Zwaan, 1982). And the further back we go in time the larger is the number of taxa we encounter, which have no Recent representatives.

When the geologist adds the dimension time by studying more than one sample from a stratigraphic section, he introduces yet another unknown factor. We can easily imagine that the successive sediments along a specific stratigraphic column were deposited in continuously changing environments, changing either unidirectionally or more often according to some fluctuation pattern. In addition we accept the result of yet another axiom; that evolution – whatever that may be – may have changed the composition by adding new taxa and dropping others. It is hard to distinguish between the roles played by environment and evolution in stratigraphic sections that correspond to no more than a few millions of years (this is true for most continuous sections). Commonly the effect of evolution will be negligible for the greater part of the faunas and floras.

However, the addition of the time factor improves the interpretation provided we eximine the same stratigraphic section or the same geographical area. Although we are still unable to pinpoint a single association on the "environment-map" with any accuracy, the comparison of associations, successive in time, gives us the opportunity to interpret differences in composition as relative points on one or more environmental gradients. Recognition of the type of change is often more important for the geologist-stratigrapher than an accurate paleoenvironment assignment to a single layer. We may arrive at conclusions concerning environmental changes in a vertical sense along the column, and these conclusions can then be used in stratigraphic correlations, i.e. in comparison of stratigraphic sections at different places in a region. We may ultimately arrive at a reconstruction of the region's history.

In other words, what a single sample could not tell us, we can deduce from a series of samples. We may recognize regular numerical combinations of taxa throughout the chart. Their behaviour along the column relative to other combinations is thought to give the best information for the environmental interpretation, and is likely to reveal much more than the trends of single taxa.

There are three successive stages on the way to the final interpretation of range charts (and distribution charts).

In the first step the geologist-micropaleontologist will gather the numerical data from which he will construct the samples-taxa matrix. A record of non-faunal/floral parameters is useful for the final phase, but is not absolutely necessary. A visual appreciation of the chart may already give suggestions about possible trends and correlations. It certainly gives the clues for biozonation and age assignment.

In the second stage appropriate statistical procedures must be applied to the data matrix to sort out meaningful signals of correlations and trends from the statistical point of view, if these exist. It is this second stage which is the main subject of this paper.

In the final phase the signals are interpreted in terms of environmental changes, evolution and stratigraphic correlation. All these conclusions remain subjective; it should be emphasized that none of them has really been proved in the second phase.

So far there has been very little statistical treatment of numerical range chart data in micropaleontology and stratigraphy. Certain aspects of the above review of axioms and specific characteristics determined which procedures we chose to follow or construct to arrive at our "best" solutions.

In considering a single stratigraphic section we assume that the faunal/floral succession it contains is sufficiently well reflected in our fairly equally spaced geological samples. These samples commonly give data on about 1/10 to 1/50 of the complete sediment sequence. Some detailed analyses have shown that this order of sample cover of the sections tends to simplify oscillation patterns, but the main changes along the column are sufficiently well reflected. Sudden large jumps in composition are unlikely to be missed, because accompanying sedimentary phenomena will have been recognized during the field survey and the stratigrapher will generally have adapted his sampling to such phenomena.

For our eventual interpretation we are interested in all frequency deviations, such as trends of single taxa, (and especially) groupings of taxa with

positive correlations, numerical characterizations of samples in terms of these taxa groups and the distribution of the taxa combinations along the column. If available, correlations between taxa and taxa groups with non-biological data of the sediments are helpful in the final phase.

Throughout this paper we discuss the independence or the interdependence of taxa. This is a shorthand notation for absence or presence of correlation between the scores of such taxa. Positive correlation does not mean that the taxa are biologically dependent on each other; it means that they have similar relations with some set of environmental conditions.

### I.4. Closed random variables concept

The major part of the statistical analyses concerns the derivation of the so-called open variables from the closed variables. Chapters II–IV are devoted to this theme. The open variables concept will be explained briefly in the following section I.5, the closed variables concept in this section.

We are dealing with a distribution chart (or range chart) consisting of a set of N counts, each count coming from a sample from the geographical area (or stratigraphic section) under consideration. A number of M taxa (or groups of taxa) is discerned according to which each specimen counted is classified. Hence, the distribution chart is an $M \times N$ matrix of scores ($x_{ij}$), $x_{ij}$ being the score (number of specimens) of taxon i in count j from sample j. The sum of the scores $x_{ij}$ over all taxa i in count j is the size $n_j$ of that count. In general a fixed number of specimens is counted from each sample j. In most cases $n_j = 200$ or 300, but from the statistical point of view there is no need for these totals $n_j$ to be equal.

Every count can be seen as a statistical sample from a statistical population. In our investigation this population cannot be defined sharply. It can be defined as the collection of individuals of the specific group of microfossils that is present in the geological sample from which the count comes. However, when microfossils are scarce the counts may be identical to the entire population, i.e. the whole geological sample was needed to complete the count. A better definition therefore seems to be that the statistical population from which count j comes is the collection of individuals present in the vicinity of the geological sample j. For range charts from stratigraphic sections one should consider only a vicinity in the "horizontal" direction within the layer from which sample j comes.

If one defines the statistical population this way one assumes that the population is homogeneous in that vicinity. This assumption may be incorrect,

however. Brolsma (in Zachariasse et al., 1978) considered five lateral samples from the same layer and could not confirm the homogeneity of the content of benthonic foraminifera over a distance of some five metres. We must conclude that the vicinity of the sample, as mentioned in the definition, may be very limited.

While each count only represents the microfauna in the sample from which it comes and the fauna in a restricted area around that sample, the set of N samples/faunas itself is a statistical sample from the microfauna in the geographical area (or in the stratigraphic section) involved.

One statistical hypothesis seems to be convenient for describing the relation between these microfaunas in the vicinities of the N samples and the microfaunas from the "total" geographical area (or stratigraphic section). This hypothesis is very simple: the fauna is homogeneous over the entire geographical area/stratigraphic section, i.e. any fauna in (the vicinity of) some sample is representative for the whole geographical area (stratigraphic section) and all counts can be seen as drawn from the same statistical population. In this paper this hypothesis is called the multinomial model. See Mosimann (1962), and our chapter II.

Before we describe the open variables concept in the following section we mention another approach proposed by Mosimann (1962). Instead of the population proportions $P_i$ of all M taxa (i = 1, 2, . . ., M) being constant all over the geographical area (or stratigraphic section), as in the above hypothesis, the population proportions $P_i$ are considered to be random variables, the so-called closed variables, because they have a "closed sum", i.e.

$$\sum_{i=1}^{M} P_i = 1$$

The random vector $P \equiv (P_1, P_2, \ldots, P_M)$ is characterized by a mean vector E(P) and by a covariance matrix C(P). The series of proportions of the taxa in the fauna of each geological sample taken from the geographical area/stratigraphic section is a realization of the vector $P = (P_1, P_2, \ldots, P_M)$.

Ignoring the counting errors for a while, the mean vector E(P) and the covariance matrix C(P) are estimated from the set of N counts. This vector E(P) and covariance matrix C(P) are used to determine the open variables, as explained in the following section.

## I.5. The open variables concept

Actually we wish to make an analysis of the relations between "real num-

bers" of individuals of the taxa. Therefore we are not interested in the covariance matrix C(P) and the related correlation matrix R(P) alone, but from the C(P) matrix we wish to draw inferences concerning the relations between the "real numbers" of individuals of the series of taxa. First we have to give an explanation of the concept "real number of individuals of a taxon i", which will be denoted as the open variable $X_i$.

The open variable $X_i$ is thought to be the number of individuals of taxon i per fixed area (of the bottom of the sea, benthonic fauna), or per fixed volume (of sea water, planktonic fauna/flora) in the biocoenosis (living community). The choice of the size of the area or volume is completely irrelevant. Of course we admit that every geological sample presents a taphocoenosis (grave community) the contents of which may be completely different from the original biocoenosis due to differential dissolution or to transport. Nevertheless such distortion is thought not to invalidate this definition.

It should be emphasized that there should be no confusion between our "real numbers" and the absolute quantities such as "number of individuals per gram of sediment" or "number of individuals per $cm^3$ of sediment". These "total numbers" are not practical because they are inversely proportional to sedimentation rate and therefore are a poor reflection of our total or real numbers in the biocoenoses.

Since we are interested in the relations between the real numbers $X_i$ of individuals per unit of area or volume of the taxa i = 1, 2, . . ., M, in the series of N counts, it is not necessary – it is not possible either – to determine the real sizes of the $X_i$ per sample. For this reason we are free to define, for each geographical area or stratigraphic section and for every microfossil group, that the mean value of T, the sum of the $X_i$ of all taxa, considering all geological samples possible, is equivalent to the convenient number one, just by choosing the appropriate unit for the fixed area or volume. Hence,

$$T = X_1 + X_2 + \ldots + X_M ; \quad E(T) = 1$$

It should be remembered that for each geological sample the population proportions $P_i$ (closed variables) add up to one;

$$P_1 + P_2 + \ldots + P_M = 1,$$

but that the hypothetical open variables $X_i$ for the individual sample need not add up to one. Hence,

$T = X_1 + X_2 + \ldots + X_M$ need not be equal to one.

Chayes and Kruskal (1966) introduced the open variables concept (in geochemistry). They state the null hypothesis that the open variables of the separate items are independent, i.e. that the covariances between $X_i$ and $X_k$ are zero for all pairs (i, k). After another introduction of the open variables concept in section II.8 we present in chapter III the procedure of Chayes and Kruskal, which we call the zero-open-covariances model, with our comments and additions. A short description of our alternative open variables model, the free-open-covariances model, is given below and elaborated in chapter IV.

The "open total" T introduced above can be seen as the factor with which the proportion $P_i$ of each taxon i in a geological sample has to be multiplied to get $X_i$,

$$X_i = T \cdot P_i$$

It is noted that T and some $P_i$ may be (and generally will be) interrelated. One can argue that one is free to substitute values at random for T for each geological sample. However, highly fluctuating values in a series of samples obviously result in a large number of significantly positive correlation coefficients $R(X_i, X_k)$. We prefer series of T-values that result in the largest possible number of correlation coefficients $R(X_i, X_k)$ near zero. Admittedly this is a subjective choice, but we believe that this choice leads to the most likely solution.

Apparently it is not efficient to find a solution by first establishing values for T for all geological samples. It is better to use the description of the collection of solutions for the covariance matrix C(X) of the open variables directly from the given covariance matrix C(P) of the closed variables (proportions). A drawback may be that a linear approximation is used to describe a solution of C(X) in terms of the known C(P). Better non-linear approximations lead to formulae that are not practical to handle.

In this linear approximation it is sufficient to add values for the variance var(T) and for the covariance $\text{cov}(T, P_i)$, $\text{cov}(T, P_2)$, . . ., $\text{cov}(T, P_M)$ to the known matrix C(P) of proportions, in order to find a solution for the matrix C(X) of open variables. Since

$$\sum_{i=1}^{M} \text{cov}(T, P_i) = 0$$

M values have to be chosen. It seems odd at first sight that the values of var(T), of $\text{var}(X_i)$ for each i and of $\text{cov}(X_i, X_k)$ for each pair (i, k), and therefore also the values of $R(X_i, X_k)$ are determined by the addition of a

set of only M numbers to the closed matrix C(P) and to the mean vector E(P).

Our intention thus is to find the series of M numbers that gives the solution for C(X) and for the related correlation coefficient matrix R(X) that has as many correlation coefficients $R(X_i, X_k)$ as possible near zero.

In our own procedure we calculate for each pair (i, k) of taxa a value for the correlation coefficient between the open variables, $R(X_i, X_k)$. This procedure seems to be more straightforward than that used by Chayes and Kruskal. According to their procedure expected values for $R(P_i, P_k)$ are calculated from the null hypothesis that the open variables of the separate taxa are mutually independent. These values are compared to the actual values for $R(P_i, P_k)$ in the chart, leading to an inference about the corresponding $R(X_i, X_k)$. Although such inferences are not correct from the point of view of mathematical statistics, we did not find that the results obtained using the procedure of Chayes and Kruskal were inferior to those obtained with our own procedure. In our experience often we were even able to make inferences concerning an $R(X_i, X_k)$ from the comparison of the value of the corresponding $R(P_i, P_k)$ in the chart to the expected value of $R(P_i, P_k)$ according to the multinomial model.

Therefore we decided to describe the multinomial model, the open variables model of Chayes and Kruskal and our own open variables model in extenso. Moreover, on the basis of the multinomial model, problems with the outside counting technique can easily be explained (section II.4), and the failure of "eliminating the closed sum character" of the proportions by means of the partial correlation technique can easily be demonstrated. The open variables model of Chayes and Kruskal is used in section III.7 to demonstrate that the proportion-ratio correlation coefficients $R(P_i, P_k/(1 - P_i))$ need not have zero as expected values under the null hypothesis that all open variables are mutually independent. We found that the technique of making inferences concerning $R(X_i, X_k)$ by comparing the values of the corresponding proportion-ratio correlation coefficients $R(P_i, P_k/(1 - P_i))$ and $R(P_k, P_i/(1 - P_k))$ to zero, gave less satisfactory results than the open variables techniques. This fact can be explained satisfactorily (see III.7).

### I.6. Time series

For range charts consisting of counts from samples derived from one stratigraphic section problems may arise which are inherent to all time series. The counts have a logical order in the range chart because their samples have

a fixed order in the stratigraphic section. This order is related to time. Because of the fixed succession of the samples, one must realise that the composition of a sample may be dependent on those of previous samples.

Consequently the definition of the random vector $P = (P_1, P_2, P_3, \ldots, P_M)$ in the previous sections needs to be modified. The vector P should be regarded as a continuous function of time t. Hence,

$$P(t) = (P_{1t}, P_{2t}, P_{3t}, \ldots, P_{Mt})$$

i.e. $P_{it}$ is the proportion in the population of taxon i at time t. Then the N samples from the stratigraphic section can be seen as N realizations of the vector P at successive "moments" $t_1, t_2, t_3, \ldots, t_j, \ldots, t_N$. The proportion in the population of taxon i in sample j is

$$P_{ij} = P_{(i,t_j)}$$

so the series $(j \mapsto P_{ij})$ is a time series for each taxon i. If these time series $(j \mapsto P_{ij})$ for several taxa i are autocorrelated (i.e. $P_{i,j+1}$ is not independent of $P_{ij}$), problems arise in the statistical interpretation of the "open" correlation coefficients $R(X_i, X_k)$. See sections V.2 and V.4-8.

Another feature is that for some taxon i the series $(j \mapsto P_{ij})$ may show a trend: the population proportions $P_{ij}$ may tend to increase with increasing j (positive trend), or may tend to decrease with increasing j (negative trend). Due to the closed sum, however, a trend in the series $(j \mapsto P_{ij})$ of the proportions of taxon i induces a trend of opposite sign in the series

$$(j \mapsto \sum_{k \neq i}^{M} P_{kj}).$$

Hence, the trends in the proportions of the taxa, if present, are interrelated, and some trend in the proportions of one taxon may be induced by a trend in the proportions of another taxon.

Therefore, in addition to these "closed" trends, we considered trends in the "real" numbers of taxa. A trend in the "real" numbers of taxon i is a trend in the series $(j \mapsto X_{ij})$ of the open variable $X_i$. "Closed" trends are interrelated due to the closed sum, whereas "open" trends do not have such an interrelation. See sections V.2 and V.9.

## I.7. Other topics

The analyses of three range charts are presented in chapters VIII, IX and X in order to illustrate all these techniques.

Since we wanted to compare the results of the methods outlined above with those of multivariate methods, a principal components analysis was also performed on each of the three range charts. A short explanation of principal components analysis and statistical considerations relating to this technique are given in chapter VI.

A simple Q-mode analysis has been applied to one range chart (chapter X) in order to acquire some insight into the relations between the samples according to the scores of the taxa they contain. This technique should enable us to discern groups of samples, each group consisting of samples that are mutually "similar" in taxa composition. The Q-mode technique applied to sets of samples is described in chapter VII.

In cases of extreme differences within a set of counts "similarity" between taxa frequencies (≈ similar biotope) may be more convenient than "correlation" between these frequencies. A series of similarity coefficients for such "heterogeneous" charts is presented in chapter XII.

## I.8. Final remarks

The range charts subjected so far to the procedures of the different models described above (charts published here, and others published elsewhere (e.g. Hageman, 1979; Tsapralis, 1981; Van der Zwaan, 1982 or as yet unpublished) show a wide variation in eventual results. At one extreme end there are charts in which next to nothing is found to pierce the noise; at the other end there are range charts (e.g. the Dingden example presented in chapter IX) in which nearly all taxa are involved in correlations and trends.

Whether these results lead to a better interpretation remains a matter of subjective judgment in the final phase. We maintain that our procedures have a better statistical background than the various multivariate analyses used in the literature. Among these principal components analysis seems to be the most attractive type of analysis to authors.

Chapter II

# INFERENCES ABOUT TAXA FROM A SET OF COUNTS

## II.1. Introduction

Having performed counts on the taxa of a particular group of microfossils (e.g. benthonic foraminifera, calcareous nannofossils) obtained from a set of samples, one may wish to have objective descriptions of the behaviour of the numbers of every taxon, and of the mutual numerical behaviour of every pair of taxa.

This chapter, as well as chapters III and IV, deal with the major problem that inferences concerning the "real numbers" of the taxa are wanted, whereas only data on the proportions of the taxa are available. Proportions have the closed sum property, i.e. the proportions of the taxa add up to one (or 100%) and are therefore interrelated. In section II.2 the closed sum property is expressed in the variances and the covariances of the proportions.

This interrelation between the proportions of the taxa need not be present in the "real numbers" or "open variables" of the taxa in the series of counts. Denoting the proportion variable of taxon i by $P_i$ and the open variable of taxon i by $X_i$, we are interested in testing for independence of the open variables $X_i$ and $X_k$ for each pair (i, k) of taxa, or in estimating the value of the correlation coefficient $R(X_i, X_k)$ with the help of the data on the proportions $P_i$.

It is not correct to use the deviation of the value of the correlation coefficient $R(P_i, P_k)$ from zero in testing for independence of $X_i$ and $X_k$. It is better to use, in the first instance at least, the deviation from the expected value $R_m(P_i, P_k)$ according to the multinomial model, which is described in section II.3. The multinomial hypothesis – that all counts are from identical assemblages – must generally be rejected in our experience. Nevertheless we are inclined to accept some preliminary conclusions, namely that $R(X_i, X_k) > 0$ if $R(P_i, P_k)$ is "significantly" greater than $R_m(P_i, P_k)$ and that $R(X_i, X_k) < 0$ if $R(P_i, P_k)$ is "significantly" less than $R_m(P_i, P_k)$. Theoretically such conclusions can hardly be defended. In practice these conclusions often appeared to be correct after they were compared with the outcome of the open variables models.

Other methods – less successful ones in our opinion – of making inferences concerning the open variables will be discussed in this chapter. The

use of proportion-ratio correlation coefficients is dealt with in sections II.3 and II.8 (and in III.7). Partial correlation is discussed in II.9.

In section II.4 the technique of outside counting is discussed in connection with the multinomial model. If the model is true, the outside scores are positively correlated. The features of the outside counting technique become very complex if the multinomial model has to be rejected. It is possible to analyse the technique of outside counting under the assumption that the open variables model of Chayes and Kruskal (chapter III) is true, but we decided not to present such an analysis, as this technique is not frequently used.

In sections II.5, II.6 and II.7 the numerical analysis and related problems are presented. This analysis is the basis of our Fortran computer program DISTUR.

In the discussion of section II.8 the open variables concept is proposed as a better alternative for making inferences from the proportions.

## II.2. Inside counts

The number of counts (samples, or repeated counts of one sample) that have been performed is denoted by N, the number of the categories (taxa) of the group of microfossils considered is denoted by M. In the following we shall speak simply of the M taxa, regardless of whether such categories are genera, species or groups of such taxa.

The number or score of taxon i (= the i-th taxon) in count j (= the j-th count) is denoted by $x_{ij}$. If all counts have the fixed size n, we can write

$$\sum_{i=1}^{M} x_{ij} = n \text{ for each } j. \tag{2.1}$$

If the size of the counts is not fixed, the size of count j is called $n_j$, so

$$\sum_{i=1}^{M} x_{ij} = n_j \text{ for each } j. \tag{2.2}$$

The proportion of taxon i in count j is defined as

$$\hat{p}_{ij} = x_{ij}/n_j. \tag{2.3}$$

This proportion $\hat{p}_{ij}$ is an estimate of the proportion $p_{ij}$ of taxon i in assemblage j from which count j comes. Obviously

$$\sum_{i=1}^{M} p_{ij} = 1 \text{ and } \sum_{i=1}^{M} \hat{p}_{ij} = 1 \text{ for each count } j. \tag{2.4}$$

For a set of counts of fixed size n, we regard the series of scores ($x_{1j}$, $x_{2j}$, $x_{3j}$, . . ., $x_{Mj}$) of a count j as a realization of the random vector ($x_1$, $x_2$, $x_3$, . . ., $x_M$). It is intuitively felt that the random variables $x_i$ of taxon i and $x_k$ of taxon k must be interdependent. If in one of the counts j $x_{ij}$ happens to be relatively large, the score of all other taxa together in that count, $n - x_{ij}$, must be relatively small. So the score $x_{kj}$ of the taxon k, being a part of the score $n - x_{ij}$, will tend to be relatively small. We conclude that $x_{ij}$ and $x_{kj}$ tend to be more or less negatively correlated for each pair of taxa i and k.

This intuitive reasoning has a foundation. We recall that the expected value of a random variable A is denoted by E(A). The variance of a random variable A is defined by

$$\mathrm{var}(A) = E((A - E(A))^2). \tag{2.5}$$

The covariance between two random variables A and B is defined by

$$\mathrm{cov}(A, B) = E((A - E(A)) \cdot (B - E(B))). \tag{2.6}$$

Correlation between two random variables is usually expressed by Pearson's product moment correlation coefficient, later on simply called the correlation coefficient:

$$R(A, B) = \frac{\mathrm{cov}(A, B)}{\sqrt{\mathrm{var}(A) \cdot \mathrm{var}(B)}} \tag{2.7}$$

Independence of two random variables implies that their covariance is equal to zero, and as a consequence that their correlation coefficient is equal to zero as well:

$$A \text{ and } B \text{ mutually independent} \rightarrow \mathrm{cov}(A, B) = 0;\ R(A, B) = 0 \tag{2.8}$$

The interdependence of the variables in the random vector ($x_1$, $x_2$, $x_3$, . . ., $x_M$) is easily demonstrated (Chayes, 1960). For each taxon i the equation

$$\mathrm{var}(x_i) + \sum_{k \neq i}^{M} \mathrm{cov}(x_i, x_k) = \mathrm{cov}(x_i, \sum_{k=1}^{M} x_k) = \mathrm{cov}(x_i, n) = 0$$

holds. Hence,

$$\sum_{k \neq i}^{M} \mathrm{cov}(x_i, x_k) = -\mathrm{var}(x_i) \text{ for each taxon i.} \tag{2.9}$$

As variances have positive values (except for random variables A that are

constant, then var(A) = 0), it is concluded that the sum of the covariances between the score of taxon i and the score of any other taxon k gives a negative value. Therefore it is impossible that all (M – 1) covariances cov($x_i$, $x_k$) are equal to zero. The score $x_i$ cannot be independent of all other scores.

Identical statements can be made for the proportions $\hat{p}_{ij}$ of (2.3). The advantage over the scores $x_{ij}$ is that we are no longer tied to equal sizes of the counts. Ignoring the variable sizes of the counts we regard the series of proportions ($\hat{p}_{1j}$, $\hat{p}_{2j}$, $\hat{p}_{3j}$, . . ., $\hat{p}_{Mj}$) of a count j as a realization of the random vector ($P_1$, $P_2$, $P_3$, . . ., $P_M$). The interdependence of the variables in this random vector is demonstrated in the same way as in (2.9):
For each taxon i the equation

$$\mathrm{var}(P_i) + \sum_{k \neq i}^{M} \mathrm{cov}(P_i, P_k) = \mathrm{cov}(P_i, \sum_{k=1}^{M} P_k) = \mathrm{cov}(P_i, 1) = 0$$

holds, so

$$\sum_{k \neq i}^{M} \mathrm{cov}(P_i, P_k) = -\mathrm{var}(P_i) \text{ for each taxon i} \qquad (2.10)$$

(See Chayes, 1960, p. 4185, who considered proportions of chemical elements). It is concluded that the sum of the covariances between the proportion of taxon i and the proportion of every other taxon k gives a negative value. Therefore it is impossible that all (M – 1) covariances cov($P_i$, $P_k$) are equal to zero. The proportion $P_i$ cannot be independent of all other proportions.

Being interested especially in the mutual numerical behaviour of various pairs of taxa, we realisé of course that the correlation coefficients R($P_i$, $P_k$) of such a pair of taxa i and k must not be compared to the value zero in testing for independence of the "real numbers" of the pair of taxa.

Since about 1960 several attempts have been made to make inferences about these "real numbers" from the closed data. The most successful attempt so far is, in our opinion, the open variables concept, which will be discussed in chapters III and IV. In the next section we first present the multinomial model.

## II.3. The multinomial model

We give the names "multinomial hypothesis" and "multinomial model" to the hypothesis that all N counts are from one single assemblage, or rather

from N identical assemblages, i.e. all mutual differences between the counts are completely due to random effects during the counting procedure (counting "errors" or "noise").

More precisely one can say that count j comes from assemblage j which is characterized by the series $(p_{1j}, p_{2j}, p_{3j}, \ldots, p_{Mj})$, in which $p_{ij}$ is the numerical proportion of taxon i in assemblage j. Then the multinomial hypothesis can be formulated as follows: for each taxon i there is a number $p_i$ ("the" proportion of taxon i) such that

$$p_{ij} = p_i \text{ for each assemblage j.} \tag{2.11}$$

Assuming that this hypothesis is true, and regarding scores in fixed counts (see 2.1), the score $x_i$ of taxon i, as a random variable, has the binomial probability distribution with parameters n and $p_i$. The expected value and the variance of $x_i$ are

$$E(x_i) = n \cdot p_i \text{ and } \operatorname{var}(x_i) = n \cdot p_i \cdot (1 - p_i) \tag{2.12}$$

Considering the pair of scores $x_i$ of taxon i and $x_k$ of taxon k, it can be shown quite easily (Lebart & Fénelon, 1973) that $\operatorname{cov}(x_i, x_k)$ is negative:

$$\operatorname{cov}(x_i, x_k) = -n \cdot p_i \cdot p_k \tag{2.13}$$

From (2.12) and (2.13) it follows that

$$R_m(x_i, x_k) = -\sqrt{\frac{p_i \cdot p_k}{(1 - p_i) \cdot (1 - p_k)}} \tag{2.14}$$

The symbol m indicates that it is a hypothetical correlation coefficient valid under the multinomial hypothesis.

A remarkable feature of expression (2.14) is the absence of n, the size of the counts; $R_m(x_i, x_k)$ is dependent only on the assemblage proportions $p_i$ and $p_k$. If these proportions are both small, $R_m$ will be close to zero; if they are large, $R_m$ will be unmistakably negative and different from zero, as if there were a negative relation between the taxa i and k. In reality, only the numbers $x_i$ and $x_k$ are interdependent, or (negatively) correlated, but this correlation is completely "due to the closed sum": the scores $x_i$ for all taxa i must add up to n and are therefore dependent on each other. This negative "bias" of such a hypothetical correlation coefficient $R_m$ is often referred to as a false correlation or a spurious correlation.

The hypothesis (2.11) gets its name multinomial from the fact that, if it is correct, the series $(x_1, x_2, x_3, \ldots, x_M)$ with fixed sum n, in which each taxon (each pair of taxa) has the properties (2.12), (2.13) and (2.14), is said

to have a multinomial distribution with the parameters n, $p_1$, $p_2$, $p_3$, . . ., $p_M$.

Skipping the assumption that the counts have equal sizes n (count j has size $n_j$) we again consider the proportions $\hat{p}_{ij} = x_{ij}/n_j$ and present similar formulae for the corresponding random vector of proportions ($P_1$, $P_2$, $P_3$, . . ., $P_M$), under the assumption that the multinomial hypothesis (2.11) holds.

The expected value and the variance of $P_i$ of taxon i are

$$E(P_i) = p_i \text{ and } var_m(P_i) = p_i \cdot (1 - p_i)/n^* \tag{2.15}$$

in which n* is the harmonic mean of the $n_j$:

$$n^* = \left(\sum_{j=1}^{N} n_j^{-1}\right)^{-1} \tag{2.16}$$

Considering the pair of proportions $P_i$ of taxon i and $P_k$ of taxon k, it can be shown that

$$cov_m(P_i, P_k) = -p_i \cdot p_k/n^* \tag{2.17}$$

and as a consequence

$$R_m(P_i, P_k) = -\sqrt{\frac{p_i \cdot p_k}{(1 - p_i) \cdot (1 - p_k)}} \tag{2.18}$$

It is noted that the expressions for $var_m(P_i)$ of (2.15) and $cov_m(P_i, P_k)$ of (2.17) contain n*, but that $E(P_i)$ of (2.15) and $R_m(P_i, P_k)$ of (2.18) are independent of the sizes $n_j$ of the counts. We also note that the expressions for $var_m(P_i)$ of (2.15) and for $cov_m(P_i, P_k)$ of (2.17) do indeed fulfil the closed sum property (2.10).

Let us assume that the correlation coefficient $R(P_i, P_k)$ calculated from a range chart deviates significantly from the value $R_m(P_i, P_k)$ of (2.18). The only correct conclusion from this fact is that the multinomial model must be rejected for the range chart under consideration. We then need the open variables approach to obtain further results.

The series of open variables ("real numbers"), denoted by the vector ($X_1$, $X_2$, . . ., $X_M$), is related to the series of closed variables ("proportions"), denoted by the vector ($P_1$, $P_2$, . . ., $P_M$), by means of the relation

$$X_i = T \cdot P_i \text{ or } P_i = X_i/T;\ T = X_1 + X_2 + \ldots + X_M.$$

We are inclined to accept yet another preliminary conclusion about this

series of open variables, namely that $R(X_i, X_k) > 0$ if $R(P_i, P_k)$ is significantly greater than $R_m(P_i, P_k)$ and that $R(X_i, X_k) < 0$ if $R(P_i, P_k)$ is significantly less than $R_m(P_i, P_k)$. It is emphasized that this preliminary conclusion may be entirely wrong; from the point of view of mathematical statistics this conclusion is incorrect. Nevertheless such preliminary conclusions are drawn by Mosimann (1962) in his example (p. 79, 80) as well as by the present author, but they are liable to be rejected in a later phase of the investigation. See also II.8.

Another attempt to solve the "closed sum problem" is mentioned next because it appears to be tied to the multinomial model. Darroch and Ratcliff (1970, 1978) suggest that one should consider not only $R(P_i, P_k)$, but also the correlation between the proportion of element (taxon) i and the part element (taxon) k covers in the proportion of all taxa with the exception of element (taxon) i, thus the correlation between $P_i$ and $P_k/(1 - P_i)$. They argue that the ratio $P_k/(1 - P_i)$ is independent of $P_i$ if "the elements are independent", or in the words of Darroch and Ratcliff (1978, p. 362): the elements exhibit "no-association". In our opinion there is a hidden system of "real numbers" in this reasoning, because of their hypothesis that the proportion of element i and (the mean value of) the proportion that element k has amongst the sum of all other elements together are independent (Darroch & Ratcliff, 1970, p. 308).

We observe that for the closed variables $P_i$ and $P_k$

$$R(P_i, \frac{P_k}{1 - P_i}) = 0 \quad \text{and} \quad R(P_k, \frac{P_i}{1 - P_k}) = 0 \tag{2.19}$$

must hold if the multinomial model (2.11) is valid. If the multinomial model is replaced by the hypothesis of mutual independent open variables of Chayes and Kruskal, these two expressions need not yield zero for any pair of taxa; they can differ considerably. See sections II.8 and III.7. The latter section is devoted to a more extensive analysis of this subject.

Another attempt by Chave and Mackenzie (1961) to "eliminate" the closed sum property (2.10) is discussed in section II.9.

## II.4. Outside counts

Before we start to describe the statistics in connection with the multinomial model, some lines must be devoted to the technique of outside counting, also called inverse counting. The fact that the statistical properties

of outside counts deviate from the properties described above for inside counts may cause some confusion about interpretations if the technique of outside counts is used.

We shall discuss the outside counts technique by means of an example. A, B, C, D and E are five taxa with proportions $q_1$, $q_2$, $q_3$, $q_4$ and $q_5$, such that $(q_1 + q_2 + q_3 + q_4 + q_5) = 1$. We assume that hypothesis (2.11) is valid for a series of counts, for instance by considering that the counts come from the same assemblage characterized by $(q_1, q_2, q_3, q_4, q_5)$. The inside counting technique means that the j-th count is finished as soon as the fixed number n has been reached: $x_{1j} + x_{2j} + x_{3j} + x_{4j} + x_{5j} = n$.

Another way to perform count j is to continue counting the five taxa until $x_{1j} + x_{2j} + x_{3j} = n$. This means that the scores of the taxa D and E are "outside" the fixed sum n; the count is stopped when n specimens of the taxa A, B and C together have been found, without taking the numbers of the taxa D and E at that moment into consideration. An important point is that if this procedure is performed the proportions $q_i$ have to be transformed:

$$p_1 = \frac{q_1}{(q_1 + q_2 + q_3)}; \quad p_2 = \frac{q_2}{(q_1 + q_2 + q_3)}; \quad p_3 = \frac{q_3}{(q_1 + q_2 + q_3)};$$

$$u_4 = \frac{q_4}{(q_1 + q_2 + q_3)}; \quad u_5 = \frac{q_5}{(q_1 + q_2 + q_3)}$$

Each of the three taxa A, B and C have scores $x_i$ $(i = 1, 2, 3)$ which have the binomial distribution with parameters n and $p_i$; the series $(x_1, x_2, x_3)$ has a multinomial distribution with the parameters n, $p_1$, $p_2$ and $p_3$, and the properties mentioned in the previous section. The scores $x_i$ $(i = 4, 5)$ of each of the two taxa D and E have different probability distributions: the negative binomial distribution with parameters n and $u_i$ which has the following properties:

$$E(x_i) = n \cdot u_i \qquad \text{var}(x_i) = n \cdot u_i \cdot (1 + u_i) \qquad (i = 4, 5) \tag{2.20}$$

It appears that the value of the variance is greater than the expected value $n \cdot u$ itself. This is in contrast to the inside scores $x_i$ $(i = 1, 2, 3)$ which have a binomial distribution with parameters n and $p_i$ and have the properties $E(x_i) = n \cdot p_i$ and $\text{var}(x_i) = n \cdot p_i \cdot (1 - p_i)$. The latter variance is less than the corresponding expected value. If an "inside" proportion $p_i$ is close to zero, at least much less than one, then

$$\text{var}(x_i) = E(x_i). \tag{2.21}$$

Similarly, if the "outside ratio" $u_i$ is close to zero or at least much less than one, (2.21) will also be valid, as can be deduced from (2.20). It appears that both a binomial distribution with small $p_i$ and a negative binomial distribution with small $u_i$ are close to Poisson distributions with parameters $n \cdot p_i$ and $n \cdot u_i$, respectively. Both distributions describe haphazard occurrences (M. M. Drooger, in Zachariasse et al., 1978, p. 23–24) for which var(A) = E(A) is correct.

Another, more striking difference between inside and outside scores is that

$$\mathrm{cov}(x_4, x_5) = +n \cdot u_4 \cdot u_5, \tag{2.22}$$

so according to (2.7)

$$R_m(x_4, x_5) = +\sqrt{\frac{u_4 \cdot u_5}{(1+u_4)\cdot(1+u_5)}} \tag{2.23}$$

Evidently the correlation between the outside scores of two taxa is positive, in contrast to (2.14). Expression (2.23), like (2.14), does not contain n. Table 1 gives values of $R_m(x_1, x_2)$ for several pairs $(p_1, p_2)$ according to (2.14), and of $R_m(x_4, x_5)$ for similar pairs $(u_4, u_5)$ according to (2.23).

TABLE 1

| $p_1$ | $p_2$ | r | $u_4$ | $u_5$ | r |
|---|---|---|---|---|---|
| 0.02 | 0.02 | −0.02 | 0.02 | 0.02 | +0.02 |
| 0.02 | 0.20 | −0.07 | 0.02 | 0.40 | +0.07 |
| 0.02 | 0.50 | −0.14 | 0.02 | 2.00 | +0.11 |
| 0.10 | 0.10 | −0.11 | 0.20 | 0.20 | +0.17 |
| 0.10 | 0.50 | −0.33 | 0.20 | 1.00 | +0.29 |
| 0.25 | 0.25 | −0.33 | 0.20 | 4.00 | +0.37 |
| 0.25 | 0.50 | −0.58 | 1.00 | 1.00 | +0.50 |
| 0.40 | 0.40 | −0.67 | 1.00 | 4.00 | +0.63 |
| 0.50 | 0.50 | −1. | 4.00 | 4.00 | +0.80 |

Each series $(y_1, y_2 \ldots)$ of outside scores under the assumption that hypothesis (2.11) holds (so that for each pair $(y_i, y_k)$ properties identical to (2.20), (2.22) and (2.23) of $(x_4, x_5)$ are valid) is said to have a negative multinomial distribution (Mosimann, 1965). Mosimann (1962, 1963, 1965) has extensively investigated inside and outside counts and the related probability distributions.

Another, less well known feature is that, provided hypothesis (2,11) is true, and still following the numbers of our example

$$R_m(x_i, x_k) = 0 \text{ for any pair } (i \leqslant 3, k \geqslant 4) \tag{2.24}$$

In words: the score of any taxon counted inside and the score of any taxon counted outside are independent. This can be argued by stating that $x_{kj}$, $k \geqslant 4$, is the score of the k-th taxon in count j at the moment when $x_{1j} + x_{2j} + x_{3j} = n$, whatever the values of $x_{1j}$, $x_{2j}$ and $x_{3j}$. If the three taxa A, B and C were lumped into one category, this would not affect the scores $x_{kj}$, $k \geqslant 4$.

Another, much more important feature of this series of scores ($x_{1j}$, $x_{2j}$, $x_{3j}$, $x_{4j}$, $x_{5j}$) of the taxa A, B, C, D and E is that one is allowed to "forget" the way these counts were realised (namely A, B and C counted inside with sum n, D and E counted outside the sum) and to consider them as counts with variable sizes $n_j = x_{1j} + x_{2j} + x_{3j} + x_{4j} + x_{5j} = n + x_{4j} + x_{5j}$, in which the random variables $P_i$ with realizations $\hat{p}_{ij} = x_{ij}/n_j$, $i = 1, 2, 3, 4, 5$, have the properties mentioned in (2.15), (2.17) and (2.18), the $p_i$ being identical to the $q_i$ defined at the beginning of this section.

In practice taxa with very large proportions will be counted outside, in order to prevent all other taxa from being represented by very small numbers. If the multinomial model (2.11) is rejected for some reason, we get problems similar to those mentioned in section II.3. Values of $R(x_i, x_k)$ calculated from the data set, which are distinctly different from the "multinomial" values from (2.14), (2.23) or (2.24) should not be used to make inferences about the real numbers, for instance to draw the conclusion that the correlation coefficient between the corresponding open variables $X_i$ and $X_k$ deviates from zero.

Nevertheless it is sometimes clear what the conclusion should be. An example is given by counts on calcareous nannofossils by Schmidt (in Meulenkamp et al., 1978, p. 347, 348) in which the two groups counted outside *Pseudoemiliana lacunosa* and the "small coccoliths" seem to have a strong, negative correlation with the "inside counted" *Helicosphaera carteri* in the lower and middle part of the section Prassa (fig. 1). Although this feature has not been considered in detail, it is believed that the strongly fluctuating patterns of the two series of outside scores are due to strongly fluctuating "real numbers" of *H. carteri* (squeezing effect, see also II.8).

## II.5. Statistics concerning the multinomial model

The multinomial model, defined by (2.11) and characterized by the properties (2.15), (2.17) and (2.18), is a good starting point to describe a series of counts of a group of microfossils. An R-mode computer program in Fortran language, called DISTUR, has been developed (R-mode means: taxon-taxon in this context). In this program statistics concerning the mul-

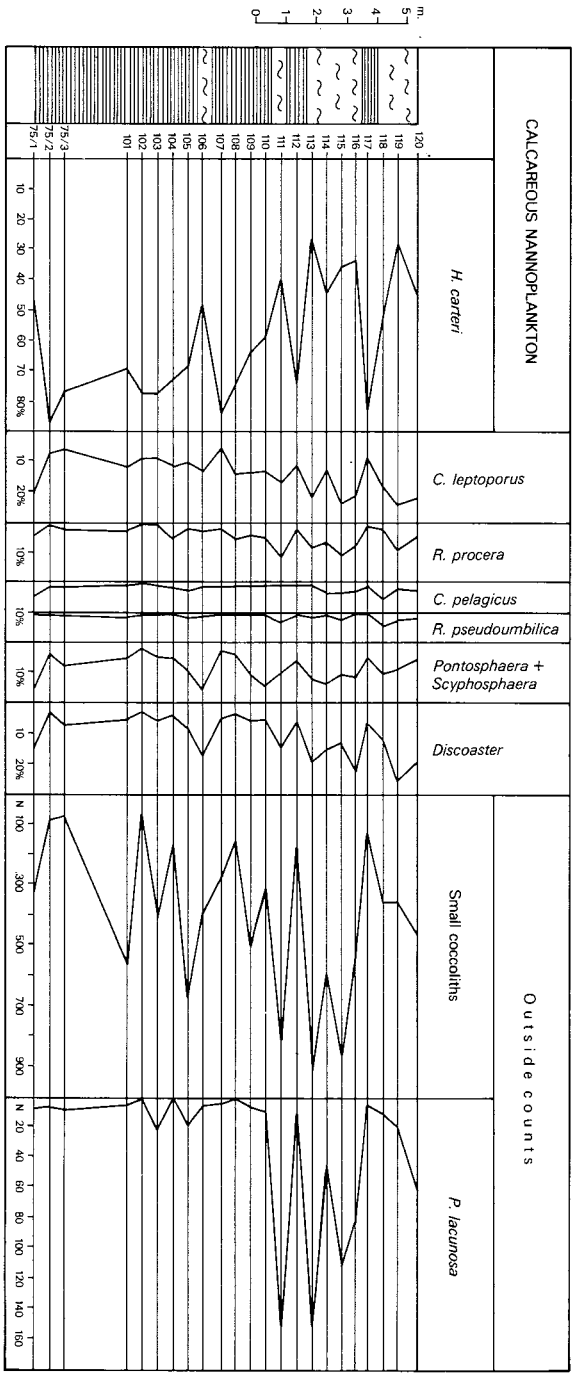



Fig. 1 Frequency distribution pattern of calcareous nannoplankton in part of the Prassa section (after Meulenkamp et al., 1978).

tinomial model are calculated. These statistics are dealt with below. It is emphasized that they even provide useful information if the multinomial model is not correct.

Hypothesis (2.11) says that for each taxon i, the proportions $p_{ij}$ in any assemblage j have a fixed value, namely the value $p_i$. Then $p_i$ is estimated by

$$\hat{p}_i = \frac{\sum_{j=1}^{N} x_{ij}}{\sum_{j=1}^{N} n_j} = \frac{\sum_{j=1}^{N} n_j \cdot \hat{p}_{ij}}{\sum_{j=1}^{N} n_j} \tag{2.25}$$

The last expression in (2.25) demonstrates that any proportion $\hat{p}_{ij}$ in count j is weighted according to $n_j$, the size of the count j. Therefore $\hat{p}_i$ is called here the weighted estimate of the proportion of taxon i. If all counts have equal sizes (2.25) reduces to

$$\bar{p}_i = (\sum_{j=1}^{N} \hat{p}_{ij})/N \quad \text{or} \quad \bar{x}_i = (\sum_{j=1}^{N} x_{ij})/N \tag{2.26}$$

The symbol "$\bar{\ }$" will be used here to indicate unweighted estimates (in this case unweighted averages). The first expression of (2.26) can also be used in the case of counts of variable size (see next section).

The multinomial hypothesis (2.11) – that there is a common proportion $p_i$ of taxon i so that for each assemblage proportion $p_{ij}$ of assemblage j $p_{ij} = p_i$ holds – can be tested by means of the chi square statistic

$$X^2 = \frac{\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{p}_i)^2}{\hat{p}_i \cdot (1 - \hat{p}_i)} \tag{2.27}$$

which has a chi square distribution with (N–1) degrees of freedom, if the hypothesis holds. If all counts are of equal size, the statistic reduces to

$$X^2 = \frac{n \cdot \sum_{j=1}^{N} (\hat{p}_{ij} - \bar{p}_i)^2}{\bar{p}_i \cdot (1 - \bar{p}_i)} = \frac{n \cdot \sum_{j=1}^{N} (x_{ij} - \bar{x}_i)^2}{\bar{x}_i \cdot (n - \bar{x}_i)} \tag{2.28}$$

See Cochran (1954) for a more elaborate treatment.

Assuming that the multinomial null-hypothesis (2.11) is valid, the correlation coefficient value between the proportions $P_i$ and $P_k$ of taxa i and k in the counts is estimated by substituting $\hat{p}_i$ and $\hat{p}_k$ according to (2.25) into $R_m(P_i, P_k)$ of (2.18):

$$\hat{R}_m(P_i, P_k) = -\sqrt{\frac{\hat{p}_i \cdot \hat{p}_k}{(1-\hat{p}_i)\cdot(1-\hat{p}_k)}} \qquad (2.29)$$

## II.6. General statistics concerning a set of counts

The proportions $\hat{p}_{ij} = x_{ij}/n_j$ are the realizations of the closed proportions $P_i$ for a taxon i in the counts. In contrast, the proportions $p_{ij}$ are the realizations of the random variable $Q_i$ for each taxon i in the assemblages. The two formulae (2.25) and (2.26), now written

$$\overline{P}_i = \frac{\sum_{j=1}^{N} \hat{p}_{ij}}{N} \quad \text{and} \quad \hat{P}_i = \frac{\sum_{j=1}^{N} x_{ij}}{\sum_{j=1}^{N} n_j} \qquad (2.30)$$

are thought to be convenient estimates of $E(P_i)$. Obviously the mean values of the assemblage proportion and of the count proportion of a taxon are equal:

$$E(Q_i) = E(P_i) \qquad (2.31)$$

One might well ask which is the best estimate of $E(P_i)$. We write for each count/assemblage j

$$\hat{p}_{ij} = p_{ij} + d_{ij}$$

in which $d_{ij}$ is the counting error (binomial error). Let $D_{ij}$ be the random variable associated with $d_{ij}$, then

$$E(D_{ij}) = 0; \quad \text{var}(D_{ij}) = \frac{E(Q_i)\cdot(1-E(Q_i))}{n_j}$$

The best estimate of $E(P_i)$ is then the expression

$$\sum_{j=1}^{N} a_j \cdot \hat{p}_{ij} \quad \text{with each } a_j > 0 \quad \text{and with} \quad \sum_{j=1}^{N} a_j = 1$$

in which all $a_j$ are chosen in such a way that $\text{var}(\sum_{j=1}^{N} a_j \cdot \hat{p}_{ij})$ has the smallest possible value. It follows that

$$\left(\sum_{j=1}^{N} a_j^2\right) \cdot \text{var}(Q_i) + \left(\sum_{j=1}^{N} \frac{a_j^2 \cdot E(Q_i)\cdot(1-E(Q_i))}{n_j}\right)$$

has the smallest possible value. With the extra condition $\sum_{j=1}^{N} a_j = 1$ it follows that there is a constant c such that

$$a_j \cdot \mathrm{var}(Q_i) + \frac{a_j \cdot E(Q_i) \cdot (1 - E(Q_i))}{n_j} = c.$$

It appears that the general expression of the constant c and of each $a_j$ in terms of $\mathrm{var}(Q_i)$, $E(Q_i)$ and $n_j$ is very complex. However, if $\mathrm{var}(D_{ij})$ is negligible compared to $\mathrm{var}(Q_i)$, we have

$$a_j \cdot \mathrm{var}(Q_i) = c$$

which leads to $c = \mathrm{var}(Q_i)/N$ and to $a_j = 1/N$. In that case $\overline{P}_i$ is the best estimate of $E(P_i)$. If $\mathrm{var}(Q_i)$ is negligible relative to $\mathrm{var}(D_{ij})$, we have

$$\frac{a_j \cdot E(Q_i) \cdot (1 - E(Q_i))}{n_j} = c$$

which leads to $c = E(Q_i) \cdot (1 - E(Q_i))/(\sum_{j=1}^{N} n_j)$ and to $a_j = n_j/(\sum_{j=1}^{N} n_j)$, so in this case $\hat{P}_i$ is the best estimate of $E(P_i)$.

In $\hat{P}_i$ the sizes $n_j$ of the counts appear as weights, whereas in $\overline{P}_i$ they do not. The estimate $\hat{P}_i$ is called the "weighted" estimate of $E(P_i)$ (and of $E(Q_i)$), $\overline{P}_i$ is called the "unweighted" estimate of $E(P_i)$.

For all other statistics both an unweighted form and a weighted form were used in our investigation. The mathematical foundation given below for the weighted statistics (2.33, 2.36, 2.39 and 2.43) is admittedly poor, however. These weighted statistics are based on the idea that large counts should contribute with more weight. However, if the multinomial model is far from true, i.e. if all $\mathrm{var}(P_i)$ are very large, such overweight can theoretically hardly be defended. In practice the values of the unweighted forms and of the corresponding weighted forms usually rarely differ. In our experience the variability of both the total numbers of the counts and of the series of proportions of the taxa is fairly limited. Extremely low total numbers evidently had already been eliminated by the micropaleontologist during the counting procedure.

### Unweighted and weighted estimates

Although in our experience weighted and unweighted estimates rarely yield different results, it is thought appropriate to give the respective lines of reasoning and the formulae.

The straightforward estimate of the variance $\mathrm{var}(P_i)$ is

$$\bar{var}(P_i) = \frac{\sum_{j=1}^{N} (\hat{p}_{ij} - \bar{P}_i)^2}{N - 1} \tag{2.32}$$

The symbol "$\bar{\ }$" indicates that the statistic is unweighted; the sizes $n_j$ are not considered. Obviously $\bar{var}(P_i)$ is connected to $\bar{P}_i$ of (2.30).

The weighted variance estimate is

$$\hat{var}(P_i) = \frac{\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{P}_i)^2}{\sum_{j=1}^{N} n_j} \tag{2.33}$$

which is connected to the weighted estimate $\hat{P}_i$. It is noted that $\hat{var}(P_i) = (\frac{N-1}{N}) \cdot \bar{var}(P_i)$ if all counts are equal in size.

It is emphasized that the variance of the assemblage proportions $Q_i$ for each taxon i is not equal to the variance of the proportions $P_i$ for that taxon i in the counts, but that

$$var(P_i) = var(Q_i) + \frac{E(Q_i) \cdot (1 - E(Q_i))}{n^*} \tag{2.34}$$

in which $n^*$ is the harmonic mean of the $n_j$. See (2.16).

The formulae for the unweighted and for the weighted estimate of the covariance $cov(P_i, P_k)$ between the proportion $P_i$ of taxon i and the proportion $P_k$ of taxon k are

$$\bar{cov}(P_i, P_k) = \frac{\sum_{j=1}^{N} (\hat{p}_{ij} - \bar{P}_i) \cdot (\hat{p}_{kj} - \bar{P}_k)}{N - 1} \tag{2.35}$$

and

$$\hat{cov}(P_i, P_k) = \frac{\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{P}_i) \cdot (\hat{p}_{kj} - \hat{P}_k)}{\sum_{j=1}^{N} n_j} \tag{2.36}$$

respectively.

Again the covariance between the assemblage proportions $Q_i$ of the taxon i and $Q_k$ of the taxon k is not equal to the covariance between the proportions $P_i$ and $P_k$ in the counts, but

$$\operatorname{cov}(P_i, P_k) = \operatorname{cov}(Q_i, Q_k) - \frac{E(Q_i) \cdot E(Q_k)}{n^*} \tag{2.37}$$

As far as the expressions (2.34) and (2.37) are concerned, see Mosimann (1962).

According to (2.7), the formulae for the unweighted and for the weighted estimate of the correlation coefficient $R(P_i, P_k)$ between $P_i$ and $P_k$ are

$$\bar{R}(P_i, P_k) = \frac{\sum_{j=1}^{N} (\hat{p}_{ij} - \bar{P}_i) \cdot (\hat{p}_{kj} - \bar{P}_k)}{\sqrt{\left( \sum_{j=1}^{N} (\hat{p}_{ij} - \bar{P}_i)^2 \right) \cdot \left( \sum_{j=1}^{N} (\hat{p}_{kj} - \bar{P}_k)^2 \right)}} \tag{2.38}$$

and

$$\hat{R}(P_i, P_k) = \frac{\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{P}_i) \cdot (\hat{p}_{kj} - \hat{P}_k)}{\sqrt{\left( \sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{P}_i)^2 \right) \cdot \left( \sum_{j=1}^{N} n_j \cdot (\hat{p}_{kj} - \hat{P}_k)^2 \right)}} \tag{2.39}$$

respectively.

According to a model such as the multinomial model (2.11) the expected values of the expressions in (2.38) and (2.39) need not be equal to zero. Testing for correlation in such cases is usually performed by means of Fisher's z-transformation, which is described briefly below.

The Fisher's z-transformation is a function that transforms every correlation coefficient value R into the value

$$z(R) = \tfrac{1}{2} \cdot \ln\left(\frac{1+R}{1-R}\right) = \left(R + \frac{R^3}{3} + \frac{R^5}{5} + \frac{R^7}{7} + \ldots\right) =$$

$$= \int_0^R \frac{dx}{1 - x^2} = \tanh^{-1}(R) \tag{2.40}$$

In the literature it is usually only the first or the last expression which is mentioned. According to the last one the transformation z is the inverse function of the hyperbolic tangent:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

From the second expression of (2.40) it can be deduced that $z(R) \approx R$ for any R in the neighbourhood of zero, and that

$$\lim_{R \to 1} z(R) = +\infty \quad \text{and} \quad \lim_{R \to -1} z(R) = -\infty$$ (see fig. 2).

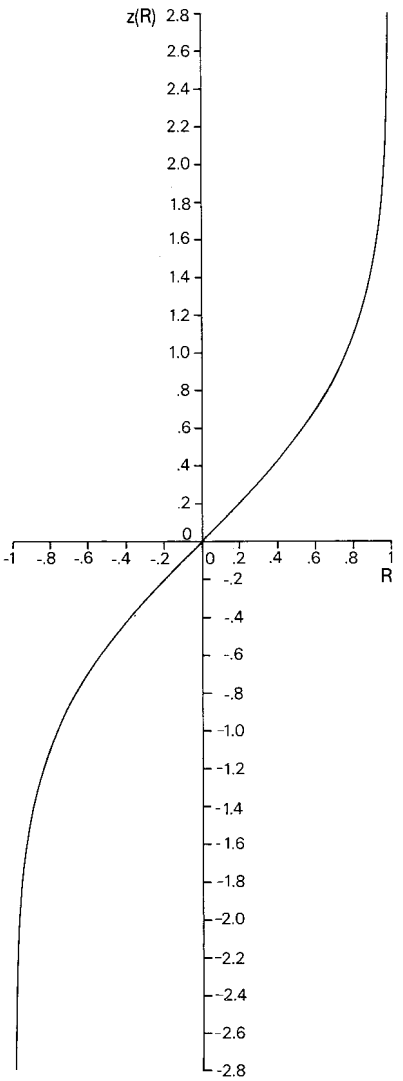



Fig. 2 Graph of $z(R) = \frac{1}{2} \cdot \ln\left(\frac{1+R}{1-R}\right)$, the transformation of Fisher.

If some model leads to a hypothetical correlation coefficient value $R_h$, for instance the multinomial model leads to $\hat{R}_m$ of (2.29), then the model can be tested by means of the statistic

$$(z(\bar{R}) - z(R_h)) \cdot \sqrt{N - 3}$$

and/or by means of the statistic (2.41)

$$(z(\hat{R}) - z(R_h)) \cdot \sqrt{N - 3}$$

which have a standard normal distribution if the hypothesis (model) is true. The probability distributions of $\bar{R}$ and of $\hat{R}$ themselves are markedly skewed if the hypothesis is true and if $R_h$ is far from zero.

Finally, the proportion-ratio correlation coefficients as defined in (2.19) are also calculated in the DISTUR program. Both the unweighted and the weighted estimate are given.

$$\bar{R}(P_i, Q_{ki}) = \frac{\sum_{j=1}^{N} (\hat{p}_{ij} - \bar{P}_i) \cdot (\hat{q}_{kij} - \bar{Q}_{ki})}{\sqrt{\left(\sum_{j=1}^{N} (\hat{p}_{ij} - \bar{P}_i)^2\right) \cdot \left(\sum_{j=1}^{N} (\hat{q}_{kij} - \bar{Q}_{ki})^2\right)}} \qquad (2.42)$$

in which the $\hat{q}_{kij} = \hat{p}_{kj}/(1 - \hat{p}_{ij})$ are considered to be realizations of the random variable $Q_{ki} = P_k/(1 - P_i)$, and in which $\bar{Q}_{ki} = (\sum_{j=1}^{N} \hat{q}_{kij})/ N$.

$$\hat{R}(P_i, Q_{ki}) = \frac{\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{P}_i) \cdot (\hat{q}_{kij} - \hat{Q}_{ki})}{\sqrt{\left(\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{P}_i)^2\right) \cdot \left(\sum_{j=1}^{N} n_j \cdot (\hat{q}_{kij} - \hat{Q}_{ki})^2\right)}} \qquad (2.43)$$

in which $\hat{q}_{kij}$ and $Q_{ki}$ are defined above, and

$$\hat{Q}_{ki} = \left(\sum_{j=1}^{N} n_j \cdot \hat{q}_{kij}\right) \Big/ \left(\sum_{j=1}^{N} n_j\right).$$

## II.7. A note about algorithms

The formulae presented in the previous section are given in such a form that one gains insight into their structure. They are not the forms that are easiest to use in computation, however. The expressions suitable for computation are presented below on the right-hand side of the equality sign.

$$\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{P}_i)^2 = \left( \sum_{j=1}^{N} \hat{p}_{ij} \cdot x_{ij} \right) - \hat{P}_i \cdot \left( \sum_{j=1}^{N} x_{ij} \right)$$

see (2.27), (2.33), (2.39) and (2.43).

$$\sum_{j=1}^{N} (\hat{p}_{ij} - \overline{P}_i)^2 = \left( \sum_{j=1}^{N} \hat{p}_{ij}^2 \right) - \overline{P}_i \cdot \left( \sum_{j=1}^{N} \hat{p}_{ij} \right)$$

see (2.32), (2.38) and (2.42).

$$\sum_{j=1}^{N} (\hat{p}_{ij} - \overline{P}_i) \cdot (\hat{p}_{kj} - \overline{P}_k) = \left( \sum_{j=1}^{N} \hat{p}_{ij} \cdot \hat{p}_{kj} \right) - N \cdot \overline{P}_i \cdot \overline{P}_k$$

see (2.35) and (2.38).

$$\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{P}_i) \cdot (\hat{p}_{kj} - \hat{P}_k) = \left( \sum_{j=1}^{N} n_j \cdot \hat{p}_{ij} \cdot \hat{p}_{kj} \right) - \hat{P}_i \cdot \hat{P}_k \cdot \left( \sum_{j=1}^{N} n_j \right)$$

see (2.36) and (2.39).

As far as the remaining expressions in the proportion-ratio correlation coefficients (2.42) and (2.43) are concerned, their short-cut formulae are obtained by replacing $\hat{p}_{ij}$ or $\hat{p}_{kj}$ by $\hat{q}_{kij}$, $\overline{P}_i$ or $\overline{P}_k$ by $\overline{Q}_{ki}$, and $\hat{P}_i$ or $\hat{P}_k$ by $\hat{Q}_{ki}$ in the formulae above.

## II.8. Discussion

The author has tried to make analyses of range charts (especially R-mode, i.e. taxon-taxon analyses) with the CDC computer of the Academic Computer Centre Utrecht. Initially a simple computer program was written suitable for range charts consisting of sets of counts of fixed size and based on the multinomial model. At that time a fair number of 200-counts on planktonic foraminifera, on benthonic foraminifera and on calcareous nannofossils were already available in Utrecht, so sufficient experience could be obtained with the statistics of range charts.

Soon it appeared that the multinomial model was not very good for describing range charts and distribution charts. As mentioned in section II.3 we wanted to derive more information from a set of significant values of (2.41)

$$( z(R(P_i, P_k)) - z(R_m(P_i, P_k)) ) \cdot \sqrt{N - 3}$$

on a set of pairs of taxa i and k, than the mere fact that the multinomial model should be rejected. Conclusions concerning the open variables ('real numbers') in the sense that $R(X_i, X_k) > 0$ of the expression (2.41) is signific-

antly positive and that $R(X_i, X_k) < 0$ if the expression (2.41) is significantly negative may be entirely wrong, as is demonstrated in the following imaginary example, illustrated in figure 3.

In the middle of this figure we show eight assemblages (I – VIII) in stratigraphic order, with the closed variables $P_1$, $P_2$, $P_3$ and $P_4$ of four taxa. The proportion $P_1$ fluctuates markedly along the column. The closed variables $P_2$, $P_3$ and $P_4$ each take up more or less equal parts of the remaining space in the closed sum. Each of the triple set $P_2$, $P_3$ and $P_4$ is negatively correlated with the highly fluctuating $P_1$, whereas each pair within that triple set is positively correlated. We assume that the counts are so large that the expressions (2.41)

$$( z(R(P_i, P_k)) - z(R_m(P_i, P_k)) ) \cdot \sqrt{N - 3}$$

are significantly negative for the pairs (1, 2), (1, 3) and (1, 4), and significantly positive for the pairs (2, 3), (2, 4) and (3, 4).

The closed variables $P_1$, $P_2$, $P_3$ and $P_4$ are considered to be derived from a set of open variables $X_1$, $X_2$, $X_3$ and $X_4$, being the ratios $P_{ij} = X_{ij}/(X_{1j} + X_{2j} + X_{3j} + X_{4j}) = X_{ij}/T_j$ for i = 1, 2, 3, 4 and for j = I, II, III, IV, V, VI, VII and VIII.

It appears that there is a "large" collection of solutions for the series of open variables that fits to a specific closed data set. In the lower part of figure 3 one alternative open variables solution based on interdependence of the four taxa is given as an example for the set of four imaginary closed variables. This set of open variables is almost identical to the set of closed variables, with some "arbitrary" small deviations.

However, we prefer a solution for the series of open variables in which there is the least possible interdependence of the taxa. This type of solution is given in the upper part of figure 3. This solution has been constructed on the basis of independence between each pair of taxa. This leads to the open variable $X_1$ with a large relative mean value and a relatively very large variance compared to the three other open variables $X_2$, $X_3$ and $X_4$. If we accept the "independent" set of open variables as the most correct solution it would be entirely wrong to conclude that $R(X_1, X_2)$, $R(X_1, X_3)$ and $R(X_1, X_4)$ are negative and that $R(X_2, X_3)$, $R(X_2, X_4)$ and $R(X_3, X_4)$ are positive.

According to our preferred solution the "real numbers" of taxon 1 "disturb" the frequency pattern of the others. "Disturbers" such as taxon 1 in this imaginary example do occur in range charts; there is often at least one taxon suspected of causing disturbance. During the last few years the

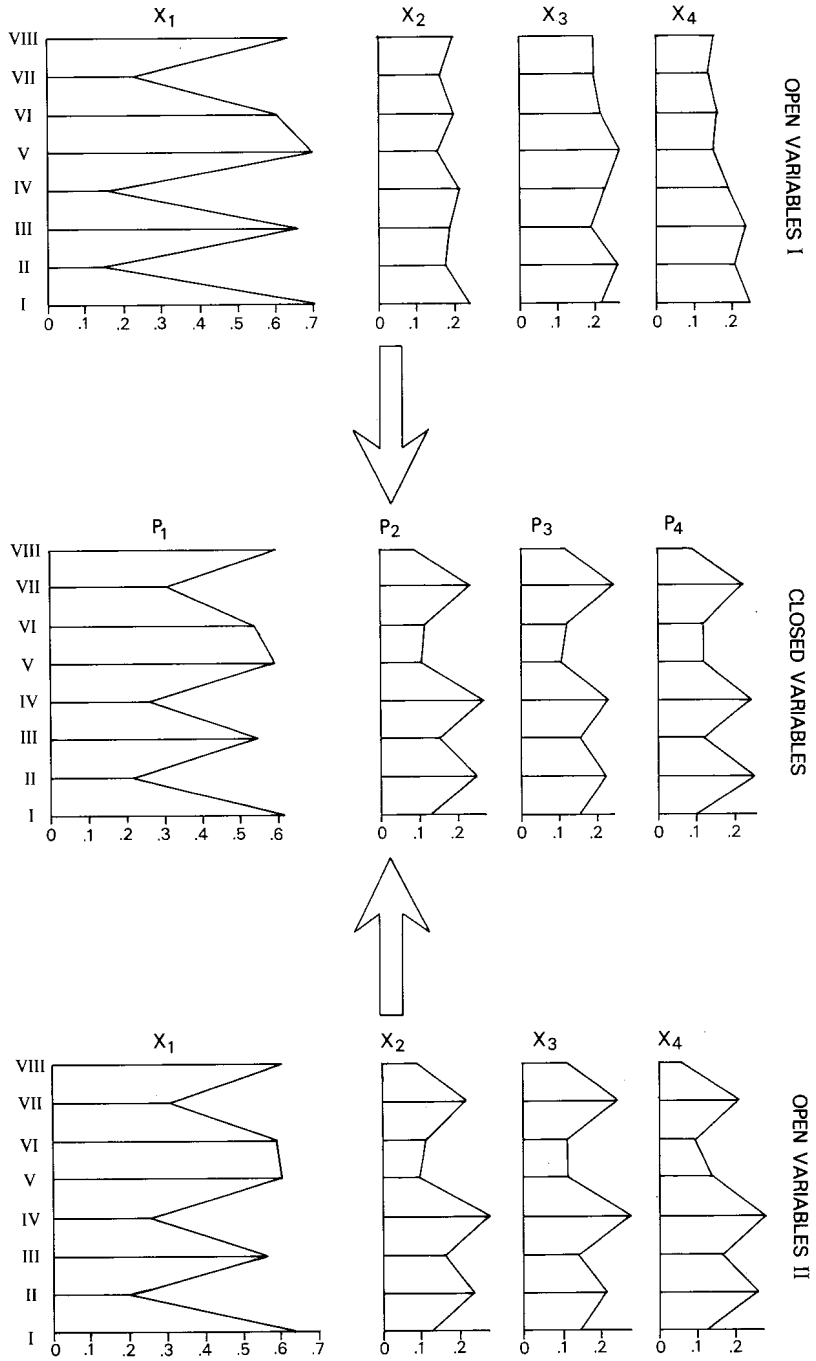



Fig. 3 Imaginary example of eight samples and four taxa to show the relation between open and closed variables (see text).

author has made several attempts to "eliminate" the effect of such disturbing taxa.

As soon as one taxon (or more taxa) with greatly fluctuating proportions is (are) recognized in the series of counts, one can delete the scores of this taxon (these taxa) from the counts as if this taxon (these taxa) had not been considered during the counting. All other taxa together form a new "revised" chart, now consisting of counts of variable size. Although our intention was to count fixed and sufficiently large numbers from all samples, this procedure is not really necessary for the statistical analysis. Anyway, one can hardly avoid having counts of variable size if the group of microfossils under consideration is rare in some of the samples from a stratigraphic section.

For some time we thought that the application of the proportion-ratio correlation coefficients (2.19) was a correct way of handling – and of recognizing – "disturbing" or "squeezing" taxa. Although taxon 1 in figure 3 is disturbing and the multinomial model therefore is certainly not valid, $Q_{21} = P_2/(1 - P_1)$ can be independent of $P_1$, if the four open variables are mutually independent. $P_1$ and $Q_{21}$ being written

$P_1 = X_1/(X_1 + X_2 + X_3 + X_4)$ and
$Q_{21} = X_2/(X_2 + X_3 + X_4)$ respectively, the variables

$X_1$, $(X_2 + X_3 + X_4)$ and $X_2/(X_2 + X_3 + X_4)$ must be supposed to be mutually independent. Then there is no relation between the proportion of taxon 1 and the share that taxon 2 has in the sum of all other taxa together, so

$$R(P_1, P_2/(1 - P_1)) = R\left(\frac{X_1}{X_1 + X_2 + X_3 + X_4}, \frac{X_2}{X_2 + X_3 + X_4}\right) = 0$$

However, $Q_{12} = P_1/(1 - P_2)$ is not independent of $P_2$, because both $P_2 = X_2/(X_1 + X_2 + X_3 + X_4)$ and $Q_{12} = X_1/(X_1 + X_3 + X_4)$ are largely determined by the highly fluctuating variable $X_1$:

$$R(P_2, P_1/(1 - P_2)) < 0$$

Considering the three stable taxa 2, 3 and 4, it appears that for instance $Q_{23} = P_2/(1 - P_3)$ is not independent of $P_3$ either, because $Q_{23} = X_2/(X_1 + X_2 + X_4)$ and $P_3 = X_3/(X_1 + X_2 + X_3 + X_4)$ both are largely determined by the highly fluctuating $X_1$:

$$R(P_3, P_2/(1 - P_3)) > 0$$

Our simple imaginary example shows that the proportion-ratio correlation coefficient $R(P_i, P_k/(1 - P_i))$ may be different from zero notwithstanding the fact that the open variables are mutually independent. If one of a pair of

taxa is clearly disturbing, we shall obtain strongly different values for $R(P_i, P_k/(1 - P_i))$ and $R(P_k, P_i/(1 - P_k))$. For some time we thought that such different values would be useful indicators of a disturbing or squeezing taxon. This procedure has indeed been successful in cases where there was only a single distinct squeezer. However, if there is more than one taxon that disturbs, the pattern soon becomes too complex to be solved with proportion-ratio correlation coefficients.

In section III.7 we shall give more details to support the statements made above concerning these correlation coefficients.

After these moderately successful attempts we started to use the open variables concept. This concept was introduced by Chayes and Kruskal (1966) and by Chayes (1971) for the handling of geochemical data. Their hypothesis that all open variables are independent, results in a set of expected values of the correlation coefficients between the closed variables of every pair of taxa ($R_c(P_i, P_k)$); the hypothesis therefore can be tested. This procedure causes problems similar to those we encountered already when considering the multinomial model. This "zero-open-covariances" model, as we call it, is described in chapter III and has been incorporated in our Fortran program DISTUR. In chapter IV we present another approach to the open variables concept. This "free-open-covariances" model leads to a "best" solution in which the open variables may still show some interdependence. This model has been elaborated in a separate computer program called BALANC.

## II.9. Partial correlation and the closed sum problem

Some lines will be devoted to the use of partial correlation coefficients in connection with closed sum data. Chave and Mackenzie (1961) apply this statistical technique to chemical data of major and trace elements in pelagic muds ($\hat{p}_{ij}$ represents the proportion of chemical element i in sample j). Chave and Mackenzie try to "eliminate the correlation due to the closed sum" by calculating partial correlation coefficients with respect to the major element.

Chayes (1962, p. 451) already showed by means of a strict proof that the closed form of the data is not eliminated by applying such a partial correlation. This last fact is shown here in another way by means of a simple example from the multinomial model.

$Y_1$, $Y_2$ and $Y_3$ are three random variables that are mutually correlated. The correlation coefficients between $Y_1$ and $Y_2$, between $Y_1$ and $Y_3$, and between $Y_2$ and $Y_3$ are denoted by $R_{12}$, $R_{13}$ and $R_{23}$ respectively. Observations are labelled j = 1, 2, 3, . . . . We are interested in those observations ($Y_{1j}$, $Y_{2j}$, $Y_{3j}$) in which $Y_{3j}$ is a fixed value c, so that we can consider the

correlation $R_{12.3}$ between the two variables $Y_1$ and $Y_2$ that is not affected by the third variable $Y_3$ because $Y_3$ is kept constant. It can be proved that

$$R_{12.3} = \frac{R_{12} - (R_{13} \cdot R_{23})}{\sqrt{(1 - R_{13}^2) \cdot (1 - R_{23}^2)}} \tag{2.44}$$

$R_{12.3}$ is called the partial correlation coefficient between $Y_1$ and $Y_2$ with respect to $Y_3$. See e.g. Simpson, Roe & Lewontin (1960). If $Y_3$ were independent of $Y_1$ and of $Y_2$, then $R_{13} = 0$ and $R_{23} = 0$, so that according to (2.44), $R_{12.3} = R_{12}$. If on the other hand $Y_3$ had a strong correlation both with $Y_1$ and with $Y_2$, i.e. $R_{13}$ and $R_{23}$ are both far from zero, then the correlation $R_{12}$ between $Y_1$ and $Y_2$ would be partly a reflection of the dependence of these variables on $Y_3$. $R_{12.3}$ is the form in which this dependence on $Y_3$ is eliminated.

Chave and Mackenzie (1961) use formula (2.44) for the purpose of eliminating false correlations between trace elements caused by the fluctuations in the proportions of the major chemical element. This procedure does not eliminate the closed sum character, however.

Let a series of counts on a group of microfossils fit the multinomial model, and, by analogy with the geochemical example, let taxon 1 have a very large proportion $p_1$, so that all other taxa have small proportions $p_2$, $p_3$, . . ., $p_M$. According to (2.18) the correlation coefficient between the proportions of taxon i and those of taxon k is

$$R_{ik} = R_m(P_i, P_k) = -\sqrt{\frac{p_i p_k}{(1 - p_i) \cdot (1 - p_k)}}$$

Substituting such forms in (2.44), we arrive at

$$R_{ik.1} = -\sqrt{\frac{p_i p_k}{(1 - p_1 - p_i) \cdot (1 - p_1 - p_k)}} = -\sqrt{\frac{q_i \cdot q_k}{(1 - q_i) \cdot (1 - q_k)}} \tag{2.45}$$

in which $q_i = p_i/(1 - p_1)$, $q_k = p_k/(1 - p_1)$. This means that the partial correlation coefficient between the taxa i and k with respect to taxon 1 is equal to the correlation coefficient between the taxa i and k, as if all scores of taxon 1 had been deleted from all counts. For the revised series of counts the multinomial model still holds and the proportions of the taxa i and k have mean values $q_i$ and $q_k$, respectively. The closed sum character is still present. Deleting the scores of a disturbing taxon and performing calculations on the revised series of counts is a more straightforward technique than calculating partial correlation coefficients. Therefore the latter technique has not been considered any further in our investigation.

Chapter III

# THE ZERO OPEN COVARIANCES MODEL

## III.1. Introduction and definitions

Chayes and Kruskal (1966) introduced the open variables concept and built a model, incorporating the extra assumption (hypothesis) that all open variables are mutually independent, i.e. that $\mathrm{cov}(X_i, X_k) = 0$ for each pair of taxa i and k. We call their model the zero open covariances model because of this extra assumption. We describe the numerical part and related problems below in section III.1, III.2 and III.3, giving our comments and modifications. The model which we present in these first three sections is somewhat different from that given by Chayes and Kruskal, but not essentially so. We start from our micropaleontological definition, and not from their geochemical definition.

It is not our aim to accept or reject the hypothesis of Chayes and Kruskal if applied to some range chart. The hypothesis that all open variables are mutually independent has to be rejected as soon as one "open variance" $\mathrm{var}(X_i)$ turns out to be negative. According to our experience the occurrence of a negative open variance is rare, however. Yet the hypothesis is rejected in most cases because "significant" deviations are generally found between values of correlation coefficients $R(P_i, P_k)$ and corresponding expected values $R_c(P_i, P_k)$. It is our aim to make inferences for individual correlation coefficients between open variables $R(X_i, X_k)$ from the differences between the values of the corresponding $R(P_i, P_k)$ and the expected values $R_c(P_i, P_k)$. This theoretically "illegal" procedure is defended at the end of section III.3. From the comparison of the results with the outcome of our own free open covariances model (see chapter IV) we cannot conclude which of the two models is the better one in practice.

In section III.4 we introduce new parameters that are more efficient for calculations and that provide a better insight into the structure of the zero open covariances model of Chayes and Kruskal. The open variance-mean ratio $H_i$ for each taxon i provides a particularly good explanation of the relation between the zero open covariances model and the multinomial model (III.5). It appears to be an indicator for taxa that disturb the closed system because of their highly fluctuating proportions (III.6). In section III.7 the open variance-mean ratios play an important part in the proof that

the expected values of the proportion ratio correlation coefficients $R_c(P_i, P_k/(1 - P_i))$ need not be equal to zero according to the zero open covariances model, except in special situations comparable to the multinomial model.

As we mentioned already in sections I.4 and II.8 the open variable $X_i$ is the number of individuals of taxon i per fixed area or per fixed volume in the biocoenosis (micropaleontological definition). The open $X_{ij}$ of taxon i in assemblage j is thought to be achieved by multiplying the proportion $\hat{p}_{ij}$ of taxon i in count j by the factor $T_j$. In terms of random variables

$$X_i = T \cdot P_i$$

It follows that T is the number of individuals of the microfossil group under consideration per fixed area or per fixed volume:

$$T = \sum_{i=1}^{M} X_i$$

We choose for our approach the fixed area or volume in the biocoenosis in such a way that

$$E(T) = \sum_{i=1}^{M} E(X_i) = 1 \tag{3.1}$$

In the literature E(T) is considered to be equal to a constant number $\tau$, which causes trouble in the formulations. Our replacing of E(T) by one (3.1) simplifies the expressions to a large extent.

Chayes and Kruskal write the basic relation between the open and the closed variables, $X_i = T \cdot P_i$, as

$$P_i = \frac{X_i}{X_1 + X_2 + X_3 + \ldots + X_M} \tag{3.2}$$

i.e. the proportion $\hat{p}_{ij}$ of any chemical element i in sample j is considered to be obtained by a realization of the random vector $(X_1, X_2, X_3, \ldots, X_M)$:

$$\hat{p}_{ij} = \frac{X_{ij}}{X_{1j} + X_{2j} + X_{3j} + \ldots + X_{Mj}}$$

We are not going to dwell on the geochemical definition of the open variables. The definition of Chayes and Kruskal (3.2) and ours

$$\sum_{i=1}^{M} P_i = 1;\ X_i = T \cdot P_i;\ \text{(and as a consequence)}\ T = \sum_{i=1}^{M} X_i \tag{3.3}$$

are equivalent.

The relation (3.2) between the open variables and the closed variables is replaced by the first order approximation:

$$P_i = \frac{X_i}{\sum_{k=1}^{M} X_k} = \frac{E(X_i) + D_i}{1 + \sum_{k=1}^{M} D_k} \approx E(X_i) + D_i - \left( E(X_i) \cdot \sum_{k=1}^{M} D_k \right) =$$

$$= E(X_i) + (1 - E(X_i)) \cdot D_i - E(X_i) \cdot \sum_{k \neq i}^{M} D_k \qquad (3.4)$$

in which $D_i$ is defined as the deviation $(X_i - E(X_i))$. The symbol $\sum_{k \neq i}^{M}$ means that the sum must be taken for k = 1, 2, 3,. . ., M with the exclusion of k=i. The expression (3.4) is regarded as if it were a perfect equality. As a consequence we may write

$$E(P_i) = E(X_i) = p_i \qquad (3.5)$$

(this $p_i$ is not to be confused with the "multinomial" $p_i$ of (2.11)) and

$$P_i - p_i = (1 - p_i) \cdot D_i - p_i \cdot \left( \sum_{k \neq i}^{M} D_k \right) \qquad (3.6)$$

From (3.6) and from the assumption of Chayes and Kruskal:

$$s_{ik} \equiv \mathrm{cov}(X_i, X_k) = 0 \quad \text{for each pair of taxa i and k} \qquad (3.7)$$

it follows that

$$s_{ii} \equiv \mathrm{var}(P_i) = (1 - p_i)^2 \cdot S_{ii} + p_i^2 \cdot \left( \sum_{k \neq i}^{M} S_{kk} \right) \qquad (3.8)$$

for each taxon i ($S_{kk} \equiv \mathrm{var}(X_k)$), and that

$$s_{il} \equiv \mathrm{cov}(P_i, P_l) =$$

$$= - p_l \cdot (1 - p_i) \cdot S_{ii} - p_i (1 - p_l) \cdot S_{ll} + p_i \cdot p_l \cdot \left( \sum_{k \neq i,l}^{M} S_{kk} \right) \qquad (3.9)$$

for each pair of taxa i and l. The symbol $\sum_{k \neq i,l}^{M}$ means that the sum must be taken for k = 1, 2, . . ., M with the exclusion of k = i and k = l. Instead of $\mathrm{var}(P_i)$ and $\mathrm{cov}(P_i, P_l)$ we shall write in the following $s_{ii}$ and $s_{il}$, respectively, in order to discriminate these expressions of the zero open covariances model from those of other models.

Because of (3.7) the variance of T is the sum of all open variances of the $X_i$:

$$S_{tt} \equiv \text{var}(T) = \text{var}\left(\sum_{i=1}^{M} X_i\right) = \sum_{i=1}^{M} \text{var}(X_i) = \sum_{i=1}^{M} S_{ii} \qquad (3.10)$$

With the help of the expression (3.10), the expressions (3.8) and (3.9) reduce to

$$s_{ii} = (1 - 2p_i) \cdot S_{ii} + p_i^2 \cdot S_{tt} \quad \text{for each i} \qquad (3.11)$$

and

$$s_{il} = -p_l \cdot S_{ii} - p_i \cdot S_{ll} + p_i \cdot p_l \cdot S_{tt} \text{ for each pair (i,l)} \qquad (3.12)$$

respectively.

It should be pointed our once more that in this investigation the open variables models are based on the closed variables $P_i$, i.e. on the proportions in the counts, and not on the proportions $Q_i$ in the assemblages from which the counts come.

### III.2. Estimates of the open variables

Equation (3.11) expresses the relation between the "open" variances and the "closed" variances. The closed variances can be estimated by $\bar{\text{var}}(P_i)$ of (2.32) or by $\hat{\text{var}}(P_i)$ of (2.33). The decision whether to use the unweighted or the weighted closed form appeared to be of no practical importance. In the DISTUR program the unweighted form is used, so

$$s_{ii} = \bar{\text{var}}(P_i) \qquad (3.13)$$

is substituted in accordance with the discussion in section II.6 between the formulae (2.31) and (2.32). Therefore $\bar{P}_i$ of (2.30) is substituted for $p_i$ of (3.5) in the computer program. The decision about whether to use unweighted or weighted is only important in the rare cases where the unweighted closed statistics and the weighted closed statistics are markedly different for one or more taxa. In practice this hardly ever occurred.

First we shall now consider the solution of $S_{tt}$ and the $S_{ii}$ from (3.11), the $p_i$ and the $s_{ii}$ being known. (3.11) can be written

$$S_{ii} = \frac{s_{ii}}{(1 - 2p_i)} - \frac{p_i^2}{(1 - 2p_i)} \cdot S_{tt} \quad \text{so}$$

$$S_{tt} = \sum_{i=1}^{M} S_{ii} = \left( \sum_{i=1}^{M} \frac{s_{ii}}{(1 - 2p_i)} \right) - \left( \sum_{i=1}^{M} \frac{p_i^2}{(1 - 2p_i)} \right) \cdot S_{tt}$$

Hence, $S_{tt}$ can be expressed in the $p_i$ and the $s_{ii}$:

$$S_{tt} = \left( \sum_{i=1}^{M} \frac{s_{ii}}{(1 - 2p_i)} \right) \Big/ \left( 1 + \sum_{i=1}^{M} \frac{p_i^2}{(1 - 2p_i)} \right) =$$

$$= \left( \sum_{i=1}^{M} \frac{s_{ii}}{(1 - 2p_i)} \right) \Big/ \left( \sum_{i=1}^{M} \frac{p_i \cdot (1 - p_i)}{(1 - 2p_i)} \right). \qquad (3.14)$$

$S_{tt}$ thus being calculated, each $S_{ii}$ is:

$$S_{ii} = \frac{s_{ii} - p_i^2 \cdot S_{tt}}{1 - 2p_i} \qquad (3.15)$$

Chayes (1971, p. 48) dwells extensively on the case where one of the proportions $p_i$ is equal to $\frac{1}{2}$, so the expression for $S_{tt}$ of (3.14) is undefined. This peculiar situation is not considered here, because it is very unlikely that it would occur in practice, and it does not lead to a better understanding.

The second remark is that $S_{tt}$ of (3.14) is not defined if M = 2, i.e. in the case of two taxa, because then $p_1 = 1 - p_2$, and the denominator in (3.14) will have the outcome zero.

Chayes (1971) shows that the problem of solving the open variances $S_{ii}$ from the closed variances $s_{ii}$ and the proportions $p_i$ can be formulated in terms of matrix algebra: expression (3.8) can be written as

$$s = P\,\sigma \qquad (3.16)$$

in which s is a column vector with i-th coordinate $s_{ii}$, $\sigma$ a column vector with i-th coordinate $S_{ii}$, P a matrix having in the j-th row and k-th column

$$P_{jk} = \begin{cases} (1 - p_j)^2 & \text{if } j = k \\ p_j^2 & \text{if } j \neq k \end{cases}$$

It can be proved that P is non-singular for $M \geqslant 3$, so there is an inverse $P^{-1}$ and $\sigma$ can be deduced from s:

$$\sigma = P^{-1}\,s \qquad (3.17)$$

The deduction formulated in (3.14) and (3.15) is more straightforward, however.

For M = 2, (3.16) leads to

$$\begin{pmatrix} s_{11} \\ s_{22} \end{pmatrix} = \begin{pmatrix} (1-p_1)^2 & p_1^2 \\ p_2^2 & (1-p_2)^2 \end{pmatrix} \begin{pmatrix} S_{11} \\ S_{22} \end{pmatrix} = \begin{pmatrix} (1-p_1)^2 & p_1^2 \\ (1-p_1)^2 & p_1^2 \end{pmatrix} \begin{pmatrix} S_{11} \\ S_{22} \end{pmatrix} \quad (3.18)$$

Expression (3.18) confirms that $s_{11} = s_{22}$. This results in only a single equation for the open variances $S_{11}$ and $S_{22}$:

$$(1 - p_1)^2 \cdot S_{11} + p_1^2 \cdot S_{22} = s_{11}.$$

Hence, there is more than one solution for M = 2.

Another important item, discussed in the literature is that the expression for $S_{ii}$ of (3.15) cannot be guaranteed to yield a desired, positive number for every taxon i. If for some taxon i $S_{ii}$ appears to be negative or zero, we can conclude that the hypothesis (3.7) concerning mutually independent open variables must be rejected. This occurred frequently in the investigation carried out by Saha, Bhattacharyya and Lakshmipathy (1974), who however considered M = 4 cases only. So far a negative open variance has been found only once in the considerable number of data sets we have investigated, but in our data sets M is much larger (between 8 and 20).

### III.3. Closure correlation according to the zero open covariances model

The open variables $S_{ii}$ and $S_{tt}$ having been calculated, the next step is to test the zero open covariances model by calculating the correlation coefficient values between each pair of hypothetical closed variables which are deduced from the open variables, and by comparing these hypothetical correlation coefficients with the corresponding correlation coefficients $\bar{R}(P_i, P_k)$ of (2.38) or $\hat{R}(P_i, P_k)$ of (2.39) calculated directly from the closed data.

The hypothetical closed covariances $s_{il}$, expected according to the zero open covariances model, are calculated from (3.12):

$$s_{il} = p_i \cdot p_l \cdot S_{tt} - p_l \cdot S_{ii} - p_i \cdot S_{ll}.$$

The correlation coefficient between the proportions of taxon i and those of taxon l, expected according to the zero open covariances model, is:

$$R_c(P_i, P_l) = \frac{p_i \cdot p_l \cdot S_{tt} - p_l \cdot S_{ii} - p_i \cdot S_{ll}}{\sqrt{s_{ii} \cdot s_{ll}}} \quad (3.19)$$

It is emphasized that $R_c$ is a hypothetical value, just like $\hat{R}_m$ of (2.18) of the multinomial model. This $R_c$ is compared to $\bar{R}(P_i, P_k)$ of (2.38) and to $\hat{R}(P_i, P_k)$ of (2.39) by means of the statistic (2.41).

Chayes (1971) remarks that this test (2.41) is only valid for $R_c$ if $M \geqslant 4$. For $M = 3$ it turns out that $R_c(P_i, P_k) = \bar{R}(P_i, P_k)$ if $\bar{var}(P_i)$ is substituted for $s_{ii}$ as we decided on the basis of (3.13). If $\hat{var}(P_i)$ is substituted for $s_{ii}$, then $R_c(P_i, P_k) = \hat{R}(P_i, P_k)$. Below we give the proof for the unweighted case; the same proof holds for the weighted case.

If M, the number of taxa, is three, we have

$$\bar{var}(P_3) = \bar{var}(P_1 + P_2) = \bar{var}(P_1) + \bar{var}(P2) + 2\,\bar{cov}(P_1, P_2).$$

Substituting $\bar{var}(P_i) = s_{ii}$ according to (3.13), we have

$$\bar{cov}(P_1, P_2) = \frac{\bar{var}(P_3) - \bar{var}(P_1) - \bar{var}(P_2)}{2} = \frac{s_{33} - s_{11} - s_{22}}{2}.$$

Substituting (3.8) in $s_{11}$, $s_{22}$ and $s_{33}$ in the last expression will yield the expression of (3.9) for $s_{12}$, so:

$$\bar{cov}(P_1, P_2) = s_{12}.$$

Hence, not only $\bar{var}(P_i) = s_{ii}$, but also $\bar{cov}(P_i, P_1) = s_{il}$, so

$$\bar{R}(P_i, P_l) = R_c(P_i, P_l).$$

We conclude that if $M = 3$ the open variances $S_{ii}$ can be determined without any problem, according to (3.14) and (3.15). The zero open covariances model cannot be tested by means of the correlation coefficients, however. From now onwards in this chapter we consider the number of taxa to be more than three.

Different interpretations of significant values of $(z(R) - z(R_c)) \cdot \sqrt{N - 3}$ of (2.41) are found in the literature, which are similar to the different interpretations of $(z(R) - z(R_m)) \cdot \sqrt{N - 3}$ in the multinomial model (see II.3).

Miesch (1969) follows the only correct way from the point of view of mathematical statistics by stating that as soon as one significant value of (2.41) is found, the zero open covariances model must be rejected; no inference is allowed about open covariances $S_{ik} = cov(X_i, X_k)$ and open correlation coefficients $R(X_i, X_k)$. He shows by means of an imaginary example with $M = 4$ that such inferences can be entirely wrong.

Chayes (1970) maintains that inferences can be made about $R(X_i, X_k)$ from the statistic (2.41):

$$(z(R(P_i, P_k)) - z(R_c(P_i, P_k))) \cdot \sqrt{N - 3}$$

if the number M is sufficiently "large" (eight or more) and if N is greater

than 30. He states that it is permissible to use the statistic to test the hypothesis that $R(X_i, X_k) = 0$ for the pair of taxa i and k involved. In contrast, Miesch wishes to conclude that a significant value of (2.41) leads to accepting the hypothesis that one or more open covariances (not necessarily the one between the taxa i and k) are different from zero.

Our own opinion is somewhere in between; our experience with range charts of microfossils has shown from the comparison with the results from our BALANC computer program (see chapter IV) that testing the hypothesis $R(X_i, X_k) = 0$ by means of the statistic

$$( z(\overline{R}(P_i, P_k)) - z(R_c(P_i, P_k)) ) \cdot \sqrt{N - 3}$$

is quite permissible if $M \geqslant 8$. If the outcome of the statistic is significantly positive, $R(X_i, X_k) > 0$ is accepted. If the outcome is significantly negative, $R(X_i, X_k) < 0$ is accepted.

### III.4. Choosing appropriate parameters

After some experience with the model of Chayes and Kruskal we realized that the above expressions concerning the closed forms $s_{ii}$ and $s_{ik}$, the two open forms $S_{ii}$ and $S_{tt}$, and the mean proportions $p_i$ are not efficient for calculations. Other expressions given below are more efficient and give a better insight into the structure. The basic formulae (3.11) and (3.12) are written more concisely by defining some convenient parameters.

$$c_{ii} \equiv s_{ii}/(p_i^2) \tag{3.20}$$

is the first parameter of this kind, being the square of the coefficient of variability of the closed variable $P_i$. The second is

$$u_i \equiv (p_i^2)/(1 - 2p_i) \tag{3.21}$$

which is a useful transformation of the proportion $p_i$. The third is

$$c_{ik} \equiv s_{ik}/(p_i \cdot p_k) \tag{3.22}$$

which we call the coefficient of covariability of the pair of closed variables $P_i$ and $P_k$. Like $s_{ik}$, the parameter $c_{ik}$ is a hypothetical parameter giving an expected value according to the zero open covariances model. Finally we define

$$H_i \equiv S_{ii}/p_i \tag{3.23}$$

as the open variance-mean ratio, abbreviated to **v.m.r.** This ratio can be given a logical foundation, which will be presented in the following section.

Here we restrict ourselves to the metamorphoses of the different formulae.

The basic equation (3.11) reduces to

$$c_{ii} = (S_{ii}/u_i) + S_{tt} \tag{3.24}$$

and (3.12) reduces to

$$c_{il} = -H_i - H_l + S_{tt} \tag{3.25}$$

The expression (3.14) for $S_{tt}$ reduces to

$$S_{tt} = \frac{\sum_{i=1}^{M} u_i \cdot c_{ii}}{1 + \sum_{i=1}^{M} u_i} \tag{3.26}$$

The expression (3.15) for $S_{ii}$ changes into

$$S_{ii} = u_i \cdot (c_{ii} - S_{tt}) \tag{3.27}$$

Finally, the expression (3.19) for $R_c(P_i, P_l)$ is modified to

$$R_c(P_i, P_l) = \frac{c_{il}}{\sqrt{c_{ii} \cdot c_{ll}}} = \frac{S_{tt} - H_i - H_l}{\sqrt{c_{ii} \cdot c_{ll}}} \tag{3.28}$$

We end this section with some remarks on the parameter $u_i$. The function $u : p \mapsto u(p) = (p^2)/(1 - 2p)$ has the properties $u(p) > 0$ for $0 < p < \frac{1}{2}$; $u(p) < 0$ for $\frac{1}{2} < p \leqslant 1$; $u(0) = 0$; $\lim_{p \uparrow \frac{1}{2}} u(p) = +\infty$; $\lim_{p \downarrow \frac{1}{2}} u(p) = -\infty$; $u(\frac{1}{2})$ is not defined; $u(1) = -1$.

### III.5. The open variance-mean ratios

In this section we postulate that the zero open covariances model is valid for the series of counts. Starting from the definition (3.23) of the open variance-mean ratios and from (3.10) we deduce that

$$S_{tt} = \sum_{i=1}^{M} p_i \cdot H_i \tag{3.29}$$

i.e. the variance of the sum of the open variables is the weighted mean of the open v.m.r.'s, weighted according to the proportions $p_i$.

We wish to consider the case where the open v.m.r.'s of all taxa are equal to some constant positive number m, so $H_i = m$ for all i. Then it immediately

follows from (3.29) that $S_{tt} = m$. This special case implies that each taxon contributes its open variance $S_{ii}$ to the total open variance $S_{tt}$ according to its proportion $p_i$, because

$$S_{ii} = m \cdot p_i \text{ and } S_{tt} = \sum_{i=1}^{M} S_{ii} = m.$$

If all open v.m.r.'s $H_i$ are equal to the constant m, it follows from (3.11), (3.24), (3.12), (3.25) and (3.28) that

$$s_{ii} = m \cdot p_i \cdot (1 - p_i); \quad c_{ii} = \frac{m \cdot (1 - p_i)}{p_i}; \quad s_{il} = -m \cdot p_i \cdot p_l;$$

$$c_{il} = -m \text{ and } R_c(P_i, P_l) = -\sqrt{\frac{p_i \cdot p_l}{(1 - p_i) \cdot (1 - p_l)}} \qquad (3.30)$$

respectively.

From the comparison of (3.30) with (2.18) it appears that the condition of mutually equal open v.m.r.'s results in a resemblance with the multinomial model (see Chayes & Kruskal, 1966, p. 697). The expected values of the correlation coefficients between the closed variables have become identical for both models. If in addition $m = 1/n^*$ is substituted in (3.30), $n^*$ being the harmonic mean of the sizes $n_j$ of the counts, then the expressions (2.15) and (2.17) reappear and it turns out that the multinomial model can be seen as a special case of the zero open covariances model.

The construction of the multinomial model by means of open variables has been considered by Mosimann (1962) and by Connor and Mosimann (1969). They suppose that the variables $X_i$ have specific probability distributions, namely gamma distributions. If each $X_i$ has a gamma distribution with parameters $\alpha = p_i/m$ and $\beta = m$ (translated into our terms), and the $X_i$ are mutually independent, then the series $(P_1, P_2, P_3, \ldots, P_M)$ defined by

$$P_i = \frac{X_i}{X_1 + X_2 + X_3 + \ldots + X_M} \qquad (3.31)$$

has a Dirichlet distribution, also called multivariate beta distribution with parameters $\alpha_1 = p_1/m$, $\alpha_2 = p_2/m$, $\alpha_3 = p_3/m$, $\ldots$, $\alpha_M = p_M/m$. It has properties almost identical to (3.30). The mutually independent $X_i$ with the gamma distributions described above have the properties

$$E(X_i) = p_i; \quad \text{var}(X_i) = m \cdot p_i; \quad \text{cov}(X_i, X_k) = 0.$$

The $P_i$ of (3.31) obtained by "closure" of these $X_i$ have the properties

$$E(P_i) = p_i;\ \mathrm{var}(P_i) = \frac{m \cdot p_i \cdot (1 - p_i)}{(1 + m)};\ \mathrm{cov}(P_i, P_k) = \frac{-m \cdot p_i \cdot p_k}{(1 + m)};$$

$$R(P_i, P_k) = -\sqrt{\frac{p_i \cdot p_k}{(1 - p_i) \cdot (1 - p_k)}}.$$

For the theory of these probability distributions the reader is referred to Mosimann (1962) and Connor and Mosimann (1969).

A remarkable result is that the substitution of the latter "closed" expressions in (3.14) and (3.15) results in

$$S_{tt} = \frac{m}{1 + m} \quad \text{and} \quad S_{ii} = \frac{m \cdot p_i}{1 + m}.$$

So with the procedure of Chayes and Kruskal the "original" open statistics $\mathrm{var}(T) = m$ and $\mathrm{var}(X_i) = m \cdot p_i$ are underestimated by a factor $1/(1 + m)$. This is only harmless if the positive number m is much less than one. If $\mathrm{var}(T) = m$ is not much less than one, the amount of underestimate will be considerable. As $n^*$ is supposed to be large (e.g. 50 or more), the substitution $m = 1/n^*$ mentioned earlier will not cause any problem.

It is emphasized that this result is only valid if the open variables $X_i$ have gamma distributions as described above. We have no idea about the value of this underestimate factor if the $X_i$ have no gamma distributions.

### III.6. Recognizing disturbing taxa

Another interesting property of the open v.m.r.'s

$$H_i = S_{ii}/p_i$$

is that they provide a tool to recognize taxa that disturb the closed system because of their highly fluctuating proportions (see II.8). In the previous section we considered the case where each taxon contributes its open variance $S_{ii}$ to the total open variance $S_{tt}$ according to its mean proportion $p_i$, i.e.

$$S_{ii} = p_i \cdot S_{tt} \quad \text{for each taxon } i.$$

Now we consider some taxon k that has

$$S_{kk} \gg p_k \cdot S_{tt} \quad \text{or equivalently} \quad H_k \gg S_{tt} \tag{3.32}$$

which means that its open variance $S_{kk}$ contributes much more to the total

open variance $S_{tt}$ than it should do according to its mean proportion $p_k$. Moreover, if this taxon has quite a large mean proportion $p_k$ it will disturb the correlation coefficients between the closed variables. By disturbing we mean that there are several taxa i for which the correlation coefficient $R_c(P_i, P_k)$ of (3.28), which is to be expected according to the zero open covariances model, is substantially different from the value given in (3.30),

$$-\sqrt{\frac{p_i \cdot p_k}{(1-p_i)\cdot(1-p_k)}}$$

based on the multinomial model. Substituting (3.24)

$$c_{ii} = \frac{S_{ii}}{u_i} + S_{tt} = \frac{H_i \cdot (1-2p_i)}{p_i} + S_{tt}$$

and a similar expression for $c_{kk}$ into (3.28), we get

$$R_c(P_i, P_k) = \frac{S_{tt} - H_i - H_k}{\sqrt{\left(\frac{(1-2p_i)}{p_i}\cdot H_i + S_{tt}\right)\cdot\left(\frac{(1-2p_k)}{p_k}\cdot H_k + S_{tt}\right)}}$$

Writing $y_i = H_i/S_{tt}$ and $y_k = H_k/S_{tt}$, this form reduces to

$$R_c(P_i, P_k) = (1 - y_i - y_k)\cdot f(p_i, y_i)\cdot f(p_k, y_k) \qquad (3.33)$$

in which f is the function defined by

$$f(p, y) = \left(1 + \left(\frac{1-2p}{p}\right)\cdot y\right)^{-\frac{1}{2}} \qquad (3.34)$$

If $H_i = H_k = S_{tt}$, then $y_i = y_k = 1$ and the expression (3.33) will reduce to the "multinomial" value (3.30). If $H_k \gg S_{tt}$, then $y_k \gg 1$ and the factor $(1 - y_i - y_k)$ in (3.33) will yield a value much less than minus one.

From figure 4 that shows graphs of the function $(p \mapsto f(p, y))$ for a series of values of y, it can be read that the factors $f(p_i, y_i)$ and $f(p_k, y_k)$ are large, provided that $p_i$ and $p_k$ are not too small. Hence, substantial deviations from the multinomial value of $R_c(P_i, P_k)$ are to be expected, if taxon k (the disturbing taxon) has $H_k \gg S_{tt}$ and not too small a value for $p_k$ and taxon i neither has too small a value for $p_i$.

If for taxon k $H_k$ is very large compared to $S_{tt}$, it can be expected for several taxa that $0 < H_i, H_l < S_{tt}$, because of (3.29). Then the first factor $(1 - y_i - y_l)$ will be close to zero. If $p_i$ and $p_l$ both are not too small, $R_c(P_i, P_l)$ also shows a substantial deviation from the multinomial value, but now in a positive direction.

The expressions (3.33) and (3.34) appear to help in giving a good foundation to the concept of disturbing taxa, leading to statements that were in accordance with intuition. In the following section a similar foundation will be presented concerning the proportion-ratio correlation coefficients.

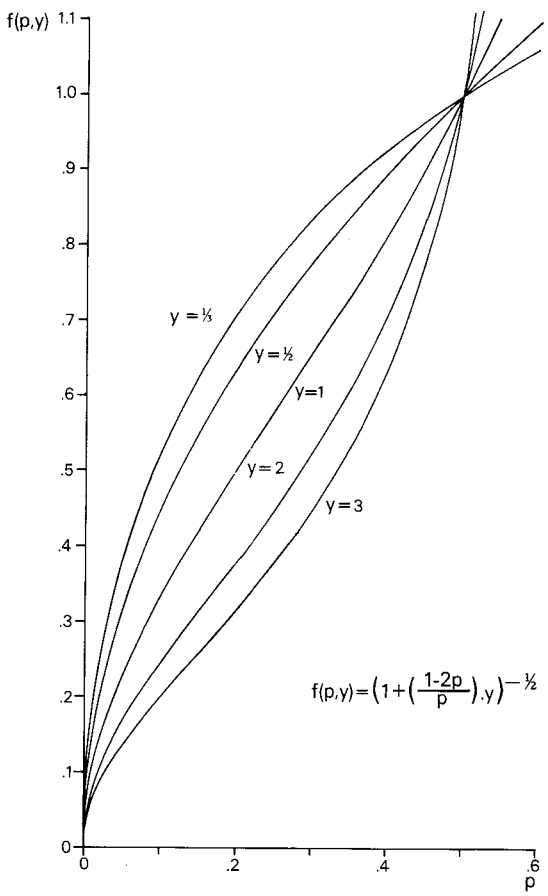



Fig. 4 Graphs of f(p, y) for y = 1/3, 1/2, 1, 2 and 3 (see (3.34)).

## III.7. Expected values of the proportion-ratio correlation coefficients

In these paragraphs statements that were made in earlier sections concerning the proportion-ratio correlation coefficients will be proved with the help of the zero open covariances model. These coefficients have been defined in (2.19) and are calculated in the DISTUR computer program

according to (2.42) (unweighted) and according to (2.43) (weighted). The part these coefficients play is described in the discussion of section II.8.

We shall consider the correlation between the proportion of taxon 1 and the part taxon 2 covers in the proportion of all other taxa. The correlation coefficient value that is expected according to the zero open covariances model is denoted by

$$R_c(P_1, \frac{P_2}{1 - P_1}) \tag{3.35}$$

The proportions of the remaining taxa 3, 4, 5, . . ., M are lumped together, forming the proportion $P_R$ of taxon R. Next, $\sum_{k=3}^{M} X_k \equiv X_R$, so that, because of the mutual independence of the open variables

$$S_{RR} = \sum_{k=3}^{M} S_{kk}$$

and

$$H_R \equiv \frac{S_{RR}}{p_R} = \frac{\sum_{k=3}^{M} S_{kk}}{\sum_{k=3}^{M} p_k} = \frac{\sum_{k=3}^{M} p_k \cdot H_k}{\sum_{k=3}^{M} p_k} \tag{3.36}$$

Hence, the open v.m.r. $H_R$ is the weighted mean of $H_3, H_4, H_5, \ldots, H_M$, weighted according to the mean proportions $p_3, p_4, p_5, \ldots, p_M$. Again we apply the approximation (3.6):

$$(P_i - p_i) = (1 - p_i) \cdot D_i - p_i \cdot \left( \sum_{k \neq i}^{M} D_k \right)$$

in which $D_i = X_i - E(X_i) = X_i - p_i$. We approximate

$$\frac{P_2}{1 - P_1} = \frac{X_2}{X_2 + X_R} = \frac{p_2 + D_2}{p_2 + D_2 + p_R + D_R} = \frac{p_2 + D_2}{(1 - p_1) + D_2 + D_R} \approx$$

$$\approx \frac{p_2}{(1 - p_1)} + \frac{p_R \cdot D_2}{(1 - p_1)^2} - \frac{p_2 \cdot D_R}{(1 - p_1)^2}$$

so that

$$\frac{P_2}{1 - P_1} - \frac{p_2}{1 - p_1} \approx \frac{p_R \cdot D_2 - p_2 \cdot D_R}{(1 - p_1)^2} \tag{3.37}$$

From the approximations (3.6) and (3.37) we deduce

$$\text{var}\left(\frac{P_2}{1-P_1}\right) = \frac{p_R^2 \cdot S_{22} + p_2^2 \cdot S_{RR}}{(1-p_1)^4} \tag{3.38}$$

and

$$\text{cov}\left(P_1, \frac{P_2}{1-P_1}\right) = \frac{p_1 \cdot p_2 \cdot S_{RR} - p_1 \cdot p_R \cdot S_{22}}{(1-p_1)^2} \tag{3.39}$$

On the basis of (3.36), (3.39) can be written:

$$\text{cov}\left(P_1, \frac{P_2}{1-P_1}\right) = \frac{p_1 \cdot p_2 \cdot p_R}{(1-p_1)^2} \cdot (H_R - H_2) \tag{3.40}$$

Analogous to the definition (3.20) of $c_{ii}$ for the closed variables, we define the square of the coefficient of variability of the open variable $X_i$ as

$$C_{ii} \equiv S_{ii}/(p_i^2) \tag{3.41}$$

so (3.38) can be written as:

$$\text{var}\left(\frac{P_2}{1-P_1}\right) = \frac{p_2^2 \cdot p_R^2}{(1-p_1)^4} \cdot (C_{22} + C_{RR}) \tag{3.42}$$

Keeping $\text{var}(P_1) = s_{11}$ in its closed form, the correlation coefficient value (3.35) expected according to the zero open covariances model results in:

$$R_c\left(P_1, \frac{P_2}{1-P_1}\right) = \frac{H_R - H_2}{\sqrt{c_{11} \cdot (C_{22} + C_{RR})}} \tag{3.43}$$

Chayes (1971) presents formulae for the same expressions (his (8.6), (8.7), (8.8) and (8.9)), but he did not transform them as we have done above. From our expression (3.43) one can easily make the following inferences.

In section III.5 it has been shown that the multinomial model is a special case of the zero open covariances model. If all open v.m.r.'s $H_i$ are equal to a constant m, then according to (3.30) the correlations between the closed variables are equal to the correlations expected using the multinomial model. However, if $H_i = m$ for each taxon i, then it appears from (3.36) that $H_R = m$, and from (3.43) that

$$R_c\left(P_1, \frac{P_2}{1-P_1}\right) = 0.$$

Secondly, the numerator in expression (3.43) does not contain the open v.m.r. $H_1$. This means that the possible disturbing behaviour of taxon 1, i.e. $H_1 \gg S_{tt}$ according to (3.32), does not affect the value of $R_c$ of (3.43), which is a confirmation of the statement made in the discussion of II.8.

Thirdly, if taxon 2 is disturbing, then $H_2 \gg S_{tt}$ according to (3.32). Substituting $(p_1 \cdot H_1 + p_2 \cdot H_2 + p_R \cdot H_R)$ for $S_{tt}$, and assuming that $(p_2 + p_R) \cdot H_2$ is still greater than $S_{tt}$ we get $(p_2 + p_R) \cdot H_2 > p_1 \cdot H_1 + p_2 \cdot H_2 + p_R \cdot H_R$. As we suppose the zero open covariances model to be valid, $p_1 \cdot H_1 = S_{11}$ is greater than zero. It follows that $H_2 > H_R$ and that $R_c$ of (3.43) has a negative value.

Finally, we consider the case where one of the taxa of the R-group disturbs in such a way that not only $H_R \gg S_{tt}$ but even $(p_2 + p_R) \cdot H_R > S_{tt}$. Again $(p_1 \cdot H_1 + p_2 \cdot H_2 + p_R \cdot H_R)$ is substituted for $S_{tt}$, so $(p_2 + p_R) \cdot H_R > p_1 \cdot H_1 + p_2 \cdot H_2 + p_R \cdot H_R$. $p_1 \cdot H_1 = S_{11}$ being greater than zero, it follows that $H_R > H_2$ and that $R_c$ of (3.43) has a positive value.

To make it easier for us to survey the output we decided not to include the calculations of the expected values of alle proportion-ratio correlation coefficients according to the zero open covariances model (3.43) in our DISTUR program. If desired, selected expected values can easily be calculated. Using the equalities

$$C_{ii} = S_{ii}/(p_i^2),$$

$$C_{RR} = (S_{tt} - S_{ii} - S_{kk})/((1 - p_i - p_k)^2), \text{ and}$$

$$H_R = (S_{tt} - S_{ii} - S_{kk})/(1 - p_i - p_k), \text{ we can calculate}$$

$$R_c\left(P_i, \frac{P_k}{1 - P_i}\right)$$

according to (3.43). $S_{tt}$ and the $p_i$, $c_{ii}$, $S_{ii}$ and $H_i$ are presented by the output of the computer program DISTUR for all taxa.

## III.8. Statistics concerning the zero open covariances model in the DISTUR program

At the end of the previous section we already made some remarks about the calculation of the different statistics of the zero open covariances model in the DISTUR program. A review of these statistics is presented below.

Substituting the unweighted variance estimate $\overline{var}(P_i)$ of (2.32) for $s_{ii}$, and the unweighted mean proportion estimate $\overline{P}_i$ of (2.30) for $p_i$ in the expres-

sion of (3.20), the square of the coefficient of variability of $P_i$ is:

$$c_{ii} = \overline{var}(P_i)/(\overline{P}_i^2).$$

The statistic $c_{ii}$ is called C(I) – "CLOSED" in the DISTUR program.

The variance of T, the sum of all open variables $X_1, X_2, \ldots, X_M$ is

$$S_{tt} = \left( \sum_{i=1}^{M} u_i \cdot c_{ii} \right) \Big/ \left( 1 + \sum_{i=1}^{M} u_i \right)$$

according to (3.26), in which $u_i = (\overline{P}_i^2)/(1 - 2\overline{P}_i)$, if $\overline{P}_i$ is substituted for $p_i$ in the expression (3.21). In the program $S_{tt}$ is called S(T) – "OPEN".

According to (3.27) the variance of the open variable $X_i$ is

$$S_{ii} = u_i \cdot (c_{ii} - S_{tt})$$

and is called S(I) – "OPEN" in the DISTUR program.

According to (3.23) the open variance-mean ratio is

$$H_i = S_{ii}/\overline{P}_i,$$

if the unweighted mean proportion estimate $\overline{P}_i$ is again substituted for $p_i$. The v.m.r. $H_i$ is called H(I) in the program.

In this program the Fisher's z-transformation of the correlation coefficient between the proportions $P_i$ and $P_k$ of any pair of taxa i and k, which is expected according to the zero open covariances model, is

$$z(R_c(P_i, P_k)) = z\left( \frac{S_{tt} - H_i - H_k}{\sqrt{c_{ii} \cdot c_{kk}}} \right).$$

In the program it is called ZRO(I, K). See the expressions (2.40) and (3.28).

In the DISTUR program some statistics are calculated concerning the sizes of the counts.

The unweighted estimate of the mean value of n is simply

$$\overline{n} = \left( \sum_{j=1}^{N} n_j \right)/N, \tag{3.44}$$

in which N is the number of counts that have been performed. The estimate $\overline{n}$ has been called M(N) in the computer program.

The unweighted estimate of the variance of n,

$$\overline{var}(n) = \left( \sum_{j=1}^{N} (n_j - \overline{n})^2 \right)/(N - 1) \tag{3.45}$$

has been called VAR(N) in the program.

The unweighted estimate of the correlation coefficient between $P_i$ and n is

$$\bar{R}(P_i, n) = \frac{\sum_{j=1}^{N} (\hat{p}_{ij} - \bar{P}_i) \cdot (n_j - \bar{n})}{\sqrt{\left( \sum_{j=1}^{N} (\hat{p}_{ij} - \bar{P}_i)^2 \right) \cdot \left( \sum_{j=1}^{N} (n_j - \bar{n})^2 \right)}} \tag{3.46}$$

In the DISTUR program $\bar{R}(P_i, n)$ has been called R(PI, N).

As to the latter expression one would not expect the estimate of any proportion ($P_i$) and the size of the count to be interdependent. Nevertheless, in practice it may happen that there is a relation between the numerical proportion of some taxon and the number of microfossils that are (or rather that can be) counted from a sample. For instance, in the ostracode data from Prassá section (Crete, Greece) described by Tsapralis (1976) positive correlations were found between the sizes of the counts and the proportions of *Xestoleberis* individuals, and between the sizes of the counts and the proportions of the Rest group. Negative correlations were found by means of the DISTUR program between the sizes of the counts and the proportions of *Callistocythere*, and between the $n_j$ and the proportions of the Cytherurinae (significance level $\alpha = 0.01$). These correlations certainly have a micropaleontological meaning, but we do not consider this in the context of this paper.

Chapter IV

# THE FREE OPEN COVARIANCES MODEL

## IV.1.1. Introduction

From the definition of the open variables given in sections I.5 and III.1, and repeated once more below, one might wish to establish the value of $X_{ij}$ for each taxon i and for each assemblage j from the closed data set. Since

$$X_{ij} = T_j \cdot \hat{p}_{ij}$$

it would be sufficient to establish the value of $T_j$ for each assemblage j. In the concept of our free open covariances model the series $(T_1, T_2, \ldots, T_N)$ should be given values such that the values of the correlation coefficients $R(X_i, X_k)$ for each pair (i, k) of taxa are "near zero".

In this section it will be argued that it is practically impossible to single out one solution from the collection of solutions $(T_1, T_2, \ldots, T_N)$, which belongs to a given correlation matrix R(X) between open variables, except in irrelevant cases where R(X) is singular.

It is better to describe the collection of solutions for the covariance matrix C(X) of open variables directly from the given covariance matrix C(P) of proportions. A linear approximation is used to obtain a useful description of solutions for C(X) in terms of the known C(P). Following the linear approximation the addition of values to the parameters var(T), $\mathrm{cov}(P_1, T)$, $\mathrm{cov}(P_2, T)$, . . ., $\mathrm{cov}(P_M, T)$ is sufficient to determine C(X) from the closed data set.

In sections IV.2, IV.3 and IV.4 numerical aspects of the relation between the "open" covariance matrix C(X) and the "closed" covariance matrix C(P) are presented.

In section IV.5 a criterion is proposed for choosing one solution C(X) from the collection of solutions for some given closed data set. This criterion

$$\sum_{k \neq i}^{M} R(X_i, X_k) = 0 \text{ for each taxon i}$$

bears some resemblance to the hypothesis of mutually independent open variables as stated by Chayes and Kruskal (1966).

A Fortran computer program called BALANC finds our "balanced" solution by means of an iterative process. The properties of such balanced solutions are considered in section IV.6.

In section IV.7 an alternative – in our opinion less successful – balanced solution is presented based on the assumption that $\sum_{k \neq i}^{M} R^2(X_i, X_k)$ has the smallest possible value for each taxon i. This solution can also be calculated by means of the computer program BALANC.

Although the procedures described in this chapter are more direct than those of Chayes and Kruskal (1966), our procedures cannot be shown to be better in practice than theirs.

### IV.1.2. Open variables definition

The approach to our model opens with a repetition of the micropaleontological open variables definition, already presented in section III.1. The open variable $X_i$ is thought to be the number of individuals of taxon i in the biocoenosis per fixed area or per fixed volume. The open variable $X_{ij}$ of taxon i in assemblage j is thought to be achieved by multiplying the proportion $\hat{p}_{ij}$ of taxon i in count j by the factor $T_j$. In terms of random variables $X_i = T \cdot P_i$. Again $\sum_{i=1}^{M} P_i = 1$. It follows that T is the total number of individuals of the microfossil group under consideration per fixed area or per fixed volume:

$$T = T \cdot \sum_{i=1}^{M} P_i = \sum_{i=1}^{M} T \cdot P_i = \sum_{i=1}^{M} X_i$$

The fixed area or volume is chosen in such a way that the mean value of the total number T of individuals, considering all geological samples which might be taken from the stratigraphic section involved, is equal to one:

$$E(T) = 1$$

This open variables problem would be solved completely if in some way a value could be established for $T_j$ of each count j so that a "logical" solution appears. If some value is substituted for the total number $T_j$ the number of individuals $X_{ij}$ of taxon i in assemblage j can then be derived directly from $X_{ij} = T_j \cdot \hat{p}_{ij}$. Further on in this section (IV.1.6) we shall present an example of such a formula for $T_j$. We mentioned already that any such formula will lead to illogical solutions, however. This will also be argued below.

In a few sentences based on an imaginary example it is shown that it may be even completely impossible to trace back the value of $T_i$ from the closed data, notwithstanding the fact that the original open data from which the closed data have been derived form an acceptable system.

It is supposed that all $X_i$ have some gamma distribution (see section III.5) and that they are mutually independent. The only restriction is that the beta parameters of all gamma distributions have the same value. Then it follows from Lukacs' theorem (Lukacs, 1955; see also Mosimann, 1962, p. 75) that $T = \sum_{i=1}^{M} X_i$ is independent of the vector $(P_1, P_2, P_3, \ldots, P_M)$ in which $P_i = X_i/T$ for each i. Hence, it is not possible to give a non-trivial estimate of $T_j$ from any realization $(P_{1j}, P_{2j}, \ldots, P_{Mj})$; the trivial estimate is of course $\hat{T}_j = 1$.

This imaginary example has the property $\mathrm{cov}(T, P_i) = 0$ for each i. If for real data these covariances have values that are not all equal to zero, it is theoretically possible to give an estimate of $T_j$ (see IV.3).

### IV.1.3. The covariance matrices C(X) and C(P)

Since it is practically impossible to establish values for $T_j$ for each j, one aims at describing the relation between C(X), the covariance matrix of the open variables, and C(P), the covariance matrix of the closed variables. Unfortunately, we could not establish the exact relation between these two covariance matrices. The linear approximation (3.4) takes only the three most important terms of the expression

$$(E(X_i) + D_i) \cdot \left(1 + \sum_{p=1}^{\infty} \left(-\sum_{k=1}^{M} D_k\right)^p\right)$$

into account. This expression is moreover only valid for describing

$$P_i = (E(X_i) + D_i)/\left(1 + \sum_{k=1}^{M} D_k\right)$$

if there is no realization of the sum of the $D_k = X_k - E(X_k)$ over all k that has a value equal to or greater than one. If one takes more terms of the expression into account than has been done in (3.4) this results in better approximations, which are too complex to be handled, however. The expression (3.4)

$$P_i \approx E(X_i) + (1 - E(X_i)) \cdot D_i - E(X_i) \cdot \sum_{k \neq i}^{M} D_k$$

is regarded as if it were a perfect equality. As a consequence we can write

$$E(P_i) = E(X_i) = p_i \qquad (3.5)$$

(do not confuse this $p_i$ with the "multinomial" $p_i$ of (2.11))
and

$$P_i - p_i = (1 - p_i) \cdot (X_i - p_i) - p_i \cdot \sum_{k \neq i}^{M} (X_k - p_k) \tag{4.1}$$

In matrix notation this is

$$y = P\,x$$

in which y is a column vector with i-th coordinate $y_i = P_i - p_i$, x is a column vector with i-th coordinate $x_i = D_i = X_i - p_i$, and P is a matrix having in the j-th row and the k-th column:

$$P_{jk} = \begin{cases} (1 - p_j) & \text{if } j = k \\ -p_j & \text{if } j \neq k \end{cases}$$

The validity of this basic linear equation as an approximation is hard to judge theoretically. We consider the linear approximation to be rather a serious problem, but we did not succeed in finding any other, better solution.

Denoting the transposed of vector y and the transposed of matrix P as y' and P', respectively, we deduce from (4.1) that

$$\Phi = E(yy') = E(Pxx'P') = P\,E(xx')\,P' = P\,\Sigma\,P' \tag{4.2}$$

in which

$$\Phi_{kl} = E((P_k - p_k) \cdot (P_l - p_l) = \mathrm{cov}(P_k, P_l) \text{ for } k \neq l,$$

$$\Phi_{kk} = E((P_k - p_k)^2) = \mathrm{var}(P_k),$$

$$\Sigma_{kl} = E((X_k - p_k) \cdot (X_l - p_l)) = \mathrm{cov}(X_k, X_l) \text{ for } k \neq l,$$

$$\Sigma_{kk} = E((X_k - p_k)^2) = \mathrm{var}(X_k).$$

See Kork (1977). $\Phi$ and $\Sigma$ are the closed covariance matrix C(P) and the open covariance matrix C(X), respectively.

Kork (1977) states that the equation $\Phi = P\,\Sigma\,P'$, in which P and $\Phi$ are known matrices to be deduced directly from the closed data, has such a large collection of solutions $\Sigma$ that a number of "equally likely" solutions can always be indicated, which are even mutually highly different. As a consequence Kork concludes that it is impossible to solve the "closed sum problem". Miesch (1969) arrived at the same conclusion following a similar procedure, but he only considered the case M = 4.

We do not agree with Kork that it would be absolutely impossible to solve the closed sum problem. We consider most of Kork's examples of solutions of open covariance matrices from imaginary closed data to be too extreme. His solutions commonly lead to open correlation coefficient matrices that contain many extremely high coefficients (high in an absolute sense). In our opinion the solutions $\Sigma$ of a set of closed data are not of equal value, because in each collection of open covariance matrices there is a subcollection that contains solutions that lead to open correlation coefficient matrices containing more moderate values.

Our own procedures, which we shall describe in this chapter, are meant to lead to "best" solutions of this kind. We are not suggesting there is only one good solution $\Sigma$, but we postulate that all solutions leading to moderate open correlation coefficients can be considered to be close to the "best" solution. The best solution cannot be easily calculated. The Fortran computer program called BALANC does the calculations.

In our procedure we try to solve equation (4.2)

$$\Phi = P \Sigma P',$$

accepting the fact that we shall find more than one open covariance matrix $\Sigma$ as solutions. As in the case of the zero open covariances model (see (3.16), (3.17) and (3.18)), solving (4.2) with the help of matrix algebra manipulations, as does Kork (1977), is neither the shortest nor the most transparent way. Instead, we prefer to start by writing (4.2) in extenso:

$$\begin{aligned} \operatorname{cov}(P_i, P_k) = \operatorname{cov}(X_i, X_k) &- p_k \cdot \left( \sum_{h=1}^{M} \operatorname{cov}(X_i, X_h) \right) - \\ &- p_i \cdot \left( \sum_{h=1}^{M} \operatorname{cov}(X_k, X_h) \right) + \\ &+ p_i \cdot p_k \cdot \left( \sum_{h=1}^{M} \left( \sum_{l=1}^{M} \operatorname{cov}(X_h, X_l) \right) \right) \end{aligned} \tag{4.3}$$

and

$$\begin{aligned} \operatorname{var}(P_i) = \operatorname{var}(X_i) &- 2 p_i \cdot \left( \sum_{h=1}^{M} \operatorname{cov}(X_i, X_h) \right) + \\ &+ p_i^2 \cdot \left( \sum_{h=1}^{M} \left( \sum_{l=1}^{M} \operatorname{cov}(X_h, X_l) \right) \right) \end{aligned} \tag{4.4}$$

We note that $\mathrm{cov}(Y, Y)$ is identical to $\mathrm{var}(Y)$.

Recalling that $T = \sum_{h=1}^{M} X_h$, we have

$$\mathrm{cov}(X_i, T) = \mathrm{cov}(X_i, \sum_{h=1}^{M} X_h) = \sum_{h=1}^{M} \mathrm{cov}(X_i, X_h),$$

and

$$\mathrm{var}(T) = \mathrm{cov}(\sum_{h=1}^{M} X_h, \sum_{l=1}^{M} X_l) = \sum_{h=1}^{M} \left( \sum_{l=1}^{M} \mathrm{cov}(X_h, X_l) \right)$$

so (4.3) and (4.4) can be reduced to

$$\mathrm{cov}(P_i, P_k) = \mathrm{cov}(X_i, X_k) - p_k \cdot \mathrm{cov}(X_i, T) - p_i \cdot \mathrm{cov}(X_k, T) + \\ + p_i \cdot p_k \cdot \mathrm{var}(T) \tag{4.5}$$

and

$$\mathrm{var}(P_i) = \mathrm{var}(X_i) - 2p_i \cdot \mathrm{cov}(X_i \cdot T) + p_i^2 \cdot \mathrm{var}(T) \tag{4.6}$$

In the above expressions the (co)variances of the closed variables $P_i$ are expressed in the (co)variances of the open variables $X_i$ and of their sum T. With the help of an equality that is easily deduced from (4.1):

$$\mathrm{cov}(P_i, T) = \mathrm{cov}(X_i, T) - p_i \cdot \mathrm{var}(T) \tag{4.7}$$

(4.5) and (4.6) can be written in such a way that the (co)variances of the open variables $X_i$ are expressed in the (co)variances of the closed variables $P_i$ and of the "open sum" T:

$$\mathrm{cov}(X_i, X_k) = \mathrm{cov}(P_i, P_k) + p_k \cdot \mathrm{cov}(P_i, T) + p_i \, \mathrm{cov}(P_k, T) + \\ + p_i \cdot p_k \cdot \mathrm{var}(T) \tag{4.8}$$

and

$$\mathrm{var}(X_i) = \mathrm{var}(P_i) + 2p_i \cdot \mathrm{cov}(P_i \cdot T) + p_i^2 \cdot \mathrm{var}(T) \tag{4.9}$$

From these expressions it is clear that the open covariance matrix $C(X)$ is determined if values are added to the closed data for the parameters $\mathrm{var}(T)$, $\mathrm{cov}(P_1, T)$, $\mathrm{cov}(P_2, T)$, . . ., $\mathrm{cov}(P_M, T)$. Because

$$\sum_{i=1}^{M} \mathrm{cov}(P_i, T) = \mathrm{cov}(1, T) = 0,$$

it appears that M values must be "freely" chosen in order to determine $C(X)$. We return to these M "degrees of freedom" later on.

### IV.1.4. Direct derivation of the approximations (4.5) to (4.9)

The approximations (4.5), (4.6), (4.7), (4.8) and (4.9) can be derived directly from the definitions

$$X_i = T \cdot P_i; \quad \sum_{i=1}^{M} P_i = 1; \quad E(T) = 1$$

which gives some insight into the neglect of higher order terms. Denoting $E(P_i) = p_i$; $y_i = P_i - E(P_i) = P_i - p_i$; $t = T - E(T) = T - 1$ (note that $p_i$ is a constant, t a random variable), we get

$$X_i = T \cdot P_i = p_i + p_i \cdot t + y_i + y_i \cdot t$$

so that

$$E(X_i) = p_i + E(y_i \cdot t) = p_i + \mathrm{cov}(P_i, T) \tag{4.10}$$

It follows that

$$\begin{aligned} \mathrm{cov}(X_i, X_k) = E((X_i - E(X_i)) \cdot (X_k - E(X_k))) = p_i \cdot p_k \cdot E(t^2) + \\ + p_k \cdot E(y_i \cdot t) + p_i \cdot E(y_k \cdot t) + E(y_i \cdot y_k) + p_i \cdot E(y_k \cdot t^2) + \\ + p_k \cdot E(y_i \cdot t^2) + 2E(y_i \cdot y_k \cdot t) + E(y_i \cdot y_k \cdot t^2) - \\ - E(y_i \cdot t) \cdot E(y_k \cdot t) \end{aligned} \tag{4.11}$$

Neglecting the five latter terms of higher order we arrive at the approximation (4.8).

Similarly,

$$\begin{aligned} \mathrm{var}(X_i) = E((X_i - E(X_i))^2) = p_i^2 \cdot E(t^2) + 2p_i \cdot E(y_i \cdot t) + E(y_i^2) + \\ + 2p_i \cdot E(y_i \cdot t^2) + 2E(y_i^2 \cdot t) + E(y_i^2 \cdot t^2) - \\ - E(y_i \cdot t) \cdot E(y_i \cdot t) \end{aligned} \tag{4.12}$$

Neglecting the four latter terms of higher order we arrive at the approximation (4.9). Neglect of these terms is based on the fact that $\mathrm{var}(T) = E(t^2)$ is considered to be much less than one.

Next,

$$\begin{aligned} \mathrm{cov}(X_i, T) = E((p_i \cdot t + y_i + y_i \cdot t - E(y_i \cdot t)) \cdot t) = p_i \cdot E(t^2) + \\ + E(y_i \cdot t) + E(y_i \cdot t^2) \end{aligned} \tag{4.13}$$

Neglecting the last term of higher order we get

$\mathrm{cov}(X_i, T) = \mathrm{cov}(P_i, T) + p_i \cdot \mathrm{var}(T)$ of (4.7).

Substituting

$$p_i \cdot p_k \cdot E(t^2) + p_k \cdot E(y_i \cdot t) + p_k \cdot E(y_i \cdot t^2) = p_k \cdot \mathrm{cov}(X_i, T) \text{ and}$$
$$p_i \cdot p_k \cdot E(t^2) + p_i \cdot E(y_k \cdot t) + p_i \cdot E(y_k \cdot t^2) = p_i \cdot \mathrm{cov}(X_k, T)$$

into (4.11), we get

$$\begin{aligned}\mathrm{cov}(X_i, X_k) = &- p_i \cdot p_k \,\mathrm{var}(T) + p_k \cdot \mathrm{cov}(X_i, T) + p_i \cdot \mathrm{cov}(X_k, T) + \\ &+ \mathrm{cov}(P_i, P_k) + 2E(y_i \cdot y_k \cdot t) + E(y_i \cdot y_k \cdot t^2) - \\ &- E(y_i \cdot t) \cdot E(y_k \cdot t)\end{aligned} \tag{4.14}$$

and substituting

$$2p_i^2 \cdot E(t^2) + 2p_i \cdot E(y_i \cdot t) + 2p_i \cdot E(y_i \cdot t^2) = 2p_i \cdot \mathrm{cov}(X_i, T)$$

into (4.12), we have

$$\begin{aligned}\mathrm{var}(X_i) = &- p_i^2 \cdot \mathrm{var}(T) + \mathrm{var}(P_i) + 2p_i \cdot \mathrm{cov}(X_i, T) + 2E(y_i^2 \cdot t) + \\ &+ E(y_i^2 \cdot t^2) - E(y_i \cdot t) \cdot E(y_i \cdot t)\end{aligned} \tag{4.15}$$

Neglecting the last three terms of higher order in (4.14) and in (4.15), we obtain the approximations (4.5) and (4.6), respectively.

### IV.1.5. The acceptability of the hypothesis of Chayes and Kruskal

With the help of the formulae (4.5) and (4.6) one can explain why it is quite permissible to test the hypothesis $R(X_i, X_k) = 0$ by means of the statistic

$$(z(\bar{R}(P_i, P_k)) - z(R_c(P_i, P_k))) \cdot \sqrt{N - 3} \text{ if } M \geqslant 8$$

(see section III.3).

Suppose that in an open covariance matrix $C(X)$ the value of $\mathrm{cov}(X_1, X_2)$ is changed to an amount d. Then the coefficients of the closed covariance matrix $C(P)$ change as follows, according to (4.5) and (4.6):

$$\begin{aligned}
d(\mathrm{cov}(P_1, P_2)) &= ((1 - p_1) \cdot (1 - p_2) + p_1 \cdot p_2) \cdot d \\
d(\mathrm{var}(P_1)) &= -2p_1 \cdot (1 - p_1) \cdot d \\
d(\mathrm{var}(P_2)) &= -2p_2 \cdot (1 - p_2) \cdot d \\
d(\mathrm{var}(P_i)) &= +2p_i^2 \cdot d && (i \geqslant 3) \\
d(\mathrm{cov}(P_1, P_i)) &= -p_i \cdot (1 - 2p_1) \cdot d \\
d(\mathrm{cov}(P_2, P_i)) &= -p_i \cdot (1 - 2p_2) \cdot d \\
d(\mathrm{cov}(P_i, P_k)) &= +p_i \cdot p_k \cdot d && (k > i \geqslant 3)
\end{aligned}$$

It appears that the change d in $\mathrm{cov}(X_1, X_2)$ induces a change of similar size in the corresponding closed covariance $\mathrm{cov}(P_1, P_2)$, if it is supposed that all $p_i$ are much larger than one. The change in $\mathrm{var}(P_1)$, $\mathrm{var}(P_2)$ and all $\mathrm{cov}(P_1, P_i)$ and $\mathrm{cov}(P_2, P_i)$ is smaller, and the change in $\mathrm{var}(P_i)$ and in $\mathrm{cov}(P_i, P_k)$ for $k > i \geqslant 3$ may be considered negligible.

After careful consideration we conclude that deviations of the values of $\bar{R}(P_i, P_k)$ from the expected values $R_c(P_i, P_k)$ are due to deviations of the corresponding $R(X_i, X_k)$ from zero.

### IV.1.6. The values of $T_j$

One might choose the open covariance matrix C(X) in such a way that it has the property

$$\sum_{i=1}^{M} \mathrm{var}(X_i) \text{ has the smallest possible value.}$$

According to (4.9) this implies that

$$\sum_{i=1}^{M} 2p_i \cdot \mathrm{cov}(P_i, T) + p_i^2 \cdot \mathrm{var}(T) = \mathrm{cov}\left(2 \left( \sum_{i=1}^{M} p_i \cdot P_i + p_i^2 \cdot T \right), T\right)$$

has the smallest possible value. It follows that

$$T = 2 - \left( \sum_{i=1}^{M} p_i \cdot P_i \right) \Big/ \left( \sum_{i=1}^{M} p_i^2 \right),$$

which permits us to calculate $T_j$ for each assemblage j directly from the closed data matrix, and then each $X_{ij}$ by $X_{ij} = T_j \cdot \hat{p}_{ij}$. At first sight this seems to be an easy way to solve the closed sum problem. However, choosing the matrix C(X) with the property that the sum of the open variances is at its minimum leads to an extreme solution. With the linear approximation of (4.1) it is easily proved that

$$X_i = P_i + p_i \cdot (T - 1) = P_i + p_i \cdot \left(1 - \left( \sum_{i=1}^{M} p_i \cdot P_i \right) \Big/ \left( \sum_{i=1}^{M} p_i^2 \right)\right)$$

so that (approximately)

$$\sum_{i=1}^{M} p_i \cdot X_i = \sum_{i=1}^{M} p_i^2.$$

This means that the sum of $p_i \cdot X_i$ over all i is approximately a constant. This implies that in this case C(X) is a singular matrix.

From the micropaleontological point of view the consequence that the sum over all i of the open variables $X_i$ multiplied by some constant is constant is completely illogical. The criterion we introduced that the solution with

$$\sum_{i=1}^{M} \text{var}(X_i) \text{ has the smallest possible value}$$

leads to an undesired result.

If we choose as an alternative the open covariance matrix C(X) to have the property

$$\sum_{i=1}^{M} c_i \cdot \text{var}(X_i) \text{ has the smallest possible value,}$$

in which each $c_i$ is a constant greater than zero, we obtain a solution with the properties

1) $$T = 2 - \left( \sum_{i=1}^{M} c_i \cdot p_i \cdot P_i \right) \Big/ \left( \sum_{i=1}^{M} c_i \cdot p_i^2 \right)$$

2) $$\sum_{i=1}^{M} c_i \cdot p_i \cdot X_i = \sum_{i=1}^{M} c_i \cdot p_i^2 \text{ (which is a constant)}$$

3) C(X) is a singular matrix.

A special case is $c_i = 1/p_i$: the open covariance matrix C(X) that has the property

$$\sum_{i=1}^{M} \text{var}(X_i)/p_i \text{ has the smallest possible value,}$$

appears to be the closed covariance matrix C(P) itself: $T = 1$; $X_i = P_i$ for each i. The system of closed variables itself is an extreme and illogical open variables solution because it has the property

$$\sum_{i=1}^{M} X_i = T = \sum_{i=1}^{M} P_i = 1$$

It is important to note that these procedures that fix the series ($T_1$, $T_2$, $T_3$, . . ., $T_N$) result in a singular matrix C(X), the singularity of C(X) being equivalent to the property that there are constants c, $a_1$, $a_2$, $a_3$, . . ., $a_M$, so that

$$\sum_{i=1}^{M} a_i \cdot X_i = c$$

which property is again illogical from the micropaleontological point of view.

We are therefore interested only in non-singular solutions of C(X). After the expressions (4.8) and (4.9) it was mentioned already that C(X) is determined by adding only M values to the closed system, assuming that the linear approximation (4.1) can be used. If the resulting solution for C(X) is non-singular, there must be still (N – M – 1) degrees of freedom for the series ($T_1$, $T_2$, $T_3$, . . ., $T_N$). Remember that one degree of freedom is lost because $\sum_{j=1}^{N} T_j = N$ must hold.

As in our range charts and distribution charts N is generally greater than (M + 1), the choice of a non-singular C(X) does not fix the value of $T_j$ for each count j.

From these considerations and from the imaginary example at the beginning of this section IV.1 that showed that it may be completely impossible to trace back the value of each $T_j$ from the closed data, we have decided not to aim at finding solutions for the series ($T_1$, $T_2$, $T_3$, . . ., $T_N$), but to aim at finding solutions for the open covariances matrix C(X).

## IV.2. The relation between the open and the closed covariance matrices

The expressions (4.5) and (4.6) are not yet the most convenient forms for solving the equation (4.2). Both for the open variables $X_i$ and for the closed variables $P_i$ the squares of the coefficients of variability and the coefficients of covariability are introduced:

$$v_{ii} \equiv \mathrm{var}(P_i)/(p_i^2) \qquad v_{ik} \equiv \mathrm{cov}(P_i, P_k)/(p_i \cdot p_k)$$
$$V_{ii} \equiv \mathrm{var}(X_i)/(p_i^2) \qquad V_{ik} \equiv \mathrm{cov}(X_i, X_k)/(p_i \cdot p_k) \tag{4.16}$$

Because of the risk of confusion we do not use the symbols c and C of the zero open covariances model (see (3.20), (3.22) and (3.41): $c_{ik}$ and $C_{ii}$ are parameters which were defined on the assumption that the zero open covariances model is valid.

The expressions (4.5) and (4.6) can now be changed into

$$v_{ik} = V_{ik} - \frac{\mathrm{cov}(X_i, T)}{p_i} - \frac{\mathrm{cov}(X_k, T)}{p_k} + \mathrm{var}(T) \tag{4.17}$$

$$v_{ii} = V_{ii} - 2 \cdot \frac{\mathrm{cov}(X_i, T)}{p_i} + \mathrm{var}(T) \tag{4.18}$$

by dividing them by $p_i \cdot p_k$ and by $p_i^2$, respectively.

Ignoring any limitation for a while, we are allowed to assign freely some value to

$$d_i = 2 \cdot \frac{\text{cov}(X_i, T)}{p_i} - \text{var}(T). \tag{4.19}$$

Such values being chosen for each parameter $d_i$, i = 1, 2, 3, . . ., M, the solution can be written as

$$V_{ii} = v_{ii} + d_i \text{ or } \text{var}(X_i) = \text{var}(P_i) + p_i^2 \cdot d_i \tag{4.20}$$

$$V_{ik} = v_{ik} + \left( \frac{d_i + d_k}{2} \right) \text{ or } \text{cov}(X_i, X_k) = \text{cov}(P_i, P_k) + \\ + p_i \cdot p_k \cdot \left( \frac{d_i + d_k}{2} \right) \tag{4.21}$$

The statistics of the open variables can now be determined from the statistics of the closed variables, as soon as values have been assigned to the series of parameters $d_1, d_2, d_3, \ldots, d_M$.

As in the zero open covariances model we can substitute for the "closed" statistics either the unweighted forms $\overline{P}_i$ of (2.30), $\overline{\text{var}}(P_i)$ of (2.32) and $\overline{\text{cov}}(P_i, P_k)$ of (2.35) or the weighted forms $\hat{P}_i$ of (2.30), $\hat{\text{var}}(P_i)$ of (2.33) and $\hat{\text{cov}}(P_i, P_k)$ of (2.36). In the computer program BALANC the unweighted forms are substituted.

Another point worth noting is that one may think it is strange that the equation $\Phi = P \Sigma P'$ of (4.2) does not lead to a unique solution of the open covariance matrix $\Sigma$. The matrix $\Sigma$ contains $M \cdot (M + 1)/2$ mutually unrelated values, because $\Sigma_{ik} = \text{cov}(X_i, X_k) = \text{cov}(X_k, X_i) = \Sigma_{ki}$. In the singular closed covariance matrix $\Phi$, however, there are relations between the coefficients $\Phi_{ik}$, because for each i:

$$\sum_{k=1}^{M} \Phi_{ik} = \sum_{k=1}^{M} \text{cov}(P_i, P_k) = \text{cov}(P_i, \sum_{k=1}^{M} P_k) = \text{cov}(P_i, 1) = 0.$$

Hence, on the $M \cdot (M + 1)/2$ values that form $\Phi$, M restrictions are imposed, which implies that $\Phi$ is determined by $M \cdot (M - 1)/2$ mutually unrelated values, so that $\Sigma$ cannot be deduced from $\Phi$, unless M values, such as our $d_1, d_2, d_3, \ldots, d_M$ are added as "extra" information.

## IV.3. Properties of the $d_i$-parameters

The statement in the previous section that any value can be assigned to

the parameters $d_i$, i = 1, 2, 3, . . ., M, is obviously not true. Primarily, $V_{ii}$ or $\mathrm{var}(X_i)$ of (4.20) must be positive, so

$$d_i > -v_{ii} \text{ for each taxon i.} \tag{4.22}$$

Next, for each pair of taxa i and k the correlation coefficient between the open variables $X_i$ and $X_k$,

$$R(X_i, X_k) = \frac{\mathrm{cov}(X_i, X_k)}{\sqrt{\mathrm{var}(X_i) \cdot \mathrm{var}(X_k)}} = \frac{V_{ik}}{\sqrt{V_{ii} \cdot V_{kk}}} =$$

$$= \frac{v_{ik} + ((d_i + d_k)/2)}{\sqrt{(v_{ii} + d_i) \cdot (v_{kk} + d_k)}} \tag{4.23}$$

must fulfil the condition: $-1 \leqslant R(X_i, X_k) \leqslant +1$. The restrictions thus placed on the series $(d_1, d_2, d_3, \ldots, d_M)$ by the requirement that the absolute value of each of the $M \cdot (M-1)/2$ "open" correlation coefficients should be less than one, cannot be easily deduced from (4.23).

Another question to be answered is whether the open covariance matrix C(X) and the open correlation coefficient matrix R(X) are positive definite. We did not find a simple answer to this question, but we think that this failure is not important in practice.

An important equality can now be derived. From (4.19) we can deduce that

$$\sum_{i=1}^{M} p_i \cdot d_i = \sum_{i=1}^{M} \left( 2 \cdot \mathrm{cov}(X_i, T) - p_i \cdot \mathrm{var}(T) \right) =$$

$$= 2 \cdot \mathrm{cov}\left( \left( \sum_{i=1}^{M} X_i \right), T \right) - \left( \sum_{i=1}^{M} p_i \right) \cdot \mathrm{var}(T) = 2 \cdot \mathrm{var}(T) -$$

$$- \mathrm{var}(T) = \mathrm{var}(T),$$

so

$$\mathrm{var}(T) = \sum_{i=1}^{M} p_i \cdot d_i, \tag{4.24}$$

which means that the variance of the sum T of the open variables is equal to the weighted mean of the $d_i$, weighted according to the proportions $p_i$. A consequence is that $\sum_{i=1}^{M} p_i \cdot d_i$ must be greater than zero.

Formula (4.24) shows a striking resemblance to the formula (3.29) in chapter III concerning the zero open covariances model:

$$S_{tt} = \sum_{i=1}^{M} p_i \cdot H_i.$$

This resemblance is entirely misleading, however. On the assumption that the zero open covariances model is valid, we have the following relation, using (4.19):

$$d_i = 2 \cdot \frac{\operatorname{cov}(X_i, \sum_{k=1}^{M} X_k)}{p_i} - \operatorname{var}(T) =$$

$$= 2 \cdot \frac{\operatorname{var}(X_i)}{p_i} - \operatorname{var}(T) = 2 \cdot H_i - S_{tt} \qquad (4.25)$$

because $\operatorname{cov}(X_i, X_k) = 0$ for $k \neq i$. Hence, $d_i = H_i$ holds for each taxon i, if and only if $d_i = H_i = S_{tt}$ for all taxa, and the latter statement is exactly the condition for the "generalized" multinomial model (3.30).

In practice it appears to be very common for some $d_i$ to have negative values. A negative $H_i$, however, would immediately lead to the rejection of the zero open covariances hypothesis. According to (4.20) a negative $d_i$ means that the open variance $\operatorname{var}(X_i)$ is less than the closed variance $\operatorname{var}(P_i)$ for that specific taxon i; this phenomenon is quite admissible.

As soon as values have been substituted for $d_1, d_2, d_3, \ldots, d_M$, $\operatorname{var}(T)$ and for each i, $\operatorname{cov}(P_i, T)$ can be calculated according to (4.24), (4.20) and (4.9), so that the matrix

$$C^* \begin{pmatrix} \operatorname{var}(P_1) & \operatorname{cov}(P_1, P_2) & \cdots & \operatorname{cov}(P_1, P_M) & \operatorname{cov}(P_1, T) \\ \operatorname{cov}(P_2, P_1) & \operatorname{var}(P_2) & \cdots & \operatorname{cov}(P_2, P_M) & \operatorname{cov}(P_2, T) \\ \vdots & \vdots & & \vdots & \vdots \\ \operatorname{cov}(P_M, P_1) & \operatorname{cov}(P_M, P_2) & \cdots & \operatorname{var}(P_M) & \operatorname{cov}(P_M, T) \\ \operatorname{cov}(T, P_1) & \operatorname{cov}(T, P_2) & \cdots & \operatorname{cov}(T, P_M) & \operatorname{var}(T) \end{pmatrix}$$

can be established. It follows that an estimate $d\hat{T}_j = \hat{T}_j - 1$ can be made for each count j by means of multiple regression analysis from the series of values of $dP_{ij} = \hat{p}_{ij} - p_i$, for $i = 1, 2, \ldots, M$. If x is the column vector with coefficients $dP_{1j}, dP_{2j}, \ldots, dP_{Mj}$, y, then the multiple regression estimate $d\hat{T}_j$ is given the value y for which

$x' (C^*)^{-1} x$ has the smallest possible value.

It is to be noted that $\hat{T}_j$ may be a very poor estimate of $T_j$; more precisely, $\operatorname{var}(\hat{T}_j)$ may be much less than $\operatorname{var}(T_j)$. In the imaginary example at the be-

ginning of this section we have $\mathrm{cov}(T, P_i) = 0$ for each i, so for this example it follows that $d\hat{T}_j = 0$ and $\hat{T}_j = 1$, irrespective of the values for $dP_{1j}$, $dP_{2j}$, . . ., $dP_{Mj}$, and obviously $\mathrm{var}(\hat{T}_j) = 0$.

Although it is theoretically possible to calculate these $\hat{T}_j$, we decided not to perform this difficult calculation. We restricted ourselves to analysing the open covariance matrix C(X) and its related correlation coefficient matrix R(X).

## IV.4. Analysis of the open correlation coefficient formula

As we deduced in the previous section, the correlation coefficient between the open variables $X_i$ and $X_k$ is given by the formula (4.23):

$$R(X_i, X_k) = \frac{v_{ik} + ((d_i + d_k)/2)}{\sqrt{(v_{ii} + d_i) \cdot (v_{kk} + d_k)}}$$

This statistic is a function of the parameters $d_i$ and $d_k$, which can be "freely" chosen. Therefore it is necessary to analyse the structure of the function $(d_i, d_k) \mapsto R(X_i, X_k)$.

Substituting $(V_{ii} - v_{ii})$ for $d_i$ and $(V_{kk} - v_{kk})$ for $d_k$ in (4.23), we have according to (4.20)

$$R(X_i, X_k) = \frac{v_{ik} + ((V_{ii} - v_{ii} + V_{kk} - v_{kk})/2)}{\sqrt{V_{ii} \cdot V_{kk}}} =$$

$$= \frac{V_{ii} + V_{kk} - (v_{ii} + v_{kk} - 2 \cdot v_{ik})}{\sqrt{V_{ii} \cdot V_{kk}}}$$

The term $h_{ik} \equiv (v_{ii} + v_{kk} - 2 \cdot v_{ik})$ can be seen as the variance of the closed variable

$$\left( \frac{P_i}{p_i} - \frac{P_k}{p_k} \right).$$

Hence, the term $h_{ik}$, which is determined by the closed data, cannot have a negative value. It is noted that $h_{ik} = h_{ki}$, and that $i \neq k$. In addition to the definition of

$$h_{ik} \equiv v_{ii} + v_{kk} - 2 \cdot v_{ik} \tag{4.26}$$

we define

$$y_{ik} \equiv \frac{V_{ii}}{h_{ik}} = \frac{v_{ii} + d_i}{v_{ii} + v_{kk} - 2 \cdot v_{ik}} \tag{4.27}$$

(note that $y_{ik}$ and $y_{ki}$ are not identical)

and

$$g(x, y) \equiv \frac{x + y - 1}{2 \cdot \sqrt{x \cdot y}} \tag{4.28}$$

Then the open correlation coefficient statistic can be written:

$$R(X_i, X_k) = \frac{V_{ii} + V_{kk} - h_{ik}}{\sqrt{V_{ii} \cdot V_{kk}}} = g\left(\frac{V_{ii}}{h_{ik}}, \frac{V_{kk}}{h_{ik}}\right) = g(y_{ik}, y_{ki}) \tag{4.29}$$

It thus appears that the problem of analysing $R(X_i, X_k)$ can be reduced to that of analysing the function $(x, y) \mapsto g(x, y)$, in which x and y must be positive numbers, because $V_{ii} = v_{ii} + d_i$ and $V_{kk} = v_{kk} + d_k$, and therefore also $y_{ik}$ and $y_{ki}$ must be greater than zero.

Primarily, we wish to restrict the domain of the function g to

$$\{(x, y) \mid x > 0,\ y > 0, \mid g(x, y) \mid \leqslant 1\}$$

The requirement that $|(x + y - 1)/(2 \cdot \sqrt{x \cdot y})| \leqslant 1$ is equivalent to $x^2 - 2xy + y^2 - 2x - 2y + 1 \leqslant 0$. Hence the domain of g, which is illustrated in figure 5 as the white part in the first quadrant of the (x, y)-plane, has as its edge the parabola

$$x^2 - 2xy + y^2 - 2x - 2y + 1 = 0.$$

The graph of $g(x, y) = 0$ is obviously the straight line segment $x + y - 1 = 0$ within the domain g. The part of the domain of g in which $g(x, y)$ is negative is the bounded field at the lower left side of this line segment. The part in which $g(x, y)$ is positive, at the other side of the line segment, is not bounded.

The graph of $g(x, y) = \pm t$, i.e. $|g(x, y)| = t$, in which t is some number between zero and one, appears to be the ellipse

$$x^2 + (2 - 4t^2)\, xy + y^2 - 2x - 2y + 1 = 0.$$

The ellipses $g(x, y) = \pm \frac{1}{4} = \pm 0.25$, $g(x, y) = \pm \frac{1}{4}\sqrt{2} = \pm 0.354$, $g(x, y) = \pm \frac{1}{2} = \pm 0.5$ and $g(x, y) = \pm \frac{1}{2}\sqrt{2} = \pm 0.707$ have been drawn in figure 5.

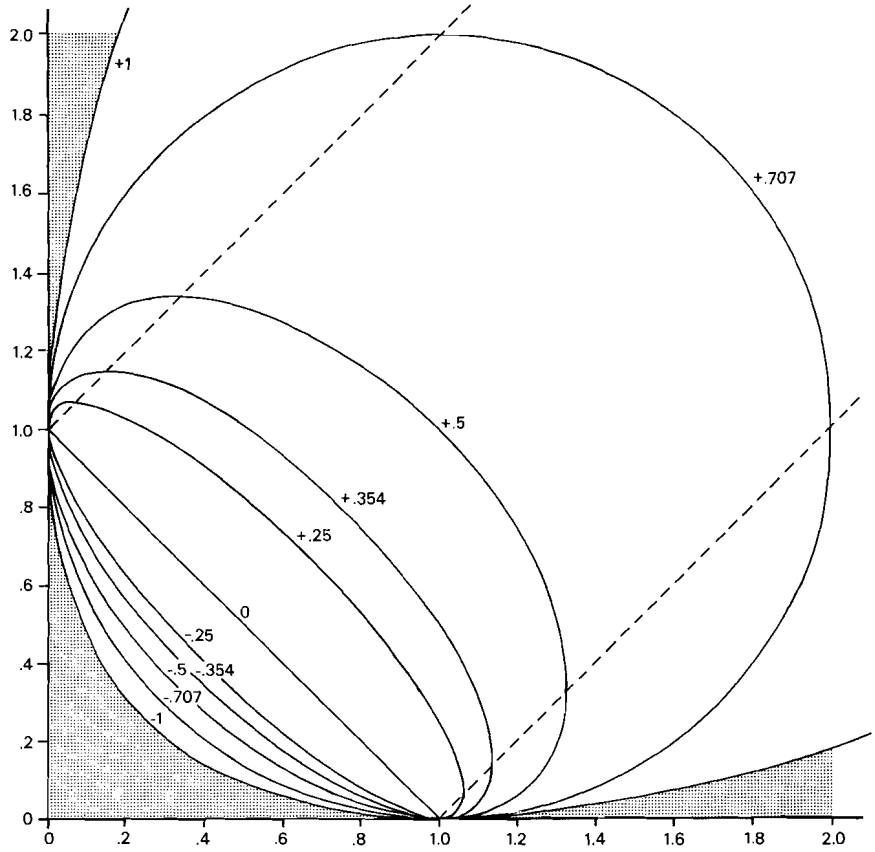



Fig. 5 Graph of $g(x, y) = (x + y - 1)/(2\sqrt{x \cdot y})$ (see (4.28)).

The last mentioned ellipse is in fact the circle $x^2 + y^2 - 2x - 2y + 1 = (x - 1)^2 + (y - 1)^2 - 1^2 = 0$, which has centre $(x = 1, y = 1)$ and radius $r = 1$.

A remarkable feature of this function g is that the two points $(x = 0, y = 1)$ and $(x = 1, y = 0)$ are the only points of the edge of the domain of g that do not belong to the domain. In words, the behaviour of g is unstable in the vicinity of these two points. This means that $R(X_i, X_k)$ is unstable in the case where $V_{ii}$ is immensely large relative to $V_{kk}$, or vice versa, as can be seen from (4.29). In practice such enormous differences in the values of the "open" squares of the coefficients of variability are rare.

A desirable property of the function g would be that $g(x, y)$ increased with increasing x, y remaining constant, and with increasing y, while x remained constant (see next section). Because of the symmetry of g in x and y,

one has only to check whether $\frac{d}{dx}(g(x,y)) > 0$ everywhere in the domain of g. It is easily shown that this derivative is negative in the part of the domain where $x < y - 1$. Due to the symmetry, $\frac{d}{dy}(g(x,y))$ is negative in that part of the domain where $y < x - 1$. The conclusion is that g has the desirable property described above only in that part of the domain where

$$| x - y | = | y_{ik} - y_{ki} | = | (V_{ii} - V_{kk})/h_{ik} | < 1 \qquad (4.30)$$

The edges of this part of the domain are parts of the following graphs: the parabola $x^2 - 2xy + y^2 - 2x - 2y + 1 = 0$, the straight line $x - y + 1 = 0$ and the straight line $x - y - 1 = 0$. These two lines have been added in figure 5.

The analysis presented above will be used in the following section.

### IV.5. A balanced solution "$\Sigma$ R = 0" of the open variables problem

In the introduction to this chapter we mentioned that among the open covariance matrices $\Sigma$ that are solutions of

$$P \Sigma P' = \Phi$$

a "best" solution can be indicated. By "best" solution we mean that all solutions that do not lead to open correlation coefficient matrices with extreme values do not deviate much from that "best" solution. Instead of the matrix expression (4.2) the expressions (4.20) and (4.21) are applied:

$$V_{ii} = v_{ii} + d_i; \quad V_{ik} = v_{ik} + \left(\frac{d_i + d_k}{2}\right)$$

A trivial solution is obtained if $d_i = 0$ is substituted for all i. Then $V_{ii} = v_{ii}$, so that $\mathrm{var}(X_i) = \mathrm{var}(P_i)$ for each taxon i, and $V_{ik} = v_{ik}$, so that $\mathrm{cov}(X_i, X_k) = \mathrm{cov}(P_i, P_k)$ for each pair of taxa i and k. This means that the statistics of the open variables are equal to the statistics of the corresponding closed variables. This is not a "likely" solution, however, because according to (4.24) this would imply that $\mathrm{var}(T) = 0$, i.e. the variance of the sum of the open variables is zero, which means that this sum is a constant from assemblage to assemblage, whereas its components $X_i$ are not.

In matrix algebra we mentioned that $\Sigma = \Phi$ is a solution of (4.2), so $P \Phi P' = \Phi$. Equation (4.2) therefore can be rewritten by stating $D \equiv (\Sigma - \Phi)$, which means that we are asking for the solutions D of

$$PDP' = P(\Sigma - \Phi)P' = P \Sigma P' - P \Phi P' = \Phi - \Phi = 0, \qquad (4.31)$$

in which 0 is the matrix consisting of zeros. Writing equation (4.31) $P\,D\,P' = 0$ in extenso leads to relations, which the coefficients $D_{ii} = \mathrm{var}(X_i) - \mathrm{var}(P_i)$ and $D_{ik} = \mathrm{cov}(X_i, X_k) - \mathrm{cov}(P_i, P_k)$ must satisfy. Obviously this leads to (4.20) and (4.21).

The examples of solutions from imaginary closed sum data given by Kork (1977) generally contain a large number of correlation coefficients $R(X_i, X_k)$ that have extreme values, of positive sign as well as of negative sign. It is remarkable, however, that the majority of these extreme coefficients have the same sign. This implies that for almost every taxon i, the sum

$$\sum_{k \neq i}^{M} R(X_i, X_k) \tag{4.32}$$

must be "very far" from zero. Evidently, taxon i "has" positive correlations with the majority of the other taxa, and there are hardly any taxa with which taxon i "has" a negative correlation, or vice versa. Such a situation may be real, but we do not consider it to be likely.

We wish to "improve" the solution by bringing the sum (4.32) closer and closer to zero for each taxon i. This balancing procedure is the basic work performed by the Fortran computer program BALANC. The idea is as follows.

Suppose that the sum (4.32) relating to taxon i is much greater than zero. Then we decrease the value of $d_i$ belonging to that solution by a small amount E. According to the text around (4.30) in the previous section, we hope that in carrying out this procedure for each taxon k, $k \neq i$, $R(X_i, X_k)$ will decrease as well, so that also the sum (4.32) will decrease. If $R(X_i, X_k)$ fulfils the condition (4.30), it is certain that each $R(X_i, X_k)$ will decrease, and as a consequence the sum. However, if one taxon, or more than one, does not satisfy (4.30), we suppose (but cannot prove) that

$$S_i : d_i \mapsto \sum_{k \neq i}^{M} R(X_i, X_k) \tag{4.33}$$

is still a monotonously increasing function in its domain. This function and its domain are dependent on all $d_k$, $k \neq i$, according to (4.23) and (4.29); these parameters are kept fixed for the moment.

Similarly, if

$$S_i(d_i) = \sum_{k \neq i}^{M} R(X_i, X_k)$$

is less than zero, then $d_i$ is increased by a small amount E, so that $S_i(d_i)$ is likely to increase as well, and to approach zero.

In the computer program BALANC an iterative process is thus performed. It starts by giving each $d_i$ the value zero: $d_1 = d_2 = d_3 = \ldots = d_M = 0$. At first $S_1(d_1)$ is calculated, and

if $S_1(d_1) \geqslant 0$, $d_1$ gets the value $-E$ in addition;
if $S_1(d_1) < 0$, $d_1$ gets the value $+E$ in addition,

in which E is the small amount of our iterative process. Similarly $d_2$ changes in value by the amount $\pm$ E, depending on the sign of $S_2(d_2)$, $d_3$ changes, and so on until $d_M$ is modified. This marks the end of the first iteration step. The second iteration step consists of calculating $S_i(d_i)$ and subtracting E from the value of $d_i$ if $S_i(d_i) \geqslant 0$ and adding E to the value of $d_i$ if $S_i(d_i) < 0$, for $i = 1, 2, 3, \ldots, M$.

After a number of iteration steps of this kind (in our experience this number is between 100 and 250, but of course the number is dependent on the size of the small number E) it is expected that an equilibrium will be reached: for each taxon i, $S_i(d_i)$ has reached zero and during the remaining number of iteration steps $d_i$ will stay close to a limit value $\hat{d}_i$. According to (4.20) and (4.21) this series $(\hat{d}_1, \hat{d}_2, \hat{d}_3, \ldots, \hat{d}_M)$ leads to a "balanced" solution, for which

$$S_i(\hat{d}_i) = \sum_{k \neq i}^{M} R(X_i, X_k) = \sum_{k \neq i}^{M} g\left(\frac{v_{ii} + \hat{d}_i}{h_{ik}}, \frac{v_{kk} + \hat{d}_k}{h_{ik}}\right) = 0 \qquad (4.34)$$

for each taxon i (see (4.29)).

It is emphasized that we did not succeed in finding a solid theoretical basis for this procedure. From our experience with the BALANC program we can say that the procedure never failed in connection with our counts of microfossils. Nevertheless, one should realize that there may be theoretical objections. In our opinion they are:

First, the property (4.34)

$$\sum_{k \neq i}^{M} R(X_i, X_k) = 0$$

for each taxon i cannot be easily defended if the number of taxa is small: $M < 6$. These cases will be dealt with in the following section.

Second, each modification of $d_i$ during the iteration process not only affects $S_i(d_i)$, but also all other $S_k(d_k)$, $k \neq i$ according to the note after

(4.33). Hence, the convergence to zero of $S_i$ might be prohibited by the changes in the values of the $d_k$, $k \neq i$.

Furthermore we did not prove that $d_i \mapsto S_i(d_i)$ is a monotonously increasing function, all $d_k$, $k \neq i$, being kept constant. We cannot even prove that there is a value d in the domain of $S_i$ that has $S_i(d) = 0$.

For the latter two reasons we cannot guarantee that the balanced solution as described above really does exist. If the computer program BALANC gives a balanced solution, one has to check whether each $V_{ii}$ or $\text{var}(X_i)$ is positive and whether for each open correlation coefficient it holds that $| R(X_i, X_k) | \leqslant 1$. In our analysis of a large number of range charts the procedure never failed and a correct balanced solution was always found. However, this may be due to the fact that we never had fewer than eight taxa in our range charts. We have no experience with smaller numbers of taxa.

For each pair of taxa i and k the BALANC program produces the value of $h_{ik}$ from (4.26), so that the expressions $x = y_{ik} = V_{ii}/h_{ik}$, $y = y_{ki} = V_{kk}/h_{ik}$ and $(x - y) = (V_{ii} - V_{kk})/h_{ik}$ can easily be calculated, if desired. In our analyses the absolute values of $(x - y)$ of the correlation coefficients $R(X_i, X_k)$ were nearly always less than one. Therefore the statement that the $S_i$ monotonously increase in $d_i$ looks fairly safe.

It should be borne in mind that all this theory is based on (3.5) and (3.6) which in fact are approximations. It is impossible to estimate whether these approximations have serious consequences for the final result.

Finally, inspection of the output of the BALANC program does not tell us whether the balanced solution $(\hat{d}_1, \hat{d}_2, \hat{d}_3, \ldots, \hat{d}_M)$ is unique for the set of closed data considered (for $M = 2$ it is not unique, as is shown in the next section). The expressions $S_i(\hat{d}_i) = 0$ (4.34) for each i form a set of M equations that the M variables $\hat{d}_i$ must fulfil. So we expect (but cannot prove) that there is only one possible solution. The BALANC program finds a solution by starting from $d_1 = d_2 = d_3 = \ldots d_M = 0$, but the $d_i$ might converge to another limit if we let the iteration process start from a different series $d_i = x_i$, $i = 1, 2, 3, \ldots, M$ (the $x_i$ of course may be mutually different). A theoretical answer to this question has not been found. In practice we examined one set of closed data in connection with this problem by substituting several different series of $d_i$ values at the beginning of the iteration process in the BALANC program. It always gave the same solution, but this fact can by no means be accepted as a proof of the uniqueness of the balanced solution.

## IV.6. Properties of the balanced solution $\Sigma R = 0$ for $M \leqslant 6$

In this section we shall discuss a property of these balanced solutions that gives rise to doubt about the justification of such so-called best solutions, when the number M of taxa is small. We still hope to convince the reader about this justification when this number M is large. For the sake of convenience in notation the open correlation coefficient $R(X_i, X_k)$ is written in this section as $R_{i,k}$.

For M = 2, the case of two taxa, (4.34) simply reads

$$R_{1,2} = 0 \tag{4.35}$$

and this is true if $\hat{d}_1 + \hat{d}_2 = -2 \cdot v_{1,2}$, as can be seen from (4.23). Hence, there is more than one balanced solution. If one used the BALANC program, one would get the solution $\hat{d}_1 = \hat{d}_2 = -v_{1,2}$.

For M = 3, the case of three taxa, (4.34) results in

$$R_{1,2} + R_{1,3} = 0;\ R_{1,2} + R_{2,3} = 0;\ R_{1,3} + R_{2,3} = 0$$

and this implies:

$$R_{1,2} = R_{1,3} - R_{2,3} = 0 \tag{4.36}$$

Hence, the requirement (4.34) of the free open covariances model is identical to the requirement (3.7) of the zero open covariances model of Chayes and Kruskal, and therefore the solutions are identical if M = 3. From (4.36) it follows that

$$\hat{d}_1 + \hat{d}_2 = -2 \cdot v_{1,2};\ \hat{d}_1 + \hat{d}_3 = -2 \cdot v_{1,3};\ \hat{d}_2 + \hat{d}_3 = -2 \cdot v_{2,3}$$

so that

$$\begin{aligned} \hat{d}_1 &= v_{2,3} - v_{1,2} - v_{1,3} \\ \hat{d}_2 &= v_{1,3} - v_{1,2} - v_{2,3} \\ \hat{d}_3 &= v_{1,2} - v_{1,3} - v_{2,3} \end{aligned} \tag{4.37}$$

In this way the open squares of the coefficients of variability acquire a remarkable property. For taxon 1 for instance

$$\begin{aligned} V_{1,1} &= v_{1,1} + \hat{d}_1 = v_{1,1} + v_{2,3} - v_{1,2} - v_{1,3} = \\ &= E\left( \left( \frac{P_1}{p_1} - \frac{P_2}{p_2} \right) \cdot \left( \frac{P_1}{p_1} - \frac{P_3}{p_3} \right) \right) \end{aligned} \tag{4.38}$$

given that $R_{1,2} = R_{1,3} = R_{2,3} = 0$, M = 3.

Before the case M = 4 is considered an important property of the balanced solution will be derived, which is valid for M greater than or equal to four.

Let A be a subcollection of the collection of M taxa. From (4.34) we can write

$$\sum_{i \in A} \left( \sum_{k \neq i}^{M} R_{i,k} \right) = 0 \text{ and this can be written}$$

$$2 \cdot \sum_{h \in A} \left( \sum_{\substack{k \in A \\ k > h}} R_{h,k} \right) = \sum_{i \in A} \left( \sum_{l \notin A} R_{i,l} \right) \tag{4.39}$$

A similar reasoning can be used for the complement of A, so

$$2 \cdot \sum_{h \notin A} \left( \sum_{\substack{k \notin A \\ k > h}} R_{h,k} \right) = \sum_{i \notin A} \left( \sum_{l \in A} R_{i,l} \right) \tag{4.40}$$

As the expressions on the right-hand sides of (4.39) and (4.40) are identical, it follows that

$$\sum_{h \in A} \left( \sum_{\substack{k \in A \\ k > h}} R_{h,k} \right) = \sum_{i \notin A} \left( \sum_{\substack{l \notin A \\ l > i}} R_{i,l} \right) \tag{4.41}$$

In words: the sum of the correlation coefficients between all pairs of taxa in a subcollection A is equal to the sum of the correlation coefficients between all pairs of taxa in the complement of A. If the first sum is far from zero, then the second sum must be too. This is an artificial property due to the imposed condition (4.34). For small M this property has undesirable consequences (see below). It is to be noted that A as well as its complement must consist of at least two taxa. That is why for (4.41) to be valid M must be at least four.

For M = 4, (4.41) results in

$$R_{1,2} = R_{3,4}; \; R_{1,3} = R_{2,4}; \; R_{1,4} = R_{2,3} \tag{4.42}$$

for A = {1, 2}, {1, 3} and {1, 4}, respectively. Hence, it is impossible to recognize single correlations between open variables if M = 4. A real correlation between $X_1$ and $X_2$ may be recognized to some extent, but this correlation is copied in the correlation between $X_3$ and $X_4$.

For M = 5, (4.41) results for instance in

$$R_{1,2} = R_{3,4} + R_{3,5} + R_{4,5} \tag{4.43}$$

taking A = {1, 2}. Again a real correlation between $X_1$ and $X_2$ may be

recognized, but this correlation induces a bias in $R_{3,4}$, $R_{3,5}$ and $R_{4,5}$. If there are more "real" correlations, the picture may become disturbed.

For M = 6, (4.41) results in expressions like

$$R_{1,2} = R_{3,4} + R_{3,5} + R_{3,6} + R_{4,5} + R_{4,6} + R_{5,6} \tag{4.44}$$

if A consists of two taxa, and in expressions like

$$R_{1,2} + R_{1,3} + R_{2,3} = R_{4,5} + R_{4,6} + R_{5,6} \tag{4.45}$$

if A, and therefore also its complement, consist of three taxa. For M = 6 the "mutual induction" as expressed by (4.41), will generally be too weak to disturb the picture in such a way that individual "real" correlations between open variables can no longer be recognized. In our opinion therefore the procedure of finding the balanced solution, executed by the Fortran program BALANC, can be used if the number of taxa is greater than or equal to six.

## IV.7. The balanced solution "Σ RdR = 0"

In the previous sections we considered the balanced solution "Σ R = 0", i.e. the solution which consists of the open covariance matrix C(X) and the associated open correlation coefficient matrix, which has

$$\sum_{k \neq i}^{M} R(X_i, X_k) = 0 \text{ for each taxon i.}$$

Another balanced solution, denoted by "Σ RdR = 0", which looks appropriate as well, is

$$\sum_{k \neq i}^{M} R^2(X_i, X_k) \text{ has the smallest possible value for each taxon i} \tag{4.46}$$

because it seems capable of reducing the number of correlation coefficients that strongly deviate from zero. We replace (4.46) by

$$\frac{\delta}{\delta\,(d_i)} \left( \sum_{k \neq i}^{M} R^2(X_i, X_k) \right) =$$

$$= \frac{\delta}{\delta\,(d_i)} \left( \sum_{k \neq i}^{M} \frac{(2v_{ik} + d_i + d_k)^2}{4\,(v_{ii} + d_i) \cdot (v_{kk} + d_k)} \right) = 0 \tag{4.47}$$

for each i, in which all $d_k$, $k \neq i$, are kept constant. See (4.23). The expres-

sion (4.47) can also be written:

$$2 \cdot \sum_{k \neq i}^{M} R(X_i, X_k) \cdot \frac{\delta\, R(X_i, X_k)}{\delta\, (d_i)} = 0$$

indicating that in the sum each correlation coefficient $R(X_i, X_k)$ is weighted according to the factor

$$\frac{\delta\, R(X_i, X_k)}{\delta\, (d_i)}$$

which in general is a positive number. See section IV.4. Recalling $V_{ii}$ of (4.20) and $h_{ik}$ of (4.26), the expression (4.47) is equivalent to

$$\frac{1}{4V_{ii}^2} \cdot \sum_{k \neq i}^{M} \frac{V_{ii}^2 - (V_{kk} - h_{ik})^2}{V_{kk}} = 0 \text{ for each i.} \qquad (4.48)$$

The expression on the left-hand side is a monotonously increasing function in $V_{ii}$, so also in $d_i$, unless $V_{kk} = h_{ik}$ would hold for each $k \neq i$. It should be remembered that $V_{ii} = v_{ii} + d_i$; $V_{kk} = v_{kk} + d_k$; $h_{ik} = v_{ii} + v_{kk} - 2\, v_{ik}$.

The computer program BALANC can be modified easily so that instead of

$$S_i\,(d_i) = \sum_{k \neq i}^{M} R(X_i, X_k)$$

the expression

$$T_i\,(d_i) = \frac{1}{4V_{ii}^2} \cdot \sum_{k \neq i}^{M} \frac{V_{ii}^2 - (V_{kk} - h_{ik})^2}{V_{kk}}$$

for each i is considered. The iteration process remains the same. From the values of $d_i$ some constant E is subtracted if $T_i\,(d_i) \geqslant 0$, and the constant E is added to the value of $d_i$ if $T_i\,(d_i) < 0$, for i = 1, 2, 3, . . ., M, repeatedly, until an equilibrium is reached in which each $d_i$ will stay in the vicinity of a limit value $\overset{*}{d}_i$, for which $T_i\,(\overset{*}{d}_i) = 0$ holds.

A solid theoretical basis for this procedure is again lacking. It cannot be proved that a "balanced solution" will be found. We only know that this procedure works in practice. Every solution presented has to be checked for errors such as negative open variances and open correlation coefficients greater than one or less than minus one. The uniqueness of such a solution has not been proved either. Obviously this procedure is also based on the expressions (3.5) and (3.6) which in fact are approximations. The reader

is referred to section IV.5 for a more elaborate description of the procedure and the problems involved.

Comparing the $\Sigma$ RdR = 0 balanced solution with the $\Sigma$ R = 0 balanced solution for a series of range charts reveals quite striking differences. The limits $\overset{*}{d}_i$ of the $\Sigma$ RdR = 0 solution tend to be larger than the corresponding limits $\hat{d}_i$ of the $\Sigma$ R = 0 solution. The open variances of the $\Sigma$ RdR = 0 solution tend to be larger. More striking is the result that in the $\Sigma$ RdR = 0 solution the majority of the significantly negative correlation coefficients according to the $\Sigma$ R = 0 solution have become non-significant. However, the significantly positive correlation coefficients according to the $\Sigma$ R = 0 solution are confirmed by the $\Sigma$ RdR = 0 solution. The latter solution produces more significantly positive correlation coefficients; the values of the significantly positive correlation coefficients tend to be somewhat greater in the $\Sigma$ RdR = 0 solution than in the $\Sigma$ R = 0 solution.

In our opinion the explanation of the differences in the correlation coefficients can be found in the graph of

$$R(X_i, X_k) = g(x,y) = \frac{x + y - 1}{2 \cdot \sqrt{x \cdot y}}$$

in which $x = V_{ii}/h_{ik}$ and $y = V_{kk}/h_{ik}$. This graph is presented in figure 5. The $\Sigma$ R = 0 solution results in more or less equal numbers of (significantly) positive values and of (significantly) negative values in the open correlation coefficient matrix. Looking at the graph of g(x, y): more or less equal numbers of points fall in the limited region in which $g < 0$ and in the non-bounded region in which $g > 0$. The $\Sigma$ RdR = 0 solution tends to make each $R^2$ as small as possible. If for instance $R < .25$ must hold for each correlation coefficient R, the "chance" of being in the region $0 < R < +.5$ seems to be much larger than the "chance" of being in the region $-.5 < R < 0$, because the latter region is much smaller than the former.

In our practice of micropaleontological data the $\Sigma$ R = 0 solution is preferred because it yields a balance between the numbers of positive and negative values of the open correlation coefficients.

# Chapter V

# TRENDS AND TIME SERIES

## V.1. Introduction

Along a stratigraphical column the proportions of a certain taxon may show an overall increase or an overall decrease; these phenomena are called trends. Considering the succession from bottom to top an increasing trend is called positive, a decreasing trend negative. Trends may be expressed by the correlation coefficient value between the successive proportions and the rank numbers of the samples ordered from bottom to top.

One meets a snag in evaluating these correlation coefficient values. Because one of the series of variables, the sample numbers, has a fixed, logical (stratigraphic) order, we are not really dealing with correlations between independent variables, on which the statistical theory, used so far was based. This logical order of the samples may have an effect on the other series of variables, the proportions. There may be an interdependence in the sense that some taxon property (e.g. its proportion) in a certain sample is strongly related to the one(s) in the sample(s) immediately below. Series with such interdependence because of the inherent logical order (in our range charts the stratigraphic order) are called time series.

Because of the possible dependence of a proportion $\hat{p}_{ij}$ on the previous one, the correlation coefficient $R(\hat{p}_{ij}, \hat{p}_{i(j+1)})$ may yield a distinctly positive value instead of the value zero that one might expect for independent variables. Such dependence in a series is referred to as autocorrelation (see V.4 and V.5).

In an earlier paper (M. M. Drooger, Raju and Doeven, 1979, section 2.8) the problem of illusory correlations arising from the logical stratigraphic order of the samples was already discussed. It should be emphasized that in that paper autocorrelation was considered in the case of the evolutionary *Planorbulinella* lineage, in which we may easily imagine that each *Planorbulinella* population derived its characteristics primarily from the immediately preceding ones. The possibility of autocorrelation is less evident for the successive faunal compositions of the range charts. However, for the range chart data too we can assume that evolution, especially in the sense of evolution of the environment, may have played a role. And it cannot be denied that faunas commonly descended directly from the earlier ones at the same spot. Hence, it seems realistic not to ignore the possible effect of

time series in the interpretation of trends calculated from the range charts.

In the present chapter we intend to give the analysis of time series a somewhat better foundation with the help of several models from probability theory (V.4, V.6). It is emphasized that in our opinion probability models rarely give a good description of reality. Yet these models are thought to give us a better understanding and to prevent us from drawing incorrect conclusions.

Recognition of the possible effect of time series may also have a bearing on the correlation coefficients between the proportions of two taxa. In testing for independence of two taxa series by means of the correlation coefficient values, the statistic (2.41)

$$(z(R) - z(\rho)) \cdot \sqrt{N - 3}$$

has been mentioned. It has a standard normal distribution if the hypothesis resulting in the value $\rho$ is true. This statement becomes inadmissible as soon as autocorrelation is present. In such cases the number of degrees of freedom $(N - 3)$ must be lower (V.6); this has consequences for the significance limits of the established correlation values.

In the closed system of proportions it is evident that the trends of different participating taxa are interdependent. A positive trend in taxon i will tend to induce negative trend effects for all other taxa. In section V.8 interdependence of trends is avoided by using trends in the open variables.

## V.2. The linear regression coefficient

So far we have considered the closed variable $P_i$ for some taxon i as a random variable Y having some probability distribution, while the realizations $Y_1, Y_2, Y_3, \ldots, Y_N$ are mutually independent, i.e.

$$\text{cov}(Y_j, Y_k) = 0 \tag{5.1}$$

for each pair of counts with rank numbers j and k. According to (2.8) it can just as well be said that all $Y_i$ are mutually uncorrelated:

$$r(Y_j, Y_k) = 0 \tag{5.2}$$

An interesting alternative hypothesis considered now is that there is a correlation between $Y_j$ and the rank number j, in other words that there is a trend in Y, starting at $j = 1$ and ending at $j = N$. Such a trend is represented by the linear regression coefficient $Y_j$ on j

$$B(Y) = \frac{\sum_{j=1}^{N} (Y_j - \bar{Y}) \cdot (j - \bar{j})}{\sum_{j=1}^{N} (j - \bar{j})^2} \tag{5.3}$$

in which $\bar{Y}$ is the mean value of $Y_1, Y_2, Y_3, \ldots, Y_N$, and $\bar{j} = (1 + 2 + 3 + \ldots + N)/N = (N + 1)/2$. $B(Y)$ is the slope of the regression line

$$Y - \bar{Y} = B \cdot (j - \bar{j}). \tag{5.4}$$

This straight line is the best fit to the data according to the principle of least squares. It gives for each rank number j an expected value of $Y_j$:

$$E(Y_j) = \bar{Y} + B(Y) \cdot (j - \bar{j}).$$

In the following we have decided to use the difference between the last expected value and the first expected value of the sequence

$$(N - 1) \cdot B(Y) = E(Y_N) - E(Y_1) \tag{5.5}$$

as the trend parameter, instead of $B(Y)$ itself.

After deduction of the fact that

$$\sum_{j=1}^{N} (j - \bar{j})^2 = \frac{(N - 1) \cdot N \cdot (N + 1)}{12}$$

it follows that

$$(N - 1) \cdot B(Y) = \frac{6 \cdot (2 \cdot (\sum_{j=1}^{N} j \cdot Y_j) - (N + 1) \cdot (\sum_{j=1}^{N} Y_j))}{N \cdot (N + 1)} \tag{5.6}$$

On the basis of the hypothesis stated at the beginning of this section, namely that the $Y_j$ are mutually independent random variables, all having the same probability distribution, the actual slope $\beta(Y)$, which is estimated by $B(Y)$, is equal to zero. This hypothesis can be tested using the statistic $B(Y)$, but we prefer to use the correlation coefficient statistic

$$R(j, Y_j) = \frac{\sum_{j=1}^{N} (Y_j - \bar{Y}) \cdot (j - \bar{j})}{\sqrt{\left(\sum_{j=1}^{N} (Y_j - \bar{Y})^2\right) \cdot \left(\sum_{j=1}^{N} (j - \bar{j})^2\right)}} \tag{5.7}$$

which has (approximately) an r-distribution with $(N - 2)$ degrees of freedom

if the hypothesis is true. As will be shown in section V.7, a significant value of $R(j, Y_j)$ need not imply that $\beta(Y)$ is different from zero. In order to prove that the $Y_j$ tend to increase with increasing j or that the $Y_j$ tend to decrease with increasing rank numbers ($\beta(Y)$ positive or negative, respectively), more facts are needed.

## V.3. The linear regression coefficient in the DISTUR program

In the Fortran computer program DISTUR both an unweighted and a weighted form are given for the trend parameter of the proportion $\hat{p}_{ij}$ for each taxon i as a function of j, and for the correlation coefficient between $\hat{p}_{ij}$ and the rank number j.

The unweighted statistic $(N-1) \cdot \overline{B}(\hat{p}_{ij})$, called $(N-1)$ B(PI) in the DISTUR program, is achieved by substituting $\hat{p}_{ii}$ for $Y_i$ in (5.6). The same substitution in (5.7) results in the unweighted correlation coefficient statistic $\overline{R}(j, \hat{p}_{ij})$, called R(J, PI) in the program. $\overline{p}_i$ of (2.26) is substituted for $\overline{Y}$. $\overline{R}(j, \hat{p}_{ij})$ can more easily be calculated from

$$\overline{R}(j, \hat{p}_{ij}) = \frac{2 \cdot \left( \sum_{j=1}^{N} j \cdot \hat{p}_{ij} \right) - (N+1) \cdot \left( \sum_{j=1}^{N} \hat{p}_{ij} \right)}{\sqrt{\frac{N \cdot (N+1)}{3} \cdot \overline{\text{var}}(P_i)}} \tag{5.8}$$

in which $\overline{\text{var}}(P_i)$ is given by (2.32).

The weighted statistic $(N-1) \cdot \hat{B}(\hat{p}_{ij})$ called $(N-1)$ BW(PI) in the DISTUR program, is given by

$$(N-1) \cdot \hat{B}(\hat{p}_{ij}) = (N-1) \cdot \frac{\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{p}_i) \cdot (j - \hat{j})}{\sum_{j=1}^{N} n_j \cdot (j - \hat{j})^2} =$$

$$= (N-1) \cdot \frac{\left( \sum_{j=1}^{N} j \cdot x_{ij} \right) - \hat{p}_i \left( \sum_{j=1}^{N} j \cdot n_j \right)}{\left( \sum_{j=1}^{N} j^2 \cdot n_j \right) - \hat{j} \cdot \left( \sum_{j=1}^{N} j \cdot n_j \right)} \tag{5.9}$$

in which $\hat{p}_i$ is the weighted estimate of (2.25), $\hat{j} \equiv \left( \sum_{j=1}^{N} j \cdot n_j \right) \Big/ \left( \sum_{j=1}^{N} n_j \right)$ and $x_{ij} \equiv n_j \cdot \hat{p}_{ij}$ is the score of taxon i in count j (see 2.3).

The weighted correlation coefficient statistic $\hat{R}(j, \hat{p}_{ij})$ called RW(J, PI) in the program, is given by

$$\hat{R}(j, \hat{p}_{ij}) = \frac{\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{p}_i) \cdot (j - \hat{j})}{\sqrt{\left(\sum_{j=1}^{N} n_j \cdot (j - \hat{j})^2\right) \cdot \left(\sum_{j=1}^{N} n_j \cdot (\hat{p}_{ij} - \hat{p}_i)^2\right)}} =$$

$$= \frac{\left(\sum_{j=1}^{N} j \cdot x_{ij}\right) - \hat{p}_i \cdot \left(\sum_{j=1}^{N} j \cdot n_j\right)}{\sqrt{\left(\left(\sum_{j=1}^{N} j^2 \cdot n_j\right) - \hat{j} \cdot \left(\sum_{j=1}^{N} j \cdot n_j\right)\right) \cdot \left(\left(\sum_{j=1}^{N} \hat{p}_{ij} \cdot x_{ij}\right) - \hat{p}_i \cdot \left(\sum_{j=1}^{N} x_{ij}\right)\right)}} \quad (5.10)$$

In the weighted statistics each count is weighted according to its size $n_j$. In this way $\hat{j}$ also is a weighted mean of the rank numbers according to the sizes $n_j$. Formula (5.9) was already presented by Cochran (1954, p. 434), but it did not include our factor $(N - 1)$.

Finally, the statistics concerning the sizes $n_j$ have to be incorporated. In (3.44), (3.45) and (3.46) we introduced the statistics $\bar{n}$, $\overline{var}(n)$ and $\bar{R}(P_i, n)$, respectively.

The trend parameter $(N - 1) \cdot B(n)$, called R(J, N(J)) in the DISTUR program, is given by substituting $n_j$ for $Y_j$ in (5.6):

$$(N - 1) \cdot B(n) = \frac{6 \cdot \left(2 \cdot \left(\sum_{j=1}^{N} j \cdot n_j\right) - (N + 1) \cdot \left(\sum_{j=1}^{N} n_j\right)\right)}{N \cdot (N + 1)} \quad (5.11)$$

The correlation coefficient statistic $\bar{R}(j, n_j)$, called R(P(I, J), N(J)) in the program, results in

$$\bar{R}(j, n_j) = \frac{\sum_{j=1}^{N} (n_j - \bar{n}) \cdot (j - \bar{j})}{\sqrt{\left(\sum_{j=1}^{N} (n_j - \bar{n})^2\right) \cdot \left(\sum_{j=1}^{N} (j - \bar{j})^2\right)}} =$$

$$= \frac{2 \cdot \left(\sum_{j=1}^{N} j \cdot n_j\right) - (N + 1)\left(\sum_{j=1}^{N} n_j\right)}{\sqrt{\frac{(N - 1) \cdot N \cdot (N + 1)}{3} \cdot \sum_{j=1}^{N} (n_j - \bar{n})^2}} \quad (5.12)$$

## V.4. Stationary random processes

The most simple theoretical model concerning random variables $Y_1$, $Y_2$, $Y_3$, . . ., $Y_N$ that have the same probability distribution and are mutually dependent (time series effect) is defined by

$$Y_j = \sum_{h=0}^{\infty} \sqrt{1 - t^2} \cdot t^h \cdot Z_{j-h} = \sqrt{1 - t^2} \cdot (Z_j + t \cdot Z_{j-1} + t^2 \cdot Z_{j-2} + \ldots) \tag{5.13}$$

The $Z_i$ for any integer value i are random variables having the same probability distribution with expected value $E(Z) = 0$; they are mutually independent. The parameter t is a number between zero and one. Expression (5.13) says that $Y_j$ is determined by the "past" values $Z_{j-h}$, but the influence decreases when the positive integer h increases, because $\sqrt{1 - t^2} \cdot t^h$ approaches zero with increasing h.

From (5.13) it immediately follows that

$$Y_j = t \cdot Y_{j-1} + \sqrt{1 - t^2} \cdot Z_j \tag{5.14}$$

and that

$$E(Y_j) = E(Z_j) = 0; \quad \text{var}(Y_j) = \text{var}(Z_j) = \text{var}(Z) \tag{5.15}$$

In the literature such a stochastic process is called an autoregressive process of the first order, with parameter t (Schwarzacher, 1975). If Z has a normal (Gaussian) distribution with mean $\mu$ and variance $\sigma^2$, then all $Y_j$ have this distribution and the process is known as a Gauss Markov process. Choosing $t = 0$, we have $Y_j = Z_j$, so again a series of mutually independent random variables.

It is possible to substitute for t a value between minus one and zero, but such a procedure results in a quite illogical model in which $Y_j$ and $Y_{j+1}$ are negatively correlated.

Choosing t between 0 and 1 results in correlation between the random variables $Y_j$, called autocorrelation.

In general terms the covariance between $Y_j$ and $Y_{j+h}$,

$$\begin{aligned} \text{cov}(h) \equiv \text{cov}(Y_j, Y_{j+h}) &= E((Y_j - E(Y_j)) \cdot (Y_{j+h} - E(Y_{j+h}))) = \\ &= E((Y_j - E(Y_j)) \cdot (Y_{j+h} - E(Y_j))) \end{aligned} \tag{5.16}$$

is called the autocovariance of Y at lag h (lag in the sense used in probability theory).

In such terms the correlation between $Y_j$ and $Y_{j+h}$ is the autocorrelation of Y at lag h and is given by

$$r(h) \equiv r(Y_j, Y_{j+h}) = \frac{cov(Y_j, Y_{j+h})}{\sqrt{var(Y_j) \cdot var(Y_{j+h})}} = \frac{cov(h)}{var(Y_j)} = \frac{cov(h)}{var(0)} \qquad (5.17)$$

The last expression of (5.16) and the two expressions on the far right in (5.17) are only valid if both $E(Y_j)$ and $var(Y_j)$ are not dependent on j. Such processes are called stationary processes.

Returning to the autoregressive process of the first order (5.13), we deduce from (5.14):

$$cov(h) = t^h \cdot var(Y); \; r(h) = t^h \qquad (5.18)$$

so $r(1) = t$, $r(2) = t^2$, $r(3) = t^3$, etcetera, and $\lim_{h \to \infty} r(h) = 0$. If t is very close to one, the autocorrelation reduces only slowly to zero with increasing lag h, which means that a value $Y_j$ determines for a large part the value of $Y_{j+h}$ even at a large distance h. Substituting $t = 1$ in (5.13) leads to nonsense ($Y_j = 0$); substitution in (5.14) would give the trivial result that all $Y_j$ are mutually equal. Yet one can define the "limit process" (t approaching 1) by replacing (5.13) by

$$Y_j = \sum_{h=1}^{j} Z_h = Y_{j-1} + Z_j \qquad (Y_0 \equiv 0) \qquad (5.19)$$

in which the $Z_h$ for any positive integer value h are random variables with the same probability distribution, which have as expected value $E(Z) = 0$, and are mutually independent, in a similar way as in (5.13). The result is a form of the Wiener process (e.g. Lamperti, 1966), but more common names are Brownian motion process or random walk process.

From (5.19) the following properties are deduced:

$$E(Y_j) = 0; \; var(Y_j) = j \cdot var(Z)$$
$$cov(h) = cov(Y_j, Y_{j+h}) = var(Y_j) = j \cdot var(Z)$$
$$r(h) = r(Y_j, Y_{j+h}) = \sqrt{\frac{j}{j+h}} \qquad (5.20)$$

It appears that $var(Y_j)$ and $r(h)$ are functions of the index j, which means that the random walk process in fact is not stationary, because its properties change with increasing j. The variance of $Y_j$ shows a linear increase with increasing j, whereas $r(h)$ approaches the value one. Especially the

first fact makes the random walk model less suitable for a theoretical description of the behaviour of numerical proportions of microfossil taxa along the stratigraphic column than the autoregressive process of the first order. See section V.7.

Yet another model concerning time series should be mentioned. It is a "signal plus noise" process, discussed by Agterberg (1974, p. 334) in a geological context.

We define c as a fixed number, $0 < c < 1$, and t also as a fixed number between zero and one, but in practice close to one. An autoregressive process of the first order $(Y_j)$ is defined according to (5.13), (5.14) and (5.15), which have this parameter t. In addition, for any integer value j the $U_j$ are random variables with the same probability distribution, which are mutually independent and which are also independent of the $Y_j$ or $Z_j$ series. Only the expected value and the variance of U are equal to the expected value and the variance of Z, respectively:

$$E(U) = E(Z);\ \mathrm{var}(U) = \mathrm{var}(Z) \tag{5.21}$$

The signal plus noise process can then be defined by

$$Y'_j \equiv \sqrt{c} \cdot Y_j + \sqrt{1-c} \cdot U_j \tag{5.22}$$

So it is the sum of an autoregressive process of the first order $(\sqrt{c} \cdot Y_j)$, being the "signal", and a process giving mutually independent random variables $(\sqrt{1-c} \cdot U_j)$, being the "noise". For instance the random counting errors in the $\hat{p}_{ij}$ can be considered as such a noise component $(U_j)$, whereas the $(Y_j)$ form the signal of the more or less autocorrelated series of "real" fluctuations in the numerical proportions.

From (5.15), (5.21) and (5.22) we deduce

$$\begin{aligned} &E(Y') = \sqrt{c} \cdot E(Y) + \sqrt{1-c} \cdot E(U) = 0 \\ &\mathrm{var}(Y') = c \cdot \mathrm{var}(Y) + (1-c) \cdot \mathrm{var}(U) = \mathrm{var}(U) = \mathrm{var}(Z) = \mathrm{var}(Y) \end{aligned} \tag{5.23}$$

The equality $\mathrm{var}(Y') = c \cdot \mathrm{var}(Y) + (1-c) \cdot \mathrm{var}(U)$ means that the variance of Y' can be split into a part due to the "signal" and another part due to the "noise". The ratio of these parts, $c/(1-c)$, is called the signal-noise ratio.

It is easily deduced from (5.16) and (5.17) that

$$\mathrm{cov}(h) \equiv \mathrm{cov}(Y'_j, Y'_{j+h}) = c \cdot \mathrm{var}(Y') \cdot t^h$$

and that

$$r(h) \equiv r(Y'_j, Y'_{j+h}) = c \cdot t^h \quad h \neq 0 \tag{5.24}$$

Some examples for different values of c and t are shown graphically in figure 6.

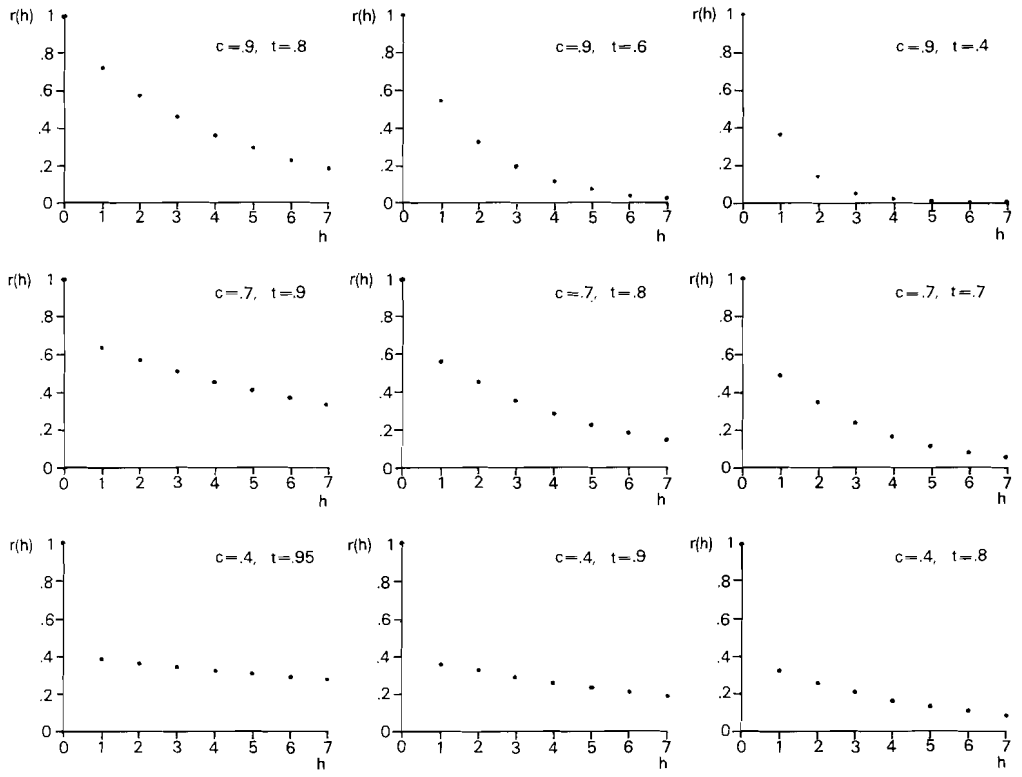



Fig. 6 The autocorrelograms of signal-plus-noise processes for a series of (c, t) combinations (see (5.24)).

It should be noted that according to the definitions (5.16) and (5.17) $\mathrm{cov}(0) = \mathrm{var}(Y_j)$, and $r(0) = 1$ hold for any series $(Y_j)$. Finally we observe that the substitutions $c = 0$ and $c = 1$ do not lead to special results. For $c = 0$ we get $Y_j' = U_j$, for $c = 1$ we get $Y_j' = Y_j$.

Before discussing the usefulness of these two models for our investigation we shall pay attention to the method of estimating autocorrelation in our series $(\hat{P}_{i1}, \hat{P}_{i2}, \hat{P}_{i3}, \ldots, \hat{P}_{iN})$.

## V.5. Estimating autocorrelation

In the literature one finds two ways of calculating the correlation coeffi-

cient between $Y_j$ and $Y_{j+h}$, $j = 1, 2, 3, \ldots, N - h$. Substituting $Y_j = \hat{p}_{ij}$ in our formulae, we shall first present the expressions in a way similar to the approaches based on stationary processes (e.g. Agterberg, 1974).

The expression cov(h) of (5.16) is estimated by

$$C_1(h) \equiv \frac{1}{N-h} \cdot \sum_{j=1}^{N-h} (\hat{p}_{ij} - \bar{p}_i) \cdot (\hat{p}_{i,(j+h)} - \bar{p}_i) =$$

$$= \frac{1}{N-h} \cdot \left[ \left( \sum_{j=1}^{N-h} \hat{p}_{ij} \cdot \hat{p}_{i,(j+h)} \right) - \right.$$

$$\left. - \bar{p}_i \cdot \left( \sum_{j=1}^{N-h} (\hat{p}_{ij} + \hat{p}_{i,(j+h)}) \right) \right] + (\bar{p}_i)^2 \qquad (5.25)$$

in which $\bar{p}_i = \left( \sum_{j=1}^{N} \hat{p}_{ij} \right)/N$ according to (2.26). The expression r(h) of (5.17) is estimated by

$$R_1(h) \equiv C_1(h)/C_1(0) = C_1(h)/\overline{var}_1(P_i) \qquad (5.26)$$

in which $\overline{var}_1(P_i) = \left( \sum_{j=1}^{N} (\hat{p}_{ij} - \bar{p}_i)^2 \right)/N$ differs from the expression $\overline{var}(P_i)$ of (2.32) by a factor $(N - 1)/N$. The assumption connected with these statistics is that the process $\hat{p}_{ij}$ as a function of j is stationary, i.e. that the mean and the variance of $\hat{p}_{ij}$ do not change with increasing j (are independent of time).

This assumption is not needed for the second approach which is used for instance by Schwarzacher (1975). The expression cov(h) is estimated now by

$$C_2(h) \equiv \frac{1}{N-h} \cdot \sum_{j=1}^{N-h} (\hat{p}_{ij} - \bar{p}_i(1)) \cdot (\hat{p}_{i,(j+h)} - \bar{p}_i(2)) \qquad (5.27)$$

in which $\bar{p}_i(1) \equiv \left( \sum_{j=1}^{N-h} \hat{p}_{ij} \right)/(N-h)$, and $\bar{p}_i(2) \equiv \left( \sum_{j=h+1}^{N} \hat{p}_{ij} \right)/(N-h)$. The expression $var(\hat{p}_{ij})$, $j = 1, 2, 3, \ldots, N - h$, is estimated by

$$V(h) \equiv \frac{1}{N-h} \cdot \left( \sum_{j=1}^{N-h} (\hat{p}_{ij} - \bar{p}_i(1))^2 \right)$$

and $var(\hat{p}_{ij})$, $j = h + 1, h + 2, h + 3, \ldots, N$, is estimated by

$$W(h) \equiv \frac{1}{N-h} \cdot \left( \sum_{j=h+1}^{N} (\hat{p}_{ij} - \bar{p}_i(2))^2 \right).$$

These expressions are needed to estimate r(h) of (5.17) by

$$R_2(h) = \frac{C_2(h)}{\sqrt{V(h) \cdot W(h)}} =$$

$$= \frac{(N-h) \cdot \left( \sum_{j=1}^{N-h} \hat{p}_{ij} \cdot \hat{p}_{i,(j+h)} \right) - \left( \sum_{j=1}^{N-h} \hat{p}_{ij} \right) \cdot \left( \sum_{j=1}^{N-h} \hat{p}_{i,(j+h)} \right)}{\sqrt{\left[ (N-h) \cdot \left( \sum_{j=1}^{N-h} \hat{p}_{ij}^2 \right) - \left( \sum_{j=1}^{N-h} \hat{p}_{ij} \right)^2 \right] \cdot \left[ (N-h) \cdot \left( \sum_{j=1}^{N-h} \hat{p}_{i,(j+h)}^2 \right) - \left( \sum_{j=1}^{N-h} \hat{p}_{i,(j+h)} \right)^2 \right]}} \qquad (5.28)$$

Our Fortran computer program called REGRES calculates for each taxon i and for a series of lags h the autocorrelation coefficient values $R_1(h)$ of (5.26) and $R_2(h)$ of (5.28). In addition to a number of features also present in the DISTUR program, such as mean values, variances and trend statistics of the $\hat{p}_{ij}$ as a function of j for each taxon i, the two forms of the autocorrelation coefficient statistic are calculated for each taxon, first for h = 1, next for h = 2, and so on up to a maximum value h = max(h). Obviously the parameter max(h), to be chosen by the user of the program, should be much less than N, the number of counts.

The hypothesis that for some taxon i the $\hat{p}_{ij}$ are random variables with the same probability distribution and are mutually independent implies that r(h) = 0 holds for each positive integer h. We can test this hypothesis by assuming that both $R_1(h)$ and $R_2(h)$ have an r-distribution with about (N − h − 2) degrees of freedom if the hypothesis is true. An r-distribution with x degrees of freedom has an expected value equal to zero, a variance equal to 1/(x + 1), and a distribution closer to a normal distribution, the larger its x.

## V.6. Reduction of the number of degrees of freedom in testing for correlation between two signal plus noise processes

We shall now consider the effect of time series on the correlation between two taxa along the column.

At the beginning of chapter V we mentioned that in testing for correlation between two series $\hat{p}_{ij}$ and $\hat{p}_{kj}$ of two taxa i and k, the statistic

$$(z(R) - z(\rho)) \cdot \sqrt{df}$$

has a standard normal distribution if the hypothesis resulting in the hy-

pothetical value $\rho$ is correct. The number of degrees of freedom, df, need not be equal to $(N - 3)$ however, as was mentioned in (2.41). The value $(N - 3)$ only may be substituted for df, if at least one of the two series $(\hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{iN})$ and $(\hat{p}_{k1}, \hat{p}_{k2}, \ldots, \hat{p}_{kN})$ is a series consisting of mutually independent values. If it is not, df is probably much less than $(N - 3)$.

Below the value of df will be estimated on the assumption that the series $(\hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{iN})$ is a signal plus noise process with parameters $c_1$ and $t_1$ and the series $(\hat{p}_{k1}, \hat{p}_{k2}, \ldots, \hat{p}_{kN})$ is a similar process with parameters $c_2$ and $t_2$. We write $X_j$ for $\hat{p}_{ij}$ and $Y_j$ for $\hat{p}_{kj}$ for the sake of easiness of notation. As a further approximation we disregard the closed sum effect in this time series analysis. The series $(X_j)$ and $(Y_j)$ are considered to be mutually independent. The influence of the closed sum property is believed to be small, except possibly in the case of clearly disturbing proportions of some taxon.

It is permissible to standardize X and Y in such a way that $E(X) = 0$; $\text{var}(X) = 1$; $E(Y) = 0$; $\text{var}(Y) = 1$; and the mutual independence is expressed by $\text{cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y) = E(X \cdot Y) = 0$.

The value of df is strictly related to the variance of the correlation coefficient statistic

$$R(X_j, Y_j) = \frac{\sum_{j=1}^{N} (X_j - \bar{X}) \cdot (Y_j - \bar{Y})}{\sqrt{\left( \sum_{j=1}^{N} (X_j - \bar{X})^2 \right) \cdot \left( \sum_{j=1}^{N} (Y_j - \bar{Y})^2 \right)}}$$

in which $\bar{X} = (\sum_{j=1}^{N} X_j)/N$ and $\bar{Y} = (\sum_{j=1}^{N} Y_j)/N$ are the mean values of the statistical sample. It is assumed that N is quite large (at least 20), and that $N \cdot (1 - t_1)$ and $N \cdot (1 - t_2)$ are both much larger than one, for instance

$$N \geqslant 20; \; N \cdot (1 - t_1) > 10; \; N \cdot (1 - t_2) > 10 \tag{5.29}$$

Then $\bar{X}$ and $\bar{Y}$ are so close to zero that they can be neglected:

$$R(X_j, Y_j) \approx \frac{\sum_{j=1}^{N} X_j \cdot Y_j}{\sqrt{\left( \sum_{j=1}^{N} X_j^2 \right) \cdot \left( \sum_{j=1}^{N} Y_j^2 \right)}}$$

The next step is to replace both factors under the square root sign by N, which is thought to be admissible under the condition (5.29);

$$R(X_j, Y_j) \approx \frac{1}{N} \cdot \sum_{j=1}^{N} X_j \cdot Y_j.$$

For the expression on the right side the variance is derived:

$$\operatorname{var}\left(\frac{1}{N}\cdot\sum_{j=1}^{N} X_j\cdot Y_j\right)=\frac{1}{N^2}\cdot E\left(\left(\sum_{j=1}^{N} X_j\cdot Y_j\right)^2\right)=$$

$$=\frac{1}{N^2}\cdot\sum_{j=1}^{N}\left(\sum_{k=1}^{N} E(X_j\cdot Y_j\cdot X_k\cdot Y_k)\right)=$$

$$=\frac{1}{N^2}\cdot\sum_{j=1}^{N}\left(\sum_{k=1}^{N} E(X_j\cdot X_k)\cdot E(Y_j\cdot Y_k)\right)=$$

$$=\frac{1}{N^2}\cdot\left[N+\sum_{j=1}^{N}\left(\sum_{\substack{k=1\\k\neq j}}^{N} c_1\cdot t_1^{|j-k|}\cdot c_2\cdot t_2^{|j-k|}\right)\right]=$$

$$=\frac{1}{N^2}\cdot\left[N+2\cdot c_1\cdot c_2\left((N-1)\cdot t_1\cdot t_2+(N-2)\cdot t_1^2\cdot t_2^2+\right.\right.$$

$$\left.\left.+(N-3)\cdot t_1^3\cdot t_2^3\cdot\ldots\right)\right]$$

by applying the property (5.24) of signal plus noise processes:

$$\operatorname{cov}(Y'_j, Y'_{j+h})=c\cdot\operatorname{var}(Y')\cdot t^h$$

which results in this special case in:

$$E(X_j\cdot X_{j+h})=c_1\cdot t_1^h \quad\text{and}\quad E(Y_j, Y_{j+h})=c_2\cdot t_2^h.$$

Due to the condition (5.29), it is admissible to approximate

$$\operatorname{var}\left(\frac{1}{N}\cdot\sum_{j=1}^{N} X_j\cdot Y_j\right)\approx$$

$$\approx\frac{1}{N^2}\cdot\left[N+2\cdot c_1\cdot c_2\left(N\cdot t_1\cdot t_2+N\cdot t_1^2\cdot t_2^2+N\cdot t_1^3\cdot t_2^3+\ldots\right)\right]=$$

$$=\frac{1}{N}\cdot\frac{1-t_1\cdot t_2+2\cdot c_1\cdot c_2\cdot t_1\cdot t_2}{1-t_1\cdot t_2}$$

which is an acceptable approximation for $\operatorname{var}(R(X_j, Y_j))$. Hence, the number of degrees of freedom is approximated by:

$$df\approx N\cdot\left(\frac{1-t_1\cdot t_2}{1-t_1\cdot t_2+2\cdot c_1\cdot c_2\cdot t_1\cdot t_2}\right)=$$

$$=N\cdot\left(\frac{1-t_1\cdot t_2}{1+(2\cdot c_1\cdot c_2-1)\cdot t_1\cdot t_2}\right) \qquad (5.30)$$

In the case of autoregressive processes of the first order, we have $c_1 = 1$ and $c_2 = 1$, so that

$$df \approx N \cdot \left( \frac{1 - t_1 \cdot t_2}{1 + t_1 \cdot t_2} \right) \tag{5.31}$$

These results (5.30) and (5.31) are special cases of a formula of Bartlett (see Box & Newbold, 1971), who stated the number of degrees of freedom of the correlation coefficient between two series (X) and (Y) to be

$$df \approx N/(1 + 2 \cdot \left( \sum_{h=1}^{\infty} r_1(h) \cdot r_2(h) \right)) \tag{5.32}$$

in which $r_1(h)$ is the autocorrelation of X at lag h, and $r_2(h)$ the autocorrelation of Y at lag h (see 5.17). This more general result can be derived in a way similar to that presented above.

Formula (5.30) is a good tool for evaluating the correlation coefficient values from the outputs of the DISTUR and BALANC programs, with the help of the output of the REGRES program. The parameters c and t can be estimated from the series R(1), R(2), R(3), . . ., R(max(h)) of any variable $\hat{p}_{ij}$. This procedure prevents us from drawing too far-reaching conclusions about mutual correlations between taxa.

Regarding the correlation between two autoregressive processes of the first order, the number of degrees of freedom is extremely low if the parameters $t_1$ and $t_2$ are both close to one, according to (5.31). The same thing happens for two signal plus noise processes, provided the parameters $c_1$ and $c_2$ are larger than $\frac{1}{2}$. Hence, especially for strongly autocorrelated processes it appears to be impossible to obtain "proof" of mutual correlations. However, if one of the processes (series of proportions) contains random variables with the same probability distribution which are mutually independent (so $c_1 = 0$, or $t_1 = 0$), then there is no reduction in the number of degrees of freedom (N – 3).

## V.7. Discussion

For the series of proportions of each taxon the REGRES program gives a series of autocorrelation coefficient values R(1), R(2), R(3), . . ., R(max(h)). With these data it is possible to check whether the graph of this series (generally named autocorrelogram or correlogram) fits the series $r(h) = c \cdot t^h$ of (5.24) for some combination of values of c and t. See figure 6. If some correspondence is found this does not mean that the series under

consideration is a signal plus noise process with these parameters c and t according to (5.22), but it means that this probabilistic model might give a good description of the actual series. One property of this theoretical series is that $\lim_{h \to \infty} r(h) = \lim_{h \to \infty} c \cdot t^h = 0$, which becomes expressed in a gradual cline (fig. 6). This gradual property is hardly ever found in actual correlograms. Many correlograms show oscillating patterns, something which has given people reason to believe – in our opinion erroneously – that such series have some periodic component.

The fact that correlograms based on real data often do not fit the theoretical correlograms so well, is not a cause of worry. All these probabilistic processes are theoretical constructions. They are said to be "random", i.e. each step $X_j \rightarrow X_{j+1}$ is established by some "cause" (or by more than one cause) with a behaviour which cannot be predicted, but which is thought to be consistent. Such consistency of behaviour is not likely for real data, however.

A significant correlation coefficient value between two series of values implies that the variation in both series may be considered to be (partly) due to the same cause. Even if the correlation coefficient does not turn out to be significant after applying the formula (5.30), the variation in both series may still be partly due to the same cause, but such reasons must not be confused with inferences from mathematical statistics.

It has been noted by M. M. Drooger, Raju and Doeven (1979) that the test for a trend in a series of values according to (5.7) seems to contain an inherent contradiction. A trend implies a dependence of the series of values on the series of rank numbers, and therefore leads to the denial of the mutual independence of the values in the series.

We suggested in that paper – and we still suggest – that a significant correlation coefficient (5.7) should not be taken as implying a trend that must have been caused by a sustained change (sustained evolutionary urge or sustained ecological adaptation). Rejection of the null hypothesis implies that the parameter values in the ordered series have some mutual relation which is expressed as a more or less distinct change in the course of the stratigraphic section. The change or trend may be due either to chance, i.e. a process that has a random nature consisting of successive steps alternating in an incomprehensible way, or to a sustained change of some kind. The cause of the change or trend is never indicated by the correlation coefficient statistic (5.7), but additional reasoning is required if the cause is to be identified.

Below we make some additional comments.

Firstly, it has already been mentioned that $R(j, Y_j)$ of (5.7) has an r-distribution with $(N - 2)$ degrees of freedom if the hypothesis, that the $Y_j$ are mutually independent random variables, all having the same probability distribution, is true.

Taking a signal plus noise process (5.22) with parameters c and t as the null hypothesis, Bartlett's formula (5.32) may be used to determine df, considering j as a variable having $r(h)$ approximately equal to one for each h. We then state with some caution that

$$df = N \cdot \left( \frac{1 - t}{1 + (2 \cdot c - 1) \cdot t} \right) \tag{5.33}$$

is the number of degrees of freedom of the r-distribution of $R(j, Y_j)$ of (5.7).

If (Y) is an autoregressive process of the first order (5.14) with parameter t, then this number is

$$df = N \cdot \left( \frac{1 - t}{1 + t} \right) \tag{5.34}$$

If t approaches one, the autoregressive process of the first order will approach the Wiener process (5.19) more and more closely. From (5.34) it appears that df is extremely small in the case of Wiener processes (random walks). Raup (1977) investigated the behaviour of $R(j, Y_j)$ for Wiener processes and already arrived at the same conclusion. From a table presented in his paper it can be concluded that the distribution of $R(j, Y_j)$ very much resembles the r-distribution with one degree of freedom. A remarkable property of the $R(j, Y_j)$ of Wiener processes is that they must have the same probability distribution irrespective of the size of N, the number of "steps".

Secondly, a requirement in the independent random variables model, in the autoregressive process of the first order, and in the signal plus noise model, is the statement that all variables Y should have the same probability distribution. This would imply that the random variable is not allowed to change its mean value during the succession of realizations $Y_1, Y_2, Y_3, \ldots, Y_N$. Testing for a trend by means of the statistic $R(j, Y_j)$ of (5.7) can only lead to rejection of the probabilistic model, but it can never lead to the acceptance (in statistical sense) that there is a sustained change in the mean value of the variable Y. The alternative hypothesis belonging to the "signal plus noise" hypothesis (the weakest hypothesis among the three mentioned above) is very complex. It can be that the process (Y) changes its nature in the time series, for instance its mean, its variance or its correlogram. For the time being we concluded that such an alternative hypothesis is too difficult to handle in relation to any profit we might obtain.

Finally, we wish to emphasize and repeat that significant correlation coefficient values $R(j, Y_j)$ of (5.7), also those according to the reduced number df of (5.33), having estimated c and t from the correlogram presented by the REGRES program, cannot lead to an automatic conclusion concerning sustained changes. Rejection of the null hypothesis implies no more than that the parameter values in the ordered series have some mutual relation which is expressed as a more or less distinct change in the course of the series.

## V.8. Trends in open variables

At first sight the time series effect seems to be in conflict with the open variables models described in the chapters III and IV. There it was stated that the chart under consideration is formed by a series of open variables $(X_1, X_2, X_3, \ldots, X_M)$ according to the relation (3.2):

$$\hat{p}_{ij} = \frac{X_{ij}}{X_{1j} + X_{2j} + X_{3j} + \ldots + X_{Mj}}$$

The set of N counts is considered as a set of N mutually independent realizations of the series of open variables. However, it is quite feasible that there is an interdependence between successive samples in a range chart for the "total numbers" on which the open variables theory is based. So it is clear that the time factor may not be excluded.

In the literature the open variables concept has been applied in geochemistry to percentages of chemical elements from rock samples, which generally have no time relation. In micropaleontology one can also consider the counts of Recent faunas from samples from an area at the bottom of the sea or from some bay. In such a distribution chart there is no time relation either, nor is there any other a priori logical order. Even in such cases it may be considered doubtful, however, whether the open variables model is realistic. The open variables model should deal with open random variables, which implies that the set of N counts in micropaleontology and the set of N rock samples in geochemistry must be considered as "random" statistical samples from some statistical "population". However, the contents of a certain sample may be thought to resemble more closely ("be dependent on") the contents of the nearest other one than those much farther away. Although there is no real time series, there might be a similar interdependence.

The aim of these introductory sentences is that care should always be taken in applying the open variables concept.

Within the framework of our free open covariances model it is possible to speak of trends in the open variables $X_i$. One can imagine that a taxon increases in its $X_i$ in such a way from bottom to top that the proportions of all other taxa in the counts must decrease. In such a case one expects to find a clear trend in the open variable of the "disturbing" taxon, but the trends of all remaining open variables should be independent of this trend.

The mutual dependence of the trends in the closed variables is easily demonstrated by means of the linear regression coefficient (5.3):

$$B(Y) \equiv \frac{\sum_{j=1}^{N} (Y_j - \bar{Y}) \cdot (j - \bar{j})}{\sum_{j=1}^{N} (j - \bar{j})^2}$$

because

$$\sum_{i=1}^{M} B(\hat{p}_{ij}) = B(\sum_{i=1}^{M} \hat{p}_{ij}) = B(1) = 0 \qquad (5.35)$$

It is noted here that all properties to be deduced in this section are valid not only for the unweighted statistic $\bar{B}$ $(\hat{p}_{ij})$, but also for the weighted form of the linear regression coefficient, $\hat{B}$ $(\hat{p}_{ij})$, given in (5.9). For practical reasons the computer program BALANC only incorporates the unweighted form.

Recalling the notation $T_j$ for the sum of the open variables – see (3.3) – we have

$$\sum_{i=1}^{M} B(X_{ij}) = B(\sum_{i=1}^{M} X_{ij}) = B(T_j) \qquad (5.36)$$

From the basic relation (3.6) between the open variables and the closed variables,

$$P_i - p_i \approx (1 - p_i) \cdot (X_i - p_i) - p_i \cdot \sum_{k \neq i}^{M} (X_k - p_k)$$

it is deduced in a similar way (the sum of the regression coefficients is equal to the regression coefficient of the sum) that

$$B(\hat{p}_{ij}) = (1 - p_i) \cdot B(X_{ij}) - p_i \cdot \sum_{k \neq i}^{M} B(X_{kj})$$

which is written, incorporating (5.36), as

$$B(\hat{p}_{ij}) = B(X_{ij}) - p_i \cdot B(T_j) \qquad (5.37)$$

or

$$B(X_{ij}) = B(\hat{p}_{ij}) + p_i \cdot B(T_j)$$

Regarding (5.35), (5.36) and (5.37), it appears that the regression coefficient of the sum of the open variables, $B(T_j)$, may be chosen freely, within certain limits of course. Each choice of a value of $B(T_j)$ leads by means of (5.37) to a series of values of $B(X_{1j})$, $B(X_{2j})$, . . ., $B(X_{Mj})$. Just as in the case where values were assigned to the open correlation coefficients $R(X_i, X_k)$ of (4.23), we have to make a "best" choice. For the "open trends", however, the finding of such a best choice is much less difficult.

According to (5.7), the correlation coefficient between the rank number j and the open variable $X_{ij}$ is

$$R(j, X_{ij}) = \frac{\sum_{j=1}^{N} (X_{ij} - p_i) \cdot (j - \bar{j})}{\sqrt{\left( \sum_{j=1}^{N} (X_{ij} - p_i)^2 \right) \cdot \left( \sum_{j=1}^{N} (j - \bar{j})^2 \right)}} \tag{5.38}$$

Recalling the equality

$$\sum_{j=1}^{N} (j - \bar{j})^2 = \frac{(N-1) \cdot N \cdot (N+1)}{12}$$

we deduce from (5.3) and (5.7) the approximation

$$R(j, X_{ij}) \approx \frac{(N-1) \cdot B(X_{ij})}{\sqrt{12 \cdot \mathrm{var}(X_{ij})}} \tag{5.39}$$

which gives the relation between the linear regression coefficient $B(X_{ij})$ and the correlation coefficient $R(j, X_{ij})$. From (5.37) it follows that

$$R(j, X_{ij}) \approx \frac{N-1}{\sqrt{12}} \cdot \frac{B(\hat{p}_{ij}) + p_i \cdot B(T_j)}{\sqrt{\mathrm{var}(X_{ij})}} \tag{5.40}$$

After choosing some value for $B(T_j)$, $R(j, X_{ij})$ can be calculated, because $B(\hat{p}_{ij})$ has been calculated directly from the closed data, and $\mathrm{var}(X_{ij})$ is given by the free open covariances procedure. Each choice of the value of $B(T_j)$ leads to a series of values for $R(j, X_{1j})$, $R(j, X_{2j})$, . . . and $R(j, X_{Mj})$. We choose a value for $B(T_j)$ that leads to a "balanced" solution, for which

$$\sum_{i=1}^{M} R(j, X_{ij}) = 0 \tag{5.41}$$

The considerations leading to (5.41) are identical to those that led to the balanced solution for the open correlation coefficients $R(X_i, X_k)$ of (4.23)

(see (4.34)). In contrast to the troublesome derivation of the latter coefficients, the derivation of the value of $B(T_j)$ for which (5.41) holds, is quite simple. From (5.40) it follows that

$$\sum_{i=1}^{M} \left( \frac{B(\hat{p}_{ij})}{\sqrt{var(X_{ij})}} + B(T_j) \cdot \frac{p_i}{\sqrt{var(X_{ij})}} \right) = 0$$

so that

$$B(T_j) = -\left( \sum_{i=1}^{M} \frac{B(\hat{p}_{ij})}{\sqrt{var(X_{ij})}} \right) \Big/ \left( \sum_{i=1}^{M} \frac{p_i}{\sqrt{var(X_{ij})}} \right) \tag{5.42}$$

It appears that there is a unique solution $B(T_j)$. The solution may be incorrect, however, because it cannot be guaranteed that there is no absolute value of $R(j, X_{ij})$ exceeding one for some taxon i. Just as in the balancing procedure of (4.34) this procedure makes more sense the larger the number of taxa (M).

Our Fortran computer program BALANC contains this balanced solution. The unweighted statistics $\overline{p}_i$ of (2.26) and $\overline{B}(\hat{p}_{ij})$ are substituted for $p_i$ and $B(\hat{p}_{ij})$ in (5.42), while $V_{ii} \equiv var(X_{ij})/(\overline{p}_i^2) = v_{ii} + \hat{d}_i$ is given by balancing the correlation coefficients $R(X_i, X_k)$ according to (4.34). Hence,

$$(N-1) \cdot \overline{B}(T_j) = -\left( \sum_{i=1}^{M} \frac{(N-1) \cdot \overline{B}(\hat{p}_{ij})}{\overline{p}_i \cdot \sqrt{V_{ii}}} \right) \Big/ \left( \sum_{i=1}^{M} (V_{ii})^{-\frac{1}{2}} \right) \tag{5.43}$$

Next,

$$\overline{R}(j, T_j) = \frac{(N-1) \cdot \overline{B}(T_j)}{\sqrt{12 \cdot var(T_j)}} \tag{5.44}$$

is calculated in the computer program BALANC, in which $var(T_j) = \sum_{i=1}^{M} \overline{p}_i \cdot \hat{d}_i$ according to (4.24). Next $(N-1) \cdot \overline{B}(X_{ij})$ is calculated for each taxon i according to (5.37), substituting $\overline{B}(\hat{p}_{ij})$ and $\overline{B}(T_j)$ of (5.43). Finally $\overline{R}(j, X_{ij})$ is calculated for each taxon i according to (5.40), by means of

$$\overline{R}(j, X_{ij}) = \frac{N-1}{\sqrt{12 \cdot V_{ii}}} \cdot \left( \frac{\overline{B}(\hat{p}_{ij})}{\overline{p}_i} + \overline{B}(T_j) \right) \tag{5.45}$$

These calculations in the BALANC program have been somewhat simplified by defining for each taxon i

$$\overline{K}(\hat{p}_{ij}) = \frac{(N-1) \cdot \overline{B}(\hat{p}_{ij})}{\overline{p}_i} \tag{5.46}$$

which may be seen as a relative trend parameter, because the overall change along the column, $(N-1)\cdot\bar{B}(\hat{p}_{ij})$, is weighted according to the mean value $\bar{p}_i$. The relation between $\bar{K}(\hat{p}_{ij})$ and $\bar{B}(\hat{p}_{ij})$ is comparable to that between $V_{ii}$ and $\mathrm{var}(X_{ij})$.

The formula (5.43) changes into

$$(N-1)\cdot\bar{B}(T_j) = -\left(\sum_{i=1}^{M}\bar{K}(\hat{p}_{ij})\cdot(V_{ii})^{-\frac{1}{2}}\right)\Big/\left(\sum_{i=1}^{M}(V_{ii})^{-\frac{1}{2}}\right) \qquad (5.47)$$

and the formula (5.45) changes into

$$\bar{R}(j,X_{ij}) = (12\cdot V_{ii})^{-\frac{1}{2}}\cdot(\bar{K}(\hat{p}_{ij}) + (N-1)\cdot\bar{B}(T_j)) \qquad (5.48)$$

# Chapter VI

## PRINCIPAL COMPONENTS ANALYSIS OF THE OPEN VARIABLES

### VI.1. Introduction

Techniques of multivariate analysis are frequently applied in many scientific disciplines, also in micropaleontology. Restricting ourselves to R-mode analysis, i.e. the analysis of the relations between the investigated variables (in our case the variables $P_i$ or $X_i$ for each taxon i), multivariate analysis aims at finding meaningful axes in the multidimensional space defined by the variables. These axes must give a sufficient description of the cluster of points in that space, each point representing a "realization" of the series of variables. Such axes, which account for a large part of the variation, may lead to the recognition of groups of variables (taxa). Within each group the variables have a more or less similar behaviour. Many techniques of multivariate analysis directly aim at such a grouping and make use of a coefficient of similarity, or a coefficient of association (the correlation coefficient is one). These cluster analyses are "numerical methods" that have hardly any foundation in the theory of mathematical statistics. Usually they lack such a foundation completely. In our opinion cluster analyses can be used only as computer techniques to give rapid insight into multivariate sets of data from a visual impression by means of dendrograms. The consistency of the groups of variables should be tested by other means, however.

In the literature the interpretation of results from principal components analyses (often also named factor analyses), i.e. the interpretation of the derived axes, often contains several mistakes. Authors often ignore the fact that each axis that has been given some specific meaning must primarily be meaningful in a statistical sense. In other words, for each axis resulting from the principal components analysis one first has to evaluate whether its direction in the multidimensional space is a characteristic of the statistical population, or whether its direction is entirely dependent on the statistical sampling.

In the following sections some statistical considerations concerning the technique of principal components analysis will be presented. In this explanation the open variables $X_i$ are used because we intend to apply this type of analysis to the output of the computer program BALANC which elaborates the free open covariances model. Obviously the following sections are relevant for any set of variables.

## VI.2. A brief explanation of principal components analysis

Principal components analysis should be based on the matrix R of the correlation coefficients between the open variables. The matrix R has M rows and M columns, the number in the i-th row and the k-th column being

$$R_{i,k} = \begin{cases} 1 & i=k \\ R(X_i, X_k) & i \neq k \end{cases} \tag{6.1}$$

Instead of the correlation matrix R the covariance matrix C, defined as

$$C_{i,k} = \begin{cases} \mathrm{var}(X_i) & i=k \\ \mathrm{cov}(X_i, X_k) & i \neq k \end{cases} \tag{6.2}$$

is often used as input for the principal components analysis. In our opinion such a procedure is not correct if one's intention is to deduce relations between the variables, because in (6.2) the sizes of the variations of the variables are involved as well, which may completely distort the picture. Variables with large variances determine the directions of the larger axes in multidimensional space.

Stating matters more precisely, we start with the case of two variables $X_1$ and $X_2$. The trouble resulting from (6.2) is avoided by standardizing the variables. The variable

$$\frac{X_i - E(X_i)}{\sqrt{\mathrm{var}(X_i)}} \tag{6.3}$$

is again called $X_i$ in the following text. Therefore each variable $X_i$ has the properties

$$E(X_i) = 0; \quad \mathrm{var}(X_i) = 1 \tag{6.4}$$

and each pair of variables $X_i$ and $X_k$ has the property

$$r(X_i, X_k) = \mathrm{cov}(X_i, X_k) = E(X_i, X_k) \tag{6.5}$$

Standardizing the variables according to (6.3), giving them an expected value equal to zero and a variance equal to one (6.4), leads to the simplifying property (6.5) of the correlation coefficients.

If one considers two variables $X_1$ and $X_2$ there are three ways of expressing correlation. If the correlation between $X_1$ and $X_2$ is equal to $\rho$,

$$r(X_1, X_2) = \rho \tag{6.6}$$

there might be a logical reason for assuming that $X_2$ is dependent on $X_1$. Then we have

$$X_2 = \rho \cdot X_1 + Y \tag{6.7}$$

in which Y is a variable independent of $X_1$. From (6.4) it follows that $E(Y) = 0$ and that $\text{var}(Y) = 1 - \rho^2$. Hence $\rho$ is the linear regression coefficient in the case of regression of $X_2$ on $X_1$.

Secondly, there may be reason to believe that $X_1$ is dependent on $X_2$. Then

$$X_1 = \rho \cdot X_2 + Y \tag{6.8}$$

in which Y is now independent of $X_2$, but similarly $E(Y) = 0$ and $\text{var}(Y) = 1 - \rho^2$. It is emphasized that the regression coefficient and the correlation coefficient are identical only in the case of standardized variables $X_1$ and $X_2$.

The third way to describe correlation between $X_1$ and $X_2$, and for our purpose the most interesting one, is to define two mutually independent and standardized variables $Y_1$ and $Y_2$ (they are generally considered to be standard normal variables, but there is no strict necessity for that), because dependence of one variable on the other need not be supposed in this case. So

$$E(Y_1) = 0;\ E(Y_2) = 0;\ \text{var}(Y_1) = 1;\ \text{var}(Y_2) = 1;$$
$$\text{cov}(Y_1, Y_2) = 0 \tag{6.9}$$

which describe $X_1$ and $X_2$ with $r(X_1, X_2) = \rho$ as follows:

$$\begin{aligned} X_1 &= \tfrac{1}{2} \cdot \sqrt{2\,(1+\rho)} \cdot Y_1 + \tfrac{1}{2} \cdot \sqrt{2\,(1-\rho)} \cdot Y_2 \\ X_2 &= \tfrac{1}{2} \cdot \sqrt{2\,(1+\rho)} \cdot Y_1 - \tfrac{1}{2} \cdot \sqrt{2\,(1-\rho)} \cdot Y_2 \end{aligned} \tag{6.10}$$

The reader may check that (6.10), together with (6.9), lead to the properties (6.4) and (6.6).

Expression (6.10) is written in matrix notation as

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} +\tfrac{1}{2} \cdot \sqrt{2\,(1+\rho)} & +\tfrac{1}{2} \cdot \sqrt{2\,(1-\rho)} \\ +\tfrac{1}{2} \cdot \sqrt{2\,(1+\rho)} & -\tfrac{1}{2} \cdot \sqrt{2\,(1-\rho)} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \tag{6.11}$$

The matrix contains two column vectors:

$\begin{pmatrix} +\tfrac{1}{2} \cdot \sqrt{2\,(1+\rho)} \\ +\tfrac{1}{2} \cdot \sqrt{2\,(1+\rho)} \end{pmatrix}$ associated with the variable $Y_1$, and

$\begin{pmatrix} +\tfrac{1}{2} \cdot \sqrt{2\,(1-\rho)} \\ -\tfrac{1}{2} \cdot \sqrt{2\,(1-\rho)} \end{pmatrix}$ associated with the variable $Y_2$.

It is easy to see that the lengths of the vectors (the length of a vector is defined as the square root of the sum of the squares of the coordinates of the vector) are $\sqrt{1+\rho}$ and $\sqrt{1-\rho}$, respectively. If $r(X_1, X_2) = \rho > 0$, then the first vector (the first principal component) is the larger one. If $r(X_1, X_2) = \rho < 0$, then the second vector is the larger one.

Denoting the column vector of the $X_i$ by X, the column vector of the $Y_i$ by Y, the matrix of (6.11) by U, and the column vector from the matrix associated with $Y_i$ by $U_i$, one writes expression (6.11) simply as

$$X = U \cdot Y \tag{6.12}$$

and

$$||U_1|| = \sqrt{1+\rho}; \quad ||U_2|| = \sqrt{1-\rho} \tag{6.13}$$

where $||U_i||$ indicates the length of vector $U_i$.

The case M = 2 (two variables $X_1$ and $X_2$) is trivial: the directions of the vectors $U_1$ and $U_2$ are independent of the correlation coefficient $\rho$, and the lengths $||U_1||$ and $||U_2||$ are simple functions of $\rho$. For more than two dimensions the description by means of matrix algebra is much more essential.

Principal components analysis is based on the assumption that the series of variables $X_1, X_2, X_3, \ldots, X_M$ can be written analogously to (6.11) as

$$X_i = \sum_{h=1}^{M} u_{ih} \cdot Y_h \tag{6.14}$$

in which the $Y_h$ for h = 1, 2, 3, . . ., M, are mutually independent and standardized variables. Hence

$$E(Y_h) = 0; \ \mathrm{var}(Y_h) = 1; \ \mathrm{cov}(Y_h, Y_l) = 0 \text{ for } h \neq l \tag{6.15}$$

Analogously to the two-dimensional case, (6.14) is written

$$X = U \cdot Y \tag{6.16}$$

in which X and Y now are M-dimensional column vectors, and U is a matrix with M rows and M columns. The number in the i-th row and in the h-th column is identical to $u_{ih}$ of definition (6.14). The matrix U now contains M column vectors $U_1, U_2, U_3, \ldots, U_M$:

$$U_h = \begin{pmatrix} u_{1h} \\ u_{2h} \\ u_{3h} \\ \cdots \\ \cdots \\ u_{Mh} \end{pmatrix} \tag{6.17}$$

$U_h$ is the column vector associated with the variable $Y_h$. An important requirement for the $U_h$ is that they must form an orthogonal basis. Therefore each pair $U_h$ and $U_l$ must be mutually orthogonal. Hence,

$$U_h \cdot U_l = u_{1h} \cdot u_{1l} + u_{2h} \cdot u_{2l} + \ldots u_{Mh} \cdot u_{Ml} = 0 \tag{6.18}$$

The basic problem is to deduce the matrix U of (6.16) from the correlation matrix R of (6.1). The relation between R and U is explained in a few lines below.

According to (6.1) and (6.5) and because of the standardization (6.3) we can write

$$R_{i,k} = R(X_i, X_k) = E(X_i \cdot X_k).$$

From (6.14) it follows that

$$R_{i,k} = E((\sum_{h=1}^{M} u_{ih} \cdot Y_h) \cdot (\sum_{l=1}^{M} u_{kl} \cdot Y_l))$$

and because of the properties (6.15) this leads to

$$R_{i,k} = \sum_{h=1}^{M} u_{ih} \cdot u_{kh} \tag{6.19}$$

which is written in matrix notation as:

$$R = U \cdot U^t \tag{6.20}$$

in which $U^t$ is the transpose of the matrix U. The matrix $U^t$ is defined by

$$U^t_{i,k} \equiv U_{k,i}$$

From (6.20) it can easily be deduced that

$$R \cdot U_h = ||U_h||^2 \cdot U_h \tag{6.21}$$

which means that the vector $U_h$ is an eigenvector of the matrix R, with the eigenvalue $l_h = ||U_h||^2$.

The eigenvalue associated with the eigenvector $U_h$,

$$l_h = ||U_h||^2 = \sum_{i=1}^{M} u_{ih}^2 \tag{6.22}$$

can be considered as the amount of variance due to the "random vector" $U_h \cdot Y_h$. Therefore the sum of all eigenvalues must be equal to the total amount of variance. Indeed:

$$\sum_{h=1}^{M} l_h = \sum_{h=1}^{M} \sum_{i=1}^{M} u_{ih}^2 = \sum_{i=1}^{M} \sum_{h=1}^{M} u_{ih}^2 = \sum_{i=1}^{M} \mathrm{var}(X_i) = M \tag{6.23}$$

because $\mathrm{var}(X_i) = \sum_{h=1}^{M} u_{ih}^2$, which follows directly from (6.14).

There are several numerical methods for deriving the eigenvalues with their associated eigenvectors from some given matrix. The eigenvalues are ordered according to size, so

$$l_1 \geqslant l_2 \geqslant l_3 \geqslant \ldots \geqslant l_M \tag{6.24}$$

As we are only interested in the larger eigenvectors associated with the larger eigenvalues (see (6.22)), we used a simple computer program that performs a fast calculation of a limited number of eigenvectors and eigenvalues. In the following section it will be shown why it is useless to calculate the complete series of eigenvectors.

## VI.3. Statistical considerations

From the point of view of mathematical statistics we must find out whether the obtained eigenvectors are "meaningful", i.e. whether they are estimates of characteristic eigenvectors of the statistical population under consideration. A vector is characterized by its length and by its direction in multidimensional space. The square of the length of the eigenvector $U_h$, $l_h$, should approximate the "real" eigenvalue $\tilde{l}_h$. The direction of the estimate $U_h$ should approximate the direction of the population eigenvector $\tilde{U}_h$.

A good description of the statistical problems involved is achieved by starting from the "most negative" null hypothesis possible, namely that for the population eigenvalues

$$H_0 : \tilde{l}_1 = \tilde{l}_2 = \tilde{l}_3 = \ldots = \tilde{l}_M = 1 \tag{6.25}$$

is true. This implies that all eigenvectors are of equal length. A consequence of this hypothesis is that there is no characteristic eigenvector. If $\tilde{l}_h = \tilde{l}_{h+1}$ holds for some h, then the associated eigenvectors $\tilde{U}_h$ and $\tilde{U}_{h+1}$ are not uniquely defined. Obviously they are interchangeable. More generally it can be argued that each pair $\tilde{U}_h$, $\tilde{U}_{h+1}$ can be replaced by the pair

$$\begin{cases} \tilde{U}'_h = \tilde{U}_h \cdot \cos(x) + \tilde{U}_{h+1} \cdot \sin(x) \\ \tilde{U}'_{h+1} = \tilde{U}_h \cdot \sin(x) - \tilde{U}_{h+1} \cdot \cos(x) \end{cases}$$

for any value of x, which in fact means that the pair $\tilde{U}_h$, $\tilde{U}_{h+1}$ can be rotated in the plane in which they are embedded. Hence, the direction of $\tilde{U}_h$ is not fixed if $\tilde{l}_h = \tilde{l}_{h+1}$. From the hypothesis $H_0$ of (6.25) it follows that each vector is an eigenvector with eigenvalue l = 1. Obviously (6.25) means that there is no relation between any pair $X_i$ and $X_k$. All variables $X_1$, $X_2$, $X_3$, . . ., $X_M$ are mutually independent.

However, even if hypothesis (6.25) holds, the numerical analysis of the set of counts will yield a series of eigenvalues $l_1, l_2, l_3, \ldots, l_M$ according to (6.24) and an associated series of eigenvectors $U_1, U_2, U_3, \ldots, U_M$. The eigenvalue $l_1$ may be much larger than one, and $l_M$ much less than one, but these differences are meaningless, as are (the directions of) the eigenvectors $U_h$.

It is difficult to find a criterion for deciding whether we accept or reject the hypothesis (6.25) by means of the larger $l_h$. The series $l_1, l_2, l_3, \ldots, l_M$ is an ordered series, which makes it difficult to derive probability distributions for the $l_h$ that must hold according to the hypothesis (6.25). Such distributions must have M and N (the number of observations) as parameters.

A completely different way of reasoning is presented below. Using the hypothesis (6.25) and assuming M to be not too small ($M \geqslant 8$), we imagine that the largest eigenvector $U_1$, found by the computer method, accounts for some part, named $b_1$, of the total variance. As a consequence,

$$l_1 = b_1 \cdot \sum_{i=1}^{M} \text{var}(X_i) = b_1 \cdot M$$

in which $b_1$ is thought to have a "narrow" distribution, with parameters M and N. The second eigenvector $U_2$ is thought to account for a comparable part $b_2$ of the remaining variance:

$$l_2 = b_2 \cdot (M - l_1) = b_2 \cdot (1 - b_1) \cdot M \approx b_1 \cdot (1 - b_1) \cdot M$$

because $b_2$ has a similar "narrow" distribution. The parameters are $(M - 1)$ and N, because $U_2$ is determined in an $(M - 1)$-dimensional space. As M is large, we think that the difference between M and $(M - 1)$ is irrelevant as far as these b-distributions are concerned. Similar statements may be made for $l_3, l_4$, and so on, provided we do not approach $l_M$.

Summarizing, on the basis of (6.25) and M is large, there is a number b, for which $0 < b < 1$, so that

$$l_h \approx b^{(h-1)} \cdot (1 - b) \cdot M \qquad h \ll M \tag{6.26}$$

Identical to (6.26) is the statement that

"$\ln(l_h)$ decreases more or less linearly with increasing h for $h \ll M$". (6.27)

Expression (6.27) will be used as a criterion to test $H_0$ of (6.25). Obviously this criterion is rather vague as is expressed by "more or less linearly"; yet we consider it to be very useful.

The most simple alternative hypothesis is:

$$H_1 : \tilde{l}_1 > \tilde{l}_2 = \tilde{l}_3 = \tilde{l}_4 = \ldots = \tilde{l}_M , \tag{6.28}$$

i.e. there is only one single meaningful axis of variation. There is no more than one meaningful eigenvector $\tilde{U}_1$, accounting for the amount $\tilde{l}_1$ of the total variation.

If we find that the $\ln(l_h)$, resulting from a set of data, decrease approximately linearly with increasing h, for $2 \leqslant h \ll M$, but that $\ln(l_1)$ is larger than expected according to the linear trend, then we are inclined to accept $H_1$ of (6.28). Then the associated eigenvector $U_1$ is considered to be a meaningful estimate of $\tilde{U}_1$, although each coefficient of the vector $U_1$ may contain a large error compared to the corresponding coefficient of the theoretical $\tilde{U}_1$, and then $l_1$ is a fair estimate of $\tilde{l}_1 > 1$. An example is given in figure 7, which pertains to chapter X.

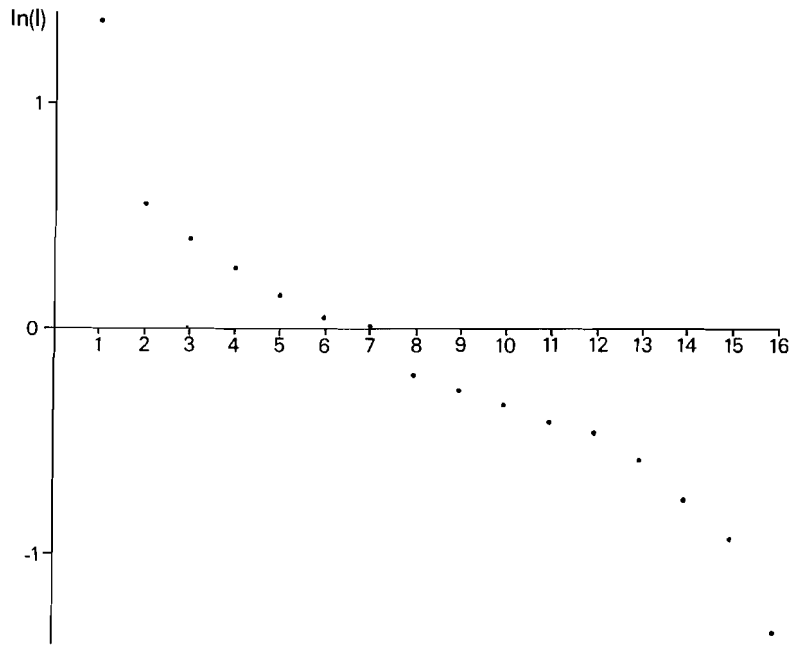



Fig. 7 The series of values of $\ln(l_h)$ for h = 1 up to and including 16, from the data of Ryan (1972, table 2), discussed in chapter X.

Muirhead and Chikuse (1975) argue that in the case of such a "single" eigenvalue $\tilde{l}_1$, the statistic

$$(N - l_1)/\tilde{l}_1 \tag{6.29}$$

has approximately a chi square distribution with N degrees of freedom, provided N is large. In chapter V it was shown that the number of degrees of freedom may be much less than N, the number of observations (counts). We translate (6.29) by saying that the standard error of $l_1$ as an estimate of $\tilde{l}_1$, is

$$SE(l_1) = l_1 \cdot \sqrt{\frac{2}{df}} \tag{6.30}$$

in which df is the number of degrees of freedom; this number may be estimated roughly from the autocorrelograms of the variables (taxa) that play an important part in the eigenvector $U_1$, according to (6.14) and (6.17).

Of course more complex alternative hypotheses are possible, such as

$$H_2: \tilde{l}_1 \geqslant \tilde{l}_2 > \tilde{l}_3 = \tilde{l}_4 = \ldots = \tilde{l}_M \tag{6.31}$$

but we believe that this hypothesis is already fairly exceptional. In the literature authors are inclined to interpret too many eigenvectors $U_h$ resulting from their data set as meaningful "principal components" (see chapter X). In our opinion accepting $H_2$ of (6.31) is only justified if both $\ln(l_1)$ and $\ln(l_2)$ are larger than expected from the more or less linear trend of the series $\ln(l_3)$, $\ln(l_4)$, $\ln(l_5)$. . . . In a similar way it might be possible that even more complex hypotheses with more than two relevant eigenvectors are acceptable. If the eigenvalues corresponding to these relevant eigenvectors are single, i.e. they are neither equal to the preceding eigenvalue nor to the following one, the standard error formula (6.30) can be used – with some care. If two of such relevant eigenvalues $l_h$ and $l_{h+1}$ differ very slightly, however, so that $\tilde{l}_h = \tilde{l}_{h+1}$ might be true, (6.30) may not be used. Accepting $\tilde{l}_h = \tilde{l}_{h+1}$ implies that $U_h$ and $U_{h+1}$ are not unique eigenvectors, because then for any real number x

$$U_h \cdot \cos(x) + U_{h+1} \cdot \sin(x)$$

is the estimate of an eigenvector with eigenvalue $\tilde{l}_h$.

# Chapter VII

# Q-MODE ANALYSIS OF SETS OF COUNTS

## VII.1. Introduction

Although our investigation mainly concerns the analysis of relations between taxa in sets of counts, some lines must be devoted to the Q-mode analysis, i.e. the analysis of the relations between the counts according to the scores of the taxa they contain. The sediment samples are compared with respect to their total taxa composition. Such a Q-mode analysis may result in discerning groups of counts (samples), each group consisting of counts that are mutually "similar" in their proportions. Obviously each group of samples is based on a certain group of taxa the composition of which was established from an R-mode analysis of the same data.

It is important to realize that Q-mode analysis and R-mode analysis are of a different nature. In R-mode analysis groups of taxa are generally established by means of the correlation coefficient, a measure which gives not only a number to each pair of taxa but also a sign (positive or negative). In other words, two "related" taxa may have a significantly positive correlation coefficient value, but other "related" taxa may have a significantly negative correlation coefficient value. In M. M. Drooger and Hageman (1979) the correlation coefficient statistic was referred to as "two-sided" for that reason.

In Q-mode analysis the groups of counts are formed on the basis of some similarity measure, which has no negative branch in its scale. Its coefficients are called similarity coefficients or distance coefficients. In our opinion all these coefficients are equivalent, because they cannot give negative values to any comparison-pair (of course similarity can also be used in R-mode analysis). Perfect match can be considered as perfectly similar or "identical", $s = 1$, or equivalently as "distance zero", $d = 0$. No match at all can be considered as entirely dissimilar, $s = 0$, or alternatively as "distance very large, infinite", $d = \infty$.

We decided to refrain from any multivariate Q-mode analysis. One reason was that the number of counts (samples) is generally (much) larger than the number of taxa (variables); this fact makes multivariate Q-mode analyses even more difficult than multivariate R-mode analyses.

In the case of distribution charts one may have several distinct groups

of samples, each group representing a typical biotope. Such groups also appear clearly when simpler methods are used like the method described in the next section. This is the second reason for refraining from multivariate Q-mode analysis.

A third reason is that in the majority of the range charts there are no distinct groups of samples. A group of samples (counts) may be found that shows some extreme characteristic, another group may show another extreme characteristic, but the majority of the samples can be described as "average" samples. A statistical treatment like the one in the following section seems more appropriate in such cases, in order to decide which samples (counts) can be considered as "average" and which as "extreme".

## VII.2. Q-mode analysis by program PQMODE

In our simple Q-mode analysis each score $x_{ij}$, for the count with index number j and size $n_j$ (2.2) is compared with the weighted mean proportion $\hat{p}_i$ defined in (2.25). If the multinomial model holds for the series of counts, then $x_{ij}$ would have approximately the expected value and the variance:

$$E(x_{ij}) \approx n_j \cdot \hat{p}_i; \quad var(x_{ij}) \approx n_j \cdot \hat{p}_i \cdot (1 - \hat{p}_i) \tag{7.1}$$

as can be deduced from (2.12). Therefore the deviation

$$ch_{ij} = \frac{x_{ij} - n_j \cdot \hat{p}_i}{\sqrt{n_j \cdot \hat{p}_i \cdot (1 - \hat{p}_i)}} \tag{7.2}$$

would have a standard normal distribution. Also if the multinomial model clearly is not valid, $ch_{ij}$ is a useful expression, weighting the deviation of $x_{ij}$ from the "expected" value $n_j \cdot \hat{p}_i$.

$$\text{As} \sum_{i=1}^{M} (x_{ij} - n_j \cdot \hat{p}_i) = 0,$$

the variables of the series $(ch_{1j}, ch_{2j}, ch_{3j}, \ldots, ch_{Mj})$ are interdependent. Suppose that taxon i in count j has a score $x_{ij}$ that deviates largely from the expected value (taxon i squeezes), then the absolute value of $ch_{ij}$ is large and induces a deviation $d(ch_{kj})$ in the value $ch_{kj}$ of any other taxon k in count j. The deviation $d(x_{kj})$ in the value of the score $x_{kj}$ of taxon k by taxon i in count j is

$$d(x_{kj}) = \left( \frac{-\hat{p}_k}{1 - \hat{p}_i} \right) \cdot (x_{ij} - n_j \cdot \hat{p}_i).$$

The relation between the deviations $d(ch_{kj})$ and $d(x_{kj})$ is easily deduced from (7.2):

$$d(ch_{kj}) = d(x_{kj}) / \sqrt{n_j \cdot \hat{p}_k \cdot (1 - \hat{p}_k)}$$

These two equations lead to

$$d(ch_{kj}) = -\sqrt{\frac{\hat{p}_i \cdot \hat{p}_k}{(1 - \hat{p}_i) \cdot (1 - \hat{p}_k)}} \cdot ch_{ij} = R_m(P_i, P_k) \cdot ch_{ij} \qquad (7.3)$$

$R_m(P_i, P_k)$ being the expected value of the correlation coefficient between the proportions $P_i$ and $P_k$ according to the multinomial model given in (2.18).

Equation (7.3) provides a means of checking whether any "significant" value of $ch_{kj}$ of some taxon k in some count j is induced by a "squeezing effect" of some taxon i in that count j. Such a squeezing effect is recognized by a very large absolute value of $ch_{ij}$. The value of $ch_{ij}$ induces a deviation $d(ch_{kj})$ in the value of $ch_{kj}$ according to (7.3).

A Q-mode computer program, called PQMODE, has been written to be used for the comparison of a set of counts. In this program the matrix of scores $(x_{ij})$ is printed, followed by the matrix $(ch_{ij})$, calculated according to (7.2). Next, a chi square test for goodness of fit to the proportions $(\hat{p}_1, \hat{p}_2, \hat{p}_3, \ldots, \hat{p}_M)$ is performed for each count. Such chi square test statistics have the basic structure

$$X^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \qquad (7.4)$$

in which $O_k$ is "the observed number in class k", and $E_k$ is "the expected number in class k". If the multinomial model is applicable, $X^2$ has a chi square distribution with $(K - 1)$ degrees of freedom.

Calling the total number of observations n, obviously

$$\sum_{k=1}^{K} O_k = \sum_{k=1}^{K} E_k = n, \qquad (7.5)$$

and $X^2$ is calculated more easily from

$$X^2 = \left( \sum_{k=1}^{K} \frac{O_k^2}{E_k} \right) - n. \qquad (7.6)$$

Another feature is that classes with too small expected numbers (e.g.

$E_k < 5$) should be lumped. Refined lumping rules have been presented by Cochran (1952) for different situations.

In our case the chi square statistic for goodness of fit results in

$$X^2 = \left( \sum_{i=1}^{K} \frac{x_{ij}^2}{n_j \cdot \hat{p}_i} \right) - n_j = \left( \left( \sum_{i=1}^{K} \frac{x_{ij}^2}{\hat{p}_i} \right) \Big/ n_j \right) - n_j \qquad (7.7)$$

for every count j. In this formula K is less than or equal to M, depending on the number of scores that are lumped. Hence, it is possible that the index i in (7.7) does not always correspond to the original rank number of the taxon.

In addition to $X^2$ and the number of degrees of freedom, $K - 1$, also the expression

$$D_j = X^2/(K - 1) \qquad (7.8)$$

is presented for each count j in the PQMODE program. $D_j$ can be considered as a distance coefficient, indicating the distance between count j and the hypothetical "average count". Formula (7.8) follows from the fact that each significant value of a chi square statistic, divided by the number of degrees of freedom, is a measure of the deviation from the null hypothesis, in this case from the hypothesis that count j has been drawn from a population with proportions $(\hat{p}_1, \hat{p}_2, \hat{p}_3, \ldots, \hat{p}_M)$.

Instead of comparing counts with the average of all counts, one might need to compare one count with another, as it was the case in the Capo Rossello investigation (U.M.B. 17). Zachariasse (1978) and Schmidt (1978) performed repeated counts, in order to investigate whether a count is representative for the sample it comes from. One can imagine other possible reasons for comparing counts with each other.

Our Q-mode program PQMODE contains such a comparison. It should be borne in mind, however, that if a number of N counts are compared with each other, the procedure results in the number of $N \cdot (N - 1)/2$ comparisons, which is the number of pairs of counts. If N = 10, we have 45 comparisons, if N = 20, the number of comparisons rises to 190, if N = 40, it already reaches 780 pairs of counts. Program PQMODE refuses this task when the number of pairs exceeds 950, which is the case when N exceeds 44.

In this mutual comparison for each pair of counts j and k, first the chi square test for homogeneity is performed, which is the test of the hypothesis that both counts are from the same statistical population. See Cochran (1952, 1954).

It should be noted that this statistical population need not be identical to the "average" population as regards the goodness of fit test statistic (7.4). The test mentioned below and the chi square test of (2.27) have an identical structure; both are equivalent to the chi square test for independence in any contingency table with two rows or with two columns.

The chi square statistic

$$X^2 = \frac{\sum_{i=1}^{K} (x_{ij} + x_{ik}) \cdot (q_{ijk} - \hat{q}_{jk})^2}{\hat{q}_{jk} \cdot (1 - \hat{q}_{jk})} = \frac{\left( \sum_{i=1}^{K} q_{ijk} \cdot x_{ij} \right) - \hat{q}_{jk} \cdot n_j}{\hat{q}_{jk} \cdot (1 - \hat{q}_{jk})} \qquad (7.9)$$

in which $q_{ijk} = x_{ij}/(x_{ij} + x_{ik})$ and $\hat{q}_{jk} = n_j/(n_j + n_k)$, has a chi square distribution with $(K - 1)$ degrees of freedom, if the hypothesis that both counts are from one statistical population is true.

Both in this statistic and in the statistic (7.7), taxa that have an expected number in one of the counts that is less than five are lumped. Therefore K may be less than M, and the index i in (7.9) does not always correspond to the original rank number of the taxon.

In addition to the $X^2$ of (7.9) and the number of degrees of freedom, $K - 1$, the distance coefficient

$$D_{jk} = X^2/(K - 1) \qquad (7.10)$$

is presented in program PQMODE for each pair of counts j and k. $D_{jk}$ can be seen as the "distance" between count j and count k, or as a measure of the deviation from the hypothesis that count j and count k are from one statistical population.

Chapter VIII

# QUANTITATIVE ANALYSIS OF THE BENTHONIC FORAMINIFERA OF LE CASTELLA

## VIII.1. Introduction

Bremer, Briskin and Berggren (1980) presented a quantitative analysis of the benthonic foraminifera of the 150 metre composite section of Le Castella (Calabria, Italy), which by definition contains the Pliocene-Pleistocene boundary. They give counts of some 60 taxonomic units of benthonic foraminifera with scores over 1% from 54 samples along the column of this boundary-stratotype section in a range chart (their appendix 2). The totals of the counts vary between 65 and 322; all data are given as percentage values. Two Q-mode principal components analyses were carried out on these data, using the computer program CABFAC (Klovan & Imbrie, 1971).

After a first run five of the samples from the Pliocene were discarded because of deviating, high percentages of *Brizalina attica* and *Bulimina exilis*. No further attention is paid to these assemblages in the later paleoecological interpretation.

The second computer run on the remaining 49 counts revealed two assemblage types, one dominated by *Cibicidoides floridanus*, the other by *Cassidulina laevigata, Uvigerina peregrina, Bulimina costata, B. aculeata + marginata* and *Hyalinea balthica*. Nothing peculiar is observed about the so-called index fossil of the Pleistocene, *Hyalinea balthica*, which occurs only in the upper half of the section.

Both associations alternate along the column in the Pliocene part of the section; the *Cassidulina* assemblage is dominant in the Pleistocene, starting a few samples above the boundary marker bed. All assemblages are thought to indicate a similar bathymetric range of 130–700 metres, corresponding to that of today's Intermediate Water Mass of the Mediterranean. The *Cassidulina* association is said to be the cooler of the two, comparable to that found in the Eastern Mediterranean today; the *Cibicidoides* assemblage indicates a warmer environment, similar to that of the Gulf of Mexico today.

For the Pliocene of Le Castella the authors assume frequent fluctuations between cooler and warmer surface waters, the Pleistocene part of the section showing a cooler surface water regime throughout. Confirmation of this assumption is found in the percentage curve of warm water planktonic

foraminifera (*Hastigerina aequilateralis* and *Globigerinoides ruber*). These percentages vary between 0 and 37 on the basis of counts of "at least" 200 specimens.

### VIII.2. Analysis of the series of counts by means of the computer programs DISTUR, REGRES, BALANC and PQMODE

Two modifications were made to the published range chart before our analysis was performed. The first alteration was to reinstate the original scores (approximately) for the percentages. The second was to reduce the number of taxonomic units to 14, which are the most frequent and include all taxa that contribute to the conclusions of Bremer et al. These 14 taxa or taxa groups are:

1 *Brizalina attica*
2 *Brizalina catanensis*
3 *Brizalina dilatata*
4 *Bulimina aculeata* + *marginata*
5 *Bulimina costata*
6 *Bulimina exilis*
7 *Cassidulina crassa*
8 *Cassidulina laevigata*
9 *Cibicidoides floridanus*
10 *Epistominella rugosa convexa*
11 *Hyalinea balthica*
12 *Trifarina angulosa*
13 *Uvigerina peregrina*
14 Rest group

The resulting 14 × 54 matrix of scores has been punched, one count per card. Then the information was stored on permanent file at the Cyber 175 computer of the Academic Computer Centre Utrecht. The data matrix was analyzed by means of the DISTUR and BALANC programs.

The output of the DISTUR program did not differ substantially from the output of the BALANC program. The essential part of the output of the BALANC program is presented in the "spider-web" diagram of figure 8, in which all significant correlation coefficients (significance level $\alpha = 0.01$) have been entered. From this figure we arrive at the following conclusions.

The taxa (1) *Brizalina attica* and (6) *Bulimina exilis* show negative trends, i.e. their open variables tend to decrease upwards along the stratigraphic column. There is a positive correlation between the open variables of these two taxa. The taxa (11) *Hyalinea balthica* and (12) *Trifarina angulosa* show

positive trends. There is a positive correlation between their open variables as well, if one takes $\alpha = 0.05$ as the significance level.

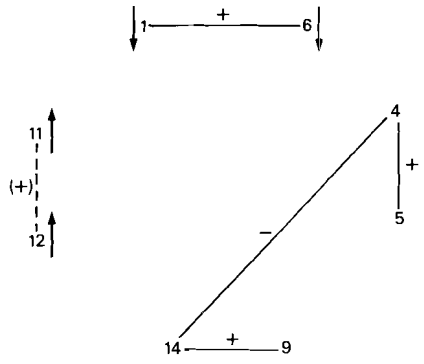



Fig. 8 Spider-web diagram for the Castella samples showing the significant correlation coefficients and trends according to the output of the BALANC program. 54 counts. Solid lines: significance level $\alpha = 0.01$; dashed lines: significance level $\alpha = 0.05$ (two-sided).

In addition there is a positive correlation between the open variables of (4) *Bulimina aculeata + marginata* and (5) *Bulimina costata.* The same holds for (9) *Cibicidoides floridanus* and the Rest group (14). There is a negative correlation between (4) *Bulimina aculeata + marginata* and the Rest group (14).

We mention some details. According to the output of BALANC the variance of the sum of the open variables appears to be $\text{var}(T) = \sum_{i=1}^{14} p_i \cdot d_i =$ .129. From the output of the DISTUR program we found for the variance of the sum of the open variables $S_{tt}$ = .123. For the Castella data the two models appear to give very similar results. Only the negative correlation between (4) *Bulimina aculeata + marginata* and the Rest group (14) has not been found in the DISTUR program.

It appears that the Rest group (14) is somewhat peculiar. Its ratio $V_{ii}/v_{ii}$ of .251 is far from one, which may be in conflict with the supposed linearity of the open variables models. For this reason the Rest group (14) was eliminated in a later computer run (see below), but this elimination did not lead to different results for the other taxa.

From the output of the REGRES program it appears that only the sequence of percentages of (12) *Trifarina angulosa* is a clearly autocorrelated series ($R_1(1)$ = .523, $R_2(1)$ = .516, see the expressions (5.26) and (5.28)).

The values of $R_1(1)$ and $R_2(1)$ of all other taxa do not reach the critical value of .349 ($\alpha$ = 0.01). It should be noted that (12) *Trifarina angulosa* shows a positive trend according to the outputs of the DISTUR and BALANC programs, but that the trends of (1) *Brizalina attica,* of (6) *Bulimina exilis* and of (11) *Hyalinea balthica* are not reflected as autocorrelated series in the REGRES output. Since it is only the series of percentages of (12) *Trifarina angulosa* which shows autocorrelation, there is no need to apply a reduction of the number of degrees of freedom concerning the outputs of the DISTUR and BALANC programs.

In a run of our PQMODE program values of $ch_{ij}$ (7.2) were calculated for each cell. If $ch_{ij}$ is greater than five, the score $x_{ij}$ is considered to be much larger than expected according to the size of the count and the mean proportion of the taxon. If $ch_{ij}$ is less than $-5$, the score $x_{ij}$ is considered to be much less than expected. It appears that (1) *Brizalina attica* has high scores in counts CA06, CA11, CA13, CA17 and CA18. The taxon (2) *Brizalina catanensis* has an extremely high score (ch = 55.3) in count CA27 (132 out of 149 specimens counted), (6) *Bulimina exilis* in count CA08: ch = 55.2 (133 among 155 specimens counted), and (10) *Epistominella rugosa convexa* in count CA26 : ch = 76.1 (133 among 178 specimens counted). In the Pleistocene part of the column there are isolated peaks of (7) *Cassidulina crassa* (ch = 31.3) in CA49, (3) *Brizalina dilatata* (ch = 29.6) in CA50, (11) *Hyalinea balthica* (ch = 21.9) in CA56 and (5) *Bulimina costata* (ch = 20.9) in CA60. It is to be noted that there are no negative $ch_{ij}$ values of such magnitude; very low scores do not occur.

### VIII.3. Quantitative analysis after elimination of five counts

After their first computer run Bremer et al. decided to eliminate samples CA06, CA11, CA13, CA17 and CA18. They considered them to be distinctly different from all other samples because they contained very high relative frequencies of *Brizalina attica* and *Bulimina exilis.* They made another analysis with the remaining 49 samples. We note that CA08 was not eliminated, although this count contains an extremely high percentage of *Bulimina exilis,* and that the five counts that they did eliminate are precisely those which contain very high percentages of *Brizalina attica* according to the output of the PQMODE program.

So that we could compare the results of Bremer et al. with ours, we eliminated the same counts CA06, CA11, CA13, CA17 and CA18 and started another analysis with the remaining 49 counts.

It appears that the results from the BALANC program (fig. 9) are strikingly similar to the corresponding results from our first analysis (fig. 8). In this new analysis the negative trend of (6) *Bulimina exilis* is no longer significant, but now the negative trend of (9) *Cibicidoides floridanus* has become significant ($\alpha = 0.01$).

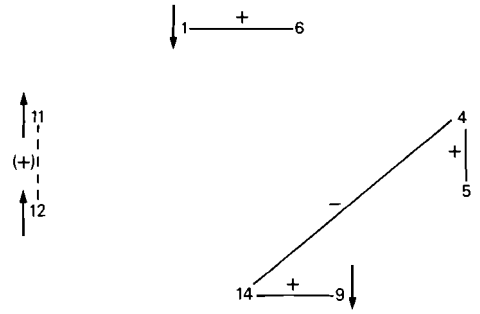



Fig. 9 Spider-web diagram for the Castella samples showing the significant correlation coefficients and trends according to the output of the BALANC program after the deletion of the counts 6, 11, 13, 17 and 18. 49 counts. See legend to figure 8.

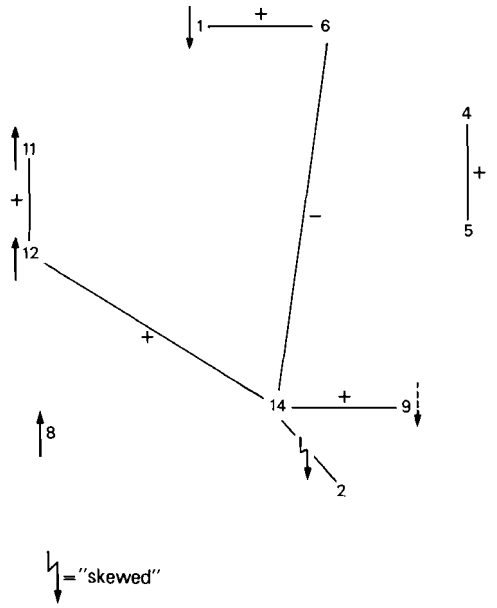



Fig. 10 Spider-web diagram for the Castella samples showing the significant proportion-ratio correlation coefficients and the significant trends according to the output of the DISTUR program using the 49 counts of figure 9. See legend to figure 8.

From the DISTUR program it appears that the taxa (2) *Brizalina catanensis,* (6) *Bulimina exilis* and (10) *Epistominella rugosa convexa* have extremely high scores for chi square and for $H_i$. Their mean proportions are small, however, and the contributions of the variances of their open variables (.017, .015 and .013, respectively) to the variance of the sum ($S_{tt}$ = .092) is not very high. Hence, it is not certain whether the scores of these three taxa disturb the analysis.

A considerable number of significant values are found in the matrix of proportion-ratio correlation coefficients (fig. 10). It has been argued in chapters II and III that these coefficients may be biased. Therefore these statistics have to be used with great care. An important reason why these matrices of proportion-ratio coefficients are maintained is that they indicate taxa that may disturb the analysis. We find the values .430 and .429 for the unweighted and for the weighted form of the correlation coefficient between $\hat{p}_{2,j}$ and $\hat{p}_{14,j}/(1 - \hat{p}_{2,j})$, respectively. For the unweighted and the weighted form of the correlation coefficient between $\hat{p}_{14,j}$ and $\hat{p}_{2,j}/(1 - \hat{p}_{14,j})$ the values are −.345 and −.329, respectively. Because of this asymmetry and similar "asymmetries" with other taxa, although to a lesser extent, the Rest group (14) might be suspected to disturb the results.

Because it was suspected that (2) *Brizalina catanensis,* (6) *Bulimina exilis,* (10) *Epistominella rugosa convexa* and the Rest group (14) were disturbing the analyses, two more computer runs were performed with the DISTUR program. In the first run the Rest group (14) was eliminated from the data matrix, in the second run (2) *B. catanensis,* (6) *B. exilis* and (10) *E. rugosa convexa* were eliminated. The outputs of these two runs do not reveal new

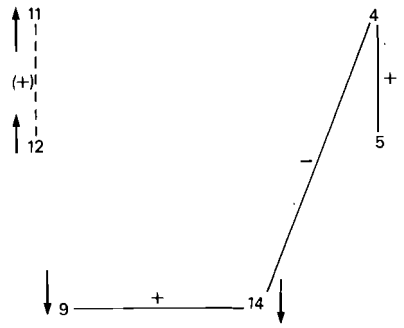



Fig. 11 Spider-web diagram for the Castella samples showing the significant correlation coefficients and trends according to the BALANC program, after elimination of the three taxa (2) *B. catanensis,* (6) *B. exilis* and (10) *E. rugosa convexa,* and using the 49 counts of figure 9. See legend to figure 8.

aspects. Elimination of the large Rest group (14) gives results that are almost identical to the results of the original DISTUR analysis on the 49 counts. The elimination of the three taxa (2) *B. catanensis*, (6) *B. exilis* and (10) *E. rugosa convexa* also gives comparable results.

Finally, another run of the BALANC program was performed with the latter three taxa eliminated. The result is presented in the spider-web diagram of figure 11. The outputs of the BALANC and DISTUR programs are identical, except for a difference in the behaviour of the Rest group (14). According to the BALANC outputs, the open variables of the Rest group (14) and of (9) *Cibicidoides floridanus* are positively correlated. There is a negative correlation between the open variables of the Rest group (14) and of (4) *Bulimina aculeata + marginata.* These correlations are not found in the DISTUR outputs.

### VIII.4. Discussion

Considering our analyses our first conclusion must be that all DISTUR and BALANC computer runs give fairly similar results. Furthermore, no taxon seems to have a notably disturbing effect for the entire sequence of samples. The reason that strongly "peaking" species have little effect is probably connected with the fact that low numbers of samples show peaks of similar character. Only in the case of *Brizalina attica* might there be a slight effect because of the five peak samples.

Apart from the occasional samples with some peak or other, the entire suite of samples gives the impression of a fairly homogeneous faunal assemblage throughout, with only minor trends and correlations. There is a basic assemblage consisting of mud-dwellers, similar in generic (and largely in specific) composition to that found in the nearby Plio-Pleistocene faunas of the Pyrgos area in Greece (Hageman, 1979). In contrast with the Greek data the elements of this basic assemblage show only a few positive correlations and no negative correlations, probably because in the Pyrgos sections we are dealing with a much wider array of habitats, including brackish, shallow marine and epiphytic. All Castella samples would fit into the open marine, offshore association of the western Peloponnesus.

Positive correlations between *Bulimina costata* and *B. aculeata + marginata,* and between *Bulimina exilis* and *Brizalina attica* are thought to be the expression of an ecological stress gradient superimposed on the regular mud-dwellers' fauna. It is thought plausible that a cline of reduced oxygen availability at the bottom caused the stress gradient, but the additional influence of differences in nutrient quality or quantity cannot be ruled out.

In this context the negative trends of *B. exilis* and *B. attica* simply mean that periods with aberrant bottom conditions were more frequent in the Pliocene than in the Pleistocene parts of the Castella section. All observed high ch values may correspond to some extra factor in the general, open marine environment. The enumeration also shows fewer and less extreme isolated peaks of different nature in the Pleistocene. The positive trend of *Hyalinea balthica* has a trivial meaning, because this so-called index fossil for the Pleistocene enters half-way along the section. The positive trend of *Trifarina angulosa* is not yet understood.

Our correlation results finally show a positive correlation between *Cibicidoides floridanus* and the Rest group. Evidently higher numbers of *C. floridanus* go together with increased frequencies of the sum of the additional 48 species, and thus possibly with faunal diversity. It can be suggested that *C. floridanus* is a representative of a more normal, or less mud-controlled, or shallower association. Because of determination differences *C. floridanus* was not recognized in the Pyrgos sediments (Hageman, 1979), but we assume that it might be comparable with Hageman's group of trochoid-biconvex taxa of which *Cibicides burdigalensis, C. ungerianus* and *C. dutemplei* are the main constituents. This trochoid-biconvex group joins the "normal" mud-dwellers' association, but it has the least affinity with sediment-type. In our analyses there are no negative correlations between *C. floridanus* and any species of the group of mud-dwellers, and the distinct role of *C. floridanus* found in the analysis of Bremer et al. is not found in our results. According to our explanation *C. floridanus* would be no more than an irregular addition from a neighbouring habitat.

Our results and explanation clearly differ from those achieved by Bremer et al. using multivariate analysis. For the sake of comparison a principal components analysis was performed on the matrix of the correlation coefficients between the open variables. Five eigenvectors were extracted by means of a Fortran computer program. They are given in figure 12.

We recall that $\|U_h\|^2 = l_h$ ; see chapter VI.

As the five eigenvalues mentioned in figure 12 form a steadily decreasing series according to the expression (6.26), one can have doubts about the statistical significance of all five eigenvectors. The individual coefficients of the eigenvectors which may be considered statistically significant have been printed in italics in figure 12. From the expression (6.20) $R = U \cdot U^t$ we consider such a coefficient to be 'significant' if its square is significant according to the table of r-distributions.

From the eigenvector $U_1$ it can be deduced that (1) *Brizalina attica* and

| | | $l_1$ | $l_2$ | $l_3$ | $l_4$ | $l_5$ | eigenvalues |
|---|---|---|---|---|---|---|---|
| | | 1.98 | 1.74 | 1.58 | 1.40 | 1.11 | |
| | | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | eigenvectors |
| taxa | 1 | *−.61* | −.47 | −.04 | −.26 | .09 | |
| | 2 | −.34 | −.28 | −.05 | −.09 | .12 | |
| | 3 | .20 | −.34 | .07 | *.68* | .30 | |
| | 4 | −.40 | *.64* | .03 | .39 | −.03 | |
| | 5 | −.27 | .36 | −.45 | .21 | −.39 | |
| | 6 | *−.56* | −.37 | .02 | −.26 | .11 | |
| | 7 | .12 | −.22 | .33 | .44 | −.20 | |
| | 8 | .09 | .40 | .12 | .04 | *.73* | |
| | 9 | .37 | .17 | *−.64* | −.35 | .13 | |
| | 10 | .01 | −.09 | −.17 | .03 | −.41 | |
| | 11 | .16 | .11 | *.60* | −.18 | −.28 | |
| | 12 | .49 | −.08 | .50 | −.33 | −.07 | |
| | 13 | .06 | .53 | .14 | −.37 | .00 | |
| | 14 | *.70* | −.36 | −.44 | .04 | −.03 | |

Fig. 12 The five largest eigenvectors with their eigenvalues of the matrix of correlation coefficients between the open variables of the Castella samples. Significant values have been printed in italics.

(6) *Bulimina exilis* form a group opposed to the Rest group (14), which is a very trivial conclusion: stress-dominated faunas versus diversified faunas. In the eigenvector $U_3$ the taxa (9) *Cibicidoides floridanus* and (11) *Hyalinea balthica* are found to be opposed, but the latter species occurs only in the upper half of the section, whereas *C. floridanus* is most frequent in the lower part.

We think that the presentation by means of the eigenvectors is much poorer than the presentation by means of the original matrix of "open" correlation coefficients, from which the eigenvectors have been extracted. Furthermore, the relations between the taxa deduced from the eigenvectors are not reflected in the original matrix, except for the relation between (1) *Brizalina attica* and (6) *Bulimina exilis.*

After elimination of the five samples with an abundance of the stress-tolerant species *Brizalina attica* and *Bulimina exilis,* Bremer et al. found an opposition along their second axis between *C. floridanus* and the *Uvigerina-Cassidulina* group. More precisely: their principal component assemblage "one" is a kind of vector mean and is therefore dominated by the abundant species. Their principal component assemblage "two" defines two end-member assemblages: one dominated by *C. floridanus* (factor score −.655

according to their table 1), the other by *C. laevigata* (factor score +.421), *U. peregrina* (+.336), *B. costata* (+.267), *B. aculeata* + *marginata* (+.266) and *H. balthica* (+.217).

Our first reaction is to wonder why the third principal component has not been considered by Bremer et al. This component has a variance equal to 6.00 according to their table 2, which is only slightly less than 7.73, the variance of the second principal component. This third component would demonstrate an opposition of *H. balthica* (factor score +.401) and *T. angulosa* (+.349) to *B. costata* (−.701) and *B. aculeata* + *marginata* (−.289).

As far as the second principal component is concerned, the two end-member assemblages would be plausible theoretically. We did not find corresponding negative correlations in our analyses, however, or rather they remain below the levels of significance in our DISTUR and BALANC programs. As far as the third principal component is concerned, we can confirm the positive correlations between (4) *B. aculeata* + *marginata* and (5) *B. costata,* and between (11) *H. balthica* and (12) *T. angulosa*, but negative correlations have not been found.

As the two end-member assemblages produced by the second principal component play an important part in the paper of Bremer et al., we checked the consistency of the *Uvigerina-Cassidulina* end member by means of three more runs of the DISTUR program.

Firstly, we lumped the scores of (4) *B. aculeata* + *marginata* and (5) *B. costata,* and the scores of (8) *C. laevigata* and (13) *U. peregrina.* Secondly, we lumped the scores of (4), (5), (8) and (13). Thirdly and lastly, we lumped the scores of these four taxa and (11) *H. balthica.*

No significant correlation coefficient value ($\alpha = 0.05$) was found between (9) *C. floridanus* and any of the constructed groups mentioned above. Hence, we do not believe that the existence of the *Cassidulina-Uvigerina* end-member assemblage of Bremer et al. can be defended from the statistical point of view.

Since the high frequencies of *C. floridanus,* which in our opinion reflect "less mud-control" in an otherwise similar habitat, leads to far-reaching conclusions on temperature fluctuations in the paper of Bremer et al., we felt obliged to make some additional remarks.

As to the temperature-dependent (in their explanation) factor loadings of their principal component II and the percentage values of their warm-water planktonic species, shown in their figure 4, one must admit that there is a positive correlation ($r = +.494$, $N = 37$, probability level $P \approx .001$) for the Pliocene part of the Castella column; this correlation remains significant

when the Pleistocene part is added ($r = +.454$, $N = 48$, $P \approx .001$). Notwithstanding the impression that the Pleistocene parts of the curves fit better than the Pliocene parts, the r value is about the same, probably because of the effect of the single aberrant topmost sample. Neither of the sequences appeared to be distinctly autocorrelated.

Another remarkable observation about both curves in their figure 4 is that if we assume that for their second principal component – 0.2 is a critical value for the factor loadings, there would be many more and wider cold spells in the bottom fauna than in the planktonic faunas at the surface. Perhaps the unmistakable fluctuations in *C. floridanus* frequencies have a more complex connection with the relative frequencies of the warm-water planktonics.

## VIII.5. Addendum

After the above text was written some additional investigations were carried out.

First of all it has to be mentioned that examination of material from Le Castella section present in the Utrecht collections, reveals differences in species determination. According to the taxonomy and nomenclature used in the Utrecht Micropaleontological Department (G. J. van der Zwaan, pers. comm.) *Cibicidoides floridanus* (Cushman) and *Brizalina attica* (Parker) are named differently. *C. floridanus* is our *Cibicides ungerianus* (d'Orbigny) and *B. attica* belongs to the *Bolivina spathulata* (Williamson) – group and is commonly named as such or as *Bolivina aenariensis* (Costa).

As Bremer et al. did not give any details in their lithological column, a new series of ten control samples was taken in the poorly exposed remains of the original Castella section in 1981. These samples were taken from different lithology types. At some places lamination was observed, a feature not recorded by earlier authors.

Counts of 100 benthonic foraminifera were performed on the larger than 125 $\mu$ fraction of these ten samples by G. J. van der Zwaan (figure 12a). In the laminated intervals (CT3 and CT6,7) we find high relative numbers of *Bolivina spathulata + dilatata* (up to 48 specimens out of 100). It is remarkable that the thin sand layer of CT6 has a faunal composition which resembles that of the overlying laminated sandy clay. The *Bolivina* peaks cause no distinct squeezing effect on the mud-dweller associations of *Cassidulina, Bulimina* and *Uvigerina.* The peak occurrences of *Bolivina spathulata + dilatata* seem to coincide with low numbers of *Hyalinea balthica, Cibicides ungerianus* and the Rest group.

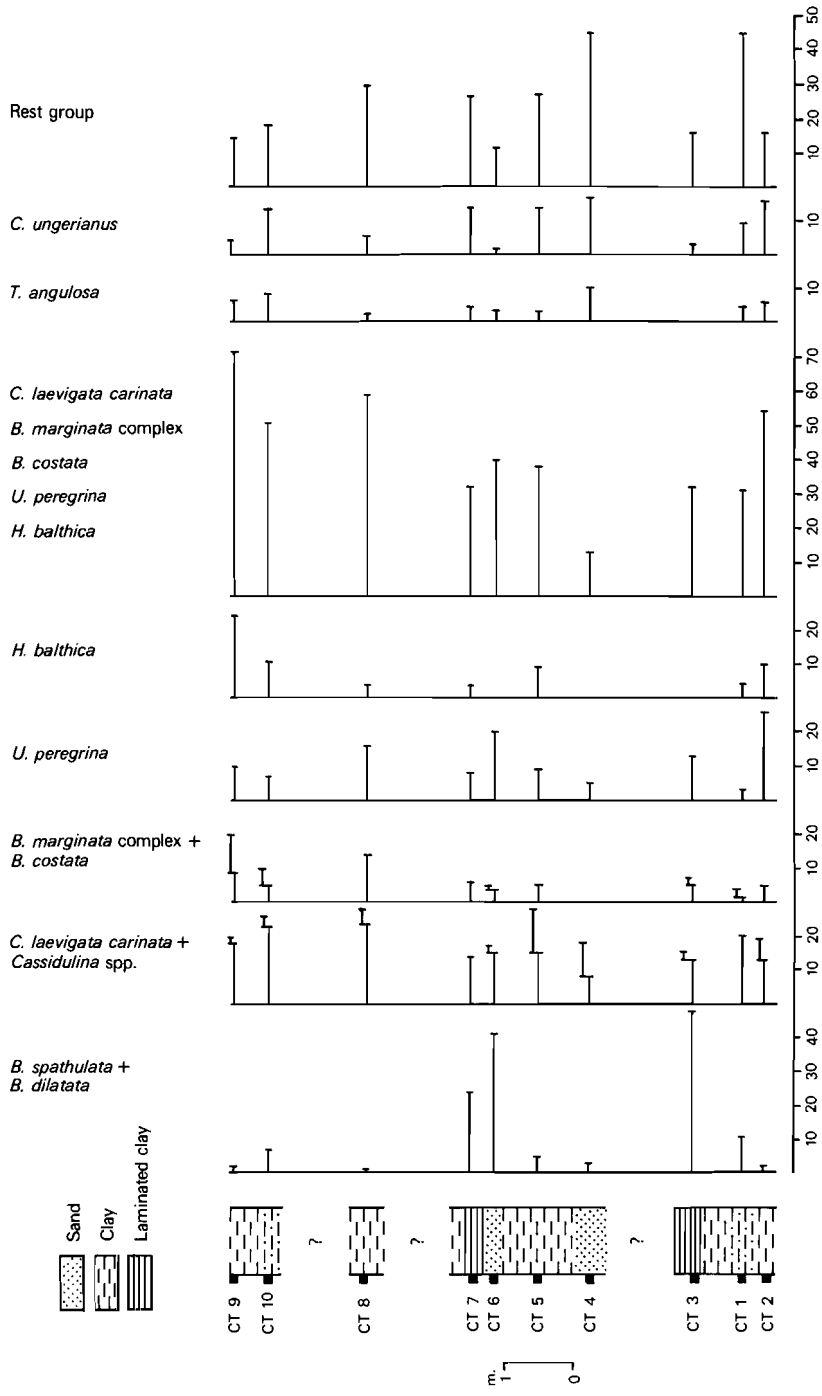



Fig. 12a Counts of 100 benthonic foraminifera from ten samples of the Castella section. See VIII.5.

Evidently the major fluctuations in the benthonic faunas are caused by stagnation of the bottom waters (lamination and *Bolivina* peaks). Since *Cibicides ungerianus* is known (G. J. van der Zwaan, pers. comm.) to be one of the first species to disappear from water with stagnant bottom conditions, it will be more frequent in homogeneous sediments. If we assume that stagnation in the Plio-Pleistocene of the Mediterranean occurred after "cold" spells, this would explain the positive correlation found by Bremer et al. between the *Cibicides* species and the "warm" planktonic foraminifera.

## Plate 1

Figs. 1, 2 — *Bolivina spathulata* (Williamson) (≈ *Brizalina difformis*, (3) *Brizalina dilatata* and (1) *Brizalina attica*).

Fig. 3 — *Bulimina marginata* d'Orbigny (≈ (4) *B. aculeata*).

Fig. 4 — *Bulimina marginata* d'Orbigny (≈ (4) *B. marginata*).

Fig. 5 — (5) *Bulimina costata* d'Orbigny.

Fig. 6 — (7) *Cassidulina crassa* d'Orbigny.

Fig. 7 — *Cibicides ungerianus* (d'Orbigny) (≈ (9) *Cibicidoides floridanus*).

Fig. 8 — *Cibicides kullenbergi* Parker (≈ (9) *Cibicidoides floridanus*).

Fig. 9 — (11) *Hyalinea balthica* (Gmelin).

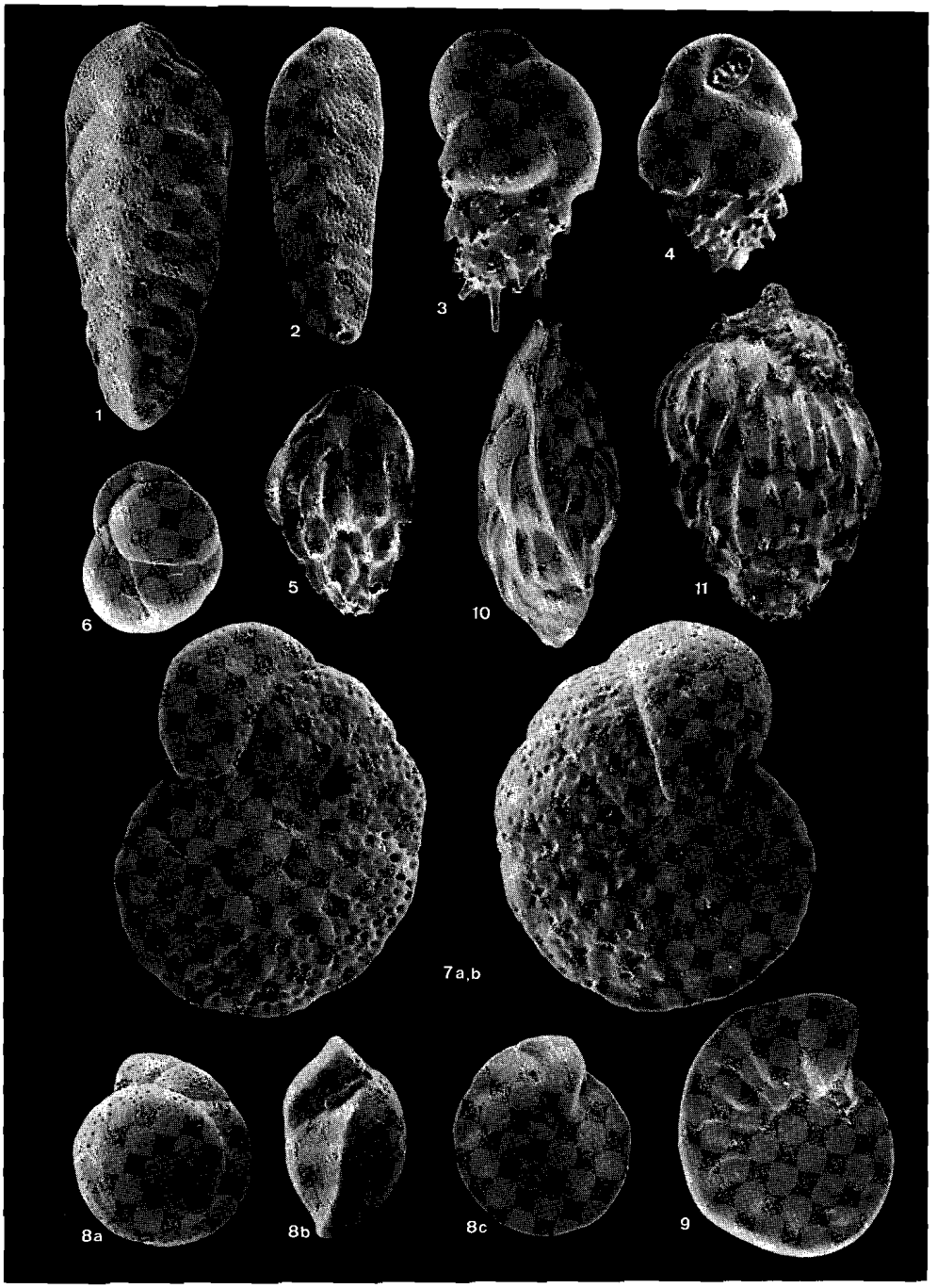
Fig. 10 — (12) *Trifarina angulosa* (Williamson).

Fig. 11 — (13) *Uvigerina peregrina* Cushman.

Figured specimens from the Plio-Pleistocene of Le Castella section (coll. Utrecht). Determinations by G. J. van der Zwaan.

All magnifications × 90, with the exception of fig. 8 (× 45).

**Plate 1**



Chapter IX

# QUANTITATIVE ANALYSIS OF THE BENTHONIC FORAMINIFERA OF DINGDEN

## IX.1. Introduction

C. W. Drooger and Felix (1961) published the counts on 49 species of benthonic foraminifera from 26 samples from a 2.60 metre section taken at the type locality of the Miocene Dingden Formation at Dingden (Western Germany). With one exception the counts contain 200 specimens.

In the authors' opinion opposed frequency fluctuation patterns of *Asterigerina gürichi* and *Spiroplectammina carinata* (plus *Martinottiella communis* and *Elphidium inflatum*) correlate with poorly understood ecological parameters (more, or less sandy). In addition, the first species shows an increasing trend, the second a decreasing one. There is practically no statistical background for these assumptions, however.

This relatively small range chart has been used for checking the successive programs for several years. At the very beginning all computer programs were only adapted to counts of fixed size. Therefore the count D463 with 60 specimens was ignored, but this is thought not to harm the analysis. In addition, all rare species were lumped into a rest group category, so that the revised species list ran as follows:

1 *Asterigerina gürichi*
2 *Bulimina elongata*
3 *Cancris auriculus*
4 *Cibicides dutemplei*
5 *Cibicides ungerianus*
6 *Elphidium inflatum*
7 *Martinottiella communis*
8 *Nonion affine*
9 *Nonion boueanum*
10 *Spiroplectammina carinata*
11 Rest group

The resulting 11 × 25 matrix of scores is presented in figure 13. It is recalled that sample D486 is lowermost in the stratigraphic column, and that D461 is uppermost. This matrix was punched and stored on permanent file before the analysis was started.

| Sample numbers | Total | Taxa | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| DL461 | 200 | 155 | 9 | 10 | 2 | 3 | 3 | 3 | 2 | 0 | 3 | 10 |
| DL462 | 200 | 156 | 14 | 0 | 6 | 2 | 7 | 1 | 5 | 2 | 0 | 7 |
| DL464 | 200 | 157 | 15 | 0 | 11 | 4 | 2 | 0 | 5 | 1 | 1 | 4 |
| DL465 | 200 | 109 | 22 | 0 | 31 | 15 | 3 | 1 | 9 | 1 | 2 | 7 |
| DL466 | 200 | 119 | 14 | 2 | 33 | 6 | 6 | 0 | 8 | 3 | 3 | 6 |
| DL467 | 200 | 107 | 15 | 2 | 32 | 14 | 6 | 1 | 12 | 1 | 5 | 5 |
| DL468 | 200 | 95 | 11 | 4 | 29 | 8 | 10 | 6 | 20 | 2 | 5 | 10 |
| DL469 | 200 | 106 | 13 | 11 | 28 | 8 | 8 | 1 | 13 | 1 | 5 | 6 |
| DL470 | 200 | 106 | 18 | 3 | 21 | 9 | 7 | 5 | 16 | 1 | 7 | 7 |
| DL471 | 200 | 75 | 16 | 5 | 24 | 3 | 7 | 8 | 17 | 6 | 29 | 10 |
| DL472 | 200 | 78 | 24 | 5 | 28 | 9 | 8 | 3 | 17 | 2 | 13 | 13 |
| DL473 | 200 | 85 | 15 | 6 | 39 | 3 | 5 | 3 | 15 | 4 | 12 | 13 |
| DL474 | 200 | 77 | 17 | 7 | 22 | 5 | 7 | 10 | 16 | 2 | 20 | 17 |
| DL475 | 200 | 95 | 18 | 4 | 15 | 5 | 8 | 5 | 15 | 5 | 17 | 13 |
| DL476 | 200 | 75 | 11 | 5 | 13 | 5 | 10 | 11 | 13 | 1 | 44 | 12 |
| DL477 | 200 | 90 | 10 | 5 | 22 | 3 | 6 | 6 | 17 | 5 | 20 | 16 |
| DL478 | 200 | 80 | 6 | 3 | 22 | 3 | 10 | 8 | 10 | 4 | 36 | 18 |
| DL479 | 200 | 82 | 12 | 5 | 22 | 2 | 7 | 5 | 16 | 7 | 20 | 22 |
| DL480 | 200 | 64 | 2 | 0 | 12 | 0 | 19 | 14 | 5 | 10 | 66 | 8 |
| DL481 | 200 | 107 | 11 | 6 | 17 | 2 | 5 | 7 | 11 | 9 | 11 | 14 |
| DL482 | 200 | 104 | 7 | 1 | 15 | 0 | 6 | 11 | 7 | 9 | 24 | 16 |
| DL483 | 200 | 66 | 3 | 1 | 22 | 0 | 19 | 13 | 3 | 5 | 57 | 11 |
| DL484 | 200 | 79 | 6 | 6 | 13 | 0 | 14 | 11 | 6 | 9 | 43 | 13 |
| DL485 | 200 | 78 | 2 | 2 | 15 | 2 | 15 | 11 | 0 | 5 | 61 | 9 |
| DL486 | 200 | 76 | 5 | 2 | 27 | 1 | 15 | 11 | 7 | 6 | 41 | 9 |

Fig. 13 Counts on benthonic foraminifera from the Dingden section (after C. W. Drooger & R. Felix, 1961).

## IX.2. Analysis of the series of counts by means of the computer programs DISTUR, BALANC and REGRES

A large part of the output from the DISTUR program is presented in figure 14. From the output and this figure we arrive at the following conclusions.

The proportions of the taxa (1) *Asterigerina gürichi* and (10) *Spiroplectammina carinata* fluctuate markedly, resulting in high chi square values (342. and 499., respectively) and high values of the parameter $H_i$ (.103 and .105, respectively). From figure 14 one can also see that there are many overall trends in the proportions. The proportions of (1) *Asterigerina gürichi*, (2) *Bulimina elongata*, and (5) *Cibicides ungerianus* tend to increase upwards along the stratigraphic column, the proportions of (6) *Elphidium inflatum*, (7) *Martinottiella communis*, (9) *Nonion boueanum*, (10) *Spiroplectammina carinata* and the Rest group (11) tend to decrease in the same direction.

It is remarkable that the variance of the open variable of (1) *Asterigerina gürichi* is the major contributor (.050) to the variance of the sum of the open variables (.066). The residue (.016) consists mainly of the variance of the open variable of (10) *Spiroplectammina carinata* (.011). Therefore *A. gürichi* in particular was suspected of disturbing the analysis (see below).

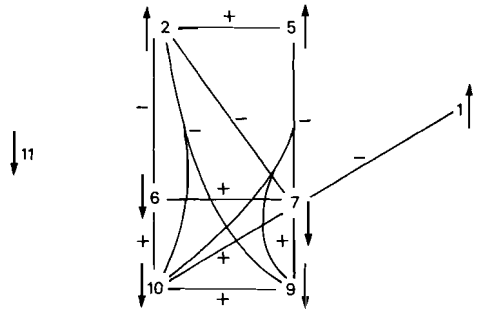



Fig. 14 Spider-web diagram for the Dingden samples showing the significant correlation coefficients and trends according to the output of the DISTUR program. Significance level $\alpha = 0.01$, two-sided.

In figure 14 the significant ($\alpha = 0.01$) correlation coefficient values between pairs of taxa are presented. There appear to be many mutual correlations, generally linked with the trends mentioned above. The taxa (2) *Bulimina elongata* and (5) *Cibicides ungerianus,* having positive trends, form a group opposite to a group of taxa having negative trends and consisting of (6) *Elphidium inflatum,* (7) *Martinottiella communis,* (9) *Nonion boueanum* and (10) *Spiroplectammina carinata.* The taxon (1) *Asterigerina gürichi* is in a somewhat isolated position, having only a negative correlation with (7) *Martinottiella communis.*

It should be noted that a correlation between two taxa is called significant if and only if both the unweighted and the weighted correlation coefficient values are significant according to the multinomial model as well as according to the zero open covariances model. If the conclusions from both models are not the same the relation between that pair of taxa is not mentioned in our conclusions. For instance: according to the multinomial model (1) *A. gürichi* and (10) *S. carinata* have a negative correlation ($\alpha = 0.01$). According to the zero open covariances model there is no such correlation ($\alpha = 0.05$) between the open variables of both taxa, however. Similarly, (4) *C. dutemplei* and (8) *N. affine* have a positive correlation according to the multinomial model, but no correlation according to the zero open covariances model. Such conflicting results are not mentioned in our conclusions.

The output from the BALANC program gives some different results (fig. 15). The taxa (1) *A. gürichi,* (4) *C. dutemplei* and (8) *N. affine* join (2) *B. elongata* and (5) *C. ungerianus.* The negative correlation between (1) *A. gürichi* and (7) *M. communis* is no longer significant. The negative trend of the Rest group (11) is no longer significant either, and the negative trend of (6) *E. inflatum* does not reach the significance level of $\alpha = 0.01$.

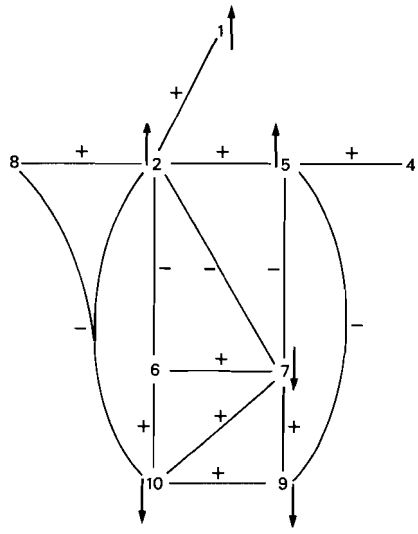



Fig. 15 Spider-web diagram for the Dingden samples showing the significant correlation coefficients and trends according to the output of the BALANC program. Significance level $\alpha = 0.01$, two-sided.

| | | $l_1$ | $l_2$ | $l_3$ | |
|---|---|---|---|---|---|
| | | 4.96 | 1.69 | 1.14 | eigenvalues |
| | | $U_1$ | $U_2$ | $U_3$ | eigenvectors |
| taxa | 1 | .40 | .11 | *−.83* | |
| | 2 | *.85* | .04 | −.15 | |
| | 3 | .27 | −.56 | −.07 | |
| | 4 | .54 | .45 | .42 | |
| | 5 | *.79* | .38 | .03 | |
| | 6 | *−.74* | .49 | −.02 | |
| | 7 | *−.88* | .04 | .07 | |
| | 8 | .64 | −.28 | .50 | |
| | 9 | *−.71* | −.13 | −.02 | |
| | 10 | *−.88* | .24 | .07 | |
| | 11 | −.28 | *−.78* | .02 | |

Fig. 16 The three largest eigenvectors with their eigenvalues of the matrix of correlation coefficients between the open variables of the Dingden samples. Significant coefficients have been printed in italics.

The values of the ratios $V_{ii}/v_{ii}$ of (1) *A. gürichi* and of the Rest group (11) are distinctly different from one, namely 2.645 and .661 respectively. From later computer runs it appears that these deviations do not harm the analyses.

The three largest principal components were extracted from the matrix of correlation coefficients between the open variables. They are presented in figure 16.

In our opinion only the first eigenvector $U_1$ has a statistical meaning, because its eigenvalue 4.96 is much greater than the other eigenvalues. The vector $U_1$ has many significant coefficients, if one uses as the critical value the square root of the 99.5-th percentile of the r-distribution: .71. From $U_1$ it can be deduced that (2) *B. elongata* and (5) *C. ungerianus* form a group opposite to a group consisting of (6) *E. inflatum,* (7) *M. communis,* (9) *N. boueanum* and (10) *S. carinata.* This result is very similar to the results obtained with the DISTUR and BALANC programs.

From the output of the REGRES program it appears that most series of proportions are autocorrelated. Another fact is that the two autocorrelation coefficients may give quite different values. For instance, for taxon (1) *A. gürichi* $R_1(4) = .260$, $R_2(4) = .544$.

For most taxa the series of proportions are autocorrelated. As a consequence, one should take care about using the critical value .506 for the correlation coefficient in the DISTUR and BALANC programs. The value of .506 is based on 23 degrees of freedom for the r-distribution and a significance level $\alpha = 0.01$. For autocorrelated series the number of degrees of freedom will be less, however. We make a rough estimate below, with help of statements made in chapter V.

We assume that for a pair of autocorrelated series

$$r(h) = (0.6) \times (0.8)^h$$

is valid (see (5.24)). In testing for independence between these two series the number of degrees of freedom is, according to (5.30)

$$df = 23 \times \frac{1 - (0.8) \times (0.8)}{1 + (2 \times (0.6) \times (0.6) - 1) \times (0.8) \times (0.8)} = 10,$$

which results in a critical value of about .71 for the correlation coefficient, using $\alpha = 0.01$.

Tests for (the absence of) a trend in such an autocorrelated series should be carried out with the number of degrees of freedom given in (5.33):

$$df = 23 \times \frac{1 - 0.8}{1 + (2 \times (0.6) - 1) \times 0.8} = 4,$$

which results in a critical value of about .92 for the correlation coefficient, using $\alpha = 0.01$.

Since such autocorrelated series are an acceptable model for many series in the Dingden range chart, namely for those of (1) *A. gürichi,* (2) *B. elongata,* (5) *C. ungerianus,* (7) *M. communis,* (8) *N. affine,* (9) *N. boueanum,* (10) *S. carinata* and the Rest group (11), according to the REGRES output, the considerations made in the section V.7 should be kept in mind. This means that the grouping of taxa shown in the spider-web diagrams of figures 14 and 15 may be caused by an overall ecological change, i.e. the trends shown by many of the taxa may be 'sustained changes', but from the point of view of mathematical statistics these suggestions cannot be proved.

The interpretation of the correlation coefficients in the matrix used for figure 15 which are still significant using the corrected critical value .71 is also problematical. There remain positive correlations between the open variables of (2) *B. elongata* and (5) *C. ungerianus,* and between those of (6) *E. inflatum,* (7) *M. communis* and (10) *S. carinata.* The conclusion from these significant values simply may be that the series under consideration should be very different from signal-plus-noise processes as defined in (5.22), because these series are far from "stationary", i.e. their probabilistic nature changes with time.

## IX.3. Elimination of taxa from the data matrix

As the scores of the taxa (1) *Asterigerina gürichi* were suspected of disturbing the analyses described in the previous section, another run of the DISTUR program was performed in which this taxon was eliminated. In figure 17 the spider-web is presented.

A positive correlation between (4) *C. dutemplei* and (5) *C. ungerianus* has appeared, which was not apparent in the previous DISTUR output (fig. 14), whereas the positive correlations between (6) *E. inflatum* and (7) *M. communis,* between (7) *M. communis* and (9) *N. boueanum,* and between (9) *N. boueanum* and (10) *S. carinata* are no longer significant at the level of $\alpha = 0.01$. Between (4) *C. dutemplei* and (7) *M. communis* a negative correlation has appeared, and the negative correlation between (2) *B. elongata* and (6) *E. inflatum* has disappeared. In the output (4) *C. dutemplei* and (8) *N. affine* show positive trends, but the negative trends of (6) *E. inflatum* and of the Rest group (11) have disappeared.

Also in this output (10) *S. carinata* has a relatively high chi square value 349. and a high H value .164. Furthermore, the variance of its open variable (.035) is the major contributor to the variance of the sum of the open

variables (.057). Therefore we also suspected that (10) *S. carinata* was disturbing the analyses and we decided to perform a third run of the DISTUR program from which both (1) *A. gürichi* and (10) *S. carinata* were eliminated.

The essential part of the output is presented in figure 18. It should be noted that by eliminating the taxa (1) *A. gürichi* and (10) *S. carinata* only about 80 specimens are left in the average count (M(N(J)) = 81.4). Furthermore,

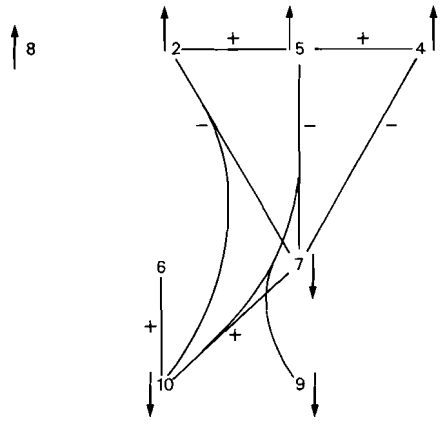



Fig. 17 Spider-web diagram for the Dingden samples showing the significant correlation coefficients and trends according to the output of the DISTUR program, after elimination of (1) *A. gürichi*. Significance level $\alpha = 0.01$, two-sided.

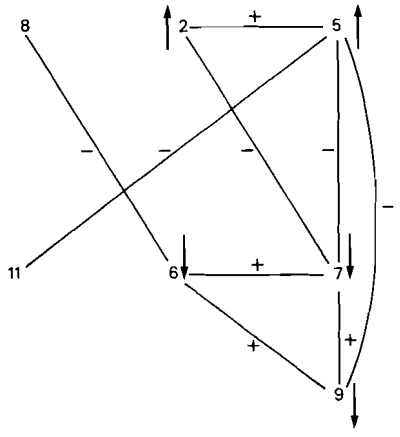



Fig. 18 Spider-web diagram for the Dingden samples showing the significant correlation coefficients and trends according to the DISTUR program, after elimination of (1) *A. gürichi* and (10) *S. carinata*. Significance level $\alpha = 0.01$, two-sided.

according to the three series of chi square values, of H values and of open variances, there is no longer any taxon disturbing the analysis. Finally, there are no distinct differences between this output and the two previous DISTUR outputs, except for the fact that amongst other things the positive trends of (4) *C. dutemplei* and of (8) *N. affine* have again disappeared.

Finally, another run of the BALANC program was performed in which the taxa (1) *A. gürichi* and (10) *S. carinata* had been eliminated in a similar way. The result is presented in figure 19. The spider-web is identical to the one of the last DISTUR program (fig. 18), except for the presence of a positive correlation between (4) *C. dutemplei* and (5) *C. ungerianus* in the BALANC output and the absence of the negative correlation between (6) *E. inflatum* and (8) *N. affine.* Such differences are often due to the strict application of the level of significance $\alpha = 0.01$ as a criterion for judging the possible significance of correlation coefficient values.

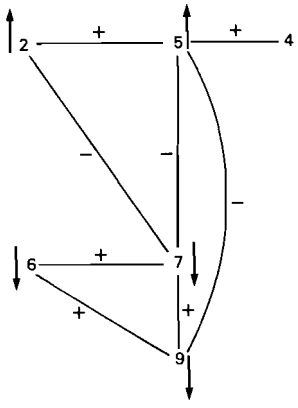



Fig. 19 Spider-web diagram for the Dingden samples showing the significant correlation coefficients and trends according to the BALANC program, after elimination of (1) *A. gürichi* and (10) *S. carinata.* Significance level $\alpha = 0.01$, two sided.

Summarizing, we conclude that two groupings can be distinguished. One consists of the taxa (1) *Asterigerina gürichi,* (2) *Bulimina elongata,* (4) *Cibicides dutemplei,* (5) *Cibicides ungerianus* and (8) *Nonion affine.* These taxa tend to increase in numbers upwards along the stratigraphic column. The other grouping consists of the taxa (6) *Elphidium inflatum,* (7) *Martinottiella communis,* (9) *Nonion boueanum,* (10) *Spiroplectammina carinata* and (probably also) the Rest group (11). It is very remarkable that (3) *Cancris auriculus* does not join either of the two groups.

## IX.4. Discussion

It is difficult to make an ecological interpretation of Miocene foraminifera of the North Sea basin because our knowledge of Recent species and their habitats in the area is very limited. As a consequence it is hard to appreciate which of our statistical procedures is the better one. An evaluation simply based on counting the notations in the columns of figure 20 is a dubious solution after all the sophisticated theory. Nevertheless we shall try to make an evaluation which seems to follow this line of reasoning.

First of all, it should be noted that some positive (*A. gürichi, B. elongata* and *C. ungerianus*) and some negative (*M. communis, N. boueanum* and *S. carinata*) trends are consistent in all analyses. We can explain this as a kind of ecology-dependent evolution, which is probably not a "sustained change" of a single factor but rather a fluctuating change of several factors together.

The assumption of Drooger and Felix (1961) that *A. gürichi* and *S. carinata* represent distinctly opposed environments does not follow clearly from the statistical analyses. It is only according to the multinomial model that there is a negative correlation between both taxa. Although data on the Oligocene and Miocene of the North Sea basin (Batjes, 1958) seem to indicate that *A. gürichi* is frequent in more sandy deposits, it seems likely that both species did not have mutually exclusive habitats, which might be interpreted in simple terms of energy or depth. Actually, *A. gürichi* seems to behave rather independently of all other taxa in our analyses, showing occasional negative and positive links with *M. communis* and *B. elongata*, respectively.

If we assume an interplay of two or more environmental gradients we had better look for the more consistent combinations in the range chart first.

One consistent combination is that of *B. elongata* and *C. ungerianus*. This combination is not illogical if we may compare both components with *B. marginata* and *C. (pseudo)ungerianus* of the Mediterranean. In the Plio-Pleistocene Pyrgos sediments (M. M. Drooger & Hageman, 1979) we would situate such a combination in the shallower part of the open marine, offshore environment. The weak tie with *C. dutemplei* would fit in very well. The occasional ties with *A. gürichi* and *N. affine* are still weaker. One might suppose that these two species had an overlap with the open marine group at its shallower and its deeper (or sandy and muddy) sides, respectively, but the explanation for the total assemblage seems to be rather artificial.

The opposite combination of *S. carinata, M. communis* and *E. inflatum* is much more problematical. Since *S. carinata* is extremely frequent in the Oligocene Boom Clay (Batjes, 1958), one might think of a (more) muddy

| | | DISTUR A | DISTUR B | BALANC | PRINCIPAL COMPONENT ANALYSIS | BALANC + REGRES | DISTUR 1 eliminated | DISTUR 1 and 10 eliminated | BALANC 1 and 10 eliminated | REGRES (AUTO-CORRELATION) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | *A. gürichi* | −10 | ↑ −7 | ↑ +2 | | | ⊘ | ⊘ | ⊘ | * |
| 2 | *B. elongata* | | ↑ +5 −(6,7,9,10) | ↑ +(1,5,8) −(6,7,10) | +5 −(6,7,9,10) | +5 | ↑ +5 −(7,10) | ↑ +5 −7 | ↑ +5 −7 | * |
| 3 | *C. auriculus* | | | | | | | | | |
| 4 | *C. dutemplei* | +8 | | +5 | | | ↑ +5 −7 | | +5 | |
| 5 | *C. ungerianus* | | ↑ +2 −(7,9,10) | ↑ +(2,4) −(7,9) | +2 −(6,7,9,10) | +2 | ↑ +(2,4) −(7,9,10) | ↑ +2 −(7,9,11) | ↑ +(2,4) −(7,9) | * |
| 6 | *E. inflatum* | | ↓ +(7,10) −2 | +(7,10) −2 | −(2,5) +(7,9,10) | +(7,10) | +10 | ↓ +(7,9) −8 | ↓ +(7,9) | |
| 7 | *M. communis* | | ↓ −(1,2,5) +(6,9,10) | ↓ +(6,9,10) −(2,5) | −(2,5) +(6,9,10) | +(6,10) | ↓ +10 −(2,4,5) | ↓ +(6,9) −(2,5) | ↓ +(6,9) −(2,5) | * |
| 8 | *N. affine* | +4 | | +2 −10 | | | ↑ | −6 | | * |
| 9 | *N. boueanum* | | ↓ +(7,10) −(2,5) | ↓ +(7,10) −5 | −(2,5) +(6,7,10) | | ↓ −5 | ↓ +(6,7) −5 | ↓ +(6,7) −5 | * |
| 10 | *S. carinata* | −1 | ↓ +(6,7,9) −(2,5) | ↓ +(6,7,9) −(2,8) | −(2,5) +(6,7,9,) | +(6,7) | ↓ +(6,7) −(2,5) | ⊘ | ⊘ | * |
| 11 | Rest group | | ↓ | | | | | −5 | | * |

Fig. 20 Diagram showing results of successive computer runs for the Dingden taxa (see text). A: significant according to the multinomial model, but not according to the zero open covariances model; B: significant according to both models.

environment possibly with some stress condition, for instance poor bottom water circulation in view of the high pyrite contents of these Oligocene sediments. The other arenaceous species, *M. communis,* is still alive but we know nothing about its ecological preferences or tolerances. The fact that it joins *S. carinata* is not thought to be anomalous. However, our ideas of the habitats of *Elphidium* species are completely different; we think they lived in shallow marine, sometimes vegetated areas. Either this typically Miocene North Sea basin *Elphidium* preferred conditions that were completely deviating, or its combination with both arenaceous species is caused by a more intricate interplay of environmental factors. *N. boueanum* is another regular addition to the group of both arenaceous species; it is an addition of equally peculiar nature. In Pyrgos it is the central element of the shallow marine group and it has a weak positive correlation with more sandy sediments. Its supposed Recent counterpart (*Nonionella atlantica,* Drooger & Kaasschieter, 1958) is frequent in the pelitic sediments of the Orinoco shelf, down to a depth of 75 metres. The combination of *N. boueanum* and *E. inflatum* would seem to be more logical than the combination with *S. carinata.* It is remarkable that the expected positive correlation between *N. boueanum* and *E. inflatum* appears only after *A. gürichi* and *S. carinata* have been eliminated as possible disturbers from the matrix.

We must conclude that both opposed groups of taxa contain queer combinations, the species of which are thought to represent different environments. If we were allowed to oppose the *Asterigerina, Elphidium* and *Nonion* species together, to the other elements we would jump to the conclusion of a simple depth or energy controlled gradient.

Since the actual groups are distinct in either positive or negative trends we must conclude that there was a time-bound local change of clusters of environments; this change fluctuated but did not evoke a complete overthrow of regimes but rather a shift in dominance. It is assumed that the sampled assemblages hardly ever reflect a single, "pure" environment, but they usually consist of mixtures from adjoining habitats, the elements of which were washed together.

We suppose that on the muddy platform corresponding to the lower part of the Dingden section there was some kind of topography with shallower, "high-energy" areas (*N. boueanum* and/or *E. inflatum*) and deeper, "low-energy" patches with slightly restricted, stagnant conditions (*S. carinata, M. communis?*). In the course of time the muddy parts lost some of their poor bottom conditions and a more open-marine, offshore environment became installed with *B. elongata* and *C. ungerianus.* The shallower parts

started to harbour another fauna in which *A. gürichi* was predominant. Although we know next to nothing about the habitat of this remarkable *Asterigerina* group which became extinct in the course of the Miocene, biogenic sedimentation may have exerted a greater influence on these shoals.

Now the explanation has become much more logical and the assumed habitats of the Miocene species are no longer in conflict with those of their Recent relatives.

This type of change in the character of the sets of subenvironments in the course of time might reflect a regressive tendency in the type section of the Dingden Formation. The parallel of such a change is much more pronounced in the Nordic Oligocene, where the pyrite-bearing, dark muds of the Rupelian Boom Clay with abundant *S. carinata* as one of its dominant elements were replaced by the sandy and calcarenitic deposits with *A. gürichi* faunas of the Late Rupelian – Chattian (Kasseler Meeressande, Asterigerinenhorizont, Zanden van Voort).

## Plate 2

Fig. 1 (1) *Asterigerina gürichi* (Franke).

Fig. 2 (2) *Bulimina elongata* d'Orbigny.

Fig. 3 (3) *Cancris auriculus* (Fichtel and Moll).

Fig. 4 (6) *Elphidium inflatum* (Reuss).

Fig. 5 (7) *Martinottiella communis* (d'Orbigny).

Fig. 6 (8) *Nonion affine* (Reuss).

Fig. 7 (9) *Nonion boueanum* (d'Orbigny).

Fig. 8 (10) *Spiroplectammina carinata* (d'Orbigny).

Figured specimens from the Miocene of the Dingden section (coll. Utrecht). Determinations by G. J. van der Zwaan, after D. A. J. Batjes, 1958.

All magnifications X 90, with the exception of figures 2, 5, 6 and 7 (X 45).

## Plate 2

Chapter X

# QUANTITATIVE ANALYSIS OF THE PLANKTONIC FORAMINIFERA OF PARKER'S CORE 189 FROM THE EASTERN MEDITERRANEAN

## X.1. Introduction

Several authors have tried to reconstruct the Late Quaternary history of the Eastern Mediterranean on the basis of the samples of Albatross core 189 from the Aegean. Parker (1958) established a warm-cold curve for the last 400,000 years on the basis of a visual impression of her countings of planktonic foraminifera. These counting data were reassessed by Ryan (1972) by means of Q-mode and R-mode factor analyses (COVAP program of Manson & Imbrie, 1964).

On the basis of the R-mode analysis Ryan discusses the meaning of the four "largest" vectors, which he calls A, B, C and D. Vector A, associated with an eigenvalue 3.953, groups the taxa *Globigerina bulloides, Globigerina pachyderma* and *Globorotalia scitula* together as having "proportionally similar relative abundances". By this vector A this group of three taxa is opposed to a group of three other taxa, consisting of *Globorotalia truncatulinoides, Globigerinella aequilateralis* and *Globigerinoides rubra,* which is in "inverse proportion" to the taxa of the first group. Ryan concludes that vector A reflects temperature and exhibits a warm-cold polarity, with the species *G. bulloides* and *G. rubra* at the cold and warm extremes, respectively.

Vector B has *Globigerina eggeri* as one extreme, which would correlate to sapropelic muds in the column, i.e. to low salinity of the upper water layer, good water stratification and stagnancy at the bottom. Vector C shows *Globigerina inflata* with *Orbulina universa* and *Globigerina digitata* versus *Globigerinita glutinata* and *Globigerinoides sacculifera,* which is interpreted as a temperature gradient indicator, i.e. warming and cooling phases. Vector D, thought to be partly an artefact of the factor analysis would correlate *Hastigerina pelagica* and *Globigerinoides conglobata* with the temperature maxima.

In our opinion Ryan's interpretation of the latter three vectors (B, C and D) is less sound, if we consider the series of eigenvalues in his table 2. Except for the first eigenvalue (= 3.953), which is considerably larger than all others, the eigenvalues (our fig. 7) form a series of steadily decreasing values

(1.744, 1.496, 1.315, 1.158, 1.049 . . .). Therefore we believe that the hypothesis (6.28) in our chapter VI must be considered, i.e. that in Ryan's analysis there is only a single meaningful axis of variation to be derived from Parker's data. In our opinion it is only his vector A which makes sense from the statistical point of view. Another drawback of Ryan's analysis is that it is based on the closed sum data directly. This is the main reason why we attempted to apply the open variables methods to Parker's set of data.

Ryan's Q-mode analysis of the 79 samples gives comparisons with reference to selected samples from the column chosen on the basis of the characteristic features warm, cold and sapropelic mud. Corresponding temperature curves for the entire core are shown; the stagnancy curve is not given but it is said to show sharp peaks in the sapropelic mud and tephra layers.

### X.2. Analysis of the series of counts using our own computer programs

We used the original data of Parker (1958) for our own computations. In her table 10 she presents percentage values for 17 species of planktonic foraminifera. Elsewhere in the publication (p. 223) it appears that about 300 specimens have been counted from each core sample. In our analysis we assumed that exactly 500 specimens had been counted, multiplying each percentage value by a factor five. In the original data set also the presence or absence of a "typical" form of *Globigerinoides rubra* is mentioned for each sample. We have added this to our data matrix by giving this typical form the score one in the case of presence and the score zero in the case of absence. Then our species list is identical to the one in Parker's data set:

1 *Globigerina bulloides*
2 *Globigerina digitata*
3 *Globigerina eggeri*
4 *Globigerina inflata*
5 *Globigerina pachyderma*
6 *Globigerina quinqueloba*
7 *Globigerina radians*
8 *Globigerinella aequilateralis*
9 *Globigerinita glutinata*
10 *Globigerinoides conglobata*
11 *Globigerinoides rubra* group
12 *Globigerinoides rubra,* typical form
13 *Globigerinoides sacculifera*

14 *Globigerinoides tenella*
15 *Globorotalia scitula*
16 *Globorotalia truncatulinoides*
17 *Hastigerina pelagica*
18 *Orbulina universa*

Seventy-five samples have been considered from core 189. Ryan (1972) put the four uppermost samples of Parker's core 194 on top of core 189 in order to create a more complete sequence from the present back into the Quaternary (Ryan, 1972, p. 152). We did likewise. Hence, the data set is a 18 X 79 matrix.

First, the data matrix was analysed using the DISTUR and BALANC programs.

In the DISTUR output only the proportions of (16) *Globorotalia truncatulinoides* show a (negative) trend upwards along the stratigraphic column (fig. 21).

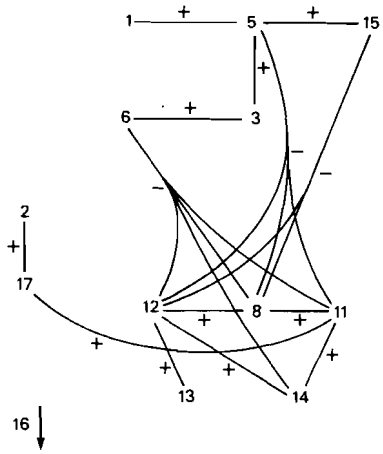



Fig. 21 Spider-web diagram for the Core 189 samples showing the significant correlation coefficients and trends according to the output of the DISTUR program. Significance level $\alpha$ = 0.01, two-sided.

The taxa (3) *Globigerina eggeri* and (11) *Globigerinoides rubra* group fluctuate markedly in their proportions, as can be concluded from the high chi square values (7996 and 7738, respectively) and from the fairly high values of the parameter $H_i$ (.217 and .243, respectively). The variance of the open variable of the (11) *G. rubra* group is a major contributor (.071) to the variance of the sum of the open variables (.131), while the variance

of the open variable of (3) *G. eggeri* contributes .030. Because of this, and because of the high chi square values mentioned above, these two taxa were suspected of disturbing the analysis (see below).

In the output of the BALANC program (fig. 22) the negative trend of (16) *G. truncatulinoides* found in the DISTUR output is confirmed by a negative trend in its open variable.

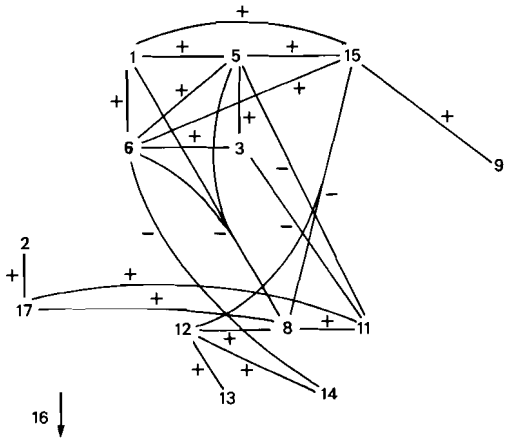



Fig. 22 Spider-web diagram for the Core 189 samples showing the significant correlation coefficients and trends according to the output of the BALANC program. Significance level $\alpha$ = 0.01, two-sided.

As far as the correlation coefficients between pairs of taxa are concerned, no essential differences were found between the DISTUR output and the BALANC output. Both show the presence of two major groups of taxa. The spider-web diagram is given for the BALANC output in figure 22, and for the DISTUR output in figure 21 (significance level $\alpha$ = 0.01). The first group consists of the taxa (1) *G. bulloides,* (3) *G. eggeri,* (5) *G. pachyderma,* (6) *G. quinqueloba,* (15) *G. scitula* and possibly (9) *G. glutinata.* The second group consists of the taxa (8) *G. aequilateralis,* (11) *G. rubra* group, (12) *G. rubra,* typical form, (13) *G. sacculifera,* (14) *G. tenella,* (17) *H. pelagica* and possibly (2) *G. digitata.* All micropaleontologists would agree that this grouping can be translated as "cold" versus "warm".

It is noted that the variance of the sum of the open variables according to the free open covariances model

$$\mathrm{var}(T) = \sum_{i=1}^{18} p_i \cdot d_i = .103$$

is considerably less then the corresponding variance according to the zero open covariances model: $S_{tt}$ = .131. It is not clear which of the two models is better.

In the output of the REGRES program it appears that the series of proportions of (4) *G. inflata,* (9) *G. glutinata* and (15) *G. scitula* are more or less strongly autocorrelated. Also the values of $R_1(1)$ and of $R_2(1)$ of (5) *G. pachyderma,* (6) *G. quinqueloba,* (11) *G. rubra* group, (12) *G. rubra,* typical form, and (17) *H. pelagica* are still significant according to the level of significance $\alpha$ = 0.01. Application of Bartlett's formula (5.32) for the number of degrees of freedom of the r-distribution reveals that the reductions in the number of degrees of freedom do not affect the above-mentioned conclusion regarding the presence of two groups of taxa.

## X.3. Elimination of taxa from the data matrix

The proportions of (3) *G. eggeri* and especially those of the (11) *G. rubra* group were suspected of disturbing the analyses described above. In a second run of the DISTUR program the (11) *G. rubra* was eliminated. It then appeared that the taxa (3) *G. eggeri,* (13) *G. sacculifera* and (16) *G. truncatulinoides* had chi square values greater than 4000, and $H_i$ values greater than .15. (3) *G. eggeri* contributes .044 to the variance of the sum of the open variables (.102), but the contributions of (13) *G. sacculifera* and (16) *G. truncatulinoides* are slight. This was the reason we made a third run of the DISTUR program and a second run of the BALANC program, from which both (3) *G. eggeri* and the (11) *G. rubra* group had been eliminated. A spider-web diagram of the DISTUR output is presented in figure 23; the diagram of the BALANC output is shown in figure 24.

There appear to be quite large differences between the two outputs. The first group consisting of the taxa (1) *G. bulloides,* (5) *G. pachyderma,* (6) *G. quinqueloba* and (15) *G. scitula* ((3) *G. eggeri* has been eliminated) is very clear in the BALANC output of figure 24, but not at all clear in the DISTUR output in figure 23. The second group consisting of the taxa (8) *G. aequilateralis* (12) *G. rubra,* typical form, (13) *G. sacculifera,* (14) *G. tenella,* (17) *H. pelagica* and possibly (2) *G. digitata* ((11) *G. rubra* group has been eliminated), shows up clearly in both outputs, but in the DISTUR output there is even one more significantly positive correlation coefficient value between the members of this group. (9) *G. glutinata* no longer shows any relation with members of the first group, but seems to form a third group together with (7) *G. radians.* In addition to the negative trend in the proportions of (16) *G. truncatulinoides,* the DISTUR output (fig. 23)

presents a positive trend in the proportions of (15) *G. scitula.* There is a striking difference between the estimates of the variance of the sum of the open variables in the two outputs. $S_{tt}$ = .089 according to the zero open covariances model, var(T) = .134 according to the free open covariances model.

Another run of the REGRES program was performed with the taxa (3) and (11) eliminated, in order to check whether this elimination influences the autocorrelation values mentioned in the previous section. The results of this second run of REGRES appeared to be nearly identical to those of the first run.

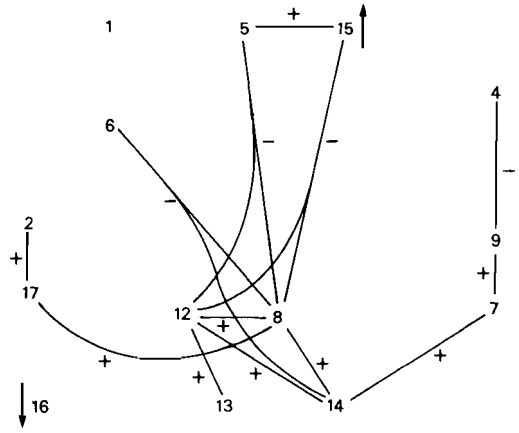



Fig. 23 Spider-web diagram for the core 189 samples showing the significant correlation coefficients and trends according to the DISTUR program, after elimination of (3) *G. eggeri* and the (11) *G. rubra* group. Significance level $\alpha = 0.01$, two-sided.

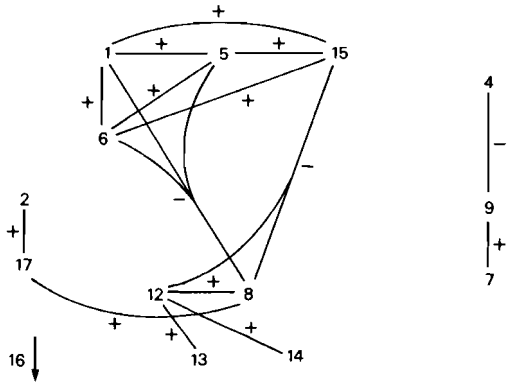



Fig. 24 Spider-web diagram for the Core 189 samples showing the significant correlation coefficients and trends according to the BALANC program, after elimination of (3) *G. eggeri* and the (11) *G. rubra* group. Significance level $\alpha = 0.01$, two-sided.

## X.4. Comparison of our results with the results of Parker

Parker (1958) selected two planktonic species groups in order to construct her temperature curves. The "warm" fauna group consists of the taxa *G. aequilateralis, G. rubra, G. sacculifera* and *H. pelagica,* the "cold" fauna group consists of *G. pachyderma* and *G. scitula.* In her figure 2 (1958, p. 235) curves are shown for the percentage values that these warm and cold fauna groups have in the counts along the column (actually, the percentages of *G. rubra* are not included in the warm fauna curve of her figure 2).

| Warm | | Cold | |
|---|---|---|---|
| Parker | this paper | Parker | this paper |
| *G. aequilateralis* | *G. aequilateralis* | *G. pachyderma* | *G. pachyderma* |
| *G. rubra* | *G. rubra* + type | *G. scitula* | *G. scitula* |
| *G. sacculifera* | *G. sacculifera* | | *G. bulloides* |
| *H. pelagica* | *H. pelagica* | | *G. eggeri* |
| | *G. tenella* | | *G. quinqueloba* |
| | *G. digitata*? | | |

As can be seen from the above table, all Parker's species appear in the correct group according to our analysis, but our "cold" group is much larger. Ryan (1972) already added *G. bulloides* to this group, but he associated the other two species (*G. eggeri* and *G. quinqueloba*) with low surface salinity and bottom stagnancy (sapropelic muds) on the basis of his vector B. Since we did not enter sediment data we did not find a separate position for *G. eggeri* and *G.quinqueloba* in our analysis. It is likely, however, that periods of lowered surface salinity are climate-controlled; this would explain why these two species join one of the temperature groups. The fact that they join the "cold" group may be due to high precipitation rates during glacial periods, but relations may be more complex. Nutrient abundance may well be more important for these species than temperature or salinity (G. J. van der Zwaan, 1982; and pers. comm.).

It is evident that the "temperature curve" based on the relative frequencies of the species of our two groups resembles the temperature curve constructed by Parker, but it differs in many details because of the high frequencies of *G. bulloides* and *G. eggeri* which are included in our "cold" group.

## X.5. Principal components analysis

For the sake of comparison with Ryan's analysis we performed an R-mode

principal components analysis on the 18 × 18 matrix of correlation coefficients between the open variables, for which the spider-web diagram was given in figure 22. The five largest principal components are presented in figure 25.

| | | $l_1$ | $l_2$ | $l_3$ | $l_4$ | $l_5$ | |
|---|---|---|---|---|---|---|---|
| | | 3.85 | 2.07 | 1.35 | 1.26 | 1.15 | eigenvalues |
| | | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | eigenvectors |
| taxa | 1 | *−.69* | .08 | .05 | .06 | .03 | |
| | 2 | .21 | *.59* | −.07 | .10 | −.13 | |
| | 3 | *−.54* | −.07 | .03 | *−.58* | −.06 | |
| | 4 | .18 | .51 | −.39 | .03 | .29 | |
| | 5 | *−.71* | .05 | .07 | .06 | −.15 | |
| | 6 | *−.69* | .02 | .11 | −.27 | .21 | |
| | 7 | .07 | −.45 | −.34 | .43 | −.17 | |
| | 8 | *.70* | −.15 | .11 | −.17 | .05 | |
| | 9 | −.34 | *−.57* | .00 | .34 | .13 | |
| | 10 | .04 | .19 | .35 | .09 | −.51 | |
| | 11 | *.57* | .10 | .18 | .26 | .23 | |
| | 12 | .43 | *−.62* | .00 | −.38 | −.03 | |
| | 13 | .29 | −.33 | .52 | −.06 | .18 | |
| | 14 | .40 | −.27 | −.43 | .08 | −.32 | |
| | 15 | *−.67* | .06 | −.04 | .34 | .12 | |
| | 16 | .26 | .14 | −.24 | −.10 | *.59* | |
| | 17 | .39 | .44 | .47 | .13 | −.10 | |
| | 18 | .11 | .28 | −.40 | −.34 | −.33 | |

Fig. 25 The five largest eigenvectors with their eigenvalues of the matrix of correlation coefficients between the open variables of the Core 189 samples. Significant values have been printed in italics.

As the eigenvalues $l_3$, $l_4$, and $l_5$ in figure 25 form a steadily decreasing series according to the expression (6.26), and only the eigenvalues $l_1$ = 3.85 and $l_2$ = 2.07 are much larger, it is only the first eigenvector $U_1$ and possibly the second $U_2$ which have a statistical meaning. Individual coefficients of the eigenvectors which may be considered statistically significant have been printed in italics in figure 25, using the square root of the 99.5-th percentile of the r-distribution, $\sqrt{.288} = .537$, as the critical value.

Then from $U_1$ it is deduced that (1) *G. bulloides*, (3) *G. eggeri*, (5) *G. pachyderma*, (6) *G. quinqueloba* and (15) *G. scitula* form a group opposed to the pair of taxa (8) *G. aequilateralis* and the (11) *G. rubra* group. According to $U_2$, which we hesitate to call a significant vector, (2) *G. digitata* is opposite to (9) *G. glutinata* and (12) *G. rubra*, typical form.

The results from our vector $U_1$ are similar to the results from Ryan's vec-

tor A. The part which (16) *G. truncatulinoides* plays in his vector A is not confirmed by our vector $U_1$; Ryan grouped this species with (8) *G. aequilateralis* and the (11) *G. rubra* group. Our vector $U_2$ cannot be retraced from Ryan's results.

The results from our vectors $U_1$ and $U_2$ of figure 25 do not match completely our own results, described in the previous sections, either. The first group described in section X.2 is recognized almost completely by vector $U_1$. The second group, however, consisting of the taxa (8) *G. aequilateralis,* (11) *G. rubra* group, (12) *G. rubra,* typical form, (13) *G. sacculifera,* (14) *G. tenella,* (17) *H. pelagica* and possibly (2) *G. digitata* is recognized by vector $U_1$ only as far as the first two taxa are concerned. Our vector $U_2$ cannot be recognized at all in the results of our previous analyses.

We prefer the results from our DISTUR and BALANC procedures to those of the principal components analyses, since the DISTUR and BALANC procedures seem to supply more understandable details.

### X.6. Q-mode analysis

As a further means of comparing our analysis with Ryan's analysis we reduced the 18 × 79 matrix to an 8 × 79 matrix. In the first column the taxa *G. bulloides, G. eggeri, G. pachyderma, G. quinqueloba* and *G. scitula* are taken together. In the second column the taxa *G. digitata, G. aequilateralis, G. rubra* group, *G. rubra* typical form, *G. sacculifera, G. tenella* and *H. pelagica* are lumped. The construction of these two groups is based on the results of the DISTUR and BALANC analyses given above. The taxa *G. inflata, G. radians, G. glutinata, G. conglobata, G. truncatulinoides* and *O. universa* are placed in the columns 3 to 8.

After a run of the PQMODE program we obtained a matrix of $ch_{ij}$ values. If $ch_{ij}$ is greater than five, the corresponding $x_{ij}$ is considered to be much larger than was expected according to the size of the count and to the mean proportion of the taxon; if $ch_{ij}$ is less than $-5$, the score $x_{ij}$ is considered to be much less than expected.

Using the data from this matrix we can subdivide the 79 core samples into three groups. Group I contains the counts/samples that have $ch_{1j} > 5$ and $ch_{2j} < -5$; group II contains the counts/samples that have $ch_{1j} < -5$ and $ch_{2j} > 5$, group III contains the remainder of the counts/samples. Figure 26 shows the list of the counts belonging to group I and those belonging to group II. In other words group I samples have "significantly" more specimens than the average of the group of taxa *G. bulloides, G. eggeri, G. pachyderma, G. quinqueloba* and *G. scitula* and have "significantly" fewer speci-

| Group I samples "cold" $ch^*_{1j} > 5, ch^*_{2j} < -5$ depth from core top | Group II samples "warm" $ch^*_{2j} > 5, ch^*_{1j} < -5$ depth from core top |
|---|---|
| | (10 cm) |
| | (20 cm) |
| | (30 cm) |
| 60 cm | 3 cm |
| 100 cm | 150 cm |
| 110 cm | 190 cm |
| 120 cm | 200 cm |
| 170 cm | 210 cm |
| 180 cm | 240 cm |
| 282 cm | 250 cm |
| 360 cm | 260 cm |
| 370 cm | 299 cm |
| 380 cm | 320 cm |
| 389 cm | 330 cm |
| 400 cm | 351 cm |
| 410 cm | 460 cm |
| 430 cm | 550 cm |
| 440 cm | 559 cm |
| 450 cm | 608 cm |
| 470 cm | 622 cm |
| 520 cm | 640 cm |
| 528 cm | 660 cm |
| 533 cm | 670 cm |
| 600 cm | 678 cm |
| 700 cm | 684 cm |
| 780 cm | 740 cm |
| 790 cm | 770 cm |

Fig. 26 "Cold" and "warm" samples from Core 189 according to our Q-mode analysis.

mens than the average of the group of taxa *G. digitata, G. aequilateralis, G. rubra* group, *G. rubra,* typical form, *G. sacculifera, G. tenella* and *H. pelagica.* For the group II samples these relations are reversed.

Formula (7.3) allows us to check whether "significant" values of $ch_{3j}$, $ch_{4j}$, $ch_{5j}$, $ch_{6j}$, $ch_{7j}$, and $ch_{8j}$ are induced by squeezing effects of $ch_{1j}$ and $ch_{2j}$, or vice versa, for any count j. The check shows that there are no such squeezing effects.

In order to check this conclusion we performed another similar run of PQMODE. The differences between this run and the previous run were that in the first column the scores of *G. eggeri* were not added and in the second column the scores of the *G. rubra* group were omitted, because these two taxa were recognized as disturbers in section X.2. The resulting

$ch_{ij}$ matrix appeared to be very similar to the previous one as far as the 3rd up to and including the 8th column are concerned.

Consequently we feel sure that our sample groups I and II may be comparable to the similar groups of somewhat different composition in Ryan's paper.

Ryan (1972, p. 156–158) also presented a Q-mode analysis of the 79 samples. He selected three vectors associated with the three largest eigenvalues 39.830, 10.432 and 4.551 (his table 4). In our opinion the third vector is not significant because the third eigenvalue 4.551 is only slightly larger than the fourth: 3.409. Therefore we shall concentrate only on his first and second vectors.

The sample from a depth of 670 cm in the core was chosen by Ryan as Reference vector A (associated with the eigenvalue 39.830). This sample appeared to come from a "warm interval". In his table 5 Ryan (p. 157–158) gives the series of values of proportional similarity of all samples to this reference sample. Comparing these values (abbreviated by us as p.s.A) to our groups I and II of figure 26, we found that our group I is best compared to the group of samples that have p.s.A values of less than .250. Only four samples do not belong to group I and yet have p.s.A $<$ .250; only five samples belonging to group I have p.s.A $>$ .250.

Group II is best compared to the group of samples that have p.s.A greater than .760. However, 12 samples belonging to group II have p.s.A $<$ .760. Five samples do not belong to group II and yet have p.s.A $>$ .760. In spite of these 17 samples these two groups match best, or differ least.

The sample from a depth of 450 cm in the core was chosen by Ryan as Reference vector B (associated with the eigenvalue 10.432). This sample appeared to be from a "cold interval". We have abbreviated the values of proportional similarity of all samples to reference sample B (Ryan, 1972, table 5, p. 157–158) to p.s.B.

Surprisingly, these p.s.B values match our groups I and II much better than the p.s.A values. We found that our group I is best compared to the group of samples that have p.s.B values greater than .650. Only one sample does not belong to group I and yet has a p.s.B $>$ .650; only four samples belonging to group I have p.s.B values $<$ .650.

Group II is best compared to the group of samples that have p.s.B values less than .250. Three samples belonging to group II have p.s.B values greater than .250, and four samples do not belong to group II and yet have p.s.B values less than .250.

Evidently the "temperature" curve along the stratigraphic column that

we might construct from our own analysis is better reflected in Ryan's second vector than in his first.

The poor results for the largest eigenvector A associated with the largest eigenvalue may be due to Ryan's computer technique. The average value of all elements outside the diagonal in the 79 × 79 cosine theta matrix was probably much larger than zero (close to one). In that case the largest eigenvector is partly an artefact of the deviation of the average cosine theta value from zero, and the second eigenvector is the best vector to describe the variations in the chart.

## Plate 3

Fig. 1 (1) *Globigerina bulloides* d'Orbigny.

Fig. 2 (2) *Globigerina digitata* Brady.

Fig. 3 (3) *Globigerina eggeri* Rhumbler.

Fig. 4 (4) *Globigerina inflata* d'Orbigny.

Fig. 5 (5) *Globigerina pachyderma* (Ehrenberg).

Fig. 6 (8) *Globigerinella aequilateralis* (Brady).

Fig. 7 (9) *Globigerinita glutinata* (Egger).

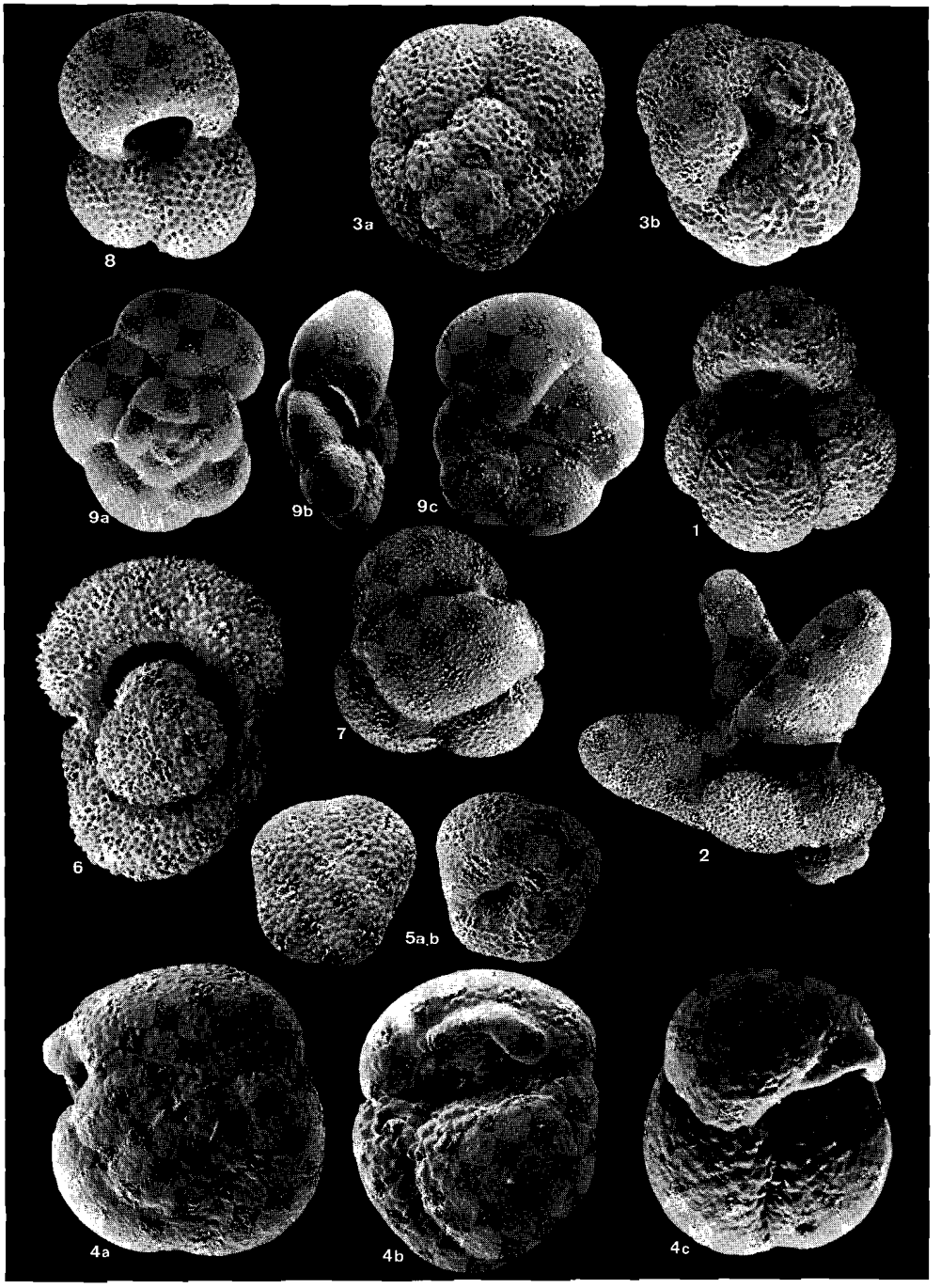
Fig. 8 (12) *Globigerinoides rubra* (d'Orbigny).

Fig. 9 (15) *Globorotalia scitula* (Brady).

Figured specimens from core 353, Quaternary of the Adriatic Sea. Determinations by G. J. van der Zwaan, after F. L. Parker, 1958.

All magnifications X 90.

## Plate 3



N.B. The text of this chapter was originally placed by the author between the present chapters I and II.

# Chapter XI

## ESTIMATING NUMERICAL PROPORTIONS

### XI.1. Estimating the numerical proportion of a frequent taxon in an assemblage

It is assumed that the binomial probability model needs little introduction. If we take a random sample (= "count") of size n from an assemblage of microfossils in which taxon A is present with a numerical proportion p, then the number X of specimens of taxon A in the random sample has a binomial distribution with parameters n and p. These two parameters n and p determine the probability distribution of X, i.e. the number of specimens of taxon A in the count. The proportion p in the assemblage is estimated by

$$\hat{p} = \frac{x}{n} \tag{11.1}$$

where X = x specimens of taxon A have been observed in the count. The standard error ("sampling error") of this estimate $\hat{p}$ is

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} = \frac{1}{n} \cdot \sqrt{\frac{x \cdot (n-x)}{n}} \tag{11.2}$$

The probability distribution of X, which depends on n and on the theoretical p, is usually formulated as

$$P(X = c) = \binom{n}{c} \cdot p^c \cdot (1-p)^{(n-c)}, \quad c = 0, 1, 2, \ldots, n \tag{11.3}$$

For our purposes the following formula, which is less well known, is more convenient (see e.g. Abramowitz & Stegun, 1964):

$$P(X < c) = P\left(\frac{c}{d} \cdot F_{(2c,2d)} > \frac{p}{(1-p)}\right) \text{ in which } d = n + 1 - c \tag{11.4}$$

In words: the probability that X, the "score" of taxon A in the count of size n, is less than the number c, is equal to the probability that a random

variable having an F distribution with $2 \cdot c$ degrees of freedom in the numerator and $2 \cdot d$ degrees of freedom in the denominator, multiplied with c/d, will exceed the value $p/(1 - p)$. With the help of a table of F distributions, which can be found in most books on mathematical statistics, the chances that X will not reach, or will exceed some number, can be calculated quite easily.

It should be noted that the complementary statement of (11.4) is

$$P(X \geqslant c) = P\left(\frac{c}{d} \cdot F_{(2c,2d)} < \frac{p}{(1-p)}\right) \tag{11.5}$$

By means of the expressions (11.4) and (11.5) a confidence interval for the assemblage parameter p can be calculated from the score $X = x$. After choosing a level of significance $\alpha$, one defines the confidence interval as containing all parameter values of p that are not rejected according to the level $\alpha$ by the outcome of X, the number of specimens of taxon A in the count.

*Example*

Five specimens of taxon A have been found in a count of size 65. We choose $\alpha = 0.01$ (one-sided), and wish to find the confidence interval for p that corresponds to this level of significance.

All values of p ($n = 65$, $x = 5$) that give

$$P(X \leqslant 5) = P(X < 6) \leqslant 0.01 \text{ must be rejected.}$$

In formula (11.4) $c = 6$, $d = 60$. The 99th percentile of $F_{12,120}$ is 2.34, so

$$P(X < 6) = P\left(\frac{6}{60} \cdot F_{12,120} > \frac{p}{1-p}\right) \leqslant 0.01.$$

It follows that $\frac{p}{1-p}$ must be greater than $\frac{6}{60} \cdot 2.34 = 0.234$, for p to be rejected. So all p values greater than 0.190 are rejected.

All values of p that give $P(X \geqslant 5) \leqslant 0.01$ must also be rejected. In formula (11.5) $c = 5$ and $d = 61$. The first percentile of $F_{10,122}$ is 0.250, so

$$P(X \geqslant 5) = P\left(\frac{5}{61} \cdot F_{10,122} < \frac{p}{1-p}\right) \leqslant 0.01.$$

It follows that $\frac{p}{1-p}$ must be less than $\frac{5}{61} \cdot 0.250 = 0.0205$ for p to be rejected. So all $p < 0.0201$ are rejected.

We conclude that for $\alpha = 0.01$ (one-sided) the confidence interval for p is [0.0201, 0.190]. All proportions p inside this interval are not rejected by the outcome $X = 5$ ($n = 65$).

It is sometimes convenient to express the confidence interval in terms of mean numbers of taxon A per count of size n (denoted by m) instead of in terms of proportions p. The relation is simply:

$$m = n \cdot p \tag{11.6}$$

in which m is the theoretical mean number of taxon A in counts of size n (from the same assemblage of course), and p is the theoretical proportion of taxon A in the assemblage.

In our example we get the confidence interval for m by multiplying the p limits with 65: [1.31 , 12.4].

It is not our intention to introduce such calculations for routine work. In general the calculation of $\hat{p} = x/n$ and its standard error $SE_{\hat{p}}$ given in (11.2) are considered to be sufficient. The underlying assumption is, however, that the confidence interval of p has the symmetrical form:

$$[\hat{p} - c \cdot SE_{\hat{p}} , \hat{p} + c \cdot SE_{\hat{p}}] \tag{11.7}$$

in which c is a positive constant, only dependent on the choice of $\alpha$. This assumption is correct if $\hat{p}$ is not too close to zero or to one. We wish to state here the rather subjective "rule" that (11.7) can be used if $c \leqslant 2.5$ and the interval $[\hat{p} - 3 \cdot SE_{\hat{p}}, \hat{p} + 3 \cdot SE_{\hat{p}}]$ is entirely within the interval [0, 1].

$$\text{Hence, } \hat{p} > 3 \cdot SE_{\hat{p}} \text{ and } 1 - \hat{p} > 3 \cdot SE_{\hat{p}}. \tag{11.7'}$$

For rare taxa that have $\hat{p} < 2 \cdot SE_{\hat{p}}$ the confidence interval is strongly asymmetrical and (11.7) is a poor approximation. For such cases the expressions (11.4) and (11.5) are useful. It will appear in the following section, however, that these expressions can be approximated by others that are much simpler and in which the confidence intervals, expressed in m (the mean number of taxon A per count of size n), are independent of n.

## XI.2. Approximation by means of the Poisson probability model

If in (11.4) and in (11.5) c is much smaller than d, the distribution of $2c \cdot F_{(2c,2d)}$ may be approximated by the chi square distribution with 2c degrees of freedom, $\chi^2_{2c}$. Equation (11.4) can be changed into

$$P(X < c) \approx P(\chi^2_{2c} > \frac{2pd}{1-p}).$$

The quantity $d/(1 - p)$ may be approximated by n. Recalling that $m = n \cdot p$, we get from (11.4):

$$P(X < c) \approx P(\chi^2_{2c} > 2m) \tag{11.8}$$

and from (11.5):

$$P(X \geqslant c) \approx P(\chi^2_{2c} < 2m). \tag{11.9}$$

It is shown by these formulae that small scores are dependent on one parameter only: the mean number of the taxon per count of fixed size.

We might consider the confidence interval of m, if X = 0 has been observed, i.e. not a single specimen of taxon A was found in the count. From (11.8) we derive ($\alpha = 0.01$):

$$P(X = 0) = P(X < 1) \approx P(\chi^2_2 > 2m) \leqslant 0.01$$

which must hold for all m to be rejected. It follows that $2m > 9.210$ and thus $m > 4.605$.

Table 2 shows the confidence intervals for m for all outcomes of X less than or equal to ten; this table has been calculated from (11.8) and (11.9) with the help of a table of distributions of chi square.

TABLE 2

| Outcome | Confidence intervals for m | |
|---|---|---|
| | $\alpha = 0.05$ | $\alpha = 0.01$ |
| X = 0 | [0 , 3.00] | [0 , 4.61] |
| X = 1 | [0.05 , 4.74] | [0.01 , 6.64] |
| X = 2 | [0.36 , 6.30] | [0.15 , 8.41] |
| X = 3 | [0.82 , 7.75] | [0.44 , 10.1 ] |
| X = 4 | [1.37 , 9.15] | [0.82 , 11.6 ] |
| X = 5 | [1.97 , 10.5 ] | [1.28 , 13.1 ] |
| X = 6 | [2.61 , 11.8 ] | [1.79 , 14.6 ] |
| X = 7 | [3.29 , 13.1 ] | [2.33 , 16.0 ] |
| X = 8 | [3.98 , 14.4 ] | [2.91 , 17.4 ] |
| X = 9 | [4.70 , 15.7 ] | [3.51 , 18.8 ] |
| X = 10 | [5.43 , 17.0 ] | [4.13 , 20.1 ] |

As can be deduced from the beginning of this section, the approximations in table 2 are better, the larger is the ratio n/X.

In the previous section the confidence interval [1.31 , 12.4] was calculated for m, for the example of X = 5, n = 65 and $\alpha = 0.01$. The approximation [1.28 , 13.1] in table 2 appears to be very close to the confidence interval calculated earlier, in spite of the small size of the count.

From table 2 several deductions can be made. Assume two specimens of

taxon A have been found in a count of size n = 200. The proportion of taxon A in the assemblage is estimated to be $\hat{p} = 2/200 = 0.01$. This estimate contains a large error, however. Table 2 tells us that the estimates $\hat{p} = 0.36/200 = 0.0018$ and $\hat{p} = 6.30/200 = 0.0315$ are still acceptable at $\alpha = 0.05$, and that the estimates $\hat{p} = 0.15/200 = 0.0008$ and $\hat{p} = 8.41/200 = 0.0421$ are still acceptable at the level of significance $\alpha = 0.01$. Note that for $\alpha = 0.01$ the largest acceptable estimate is 50 times as large as the smallest acceptable estimate.

If only one specimen of A had been observed in a count of size 200, the estimate would be $\hat{p} = 0.005$, but the estimates $\hat{p} = 0.05/200 = 0.00025$ and $\hat{p} = 4.74/200 = 0.0237$ are still acceptable at $\alpha = 0.05$, and the estimates $\hat{p} = 0.01/200 = 0.00005$ and $\hat{p} = 6.64/200 = 0.0332$ are still acceptable at $\alpha = 0.01$.

It appears that it is hardly possible to make inferences about numerical proportions from scores X = 0, 1, 2, up to 10, and even impossible for the smaller scores X among them. This conclusion led to the statistical treatment of a special counting technique for rare taxa which has been described extensively by the author (M. M. Drooger, in Zachariasse et al., 1978). A review is given in the following sections.

Two further remarks must be made. Firstly, it can be seen from table 2 that for X increasing up to ten, the confidence intervals tend to the symmetrical form given in (11.7). For X greater than ten the expression in (11.7) may be considered to be a good approximation. This is confirmed by the fact that the inequality $\hat{p} > 3 \cdot SE_{\hat{p}}$ of (11.7′) implies that $n \cdot \hat{p} = x > 9$, as can be deduced from (11.2).

The second remark concerns the nature of random variables X that fulfil for a fixed value of m the equation $P(X < c) = P(\chi^2_{2c} > 2m)$ for each positive integer c. Such random variables X must have a Poisson distribution with parameter m, defined by

$$P(X = y) = \frac{m^y \cdot e^{-m}}{y!}; \quad y = 0, 1, 2, 3, \ldots \tag{11.10}$$

in which $y! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (y-1) \cdot y$.

Such distributions are related to the Poisson probability model which describes repeated haphazard occurrences in time and space. If m is the mean number of such occurrences per unit of time, length, area or volume, the number per unit, X, has the distribution given in (11.10). It is thereby concluded that the numbers of specimens of a rare taxon in a series of counts of equal size from the same assemblage have a Poisson distribution, and that the

appearance of the rare taxon in a count can be regarded as a haphazard occurrence.

### XI.3. Logarithm-transformed estimates of proportions of rare taxa

A procedure, which yields a more accurate estimate of the numerical proportion of a rare taxon, has been elaborated by the author in Zachariasse et al. (1978). It can be summarized as follows. During a systematic examination of the contents of a tray or smear slide a number of at least ten specimens of the rare taxon is counted. This number is denoted by K. An estimate, n, is made of the total number of specimens in the searched area. From these two numbers the logarithm-transformation is given of the percentage value of the rare taxon in the assemblage:

$$L = {}^{10}\log(100 \cdot \hat{p}) = {}^{10}\log\left(\frac{100 \cdot K}{n}\right) \tag{11.11}$$

Because n is an estimate of the total number, extra errors are introduced into L, in addition to the "binomial" error. The error of n is estimated from a preceding series of counts of all specimens on fields of the picking tray or on tracks over the smear slide.

If we consider this process of calculating L as if it were repeatable on the same assemblage, our first correction is that K cannot be considered to have a binomial distribution, because n is not a constant number.

We consider first the case of microfossils on a picking tray. Let K be the number of individuals of taxon A that have been found on a number M square fields of the picking tray. The number of microfossils in a square field is denoted by d ("density"). The number present on the M fields, n', fulfils the equality

$$n' = \sum_{i=1}^{M} d_i \tag{11.12}$$

As a consequence:

$$E(n') = M \cdot E(d); \quad \mathrm{var}(n') = M \cdot \mathrm{var}(d) \tag{11.13}$$

in which E means "expected value of" and var "the variance of". The expected value of K can be considered to depend on n':

$$E(K \mid n') = p \cdot n', \tag{11.14}$$

in which p is the proportion of the rare taxon A in the assemblage. Integrating over all possible outcomes of n′ gives:

$$E(K) = E(E(K \mid n')) = p \cdot E(n') = p \cdot M \cdot E(d). \tag{11.15}$$

The calculation of var(K) requires a similar double integration:

$$\begin{aligned} \mathrm{var}(K) &= E(E((K - E(E(K \mid n')))^2 \mid n')) = \\ &= E(E((K - E(K \mid n') + E(K \mid n') - E(E(K \mid n')))^2 \mid n')) = \\ &= E(E((K - E(K \mid n') + p \cdot n' - p \cdot E(n'))^2 \mid n')) = \\ &= E(E((K - E(K \mid n'))^2 \mid n')) + E(E((p \cdot n' - p \cdot E(n'))^2 \mid n')). \end{aligned}$$

For a fixed value of n′, K can be considered to have a binomial distribution. A is a rare taxon, so this distribution may be approximated by a Poisson distribution, which implies that

$$\mathrm{var}(K \mid n') \doteq E((K - E(K \mid n'))^2 \mid n') = E(K \mid n')$$

and we come to

$$\begin{aligned} \mathrm{var}(K) &= E(E(K \mid n')) \cdot E(E((p \cdot n' - p \cdot E(n'))^2 \mid n')) = \\ &= p \cdot E(n') + p^2 \cdot \mathrm{var}(n') = p \cdot M \cdot E(d) + p^2 \cdot M \cdot \mathrm{var}(d). \end{aligned} \tag{11.16}$$

Now we must consider how the number n in formula (11.11) is estimated. In a series of N square fields which are required to be randomly chosen and to have no special relation with the M fields from which the K specimens of the rare taxon have been counted, all specimens have been counted, resulting in N "densities" $d_1, d_2, \ldots, d_N$. The number n in formula (11.11), which in fact is an estimate of n′, is calculated as:

$$n = \frac{M}{N} \cdot \sum_{j=1}^{N} d_j \tag{11.17}$$

It follows that

$$\begin{aligned} E(n) &= \frac{M}{N} \cdot \sum_{j=1}^{N} E(d) = M \cdot E(d) \\ \mathrm{var}(n) &= \frac{M^2}{N^2} \cdot \sum_{j=1}^{N} \mathrm{var}(d) = \frac{M^2}{N} \cdot \mathrm{var}(d) \end{aligned} \tag{11.18}$$

Now we are able to establish var(L). We repeat (11.11):

$$L = {}^{10}\log\left(\frac{100 \cdot K}{n}\right) = 2 + {}^{10}\log(K) - {}^{10}\log(n). \tag{11.19}$$

The number n has been established from the investigation of N fields of the tray before any rare taxon has been considered on the M fields, which M fields bear no special relation to those N fields, so n and K are mutually independent and

$$\mathrm{var}(L) = \mathrm{var}({}^{10}\log(K)) + \mathrm{var}({}^{10}\log(n)) =$$

$$= \left(\frac{{}^{10}\log(e)}{E(K)}\right)^2 \cdot \mathrm{var}(K) + \left(\frac{{}^{10}\log(e)}{E(n)}\right)^2 \cdot \mathrm{var}(n) \qquad (11.20)$$

Substitution of (11.15), (11.16) and (11.18) into (11.20) gives:

$$\mathrm{var}(L) = ({}^{10}\log(e))^2 \cdot \left(\frac{1}{p \cdot M \cdot E(d)} + \frac{\mathrm{var}(d)}{M \cdot (E(d))^2} + \frac{\mathrm{var}(d)}{N \cdot (E(d))^2}\right) \qquad (11.21)$$

At this point we have to look at the sample statistics. Out of the N densities the sample mean value is calculated:

$$\bar{d} = \left(\sum_{j=1}^{N} d_j\right)/N \qquad (11.22)$$

and the sample variance, which is the square of the sample standard deviation:

$$SD_d^2 = \left(\sum_{j=1}^{N} (d_j - \bar{d})^2\right) /(N-1). \qquad (11.23)$$

With the help of (11.17) we write (11.11) in the form

$$L = {}^{10}\log\left(\frac{100 \cdot K}{M \cdot \bar{d}}\right) \qquad (11.24)$$

So the assemblage parameter p is estimated by $K/(M \cdot \bar{d})$, $E(d)$ by $\bar{d}$, and $\mathrm{var}(d)$ by $SD_d^2$. From (11.21) we then can write the standard error of L:

$$SE_L = {}^{10}\log(e) \cdot \sqrt{\frac{1}{K} + \frac{SD_d^2}{M \cdot \bar{d}^2} + \frac{SD_d^2}{N \cdot \bar{d}^2}} =$$

$$= {}^{10}\log\left(1 + \sqrt{\frac{1}{K} + \frac{SD_d^2}{M \cdot \bar{d}^2} + \frac{SD_d^2}{N \cdot \bar{d}^2}}\right) \qquad (11.25)$$

The last equality is in fact an approximation, which is a good one if the square root expression is much less than one. In our experience this condition is satisfied.

Expressions (11.22) up to and including (11.25) form the basic formulae for estimating proportions of rare taxa from a picking tray. They were presented by the author and practised by others in the paper of Zachariasse et al. (1978). The author's presentation in that paper was rather short and not absolutely correct from a mathematical point of view; this does not devalue the applications, however. The standard error of L given in (11.25) was split in 1978 into two mutually independent components:

the "binomial" error $^{10}\log\left(1+\sqrt{\frac{1}{K}}\right)$

and the "density" error $^{10}\log\left(1+\frac{SD_d}{\bar{d}}\cdot\sqrt{\frac{1}{N}+\frac{1}{M}}\right)$

The name "density error" seems correct because it contains the sample statistics of d, namely $\bar{d}$ and $SD_d$. The name "binomial error" is not strictly correct, because K does not have a binomial distribution, as has been argued above.

Two essential points should be noted. Firstly, the distribution of microfossils over the tray is not considered to be even. In the case of evenness, the densities d would have a Poisson distribution, so that $SD_d \sim \sqrt{\bar{d}}$ and the density error would be:

$$^{10}\log\left(1+\sqrt{\frac{1}{N\cdot\bar{d}}+\frac{1}{M\cdot\bar{d}}}\right).$$

This would imply that the "binomial" error is much larger than the density error. In reality, both errors can be of equally-large size, because microfossils on a picking tray tend to cluster, so the densities d can be much more variable than expected according to the Poisson model, i.e. $SD_d \gg \sqrt{\bar{d}}$. The reader is referred to the author's earlier publication in Zachariasse et al. (1978).

The second point is that the technique only leads to reliable results if the number M of the fields searched through is much greater than the number N of the fields from which all specimens have been counted. If N fields are sufficient to make a reliable estimate of the proportion of the rare taxon, the N fields, the total numbers of which are exactly known, can be used; the simple binomial formulae given in (11.2) are then correct.

## XI.4. Logarithmic estimates in the case of calcareous nannofossils

In the case of calcareous nannofossils in smear slides, the procedure of

establishing a logarithm-transformed estimate of the proportion of a rare taxon is quite different. In order to estimate the density of nannofossils in the smear slide, one makes a track over the slide until a fixed number A of nannofossils has been counted. The traverse length needed to reach this number A, denoted by x, is recorded. A number N of such tracks are made, "randomly" distributed over the smear slide, resulting in N measurements $x_1, x_2, x_3, \ldots, x_N$ of traverse lengths.

Let us assume that one or more tracks were needed, having a total length X, to count K specimens of a rare taxon. K should again be at least ten. The number n′ of nannofossils present on the traverse length X is not described as easily as in the case of the picking tray. We need formulae for E(n′) and var(n′) like the ones in (11.13), however.

If we find A specimens on a track with mean length E(x), E(n′) can be approximated very well by

$$E(n') = \frac{X \cdot A}{E(x)}. \tag{11.26}$$

For establishing var(n′), one defines the random variable $n_x$, which indicates the number of fossils to be found on a fixed traverse length E(x), so $E(n_x) = A$ and

$$\frac{\text{var}(n_x)}{(E(n_x))^2} = \frac{\text{var}(x)}{(E(x))^2}$$

The last equality states that the coefficients of variability of x and of $n_x$ must be equal. We conclude that

$$\text{var}(n_x) = \frac{A^2 \cdot \text{var}(x)}{(E(x))^2}$$

The expression (11.26) can now be written:

$$E(n') = \frac{X}{E(x)} \cdot E(n_x),$$

which tells us that the number to be found in a track that is X/(E(x)) times longer can be expected to be X/(E(x)) times larger. Such a relation, however, is also true for the variances:

$$\text{var}(n') = \frac{X}{E(x)} \cdot \text{var}(n_x),$$

so it has been deduced that

$$\text{var}(n') = \frac{A^2 \cdot X \cdot \text{var}(x)}{(E(x))^3}. \tag{11.27}$$

The equalities (11.15) and (11.16) turn out to be:

$$E(K) = p \cdot E(n') = \frac{p \cdot X \cdot A}{E(x)}$$

and (11.28)

$$var(K) = p \cdot E(n') + p^2 \cdot var(n') = \frac{p \cdot X \cdot A}{E(x)} + \frac{p^2 \cdot A^2 \cdot X \cdot var(x)}{(E(x))^3}$$

In the logarithm-transformed proportion formula

$$L = {}^{10}\log\left(\frac{100 \cdot K}{n}\right) = 2 + {}^{10}\log(K) - {}^{10}\log(n)$$

the number n is an estimate of n', based on the traverse lengths $x_1, x_2, x_3, \ldots, x_N$ needed to count N times A nannofossils:

$$n = \frac{N \cdot X \cdot A}{\left(\sum_{i=1}^{N} x_i\right)} \tag{11.29}$$

It follows that

$$E(n) = \frac{X \cdot A}{E(x)}$$

and

$$var({}^{10}\log(n)) = var({}^{10}\log(\sum_{i=1}^{N} x_i)) = \left(\frac{{}^{10}\log(e)}{E(\sum_{i=1}^{N} x_i)}\right)^2 \cdot var(\sum_{i=1}^{N} x_i) =$$

$$= \left(\frac{{}^{10}\log(e)}{N \cdot E(x)}\right)^2 \cdot N \cdot var(x) \tag{11.30}$$

The track of length X and the N tracks in each of which A nannofossils were counted are supposed to have no part in common. Hence, the number K and the number n are mutually independent, so

$$var(L) = var(2 + {}^{10}\log(K) - {}^{10}\log(n)) = var({}^{10}\log(K)) + var({}^{10}\log(n)) =$$

$$= \left(\frac{{}^{10}\log(e)}{E(K)}\right)^2 \cdot var(K) + var({}^{10}\log(n)) \tag{11.31}$$

Substitution of (11.28) and (11.30) into (11.31) gives:

$$\mathrm{var}(L) = ({}^{10}\log(e))^2 \cdot \left( \frac{E(x)}{p \cdot X \cdot A} + \frac{\mathrm{var}(x)}{X \cdot E(x)} + \frac{\mathrm{var}(x)}{N \cdot (E(x))^2} \right) \qquad (11.32)$$

Again we have to look at the sample statistics. From the N traverse lengths we can calculate the sample mean value and the sample variance

$$\bar{x} = \left( \sum_{i=1}^{N} x_i \right)/N; \quad SD_x^2 = \left( \sum_{i=1}^{N} (x_i - \bar{x})^2 \right)/(N-1) \qquad (11.33)$$

With the help of (11.29) we write

$$L = {}^{10}\log \left( \frac{100 \cdot K \cdot \bar{x}}{A \cdot X} \right) \qquad (11.34)$$

So the assemblage parameter p is estimated by $(K \cdot \bar{x})/(A \cdot X)$, $E(x)$ by $\bar{x}$, and $\mathrm{var}(x)$ by $SD_x^2$. From (11.32) we deduce the standard error of L:

$$SE_L = {}^{10}\log(e) \cdot \sqrt{\frac{1}{K} + \frac{SD_x^2}{X \cdot \bar{x}} + \frac{SD_x^2}{N \cdot (\bar{x})^2}} =$$

$$= {}^{10}\log \left( 1 + \sqrt{\frac{1}{K} + \frac{SD_x^2}{X \cdot \bar{x}} + \frac{SD_x^2}{N \cdot (\bar{x})^2}} \right). \qquad (11.35)$$

In normal cases the square root expression, w, is much less than one, so the approximation of ${}^{10}\log(e) \cdot w$ by ${}^{10}\log(1 + w)$ is justified.

Expressions (11.33), (11.34) and (11.35) are the basic formulae for estimating proportions of rare nannofossil taxa from a smear slide, described and practised already in Zachariasse et al. (1978). The presentation in that paper is not entirely correct from a mathematical point of view (similar to the difference in the presentation of the logarithmic estimates for microfossils on a picking tray). The standard error of L given in (11.35) was split into a "binomial" part and a "density" part, which were considered to be mutually independent:

the "binomial" error ${}^{10}\log \left( 1 + \sqrt{\frac{1}{K}} \right)$

and the "density" error ${}^{10}\log \left( 1 + \frac{SD_x}{\bar{x}} \cdot \sqrt{\frac{\bar{x}}{X} + \frac{1}{N}} \right)$.

The name "density error" stems from the statistics $\bar{x}$ and $SD_x$ of the series

$x_1, x_2, x_3, \ldots, x_N$, which in fact measure the density of the nannofossils in the smear slide. The name "binomial error" is not correct, however, because K does not have a binomial distribution.

The other remarks we made for the picking tray case also apply to the smear slide case. The distribution of nannofossils over the smear slide is not considered to be even. If it were the "binomial" error would be much larger than the "density" error (M. M. Drooger, in Zachariasse et al., 1978).

It is again obvious that the technique only leads to reliable results if

$$X \gg \sum_{i=1}^{N} x_i.$$

In other cases the N "density" tracks can be used, assuming that they are "retraceable", to count a sufficient number of individuals of the rare taxon, because the total number of fossils in these tracks is known to be (approximately) N.A, so that again the binomial formulae given in (11.2) can be used.

Examples of logarithmic estimates, for calcareous nannofossils as well as for foraminifera, have been given in Zachariasse et al. (1978). That paper also included an investigation into the evenness of the distribution of foraminifera over picking trays and of calcareous nannofossils over smear slides. It appeared that for high densities of microfossils the $d_i$ measured from a picking tray and the $x_i$ measured from a smear slide have variances that are too large for the hypothesis of even distribution to be acceptable.

Once this hypothesis is rejected, the density errors can no longer be neglected relative to the "binomial" errors. The examples showed that both types of errors can be of comparable size.

N.B. The author originally placed this text between the present chapters VII and VIII.

# Chapter XII

# R-MODE SIMILARITY COEFFICIENTS TO BE USED FOR SETS OF COUNTS

## XII. Introduction

In the previous chapters sets of counts were compared (as far as R-mode analyses are concerned) by means of the correlation coefficient statistic. Such a set of counts was assumed to come from one single stratigraphic section having more or less similar lithology. This assumption gives some foundation for the open variables concept that states that the counts are "random" realisations of a multivariate probability distribution. This multivariate probability distribution may "gradually" change with time, but not in an extreme sense.

The application of this concept to sets of counts from several different stratigraphic sections taken together is of course bound to give rise to problems. We did in fact encounter a number of problems during our computer analysis of the data on benthonic foraminifera from the Pyrgos sediments (M. M. Drooger & Hageman, 1979). Some of these counts appeared to differ markedly from each other, because they correspond to completely different biotopes. The sets of counts can be split roughly into parts, each part representing one biotope. When we took the entire data set as a whole during the numerical analysis we obtained results that led us to doubt the applicability of the open variables model and question the correctness of using the correlation coefficient statistic in general, in cases where there were extreme differences within the set of counts. From the fact that many of the taxa may be absent in a large part of the set of counts we came to the conclusion that the concept of "similarity" between taxa frequencies (similar biotope) is more convenient than the concept of "correlation" between taxa frequencies in cases of such extreme variation.

For each pair of taxa i and k one must mention the number of counts in which

a) both taxa are present, $N_{i,k}$
b) taxon i is present and taxon k absent, $P_{i,k}$
c) taxon k is present and taxon i absent, $P_{k,i}$, and
d) both taxa are absent, $A_{i,k}$.

If there is no count in which taxa i and k are both present, i.e. $N_{i,k} = 0$, we shall call the relation between the taxa "completely dissimilar", regardless of the values of the other three characteristic numbers $P_{i,k}$, $P_{k,i}$ and $A_{i,k}$. The value of the correlation coefficient in such a "completely dissimilar" case may be strongly dependent on the size of $A_{i,k}$, relative to the size of N, the total number of counts. This feature is visualized in figure 27, and it was mentioned in Drooger and Hageman (1979) in a similar way. In fig. 27A the large numbers $P_{i,k}$ and $P_{k,i}$ lead to a distinctly negative value of

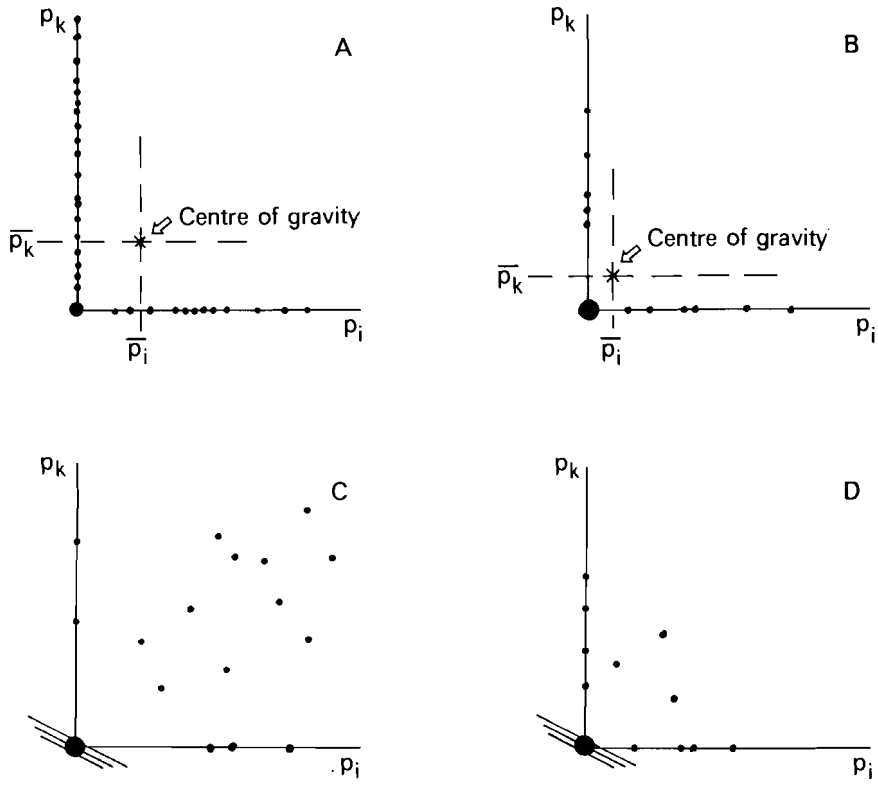



Fig. 27 Hypothetical scatter diagrams of the proportion $p_i$ of taxon i versus the proportion $p_k$ of taxon k. Examples A and B illustrate the problem concerning the correlation coefficient $r(p_i, p_k)$ in the case of "complete dissimilarity". Examples C and D illustrate similar problems concerning $r(p_i, p_k)$ in a "highly similar" case and in an "average" case, respectively.

the correlation coefficient between the proportions of taxon i and those of taxon k, if the number of counts in which both taxa are present ($N_{i,k}$) is zero and $A_{i,k}$ is small. In fig. 27B the number $A_{i,k}$ is much larger, but the correlation coefficient between the proportions is still on the negative side, although close to zero. Hence, the value of the correlation coefficient between the proportions of taxon i and those of taxon k (that are not present together in any count) is dependent on the relative number of counts in which both taxa are absent. This conclusion is not very satisfying, and it indicates that the establishment of the correlation is not a good procedure in such cases.

The problem is not solved if we take only those counts into account in which at least one of the two taxa is present in calculating the correlation coefficient. As can be seen from figures 27C and D the procedure of eliminating the scores at or near the origin still affects the correlation coefficient, but now in a negative sense. Highly "similar" taxa, i.e. taxa with $P_{i,k}$ and $P_{k,i}$ close to zero, may still show a correlation coefficient value close to zero, when the larger number of $A_{i,k}$ counts is not considered (fig. 27C). On the other hand, a pair of taxa each of which has only a few occurrences without numerical dependence on the other, will tend to show a negative correlation coefficient value, if the counts in which both taxa are absent are discarded (fig. 27D). Finally, elimination of the points near the origin is not logical from a theoretical point of view: the correlation coefficient is the expression of the relation between the values of the proportion of one taxon and the values of the proportion of another taxon. If the proportion of a taxon is allowed to be zero, the hypothesis of independence comes into conflict with our decision to eliminate counts in which both taxa have proportions equal to zero.

Similarity coefficients, already introduced in the introduction of chapter VII, and to be defined in the coming sections, are thought to be better tools than the correlation coefficient, or at least to be parameters that provide more useful information if the closed data set shows extreme variations. These similarity coefficients disregard the counts in which both taxa to be considered are absent, whether such counts are numerous or not. Only those counts are considered in which at least one of the taxa is present.

## XII.2. Cosine theta coefficients

The unweighted estimate of the correlation coefficient between the proportions $p_{ij}$ of taxon i and the proportions $p_{kj}$ of taxon k, $\bar{R}(\hat{p}_{ij}, \hat{p}_{kj})$, the formula for which is given in (2.38), can be interpreted as the cosine of the

angle between the vectors $(\hat{p}_{i1} - \bar{p}_i, \hat{p}_{i2} - \bar{p}_i, \hat{p}_{i3} - \bar{p}_i, \ldots, \hat{p}_{iN} - \bar{p}_i)$ and $(\hat{p}_{k1} - \bar{p}_k, \hat{p}_{k2} - \bar{p}_k, \hat{p}_{k3} - \bar{p}_k, \ldots, \hat{p}_{kN} - \bar{p}_k)$ in N-dimensional space. If we aim at correlation, each value $\hat{p}_{ij}$ and $\hat{p}_{kj}$ is compared to the mean values $\bar{p}_i$ and $\bar{p}_k$, respectively. If we aim at similarity, one may use the cosine of the angle between the vectors $(\hat{p}_{i1}, \hat{p}_{i2}, \hat{p}_{i3}, \ldots, \hat{p}_{iN})$ and $(\hat{p}_{k1}, \hat{p}_{k2}, \hat{p}_{k3}, \ldots, \hat{p}_{kN})$, which results in the similarity coefficient:

$$Sm_1(i,k) = \frac{\sum_{j=1}^{N} \hat{p}_{ij} \cdot \hat{p}_{kj}}{\sqrt{\left(\sum_{j=1}^{N} \hat{p}_{ij}^2\right) \cdot \left(\sum_{j=1}^{N} \hat{p}_{kj}^2\right)}} \tag{12.1}$$

Such a formula was applied by Imbrie and Purdy (1962). We do not fully agree with their mathematical interpretation of what they call the "cosine theta formula". In our opinion there is no need to worry about the meaning of the angle theta. The fact that theta may be expressed in degrees does not mean that theta has a dimension, as they assume. Degree and radian are not units of measurement as are metre, second and kilogram. The measure of a specific angle essentially is the ratio between the length of an arc in the interior of the angle and the circumference of the circle that is the extension of that arc and that has the vertex of the angle as its centre. Hence, the angle is in fact dimensionless and both the degree and the radian are specific ratios, $1/360$ and $1/(2 \cdot \pi)$, respectively.

Imbrie and Purdy suggest transforming theta by

$$\vartheta' = (45° - \vartheta)/(45°) \tag{12.2}$$

for two reasons. The first is that $\vartheta'$ would be a dimensionless parameter, the second is that this parameter would range from +1 through 0 to −1, with the three stated values corresponding to $\vartheta$-values of 0°, 45°, and 90°, respectively. Negative values would reflect dissimilarity and positive values similarity. Although the authors appear to use $\cos(\vartheta)$ (which certainly is dimensionless) instead of $\vartheta$ or $\vartheta'$ in their calculations, we feel bound to comment on their transformation (our 12.2).

Primarily, the transformation is not necessary to make the parameter dimensionless, because $\vartheta$ is already dimensionless, as we argued above. Furthermore, we think that dissimilarity should not be associated with negative values, because it suggests that dissimilarity is a "negative similarity". In our opinion dissimilarity is no similarity, which has to be associated with zero. Expression (12.2) implies that $\vartheta' = 0$ (or $\vartheta = 45°$, or $\cos(\vartheta) = \frac{1}{2} \cdot \sqrt{2}$) reflects neither similarity nor dissimilarity, but we think that $\cos(\vartheta) = \frac{1}{2} \cdot \sqrt{2}$

is not a sound boundary that makes a correct distinction between similarity and dissimilarity.

Dissimilarity is equivalent to no similarity, to be associated with zero or with a positive value close to zero for the similarity coefficient. Increasing similarity should be expressed by an increasing value of the similarity coefficient, while perfect similarity would result in a similarity coefficient value equal to one.

Returning to (12.1), it is easily deduced that $Sm_1$ can only have non-negative values. $Sm_1(i, k)$ is equal to zero if and only if

$$\sum_{j=1}^{N} \hat{p}_{ij} \cdot \hat{p}_{kj} = 0,$$

which is equivalent to

$$\hat{p}_{ij} \cdot \hat{p}_{kj} = 0 \quad \text{for each count } j. \tag{12.3}$$

Expression (12.3) means that there is not a single count (sample) in which both $\hat{p}_{ij}$ and $\hat{p}_{kj}$ are positive, i.e. there is no count in which both taxon i and taxon k are present. We already considered this case to be "completely dissimilar" in the introduction.

From the properties of the inner product it can be deduced that $Sm_1(i, k)$ cannot exceed one, and that $Sm_1(i, k) = 1$ occurs only if there is some number x, such that

$$\hat{p}_{ij} = x \cdot \hat{p}_{kj} \text{ for each count } j, \tag{12.4}$$

which property we would call "perfect similarity". Hence, $Sm_1$ as defined by (12.1) seems to be a good similarity coefficient.

More generally, for any fixed non-negative number a, the cosine of the angle between the vectors $(\hat{p}_{i1}^a, \hat{p}_{i2}^a, \hat{p}_{i3}^a, \ldots, \hat{p}_{iN}^a)$ and $(\hat{p}_{k1}^a, \hat{p}_{k2}^a, \hat{p}_{k3}^a, \ldots, \hat{p}_{kN}^a)$ can be considered as a similarity coefficient $Sm_a(i, k)$. Hence, the coefficients in the vectors need not be proportions themselves (a = 1), but may as well be the powers of the proportions, with some fixed exponent.

Three special cases will be considered. The first one leads to the similarity coefficient

$$Sm_{\frac{1}{2}}(i, k) = \frac{\sum_{j=1}^{N} \sqrt{\hat{p}_{ij} \cdot \hat{p}_{kj}}}{\sqrt{\left(\sum_{j=1}^{N} \hat{p}_{ij}\right) \cdot \left(\sum_{j=1}^{N} \hat{p}_{kj}\right)}} \tag{12.5}$$

by substituting $a = \frac{1}{2}$, which leads to $\hat{p}_{ij}^{\frac{1}{2}} = \sqrt{\hat{p}_{ij}}$ and $\hat{p}_{kj}^{\frac{1}{2}} = \sqrt{\hat{p}_{kj}}$ for the coefficients of the first and second vectors, respectively. Just like $Sm_1$, $Sm_{\frac{1}{2}}$ has the properties of (12.3) and (12.4). The similarity coefficient $Sm_{\frac{1}{2}}$ incorporates small proportions better than $Sm_1$ does. The reason is that if the ratio between $p_1$ and $p_2$ is large, the ratio between $\sqrt{p_1}$ and $\sqrt{p_2}$ is much less ($\sqrt{p_1} / \sqrt{p_2} = \sqrt{p_1/p_2}$).

Before we define $Sm_0(i,k)$ and $Sm_\infty(i,k)$, we list the numbers that provide information about the occurrences of taxon i and of taxon k in the set of counts, and also the numbers that give information about the occurrences of the two taxa relative to each other in the set of counts.

$$\begin{aligned} P_i &= \#\,(\hat{p}_{ij} > 0) \\ P_k &= \#\,(\hat{p}_{kj} > 0) \\ N_{i,k} &= \#\,(\hat{p}_{ij} > 0 \text{ and } \hat{p}_{kj} > 0) \\ P_{i,k} &= \#\,(\hat{p}_{ij} > 0 \text{ and } \hat{p}_{kj} = 0) \\ P_{k,i} &= \#\,(\hat{p}_{kj} > 0 \text{ and } \hat{p}_{ij} = 0) \\ A_{i,k} &= \#\,(\hat{p}_{ij} = 0 \text{ and } \hat{p}_{kj} = 0) \end{aligned} \qquad (12.6)$$

in which "#" means "the number of counts j that have . . .".

Substituting a = 0 into $p^a$, the a-th power of the proportion, leads to $p^0 = 1$ for proportions p greater than zero and to $p^0 = 0$ for p equal to zero, as the coordinates of the two vectors. This leads to the similarity coefficient

$$Sm_0(i,k) = \frac{N_{i,k}}{\sqrt{P_i \cdot P_k}} \qquad (12.7)$$

Obviously complete dissimilarity is equivalent to $Sm_0 = 0$, because then $N_{i,k}$, the number of counts in which both taxa i and k are present, is equal to zero. In fact, $Sm_0$ satisfies (12.3). Perfect similarity has to be equivalent to $Sm_0 = 1$. The index $Sm_0$ being equal to one is not equivalent to (12.4), however, but to

$$\begin{aligned} &N_{i,k} = P_i = P_k \quad \text{which is equivalent to} \\ &P_{i,k} = P_{k,i} = 0, \end{aligned} \qquad (12.8)$$

which implies that complete similarity according to $Sm_0$ means that in each count either both taxa are present or both taxa are absent, which is a weaker condition than (12.4).

Expression (12.7) is a form of the Otsuka coefficient (Cheetham & Hazel, 1969).

Finally, we shall consider the similarity coefficient that results from substituting $a = \infty$ in $Sm_a$. However,

$$\lim_{a \to \infty} Sm_a\,(i, k)$$

for any pair of taxa i and k appears to be unstable (not continuous). It yields unity if

$$\max_j (\hat{p}_{ij} \cdot \hat{p}_{kj}) = \max_j (\hat{p}_{ij}) \cdot \max_j (\hat{p}_{kj}) \tag{12.9}$$

and it yields zero if (12.9) does not hold. The symbol $\max_j(\text{". ."})$ means the maximum value of ". ." over all counts j. It is much more convenient to define

$$Sm_\infty (i, k) = \lim_{a \to \infty} \sqrt[a]{Sm_a(i, k)} = \frac{\max_j (\hat{p}_{ij} \cdot \hat{p}_{kj})}{\max_j (\hat{p}_{ij}) \cdot \max_j (\hat{p}_{kj})} \tag{12.10}$$

Complete dissimilarity according to (12.3) is evidently again equivalent to $Sm_\infty (i, k) = 0$. Perfect similarity according to this similarity coefficient ($Sm_\infty (i, k) = 1$) appears to be equivalent to (12.9), i.e. there is one count (or more) in which the proportions of both taxon i and of taxon k reach their maximum values. Like (12.8), (12.9) is a weaker condition than (12.4), because (12.4) requires that in each count the proportion of taxon i and the proportion of taxon k be related by some constant factor.

In addition to the numbers $N_{i,k}$, $P_{i,k}$, $P_{k,i}$ and $A_{i,k}$ of (12.6) the four similarity coefficients $Sm_0$, $Sm_{\frac{1}{2}}$, $Sm_1$ and $Sm_\infty$ described above are calculated for each pair of taxa in the Fortran computer program SMLRTY. For each taxon this computer program gives not only the number $P_i$ of counts in which taxon i is present, but also the values of

$$\sum_{j=1}^{N} \hat{p}_{ij}, \quad \sum_{j=1}^{N} \hat{p}_{ij}^2 \quad \text{and} \quad \max_j (\hat{p}_{ij}) \tag{12.11}$$

In this section the approach to similarity is based upon the inner product concept of vectors in multidimensional space, and the resulting similarity coefficients are in no way to be considered as test statistics, testing some hypothesis of "similarity". Another problem that might arise when using these coefficients is that some similarity coefficient value of a pair of taxa will be affected by the highly fluctuating proportions of a third "disturbing"

taxon. Since proportions are involved in these similarity coefficients (except for $Sm_0$), the closed sum effect may cause trouble, especially for $Sm_1$ and $Sm_\infty$, these two coefficients being most sensitive to high fluctuations in the proportions from count to count.

## XII.3. Similarity coefficients based on the hypothesis of homogeneity

In this section a different approach to similarity between proportions of taxa will be dealt with. This approach has a somewhat better statistical basis; in addition the closed sum effect is avoided. Experience should prove whether the resulting two similarity coefficients are any better.

In chapter VII the chi square test for homogeneity has been mentioned, which has the chi square statistic (7.9) in common with the chi square test for independence in every contingency table with two rows or with two columns. This statistic gave us the idea of considering the expressions

$$X_u^2 = \sum_{j=1}^{N'} (q_j - \bar{q})^2 \quad \text{"unweighted"} \tag{12.12}$$

and

$$X_w^2 = \sum_{j=1}^{N'} u_j \cdot (q_j - \hat{q})^2 \quad \text{"weighted"} \tag{12.13}$$

for each pair of taxa i and k, in which for each count j:

$$u_j = x_{ij} + x_{kj}, \quad q_j = x_{ij}/u_j \tag{12.14}$$

In words, $u_j$ is the sum of the scores of the taxa i and k in count j, and $q_j$ is the proportion of the score of taxon i in this sum. The unweighted mean value and the weighted mean value of these proportions $q_j$ are

$$\bar{q} = \left( \sum_{j=1}^{N'} q_j \right) / N' \text{ and } \hat{q} = \left( \sum_{j=1}^{N'} x_{ij} \right) \Big/ \left( \sum_{j=1}^{N'} u_j \right), \tag{12.15}$$

respectively. There may be counts that have $u_j = 0$, i.e. both taxon i and taxon k are absent in these counts. Such counts have to be deleted from the above expressions. The remaining number N' of counts in which at least one of the two taxa is present may be considerably less than N, the total number of the counts. If N' is less than N, the index j in the formulae above does not always correspond to the original rank number of the count.

Obviously the closed sum effect is avoided in (12.12) and (12.13), because only the scores of the taxa i and k are taken into account.

The hypothesis of homogeneity states that in each assemblage j, on which count j is based, the proportions of taxa i and k have the same ratio, i.e. for some positive number c

$$p_{ij}/p_{kj} = c \tag{12.16}$$

is valid for each count j. If this hypothesis is true, then it is permissible to approximate

$$E((q_j - \bar{q})^2) \approx \frac{\bar{q} \cdot (1 - \bar{q})}{u_j}; \quad E(u_j \cdot (q_j - \hat{q})^2) \approx \hat{q} \cdot (1 - \hat{q})$$

according to the binomial model. Hence we have

$$E(X_u^2) \approx \bar{q} \cdot (1 - \bar{q}) \cdot \sum_{j=1}^{N'} \frac{1}{u_j}; \quad E(X_w^2) \approx N' \cdot \hat{q} \cdot (1 - \hat{q}) \tag{12.17}$$

The hypothesis of homogeneity (12.16) can be seen as an expression of perfect similarity between the taxa i and k, leading to relatively small values of $X_u^2$ and of $X_w^2$, according to (12.17). Expression (12.16) is in line with (12.4) of the previous section.

Next, the case of complete dissimilarity between two taxa has to be considered. It is stated here (without proof) that, if the numbers $u_1, u_2, u_3, \ldots, u_N$, and the mean proportion $\bar{q}$ remain fixed, that complete dissimilarity (in the sense that both taxa are absent from all counts) is characterized by $X_u^2$ reaching its maximum value. This maximum value is

$$\max(X_u^2) = N' \cdot \bar{q} \cdot (1 - \bar{q}). \tag{12.18}$$

An identical statement holds for the weighted case, replacing $\bar{q}$ by $\hat{q}$, and $X_u^2$ by $X_w^2$. The maximum value then is

$$\max(X_w^2) = \hat{q} \cdot (1 - \hat{q}) \cdot \sum_{j=1}^{N'} u_j. \tag{12.19}$$

In order to get similarity coefficients as defined in the previous sections, we transform $X_u^2$ and $X_w^2$ in the following way:

$$TX_u^2 = 1 - \frac{X_u^2}{\max(X_u^2)}; \quad TX_w^2 = 1 - \frac{X_w^2}{\max(X_w^2)} \tag{12.20}$$

The formula for the similarity coefficient $TX_u^2$ is

$$TX_u^2 = 1 - \frac{\sum_{j=1}^{N'} (q_j - \bar{q})^2}{N' \cdot \bar{q} \cdot (1 - \bar{q})} = \left(1 - \frac{\sum_{j=1}^{N'} q_j^2}{N' \cdot \bar{q}}\right) \Big/ (1 - \bar{q}) \tag{12.21}$$

The formula for the similarity coefficient $TX_w^2$ is

$$TX_w^2 = 1 - \frac{\sum_{j=1}^{N'} u_j \cdot (q_j - \hat{q})^2}{\hat{q} \cdot (1 - \hat{q}) \cdot \sum_{j=1}^{N'} u_j} = \left(1 - \frac{\sum_{j=1}^{N'} u_j \cdot q_j^2}{\hat{q} \cdot \sum_{j=1}^{N'} u_j}\right) \Big/ (1 - \hat{q}) \quad (12.22)$$

These coefficients yield exactly zero in the case of the complete dissimilarity of taxa i and k. In the case of perfect similarity between the two taxa we assume the hypothesis of homogeneity (12.16) to be valid. This hypothesis leads to

$$E(TX_u^2) = 1 - (1/\tilde{u}); \quad E(TX_w^2) = 1 - (1/\bar{u}) \quad (12.23)$$

by application of (12.17), (12.18), (12.19) and (12.20). In these formulae

$$\tilde{u} = \left(\frac{1}{N'} \cdot \sum_{j=1}^{N'} \frac{1}{u_j}\right)^{-1} \text{ and } \bar{u} = \frac{1}{N'} \cdot \sum_{j=1}^{N'} u_j \quad (12.24)$$

are the harmonic mean and the "normal" mean of the scores $u_j$, respectively. It should be noted that the expected values $E(TX_u^2)$ and $E(TX_w^2)$ are only dependent on these mean values of the $u_j$; they are independent of the number N′ of counts in which at least one of the taxa is present. Evidently $E(TX_u^2)$ and $E(TX_w^2)$ are close to one, unless $\tilde{u}$ or (and) $\bar{u}$ is close to one. The last exception means that the scores of both taxa together are very low in the series of counts, thus impeding the recognition of similarity under the assumption that the hypothesis of homogeneity (12.16) is true. In other words, the power of $TX_u^2$ and $TX_w^2$ to recognize similarity is small if the mean value of $u_j = x_{ij} + x_{kj}$ is not much larger than one, and considering only the counts in which $u_j > 0$.

The question of whether the unweighted similarity coefficient $TX_u^2$ or the weighted similarity coefficient $TX_w^2$ is to be preferred is very similar to the question of whether the unweighted or the weighted statistics deserve preference, a question which was posed in earlier chapters. We again decided to apply both coefficients. The weighted form $TX_w^2$ seems preferable because of its close connection to the chi square statistic of (7.9). The unweighted form $TX_u^2$ may be preferred if one doubts the logic of the weighting of the deviations $(q_j - \hat{q})$ (according to the number $u_j = x_{ij} + x_{kj}$) in cases of moderate similarity, i.e. in cases where these deviations are large.

The Fortran computer program SMLRTY contains for each pair of taxa the coefficient $TX_u^2$ of (12.21), together with $\tilde{u}$ of (12.24), the coefficient

$TX_w^2$ of (12.22), together with $\bar{u}$ of (12.24), and also the number $N' = N_{i,k} + P_{i,k} + P_{k,i} = N - A_{i,k}$, according to (12.6).

N.B. The program SMLRTY was used for the data of Hageman (1979), but we found no text corresponding to the output.

# Chapter XIII

# DIVERSITY

## XIII.1. Numerical expressions for microfossil assemblage diversity

When carrying out paleoecological investigations there appear to be many measures of (species) diversity, ranging from the simple number of taxa (S) to fairly intricate formulae. Trying to find quantitative expressions for the diversities of (micro)paleontological assemblages is a laborious task, however, especially if such diversity estimates or diversity indices are to have a statistical foundation. First of all, one must define what one means by "diversity" and decide when a formula deserves the name diversity index. Next, one has to consider whether such an index is an assemblage parameter that can be estimated from a random sample (= count). Like all other statistical estimates the diversity estimate from a count should be accompanied by a standard error that gives an indication of the possible distance between the estimate and the "real" diversity parameter of the assemblage from which the count comes.

In the literature on the subject one notices that it is rare for all the obvious conditions stated above, to be fulfilled at the same time. In many publications the feature diversity is enveloped in rather dogmatic reasoning including a priori laws which state how the numerical proportions of taxa are related to each other, or how the number of taxa will increase if larger random samples are taken from some assemblage. In biology one finds references to the "organisation" and the "hierarchy" in an ecological unit in the explanation of such numerical relations. Even for biological communities one wonders whether the hierarchy between species, which in itself seems to be beyond any doubt, justifies highly restrictive hypotheses about the numerical proportions of the species in the assemblages. Such hierarchy assumptions are completely incorrect for taphocoenoses in paleontology. The species concept of paleontologists is markedly different from the species concept of biologists who are concerned with living fauna and flora elements. In paleontology we have to acknowledge the subjectivity of taxonomic units. The grouping of microfossils into different taxa is to a large extent dependent on the knowledge or experience of the investigator and on his a priori concepts of species ranges (for instance is he a "lumper" or a "splitter"?). Furthermore, part of the microfossil assemblage may be allochthonous. In such a case a diversity index of a mixture of autochthonous

and allochthonous elements can hardly be interpreted. If possible, the allochthonous taxa should be put aside and only the scores of the autochthonous taxa considered.

In this chapter we shall not dwell on the interpretation of high and low values of diversity indices, but we shall restrict ourselves to mathematical descriptions of diversity indices, making positive and negative remarks about their foundation and some comments on practical limitations from the point of view of mathematical statistics.

We accept a diversity index to be an assemblage parameter with a numerical scale in which the values increase if the number of taxa is increasing and/or if the taxa tend towards numerically equal representation.

## XIII.2. The Fisher alpha index

The diversity index that is most commonly used in paleontology was introduced by Fisher in 1943, and developed from his theory describing the apparent abundance of different species in biological samples. The index is denoted by $\alpha$. After a series of arguments, which are rather obscure to the present author, Fisher deduces that there is a group of assemblages with such qualities that each assemblage of that group has a positive constant $\alpha$ which gives for any random sample from that assemblage the expected number of species that will be present with n individuals in the random sample:

$$E(S(n)) = \frac{\alpha}{n} \cdot x^n \tag{13.1}$$

In this formula x is a number between zero and one, which is dependent only on the assemblage parameter $\alpha$ and on N, the size of the sample. So

$$0 < x(\alpha, N) < 1 \tag{13.2}$$

Without examining further the way in which Fisher reached these conclusions we might consider (13.1) and (13.2) to be postulates for any assemblage, and then look for the consequences.

The expected number of species that appear in the random count is:

$$E(S) = \sum_{n=1}^{\infty} E(S(n)) = \sum_{n=1}^{\infty} \frac{\alpha}{n} \cdot x^n = \alpha \cdot \ln\left(\frac{1}{1-x}\right) \tag{13.3}$$

The size of the random count, N, can also be expressed in $\alpha$ and x by using (13.1):

$$N = \sum_{n=1}^{\infty} n \cdot E(S(n)) = \sum_{n=1}^{\infty} \alpha \cdot x^n = \frac{\alpha \cdot x}{1-x} \tag{13.4}$$

From (13.4) the sample parameter $x(\alpha, N)$ can be deduced directly:

$$x(\alpha, N) = \frac{N}{N + \alpha}. \tag{13.5}$$

It follows that, as one makes larger and larger counts, x will approach the value one:

$$\lim_{N \to \infty} x(\alpha, N) = 1 \tag{13.6}$$

Finally, we write the relation between N and the expected number of species E(S), using (13.3) and (13.5), as:

$$\frac{E(S)}{\alpha} = \ln \frac{1}{1 - x} = \ln\left(1/\left(1 - \left(\frac{N}{N + \alpha}\right)\right)\right) = \ln\left(1 + \frac{N}{\alpha}\right) \tag{13.7}$$

This is the basic formula which states the relation between $\alpha$, E(S) and N. In practice, when the size of a count (N) and the number of species in the count (S) are given, the Fisher index is estimated by $\hat{\alpha}$, which can be solved from

$$S/\hat{\alpha} = \ln(1 + N/\hat{\alpha}). \tag{13.8}$$

Williams (1943) presented a graph from which $\hat{\alpha}$ can be read directly.

In the literature from 1943 onwards one often finds the parameter $n_1$, the expected number of species represented by single specimens in the count; this $n_1$ is used to explain Fisher's theory ($n_1 = E(S(1))$). In our opinion this parameter adds to the confusion, because then five parameters are involved: S, N, $\alpha$, x, and the superfluous $n_1$. Actually, expression (13.5) implies that the parameter x can also be eliminated, so the "axioms" (13.1) and (13.2) can be written:

$$E(S(n)) = \frac{\alpha}{n} \cdot \left(\frac{N}{N + \alpha}\right)^n \tag{13.9}$$

in which E(S(n)) is the expected number of species that have n specimens in a count of size N from an assemblage that has $\alpha$ as the value of Fisher's diversity index.

Fisher's theory has some remarkable consequences. Firstly, the expected number of species present in a count of size N with single specimens only, is

$$E(S(1)) = \frac{\alpha \cdot N}{N + \alpha}.$$

This number approaches $\alpha$ when N is made larger and larger. So during the continued counting from an assemblage the number of "singles" is expected to remain constant. This fact, expressed as

$$\lim_{N \to \infty} E(S(1)) = \alpha,$$

is presented by Fisher as a kind of law that assemblages should obey. In our opinion this "law" is quite illogical and for us it is the first reason to doubt the value of the axiom (13.9). In a similar way the axiom will restrict the expected number of species that are represented by two individuals during continued counting from an assemblage, or by three individuals, and so on:

$$\lim_{N \to \infty} E(S(2)) = \frac{\alpha}{2}; \quad \lim_{N \to \infty} E(S(3)) = \frac{\alpha}{3}$$

So another consequence of the "law" is that in large counts the number of species that are represented by two individuals is expected to be about half the number of species that have one individual counted, because $\lim E(S(2)) = \frac{1}{2} \cdot \lim E(S(1))$. Similar relations exist for all $\lim E(S(n))$ with small n values. We fail to see the logic of such relations. These relations are a second reason why we think that the axiom (13.9) is not well founded.

A third reason for doubting the value of (13.9) is based on cases of the presence of a dominant species in an assemblage of very low diversity. To us it seems reasonable to assume that in a count of size N = 1,000 about 900 specimens belong to one and the same species. If the procedure of counting 1,000 specimens from this assemblage of low diversity can be repeated, we can approximate the probability P(x) that the dominant taxon has x individuals in the count, if $x > 500$, as follows:

$$P(x) \approx E(S(x)) = \frac{\alpha}{x} \cdot \left(\frac{1000}{1000 + \alpha}\right)^x \approx \frac{\alpha}{x} \cdot e^{-\left(\frac{\alpha \cdot x}{1000}\right)}.$$

If N is a fixed number (in our example 1,000) this formula leads to strange conclusions.

Strictly speaking, Fisher did not mean the N's to be mutually equal when he mentioned that samples were "of equal size". What he really meant was samples of "equal volume", in which the total number N of individuals per sample is a random variable with a Poisson distribution. If we stick to our criterion of equal sizes of samples, meaning that N is a fixed number, the number x has a binomial distribution. As a consequence, x has a unimodal distribution in both cases because x (and $N - x$) is (are) large. However, the distribution

$$P(x) = \frac{\alpha}{x} \cdot e^{-(\frac{\alpha \cdot x}{1000})} \quad \text{(x greater than 500)}$$

has no mode. Hence, Fisher's theory leads to contradictions in descriptions of assemblages with a dominant species.

It is to be noted that substituting E(N), the expected value of N, instead of N in the formulae (13.4), (13.5), (13.6), (13.7) and (13.9) does not affect our conclusions.

Nevertheless, Fisher's $\alpha$ index has features that seem to make it a good diversity index, if one assumes that postulate (13.9) is true for a collection of assemblages. The expected number of taxa that have one specimen in any count of size N, $E(S(1)) = (\alpha \cdot N)/(N + \alpha)$ will increase if $\alpha$ is increasing, i.e. if counts of equal size are made from a series of assemblages with increasing $\alpha$ this expected number will increase as well.

From (13.7) it can easily be deduced that if from some assemblage a count is made, which is $e \approx 2.7$ times as large as the previous count, one can expect to find about $\alpha$ more species. Again increasing $\alpha$ fits in well with the idea of increasing diversity.

The index $\alpha$ cannot be an assemblage parameter, however, because assemblages exist that do not fit the postulate (13.1). An awkward feature of this postulate is that it refers to the behaviour of samples from assemblages, and not to characteristics of the assemblages proper. A consequence of $\alpha$ not being an assemblage parameter is that the standard error of $\alpha$, presented by Fisher (1943), has no real value, except perhaps for the collection of assemblages satisfying (13.1). However, we have serious doubts about the practical importance of this collection.

## XIII.3. A geometrical diversity index

It was mentioned above that Fisher's diversity index is based on the behaviour of samples from assemblages, and not on the characteristics of the assemblages themselves. As far as the diversity of an assemblage is concerned it can be said that the assemblage is characterized by the numerical proportions of the series of its species. We denote the proportion of a species with frequency rank number i as $p_i$, and we state that

$$0 \leqslant p_i \leqslant 1 \qquad \text{for every rank number i} \qquad (13.10)$$

$$p_i \geqslant p_{i+1} \qquad \text{for every rank number i} \qquad (13.11)$$

$$\sum_{i=1}^{\infty} p_i = 1 \qquad (13.12)$$

The last statement permits the presence of an infinite number of species in the assemblage, but it is also possible – and more logical – that the number of species is finite and equal to M, so

$$p_M \neq 0; \quad p_{M+1} = 0; \quad \sum_{i=1}^{M} p_i = 1 \tag{13.13}$$

A diversity index D that is based on such assemblage characteristics can be written as $D(p_1, p_2, p_3, p_4, \ldots)$.

Finding S species in a count of size N does not tell us much about their proportions in the assemblage. Nevertheless, if one wishes to use S in estimating the diversity of the assemblage, one has to make assumptions about the assemblage proportions (In practice S is often presented as the most simple indication of diversity, which is quite admissible unless the counts are of different sizes).

One such assumption is discussed in this section XIII.3. Let the proportions form a geometrical sequence, i.e. there is a real number r in such a way that

$$0 \leqslant r < 1, \text{ and for every i: } p_i = (1 - r) \cdot r^{(i-1)} \tag{13.14}$$

The parameter r, the so-called common ratio of the geometrical sequence, can be regarded as a diversity index. If r is zero, then $p_1 = 1$ and all other proportions are zero, so in fact the assemblage consists of only a single species, which obviously is the minimum possible diversity. If r approaches one, the most frequent species has $p_1 = 1 - r$, which approaches zero, and the proportions of all pairs of adjoining species in the series tend to be equal. We obviously approach maximum diversity of the assemblage.

How do we estimate r from S and N? If the diversity index r is known, it is possible to calculate the expected number of species in a count of size N; we obtain an expression equivalent to (13.7) that was given for the Fisher $\alpha$ index. The expression for our geometrical diversity index r is more complex, however, and it does not easily lead to an estimate of r when S and N are know numbers.

A simpler way of estimating is to assume that all S species, except the last one, cover the entire count, except for one single specimen, so

$$\sum_{i=1}^{S-1} p_i = 1 - \frac{1}{N} \text{ or } \sum_{i=S}^{\infty} p_i = \frac{1}{N} \quad (S \neq 1).$$

Substituting (13.14) in the last equality gives:

$$\sum_{i=S}^{\infty} (1 - \hat{r}) \cdot \hat{r}^{(i-1)} = \hat{r}^{(S-1)} = \frac{1}{N} \quad (S \neq 1).$$

So we get the following estimate of the common ratio r:

$$\begin{aligned} &\text{if } S = 1, \text{ then } \hat{r} = 0 \\ &\text{if } S \geqslant 2, \text{ then } \hat{r} = N^{(-1/(S-1))} = 1/\left(\sqrt[S-1]{N}\right) \end{aligned} \tag{13.15}$$

This diversity estimate is a rough tool that has an advantage over Fisher's $\alpha$ index in that it is easy to calculate with the help of any small microcomputer. The estimate $\hat{r}$ of the common ratio is a poor measure, however, even if the species proportions in the assemblage were to form a fairly good geometrical sequence, and especially if the count contains very few species (S).

The geometrical sequence hypothesis (13.14) cannot be considered a valid postulate for any assemblage. Hence, the estimate $\hat{r}$ is not a statistical estimate of an assemblage parameter which is determinable in any assemblage. As a consequence the standard error of $\hat{r}$ cannot be given.

All other kinds of diversity indices that are based on S and N only have the disadvantage that they are in fact based on some assumption about the assemblage considered. Such an assumption may be a tolerable approximation of the real situation, but the quality of the approximation is always unknown in individual cases. The parameter to be evaluated by the diversity estimate may be undefined, so the diversity estimate will lack any statistical foundation.

In the following sections we shall consider the only remaining alternative for arriving at estimates of diversity: we are going to use the proportions of the species in a count, instead of only the number of species.

## XIII.4. Diversity indices based on proportions of species

In the previous section we already mentioned that diversity indices D based on assemblage characteristics (without any further assumption) can be written as a function of the proportions of the taxa in the assemblage:

$$D(p_1, p_2, p_3, \ldots) \tag{13.16}$$

Each proportion $p_i$ of species i is estimated from a count from the assemblage as

$$\hat{p}_i = x_i/N \tag{13.17}$$

in which $x_i$ is the score of species i in the count of size N. The assemblage parameter D is basically estimated from the count by

$$\hat{D} = D(\hat{p}_1, \hat{p}_2, \hat{p}_3, \ldots) \tag{13.18}$$

Our first aim is now to find out what the functions D in (13.16) must look like if they are to be called diversity indices. After this a statistical analysis is necessary with respect to the estimates $\hat{D}$ in (13.18). The standard errors and possible biases of these estimates have to be established. These errors and biases may give an indication of the kind of situations in which these indices can be used. These topics will be dealt with in the sections 5 to 9.

## XIII.5. A collection of diversity indices

Expression (13.16) is given more structure by choosing

$$D = F\left(\sum_{i=1}^{\infty} \top(p_i)\right), \tag{13.19}$$

which implies that the diversity value is not dependent on the way in which the taxa have been ranked, because the summation $\sum_{i=1}^{\infty}$ does not require a specific order of the $\top(p_i)$.

It should be noted that the infinite series has to be absolutely convergent for the order of the $\top(p_i)$ to be irrelevant in the calculation of the sum of the series. In the following we will require, however, that $\top(p)$ must be greater than or equal to zero for $0 \leqslant p \leqslant 1$. Then the sum of the series will become well defined. As a consequence we drop the assumption (13.11) that $p_1 \geqslant p_2 \geqslant p_3 \ldots$, but we keep (13.13), i.e. there may be a number M so that $p_M \neq 0$ and $p_i = 0$ for any i greater than M. Furthermore, we assume that the functions $\top$ and F are continuous. Absent species obviously should not play a part, so

$$\top(0) = 0 \tag{13.20}$$

The function F is merely a transformation of the sum of the $\top(p_i)$. We shall discuss such transformations later, but now we state F to be the identity function, so

$$D = \sum_{i=1}^{\infty} \top(p_i) \tag{13.21}$$

If the assemblage consists of only one single species the diversity is regarded to be the lowest possible and D to have its lowest value, which is defined here as zero. As a consequence

$$\top(1) = 0 \tag{13.22}$$

In all other situations D must have a positive value.

We wish D to increase not only if the number of species is increasing, but also if a fixed number of species tends to have more equal proportions. Our first wish may come into conflict with the second, however.

If D has to increase if two species i and j tend towards more equal proportions, then it can be deduced from (13.22) that

$$\top((1-x)\cdot p_i + x\cdot p_j) + \top((1-x)\cdot p_j + x\cdot p_i) \geqslant \top(p_i) + \top(p_j)$$

must hold for any $0 \leqslant x \leqslant 1$. Then D can be proved to increase also for more than two species if these species tend towards mutually equal proportions. It is sufficient (but not necessary) that

$$\top \text{is a concave function in the interval } 0 \leqslant p \leqslant 1 \tag{13.23}$$

which statement (fig. 28) is equivalent to (if the second derivative of $\top$ is continuous)

$$\frac{d^2\top(p)}{dp^2} \leqslant 0 \text{ in the interval } [0, 1] \tag{13.24}$$

From (13.20), (13.22) and (13.23) it follows that $\top(p) \geqslant 0$ for all $0 \leqslant p \leqslant 1$. Hence, D in (13.21) is non-negative.

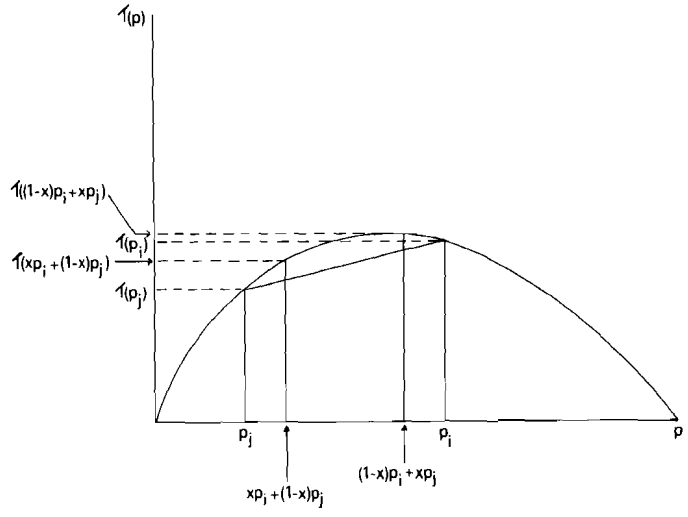



Fig. 28 Graph of a hypothetical concave function $\top$.

If there are M species with mutually equal proportions, so that $p_1 = p_2 = p_3 = \ldots = p_M = 1/M$, then $D = M \cdot \top(1/M)$. Because of the concavity of $\top$:

$$(M+1) \cdot \top\left(\frac{1}{M+1}\right) \geqslant (M+1) \cdot \left[\frac{M}{M+1} \cdot \top\left(\frac{1}{M}\right) + \frac{1}{M+1} \cdot \top(0)\right] = M \cdot \top\left(\frac{1}{M}\right),$$

from which it follows that the diversity D will increase if the number of species that have equal proportions (M) increases, just as we wished it to do.

On the other hand, if the proportions of one of the M species approaches one, so that the proportions of the other $(M - 1)$ species get infinitesimally small, then D of (13.21) will tend to $\top(1) + \top(0) + \top(0) + \ldots + \top(0) = 0$, because function $\top$ is continuous.

Summarizing, $\top$ must be a continuous and concave function, $\top(p) > 0$ for $0 < p < 1$, $\top(0) = \top(1) = 0$, for $D = \sum_{i=1}^{\infty} \top(p_i)$ to have the properties that we think are suitable to make it an acceptable diversity index. This is: D is zero if and only if the assemblage consists of only a single species; D approaches zero if one species dominates the assemblage more and more strongly even if the number of "vanishing" species is large; if all M species have equal proportions in the assemblage, D will increase with increasing M.

## XIII.6. A series of polynomial indices

A series of functions that has the properties established in the preceding section is denoted by $(\top_k^*)$, $k = 1, 2, 3, 4 \ldots$, and is given by the formula

$$\top_k^*(p) = p \cdot \sum_{j=1}^{k} \frac{(1-p)^j}{j} \tag{13.25}$$

So $\top_k^*(p)$ is a polynomial in p of degree $(k + 1)$. It is clear that $\top_{(k+1)}^*(p) \geqslant \top_k^*(p)$. For any value of p, $\top_k^*(p)$ does not increase beyond all bounds, however, because

$$\top_\infty^*(p) = \lim_{k \to \infty} \top_k^*(p) = p \cdot \sum_{j=1}^{\infty} \frac{(1-p)^j}{j} = -p \cdot \ln(p) \tag{13.26}$$

These polynomial functions $\top^*$ lead to a series of indices $D^*$ according to (13.21). Some of these diversity indices are given below:

$$D_1^* = \sum_{i=1}^{\infty} \top_1^*(p_i) = \sum_{i=1}^{\infty} (-p_i^2 + p_i) = \sum_{i=1}^{\infty} p_i (1 - p_i)$$

$$D_2^* = \sum_{i=1}^{\infty} \top_2^*(p_i) = \sum_{i=1}^{\infty} (\tfrac{1}{2} p_i^3 - 2p_i^2 + \tfrac{3}{2} p_i) = \sum_{i=1}^{\infty} \tfrac{1}{2} p_i (1 - p_i)(3 - p_i)$$

$$D_3^* = \sum_{i=1}^{\infty} \top_3^*(p_i) = \sum_{i=1}^{\infty} (-\tfrac{1}{3} p_i^4 + \tfrac{3}{2} p_i^3 - 3p_i^2 + \tfrac{11}{6} p_i) =$$

$$= \sum_{i=1}^{\infty} \tfrac{1}{6} p_i (1 - p_i)(2p_i^2 - 7p_i + 11)$$

and

$$D_\infty^* = \sum_{i=1}^{\infty} \top_\infty^*(p_i) = \sum_{i=1}^{\infty} - p_i \cdot (\ln(p_i))$$

The graphs of these $\top^*$ functions are shown in figure 29.

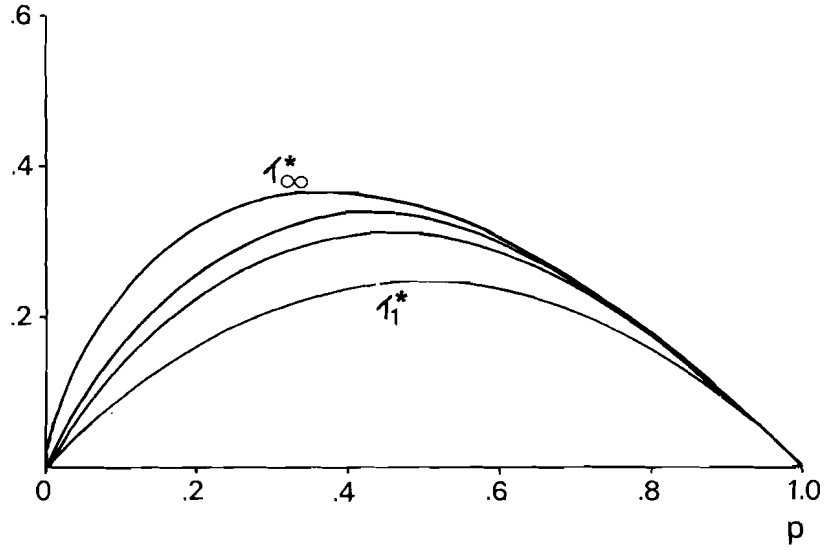



Fig. 29 Graphs of the functions $\top_1^*$, $\top_2^*$, $\top_3^*$ and $\top_\infty^*$.

Our diversity index $D_\infty^*$ is the Shannon-Weaver function $H = -\sum_{i=1}^{M} p_i \cdot \ln(p_i)$, an index often used in ecology. It is also called the "information function", a name which indicates that the index has its origin in information theory.

In that theory H is considered to be the "mean amount of information" or the "average uncertainty" of the M categories (species in our case). The term $-\ln(p_i)$ is the information or the uncertainty from an observation element of category i. It should be noted that the number of categories, M, is finite. Why it should be finite will be explained in section 8.

In information theory several properties of the function H are shown (Lebart & Fénelon, 1973, and Aczél, Forte & Ng, 1974). The most typical property of H, its so-called additivity in the case that the categories are mutually independent events, is not important in our investigation. In our opinion it causes H to lose its superiority over other measures of diversity in our field of interest.

The index $D_1^* = \sum_{i=1}^{\infty} p_i \cdot (1 - p_i) = 1 - \sum_{i=1}^{\infty} (p_i^2)$, which is equivalent to the Yule-Simpson index, can be given a much clearer meaning. If we suppose that a microfossil of type i has been recorded during the counting procedure, then the chance that the next specimen will belong to a different type is given by $(1 - p_i)$. Hence, the probability that during the counting procedure of a microfossil assemblage any following specimen will belong to another type than the specimen counted last, is the mean value of all $(1 - p_i)$, so $\sum_{i=1}^{\infty} p_i \cdot (1 - p_i)$.

The index $D_1^*$ is the only index with such a logical foundation. One might suggest defining diversity as the probability formulated above, thereby regarding $D_1^*$ as the index of diversity. Although this definition can very well be defended, it is in our opinion rather restrictive. Murray (1973) states that the disadvantage of the Yule-Simpson index is that it is controlled mainly by the abundant species. It can be stated that the contribution to the sum $\Sigma\, p_i \cdot (1 - p_i)$ by a series of species with proportions $p_1, p_2, \ldots, p_m$ such that $\sum_{i=1}^{m} p_i$ is equal to a small quantity $\epsilon$, is approximately $\epsilon$, whatever the number m of these species is. Hence, the number of rare species is practically ignored by the index $D_1^*$.

Similar statements can be made for any of the indices $D_k^*$ (except for $D_\infty^*$). From figure 29 it can be read that for an index $D_k^*$ the contribution of a frequent taxon (more than 20%, so $p \geqslant 0.20$) does not exceed the value $2p \cdot (1 - p)$. A series of m rare species with

$$\sum_{i=1}^{m} p_i = \epsilon \text{ (small) has: } \sum_{i=1}^{m} \top_k^*(p_i) \leqslant \frac{\top_k^*(0)}{dp} \cdot \sum_{i=1}^{m} p_i =$$

$$= (1 + \tfrac{1}{2} + \tfrac{1}{3} + \ldots + \tfrac{1}{k}) \cdot \epsilon \text{ whatever the size of m.}$$

The series $S(k) = (1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{k})$ increases (slowly) beyond any limit if k is increasing. For instance $S(3) = 1.83$, $S(10) = 2.93$, $S(20) = 3.60$, $S(30) = 3.99$. So another disadvantage of this series of indices is that k must be large in order to give the group of rare species a substantial weight in the index. In this respect $D^*_\infty$ would be best.

However, the index $D^*_\infty = -\sum_{i=1}^{\infty} p_i \cdot \ln(p_i)$, the Shannon-Weaver function, appears to contain an inconsistency owing to the fact that we allow the number of species to be infinite. It can be proved that for any chosen value of $\epsilon$ (small), there is a series of proportions $p_m, p_{m+1}, p_{m+2}, \ldots$, such that

$$\sum_{i=m}^{\infty} p_i = \epsilon \quad \text{and} \quad -\sum_{i=m}^{\infty} p_i \cdot \ln(p_i) = +\infty .$$

Since the proportions of rare species in an assemblage are never known exactly, this means that in practice $D^*_\infty$ is an inconsistent assemblage parameter, unless the number of species has been delimited in advance. If one places the additional rare species in one group (category), the number of groups M thus being specified in advance, then the formula $-\sum_{i=1}^{M} P_i \cdot \ln(P_i)$ ($P_i$ is the assemblage proportion of group i) becomes a well-defined assemblage parameter that still gives considerable weight to the rare groups, as can be deduced from figure 29.

## XIII.7. The series of root diversity indices

In this section we discuss another series of diversity indices which has better properties than the series $(D^*_k)$ given in the previous section, and which we think will yield the best measures of diversity. These diversity indices we call here the root diversity indices. They are defined by

$$D_k = k \cdot (1 - \sum_{i=1}^{\infty} p_i \cdot \sqrt[k]{p_i}) \quad \text{for } k = 1, 2, 3, \ldots \tag{13.27}$$

Each $D_k$ can be written as $\sum_{i=1}^{\infty} \top_k (p_i)$, in which

$$\top_k (p) = k \cdot (p - p \cdot \sqrt[k]{p}) = k \cdot p \cdot (1 - \sqrt[k]{p}) \tag{13.28}$$

Each $\top_k$ fulfils the conditions (13.20), (13.22), and (13.24), so each $D_k$ can be considered as a real diversity index. From (13.27) we can directly deduce that for any choice of assemblage proportions

$$D_k (p_1, p_2, p_3, \ldots) < k \tag{13.29}$$

Again $D_\infty$ can be defined by writing $\top_\infty = \lim_{k\to\infty} \top_k$:

$$\top_\infty(p) = \lim_{k\to\infty} \top_k(p) = p \cdot \lim_{k\to\infty} (k \cdot (1 - \sqrt[k]{p})) = -p \cdot \ln(p) \qquad (13.30)$$

We write the six diversity indices to be used in practice in the following table:

$$D_1 = 1 - \sum_{i=1}^{\infty} p_i^2 \qquad \top_1(p) = p - p^2$$

$$D_2 = 2\,(1 - \sum_{i=1}^{\infty} p_i \cdot \sqrt[2]{p_i}) \qquad \top_2(p) = 2p - 2p\sqrt[2]{p}$$

$$D_3 = 3\,(1 - \sum_{i=1}^{\infty} p_i \cdot \sqrt[3]{p_i}) \qquad \top_3(p) = 3p - 3p\sqrt[3]{p}$$

$$D_4 = 4\,(1 - \sum_{i=1}^{\infty} p_i \cdot \sqrt[4]{p_i}) \qquad \top_4(p) = 4p - 4p\sqrt[4]{p}$$

$$D_5 = 5\,(1 - \sum_{i=1}^{\infty} p_i \cdot \sqrt[5]{p_i}) \qquad \top_5(p) = 5p - 5p\sqrt[5]{p}$$

$$D_\infty = \sum_{i=1}^{\infty} p_i \cdot \ln(1/p_i) = -\sum_{i=1}^{\infty} p_i \cdot \ln(p_i) \qquad \top_\infty(p) = -p \cdot \ln(p)$$

We already encountered two of these indices in the previous section. $D_1$ is equivalent to the Yule-Simpson index, $D_\infty$ is the Shannon-Weaver function. From figure 30 it is clear that for any series of assemblage proportions $p_1$, $p_2$, $p_3$, ...

$$D_1 \leqslant D_2 \leqslant D_3 \leqslant D_4 \leqslant D_5 \leqslant \ldots \text{ and } \lim_{k\to\infty} D_k = D_\infty \qquad (13.31)$$

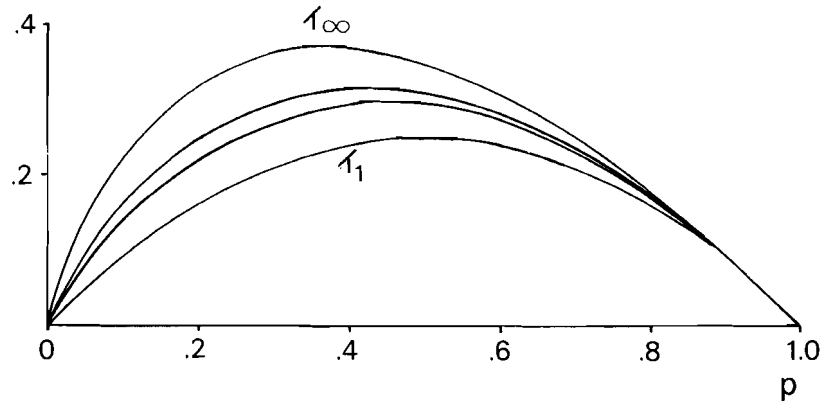



Fig. 30 Graphs of the functions $\top_1$, $\top_2$, $\top_3$ and $\top_\infty$.

The proof will not be given here. (13.31) is the reason why the factor k in each $D_k$ formula is maintained in the following text.

The root diversity indices can be equally presented as

$$(D_k/k) = 1 - \sum_{i=1}^{\infty} p_i \cdot \sqrt[k]{p_i} \quad (< 1)$$

From figure 30 it can be seen that, when k is increasing, rare species acquire relatively more weight in the calculation of $D_k$ compared to the frequent species. The indices $D_2$, $D_3$, $D_4$ and $D_5$ are thought to be less controlled by the abundant species than $D_1$ and the Yule-Simpson index and still to continue to be of consistent nature, which $D_\infty$ (the Shannon-Weaver function) is not, as was explained in the previous section.

Only a few lines will be added concerning F transformations of the diversity indices, such as presented in formula (13.19). Applying to $D_k$ the transformation

$$F_k(x) = (1 - \frac{x}{k})^{-k} \quad (0 \leqslant x < k)$$

and to $D_\infty$ the transformation

$$F_\infty(x) = \lim_{k \to \infty} F_k(x) = e^x \quad (x \geqslant 0)$$

this results in the diversity indices

$$F_k(D_k) = \left( \sum_{i=1}^{\infty} p_i \cdot \sqrt[k]{p_i} \right)^{-k} \quad k = 1, 2, 3, \ldots$$

and

$$F_\infty(D_\infty) = e^{-\left(\sum_{i=1}^{\infty} p_i \cdot \ln(p_i)\right)}$$

These indices have the property that minimum diversity (only one species in the assemblage) acquires the value one instead of zero. If there are M species in the assemblage with mutually equal proportions ($p_1 = p_2 = p_3 = \ldots = p_M = 1/M$) then $F(D) = M$. In contrast to $D_k$ the function $F_k(D_k)$ thus may acquire values beyond any limit. $F_\infty(D_\infty)$ is called the number of equally common taxa: "that would yield an uncertainty equal to that observed" (Beerbower & Jordan, 1969). A similar expression would be appropriate for any $F_k(D_k)$.

The index

$$F_1(D_1) = 1/(\sum_{i=1}^{\infty} p_i^2)$$

is the Yule-Simpson index, or stated more precisely it is the assemblage parameter that is estimated by the Yule-Simpson index (see section 8).

These F transformations can be regarded as formulations of an alternative definition of diversity. They do not give any more information than our indices and they do not provide any further clarification either. This is why we shall stick to our own way of describing diversity, which has been formulated in (13.20), (13.21), (13.22) and (13.23).

We end this section with the warning that the series of diversity indices $D_k$ does not allow us to speak of diversity in an absolute sense either. We give an example of two hypothetical assemblages:

| | $D_1 = 1 - \Sigma\, p_i^2$ | $D_2 = 2(1 - \Sigma\, p_i \cdot \sqrt[2]{p_i})$ |
|---|---|---|
| $p_1 = 0.50, p_2 = 0.50$ | $D_1 = 0.5000$ | $D_2 = 0.5858$ |
| $p_1 = p_2 = 0.15, p_3 = 0.70$ | $D_1 = 0.4650$ | $D_2 = 0.5963$ |

According to $D_1$, the two-species assemblage is more diverse than the three-species assemblage, but this is not the case according to $D_2$. One might use $D_1$ as an absolute measure, however, because of its logical foundation. Hence, in stating the diversity one always has to mention the diversity index used.

### XIII.8. Estimating the root diversity indices

In a first attempt to estimate $D_k$ of (13.27) from a count from some assemblage (the count contains m species and N individuals) one substitutes for $p_i$ the estimate $\hat{p}_i = x_i/N$, in which $x_i$ is the number in the count belonging to species i. Then we get the following series of estimates $\hat{D}_k$:

$$\hat{D}_1 = 1 - \left(\frac{\sum_{i=1}^{m} x_i^2}{N^2}\right); \quad \hat{D}_\infty = \ln(N) - \frac{\sum_{i=1}^{*} x_i \cdot \ln(x_i)}{N} \tag{13.32}$$

$$\hat{D}_k = k \cdot \left(1 - \left(\frac{\sum_{i=1}^{m} x_i \cdot \sqrt[k]{x_i}}{N \cdot \sqrt[k]{N}}\right)\right) \text{ for } k = 2, 3, 4, \text{ and } 5$$

in which $\overset{*}{\sum_{i=1}}$ means that all $x_i$ (i is one up to and including m) that are equal to zero or to one should not be entered in the sum. In section 6 it was already indicated that it is impossible to estimate $D_\infty$ from a count, if the number of species is not strictly limited. The assemblage parameter $D_\infty$ may very well be much larger than the corresponding estimate $\hat{D}_\infty$ of (13.32). $D_\infty$ might even be infinite. The estimate may only be used as an indication of diversity, but like the geometrical diversity index $\hat{r}$ discussed earlier (see 13.15), it lacks a good statistical foundation.

Also the estimates $\hat{D}_k$ appear to have a bias if compared to the assemblage parameters $D_k$, but this bias can be proved to be limited and it can even be estimated, as will be shown below.

Firstly, because of the concavity of $\top$, and making use of Jensen's inequality (see e.g. Ferguson, 1967), we have

$$E(\top(\hat{p}_i)) \leqslant \top(E(\hat{p}_i)), \quad \text{so consequently}$$

$$E(\hat{D}_k) = E(\sum_{i=1}^{\infty} \top(\hat{p}_i)) = \sum_{i=1}^{\infty} E(\top(\hat{p}_i)) \leqslant \sum_{i=1}^{\infty} \top(E(\hat{p}_i)) = \sum_{i=1}^{\infty} \top(p_i) = D_k$$

which says that the expected value of the estimate $\hat{D}_k$ is less than the assemblage parameter $D_k$.

Being more precise, we approximate

$$\top(\hat{p}_i) = \top(p_i) + \frac{d\top}{dp}(p_i) \cdot (\hat{p}_i - p_i) + \frac{1}{2} \cdot \frac{d^2\top}{dp^2}(p_i) \cdot (\hat{p}_i - p_i)^2 \qquad (13.33)$$

Bearing in mind that $E(\hat{p}_i - p_i) = 0$ and assuming that the binomial probability model is valid ($E((\hat{p}_i - p_i)^2) = \text{var}(\hat{p}_i) = p_i \cdot (1 - p_i)/N$) we arrive at

$$E(\top(\hat{p}_i)) = \top(p_i) + \frac{d^2\top}{dp^2}(p_i) \cdot \frac{p_i \cdot (1 - p_i)}{2 \cdot N}$$

As a consequence

$$E(\hat{D}_k) = \sum_{i=1}^{\infty} E(\top(\hat{p}_i)) = \sum_{i=1}^{\infty} \left(\top(p_i) - \frac{(k+1)}{2 \cdot k \cdot N} \cdot \sqrt[k]{p_i} \cdot (1 - p_i)\right) =$$

$$= D_k - \frac{(k+1)}{2 \cdot k \cdot N} \cdot \sum_{i=1}^{\infty} \sqrt[k]{p_i} \cdot (1 - p_i).$$

So a better estimate of $D_k$ seems to be:

$$\hat{D}'_k = \hat{D}_k + \frac{(k+1)}{2 \cdot k \cdot N} \cdot \sum_{i=1}^{m} \sqrt[k]{\hat{p}_i} \cdot (1 - \hat{p}_i) =$$

$$= k \cdot \left( 1 - \frac{\sum_{i=1}^{m} x_i \cdot \sqrt[k]{x_i}}{N \cdot \sqrt[k]{N}} + \frac{(k+1) \cdot \sum_{i=1}^{m} \sqrt[k]{x_i}}{2 \cdot k^2 \cdot N \cdot \sqrt[k]{N}} - \frac{(k+1) \cdot \sum_{i=1}^{m} x_i \cdot \sqrt[k]{x_i}}{2 \cdot k^2 \cdot N^2 \cdot \sqrt[k]{N}} \right)$$

This last expression is well approximated by

$$\hat{D}'_k \doteq k \cdot \left( 1 - \frac{\sum_{i=1}^{m} (x_i - d) \cdot \sqrt[k]{x_i}}{(N - d) \cdot \sqrt[k]{N}} \right) \tag{13.34}$$

in which $d = (k + 1)/(2 \cdot k^2)$.

There are reasons for doubting the validity of the approximation of $\top(\hat{p}_i)$ given in (13.33) if $p_i$ is very small, say $p_i \ll 1/N$, or $p_i \cdot N \ll 1$. For the extremely rare species we deduce:

$$E(\top(\hat{p}_i)) \approx \top(0) \cdot P(x_i = 0) + \top(1/N) \cdot P(x_i = 1) = 0 +$$
$$+ \left(\frac{k}{N} - \frac{k}{N} \cdot \sqrt[k]{\frac{1}{N}}\right) \cdot (p_i \cdot N) = k \cdot p_i - k \cdot p_i \cdot \sqrt[k]{\frac{1}{N}}$$

For such species we may approximate $\top(p_i) = k \cdot p_i$, and thus

$$E(\top(\hat{p}_i)) = \top(p_i) - k \cdot p_i \cdot \sqrt[k]{\frac{1}{N}}$$

We conclude that, using the formulae of (13.32) as estimators, each rare species ($p_i \ll 1/N$) is underestimated with an amount of

$$k \cdot p_i \cdot \sqrt[k]{\frac{1}{N}}$$

in the indices. Now, all single scores in a count, i.e. all $x_i = 1$ are assumed to come from such rare species. If the number of "singles" is denoted by t, then we should add to the diversity formulae of (13.32)

$$k \cdot \frac{t}{N} \cdot \sqrt[k]{\frac{1}{N}} = \frac{k \cdot t}{N \cdot \sqrt[k]{N}}$$

It is therefore advisable to use both the diversity estimates given in (13.32), which will result in some underestimate of the $D_k$, and the corrected estimates of (13.34), in which every score $x_i = 1$ is not entered in the sum $\sum_{i=1}^{m}$, however; we denoted this sum as $\sum_{i=1}^{*}$. The statements made above concerning rare species are not necessary for $D_1$, i.e. the case k = 1, because then (13.33) is not an approximation, but it holds exactly. Yet the scores $x_i = 1$ play no part in $\hat{D}'_1$, because d = 1 if k = 1. Because of the assumption that every score $x_i = 1$ comes from an extremely rare species, the $\hat{D}'_k$ with $\sum_{i=1}^{*}$ might overestimate the $D_k$. We give a list of these latter estimates below.

$$\hat{D}'_1 = 1 - \frac{\sum_{i=1}^{*} (x_i - 1) \cdot x_i}{(N-1) \cdot N}$$

$$\hat{D}'_2 = 2 \cdot \left(1 - \frac{\sum_{i=1}^{*} (x_i - \frac{3}{8}) \cdot \sqrt[2]{x_i}}{(N - \frac{3}{8}) \cdot \sqrt[2]{N}}\right)$$

$$\hat{D}'_3 = 3 \cdot \left(1 - \frac{\sum_{i=1}^{*} (x_i - \frac{2}{9}) \cdot \sqrt[3]{x_i}}{(N - \frac{2}{9}) \cdot \sqrt[3]{N}}\right) \qquad (13.35)$$

$$\hat{D}'_4 = 4 \cdot \left(1 - \frac{\sum_{i=1}^{*} (x_i - \frac{5}{32}) \cdot \sqrt[4]{x_i}}{(N - \frac{5}{32}) \cdot \sqrt[4]{N}}\right)$$

$$\hat{D}'_5 = 5 \cdot \left(1 - \frac{\sum_{i=1}^{*} (x_i - \frac{3}{25}) \cdot \sqrt[5]{x_i}}{(N - \frac{3}{25}) \cdot \sqrt[5]{N}}\right)$$

It is noted here that the Yule-Simpson diversity estimate is

$$F_1(\hat{D}'_1) = \frac{(N-1) \cdot N}{\sum_{i=1}^{*} (x_i - 1) \cdot x_i}$$

which estimates the assemblage parameter $F_1(D_1) = 1/(\sum_{i=1}^{\infty} p_i^2)$ (see section 7).

## XIII.9. Standard errors of the root diversity estimates

In this section we deduce standard errors of the diversity estimates given in (13.32). Bearing in mind that $\hat{D}_k = \sum_{i=1}^{\infty} \top_k(\hat{p}_i)$, and assuming that the $x_i = N \cdot \hat{p}_i$ have binomial distributions, so that $\mathrm{var}(\hat{p}_i) = p_i \cdot (1 - p_i)/N$ and $\mathrm{cov}(\hat{p}_i, \hat{p}_j) = -p_i \cdot p_j/N$ (the covariance of a pair of random variables X and Y, cov(X, Y), we again define as the expected value $E((X - E(X)) \cdot (Y - E(Y)))$, we elaborate the following approximation:

$$\mathrm{var}(\hat{D}_k) \approx \sum_{i=1}^{\infty} \left(\frac{d\top}{dp}(p_i)\right)^2 \cdot \mathrm{var}(\hat{p}_i) +$$

$$+ \sum_{\substack{i=1 \\ i \neq j}}^{\infty} \sum_{j=1}^{\infty} \left(\frac{d\top}{dp}(p_i)\right) \cdot \left(\frac{d\top}{dp}(p_j)\right) \cdot \mathrm{cov}(\hat{p}_i, \hat{p}_j) =$$

$$= \sum_{i=1}^{\infty} (k - (k+1) \cdot \sqrt[k]{p_i})^2 \cdot \frac{p_i \cdot (1 - p_i)}{N} +$$

$$+ \sum_{\substack{i=1 \\ i \neq j}}^{\infty} \sum_{j=1}^{\infty} (k - (k+1) \cdot \sqrt[k]{p_i}) \cdot (k - (k+1) \cdot \sqrt[k]{p_j}) \cdot \frac{-p_i \cdot p_j}{N} =$$

$$= \frac{(k+1)^2}{N} \left[ \left( \sum_{i=1}^{\infty} p_i \cdot \sqrt[k]{p_i^2} \right) - \left( \sum_{i=1}^{\infty} p_i \cdot \sqrt[k]{p_i} \right)^2 \right] \qquad (13.36)$$

This last expression is estimated in a simple way by substituting $\hat{p}_i = x_i/N$ for $p_i$; this time possible biases that have been dealt with in previous considerations are ignored:

$$\hat{\mathrm{var}}(\hat{D}_k) = \left( \frac{k+1}{N \cdot \sqrt[k]{N}} \right)^2 \cdot \left[ \left( \sum_{i=1}^{m} x_i \cdot \sqrt[k]{x_i^2} \right) - \right.$$

$$\left. - \frac{1}{N} \left( \sum_{i=1}^{m} x_i \cdot \sqrt[k]{x_i} \right)^2 \right] \qquad (13.37)$$

The standard error of $\hat{D}_k$ is the square root of the expression

$$\mathrm{SE}(\hat{D}_k) = \sqrt{\hat{\mathrm{var}}(\hat{D}_k)} \qquad (13.38)$$

The standard error of the $\hat{D}'_k$ of (13.35) is thought to be not very different from this $\mathrm{SE}(\hat{D}_k)$. The discarding of scores $x_i = 1$ plays a part in the establishing of the standard error, which is difficult to evaluate.

Extremely small proportions $p_i \ll 1/N$ tend to enlarge the value of $\mathrm{var}(\hat{D}_k)$

of (13.36). This might be incorporated in $\hat{var}(\hat{D}_k)$ of (13.37) by leaving all $x_i = 1$ scores out of the sums; in this way one risks overestimating $var(\hat{D}_k)$. An alternative formula for the standard error will thus be:

$$SE(\hat{D}_k) = SE(\hat{D}'_k) = \\ = \frac{k+1}{N \cdot \sqrt[k]{N!}} \cdot \sqrt{\left( \sum_{i=1}^{*} x_i \cdot \sqrt[k]{x_i^{2!}} \right) - \frac{1}{N} \cdot \left( \sum_{i=1}^{*} x_i \cdot \sqrt[k]{x_i!} \right)^{2!}} \qquad (13.39)$$

N.B. A number of drawings was found in which the calculated values of several diversity estimates are shown for series of samples from stratigraphic sections in Greece. One of these drawings is reproduced here as figure 31. The calculations had evidently been carried out with the help of a computer program DIVERS.

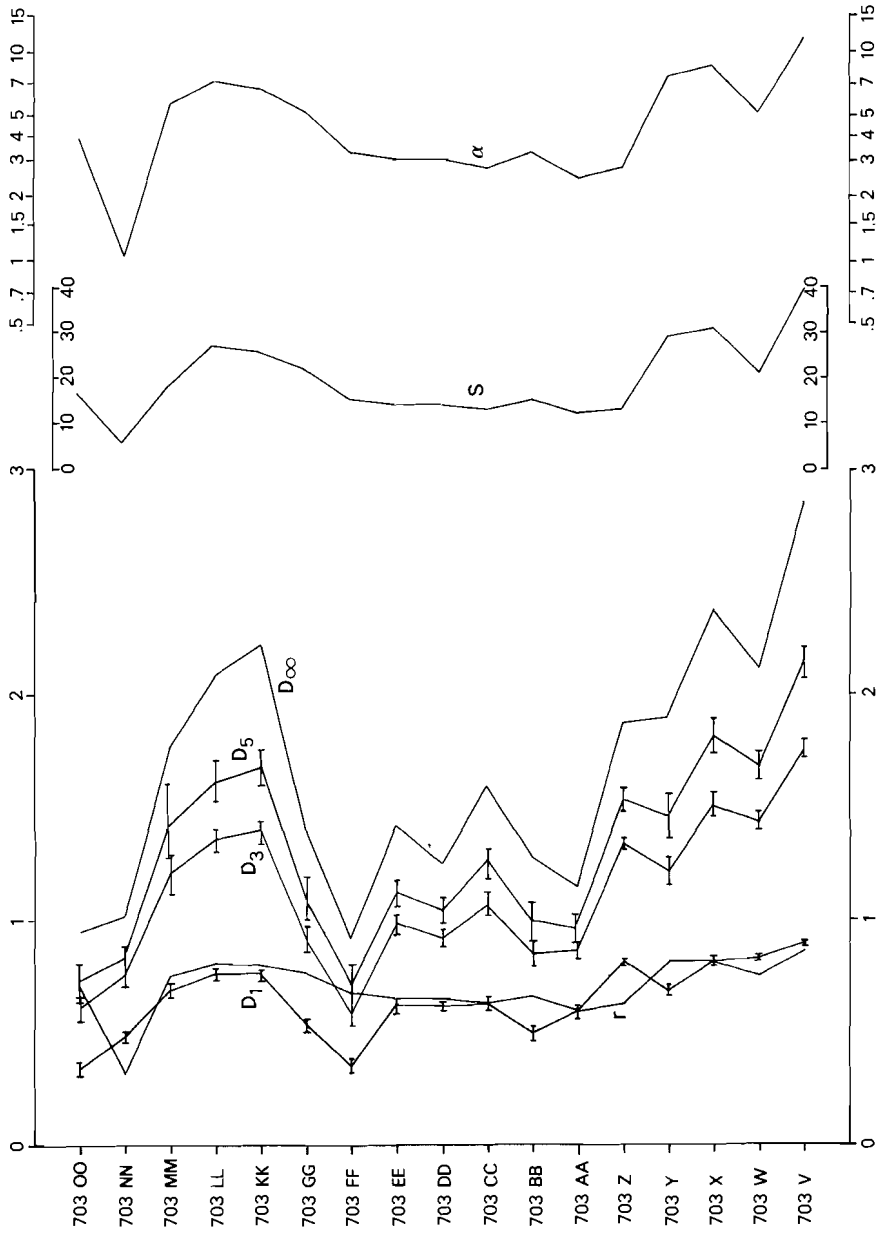



Fig. 31 Values of the diversity estimates S, α, r, $D_1$, $D_3$, $D_5$ and $D_\infty$ for a series of samples from the Plio-Pleistocene Pigadion section, Pyrgos area, Greece (after Hageman, 1979). N = 300 benthonic foraminifera.

# REFERENCES


Abramowitz, M. and I. A. Stegun (ed) (1964). Handbook of mathematical functions with formulas, graphs, and mathematical tables. Nat. Bur. Standards Appl. Math. Series 55, Washington D.C.

Aczél, J., B. Forte and C. T. Ng (1974). On the triangular functional equation and some applications, in particular to the generalized theory of information. Aequationes Math., v. 11, pp. 11–30.

Agterberg, F. P. (1974). Geomathematics. Mathematical background and geoscience applications. Elsevier Sci. Publ. Co., Amsterdam.

Batjes, D. A. J. (1958). Foraminifera of the Oligocene of Belgium. Verh. Kon. Belg. Inst. Natuurwet. 143.

Beerbower, J. R. and D. Jordan (1969). Applications of information theory to paleontological problems: taxonomic diversity. J. Paleont., v. 43, pp. 1184 1198.

Box, G. E. P. and P. Newbold (1971). Some comments on a paper of Coen, Gomme and Kendall. J. Roy. Stat. Soc., ser. A, v. 134, pp. 229–240.

Bremer, M. L., M. Briskin and W. A. Berggren (1980). Quantitative paleobathymetry and paleoecology of the Late Pliocene – Early Pleistocene foraminifera of Le Castella (Calabria, Italy). J. Foram. Res., v. 10, pp. 1–30.

Brouwer, L. E. J. (1907). Over de grondslagen der wiskunde. Ph.D. thesis Amsterdam, Mees & van Suchtelen, Amsterdam-Leipzig.

Chave, K. E. and F. T. Mackenzie (1961). A statistical technique applied to the geochemistry of pelagic muds. J. Geology, v. 69, pp. 572–582.

Chayes, F. (1960). On correlation between variables of constant sum. J. Geoph. Res., v. 65, pp. 4185–4193.

Chayes, F. (1962). Numerical correlation and petrographic variation. J. Geology, v. 70, pp. 440–452.

Chayes, F. (1970). Effect of a single nonzero open covariance on the simple closure test. In: Geostatistics, Plenum Press, New York, pp. 11–22.

Chayes, F. (1971). Ratio correlation. Univ. Chicago Press, Chicago.

Chayes, F. and W. Kruskal (1966). An approximate statistical test for correlations between proportions. J. Geology, v. 74, pp. 692–702.

Cheetham, A. H. and J. E. Hazel (1969). Binary (presence-absence) similarity coefficients. J. Paleont., v. 43, pp. 1130–1136.

Cochran, W. G. (1952). The chi square test of goodness of fit. Annals of Math. Stat., v. 23, pp. 315–345.

Cochran, W. G. (1954). Some methods for strengthening the common chi square tests. Biometrics, v. 10, pp. 417–451.

Connor, J. and J. E. Mosimann (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. J. Amer. Stat. Assoc., v. 64, pp. 194–206.

Darroch, J. N. and D. Ratcliff (1970). Null correlation for proportions – II. Math. Geology, v. 2, pp. 307–312.

Darroch, J. N. and D. Ratcliff (1978). No-association of proportions. Math. Geology, v. 10, pp. 361–368.

Drooger, C. W. and R. Felix (1961). Some variations in foraminiferal assemblages from the Miocene of the North Sea basin. Proc. Kon. Ned. Akad. Wetensch., ser. B, v. 64, pp. 316–324.

Drooger, C. W. and J. P. H. Kaasschieter (1958). Foraminifera of the Orinoco-Trinidad-Paria shelf. Verh. Kon. Ned. Ak. Wetensch., afd. Natuurkunde, ser. 1, v. 22.

Drooger, M. M. (1978). Statistics. In: W. J. Zachariasse et al., Utrecht Micropal. Bull. 17, pp. 19–46.

Drooger, M. M. and C. W. Drooger (1979). Numerical composition data based on Mediterranean Neogene microfossils. Ann. Géol. Pays Hellén., tome hors série, fasc. 1, pp. 371–374.

Drooger, M. M. and J. Hageman (1979). Computer analysis of the foraminiferal frequency data from the Pyrgos sediments. Utrecht Micropal. Bull. 20, pp. 134–147.

Drooger, M. M., D. S. N. Raju and P. H. Doeven (1979). Details of *Planorbulinella* evolution in two sections of the Miocene of Crete. Utrecht Micropal. Bull. 21, pp. 59–127.

Ferguson, T. S. (1967). Mathematical statistics, a decision theoretic approach. Academic Press, New York.

Fisher, R. A., A. S. Corbet and C. B. Williams (1943). The relationship between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol., v. 12, pp. 42–58.

Hageman, J. (1979). Benthic foraminiferal assemblages from Plio-Pleistocene open bay to lagoonal sediments of the Western Peleponnesus (Greece). Utrecht Micropal. Bull. 20.

Haq, B. U. and A. Boersma (ed.) (1978). Introduction to marine micropaleontology. Elsevier Publ. Co., New York.

Imbrie, J. and E. G. Purdy (1962). Classification of modern Bahamian carbonate sediments. Mem. Amer. Assoc. Petr. Geol. 1, pp. 253–272.

Klovan, J. E. and J. Imbrie (1971). An algorithm and FORTRAN-IV program for large scale Q-mode factor analysis and calculation of factor scores. Math. Geology, v. 3, pp. 61–77.

Kork, J. O. (1977). Examination of the Chayes-Kruskal procedure for testing correlations between proportions. Math. Geology, v. 9, pp. 543–562.

Lamperti, J. (1966). Probability, a survey of the mathematical theory. Benjamin Inc., New York.

Lebart, L. and J. P. Fénelon (1973). Statistique et informatique appliquées. Dunod, Paris.

Lukacs, E. (1955). A characterization of the gamma distribution. Ann. Math. Statistics, v. 26, pp. 319–324.

Manson, V. and J. Imbrie (1964). Fortran program for factor and vector analysis of geologic data, using an IBM 7090 or 7094/1401 computer system. Kansas Geol. Surv. Spec. Distribution, 13.

Meulenkamp, J. E., R. R. Schmidt, V. Tsapralis and G. J. van der Zwaan (1978). An empirical approach to paleoenvironmental analysis. 1. Foraminifera, calcareous nannoplankton and ostracodes from the Pliocene of section Prassá, Crete. Proc. Kon. Ned. Ak. Wetensch., ser. B, v. 81, pp. 339–363.

Miesch, A. T. (1969). The constant sum problem in geochemistry. In: Computer applications in the earth sciences; an international symposium. Plenum Press, New York, pp. 161–176.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate beta-distribution, and correlations among proportions. Biometrika, v. 49, pp. 65–82.

Mosimann, J. E. (1963). On the compound negative multinomial distribution and correlations among inversely sampled pollen counts. Biometrika, v. 50, pp. 47–54.

Mosimann, J. E. (1965). Statistical methods for the pollen analyst: Multinomial and negative multinomial techniques. In: B. Kummel and D. M. Raup (ed.), Handbook of paleontological techniques. Freeman and Co., San Francisco.

Muirhead, R. J. and Y. Chikuse (1975). Asymptotic expansions for the joint and marginal distributions of the latent roots of the covariance matrix. The Annals of Statistics, v. 3, pp. 1011–1017.

Murray, J. W. (1973). Distribution and ecology of living benthic foraminiferids. Heinemann, London.

Parker, F. L. (1958). Eastern Mediterranean Foraminifera. Swedish Deep-Sea Exped. Reports, v. 8, pp. 219–283.

Phleger, F. B. and F. L. Parker (1951). Ecology of foraminifera, Northwest Gulf of Mexico. Mem. Geol. Soc. America 46.

Raup, D. M. (1977). Stochastic models in evolutionary paleontology. In: A. Hallam (ed.), Patterns of evolution, as illustrated by the fossil record. Elsevier Publ. Co., Amsterdam, pp. 59–78.

Ryan, W. B. F. (1972). Stratigraphy of Late Quaternary sediments in the Eastern Mediterranean. In: D. J. Stanley (ed.). The Mediterranean Sea: a natural sedimentation laboratory. Dowden, Hutchinson & Ross, Stroudsburg, Penns., pp. 149–169.

Saha, A. K., C. Bhattacharyya and S. Lakshmipathy (1974). Some problems of interpreting the correlations between the modal variables in granitic rocks. Math. Geology, v. 6, pp. 245–258.

Schwarzacher, W. (1975). Sedimentation models and quantitative stratigraphy. Elsevier Publ. Co., Amsterdam.

Tsapralis, V. (1976). Ostracode associations and paleoenvironmental analysis of the Pliocene of section Prassá, Crete, Greece. Proc. Kon. Ned. Ak. Wetensch., ser. B, v. 79, pp. 300–311.

Tsapralis, V. (1981). Contribution to the study of Pleistocene of Zakynthos island, W. Greece (Ostracoda – palaeoenvironment). Ph.D. thesis Patras.

Zachariasse, W. J., W. R. Riedel, A. Sanfilippo, R. R. Schmidt, M. J. Brolsma, H. J. Schrader, R. Gersonde, M. M. Drooger and J. A. Broekman (1978). Micropaleontological counting methods and techniques – an exercise on an eight metres section of the Lower Pliocene of Capo Rossello, Sicily. Utrecht Micropal. Bull. 17.

Zwaan, G. J. van der (1982). Paleoecology of Late Miocene Mediterranean foraminifera. Utrecht Micropal. Bull. 25.

Spec. Publ. 1. A. A. BOSMA – Rodent biostratigraphy of the Eocene-Oligocene transitional strata of the Isle of Wight. 128 p., 7 pl., 38 figs. (1974) *f* 43,–

Spec. Publ. 2. A. VAN DE WEERD – Rodent faunas of the Mio-Pliocene continental sediments of the Teruel – Alfambra region, Spain. 217 p., 16 pl., 30 figs. (1976) *f* 63,–

Spec. Publ. 3. R. DAAMS – The dental pattern of the dormice *Dryomys, Myomimus, Microdyromys* and *Peridyromys.* 115 p., 5 pl., 42 figs. (1981) *f* 41,–

Sales office U.M.B.: Singel 105, 3984 NX Odijk, Netherlands
Postal account: 3028890, T. van Schaik, Odijk
Bank account: 55 89 19 855, Alg. Bank Nederland, T. van Schaik, Odijk

After *prepayment* to the sales office on one of the above accounts, the books will be sent by surface mail without further charges. Orders for these books not directly from the purchaser to the sales office may cause much higher costs to the purchaser.

Bull. 15. Z. REISS, S. LEUTENEGGER, L. HOTTINGER, W. J. J. FERMONT, J. E. MEULENKAMP, E. THOMAS, H. J. HANSEN, B. BUCHARDT, A. R. LARSEN and C. W. DROOGER – Depth-relations of Recent larger foraminifera in the Gulf of Aqaba-Elat. 244 p., 3 pl., 117 figs. (1977) *f* 53,–

Bull. 16. J. W. VERBEEK – Calcareous nannoplankton biostratigraphy of Middle and Upper Cretaceous deposits in Tunisia, Southern Spain and France. 157 p., 12 pl., 22 figs. (1977) *f* 51,–

Bull. 17. W. J. ZACHARIASSE, W. R. RIEDEL, A. SANFILIPPO, R. R. SCHMIDT, M. J. BROLSMA, H. J. SCHRADER, R. GERSONDE, M. M. DROOGER and J. A. BROEKMAN – Micropaleontological counting methods and techniques – an exercise on an eight metres section of the Lower Pliocene of Capo Rossello, Sicily. 265 p., 23 pl., 95 figs. (1978) *f* 59,–

Bull. 18. M. J. BROLSMA – Quantitative foraminiferal analysis and environmental interpretation of the Pliocene and topmost Miocene on the south coast of Sicily. 159 p., 8 pl., 50 figs. (1978) *f* 49,–

Bull. 19. E. J. VAN VESSEM – Study of Lepidocyclinidae from South-East Asia, particularly from Java and Borneo. 163 p., 10 pl., 84 figs. (1978) *f* 53,–

Bull. 20. J. HAGEMAN – Benthic foraminiferal assemblages from the Plio-Pleistocene open bay to lagoonal sediments of the Western Peloponnesus (Greece). 171 p., 10 pl., 28 figs. (1979) *f* 54,–

Bull. 21. C. W. DROOGER, J. E. MEULENKAMP, C. G. LANGEREIS, A. A. H. WONDERS, G. J. VAN DER ZWAAN, M. M. DROOGER, D. S. N. RAJU, P. H. DOEVEN, W. J. ZACHARIASSE, R. R. SCHMIDT and J. D. A. ZIJDERVELD – Problems of detailed biostratigraphic and magnetostratigraphic correlations in the Potamidha and Apostoli sections of the Cretan Miocene. 222 p., 7 pl., 74 figs. (1979) *f* 57,–

Bull. 22. A. J. T. ROMEIN – Evolutionary lineages in Early Paleogene calcareous nannoplankton. 231 p., 10 pl., 50 figs. (1979) *f* 64,–

Bull. 23. E. THOMAS – Details of *Uvigerina* development in the Cretan Mio-Pliocene. 168 p., 5 pl., 65 figs. (1980) *f* 50,–

Bull. 24. A. A. H. WONDERS – Planktonic foraminifera of the Middle and Late Cretaceous of the Western Mediterranean area. 158 p., 10 pl., 44 figs. (1980) *f* 52,–

Bull. 25. G. J. VAN DER ZWAAN – Paleoecology of Late Miocene Mediterranean foraminifera. 202 p., 15 pl., 65 figs. (1982) *f* 57,–

Bull. 26. M. M. DROOGER – Quantitative range chart analyses. 227 p., 3 pl., 32 figs. (1982) *f* 59,–