STRATEGIES IN NUMERICAL ANALYSIS OF BENTHIC MARINE COMMUNITIES

by

Eugenio FRESI, Maria Pia PONTICELLI & N. Carlo LAURO, Stazione Zoologica, Benthic Ecology Laboratory, Ischia Porto & Istituto di Statistica dell'Università, Napoli Italy)

Resumé. *Ce travail a pour but d'illustrer les phases cruciales d'une stratégie de analyse numerique des communautés benthiques marines. On prend en considération sourtout 1) les problèmes d'échantillonage; 2) la représentation des variables; 3) la sélection des critères de similarité-distance; 4) les techniques de traitement des donnés, les méthodes de classification et d'ordination en particulier.*

Introduction

In the last 20 years there has been a considerable increase in the application of mathematical and statistical methods in the field of marine biocenology. This is a positive sign of the development of this discipline from the "stage of verbal statements" to a higher degree of maturity, characterized by the mathematical formulation of hypotheses and results.
The application of statistics feeds back positively also to the theoretical ground as the better exploitation of data, not only allows the formulation of sensible hypotheses but also, as GOUNOT (1961) has said, "la statistique impose de penser clairement". As result of this, it may be claimed that some of the fundamental ideas and concepts of modern biocenology are, at least in part, issues of the methods that have been applied (BOUDOURESQUE, 1970).
In such a wide field as the application of numerical methods in the analysis of communities, choices and limitations are required. This paper will thus concentrate on a restricted number of topics, focusing problems more than their solutions. Neither it is meant as a critical review nor as a guide to the selection of methods suitable for processing bionomical data. It is more simply a share of the experience that has been gained in all the stages of the real work.

Sampling strategy

Almost by definition, a community is beyond total knowledge and complete representation. The observer must content himself to *sample some* of the many populations that might be sampled or to measure *some* of the many properties that might be measured, he feels to be significant in relation to the community overall structure and function.

Two choices must be made at this stage: a) *the selection of variables*, i.e. the taxonomic group or groups (often within a pre-determined dimensional scale) whose identification and counting are feasible; b) *the area or volume* within which the collection is to be performed (the stand). It is important to briefly discuss this second point. It has often been said that the choice of a space unit should be made on the basis of its *homogeneity*. Several interpretations have been given (and hotly debated) but none of them seems to be satisfactory. Stand delimitation should therefore rely more on intuition than on exactitude (PIELOU, 1977): the notion of *biotope* sensu PERES & PICARD (1964) could, for instance, be taken as a n operational definition in benthic community analysis (CHARDY, 1970).
Location of samples within a given stand should be at random. The same is true when a stratified sampling strategy (see, for instance, BOUDOURESQUE, cit.) is adopted. *Minimal area* (or volume) should be used in determining the sampling size for a given taxocene of whatever rank. BOUDOURESQUE (cit.) has stressed the fact that the notion of minimal area, simple though it is at a first glance, is of great theoretical complexity and its definition by rigorous statistics is not yet satisfactory. Several AA (e.g. DHONDT & COPPEJANS, 1977) have also shown that the area-species curve is saturated at very different levels according to the different computing methods. Besides this *specific minimal area* (or qualitative minimal area), one has to contemplate the *structural minimal area* that takes into account the abundance of the single species (e.g. in terms of stabilization of abundances or of their variance). An approach to structural minimal area has been proposed by MARGALEFF (1962) and applied by NIELL (1974) to marine benthic community. It takes into account not only the species richness in a sample, but also the evenness of these species, by utilizing the SHANNON's information measure of diversity. A diversity-area curve is plotted that saturates at the level of the optimum sampling size. Things are even more complicate when one examines multi-taxocene communities, although highest ranking minimal area may be assumed as containing all the lower-ranking ones. An operational approach (CINELLI et Al. 1977) may consist of considering the vegetal taxocene minimal area as it contained all the others within a given dimensional scale. The problem of minimal area, however, has so far received so little attention in marine biocenology that it might be wiser to first spread the use of simple though empirical methods to gain more information rather than to insist on sophisticate, time consuming and seldom applicable ones.

Representation of variables

As soon as a community is sampled with the above criteria, the data may consist of different types of records: most usually one has a list of species and the values of their abundance (or related quantities). A semi-quantitative or qualitative codification is also used. These data are most conveniently arranged in contingency tables of the form IxJ where I are the variables and J the observations, Xij being the score of the Ith species in the Jth sample.
The analysis can hence take many different directions according to the type of pro-

blem. We won't delve into the problems connected with the application of parametric statistics and the transformations of raw data that have to be performed. It can only be stated here that particular caution has to be devoted to this matter, although as CHARDY & Al. (1976) claim, normalization of ecological data, particularly when several zeroes are recorded, is often a vain endeavour.
Our further step will be to look for criteria that allow us to single out from the data mass those forms and structures that may serve as clues to underlying processes. This has to be done bearing in mind the principle of minimizing the information losses that are involved in the procedure. In the type of community analysis we are dealing with here, two steps must be considered: a) the selection of an *appropriate criterion for comparing samples*; b) the selection of an *appropriate technique for sorting comparisons*.

Comparison of samples

The main problem is to find an appropriate criterion of comparing samples. It is important to insist on this point because this choice will strongly influence the results yielded by the structural analysis.
Similarities between samples can be formulated by means of several *indices* that are usually arranged in three broad classes:

- proper similarity indices (e.g. SØRENSEN)
- correlation coefficients (e.g. BRAVAIS-PEARSON)
- distances (e.g. Euclidean distance)

This distinction is more formal than real as many similarity indices are directly connected with distances, as well as the correlation coefficients that fall into the category of angular distances.
The selection among the very many indices that can be found in the litterature (often the same algorythm is reported under different names) is not an easy one. It must be stressed, however, that is fundamentally the nature of the problem as well as the data code that should guide the ecologist in the selection of the most appropriate criteria. It should be considered that the range of options is very different when starting from measurements-, frequency- or logic matrices (e.g. a wide range of distances is available for the first and the last type of tables, what is not the case for the second one). It should also be considered that, in some cases, the use of different distances yields the same ordination. Of the indices that describe distances in presence/absence tables, for instance, those of JACCARD, CZEKANOWSKY, DICE-SØRENSEN, KULCZINSKI and SOKAL & SNEATH give the same ordination. To evaluate the concordance between two indices (this could be sometimes of importance), a function of the form

$$PR(d1,d2) = f(d1,d2)$$

(where f is the proximity between two distances d1,d2) such as when d1=d2, PR(d1, d2) = 0, may be looked for. Pearson's correlation coefficient, Spearman's correlation coefficient are the most used function in this respect.

Sorting comparisons

The distance matrix obtained by computing indices has now to be submitted to further processing in order to sort comparisons. What we want is to look for regularities in the data structure that may have ecological meaning. The assumption is made that, for instance, proximities or dissimilarities in species composition and abundance of two samples do reflect proximities or dissimilarities in their ecological conditions. The generalization we would like di produce may have only a descriptive purpose, but it is also hoped that some inferences may be done on the community structural aspects.
Two basic options are given that can be broadly categorized as *classification* and *ordination*. *Discrimination* may also be listed as complementary to both ordination and classification.

Classification is essentially an indentification process of discrete groups of similar objects within a given system. These groups or *clusters* are so formed, on the basis of some similarity criteria, as they are as compact and as sharply distinct as possibile. To avoid confusion, it must be stressed that allocation of objects to pre-existing groups is a *discrimination* procedure.
There is a great variety of classification methods that are known under the name of *cluster analysis*. They can be categorized as 1) *agglomerative* when clusters are formed amalgamating individual samples; 2) *subdivisive* when the clustering strategy consists of progressively subdividing a set of samples in sub-sets, varying the allocation criterion at avery stage. Agglomerative methods, somehow less efficient than the subdivisive ones, have the advantage of taking into acacount simultaneously all the relevant characteristics of the objects and require a relatively short computing time. This is the reason why, in general, we prefer them. Agglomerative procedure can either be *hyerarchical* or *nucleate*. Hyerarchical methods imply that all the groups belonging to an Ith level are parts of a Ith+1 level group, all the groups being parts of an "universal" cluster of N-1 level. It seems to us that such a strategy, although very popular in biocenological works, deserves attention only when hyerarchical structures can be inferred a priori: this could be the case of taxonomy, as it may reflect the divergent process of evolution that might be represented as hyerarchical. In other instances a hyerarchical representation can badly distort the reality of the system under study. In this respect, it is our opinion that there is no evidence for intrinsic hyerarchies in biological communities (among and, probably, within communities) and, therefore, nucleate methods should be preferred. Nucleate classification attempts to represent the structure of a data set in terms of discrete, non-superimposed clusters. These clusters may be defined as those portions of space where the N points representing the N elements of a system are best clumped, with the criterion of maximizing the homogeneity within clusters and the dissimilarites among clusters. The methods using Beale-Sparks or Diday algorythms (this latter known under the name of *dynamic clouds analysis*) seem to us very promising in synecological applications.

Ordination. Community analysis, as well as the study of any system, involves si-

tuations characterized by multiple observations over a set of variables. A deductive approach, implying the construction of descriptive models does not seem suitable in a field, such as biocenology, where the running laws are almost unknown and where these very laws are often under investigation. To overcome this difficulty, methods have been devised that allow to extract the structure of the system directly from the data. With an effective image, LEBART & FENELON (1971) compare these methods to a radiographic machine that allows the insight of a reality which cannot be otherwise observed: multi-dimensionality of data, as a bar to their penetration, is then the opacity of tissues that prevents the vision of the skeleton.

At the base of the techniques that handle multivariate statistical data - known under several comprehensive denominations (factor analysis, general analysis, inertia analysis) the most popular of which among ecologists being that of *ordination*, there is the assumption that when many variables are in mutual correlation, these correlations may be due to the presence of *underlying factors*. Underlying, or latent factors, may be more or less related to variables: measuring these relationships may lead to the identification of the factors nature. The hypothesis is made that latent factors are fewer than the original variables being linearly related to them. This means that redundant information issued from higly correlated variables is reduced by the introduction of new uncorrelated variables.

Geometrically speaking, the problem consists of projecting say the N sample points scattered in the space in the space defined by a S-species coordinate frame, onto a space of fewer dimensions, in such a way that the arrangement of the points undergoes the least possibile distortion, thus preserving the most important features of the original S-space patterns. New axes must be orthogonal as we wish the new variables to be uncorrelated (1).

The resulting *ordination model* allows an easy visualization of these patterns in 2 or 3 new dimensions (or components or factors) that are explanatory of a given portion of the system total variance. It has to be borne in mind that these factors do not correspond to the classical notion of ecological factors, but rather to a combination of highly correlated environmental characteristics. This is important when an attempt is made to identify them or when one strives to find correlations between the ordination factors and previously carried out environmental measurements. It should also be stressed here that the percentage of variance that has been attached to a given factor may be issue of accidental circumstances: the availability of a significance statistical test is therefore substantial for a critical evaluation of the analysis results.

There are several strategies for factors extractio, the selection of which depends, once again, on the nature of the system under investigation. The most widely used in synecological studies seem to be the Principal Component Analysis (PCA), the Principal Coordinate Analysis (PCoA) and the Factorial Analysis of

---

(1). These are linear ordinations and may be assumed to perform well only when variables are, at least approximately, linear. In the case of biocenology, linearity of data is an exception rather than a rule: transformations should then be made to restore linearity. Otherwise, curvilinear ordination may be attempted, using e.g. *catenation* methods (NOY MEIR, 1974).

Correspondences (FAC) that some AA consider as variants of a more general method the General Analysis of LEBART & FENELON (cit.). Of these, the FAC seems to us the most attractive method in biocenological research for the following reasons:

- it affords a perfect symmetry in contemporaneous representations of observations points (samples) and variable-points (species). This symmetry is not perfect in the case of PCA (hence the double resolution in R mode (interspecies distance matrix) or in Q mode (intersample distance matrix)) and impossible for PCoA. The simultaneous projection in the factorial space of sample- and species-points, whose proximity brings into evidence real affinities of sets of species to a given sample (or among samples), is of great help in the identification of factors;
- for its particular metrics (χ^2weighted distance), the structures it yields remain unaffected by "influences" such as the abundance of species, richness of samples, double absence of species, etc. (CHARDY & Al. 1976). Another advantage consists of the fact that distortions due to variables codification and transformation are avoided (as a matter of fact, FAC is performed on frequencies and therefore the code is univocal). This is very important in biological system analysis when data have a qualitative or, more often, a boolean code. Boolean data are of particular interest when the structure of a multi-taxocene community is studied, as the different taxocene elements are often recorded under different codes. These facts are substantiated by the fairly consistent identity of ordination models we have obtained starting from either abundance or presence-absence tables. Besides the obvious pragmatical interest of this property, there are several important theoretical implications that are out of the scope of this paper.
- FAC seems, in general, more sensitive than other methods in uncovering ecological relationships between samples as well as in defining statistical groups of species.

Some AA (CASSIE, 1969; ALLEN & SKAGEN, 1973; THURLOW, 1975; CHARDY & GLEMAREC, 1977; BARTELL & Al., 1978) have used multivariate analysis also in the study of time series in marine biological communities. In our own research, we extensively use such an application of the method for the study of community variations in space and time and their possible causes. We have found it useful, for instance, in studying the time stability of a community "polarizing factor" such as an environmental gradient. We attach much importance to time series analysis in pointing out possible "invariant nuclear portions" of biocenotical complexes.

In a previous paragraph we mentioned the necessity of a statistical test to evaluate the significance of an ordination. Various methods have been devised (e.g. *eigenvalue-one* method, *screening test* method, etc.) some of which seem unsatisfactory. The most interesting one, to our opinion, is the so-called *simulation* method, proposed and tested by LINN (1968), LEBART & FENELON (cit.), LAURO & MONGELLUZZO (1976), and by ourselves in several papers.
The procedure consists of confronting the eigenvalues extracted by the analysis of the original samples to a sufficient number of new eigenvalues extracted by

the analysis of simulated samples. Simulation is obtained, for instance, by random permutations of the column vector elements in the data matrix. Under the null-hypothesis that the original eigenvalues are not issued from structural conditions, if we perform N simulations, the original eigenvalues have 1/N probabilities of being higher than the new ones. Performing, say 99 simulations, p=A/100, A being the number of new eigenvalues higher than the original ones.

It would be incorrect to conclude this methodological discussion overlooking the stage of development of biocenology as a whole. In fact, in a discipline that still strives to overcome the phase of "natural history", premature insistence on rigor, objectivity and exactitude may lead to a methodology based on illusions or to the sterilization of the research (WHITTAKER, 1962). As BOUDOURESQUE (cit.) has justly stressed, methodology should rely on a "equilibre aussi judicieux que possible entre le rigueur statistique et le rendement scientifique optimum".

LITERATURE

ALLEN, T.H.F. & S. SKAGEN, 1973. *Br. Phycol. J.*, 8, 267-287

BARTELL, S.M., T.H.F. ALLEN & J.F. KOONCE., 1978. *Phycologia*, 17(1), 1-11

BOUDOURESQUE, C.F., 1970. *Thèses Fac. Sci. Marseille*, 5-623

CASSIE,R. M., 1969. *Mem. Ist. Ital. Idrobiol.*, 25, 33-48

CHARDY, P., 1970. *Thèses Fac. Sci. Paris*, 1-77

CHARDY, P., GLEMAREC, M. & A. LAUREC, 1976. *Estuar. Coastal Mar. Sc.*, 4, 179-205

CHARDY, P. & M. GLEMAREC, 1977. in *Biology of Benthic Organisms (Keegan, O'Ceidigh & Boaden Ed.)*, 165-172, Pergamon Press, London & New York.

CINELLI, F., FRESI, E., MAZZELLA, L., PANSINI, M., PRONZATO, P. & A. SVOBODA, 1977. *ibidem*, 173-183

DHONDT, F. & E. COPPEJANS, 1977. *Rapp. Comm. int. Mer Médit.*, 24(4), 141-142

GOUNOT, M., 1961. *Bull.Serv.Carte Phytogéogr.*, série B, 6(1), 7-64

LAURO, N.C. & R. MONGELLUZZO, 1976. *Atti Simp."Ingegneria dei Sistemi", Roma, dec. 1975*. Ass. Elettrotecnica e Elettronica Italiana, 1-18

LEBART, L. & J.P. FENELON, 1971. *Statistique et informatique appliquées*. Dunod, Paris.

LINN, R. L., 1968. *Psychometrika*, 33, 37

MARGALEFF, R., 1962. *Comunidades naturales*. Mayguez, i-469

NIELL, X., 1974. *Bull. Soc. Phycol. Fr.*, 19, 238-254.

NOY MEIR,I.,1974. *Vegetatio*, 29, 89-99

PERES, J.M. & J. PICARD, 1964. *Rec. Trav. St. Mar. Endoume* (31-47), 1-137

PIELOU,E.C.,1977. *Mathematical Ecology*. 1-385. Wiley Interscience, New York

THURLOW, D.L., DAVIS, R.B. & D.R. SASSEVILLE, 1975. *Verh. Internat. Verein.Limnol.*, 19, 1029-1036.

WHITTAKER, R.H., 1962. *BotanicalReview*.28 (1), 1-239.