CABIOS

Vol.8. no.6. 1992
Pages 587–590

A program for computer-aided analyses of ecological field data

Dieter Piepenburg and Uwe Piatkowski[1]

Abstract

A program for IBM-compatible microcomputers is introduced which combines several complementary analyses of species-station-tables generated in ecological field investigations. The scope of the program encompasses table editing functions, routines for community delimitation by cluster analysis and procedures for the analysis of properties related both to stations (e.g. diversities) and species (e.g. abundance statistics and association indices). The essential reasoning behind the application of community studies is presented briefly, as well as the multi-step analytical approach implemented in the program.

Introduction

One main objective in ecological field studies is the delimitation of species assemblages and/or faunistic zones, and the description of their spatial (and/or temporal) distribution and composition. These investigations, widely known as community analyses, are often based on extensive species-station-tables, i.e. data arrays typically in the form of abundance measures of the various species represented in a series of stations or collections. Based on experience in studying marine zooplankton and benthos communities (Piepenburg, 1988; Piatkowski, 1989) a computer program was developed to assist in performing analyses of species-station-tables.

The conceptual basis of species-station-table analyses is the 'community' of co-occurring species. This concept is one of the most important rationales of ecological research (Odum, 1973), but there are various definitions reflecting different theoretical approaches (see Gray, 1981). The definition also applied in the present paper was formulated by Mills (1969); 'Community means a group of organisms occurring in a particular environment, presumably interacting with each other and with the environment, and separable by means of ecological survey from other groups.' This pragmatic approach is based on methodology and avoids an ecological interpretation which may be controversial (see Petersen, 1989). The differences of the various theories refer mainly to the degree of integration of communities and the role of the biological self-regulation of community structures, i.e. the question whether the spatial distribution of species is mainly influenced by biological interactions or rather by abiotic gradients (Gray, 1981).

Following a multi-step strategy for the study of multi-species distribution patterns (see Bölter *et al.*, 1980; Field *et al.*, 1982), which encompasses a suite of multivariate and analytical techniques, quite a number of computations are usually involved (Legendre and Legendre, 1987). For computer-aided analyses a variety of spreadsheet and statistical software is available, each offering a subset of the tools necessary to perform the various tasks, such as table editing, data handling, differentiation of intrinsic data fractions by cluster analyses, and computation of station-specific and species-specific parameters. A computer program, however, which combines the whole suite of these procedures is harder to find. Here, we introduce a comprehensive computer program, integrating the different parts of a multi-species data analyses, in order to conduct this often time-consuming work in a quick and safe way.

Analytical approach

Ecological field investigations usually result in complex sets of biotic (and/or environmental) data from which patterns or relationships are to be extracted. For instance, community studies are based on a faunistic inventory in a well-defined area by collecting or observing organisms at several stations, i.e. at certain locations and/or at certain time intervals. The methodological rationale of the analysis of the resulting species-station-table may be stated as follows: the distribution of the various species represented in a series of stations is not random, but displays a certain pattern, i.e. certain species groups (assemblages) are limited or at least concentrated in their occurrence to certain station groups (faunistic zones). This 'multispecies distribution pattern' is discernible from species-station-tables by multivariate analytical techniques. Field *et al.* (1982) have provided an overall strategy for the analyses of multispecies data sets. Most of its various stages can be performed by the program introduced in the present paper.

Classification techniques are methods to extract patterns from complex data sets, e.g. to delimitate station groups and species assemblages from species-station-tables and to discern their resemblance structure. They are not the end of a community study but rather the starting point for further data analyses and ecological interpretations. The 'intrinsic' structures of communities may be described, for instance, by certain station-specific and species-specific properties (e.g. diversities and

Institut für Polarökologie der Universität Kiel, Olshausenstrasse 40 and [1]Institut für Meereskunde an der Universität Kiel, Düsternbrooker Weg 20, W-2300 Kiel, Germany

faunistic composition with regard to stations; abundances, frequencies and association indices with regard to species).

The station-specific parameters include measures to express quantitatively the species composition of samples, both in terms of (i) the number and/or (ii) the relative abundance of the species occurring in a sample. These two sample properties, species richness and dominance pattern (evenness), are combined in the concept of 'species diversity'. Its theoretical basis, ecological meaning and measurement are among the most controversial topics in ecology (see Hurlbert, 1971).

Distribution and internal structure of the communities are related to a complex set of interacting environmental ('extrinsic') factors. The community–environment relationship may be examined by several approaches, ranging from plotting the spatial distribution of station groups or species assemblages on a map of the study area to an *a-posteriori* statistical analysis of the environmental data, i.e. the comparison of the extrinsic factors corresponding to the groups. The major goal of all these analyses is to determine the environmental factors responsible for the community patterns found.

There are a large number of textbooks and reviews on the various analytical techniques implemented in the program. Synoptic introductions to multivariate analyses are given, for instance, by Backhaus *et al.* (1990), Clifford and Stephenson (1975) and Romesburg (1984). Comprehensive presentations of the rationale of multivariate analyses as well as descriptions of biological and ecological applications were published by Sneath and Sokal (1973), Jongman *et al.* (1987) and Legendre and Legendre (1983).

Implementation

The computer program COMM introduced in the present paper provides ecologists with a suite of tools for the analysis of species-station-tables with a maximum size of 13 000 cells. It features a window environment, a hierarchical menu structure, routines for file management, printing and plotting, and a screen-oriented editor to edit, sort and subsample data sets. The focus of the program is, however, on cluster analytical techniques and the examination of station-specific and species-specific properties.

Cluster analysis

The program offers various procedures to perform the classification of stations or species applying the multi-step strategy given by Field *et al.* (1982):

(i) transformation/standardization of raw data (optional);
(ii) computation of resemblances between objects;
(iii) grouping of objects using hierarchical-agglomerative cluster strategies.

A transformation of the raw data prior to the cluster analysis *sensu strictu* is recommended because of several reasons: (i) samples from ecological field investigations usually comprise few species with large numbers (or biomass) and many species with small numbers; (ii) the statistical distribution of the data values tends to be non-normal; and (iii) sampling efficiency and/or effort may differ between stations (see Clifford and Stephenson, 1975). The program offers several commonly used data transformations and standardizations. The transformations available are log transformation [log $(x + 1)$], root transformation ($x^{0.5}$), and root–root transformation ($x^{0.25}$). Percentage-standardization may be performed by station or by species. *z*-standardizations, either by station or by species, alter the data values in deviates from the mean of the values of each station and each species respectively. Interval or ratio scale data, e.g. abundance or biomass values, may be converted to codes of lesser information content: ranks, dominance classes and logarithmic abundance classes.

The cluster analysis *sensu strictu* encompasses two principal steps following the optional data transformation: (i) computation of resemblances between each possible pair of objects [i.e. between stations (Q-analysis) or between species (R-analysis)]; and (ii) subsequent classification of the objects on the basis of these resemblances.

A large number of different resemblance measures have been proposed from which many have been restricted to certain disciplines. Eight measures commonly applied in marine studies are available in the program. Three of these are qualitative similarity coefficients comparing objects on the basis of binary data (absence – presence): the Jaccard Coefficient (after Jaccard, 1902); the Sörenson Coefficient (after Sörenson, 1948, synonymous with the Dice Coefficient and the Czekanowski Coefficient); and the Simple Matching Coefficient. The other indices are often referred to as quantitative coefficients, i.e. they measure the resemblance between objects not only with regard to presence or absence, but also consider differences measured on ordinal, interval or ratio scale: Percentage-Similarity Coefficient, Canberra-Metric, Bray–Curtis Index (after Bray and Curtis, 1957), Average Euclidean Distance, Pearson Product–Moment Correlation Coefficient, and Spearman Rank Correlation Coefficient. The measures stored in a resemblance matrix may be standardized, printed and saved to an ASCII file. Optionally, the minimum-spanning tree may be calculated. Field *et al.* (1982) recommended the use of the Bray–Curtis coefficient for multi-species data generated in marine ecological studies. Other measures frequently applied in biological investigations are the Jaccard Index and the Canberra-Metric (Legendre and Legendre, 1983).

The classification of objects, i.e. the objective grouping of objects based on their resemblances, can be performed with COMM by applying numerical procedures known as hierarchical-agglomerative cluster strategies. In an iterative process these methods fuse pairs of similar objects, then pairs of these groups, etc., until all objects are fused. The strategies differ in the computation of similarities between any two clusters during the clustering process: Single Linkage (syn. Nearest-Neighbour Method), Complete Linkage (syn. Farthest-Neighbour Method), Unweighted Pair-Group Method using Arithmetric Averages (UPGMA), Weighted Pair-Group Method using Arithmetric Averages (WPGMA), Centroid Linkage (syn. Unweighted Pair-Group Method using Centroids, UPGMC), Median Linkage (syn. Weighted Pair-Group Method using Centroids, WPGMC), and the Flexible Strategy suggested by Lance and Williams (1967). For a graphic representation of the inter-object resemblance structure th eprogram produces a dendrogram. For more information on the various clustering procedures and their effect on classification results see Clifford and Stephenson (1975), Legendre and Legendre (1983) or Romesburg (1984). The strategies most commonly applied in ecological studies are UPGMA and Complete Linkage (Field *et al.*, 1982).

Station properties

For the description of the intrinsic structure of stations or groups of stations, COMM computes ten often-used, station-specific parameters which may be printed and saved to an ASCII file. As a measure of total abundance per station the sum of species scores is calculated for each station. The set of diversity parameters encompasses: the number of species per station; the number of species comprising 90% of the ranked species' scores; the percentage dominance of the commonest species per station; the diversity measures after Margalef (1951) and Shannon and Weaver (1963); and the evenness parameters after Pielou (1963), Heip (1974) and Simpson (1949). These coefficients share the disadvantages of being rather dependent on sample size (Hurlbert, 1971). Therefore, another diversity measure is provided in addition: the expected number of species, calculated with the rarefaction method of Sanders (1968). It equals the theoretical number of species the samples would contain if they were subsampled to a standard size, here: the size of the smallest sample.

An alternative approach for the comparative analysis of diversities is to plot the relative abundances of the species occurring in a sample as a curve. These 'graphical/distributional methods' (Warwick and Clarke, 1991) retain more information about the dominance pattern then the univariate approach of reducing the distribution to a single index. Two methods are implemented in COMM: (i) a plot of the percentage cumulative abundances of the ranked species against the species rank (Lambshead *et al.*, 1983); and (ii) a rarefaction plot, i.e. a plot of the number of species for different subsample sizes using the rarefaction method of Saunders (1968), which allows the comparison of species richness of samples of different size.

Species properties

In order to assist in the identification of important species in terms of frequency, abundance and association degree, several species-specific parameters are calculated for each (selected) species. This parameter set comprises the sum and mean of species scores over all (selected) stations, minimum, median and maximum of scores greater than zero, percentage frequency of species occurrence, and percentage dominance referred to total sum of species scores. In addition, the 'Biological Index' (BI) after McCloskey (1970) is computed. This is a measure of species importance combining aspects of frequency and dominance. It is computed for each species over j (selected) stations by summing scores in the range from 1 to 10 which are attributed to each species according to its abundance rank at each of the j stations concerned.

In case of station selection two parameters of association degree are computed for each species occurring in the station group (Salzwedel *et al.*, 1983). The 'Degree of Association regarding Individuals' (DAI) is calculated as the number of individuals of the species concerned in the station group divided by the total number of individuals from all stations. The 'Degree of Association regarding Stations' (DAS) is equal to the number of stations within the station group at which the species occurs divided by the total number of stations where the species is present. These association coefficients allow the identification of characterizing species of station groups.

The 'Index of Patchiness' after Lloyd (1967) is calculated, in addition, as a parameter of the distribution of individuals of each species over the (selected) stations. This dispersion coefficient 'purports to measure how many times as crowded an individual is, on the average, as it would have to be if the same population had a random distribution' (Lloyd, 1967). Values >1 indicate aggregation at certain stations; values <1 suggest a uniform distribution. The significance of departure from the expected random distribution (index equals 1) may be tested by a two-tailed chi-square statistic.

Systems and methods

The COMM program runs on personal computers under MS-DOS (v. 3.0 or later). It is written using Borland's Turbo BASIC compiler (v. 1.00e). Its user interface is implemented employing the window tools provided by Baloui (1988). The size of the compiled program (INTEL-8088 code) is 251 kbytes, the source text modules encompass ~8000 lines in total. A 808×7 coprocessor and a hard disk are not required but may speed up the computations considerably. The maximum size of the ASCII-coded data files is ~150 kbytes (i.e. 13 000 data cells). Supported output devices are an Epson-compatible lineprinter and a HP-compatible plotter. A mouse is optionally supported. The program can be obtained from D.Piepenburg on request.

Acknowledgements

We thank Michael Breuer and Werner Heitland for their assistance in programming. Thanks also to Manfred Bölter and two unknown referees for reviewing the manuscript.

References

Backhaus,K., Erichson,B., Plinke,W. and Weiber,R. (1990) *Multivariate Analysemethoden.* Springer-Verlag, Berlin.

Baloui,S. (1988) *Profi-Tools TurboBASIC.* Markt-u.-Technik Verlag, München.

Bölter,M., Meyer,M. and Probst,B. (1980) A statistical scheme for structural analysis in marine ecosystems. *Ecol. Modell.*, **9**, 143–151.

Bray,J.R. and Curtis,J.T. (1957) An ordination of the upland forest communities of Southern Wisconsin. *Ecol. Monogr.*, **27**, 325–349.

Clifford,H.T. and Stephenson,W. (1975) *An Introduction to Numerical Classification.* Academic Press, New York.

Field,J.G., Clarke,K.R. and Warwick,R.M. (1982) A practical strategy for analysing multispecies distribution patterns. *Mar. Ecol. Prog. Ser.*, **8**, 37–52.

Gray,J.S. (1981) *The Ecology of Marine Sediments.* Cambridge University Press, Cambridge.

Heip,C. (1974) A new index measuring evenness. *J. Mar. Biol. Assoc. UK,* **54**, 555–557.

Hurlbert,S.H. (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology,* **52**, 577–586.

Jaccard,P. (1902) Lois de distribution florale dans la zone alpine. *Bull. Soc. Vaud. Sci. Natur.*, **38**, 69–130.

Jongman,R.H.G., ter Braak,C.J.F. and van Tongeren,O.F.R. (1987) *Data Analysis in Community and Landscape Ecology.* Pudoc, Wageningen.

Lambshead,P.J.D., Platt,H.H. and Shaw,K.M. (1983) The detection of differences among assemblages of marine benthic species based on an assessment of dominance and diversity. *J. Nat. Hist.*, **17**, 859–874.

Lance,G.N. and Williams,W.T. (1967) A general theory of classificatory sorting strategies. I. Hierarchical systems. *Comput. J.*, **9**, 373–380.

Legendre,L. and Legendre,P. (1983) *Numerical Ecology.* Elsevier, Amsterdam.

Legendre,L. and Legendre,P. (1987) *Developments in Numerical Ecology.* Springer-Verlag, Berlin.

Lloyd,M. (1967) Mean crowding. *J. Anim. Ecol.*, **36**, 1–30.

Margalef,R. (1951) Diversidad de especies in las communidades naturales. *Publ. Inst. Biol. Apl. Barcelona,* **9**, 5–27.

McCloskey,L.R (1970) The dynamics of the community associated with a marine slceratinian coral. *Int. Rev. Ges. Hydrobiol.*, **55**, 13–81.

Mills,E.L. (1969) The community concept in marine zoology, with comments on continua and instability in some marine communities: a review. *J. Fish. Res. Board Can.*, **26**, 1415–1428.

Odum,E.P. (1973) *Fundamentals of Ecology*, 3rd ed. W.B.Saunders, Philadelphia.

Petersen,G.H. (1989) Benthos, an important compartment in northern aquatic ecosystem. In Rey,L. and Alexander,V. (eds), *Proceedings of the Sixth Conference of the Comité Arctique International.* Brill, Leiden, pp. 162–176.

Piatkowski,U. (1989) Macroplankton communities in Antarctic surface waters: spatial changes related to hydrography. *Mar. Ecol. Prog. Ser.*, **55**, 251–259.

Pielou,E.C. (1966) Shannon's formula as a measure of specific diversity: its use and disuse. *Am. Nat.*, **100**, 463–465.

Piepenburg,D. (1988) Zur Zusammensetzung der Bodenfauna in der westlichen Fram-Strasse. *Ber. Polarforsch.*, **52**, 1–118.

Romesburg,C.H. (1984) *Cluster Analysis for Researchers.* Wadsworth, Belmont.

Salzwedel,H., Rachor,E. and Gerdes,D. (1985) Benthic macrofauna communities in the German Bight. *Veröff. Inst. Meeresforsch. Bremerhaven,* **20**, 199–267.

Sanders,H.L. (1968) Marine benthic diversity: a comparative study. *Am. Nat.*, **102**, 243–282.

Shannon,C.E. and Weaver,W. (1963) *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Simpson,E.H. (1949) Measurement of diversity. *Nature,* **163**, 688.

Sneath,P.H. and Sokal,R.R. (1973) *Numerical Taxonomy.* W.H. Freeman, San Francisco.

Sörenson,T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. *Biol. Skr.*, **5**, 1–34.

Warwick,R.M. and Clarke,K.R. (1991) A comparison of some methods for analysing changes in benthic community structure. *J. Mar. Biol. Assoc. UK,* **71**, 225–244.

Received on January 28, 1992; accepted on May 1, 1992