Ecological Computer Programs: The Importance of Being Friendly

JOHN W. McMANUS
University of Rhode Island
and
Marine Science Institute
University of the Philippines
Diliman, Q.C. 1101, Philippines

Tropical fishery resources are usually embedded in diverse tropical communities of fish and invertebrates. The daring ecologist who attempts to make some sense of these systems must wade through thick morasses of data representing hundreds of species and hundreds of ways by which they are distributed by nature or perturbed by humans. Theoretically at least, the paths to take are clearly marked. There are, for instance, at least seven recent books on how to simplify ecological data analysis through techniques such as cluster analysis and ordination. The potential analyst quickly learns that with the right computer programs, one can sort huge tables of data into logical blocks of sites and species, or peer at the data points in multivariate space through an inviting array of ordination windows. Somewhat later, the intrepid explorer usually finds that the programs he or she wants run on the machine he/she has to work with, or that it might take a few months, or years, for some program offered as "available from the author" to be mailed (if ever), or that the program is available for $300 nobody budgeted for, or that he or she must now spend the next six months learning yet another computer language and tediously convert hundreds of pages of data tables from one format to another through programming or retyping. The paths are there alright, but they are overgrown with dense vines and undermined with quicksand.

The Statistical Applications System (SAS) and similar leading program packages do not do clustering and ordination to suit the needs of the average ecologist. Certainly, they have options for clustering and ordination. However, those are usually the older methods, the ones that have come under sharp criticism for assuming linearity, normality, or a myriad of other things with which an ecologist's data are rarely blessed. Just try looking for TWINSPAN or Detrended Canonical Correspondence Analysis in a popular ecological package. Even worse, these methods are not just computational, they are algorithmic. That means that it is very unrealistic to attempt to do them with a general mathematics package such as MATLAB. It is easier to simply reprogram them. That way, one merely has to reinvent the wheel, and spend several months making sure that it will roll correctly.

The potential analyst quickly learns that with the right computer programs, one can sort huge tables of data into logical blocks of sites and species, or peer at the data points in multivariate space through an inviting array of ordination windows.

Packages such as SAS are by no means useless to an ecologist. On the contrary, analytical methods such as ANOVA, MANOVA, and multiple regression are essential in modern ecology and should be utilized far more than they are. Commercial statistical packages are generally oriented towards inferential statistics, analyses which follow an experiment or sampling program wherein a closed set of hypotheses are being tested. Many have argued that ecology, and its estranged offspring fisheries biology, would progress more efficiently if more research was oriented toward inferential statistics. However, this is usually only practical in fairly well-studied systems. There is an excellent book by Green (1979, see p. 7) that is heavily oriented towards inferential statistics, particularly for environmental impact assessment, and SAS is the computer package he recommends most highly.

For those having to break new ground to obtain the answers being sought, some degree of exploratory data analysis or pattern analysis is necessary. In these approaches, the set of hypotheses is not closed, and many inferential methods are inappropriate or even misleading. Because of the multivariate nature of the field, ecologists often must rely on cluster and ordination methods. Both approaches involve a substantial degree of subjectivity, particularly in the final stage of interpreting the results. They are also prone to a broad variety of biases which, although probably not less than those of inferential statistics, are certainly less well known. There is currently a rapidly growing effort to develop ordination and clustering methods which involve less subjectivity, reduced bias, and probability determinations of both. Two areas of considerable promise are "fuzzy mathematics", wherein the false cloak of certainty is removed, and computer intensive validification, wherein assessment of errors is done by

Two areas of considerable promise are "fuzzy mathematics", wherein the false cloak of certainty is removed, and computer intensive validification, wherein assessment of errors is done by systematic trial and error through many repetitions.

systematic trial and error through many repetitions. The ultimate effect of this research will be to dissolve the distinction between inferential and noninferential approaches. Much work has already been done along these lines, but little of it has had any impact on ecological analyses. Two major reasons for this are the availability of the programs, and the user-friendliness of the programs and documentation.

Few scientists in developing countries can take time out to learn computer languages. Fewer still have research funds to devote to maintaining a highly paid programmer merely to broaden the range of analyses available from the immediately gratifying to the ultimately important.

Some packages of programs have in fact been produced with the ecologist in mind. Two of these are the Cornell Ecology Package and PATN. Both of these packages are expensive enough (a few hundred dollars) to require specific prior budgeting. Both packages are now available for microcomputers, after having evolved for more than a decade on mainframe computers. Neither is particularly user-friendly, when compared with such software as word processors, spreadsheets, and graphics programs. However, each is learnable and useable by the statistically inclined ecologist. Both packages are worth getting, in that their programs have no counterparts in the more popular statistical packages, and there is little overlap between the two. However, don't expect to see last year's breakthroughs in this year's package. Each package is a few years behind the state of the art. Nonetheless, I applaud the maintainers of these and similar packages for persevering in the face of little-known hardships to keep ecology a productive science.

Why Don't Ecologists Make Friendly Programs?

There are some general rules of programming which are helpful to know:

1. The time it takes to build a program rises exponentially with the number of lines in the program. That is, a programmer may take one day to write and debug a one hundred line program, and three months to do a program of a thousand lines. This assumes that all the mathematics has been worked out previously, and it is only the program, rather than the general algorithm which must be tested. The latter could quadruple the development time. Most unfriendly ecology programs are about 1,000-10,000 lines. The time will be much greater if the arrays of data are large (as is often our case), and if one is trying to fit a big program into a small memory.
2. It takes at least twice as long to make a program user-friendly as it does to make it functional. Surprisingly, the easiest part of a program to make is often the computational part. The most difficult are the parts of the program which involve data input, data screening, graphical output, and most of all the parts which tie the other parts together.
3. Programming is not a fruitful part-time activity. Serious programming is an all-consuming activity in which the programmer must be totally immersed in order to make headway. Ecologists often know what needs to be done, and can sometimes spare a few months to do it in a minimally workable form. However, an ecologist who spends a year or more making programs user-friendly has crossed the line from scientist to technician. There are very few working ecologists who can set aside a year or more of their careers for the sake of writing a set of programs. To do so would mean foregoing the four or five papers each year that a good career demands. Programming beyond the development stage should be done by technicians, under the advice or supervision of working ecologists.
4. After a year of development and in-house testing, the program must be sent out for use among those who invariably find errors when no programmer can. This is unavoidable. All major programs need improvements. That is why this is being drafted on Word Perfect 5.1 instead of just Word Perfect. One programmer even chose to call his initial product dBASE II rather than dBASE I, just so potential buyers would assume that the bugs had been already identified and corrected. After a year of field use, a program must be revised and distributed again. As research continues, the program must be updated accordingly. Witness, for example, the ten-year evolution of ICLARM's ELEFAN package (Naga, this issue).
5. The bottom line — there is no funding for such activity. Occasionally some funds are scraped up to turn an idea into a program, or more commonly to assemble some extant programs into a semi-convenient package. However, what funding agency will commit itself to fund a serious effort to produce, distribute and update a useful package of programs?

Obviously, what is needed is a farsighted funding agency which is willing to respond to a serious need in applied ecology.

The Need

Around the world, research on ecological communities is being held up by these problems. Practising community ecologists spend up to half of their research time each year keeping up with the technology of programming or employing this knowledge to manipulate data into analyzable forms and programs into workable shape. Hundreds of would-be ecologists are shelving projects before they begin — or worse, after completion, because of the complexity of doing simplifying analyses. Millions of dollars of data are molding in file cabinets because of the discouraging nature of ecological data analysis.

Developing countries are the hardest hit. Most developing countries are tropical, and virtually all ecosystems in

tropical countries are diverse. The need for simplifying techniques is greater in such countries than elsewhere. However, active scientists are a valuable resource in a developing country. A scientist in such a place must be concerned with cost-effective use of research time. Usually, cost effectiveness is judged in terms of immediate, short-term goals. Thus, a trawl study of a multispecies fishery will generally result in an analysis of distributions of biomasses of selected commercial species. It will rarely result in an analysis of what has happened to the community structure once fishing has commenced. Few scientists in developing countries can take time out to learn computer languages. Fewer still have research funds to devote to maintaining a highly paid programmer merely to broaden the range of analyses available from the immediately gratifying to the ultimately important.

A Way Out

Obviously, what is needed is a far-sighted funding agency which is willing to respond to a serious need in applied ecology. The agency could fund ecologists involved in developing analytical methods to hire personnel to make the programs user-friendly, distribute the programs, evaluate and debug them, and to update them as necessary. Funding for program development should be in terms of three to five years, and longer for packages of programs. As an interim measure, efforts should be directed at locating useful, tested routines and converting them into user-friendly programs for immediate distribution. The rewards for such minimal investment would be substantial. One could expect a high probability of success in producing a useful program, a very rapid transfer of technology, and noticeable upgrading of the level of ecological research being undertaken in the context of such priority areas as coastal zone management, biodiversity studies, environmental impact assessment, tropical forestry, red tide research and multispecies fisheries analysis.

Suggested Materials

Two books, out of the many currently available, present a well-rounded view of ecological data analysis. The first, by Green (1979), explains only the approach and refers the reader to other literature for details. However, it is the only good text I have seen which clearly explains how to approach problems, such as environmental impact assessment, where a clear and closed set of hypotheses is at issue.

For those working in less well-charted ecological waters, several recent books emphasizing pattern analysis might be helpful. However, none I have seen is as concise and instructive as the one by Pielou (1984). While many of the routines necessary for the approach of Green are available on advanced statistical packages such as SAS, very little of the work described in Pielou or at least five similar texts can be done with the major packages.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley and Sons, New York. 257 p.

Pielou, E.C. 1984. The interpretation of ecological data: a primer on classification and ordination. John Wiley and Sons, New York. 263 p.

Several colleagues and I have found that a good general approach to ecological community data analysis involves classification (divisive clustering) and tabular sorting on TWINSPAN, analysis of species distributions against environmental variables with CANOCO, and visual ordination assessment with the Multidimensional Scaling (MDS) program in PATN. TWINSPAN is reportedly now available in a microcomputer version of the Cornell Ecology Package, being distributed by Hugh Gauch. Neither the MDS nor TWINSPAN programs represent the current state-of-the-art, but they are at least available. I have a program similar to TWINSPAN which may be preferable in some circumstances, but it may be a few years before funding is found to make it user-friendly. Here are the references:

Belbin, L. 1987. PATN: Pattern analysis package reference manual. Commonwealth Scientific Industrial Research Organisation, Division of Wildlife and Rangelands Research, Canberra. 353 p.

Hill, M.O. 1979. TWINSPAN — A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of individuals and attributes. Cornell Unversity, Ithaca, New York. 90 p.

ter Braak, C.J.F. 1988. CANOCO — A FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (Vers. 2.1). Agricultural Mathematics Group, Wageningen, The Netherlands.

Finally, I include some references wherein these programs were used in the contexts of tropical fisheries and coastal zone management:

Gomez, E.D., W.Y. Licuanan and V.V. Hilomen. 1988. Reef fish-benthos correlations in the Northwest Philippines. Proc. 6th Internat. Coral Reef Symp. 3:245-249.

Hilomen, V.V. and E.D. Gomez. 1988. Distribution patterns of fish communities in some Philippine reefs. Proc. 6th Internat. Coral Reef Symp. 3:257-262.

Licuanan, W.D. and E.D. Gomez. 1988. Coral reefs of the Northwestern Philippines: a physiognomic approach. Proc. 6th Internat. Coral Reef Symp. 3:245-249.

McManus, J.W. 1986. Depth zonation in a demersal fishery in the Samar Sea, Philippines, p. 483-486. *In* J.L. Maclean, L.B. Dizon and L.V. Hosillos (eds.) The First Asian Fisheries Forum. Asian Fisheries Society, Manila, Philippines. 727 p.

McManus, J.W. 1989. Zonation among demersal fishes of Southeast Asia: the southwest shelf of Indonesia, p. 1011-1022. *In* Proc. 6th Symp. on Coastal and Ocean Management/ASCE, July 11-14, 1989, Charleston, SC.

Nañola, C.L. Jr., J.W. McManus, W.L. Campos, A.G.C. del Norte, R.B. Reyes, Jr., J.B.P. Cabansag and J.N.D. Pasamonte. 1990. Spatio-temporal variations in community structure on a heavily fished forereef slope in Bolinao, Philippines, p. 377-384. *In* R. Hirano and I. Hanyu (eds.) The Second Asian Fisheries Forum. Asian Fisheries Society, Manila, Philippines. 991 p.

Clustering and Ordination in Coastal Resource Management

About 50 years ago, systematic methods were developed for converting a raw data table listing sample units from a community against the abundances of species, into sets of sites and species representing somewhat uniform subcommunities. The technique is extremely tedious, fairly subjective, and requires considerable experience to insure that rational choices are made at each step. Some ecologists continue the same approach today. A decade later, techniques such as Principal Components Analysis were being used to produce scatter diagrams of sample units to aid in their interpretation. Cluster analysis (sometimes called classification analysis) was soon to follow, with routines to sort

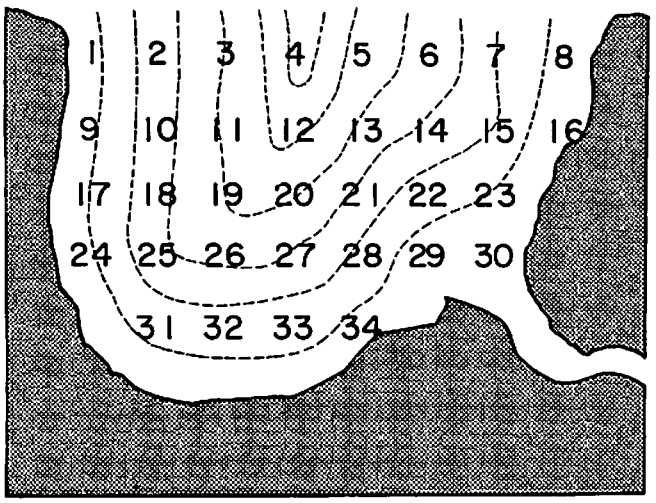

Fig. 1. A hypothetical estuarine bay with a grid of trawl stations. Depth contours could be in tens of meters.

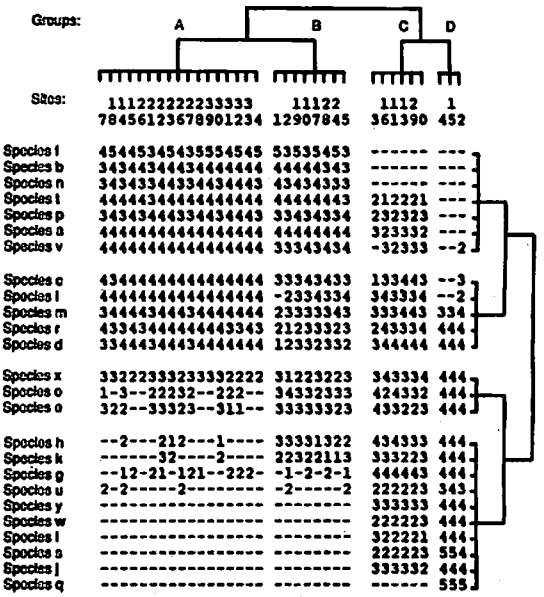

Fig. 2. TWINSPAN sorted table of species abundances at each trawl station in a hypothetical bay. Sample numbers are read vertically (7, 8, 14, 15, etc.). Each digit in the table represents a species abundance in a station converted to a scale of 1 to 5. Hierarchical relationships among station and species groups are indicated here with tree diagrams, wherein higher (or rightmost) bars indicate less similarity.

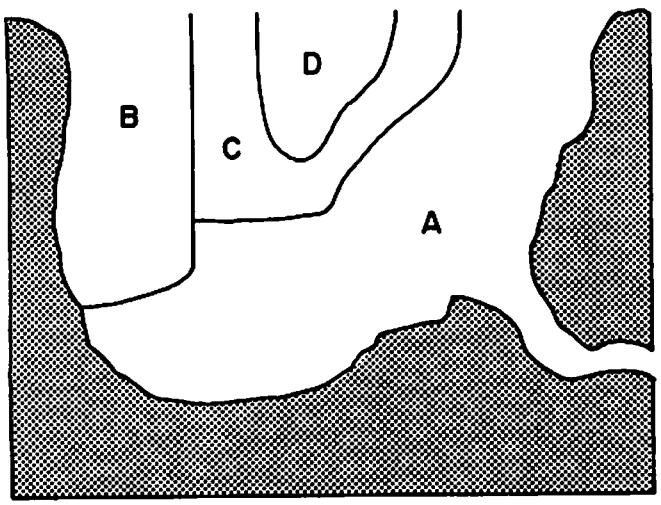

Fig. 3. Site clusters mapped for geographic interpretation. Notice the combined effects of depth and salinity in determining the way similar sites are distributed.

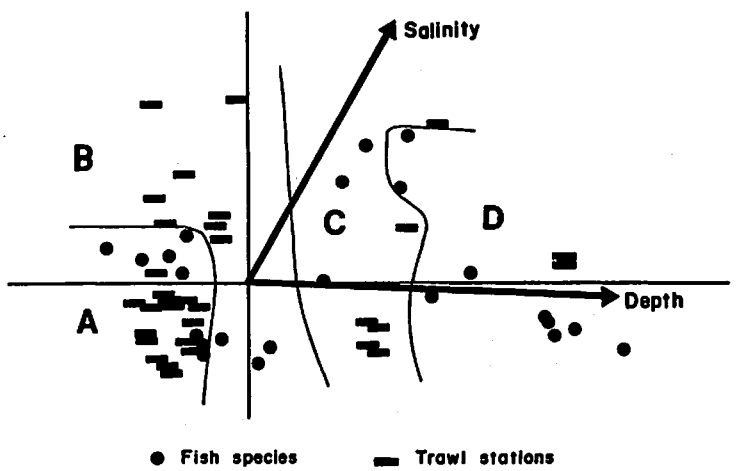

Fig. 4. Detrended canonical correspondence analysis (DCCA) of species, sites and environmental variables. Site groups from the TWINSPAN analysis (A-D) are separated by fine lines. The lengths of the arrows for environmental factors indicate their influence as variables correlated with site and species distributions. Subsequently, sets of species can be described in terms of their environmental preferences.

the sample units or species into levels of similarity in tree diagrams. Finally, in 1979, the TWINSPAN program was developed by M.O. Hill, which used ordination and clustering methods as a basis for sorting raw data tables into groups. Now ecologists can do what they set out to do initially, but they can do it more quickly, more objectively, and with biases which are much better understood. The next step would be to quantify these biases and potential errors, and work along those lines is underway.

Most community ecologists use a combination of clustering and ordination methods. Because the objective is often to explain distributions and abundances in terms of environmental (or human) factors, methods which compare the species abundance data with other variables are often helpful. As an example, imagine a bay with a broad range of depths and salinity values (Fig. 1). Abundance values for each species at each of a grid of trawl stations can be sorted in a table indicating groups of similar sites and species (Fig. 2). The site groups can be mapped for geographical interpretation (Fig. 3). Finally, the relationships between species, sites and environmental parameters can be illustrated in an ordination diagram (Fig. 4). In this case, it is clear that depth is the major factor differentiating the site groups, and that these groups are not particularly discrete.