The Role of Data in Scientific Progress, P.S. Glaeser (ed.)
Elsevier Science Publishers B.V. (North-Holland)

COMPUTERS IN PHYTOSOCIOLOGY

Mark V. Wilson

Department of Botany and Plant Pathology, Oregon State University
Corvallis, Oregon 97331-2902, U.S.A.

The use of computers in phytosociology has been stimulated in the United States by several historical developments: the application of multivariate analyses, the International Biological Program, phytosociological data processing packages, and large-scale and long-term ecological studies.
The advantages of computer use in phytosociology--speed, flexibility and the ability to store and retrieve data--are often balanced by drawbacks or difficulties. These include the limitations of computer packages, the mystique of computer-generated results and the use of diverse field methods.
The Oregon State University Forest Science Data Bank is discussed as a new but important ecological data storage and exchange system in the United States.

INTRODUCTION

Computer processing enhances phytosociological analyses, but also poses many challenges. This report outlines the benefits, drawbacks and possible new directions of computer use in phytosociology. I will concentrate almost exclusively on plant ecology in the United States. Related topics with a European (and broader) perspective are presented in van der Maarel et al. (1980).

HISTORICAL DEVELOPMENTS IN THE USE OF COMPUTERS IN ECOLOGY

Historical developments have stimulated computer use in ecology in the United States. The first development started with the importance placed in the 1940s and 1950s on gradient analyses of phytosociological data. Early researchers (e.g., Curtis and McIntosh 1951, Whittaker 1956) did their calculations manually, but as computers became more widely available in the 1960s, these same techniques became automated. The second development, the International Biological Program (IBP) during the 1960s, was a great stimulus to North American plant ecology. The emphasis of the IBP on intensive, ecosystem-level studies dictated the use of several forms of computer processing, including computer simulation modeling and data storage. Futhermore, the interaction of separate projects existing under the IBP umbrella indicated the potential usefulness of data sharing.

The third step in the evolution of computer use in ecology was the development in the 1970s of data processing packages specifically designed for phytosociological analysis. The most influential of these in the United States was the Cornell Ecology Program series (e.g., Gauch 1977), under the direction of R.H. Whittaker and H.G. Gauch. The availability of useful and well-documented programs brought computer processing of multivariate data into the mainstream of ecology in the United States. Moreover, these packages, simply by their widespread use, imposed a certain uniformity of data collection and analysis. It became possible to compare the results of separately conducted studies.

The fourth impetus, starting in the 1970s and continuing today, is the emergence of regional and global perspectives on both basic and applied ecological problems (Botkin 1980). For example, land use patterns, rates of deforestation, and increasing levels of atmospheric carbon dioxide became critical in understanding the relationship of the world's biota to global ecological systems. Although large-scale data collection efforts have been undertaken, most regional or global studies have relied on disparate sources of data. These experiences have stimulated a recognition of the importance of data storage and exchange among independent researchers. In a similar fashion, recent emphasis on long term ecological research (Callahan 1984) has stimulated concurrent efforts in efficient data management and exchange (Gorentz et al. 1982).

USES OF COMPUTERS IN PHYTOSOCIOLOGY

Current major uses of computers in phytosociology include data analysis, computer modeling, and data storage and exchange. In the following sections I will discuss the benefits and drawbacks of each of these uses, and some possible future developments.

Data analysis

Both the advantages and drawbacks of computer processing of data stem from the type of data collected in phytosociology. Typical phytosociological data sets are observational and consist of matrices of species abundances within individual samples, and perhaps of environmental attributes within the same samples. Therefore, phytosociology data are multivariate data. The goals of analysis include the recognition of natural groups of species and/or samples, the recognition of underlying vegetation gradients, and the exploration of the relationships of vegetation to environment. The complexity of vegetation has several implications for phytosociological analyses. First, data sets are nearly always study-specific, that is, the results of one study are not repeatable in detail in other systems. (This contrasts with the repeatability of laboratory results common in the physical sciences). Second, single environmental or other factors are seldom sufficient to explain vegetation patterns. Instead, several explanatory factors are important and interact in complex ways. Third, single techniques of analysis are seldom universally applicable to phytosociological cata collected from different vegetation. Existing techniques (including options within techniques) are variously sensitive to changes in such vegetation attributes as the number of species present, the degree of spatial patterning and the extent of vegetation change.

The analysis of complex vegetation data usually requires sophisticated processing tools implemented on computer systems. The advantages of data processing in phytosociological analysis are basically speed and flexibility. The speed of computer analysis compared to manual analysis allows the search for ecological patterns that would have been otherwise untractable. Once data have been entered into a computer system, the researcher has the flexibility to try different types of analysis. Since no single technique is universally applicable to all phytosociological data, the speed and flexibility of computer processing can greatly enhance the success of research.

The advantages of computer processing in phytosociological analysis are balanced by several drawbacks. One drawback, perhaps more common when computers were first being used, was the limited availability of adequate computer programs. Some techniques may have been used more because they were available on the researcher's computer system than because the technique was suitable for phytosociological analyses. A second drawback is the apparent significance of some computer-generated results. This problem is part methodological and part psychological. Computer analyses, especially involving techniques poorly understood by the general scientific audience (or, unfortunately, by the researcher), are sometimes not subjected to the healthy scepticism often directed at other results. This mystique will disappear as first-hand knowledge of computer processing techniques becomes more widespread.

Phytosociological analyses, because of their multivariate nature, can also produce apparently meaningful patterns where none in fact exist. For example, eigenanalysis of randomly generated species-by-samples data sets can produce a concentration of information on the first derived axes, even though these axes have absolutely no ecological significance (H.G. Gauch, pers. comm.). A solution to this methodological problem is the application of statistical tests (e.g., Wilson 1981) to phytosociological analysis.

Computer modeling

Computer modeling can be an important tool in phytosociology. Models serve to make explicit our assumptions about vegetation, and can point out critical yet unexplored types of information. Moreover, because the slowly changing nature of vegetation often makes conventional field and analytic techniques impractical, computer modeling can be the only feasible form of analysis, especially for prediction. Forest succession research, in particular, benefits from the application of computer models to field data (e.g., West et al. 1981, Means 1982). On the other hand, vegetation is complex, poorly controlled, subject to numerous exogenous forces, and hard to replicate. Therefore, vegetation is very difficult to model, and the resulting models are often difficult to validate and to generalize. The solution to these problems is simply to continue using modeling in understanding vegetation, and to continue using vegetation studies in improving computer models.

Data storage and exchange

A simple form of data storage is the result of entering phytosociological data for computer analysis. This type of storage is often useful, however, only for the original researchers. Efficient data management, in contrast, can provide for the primary needs of the primary researchers, while also providing for data retrieval by colleagues in the same research project and for data capture beyond the tenure of the project.

The exchange of stored data benefits ecological research in many ways. Phytosociology is a field science; replicated observations and controlled experiments are often impossible to conduct. The alternative is to repeat measurements, experiments and analyses at many sites and to search for patterns. Efficient computer-based exchange of primary data would greatly facilitate and increase the scope of such comparative studies.

Some ecological phenomena are inherently large scale or long term and must involve data exchange. Examples of large-scale phenomena include the geographic variability of species or of vegetation, the distribution and effects of acid precipitation, and the atmospheric carbon dioxide enhancement. Examples of long-term phenomena include trends in ecosystem productivity, cycles of tree seed production, and succession. Except in unusual cases, studies at these large scales or long term cannot be conducted successfully by single investigators or even by single teams of investigators. Therefore it is imperative for phytosociological research that efficient systems of data storage and exchange be developed.

Confronting the obvious benefits of computerized data storage and exchange are several difficulties. First, there is little uniformity in the United States in methods of collection of phytosociological and other ecological data. The investigation of some phenomena requires the use of identical methods. For example, Glenn-Lewin (1977) surveyed geographical and landscape trends in species richness (the number of species in a sample). Species richness is thought to be function of climate, soil fertility, disturbance history and other factors of interest. But species richness also increases simply with sample area, and the form of this relationship is unpredictable. Therefore Glenn-Lewin (1977) was limited to studies using identical sample areas, conducted by R.H. Whittaker and associates. There is no simple general answer to the problem of disparate field methods. At the least, primary data for computerized exchange must be accompanied by thorough documentation of the field methods used. As a longer-term solution, plant ecologists should agree on a few acceptable and flexible field methods, to which individual researchers can adhere if they choose. That is, I am proposing a form of voluntary standardization that could expand the usefulness of field data beyond the original intentions.

A second difficulty is the problem with inadequate attribution. Researchers are often reluctant to share their hard-earned phytosociological data. Several solutions are possible, including access to centrally-stored data by permission only, the requirement of prominent acknowledgement if exchanged data are used, and possible co-authorship.

A third and related problem is the assurance of the cooperation of primary researchers, the suppliers of data. Researchers, including phytosociologists, often work independently and are under pressure to publish; sometimes the extra effort to place their data into a data bank does not seem worthwhile. One solution to this problem is for a higher authority to mandate the contribution of data to data banks. In the United States, such authorities might be funding agencies, such as the National Science Foundation, or employers of large numbers of researchers, such as the Forest Service. A more generous solution would be to lure the contribution of data to a data sharing system with the reward of efficient data management, and the availability of sophisticated tools of analysis and graphical display.

THE OREGON STATE UNIVERSITY FOREST SCIENCE DATA BANK.

One of the major ecological data base management systems in the United States is the Oregon State University Forest Science Data Bank (FSDB). The FSDB is administered by the O.S.U. Department of Forest Science, and serves the entire College of Forestry, the Forestry Sciences Laboratory (the local research arm of the United States Forest Service) and individual researchers. The goal of the FSDB is to make available past and current data sets collected by and of interest to its clientele (Stafford et al. 1984). There have been two important stimuli to the development of the FSDB. First, many data sets of interest were inherited from the International Biological Program Coniferous Biome Project. A major effort of the FSDB has been the machine entry of documentation for both the field methods and the data set structure for the IBP data. Second, Oregon State University was chosen in 1980 as an administration site for the National Science Foundation's Long Term Ecological Research program, which focused additional attention on efficient ecological data base management. As a response, in part, to these developments, resources at Oregon State University were pooled during 1980-1981 to form the Quantitative Services Group, a major task of which was the creation of the Forest Science Data Bank system.

The FSDB now contains data sets and supporting documentation for phytosociological, ecosystem, sivicultural, genetic and other types of studies. Currently the FSDB contains approximately 1425 individual data sets, of which about 300 are vegetation data sets. (Each vegetation data set typically contains numerous individual samples within a given site.) The FSDB is organized in a two-tier fashion. An extensive relational data base management system contains separate data bases of study abstracts, detailed study documentation, references, formats, variable explanations, and data file organization. These data bases are used (1) to identify a particular data set of interest, (2) to verify, via on-line scanning, the relevance of the data set, (3) to determine the archival location of the data set, and (4) to retrieve formats and explanations that allow the use of the data set. The complete data sets themselves are stored on magnetic tape under a separate system. The incorporation of data sets into the data bank is facilitated by the use of a standard regional species list (Garrison et al. 1976), standard data forms, and data entry services.

One of the biggest challenges of the FSDB in its short history has been the collection of adequate documentation from disparate sources. Documentation is required for both field methods and the structure of data sets. A second major challenge is the maintenance of the cooperation of primary researchers. These investigators are the sources of new data for the FSDB. Both challenges are addressed by encouraging use of the system. The FSDB managers feel that researcher cooperation, both in supplying new data sets and in providing adequate documentation, is best stimulated by providing an easily accessed and thoroughly useful system. Towards this end two tactics have been employed. One, the FSDB is simple and accessible: it is available for direct use by researchers without technical computer backgrounds. Two, the data base management system is fast, flexible and well-designed, and statistical analyses and graphics capabilities for data in the FSDB are available through the parent Quantitative Services Group. In this way FSDB use and FSDB quality feed back upon one another, for the improvement of both (see Figure 1). The result is an ecological data base that grows in its service to both local and national researchers, and thus improves the quality of ecological research.

Data Bank Design Goals

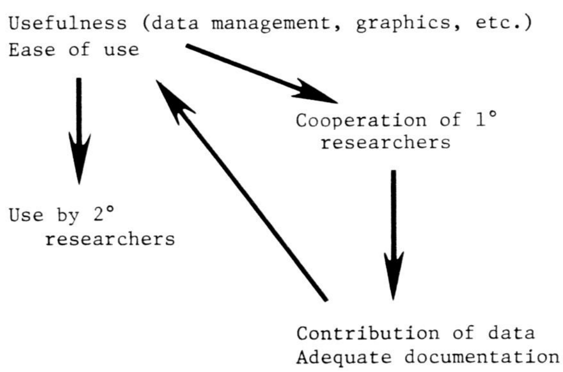

Figure 1. The interaction of data bank usefulness and use by researchers.

REFERENCES:

(1) Botkin, D. B., Life from a Planetary Perspective: Fundamental Issues in Global Ecology. (Final report NASA Grant NASW-3392, 49 pp., 1980).

(2) Callahan, J. T., Long-term ecological research, BioScience 34 (1984) 363-367.

(3) Curtis, J. T. and McIntosh, R. P., An upland forest continuum in the prairie-forest border region of Wisconsin, Ecology 32 (1951) 476-496.

(4) Gauch, H. G., ORDIFLEX -- A Flexible Computer Program for Four Ordination Techniques: Weighted Averages, Polar Ordination, Principal Components Analysis, and Reciprocal Averaging, Release B, Cornell University (1977).

(5) Garrison, G. A., Skovlin, J. M., Poulton, C. E. and Winward, A. H., Northwest plant names and symbols for ecosystem inventory and analysis, United States Forest Service General Technical Report PNW-46 (1976).

(6) Glenn-Lewin, D. C., Species diversity in North American temperate forests, Vegetatio 33 (1977) 153-162.

(7) Gorentz, J., Alaback, P., Dyer, M., Farrell, M., Koerper, G., Maroses, M., Marzolf, G. R. and Weiss, S., Data Management at Biological Field Stations, Report of a Workshop at Kellogg Biological Station (May 1982).

(8) van der Maarel, E., Orloci, L. and Pignatti, S., Data-processing in Phytosociology (Junk, The Hague, 1980).

(9) Means, J. E. (ed.), Forest Succession and Stand Development Research in the Northwest, Proceedings of the Symposium held at Corvallis, Oregon (1982).

(10) Stafford, S. G., Alaback, P. B., Koerper, G. J. and Klopsch, M. W., Evolution of a forest science data bank, Journal of Forestry (in press).

(11) West, D. C., Shugart, H. H. and Botkin, D. B., Forest Succession (Springer-Verlag, 1981).

(12) Whittaker, R. H., Vegetation of the Great Smoky Moutains, Ecological Monographs 30 (1956) 1-80.

(13) Wilson, M. V., A statistical test of the accuracy and consistency of ordinations, Ecology 62 (1981) 8-12.