Teaching Multivariate Data Analysis in the Fields of Biology and Ecology

Hans-Peter Bäumer
Carl von Ossietzky University of Oldenburg
HRZ-Applied Statistics
P.O.B. 2503
D-26111 Oldenburg, FRG
E-mail: baeumer@hrz.uni-oldenburg.de

1. Introduction

The increasing demand for statistical literacy combined with the dissatisfaction concerning the skills in applied statistics of students majoring in other fields [see e.g. Romero et al. (1995), par. 2 ff.] has led to a "widespread consensus among statisticians that a beginning course should emphasize practical understanding" [Velleman and Moore (1996), p. 218]. In relation to the content, the "key word here is data" [Joiner (1986), p. 30] since many years. Consequently, the focus has turned to multivariate data analysis (MDA). But teaching MDA should not be separated from the at least emulated solving of real world problems [cf. Yilmaz (1996), par. 4 ff.]. Furthermore, an applications-oriented approach why and how to collect efficiently experimental data as well as how to analyse reliably multivariate data obtained in laboratory experiments and/or in field surveys or experiments meets the requirements Biologists and Ecologists are confronted with in their profession. Therefore, teaching the subject area in the fields of Biology and Ecology is embedded in the common circle of scientific problem solving exemplified by selected substantive research which is henceforth called the emulated research approach. As far as these embedding is generalizable to the teaching of MDA in other fields it will be presented and discussed in some detail below.

2. Prerequisites of Teaching MDA in the Fields of Biology and Ecology

MDA encompasses interpretative activities [cf. Cobb (1998), p. xviii]. Contrary to the teaching in deductive sciences as mathematical statistics a substantive context provides meaning. Regularly, students in Biology and Ecology have no previous background in statistics. At university level they are usually faced with multivariate problems including space-time variability. With spatiotemporal techniques in confirmatory statistics - as spatiotemporal random fields - strong mathematical skills are required. In exploratory multivariate data analysis, techniques are available to analyse space-time dependent variables - as nonmetric multidimensional scaling (NMDS) or canonical correspondence analysis (CCA) – which may be introduced without much technical efforts and mathematical derivations. Techniques to present in the context of teaching MDA should already have been successfully applied to real world problems in Biology and/or Ecology [see e.g. Clarke and Warwick (1998)].

3. Aims

Objectives in teaching MDA in the fields of Biology and Ecology are to motivate students to change their attitude about MDA and to locate its place in the context of scientific research processes, to foster students in grasping of some concepts basic to MDA like population and sample, randomness and the frequency model behind, independence, spatial, temporal and spatiotemporal dependence, variability [cf. Moore (1997), p. 127] as well as some principles in experimental design like scale of measurement, randomization, blocking, replication vs. pseudoreplication, to promote students to understand some core concepts in MDA like the difference between association and cau-

sality, between exploratory and confirmatory analysis, between data reduction or visualization and statistical inference and the distinction between univariate, multiple and multivariate analysis, to assist students in developing a flexible ability to apply a few basic concepts – as choosing a similarity measure in analogy to substantive ideas on similarity, to encourage students to work cooperatively and present substantive conclusions drawn from the results obtained in applying MDA, and last but not least to qualify students to communicate fruitfully with a consultant statistician.

4. What Kind of Software to Incorporate?

First of all, the intended educational aims will determine the kind of software to employ in teaching MDA. Therefore, a full-function statistics software system takes priority of other kinds to ensure primarily the transfer of the skills in data handling, data analysis and representation of results applying the system to tackle problems outside the lecture room. In fact, the professional statistics software systems which are widely accepted in the biological and ecological research community determine the range of tools to consider. To choose in the universe of existing systems it should not be overlooked that the statistical analysis of data only covers a small part in the process of practising MDA. Evidently, some standard techniques applied in Biology and/or Ecology must be implemented in the favoured easy-to-use and easy-to-learn statistics software system. Furthermore, some interfaces for the import of data stored in a common format as well as their export, saving the actual analysis in a command file, strong graphical features and last but not least numerical reliability [see e.g. Sawitzki, G. (1994)] are required. Insofar, the experiences with the medium sized system SYSTAT are satisfying. But, dependent on the substantive problem under investigation the working environment for MDA has to be supplemented by programmes which are accepted tools in biological and/or ecological practice like BMDP which is accessible through SYSTAT, PRIMER or CANOCO.

The statistics software applied is not an integral part of an appropriate multimedia system. Actual multimedia systems which support to teach statistics will “not necessarily promote discussing problems, working cooperatively, and communicating conclusions” [Velleman and Moore (1996), p. 219]. Further pros and cons are discussed in the last mentioned paper. From the pedagogical perspective, some criteria software systems incorporated in teaching statistics should fulfil are extensively discussed in the literature [see e.g. Biehler (1997)].

5. The Emulated Research Approach

The most important steps in the emulated research approach to teach MDA are represented in Fig. 1. Furthermore, the following example will provide an illustration how to proceed from the substantive question to answers via the explicit research hypothesis to applying MDA and the substantive interpretation of results back to the substantive problem under investigation.

Let us begin with the question whether drastic disturbances lead to changes over time in the vegetation of typical Northwest German lowland rivers. Succession models and the corresponding basic processes, like invasion, maintenance, decline and extinction, constitute the theoretical background. For further details see Wiegleb et al. (1989). Then, the starting question is reduced to the manageable research hypothesis that the dynamics of some hydrophytic and helophytic species are influenced by strong disturbances in the upper course of one of the rivulets called Lethe. The spatial and temporal scale of the investigation, site selection, the variables under consideration as well as the data collection are criticized and an appropriate sampling scheme is worked out. In a first step the maximal percentage cover per year of 14 species for the years 1978 to 1988 are proposed to be analysed. Some strategies in MDA to deal with spatiotemporal problems in Biology and Ecology are characterized. Consequently, the (dis)similarity concept is preferred to quantify the (dis)similarity between pairs of years and some (dis)similarity measures are introduced. The construction ideas behind these measures as well as their properties are discussed [see e.g. Pfeifer et al. (1998)]. (Dis)similarity matrices are computed and the outcomes are tried to interpret with special regard to

the implications of pairwise comparison. After a description of the advantages and disadvantages of Kruskal's model in NMDS, multivariate analysis is then started by applying this approach. The results which are documented and orally presented lead to the substantive conclusion of a directional variation in the percentage cover of some hydrophytic and helophytic species over time in the upper course of the rivulet under scrutiny. To proceed with the findings obtained so far species and environment relationships are taken into account. The variable disturbance and some meteorological variables are included in the further analysis. CCA is introduced and its assumptions, limitations and special drawbacks in the context of ordination are discussed [see e.g. Jongman, R.H.G. et al. (1995), p. 91 ff.]. CANOCO, now available in version 4, is employed to practise MDA and the emulated research approach is continued. Splitting the assemblage of participants into small groups active learning is inherent in each step to foster understanding of basic concepts in MDA.

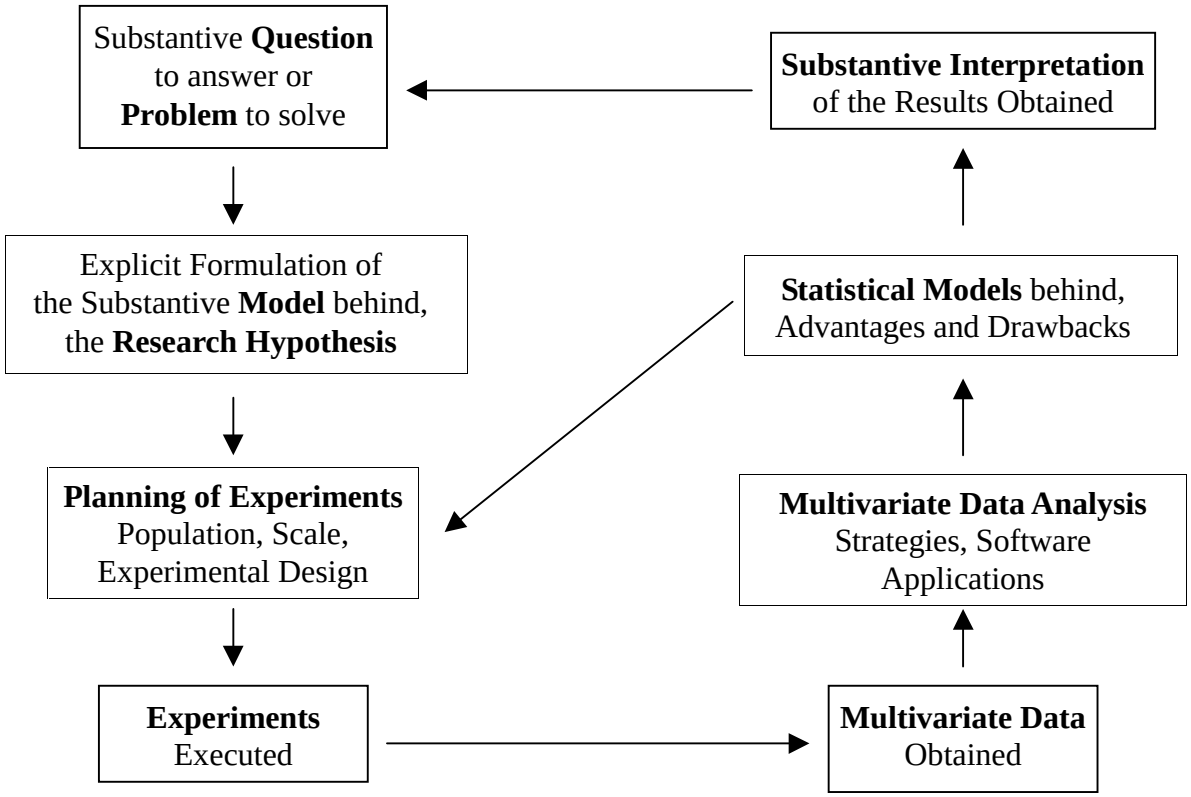

Figure 1. Steps in the Emulated Research Approach to Teach Multivariate Data Analysis

6. More Requests Than Answers

Evidently, in the statistical profession there does not exist a consensus on the content to teach to non-specialists in an introductory course of MDA [see e.g. Discussion to Moore (1997), p. 143 ff.]. Teaching in the framework of the emulated research approach allows to some extent to decide rationally on this content with regard to the substantive research hypotheses under scrutiny.

Incorporating a full-function statistics software system supplemented by further software tools confronts students with "the complexity of tool problem" as well as with "the variety problem"[Biehler, R. (1997), p. 169 ff.]. On the one hand it is the rule rather than the exception that students in the fields of Biology and Ecology are familiar with software applications in other contexts. Furthermore, preparing small command files as stepping-stones for the students as well as employing the command log after interactive sessions may mitigate these problems to some extent. On the other hand the skills gained in applying a full-function statistics software system may soon be obsolete. Consequently, the software is looked upon as merely a tool in dealing with substantive problems and the emphasis is on concepts with regard to MDA.

Assessment is one of the most crucial topics in learning and teaching [see e.g. Garfield (1994)]. Assessment should reflect the content, pedagogy and technology as well as their interrelations and should primarily aim to improve the learning as well as the teaching. In the framework of the emulated research approach a small group's solution to a substantive problem or question and its presentation comprising substantive conclusions is the preferred technique to evaluate learning and teaching.

The syllabus of students in Biology or Ecology often includes several practical courses in laboratory as well as field experiments. In the run of such courses problems are often encountered which could adequately be tackled applying MDA. As I plead for teaching MDA embedded in substantive content statistics lecturers should be involved in developing the curricula for such practical sequences. Furthermore, a statistics lecturer should take part in their carrying out in a more interdiscipline-oriented framework. Then, students as well as colleagues will be convinced of the necessity and merits of MDA with regard to their daily laboratory and field work.

At last, different concepts in pedagogy applied in teaching MDA should not degenerate to tenets. Therefore, pedagogical research projects should accompany the implementation of a new concept to control for its impact on learning and teaching.

REFERENCES

Biehler, R. (1997): Software for Learning and Doing Statistics. International Statistical Review, 65, 167-189

Clarke, K.R. and Warwick, R.M. (1998): Quantifying structural redundancy in ecological communities. Oecologia, 113, 278-289

Cobb, G.W. (1998): Introduction to Design and Analysis of Experiments. Springer. New York etc..

Garfield, J.B. (1994): Beyond Testing and Grading: Using Assessment to Improve Student Learning. Journal of Statistics Education, 2 (1); URL: http://www.stat.ncsu.edu/info/jse/v2n1/garfield.html

Joiner, B.L. (1986): Transformation of the American Style of Teaching Statistics. In The Next 25 Years in Statistics (eds W.J. Hill and W.G. Hunter), 30-33. Report No. 10. University of Wisconsin-Madison, Center for Quality and Productivity Improvement, Madison.

Jongman, R.H.G., ter Braak, C.J.F. and van Tongeren, O.F.R. (eds.) (1995): Data Analysis in Community and Landscape Ecology. Cambridge University Press. Cambridge etc..

Moore, D.S. (1997): New Pedagogy and New Content: The Case of Statistics. International Statistical Review, 65, 123-165

Pfeifer, D., Bäumer, H.-P., Dekker, R. and Schleier, U. (1998): Statistical Tools in Monitoring Benthic Communities. Senckenbergiana maritima, 29, 63-76

Romero, R. Ferrer, A., Capilla, C., Zunica, L., Balasch, S., Serra, V. and Alcover, R. (1995): Teaching Statistics to Engineers: An Innovative Pedagogical Experience. Journal of Statistics Education, 3 (1); URL: http://www.stat.ncsu.edu/info/jse/v3n1/romero.html

Sawitzki, G. (1994): Report on the Numerical Reliability of Data Analysis Systems. Computational Statistics & Data Analysis, 18, SSN, 289-301

Velleman, P.F. and Moore, D.S. (1996): Multimedia for Teaching Statistics: Promises and Pitfalls. The American Statistician, 50, 217-225

Yilmaz, M.R. (1996): The Challenge of Teaching Statistics to Non-Specialists. Journal of Statistics Education, 4 (1); URL: http://www.stat.ncsu.edu/info/jse/v4n1/yilmaz.html

Wiegleb, G., Herr, W. and Todeskino, D. (1989): Ten years of vegetation dynamics in two rivulets in Lower Saxony (FRG). Vegetatio, 82, 163-178

RÉSUMÉ

Il faut tenir compte de nombreuses limites, quand il s'agit d'établir un programme d'études pour un cours introductoire d'analyse des données multivariées en biologie et en écologie. Il s'agit, ici, d'un accès interdisciplinaire à l'enseignement de cette matière qui s'intègre dans le domaine usuel de solutions scientifiques aux problèmes qui peuvent surgir, et cet accès sera discuté en détail.