0

Introduction

(What this book is all about, how and why?)

For the researcher, it is a source of both pleasure and difficulty that biological objects can often be characterized simultaneously by numerous – occasionally by hundreds of – properties: *the subject matter is of multivariate nature.* Biological surveys commonly produce substantial amounts of information. These form a seemingly incomprehensible and inpenetrable mass of data and hide deeper relationships among objects or properties. Even though the biologist is familiar with some features of the objects from personal experience gained during data collection, their presentation in an easily understandable format is impossible without *multivariate data analytical techniques.* Currently, these methods have become everyday tools of data evaluation in biology, and in most areas, especially in ecology and taxonomy, knowledge of the underlying principles is essential.

If we consider the main objectives of multivariate analysis, two areas of application can be separated sharply. Some of the methods are obvious extensions of uni- and bivariate statistical procedures thoroughly discussed in the biometric literature. As such, they allow significance tests of statistical hypotheses. Typical examples are multivariate analysis of variance (MANOVA) in which the effect of treatments is measured on several variables simultaneously and multiple regression, which seeks functional relationships between a dependent variable and several 'independent' variables. A natural element of these statistical tests is *estimation* of some *parameter* of the '*population*' (understood as a statistical term, which should not be confused with the same word as used in the biological sciences). The estimates can serve as a basis for the detection of causal relationships and for the construction of models suitable for prediction. For example, the estimated regression coefficients can be used to predict the values of the dependent variable even for unobserved combinations of the independent variables. These methods may be simply referred to as *multivariate statistical procedures.*

A quick overview of the history of biology shows that the alternative to estimation, the pattern detecting and data structure exploring function of multivariate methods, is equally if not more important than hypothesis testing. In this case, the objectives include the extraction of essential features of the data, the discovery of latent structural relationships or a mere summarization, visualization and description of biological pattern in the form of mathematical constructs (artificial dimensions in ordination scattergrams, trees, partitions, etc.). The main purpose is thus *data exploration* via procedures commonly labeled as "*exploratory data analysis*" in the literature, referring to methods of cluster analysis and ordination (e.g., multidimensional scaling). Estimation and subsequent statistical inference are in this case of secondary importance or even negligible, although there are areas in which hypothesis testing is an integral part of data exploration (e.g., comparison of results, Chapter 9).

This book lays emphasis on the second group of multivariate methods, so that statistical significance tests will only be used occasionally as tools supplementary to data exploration. The reader may have the feeling that many concepts and terms so familiar from conventional biometric studies (e.g., distribution, significance level, estimation, null-hypothesis, test, error, parameter) are much neglected in this book, but this only underlines the contrast between the two principal objectives of multivariate methods.

The English language literature abounds with books that focus on the application of exploratory multivariate analysis to biological problems. Most of these books, however, are specialized to a certain area of biology, and do not place the contents into a general context. Much useful information is dispersed over the vast biological and mathematical literature, leading to an undesirable isolation of different fields. To give an exhaustive treatment of the subject matter in a single book is obviously a difficult if not impossible task, but I attempt to collect at least the major aspects in a format somewhere between the reference books and postgraduate texts. The chapters were written to illustrate the high diversity of the topic and to illuminate as many approaches to data exploration as possible. The literature reviews following each chapter, and the extensive bibliography at the end of the book, will facilitate further orientation for readers wishing to get more insight into a particular problem. Selected computer program packages, without which data exploration would be an impossible adventure, are also discussed and characterized in brief.

Attention is focused on the 'supraindividual' biological level, for example plant ecology, phytosociology and taxonomy, showing that I am somewhat predisposed towards these typically multivariate subjects. Of course, the 'multivariate situation' appears much more often in biology, as shown by Table 0.1. Fortunately, the contents of the book can be easily 'translated' to fit any other areas of biological sciences. It is the reader who should take the (hopefully simple) job of adapting the jargon to his/her own field of interest. For example, if quadrats or other sampling units and species occurring in them are mentioned in the context of vegetation science, then these terms should be replaced in your mind by the most appropriate type and name of object and variable.

The biological significance of multivariate data exploration has been emphasized by many authors. A relatively fresh overview by James & McCulloch (1990) argues – with some reservations – that "it is no longer possible to gain a full understanding of ecology and systematics without some knowledge of multivariate analysis. Or, contrarywise, misunderstanding of the methods can inhibit advancement of the science." This statement is supported by the thematic evaluation of seven widely appreciated taxonomical and ecological journals for the years of

Table 0.1. Multivariate situations in various fields of biology and related disciplines.

Discipline	Objects	Variables
Ethology	species	behavioral characters
Paleontology	layers, strata	species
Anthropology	findings	morphological traits
Biogeography	species	distributional properties
Medicine	diseases	symptoms
Genetics	populations	allele frequencies
Molecular biology	proteins	amino acid per position
Ecophysiology	species	photosynthesis types
Agronomy	cultivars	crop features
Forest science	tree species	age classes
Hydrobiology	water courses, lakes	water quality indicators
Psychology	test patients	answers to questions
Microbiology	bacteria	substrates
Soil science	soil profiles	particles
Bioclimatology	habitats	climatic features

1983-1988. The authors found more than 500 papers with applications of multivariate methods, the first three most frequently used procedures being 1. principal components analysis, 2. discriminant analysis (which involves tests and predictive elements as well) and 3. cluster analysis.

In this book, the term ‘multivariate method’ will be understood in the most general sense. The usage is definitely wider than in mathematics, for example, because cladistics is also included. Procedures of numerical classification receive much more attention than in several standard texts on multivariate analysis (e.g., Mardia et al. 1979); the two major approaches to data exploration, clustering and ordination are treated in a more balanced way. Since the biologically-oriented reader is assumed to be of the ‘visual’ type, the number of figures and diagrams is more than usual in biomathematical texts. The very first illustration (Figure 0.1) serves as a summary of the major methodological pathways and choices discussed in the book[1]. The scheme does not – and cannot – show all the possibilities, but it illustrates what you can expect to find in this book. It is unlikely that this chart will ever be used to select the particular methods to be applied, yet it is a reasonable summary. The main pathway of the flowchart is the ‘*statistical population* → *data matrix* → *distances...*’ route, which almost always occurs. Then follow some more concrete choices towards classification or ordination.

[1] I admit that I am not very enthusiastic about such schemes and flowcharts, because they are rarely successful. Often, the diagrams are too detailed and incomprehensible, hence useless, and in other cases they are so simple that they add nothing to the discussion. In this case, however, I felt that this diagram composed of small pictograms will facilitate a quick overview of the subject matter treated in this book.

Figure 0.1. Scheme illustrating major pathways of multivariate data exploration in biology (opposite page).

To the bottom of the figure many more arrows could have been directed (there are only three symbolic ones); this reflects my contention that we can never be satisfied with the results of ordination and classification, and further evaluation (e.g., comparison of alternative results for the same set of objects) is almost always necessary.

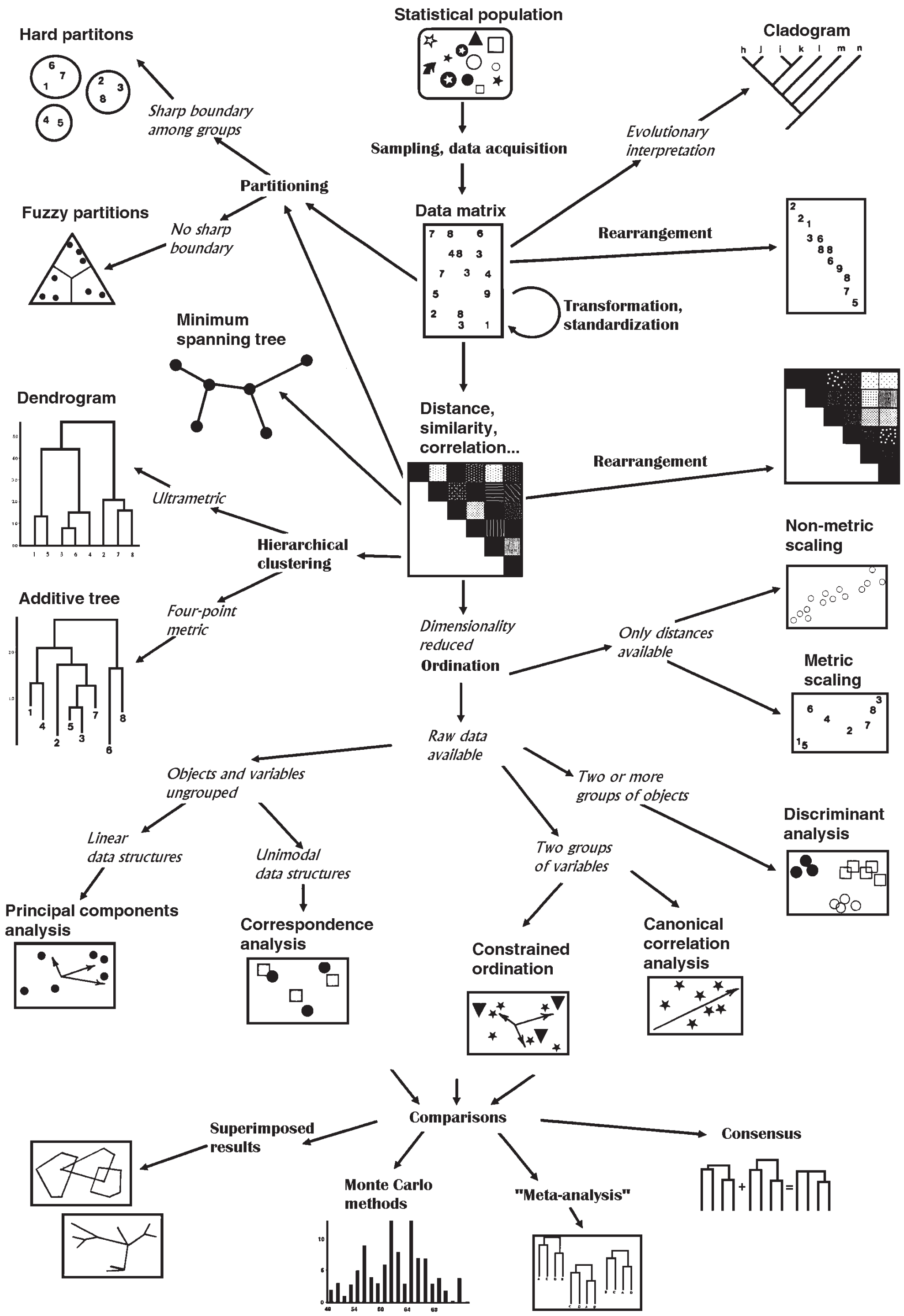
The structure of the book

This introduction is followed by nine major chapters, discussing the multivariate procedures in a division thought to be didactically most advantageous. These chapters need not be read in the order they are numbered, however. Although there are many cross-references in the text, almost all chapters can be considered as a separate reading. For those interested in cladistics, for example, Chapters 1-5 will provide little information except for a few paragraphs. To understand ordination methods, full knowledge of the third chapter is not essential, and so on. It is Chapter 9 that relies the most on previous chapters, and this is understandable because it deals with the comparison and evaluation of results. The internal structure is identical in all chapters: the methodological discussion is followed by a short literature/software review and the relatively dry subject matter is refreshed by an imaginary dialogue in the closing subsections. Three appendices are supplied containing the sample data matrices, information on the availability of software and a summary of matrix algebra. The list of references is not merely a bibliography but also an index of authors. Therefore, I must apologize that references according to second and subsequent authors in the text cannot be found based on this list. The book is closed by a subject index.

Acknowledgements

The text of the Hungarian edition of this book was scanned carefully by many colleagues and friends, contributing considerably to the elimination of errors and misunderstandings. I am especially grateful to Gy. Kontra for his thorough citicism and for pointing out several inconsistencies appearing in early versions of the manusrcipt. Encouragement and the valuable comments by B. Tóthmérész, J. Garay, P. Ódor, A. Demeter, K. Kontra, L. Peregovits, T. Czárán, I. Scheuring, and J. Tamás are also appreciated. I thank my students for their attention and moderate scepticism expressed during regular and special courses I gave on this topic, and that they had been for some time the ‘victims’ of unfinished parts of the book. I am indebted to G. Copp (University of Hertfordshire) and S. S. Talbot (U.S. Wildlife Service) for checking and ‘polishing’ the English translation of some chapters. To be honest, this work could not have been completed without the indirect contribution of many biologists and mathematicians working in this interdisciplinary field all over the world. Nonetheless, all errors appearing in the book are mine.

Thanks are due to the developers and dealers of some program packages mentioned in the book, for placing the software free of charge at my disposal: **Statistica** (StatSoft Inc., Tulsa,


Statistical population
Sampling, data acquisition
Data matrix
Transformation, standardization
Rearrangement
Evolutionary interpretation
Cladogram
h j i k l m n
Hard partitons
Sharp boundary among groups
Partitioning
No sharp boundary
Fuzzy partitions
Minimum spanning tree
Distance, similarity, correlation...
Rearrangement
Dendrogram
Ultrametric
Hierarchical clustering
Additive tree
Four-point metric
Dimensionality reduced
Ordination
Only distances available
Non-metric scaling
Metric scaling
Raw data available
Objects and variables ungrouped
Two or more groups of objects
Discriminant analysis
Linear data structures
Unimodal data structures
Two groups of variables
Principal components analysis
Correspondence analysis
Constrained ordination
Canonical correlation analysis
Comparisons
Superimposed results
Monte Carlo methods
"Meta-analysis"
Consensus
+
=

Oklahoma, USA), **BMDP** (Statistical Software Ltd., Cork, Ireland) and **PHYLIP** (J. Felsenstein, University of Washington, Seattle, USA).

Financial support for preparing the Hungarian edition of this book came in form of a Hungarian National Scientific Research Grant (OTKA) #T6032, whereas the costs of the first printing were covered by the OTKA special grant #P18941. Without this, I could not even think of writing the book, and I would like to express my sincerest thanks for support. Most of the English translation was prepared during my fellowship at the Institute of Advanced Study, Collegium Budapest. I am most grateful to my host for research facilities and the unique working environment.

Important note

Despite all effort, this book, as most other books, is not error free. The author welcomes any corrections, suggestions and questions sent to his email address,

podani@ludens.elte.hu.

An updated list of potential errors, a summary of problems raised by the readers, and additions to the imaginary dialogues will be found at the web site

http://ramet.elte.hu/~podani.