BIOSTATISTICS

G. Welzl, T. Faus-Kessler, H. Scherb and **K. Voigt**
Institute of Biomathematics and Biometry, Neuherberg, Germany

Keywords: biostatistics, ecological-environmental structure, spatio-temporal structure, visualization techniques, clustering, ordination, geostatistics, point processes, time series

Contents

Summary

Biostatistical methods are essential in environmental sciences and ecology. In these fields, complex processes and structures, usually indexed by space and time, have to be analyzed. A great variety of statistical methods are available to investigate the environmental–ecological and the spatio-temporal structure of empirical data. In addition to a general view, more detailed methodological information and examples are given, elaborated in the working group “Biostatistics” of the Institute of Biomathematics and Biometry at the GSF, National Research Center for Environment and Health. We believe that all biostatistical groups have to influence the interpretation of environmental data and ecological findings actively and that they are responsible for

selecting sound methods and for correctly applying the established techniques. There are some crucial problems of environmental contamination and alteration, which urgently need new statistical methods. This paper may provide some advice and guidance in the analysis, interpretation and integration of environmental data.

1. Introduction—Overview

1.1. Quantitative Methods—Statistics–Biostatistics

Life is a highly complex set of interactive systems with a jeopardized future on our planet. At present, the use of land, water, air, energy, and other resources has turned into a matter of grave concern and needs careful observation. It is the objective of environmental sciences and ecology to analyze these systems theoretically and quantitatively. In recent years, some attempts have been made towards increasing the part quantitative methods play. It makes sense to conduct quantitative studies of the environmental system to describe the condition of soil, water, and atmosphere. Statistical or numerical methods are used to improve the precision of both design and analysis.

Quantitative methods in ecological and environmental sciences are based on two components: the first component has biological and technological features, i.e., good sampling devices for accurate and precise measurements. The second component concerns methodological aspects, i.e., good experimental design as basis for an adequate statistical or numerical analysis. The term experimental needs extension: in addition to manipulative experiments, which involve assigning some treatment to the experimental unit, mensurative experiments have to be taken into account.

There are several general principles and types of experimental design, described in many excellent textbooks. Methods for analyzing data are discussed both in statistical and in environmental–ecological texts. Complex ecological and environmental studies require special methods for measuring, collecting, registering, editing, storing, and analyzing the data. Due to the complexity involved in formulating the problem, many variables must be considered and various dependencies are present. A specialty of such data is their temporal and spatial variability.

Often it is necessary to analyze large data sets entailing special requirements for statistical software and techniques of visualization. The heterogeneous origin of data calls for methods that allow consideration of data about data (meta-data). In ecological literature the components deterministic modeling, statistics, and numerical methods are integrated under the header of quantitative methods. Whereas statistics deals with classical descriptive and inferential statistics (e.g., test of significance), numerical methods are derived from the fields of discrete mathematics, mathematical physics, parametric and nonparametric statistics, numerical taxonomy, psychometry, econometry, and others.

Multidimensional statistical methods of particular relevance are systematically combined with non-statistical techniques (e.g., methods from discrete mathematics). Another criterion for differentiation is the relationship between the natural conditions and the outcome of an observation. In deterministic modeling, there is only one possible result according to the deterministic relationship. A common attribute of numerical methods is the random relationship between the natural conditions and the outcome of an observation. As a consequence, many results with predictable probability function are possible. If the results of an outcome depend on the specific strategies of the organisms and the environment, then the methods adequate for data analysis come from game theory. Many possible but unpredictable results induce special forms of uncertainties; here it is best to apply methods developed in chaos theory.

Statistics as a discipline in general, does not have a good reputation. “Lies, damned lies, statistics,” or “You can prove whatever you like with statistics,” are only two examples of popular opinion. Many people meet statistics with suspicion. They distrust statistics and imagine statisticians to assiduously collect irrelevant facts and use mysterious mathematics to manipulate society. There are various reasons for this dilemma. First, statisticians must share some of the blame.

In many countries, statistics has been developed in the shadow of pure mathematics. Therefore, often disguised in esoteric language, mathematics may obscure the natural phenomena that scientists are trying to elucidate. Also, statisticians are involved in an everlasting internal debate about foundations, Bayesian versus frequentists, which further weakens their authority. Second, many people are uninterested in natural sciences, and most of those interested have the idea that every event has a simple, direct cause. In public opinion there is great discomfort against a science based on uncertainty instead of certainty and on complexity instead of simplicity. But good statistical thinking is not that much more than good common sense.

The object of statistics is information; the objective is the understanding of information contained in data. A statistical problem is characterized by variability and uncertainty. A modern definition of statistics is: “Statistics is a straightforward discipline designed to amplify the power of common sense in the discernment of order amid complexity.” It is this vision of statistics that reveals the usefulness of statistical methods in environmental and ecological sciences. Statistics is concerned with understanding the real world through information by experiment and observation.

This striving is hampered by the fact that many peoples’ view of statistics is too narrow. Bartholomew broadened this view by defining four types of statistics. Type I is related to the effective collection and presentation of data. Statistics of type II is what is taught in lectures in mathematical statistics. Random samples, drawn from well-defined populations, are the basis for inference statistics. Whereas the school of inference, Bayesian or frequentists, is not so important as some of their devotees imagine, the difficulty with statistics of type II is that it rarely fulfills the conditions to be met when the ecological and environmental problems of our time are analyzed.

Observational studies of large and complex systems are typical of ecology and environmental sciences. For their investigation, statistics of type III is necessary. Here the data are primarily obtained by observation, and any realistic model will be highly complex, involving many parameters and few or no replications. Relationships that also consider space and time are of principal interest. Statistical methods in this field include time series, Markov chains, point processes, stochastic geometry, geostatistics, etc. The access to the territory of politics and management, of science and philosophy gives rise to statistics of type IV. In this field, many arguments tend to be from non-numerical information and personal judgments, but some of the arguments will be statistical decisions. Statistics can and should contribute to this level of decision making.

In the following article, we integrate all statistical procedures in the wide sense of Bartholomew (especially statistical methods of type III) and include numerical methods derived from the fields of mathematical physics, information theory, discrete mathematics, chemometry, econometry, and others under the header of "Biostatistics."

The challenging aspect is to combine various methods to an iterative procedure and interlock effective methods of different fields. We reason that numerical and statistical approaches can never dispense ecologists from reflection on observations. But we believe that the complex multivariate structures of data in ecology and environmental science can be profitably analyzed by numerical and statistical methods in order to find a clear and relatively simple description of scientific relations.

-
-
-

Bibliography

Bailey T. C., Gatrell A. C. (1995). *Interactive Spatial Data Analysis*, pp. Essex: Longman Scientific and Technical. [Both the analysis of point patterns and the analysis of spatially continuous data are treated. The book provides an introduction to spatial data analysis and gives a survey of the subject.]

Bartholomew D. J. (1995). What is Statistics? *Journal of Royal Statistical Society A* **158**, 1–20. [This lecture of the President of the Royal Statistical Society describes the role of statistics in our time. Definitions of four types of statistics broaden the general concept of statistics.]

Cressie N. A. C. (1991). *Statistics for Spatial Data*, pp. 900. New York: John Wiley and Sons. [Theory and practice for all types of spatial data: geostatistical data, lattice data and spatial patterns are presented.]

Chilès J-P., Delfiner P. (1999). *Geostatistics, Modeling Spatial Uncertainty*, pp. 695. New York: John Wiley and Sons. [Based on the theory of geostatistics developed by Georges Matheron, this book is a modern view of the subject where theory and practice are reconciled. Examples are discussed in full detail.]

Diggle P. J. (1990). *Time Series: A Biostatistical Introduction,* pp. 257. New York: Oxford University Press. [The book presents a very comprehensibly introduction to the classical methods.]

Einax J., Zwanziger H. W. and Geiss S. (1997). *Chemometrics in Environmental Analysis*, pp. 384. Weinheim: VCH, Wiley company. [Analysis of environmental data with methods from statistics and chemometrics are discussed.]

Kaufman L., Rousseeuw P. J. (1990). *Finding Groups in Data*, pp. 342. New York: John Wiley and Sons. [Cluster analysis is interpreted as the art of finding groups in data. This book describes a lot of partitioning and hierarchical methods realized in S-PLUS.]

Legendre P.and Legendre L. (1998). *Numerical Ecology*, pp. 853. Amsterdam: Elsevier. [Introducing the book with comments on the part data analysis plays in ecology the authors give a comprehensive survey of biostatistical methods. Main topics are concerned with the analysis of abundance data. The definition of the term "numerical methods" corresponds to definition of statistics type III in Bartholomew to a large extent].

Biographical Sketches

The authors all work as scientists in the working group 'biostatistics' at the Institute of Biomathematics and Biometry of the National Research Center for Environment and Health (GSF). Gerhard Welzl studied mathematics at the Technical University of Munich and is now head of the working group. Theresa Faus-Keßler studied mathematics and physics in Regensburg and Munich and is engaged in statistical analysis of medical and environmental data. Hagen Scherb studied mathematics and physics in Saarbrücken and is occupied as senior scientist in mathematics and statistics. Kristina Voigt graduated in food chemistry at the Technical University of Berlin and is concerned with environmental statistics and chemometrics.