ENVIRONMENTAL DATA AND STATISTICS

S. Heiler
Department of Mathematics and Statistics, University of Konstanz, Germany

Keywords: environmental data, multivariate data, time series analysis, geostatistics, Space-time models.

Contents

Summary

This review is a contribution on applications of statistical procedures to environmental data. Since this has become a very vast field, a comprehensive presentation seems to be impossible. Therefore, only some important areas were chosen, starting with modern tools for graphical presentation of univariate and multivariate environmental data sets. Understanding correlation is supported by an example of heavy metal content of sediments in waters.

Linear and non-linear nonparametric regression is an interesting and important technique in all kinds of statistical applications. Principal components- and factor analysis are illustrated with the same heavy metal data set. Cluster analysis and analysis of variance conclude the chapter on multivariate procedures. Time series analysis has become an important area also for environmental data. Trend- and seasonal estimation are exemplified with monthly measurements of phosphorus contents in the lake Constance. Exponential smoothing as a simple forecasting device is applied to a similar data set and ARMA modelling is illustrated with monthly measurements of sodium content of Lake Constance. The contribution finishes with a glance at geostatistics.

1. Introduction

Statistics is an indispensable means of environmental research. It is used to analyse and to interpret the increasing flood of vast data from environmental areas, which are often of heterogeneous nature and show high variability. Many important results and statements concerning our environment are based on statistical investigations, such as changes of the ozone layer, climate changes etc.

But also less spectacular results about the influence of various human activities on certain environmental parameters which are not obvious at first glance and superimposed by considerable random variations are important findings of statistical analysis. In official statistics environmental monitoring has become a serious tool for political consulting. For environmental research also statistical methods for the design and analysis of experiments play an important role.

A specific scientific discipline of environmental statistics does not exist. The whole statistical methodology may be used in investigating environmental questions. But it is also true that certain statistical techniques are of particular importance because of the special nature of the question.

Particularly, procedures where the position of the data in time and / or space is taken into account are of importance, such as methods of time series analysis, the theory of point processes and geostatistics. For the investigation of phenomena that consist of several simultaneously interacting variables procedures of multivariate statistical analysis are of interest.

For the analysis of exceptional natural phenomena extreme values theory may be helpful and sampling theory is useful in answering questions like “on which spots and

how frequently should certain environmental parameters be measured in order to get a satisfactory picture of the general situation".

2. Environmental Data

In environmental research quite often the measurement, collection, storage, processing and analysis of data are not carried out by the same institution. Only in exceptional cases there is one person who knows about all the details of collection and analysis of the data, who knows about the scientific environmental background and at the same time also about the mathematical procedure and algorithms for the statistical evaluation and the presentation of the results.

There is a great complexity in environmental statistics. Heterogeneous data from different sources and collection principles are analysed simultaneously. There are dependencies between the measured quantities that usually do not follow fixed laws or rules, but reveal random variability. Besides this variability in the nature of the environmental problem there is variability in time and space. And measurements in time are not repeatable.

From time to time disturbances in the measurement device lead to a loss of some data and lacunas in the data series. Sometimes one realises during the exploration process that measurements at intermediate points would be helpful. Therefore a learning process goes along with the statistical analysis. Also different time scales have to be taken into account. Some measurements are taken half-hourly, others only daily or once a month.

Another problem lies in the abundance of data in environmental statistics. Although modern computer technique can cope with this, problems of compatibility, standardisation and data harmonisation arise as obstacles. Storing of data is often connected with coding. But the description of the coding principles is sometimes insufficient and this may make it difficult to join data from different sources together.

If you have not done the measurements yourself then you may lack information about the measurement process itself. This may start with lack of information about the definitions of what is being measured (exact objective, temporal and spatial description), the accuracy and the liability of the data. Closed information systems where all stages of data collection follow a unified principle are rather the exception in environmental statistics.

2.1 Data structures

The basic structure of an environmental data set is a matrix, where usually the rows correspond to the individual objects (measurements, time units or measuring spots) and the columns contain the series of readings for the corresponding variable. The units in the columns may be logical characters (true = 1 or false = 0), ordered or unordered categories, integers (count data) or reals (measurements) – they may also contain a time information or a coding of the measurement spots. The coding of missing data and of censored data (for extreme values) has to be fixed. Of course some describing or classifying text may be contained as well in the rows.

In environmental statistics most data are the result of a measurement process, where the measuring instruments have a certain degree of precision and a limited range of scale. Both have to be taken into account in the analysis of the data, as well as their liability. Unintentional failure of measurement instruments and disturbances of a transmission channel may lead to false or missing values. If the deviations are big enough, the corresponding data are detected as outliers. There are statistical procedures to perform this. Supplementary aetiology may even lead to a correction of the values.

Intentional modification of data cannot be excluded if interests of persons involved are touched. For example, unwelcome measurement results are discarded or transformed. Also a convenient choice of measurement times and spots may influence the result. In environmental data threshold values play an important role. Short exceeding or large averages express completely different things and may be chosen according to the emittents discretion.

Often it turns out only in the course of the research process which data are relevant for the question at hand. Consequences can be drawn from this only for future research. Preliminary examinations may be helpful to give evidence about which data are of importance.

Data protection rules may also play a role in environmental statistics although one thinks that observance of secrecy should be kept on a low level for environmental data. But economic and personality protection rules have to be allowed for. In medical research the personality rights of the examined patients have to be observed.

The choice of the time scale is also of importance. Often there is no choice and one is obliged to work with a given time scale. The intervals are given by social convention or technical circumstances. Sometimes it turns out only later which scale would have been reasonable or optimal for the project at hand.

Usually the time needed for an environmental analysis is not limited. But particularly in environmental monitoring with the aim of defence of danger or damage for the environment the measurement results have to be transformed into warnings and announcements without delay. Of course statistical procedures have to take this into account. They have to work more or less automatically with human interactions only in exceptional cases.

-
-
-

Bibliography

Armstrong M., ed. (1989). Geostatistics, 1, 2. Dordrecht: Kluwer Academic Publishers.

Barnett V. and Turkman F.eds. (1993). Statistics for the Environment. Chichester: J. Wiley and Sons. [Proceedings of a conference containing contributions to environmental monitoring and sampling, measurement levels, consequences of pollution, climatological and meteorological issues, to water resources, dynamics of fish population and forestry.]

Bergmann E., Bender J., and. Weigel H.-J (1997). Relative Empfindlichkeit und Reproduktionsverhalten von Wildpflanzen gegenüber Ozonstreß. Umweltbundesamt, Berlin. [Research report on the impact of ozone on plant growth.]

Box G. E. P. and Jenkins G. M. (1976). Time Series Analysis, Forecasting and Control. Oakland: Holden-Day. [Classical time series book introducing ARMA- and ARIMA models.]

Brockwell P. J. and Davis R. A. (1996). Time Series: Theory and Methods. Berlin, Heidelberg, New York: Springer-Verlag. [Time series book for practitioners with many examples which can be carried out by the reader using an enclosed diskette.]

Chilès J.-P. and Delfiner P. (1999). Geostatistics. Modeling Spatial Uncertainty. New York: J. Wiley and Sons. [This book is based on the theory of geostatistics by G. Matheron. It presents a modern unified view of the subject and includes detailed examples.]

Cressie N. (1993). Spatial Data Analysis. New York: J. Wiley and Sons. [Book containing theory and practice for all types of spatial data.]

Draper N. R. and Smith H. (1981). Applied Regression Analysis. New York: J. Wiley and Sons. [Comprehensive textbook on linear and nonlinear regression with many examples using matrix approach.]

Fahrmeir L., Hamerle A., and Tutz G. eds. (1996). Multivariate Statistische Verfahren. Berlin: Walter de Gruyter and Co. [Comprehensive description of many multivariate procedures with examples, oriented at practitioners.]

Härdle W. (1990). Applied Nonparametric Regression. Cambridge: Cambridge University Press. [Applied textbook discussing kernel-, k-nearest neighbor- and spline smoothing.]

Heiler S. and Feng Y. (2000). Data-driven Decomposition of Seasonal Time Series. Journal of Statistical Planning and Inference 91, 351-363. [A nonparametric procedure of analysing seasonal time series is discussed.]

Heiler S. and Michels P. (1994). Deskriptive und Explorative Datenalanlyse. München: Oldenbourg Verlag. [Textbook on descriptive and exploratory data analysis.]

Hoaglin D. C., Mosteller F., and Tukey I. W. (1983). Understanding Robust and Exploratory Data Analysis. New York: J. Wiley and Sons. [An introduction to methods of modern exploratory data analysis.]

Jobson J. D. (1991, 1992). Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design. Berlin, Heidelberg, New York: Springer Verlag. Volume II: Categorial and Multivariate Methods. Berlin, Heidelberg, New York: Springer Verlag. [Comprehensive compendium of applied multivariate data analysis including regression, experimental design, ANOVA, MANOVA, contingency tables, principal components- and factor analysis, cluster analysis, multidimensional scaling and correspondence analysis. Much emphasis is put on the use of statistical computing packages and the interpretation of numerous real data examples.]

Müller H. (2000). Limnologischer Zustand des Bodensees Nr. 26. Jber.Int.Gewässerschutzkomm.Bodensee: Limnol.Zust.Bodensee 26. [Annual report on the limnological conditions of Lake Constance.]

Patil G. P. and Rao C. R. eds. (1994). Handbook of Statistics 12. Environmental Statistics. Amsterdam: North Holland. [An overview of statistical issues related to environmental problems: environmetrics, environmental monitoring and sampling, statistical ecology, environmental biometrics, systems modeling and analysis, compartment models, spatial statistics and statistics in environmental regulation.]

Pena D., Tiao G. C. and Tsay R. S. eds. (2001). A Course in Time Series Analysis. New York: J. Wiley and Sons. [Textbook on modern time series analysis oriented at practitioners.]

Polasek W. (1988). Explorative Datenanalyse. Berlin, Heidelberg, New York: Springer-Verlag. [Textbook giving an overview on descriptive and exploratory data analysis.]

Stoyan D., Stoyan H. and Jansen U. (1997). Umweltstatistik. Stuttgart, Leipzig: B.G. Teubner Verlagsgesellschaft. [An excellent textbook covering a broad area of environmental statistics with the help of many illustrative examples.]

Tukey I. W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley. [This is one of the classical contributions to exploratory data analysis, looking at data from a non-classical view point, with many meaningful examples.]

Vellemann P. F. and Hoaglin D. C. (1981). Applications, Basics and Computing of Exploratory Data Analysis. North Situate, MA: Duxburry Press.

Wackernagel H. (1997). Multivariate Geostatistics. Berlin, Heidelberg, New York: Springer-Verlag. [Textbook with many details in multivariate geostatistical problems.]

Webster R. and Oliver M. (1997). Geostatistics for Environmental Scientists. Chichester: J. Wiley and Sons. [Comprehensive textbook.]

Biographical Sketch

S. Heiler was born in 1938 in Neuravensburg/Wangen im Allgäu, Germany. He studied economics at the universities of Tübingen, Hamburg and Munich from 1959–1963 (Ph.D. in economics in 1966). In 1971, he habilitated in statistics and econometrics at Berlin TU. From 1972–1987 he was professor for Statistics at the Department of Statistics, Dortmund University and after at Konstanz University (from 1987 to 1998 Faculty of Economics and Statistics, since 1993 co-opted by Faculty of Mathematics and Computer Science; since 1998 he has been member of the Departments of Economics and of Mathematics and Statistics). *Memberships:* Member of the International Statistical Institute (enrolment 1979), president of the German Statistical Society 1988-1992, vice-president 1992-1996, founding member and member of the Advisory Board of ECAS (European Courses of Advanced Statistics), an organisation of Statistical Societies in Europe. President of ECAS 1993-1997, since 1997 Vice-President, member of a Scientific Advisory Board at the States Secretary of Environmental Affairs. International academic experience through several research and teaching sojourns at the Australian National University, Canberra, the Laboratoire d'Informatique et de Mathématiques Appliquées de l'Universitée Scientifique et Médicale de Grenoble, the Department of Probability and Statistics at the University of Sheffield, the Department of Mathematics at Charles University in Prague, Departamento de las Matematicas y de la Estadística/Universidad Nacional de Colombia, Bogota, and at the Department of Applied Mathematics of the Jiao Tong University, Shanghai.