Environmental Statistics in the Information Age

Sylvia R. Esterby

Abstract

To discuss the topic of environmental statistics in the information age, the two parts of the topic need to be considered separately. The meaning of the term Information Age and the implications for statistics, and more specifically environmental statistics and environmetrics, will be explored. In the course of the discussion, the role of computational capabilities and the collaborative aspects with other disciplines, particularly informatics, will be considered.

1. Introduction

Rapid and far-reaching dissemination of verbal and visual (text and graphical) material and other components of the computer technology that has made this possible are the aspects of the Information Age that will be considered. The implications of the dissemination and of the enabling technological capabilities affect us in our daily lives and in our professional lives. In turn, there are a number of implications for the statistical sciences. First, there are the challenges and opportunities associated with the technological advances and the commitment to the provision of information, and secondly there is the issue of defensible reporting to the public and the scientific community at large.

Statistics has a role at two general levels: 1) the scientific or primary level, and 2) provision of information. These can be thought of as contributing to answering the question and then conveying the answer to parties outside our discipline or our interdisciplinary team.

It will be useful to clarify what is meant by information herein. Clearly there is a difference between data and information. In an article on environmental information systems, Tochtermann (2002) made the distinction between primary data, and aggregated and analyzed data. The statistician tends to think of primary data as including not only the numbers but also the supporting documentation that allows one to analyze the numbers. For clarity, let us think about information in the context

1 Department of Mathematics & Statistics, Okanagan University College, 3333 College Way, Kelowna, BC Canada V1V 1V7. Tel. +1 (250) 762 5445 local 7536, Fax: +1 (250) 470 6004, email: sresterby@ouc.bc.ca
of reporting to the public on environmental quality. Information about environmental quality might then be a quantitative statement about status or change of an indicator of quality, or graphical display from which a quantitative statement could be made, and would include a measure of the level of uncertainty, all of this set within the scenario of the environmental system.

Interconnections between the stages of collecting data, modeling and analysis, and the drawing of conclusions is shown schematically below. Information provided to a general audience would then be framed on the basis of the last column of boxes together within the context of the environmental system. These topics follow in the next two sections. The approach taken here applies more to environmental monitoring and assessment of ambient conditions than to regulation and compliance, although many of the considerations apply to the latter topics.

![Fig. 1: Interconnections between the design, modelling and reporting stages](image)

2. Environmental statistics and the primary role

Environmetrics has been defined as the “discipline that focuses on the development and use of quantitative methods to understand and solve environmental problems” (El-Shaarawi/Hunter 2002). The objectives of The International Environmetrics Society further define the activities as fostering the development and use of statistical and other quantitative methods in the environmental sciences, environmental engineering and environmental monitoring and protection. The
objectives also include the promotion of collaboration and communication, both interdisciplinary and between researchers and practitioners (http://www.nrcse.washington.edu/ties/). The role of environmental statistics sits within this broad scope. The aspects of study design and of modeling and analysis will be considered here.

Models are an inherent part of all environmental assessments, even if they are not elaborated explicitly. Studies are often initiated in response to a general statement about the health of an environmental system. To plan the study, many explicit assumptions and decisions must be made such as: what is meant by health or quality, what indicator variables are to be used, the appropriate time and space scales, measurement instrumentation, and data management methods. Thus a model of the system is specified and statistical design considerations can be taken into account in the study plan. Examples of active areas of research that are addressing issues in the design of environmental studies are grid-based sampling designs (Stevens 1997), optimal spatial designs (Zimmerman 2002) and temporal sampling (JABES 1999).

The crucial subsequent step is to address the original question, now in a more explicit form, using the data that have been collected and a method that accounts for the features of the design. This sounds transparent, yet with the complexities of many data sets, it may actually be a very difficult task. Examples of areas of statistical research that are being pursued to help achieve this goal are hierarchical Bayesian methods (Berliner 2002) and space-time covariance models including physical models (Gneiting/Schlather 2002).

Advances in computational capabilities have dramatically advanced statistical methodology largely through the ability to simulate and iterate where large dimensions and/or a large number of repetitions are needed. The introduction to the monograph “Statistics in the 21st Century” (Raftery/Tanner/Wells 2001) states that the computer revolution has transformed statistics, and cites early software such as GLIM (specialized package for generalized linear models, McCullagh/Nelder 1989) in the 1970s, the bootstrap (Efron/Gong 1983) and software such as S in the 1980s, and the Bayesian revolution and MCMC (Kass/et.al 1998) of the 1990s. Along with this, there have been considerable advances in the statistical modeling of environmental systems and in the accompanying inferences drawn about the status and change of environmental quality. The additional capabilities that have made this possible include the ability to process very large data sets, new software that facilitate modeling of data measured on different scales, GIS and mapping.

Complex modeling of environmental systems would be expected to benefit from collaboration among specialists such as: 1) the scientist whose primary area of study is a particular type of environmental system (such as the atmosphere or watersheds), 2) the analyst/modeller which includes the statistician, and 3) the information-system specialist. Spatial-temporal statistical modeling provides an example wherein there are mutual benefits. The statistician requires adequate data bases and supplementary documentation, software to facilitates modeling, and the geographic reference and
the visual displays contributed by GIS and mapping. The models must have a basis in the knowledge that already exists about the physical system, ie the environmental component being studied. In turn, the design of the data collection program can benefit from the ideas of partitioning variability and randomness, inherent in statistical modeling. Design of data bases, the means of linking platforms and efficient software require understanding of the environmental system at least in terms of the way in which data are collected and the ways in which data are used in modeling and analysis. Simply stated, each group knows the needs of the other groups and some basics in the knowledge base of the other groups.

The current directions of making data from many sectors available on the web and of web-based applications for modelling and prediction from environmental data sets, although clearly innovative and promising, pose many challenges as well. The comments of Stockwell/et al. (1999) regarding biodiversity informatics and ecological computing apply more generally. There are several issues that are often not emphasized, but are issues that statisticians and environmental scientists have experience with in simpler data sets. These include the need to ensure the quality of the data that is made available, to provide sufficient documentation to allow compatibility to be assessed before pooling and to raise awareness of these issues among potential users. Relevant to the groups of potential users, there is the need to make both primary data available for modeling activities of researchers as well as information suitable for the science community at large, decision makers and the general public, sometimes called multi-stage access.

3. Provision of information

There are two sides to this theme. The idea of providing defensible information can be taken to mean that we must convey the current state of knowledge as accurately as possible in order to maintain the integrity of science. Coupled with this is the difficulty of conveying unfamiliar ideas in an understandable and accurate way. Accurate statements about partial knowledge seem logically to include a measure of our level of confidence. The degree of difficulty involved in presenting information depends upon the level of comfort our audience has with the idea of uncertainty. Therein lies the other aspect, that is quantitative literacy and our role in education.

Maintaining integrity will also be better served by information for non-specialists that includes names of contacts or technical references as footnotes or bibliography. This enables the reader to substantiate the information to his/her level of understanding and also helps to educate readers to expect documented evidence in addition to well-written text and illustrative figures. Fact sheets published by Environment Canada illustrate two extremes: 1) a fact sheet on smog (http://www.msc.ec.gc.ca/ed/factsheets/smog/index_e.cfm/) is easily understood but does not substantiate claims, and 2) a fact sheet on environmental implications of the
automobile contains more facts and provides the sources for the numbers quoted (http://www.ec.gc.ca/soer-ree/English/products/factsheets/93-1.cfm). These two examples do not address the issue of level of confidence of estimates.

One challenge to the statistical sciences that results from the extensive availability of statistical software is to try to bring an appreciation for the applicability of the methods and interpretation of results to a level comparable to the ability to compute. Since a corresponding statement can be made about empirical modeling by statisticians, without adequate attention to the processes operating in the system, a case can again be made for collaboration.

Bibliography