Environmental Statistics

Peter Guttorp

NRCSE

Technical Report Series

NRCSE-TRS No. 032

The **NRCSE** was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

EPA

Environmental Statistics

Peter Guttorp
National Research Center for Statistics and the Environment
University of Washington
Box 351720
Seattle, WA 98195-1720
peter@stat.washington.edu

ABSTRACT

The field of environmental statistics is one of rapid growth at the moment. Environmental decision making is prevalent in much of the world, and politicians and other decision makers are requesting new tools for understanding the state of the environment. In this vignette some areas of the field are described, and a personal view of important directions is outlined.

Key words: environmental monitoring; design; compositional data; risk assessment; spatial covariance; visualization.

1. Introduction

The field of environmental statistics is relatively young. The term "environmetrics" was apparently introduced in an NSF proposal by Philip Cox in 1971 (Hunter, 1994). During the last decade, the field has achieved some recognition, in that there now are three journals wholly or partially devoted to the field (*Environmetrics* published by the International Environmetrics Society and Wiley; *Ecological and Environmental Statistics* published by Kluwer, and *Journal of Agricultural, Biological and Environmental Statistics* published by the American Statistical Association). The ASA has a section on Statistics and the Environment, and the International Statistical Institute is currently discussing such a section. Volume 12 of the series *Handbook of Statistics* (Patil and Rao, 1994) was devoted to the topic of environmental statistics. Its 28 chapters constitute an interesting overview over the field.

In this vignette I present some of the current areas of research in environmental statistics. This is of course by no means an overview of the field as it stands, rather, it is a list of areas in which I can see the need for, and largely also the tools for, methodological developments.

2. Environmental monitoring

Environmental monitoring design deals mainly with two quite different sorts of design problems: monitoring for trend, where spatial and temporal dependence is of importance, and monitoring for "hot spots", or regions of local high intensity, which is often used for monitoring compliance with pollution regulations. The basic theory of optimal design for spatial random fields is outlined in Ripley (1981, Chapter 3). Among the popular designs are systematic random sampling designs, in which a point is chosen uniformly over the study area, and a regular design (consisting of squares, triangles, or hexagons) is put down starting at the chosen point. When the sample mean is used to estimate the spatial mean of an iso-

tropic random field over a region, the regular sampling plans are most efficient (Matérn, 1960, Chapter 5). The hexagonal design requires fewer sampling sites than a square or triangular one to cover the same area, but does not take into account spatial covariance heterogeneity or temporal nonstationarity. The covariance mapping technique mentioned below in section 3 can be used to deal with spatial heterogeneity, by implementing a spatial design in the transformed space.

Zidek and coworkers (e.g. Caselton et al., 1992, Guttorp et al., 1993) have developed an approach to network design which can deal with heterogeneous random fields. The basic idea is to consider a number of potential monitoring sites, some of which are gauged and some ungauged. In a multivariate normal setting, the design maximizes the amount of information (of Kullback-Leibler type) about the ungauged sites that can be obtained from the gauged sites. This can be particularly useful when trying to redesign a current network, by adding and removing stations.

It is frequently the case that data from a monitoring network will serve more than one purpose. For example, in analyzing trends in tropospheric ozone (Reynolds et al., 1998), the data were collected by the state of Washington to monitor compliance with the Clean Air Act. Consequently, the network was aimed at finding areas of high air pollution, and was changing over time. Statistical methods for analyzing data from a network adaptively designed to find the extremes of a random field need to be developed.

The US Environmental Protection Agency (EPA) started in 1989 an ambitious monitoring program called EMAP (Environmental Monitoring and Assessment Program). This was intended to create a “report card” for the state of the US environment. The basic design of the EMAP study (Overton et al., 1990) is a hexagonal grid, with a random starting

point and a side of 27 km, resulting in 12,600 grid points over the continental United States, of which 25 fall in the Delaware Bay on the US East coast, where EMAP has an ongoing study of benthic invertebrates. The EMAP protocol required revisiting some of the sites on a rotating 3-year basis. The measurements made at each site (three times each summer) included a bottom grab sample of benthic organisms, together with measurements of covariates such as temperature, depth, and salinity.

The basic biological tenet behind this sampling scheme is that environmental insults affect the distribution of organisms, in that pollution tolerant species tend to get a larger proportion of the sample than do pollution sensitive species. In order to deal with species composition data, Aitchison (1986) developed a methodology based on transforming the proportions from the unit simplex to Euclidean space. The proportions are then treated as multivariate normal data in the transformed space. Billheimer (1995) extended this model to space-time data, showed how to estimate parameters using Markov chain Monte Carlo techniques, and how by backtransforming to the simplex the parameters can be given a natural interpretation as proportions (Billheimer et al., 1999). In fact, it is possible to develop an algebra of proportions which allows the statement of common models, such as regression, in terms of proportions. In order to account for counts of species from samples, the proportions are thought of as hidden state variables, and the counts are, e.g., conditionally multinomial, given the (unobserved) proportions. Billheimer et al. (1997) analyze the spatial distribution of EMAP data from Delaware Bay, and Silkey (1998) looks at changes over time.

Another example of compositional data in environmental statistics deals with particulate matter air pollution. Here the chemical analysis determines the distribution of chemical species among the particles. Regression on known pollution profiles enables identi-

fication of sources (Park et al., 1999), but the compositional analysis approach mentioned above may yield additional insight, particularly into seasonal patterns.

3. Spatial prediction

Environmental monitoring data are often used to develop regional summaries of pollution fields. In order to do so, values at unobserved sites have to be predicted. Geostatistical methods, such as kriging, were originally developed to do spatial prediction from a single observation of a network of sites. The main difference in the environmental context is that we generally have a time series of observations. Where ordinary geostatistical methods are forced to make strong assumptions on the spatial covariance structure, such as isotropy, these are not needed , and often not warranted, in the environmental context. Methods are available to study spatially heterogeneous covariance structures (Guttorp and Sampson, 1994). Such methods are needed, e.g., when the covariance structure is determined by hydrology or meteorology.

Our preferred approach is to use the class of covariance functions of the form $c(x, y) = c_0(\| f(x) - f(y) \|)$, where c_0 is an isotropic covariance function and f is a smooth mapping taking the geographic coordinates (x,y) into a different space in which covariances are isotropic (some facts regarding this class of covariance functions can be found in Perrin and Meiring, 1999). The mapping f can be estimated nonparametrically, and current work involves implementing the fitting procedure using Markov chain Monte Carlo (MCMC) techniques.

Given a covariance model, spatial prediction traditionally proceeds as a generalized least squares problem. The standard error of the least squares prediction has three components: one due to the uncertainty about the random field, one due to the uncertainty in

the covariance estimation, and one due to the choice of c_0 and *f*. Traditional geostatistical work ignores the second and third components. The use of MCMC estimation of the covariance function allows direct estimation of the second component. Model uncertainty calculations (e.g., Clyde 1999) can be used to estimate the overall uncertainty by estimating the support of the data for each of several potential covariance models c_0.

4. Risk assessment

The US Environmental Protection Agency is committed to assessing environmental problems using risk analysis. Traditionally, this has been done by putting down a deterministic model of the relationship between level of pollutant and effect. The typical risk function is a differential equation, with parameters that are determined from a variety of sources, such as laboratory experiments, measurements on exposed individuals, or scientific consensus. When the model has to do with human health effects, the basis for the risk function is more often than not experiments on animals, which are then rescaled to provide a risk functions for humans using a fairly arbitrary scaling factor. For further discussion of health effects estimation, see the vignette by Thomas (2000).

Recently much emphasis has been put on uncertainty analysis of these risk assessments. Primarily it has been noted that the values of the parameters in the model are subject to uncertainty, which then propagates through the whole assessment and results in uncertainty about the final risk. The method of probabilistic risk analysis (Cullen and Frey, 1998) assigns what a statistician would call a prior distribution to each of the parameters. Typically the parameters are treated as independent *a priori,* with simple marginal distributions such as uniform or normal. The analysis is done by simulating values from the prior distributions and summarized by producing simulated confidence intervals for quantiles from the resulting risk distribution. Current work aims at assessing the uncertainty more accurately by looking

at the entire model uncertainty (e.g., Givens et al., 1995, Poole and Raftery, 1999). This includes, in addition to the uncertainty of the parameters mentioned above, uncertainty of the data used to fit and/or assess the model, and uncertainty of the model itself.

5. Environmental standards

The detailed understanding of the health effects of a pollutant is one of the tools needed for setting scientifically valid standards for environmental compliance. As an example, the US standard for ozone requires that all sites in a region have an expected number of annual maximum daily 1-hour exceedances of 120 ppb of not more than one. Such a standard is not enforceable, since the expected number of exceedances is not directly measurable, and measurements cannot be taken everywhere in the region. Rather, it describes an ideal of campliance, and may be termed an ideal standard. The standard is implemented by requiring that each site in an approved monitoring network is to have no more than 3 exceedances in 3 years. In effect, this rule applies the law of large numbers to n=3.

The concept of statistically realizable ideal standards was introduced by Barnett and O'Hagan (1997). Their idea is to combine an ideal standard with a statistically based rule of implementation. A simple approach to the problem of setting scientifically defensible environmental standards uses very traditional statistical tools, namely the Neyman-Pearson approach to hypothesis testing. The basic null hypothesis to be tested is that the region is in violation of the regulation, i.e., in the ozone case that the expected number of exceedances is more than one per year. Type I errors are more serious, since they indicate unacceptable health risks to the population, while type II errors can have serious consequences for the state environmental administrators in having to develop control strategies that are not strictly speaking needed. When viewing the EPA regulations from this point of view, they entail type I error probabilities that would be viewed as unacceptable by statisticians (Carbonez et

al. 1999, Cox et al., 1999). In addition, a statistical approach to testing the basic null hypothesis would use test statistics different from the number of exceedances.

In air quality data, measurements are generally made on multiple pollutants. Standards are, however, set on individual pollutants (in the United States these are the *criteria pollutants*, carbon monoxide, ozone, particulate matter, sulfur dioxide, and nitrous oxides). It is an open problem how to set multivariate standards, taking into account the joint health effect of several correlated primary and secondary pollutants.

6. Graphical methods

An area of considerable importance in all of modern statistics is the management, display and analysis of massive data sets. Land use data from satellite-based sensors, automated air quality sensors, continuous water flow meters are among a variety of new measurement devices producing vast amounts of data. We are lacking tools for displaying spatially expressed data with uncertainty measures (see however Lindgren and Rychlik, 1995 and Polfeldt, 1999 for two approaches). Recent advances in three-dimensional visualization (virtual reality) allows a viewer to immerse herself in spatially expressed multivariate data (Cook et al., 1998).

As in all visualization of multivariate data, the tools of linked plots and brushing are extremely useful. There are promising developments in multi-platform graphical systems design using Java-based tools (e.g., the ORCA system, Sutherland et al. 1999). In particular, views of projections of multiple multivariate time series can yield valuable insights into the temporal structure of values that are multivariate outliers originating in a particular temporal part of the data, something that may not be visible in a rotating scatter cloud, and even less in a bivariate scatter plot matrix.

7. The future of environmental statistics

Many of the important environmental problems directly involve multi-dimensional, spatially heterogeneous, and temporally non-stationary random fields. My personal belief is that the development of statistical research tools for classes of such processes may prove to be the most useful development in the field of environmental statistics. The multivariate aspect, in particular, is very important, in that there are few symmetries in space and time that can be used in setting up models for realistic situations. As an example, if we are studying the joint distribution of SO_2 and SO_4 during situations of similar meteorology, we will find different space-time correlations for positive and negative time lags, since most of the SO_4 is produced from SO_2 emissions. As pointed out earlier, tools for looking at the joint behavior of several pollutants and for developing control strategies for their behavior are currently the focus of intensive research.

This vignette has focused on examples from the air quality arena. There are equally important, and often more complex, issues in water quality, and more generally in ecological assessment of natural resources. In the long run, a battery of tools for describing, analyzing, and controlling the state of ecological systems must be developed. There are significant challenges ahead for environmental statisticians.

Acknowledgments:

The assistance of numerous colleagues in working groups at the National Research Center for Statistics and the Environment is gratefully acknowledged. Although the research described in this article had partial support from the U. S. Environmental Protection Agency under cooperative agreement CR 825173-01-0 with the University of Washington, it has not

been subjected to the Agency's required peer and policy review, and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

References:

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, London: Chapman & Hall.

Barnett, V. and O'Hagan, A. (1997), *Setting Environmental Standards*, London: Chapman & Hall.

Billheimer, D. (1995), *Statistical Analysis Of Biological Monitoring Data: State-Space Models For Species Composition,*. Ph.D. dissertation, Department of Statistics, University of Washington.

Billheimer, D. , Cardoso, T., Freeman, E., Guttorp, P., Ko, H., and Silkey, M. (1997), "Natural variability of benthic species composition in the Delaware Bay", *Environmental and. Ecological Statistics,* 4, 95-115.

Billheimer, D., Guttorp, P. and Fagan, W. F. (1997), "Statistical Analysis and Interpretation of Discrete Compositional Data", NRCSE Technical Report Series, 11, University of Washington. Available at URL http://www.nrcse.washington.edu/research/reports/papers/trs11_interp/trs11_interp.pdf

Carbonez, A., El-Shaarawi, A. H. and Teugels, J. L. (1999), "Maximum microbiological contaminant levels", *Environmetrics*, 10, 79-86.

Caselton, W. F., Kan, L. and Zidek, J. V. (1992), "Quality data networks that minimize entropy", in *Statistics in the Environmental & Earth Sciences*, A. T. Walden and P. Guttorp (eds.), 10-38, London: Edward Arnold.

Clyde, M. (1999), "Bayesian Model Averaging and Model Search Strategies (with discussion)", in *Bayesian Statistics,* 6, J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith (eds.), 157-185, Oxford: Oxford University Press.

Cook, D., Cruz-Neira, C., Kohlmeyer, B. D., Lechner, U., Lewin, N., Nelson, L., Olsen, A., Pierson, S., and Symanzik, J. (1998), "Exploring Environmental Data in a Highly Immersive Virtual Reality Environment", *Environmental Monitoring and Assessment,* 51, 441-450.

Cox, L. H., Guttorp, P., Sampson, P. D., Caccia, D. C. and Thompson, M. L. (1999), "A Preliminary Statistical Examination of the Effects of Uncertainty and Variability on Environmental Regulatory Criteria for Ozone", in *Environmental Statistics: Analyzing Data for Environmental Policy*, Novartis Foundation Symposium 220, 122-143, Chichester: Wiley.

Cullen, A.C. and Frey, H. C. (1999), *Probabilistic Techniques in Exposure Assessment*, New York: Plenum Press.

Givens, G. H., Raftery A. E., and Zeh, J. E.(1995), "Inference from a deterministic population dynamics model for bowhead whales (with discussion)". *Journal of the American Statistical Association,* 90, 402 - 430.

Guttorp, P., Le, N. D., Sampson, P. D. and Zidek, J. V. (1993), "Using entropy in the redesign of an environmental monitoring network", in *Multivariate Environmental Statistics,* G. P. Patil and C. R. Rao (eds.), 175-202, Amsterdam: North-Holland.

Guttorp, P. and Sampson, P. D. (1994),"Methods for estimating heterogeneous spatial covariance functions with environmental applications", in *Handbook of Statistics, vol. XII: Environmental Statistics*, G.P.Patil and C.R.Rao (eds.) , 661-690, Amsterdam: North-Holland.

Hunter, S. (1994), "Environmetrics: an emerging science" ", in *Handbook of Statistics, vol. XII: Environmental Statistics*, G.P.Patil and C.R.Rao (eds.) , 1-8, Amsterdam: North-Holland.

Lindgren, G. and Rychlik, I. (1995), "How reliable are contour curves? Confidence sets for level contours", *Bernoulli*, 1, 301-319.

Matérn, B. (1960), *Spatial Variation*, Meddelanden fran Statens Skogsforskningsinstitut, 49, vol. 5. Republished in *Lecture Notes in Statistics*, vol. 36, New York: Springer.

Overton, W. S., White, D. and Stevens, D. K. (1990), *Design report for EMAP: Environmental Monitoring and Assessment Program*,. EPA/600/3-91/053, Washington: Environmental Protection Agency.

Park, E. A., Spiegelman, C. H. and Henry, R. C. (1999), "Bilinear estimation of pollution source profiles in receptor models", NRCSE Technical Report Series 19, University of Washington. Available at URL
http://www.nrcse.washington.edu/research/reports/papers/trs19_vertex/trs19_vertex.pdf

Patil, G. P., and Rao, C. R. (1994), *Handbook of Statistics, vol. XII: Environmental Statistics*, Amsterdam: North-Holland.

Perrin, O. and Meiring, W. (1999), "Identifiability for non-stationary spatial structure", to appear, *Journal of Applied Probability.*

Polfeldt, T. (1999), “On the quality of contour lines”, to appear, *Environmetrics.*

Poole, D. and Raftery, A. E. (1999), “Inference for Deterministic Simulation Models: The Bayesian Melding Approach”, Department of Statistics Technical Report 346, University of Washington. Available at URL http://www.stat.washington.edu/tech.reports/tr346.ps

Reynolds, J. H., Das, B., Sampson, P. D. and Guttorp, P. (1998), “Meteorological Adjustment of Western Washington and Northwest Oregon Surface Ozone Observations with Investigation of Trends”, NRCSE Technical Report Series 15, University of Washington. Available at URL http://www.nrcse.washington.edu/research/reports/papers/trs15_doe/trs15_doe.pdf

Ripley, B. D. (1981), *Spatial Statistics*, New York: Wiley.

Silkey, M. (1998), *Evaluation of a model of the benthic macro invertebrate distribution of Delaware Bay, Delaware*, M. Sc. thesis, Graduate program in Quantitative Ecology and Resource Management, University of Washington.

Sutherland, P., Cook, D., Lumley, T. and Rossini, T. (1999), “ORCA-A toolkit for statistical visualization”. Web page at URL http://pyrite.cfas.washington.edu/orca/

Thomas, D. C. (2000), “Some contributions of statistics to environmental epidemiology”, *Journal of the American Statistical Association”*, this volume.