EPA/600/A-96/018

STATISTICAL ISSUES IN ENVIRONMENTAL MONITORING AND ASSESSMENT OF ANTHROPOGENIC POLLUTION

LAWRENCE H. COX
SENIOR MATHEMATICAL STATISTICIAN
OFFICE OF RESEARCH AND DEVELOPMENT
N. PHILLIP ROSS
DIRECTOR
ENVIRONMENTAL STATISTICS AND INFORMATION DIVISION
U.S. ENVIRONMENTAL PROTECTION AGENCY[1]

INTRODUCTION

Environmental data are often collected to assist in assessment of the impact of anthropogenic pollution on the natural environment, to determine the effects of pollution on human and ecological health, for enforcing compliance with environmental regulations and standards, and to assess the state of the environment. Since primary environmental data collection is costly, data sets are often used for multiple purposes. In addition, environmental information is collected by many diverse and independent organizations. This results in a patchwork of spatially and temporally different data sets that profess to measure the same phenomenon, but defy the use of classical statistical approaches for their integration and analysis. This duality leads to a number of statistical issues relating to the monitoring, measurement, use, and analysis of environmental data. Statisticians are being asked to convert the proverbial "sow's ear" into a "silk purse". In this keynote paper to the Third SPRUCE Conference, we address some statistical issues at the interface of environmental monitoring, measurement and regulatory decision making. Some of the areas explored are: environmental monitoring and measurement, environmental indicators, sampling approaches, use of environmental models, use of "encountered" or "found" environmental data, environmental decision making and public policy, and environmental reporting.

STATISTICAL ISSUES IN ENVIRONMENTAL MONITORING AND MEASUREMENT: THE COLLECTION OF PRIMARY ENVIRONMENTAL DATA

Monitoring

The monitoring and collection of environmental data is a costly undertaking. Environmental data collection involves complex measurement processes that often require the development of new technology for the direct or indirect measurement of environmental variables. In addition to instrumentation, additional cost and consideration must be given to

---

[1] The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency. It has been subjected to Agency review and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

the placement of the measurement instruments and the timing of measurements so that the data collected are representative. Statistical issues related to defining the universe to be measured, designing sampling strategies that are cost effective and representative, and integration of additional monitoring information with primary data for assessing the state of the environment are among the problems associated with environmental monitoring.

In addition, new methods and strategies need to be developed to deal with the collection of multi-media information at a single site. Initiatives like the U.S. Environmental Monitoring and Assessment Program are attempts to establish monitoring systems based on sound statistical design and capable of collecting multi-media information on a site-specific basis. Although the collection of a number of variables from a single monitoring site provides a cost effective approach to measuring the state of the environment across a number of dependent variables, it brings with it new statistical problems of data interpretation.

Spatial and Temporal Variability

Spatial and temporal aspects of the environment must be considered when monitoring and measuring anthropogenic pollutants: environmental measures have inherent spatial and temporal characteristics and interrelationships. Many of the assumptions underlying classical statistical time series and geographic data methods such as multivariate normality, i.i.d. observations, replication, and simple random sampling of points in space and time are often invalid or not recoverable (e.g., by logarithmic transformation, weighting or adjustment for autocorrelation or trend) for environmental data sets. This lack of conformance to assumptions underlying classical methods raises a number of interesting issues for statistical application in the spatial and temporal analysis of environmental data.

Ecological monitoring and assessment illustrates these problems. The sampling design for the Environmental Monitoring and Assessment Program is based on selecting monitoring locations probabilistically using a randomly situated hexagonal tessellation of the area of interest. The tessellation can be as fine as needed (i.e., involving smaller and smaller hexagons), provided that the sample hexagons remained i.i.d. Within each sampled hexagon, sampling methods based on less restrictive assumptions (e.g., adaptive or other sequential methods) may be employed. To capture temporal trends, geographic sampling is accompanied by a temporal sampling design analogous to "rotation group" designs familiar to socio-demographic statistics. Sample collection for a single period can take several weeks or months, but typically is confined to a single season. Thus, space is considered to be homogeneous at sufficiently large scales, time is considered homogeneous at sufficiently small scales, and space and time are considered separable.

Different problems are encountered in hazardous waste site characterization. Time is regarded as fixed, and space as inhomogeneous. Simple random sampling schemes (e.g., along a regular grid), though in widespread use, may fail to capture spatial trends, and "hotspot" sampling may overestimate the magnitude or extent of pollution. Methods based on prior knowledge of spatial structure are preferred, but in their current form are often too

complex for the average user. Sequential methods have been shown to be effective and are useful when prior information is unavailable.

One area where simple spatial analysis is being applied is through the use of Geographic Information Systems (GIS). Different types of environmental data can be represented on the same plane. Traditional representation is relatively unsophisticated in that different color schemes or shadings are superimposed on the plane under the visual or intuitive judgement of the analyst. Such methods can provide confirmatory information (e.g., for identifying potential environmental justice) but are not trenchant. The next step (commonly not employed) is the use of statistical methods for defining and comparing relationships and cluster groupings. Such approaches would provide less subjective assessment of relational patterns and interactions of the environment with anthropogenic stress factors. This would enable site assessments based on risk as well as provide a mechanism for stratification of areas for focused environmental monitoring. Combination of the ability to automatically manipulate and visualize spatial data provided by GIS with algorithms and methods of spatial statistics that capture local spatial structure is emerging and should result in statistical software for spatial analysis comparable in power and ease of use to that already available for time series analysis, demographic studies, etc.

ENVIRONMENTAL INDICATORS

In most countries, environmental monitoring systems are put in place to assist in the enforcement of standards aimed at protecting the environment and the health and well being of its inhabitants. Direct measurement of environmental quality is generally not feasible and indirect measures, or indicators, are used to provide some assessment of the state of the environment. Indicators are typically closely tied to environmental measurements, e.g., concentration (parts per billion) of ozone in ambient air or of dissolved oxygen (mg/l) in estuaries. Such values, meaningful to scientists, may communicate little to policy makers or the public, and do not effectively characterize total environmental condition. For these purposes, environmental indices, i.e., (unitless) values based on several different environmental measurements, are preferred. Examples of environmental indices are the Pollution Standard Index (PSI), which reports daily air quality based on separate measured concentrations of five pollutants in ambient air, and "benthic" indices, which combine biological and chemical measurements taken in estuarian sediment. The types and measurements used for indicators are of critical importance: resolution of these questions gives rise to problems in both environmental and statistical science.

Indices present a special challenge to the statistician. The tendency to aggregate a number of disparate measures into a single variable that will tell us how the environment is doing at any point in time and over time is appealing. However, the process of integrating multi-variate measures into an index for the environment is not simple. The types and measurements used for indicators are of critical importance. Environmental structure and function are filled with uncertainties and predicted linkages between observed measures and actual state of environmental health is difficult if not impossible to show. Modern

environmental indicator and index development does not account fully for these uncertainties.

STATISTICAL ISSUES IN ENVIRONMENTAL SAMPLING

The most reliable answers to scientific and policy questions are those based on data of known quality and data collection and estimation or inferential methods of known precision and accuracy. In the environmental arena, lack of conformance of data to i.i.d. and other assumptions raise the need for new statistical design and estimation approaches.

Environmental remediation is an expensive undertaking with significant public health consequences. The development of reliable, statistically powerful spatial sampling designs for site characterization is therefore important. Issues include inference and modelling of spatial structure, number of samples, choosing or balancing among random grid-based designs and sequential sampling methods, assessing and ensuring statistical power of tests, and accounting for uncertainty due to physical properties (e.g., granularity) of samples. Each of these questions is of current research interest. At the other end of the spectrum, current practice is often wrong statistically (e.g., use of t-tests based on faulty assumptions of normality--or failure to be cognizant of such assumptions), and it is incumbent upon statisticians to address these deficiencies and offer useful (ideally, simple) alternatives to practitioners.

Ecological sampling offers opportunities for development and application of sequential, adaptive sampling methods. Most wildlife and fish populations move in groups, often in response to local conditions, and plant life develops and flourishes or decays within local ecosystems. Ecological sampling must take these factors into account in defining units for sampling and measurement and in developing and using statistical sampling strategies for the environment. Adaptive sampling (e.g., oversampling areas where units of interest are found) offers a suite of statistically reliable methods for doing so.

An emerging arena in statistical survey methodology is the design of human (and, potentially, ecological) exposure surveys. Human exposure surveys may be population based or conducted within cohorts (e.g., occupation groups) or regions of interest (e.g., in proximity to hazardous waste sites). In addition to questionnaire data and data from administrative sources, these surveys typically involve taking measurements from subjects' environments and/or from subjects themselves. This forces both per-subject costs and respondent burden to be high, raising serious statistical reliability and response bias issues that statisticians need to solve through the development of new or refined survey methodology. As a paradigm for sampling design, adaptive sampling methods offer the potential to improve both statistical efficiency and cost effectiveness of these surveys.

USE OF ENVIRONMENTAL MODELS

Much environmental decision making scientific study is based on computer models of environmental phenomena. Some of these models are statistical (e.g., regression models that adjust ambient ozone concentration data for local meteorological effects, or that adjust daily

nonaccidental death counts for ambient concentrations of particulate matter) but most are not (e.g., regional models of ozone formation and transport based on atmospheric chemistry such as the USEPA Regional Oxidant Model, and pharmacokinetic models to assess dose-response based on biochemistry). Both types of models pose interesting and important statistical questions and applications.

Regression based statistical models are being more frequently used to identify relationships and account for uncertainty in environmental observations. A classical issue is the overinterpretation of regression results and statistical significance, usually by nonstatisticians: At what point are statistical methods being used to overinterpret data, confusing quantitative artifacts for true relationships, or missing the larger, process-driven picture in favor of isolated numerical findings lacking scientific plausibility?

An important issue in statistical modelling of environmental phenomena is the transportability of model findings and of models themselves. If, as is often the case, several different regression models are developed to assess the effects of particulate matter on mortality in each of several cities, then how can these results be combined statistically to draw more general (and, presumably, more powerful) conclusions? This calls for the development of meta-analytic methods outside the realm of controlled experiments. In addition, is it possible to develop "meta-models" that can be calibrated and transported from one situation (viz., set of local conditions such as climate, presence of co-pollutants, etc.) to another? This would enable the comparison of environmental problems such as mortality due to particulate matter between different cities within a common, representative framework.

Non-statistical models of environmental processes typically have little or no ability to account for uncertainty. Often reliant on numerical solutions to complex systems of differential equations, these models are wholly deterministic and unable to account for sensitivity of outputs to inputs. Sensitivity is often not examined at all, even empirically, nor are models validated to identify and account for systematic biases. Often this is due to the high cost of running models. Statistical designs for cost-effective model validation experiments are needed. Also needed are statistical methods incorporated in the models themselves to estimate uncertainty and the effects of propagation of uncertainty. Designs such as for process optimization would be useful to optimize model performance.

USE OF ENCOUNTERED DATA

One of the most challenging areas for environmental statistics is the development of methods that facilitate reuse of existing data. How do we use information for purposes other than what it was originally collected for? How can one data set be used to validate another? How and when should missing or faulty environmental data be replaced by imputed values? How do we take spatially and temporally disparate data sets that purport to measure the same things and combine them into "synthetic" data sets for use in decision making and regulatory standard setting?

For example, the USEPA and the University of Maryland have been working on the evaluation of the attainment of restoration goals for dissolved oxygen (DO) in the Chesapeake Bay using a statistical method to combine monitoring station and buoy data. Dissolved oxygen is an essential element in maintaining viable conditions for living resources. The Chesapeake Bay Executive Council, comprising representatives from EPA, The Chesapeake Bay Commission, the District of Columbia, and the states of Maryland, Pennsylvania and Virginia, established goals for restoration of the Bay. The standards for dissolved oxygen were set on the basis of extensive laboratory and field research as:

Target DO Concentration	**Time and Location**
DO≥ 1.0 mg/l	At all times, everywhere
1.0 mg/l ≤ DO ≤ 3.0 mg/l	For no longer than 12 hours; interval between excursions at least 48 hours everywhere
Monthly mean DO ≥ 5.0 mg/l	At all times, throughout upper layer waters
DO ≥ 5.0 mg/l	At all times, throughout upper layer, in spawning reaches, spawning rivers, and nursery areas

The restoration standards were time dependent; however, most of the monitoring data being collected for DO on the Bay is done monthly or weekly at fixed sites. During the summer months continuous monitoring of DO (every 15 minutes) was conducted at selected sites during summer months. Continuous monitoring of DO is extremely expensive and cannot be done year round. The challenge to the University of Maryland and EPA statisticians was to develop an approach in which station data (intermittent monthly data) and limited buoy data (continuous) could be combined into a single synthetic data base containing the long term trend properties from the station data and the short term behavior similar to the buoy data. The method of combining used spectral analysis techniques. Work is continuing on this synthesis process to develop a data set that can be used to assess progress towards achieving the restoration goals.

Environmental science and decision making need proper methods for combining environmental data, for several reasons. Direct data collection is time consuming and expensive. It can be prohibitively expensive, meaning that, absent the ability to combine existing or partial data, important phenomena may go unstudied. Often, situations that occur in one place in space and time cannot be reliably replicated elsewhere, requiring the combination of actual data from one place or time with "surrogate" data from another. This raises issues of the "transportation" of data, analogous to that of transportability of models previously discussed. The benefits here will come both in terms of cost savings and increased knowledge and reliability and weight of evidence of conclusions. Related issues include combining administrative records data and monitoring data to assess socioeconomic impacts of pollution and environmental restoration, and using quality assurance data to validate monitoring data and adjust it for embedded systematic errors.

Most environmental data are not design-based, i.e., collected according to a

probabilistic sampling design. For example, lake data might be collected from lakes within selected areas, large lakes, lakes regarded as being in the most degraded environmental condition, the most accessible lakes, or from lakes without specification as to how they were chosen. Such data are known as encountered or "found" data. Environmental scientists use encountered data effectively to study environmental processes. However, their use for environmental assessment is limited due to lack of quantifiable knowledge of selection bias, sampling variability, etc.

Prior to the advent of design-based approaches, environmental data were often modelled statistically using regression, spatial, or time series methods. It was unclear how to combine of such data (e.g., combining lake data between states within a geographic region), and combination was often not attempted. However, we are now in a situation where considerable resources have been expended by society to amass volumes of environmental data, some of which now is design based. Statistical methods are needed to combine encountered and design-based data among themselves and each other. Various statistical approaches are available for these problems, each with its own strengths and weaknesses. Under appropriate circumstances, probability samples can be combined directly. If probability and encountered samples share frame variables, regression can be used to predict sample values for variables observed in one sample but not in the other. Synthetic "pseudo-units" can be formed by statistically matching sample units across two data sets. Perhaps most reliable in general, methods such as dual frame estimation or minimum variance weighting can be used to combine estimates (in lieu of direct combination of data) between two data sets. And, weighted distribution functions can be used to adjust for bias and normalize data for combination.

ISSUES FOR ENVIRONMENTAL STATISTICS IN DECISION MAKING AND PUBLIC POLICY

Public policy requires decision making which incorporates a number of variables ranging from objective measures of the state of the environment to social and political aspects of the policy outcome. The statistical sciences provide the objective measure of the state of the environment; the quantitative bases for public policy decision making. In the public policy arena, decision making uses objective information, but is not necessarily driven by it. An environmental manager is charged with making a decision about the construction of a dam on a major river. The decision needs to be made within a few months. To determine the impact of such a decision on the environment would require the collection and analysis of large amounts of information. This process, if not underway, could require years. The decision needs to be made in three months. The challenge to statisticians is to look for ways to use what is available-- good, bad or indifferent, to the best advantage possible. To provide the decision makers with the best information within the needed time frame.

Environmental risk assessment

Public policy and environmental decision making requires that some form of risk assessment be done to provide a quantitative basis for cost/benefit and decision making.

Indeed, the limited funding for environmental protection leads environmental managers to rely more and more on what are called "comparative risk assessments". Assessment of environmental risk is a multi-disciplinary approach involving information from ecological studies, chemistry, meteorology, statistics, biology, etc.

Current methods for estimating risks and defining safe levels of exposure do not take full advantage of the data and information from the different disciplines. Indeed, at the local community level comparative risk projects give little attention to the use of statistical methods as a means to organize and analyze information provided from the different disciplines. Statistical consideration of methods of sampling, predictive correlations using appropriate stochastic models, and use of multivariate models for assigning risk measurement need to be developed and incorporated into the comparative risk process. Uncertainty in risk analysis must be addressed. The multiple stages in assessing risk give rise to a cascading of uncertainty. However, in most studies on environmental risk the endpoint is presented as a point estimate without any associated uncertainty analysis. Statistical approaches to uncertainty analysis incorporating the cascading effect need to be developed and applied.

REPORTING ON THE STATE OF THE ENVIRONMENT

Environmental managers and policy makers would like to have a crystal ball that summarizes ecosystem status and predicts future states. What ever the immediate practicality of diverse expectations, we need much better approximations of environmental knowledge--environmental indicators. Statisticians must consider the community which will use these indicators. Environmental indicators (like economic indicators) are also useful to a variety of individuals: political officials, their staffs, program planners and assessors, contractors, researchers, environmentalists, educators, market analysts, students and the general public. Most of the audience lacks formal training in statistics or the environmental sciences. This has implication for statisticians in the design and presentation of indicators.

Public Health

The link between human health and the environment has become an important issue. With the occurrences for Love Canal, Times Beach, secondary smoke and deterioration of the stratospheric ozone level, the public has become keenly aware that continued degradation of the environment will lead to serious health problems for their and future generations. Statistical and epidemiological methods and research are needed to obtain a better understanding of the complex relationships between human health, ecological health and pollution. These relationships are not based on standardized sets of observations or easily obtainable data. The present tools of biostatistics and epidemiology are inadequate to deal with many problems of environmental health. These areas pose unusual sampling problems, produce data that often are not normally distributed, or pose problems for which adequate models have not been developed. Standard multivariate analysis does not provide a sampling frame to account for fine mesh distribution.

Public Access

With the advent of the "information highway," the public is being provided unprecedented access to environmental data collected by Federal, State and Local organizations. Unfortunately, the free economy philosophy of "caveat emptor" cannot hold. Much of the raw data that is becoming available has a number of serious problems relating to data quality and definition. If these data are to be made available to the public, then it is the responsibility of environmental statisticians to provide the public with the capability to make the data into information or to make appropriate judgements on the correct use of the data. The release of environmental data/information under the "buyer beware" principle is irresponsible and will lead to misinformation and costly mistakes in assessing the state and health of the environment. Access needs to be given to the public; however, the public must be educated on how to use and understand data which is uncertain and often biased. Environmental data providers must ensure that appropriate "meta-data" are available to allow this "educated" public to appropriately use and interpret the data/information being released.

This conundrum and the associated statistical issues are exemplified by the USEPA Toxic Release Inventory (TRI). Through the Superfund Reauthorization Amendments (SARA) Title 313, in 1987 the U.S. Congress passed legislation requiring companies who employ more than ten employees and who produce more than 25,000 pounds of the TRI's list of substances, or firms that use more than 10,000 pounds of these substances per year, are required to report annual releases and transfers of TRI chemicals to the USEPA. In turn, the USEPA is required to make this information available to the public on a site identifiable basis.

TRI data are now available to the public, but only in their "raw" form with no meta information. A number of information services have downloaded the TRI data bases and are providing summary statistics, time series and interpretation of the changes as if the data were of known quality. In fact, the quality of the data is unknown: TRI data are self-reported and there are no standard for reporting. Some of the data is observational, some is model generated, and some are "best guess". The public has no way of knowing which is which or what comparisons are legitimate, if any. With all these problems, the release of TRI has been an environmental information success. The public is using the information to effect change. Companies are beginning to realize that the data they provide will be used and that they need to be more careful in data measurement and generation. Statisticians can play an important role in the development of appropriate methods to use, and in the display and visualization of this type of data in a manner that allows the public to make more informed decisions.

SUMMARY

We have discussed several areas where statistical methods are central to environmental science and decision making. Solutions to these problems and proliferation of the use of these methods will improve the quality and usefulness of environmental data and decision making. The interface of statistics and regulatory policy requires the development and use of

new and innovative methods which can provide environmental managers with the quantitative component of their decision making process. Challenges are in the application of sound statistical methods to combine existing environmental data and the development of cost effective methods for the monitoring, analysis and display of primary data.

TECHNICAL REPORT DATA

1. REPORT NO.	2.	3
EPA/600/A-96/018		

4. TITLE AND SUBTITLE	5. REPORT DATE
Statistical Issues in Environmental Monitoring and Assessment of Anthropogenic Pollution	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Lawrence H. Cox N. Phillip Ross	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Environmental Protection Agency National Environmental Research Laboratory Research Triangle Park, NC 27711 U.S. Environmental Protection Agency ESID/OPPE, Washington, DC 20460	10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO. N/A
12. SPONSORING AGENCY NAME AND ADDRESS U.S. Environmental Protection Agency National Environmental Research Laboratory Research Triangle Park, NC	13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

Spruce III Conference keynote paper to be published as a book chapter

16. ABSTRACT

Environmental data are often collected to assist in assessment of the impact of anthropogenic pollution on the natural environment, to determine the effects of pollution on human and ecological health, for enforcing compliance with environmental regulations and standards, and to assess the state of the environment. Since primary environmental data collection is costly, data sets are often used for multiple purposes. In addition, environmental information is collected by many diverse and independent organizations. This results in a patchwork of spatially and temporally different data sets that profess to measure the same phenomenon, but defy the use of classical statistical approaches for their integration and analysis. This duality leads to a number of statistical issues relating to the measurement and monitoring, use, and analysis of environmental data. Statisticians are being asked to convert the proverbial "sow's ear" into a "silk purse". In this introductory paper to the Third Spruce Conference, we address some statistical issues at the interface of environmental measurement, monitoring and regulatory decision making. Some of the areas explored are: environmental measurement and monitoring, environmental indicators, sampling approaches, use of environmental models, use of "encountered" or "found" environmental data, environmental decision making and public policy, and environmental reporting.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/ OPEN ENDED TERMS	c. COSATI

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
Unclassified	Release to Public	
	20. SECURITY CLASS *(This Page)*	22. PRICE
	Release to Public	