STATISTICAL ANALYSES´ DESIGN

Vijay P. Singh
Louisiana State University, U.S.A

Keywords: Correlation, design, entropy, errors, extreme value analysis, goodness of fit, parameter estimation, periodicity, probability distribution, reliability, sampling, statistics, stochastic process, time series, trend.

Contents

Summary

Environmental problems can be classified into five groups: (1) prediction, (2) forecasting, (3) identification, (4) modeling or simulation, and (5) detection. A statistical analysis design depends on the problem to be tackled, the quantity and quality of data available, and the objective for which the solution is to be used. Statistical techniques are briefly surveyed to address different groups of problems, and a design strategy for statistical analyses is formulated.

1. Introduction

A physical system (e.g., an environmental or a water resources system) is comprised of

(1) geometry; (2) forcing functions, including sources (input) and sinks (abstractions); (3) scientific laws governing the system behavior; (4) initial and boundary conditions; and (5) system response (or output) (Singh, 1996). Consider, for example, riverine pollution. The river reach under consideration constitutes the system. Fluvial geomorphology of the reach specifies the system geometry. The reach receives pollutants either at its head or along its sides or both. Flow and pollutants entering the reach from its sides define the input or source (of water and pollutants). The equations governing the flow of water and pollutants are the equations of continuity and momentum, (or flux) for water and the continuity equation, and flux law for pollutants. These governing equations can also be expressed in simpler forms. The condition of the reach (both in terms of flow and pollutant concentration) prior to the introduction of pollutants defines the initial condition. The flow and pollutant concentration at the head of the reach define the upstream boundary condition. If there is a downstream control, then the flow and pollutant concentration will define the downstream condition. Depending on the type of the governing equations to be employed and the type of flow (e.g., subcritical or supercritical), the downstream boundary condition(s) may be needed, regardless of the existence of the downstream control.

All approaches to modeling depend on the detail with which the system components or a combination thereof is taken into account. For example, a statistical regression approach lumps system geometry, governing laws and initial and boundary conditions; and establishes a relation between the forcing functions (input) and system response (output). This also is more or less true of a time series approach. In a phenomenological approach, governing equations are replaced by simple axioms or hypotheses. In a physically-based approach scientific laws are preserved as much as data and practical justification will permit.

Environmental systems are inherently spatial and complex, and our understanding of these systems is less than complete. Many of the systems are either fully stochastic, or part stochastic and part deterministic. Their stochastic nature can be attributed to randomness in one or more of the aforementioned components that constitute them and influence their behavior. As a result, a stochastic description of these systems is needed, and statistics enables development of such a description.

There is a rich diversity of statistical techniques that can be employed to address environmental problems. The objective of this note is to present a design strategy for statistical analyses. The description, however, depends on the environmental problem to be tackled, the quantity and quality of data available, the objective to be achieved and the constraints to be satisfied.

2. Classification of Environmental Problems

Most environmental problems can be broadly classified into three groups (Singh, 1988): (1) direct problems, (2) indirect (or inverse) problems, and (3) simulation problems. Direct problems are characterized by a complete specification of the system in terms of a mathematical equation. The objective is to study or predict the system response for a specified input. Direct problems can be of five types: (a) analysis problems, (b) estimation problems, (c) prediction problems, (d) forecasting problems, and (e)

frequency problems.

An estimation problem consists in determination of the system response, given the system, the input, and the initial and boundary conditions. An estimation problem involves determination of the past, future, or present states of a system in the presence of noise. Because of noise, the determination is less than perfect. An example is determining pollution in a river reach under specified conditions and input. The input may be noisy. A prediction problem may involve either the determination of the future state of the system at a specified time and location, or specification of the system response at any time represented by past, future or present. Given input and river morphology, determining pollution in a river reach for specified initial and boundary conditions will exemplify a prediction problem. A forecasting problem entails determination of the future state of the system with a certain probability. Examples include weather forecasting, forecasting river pollution, river forecasting, forecasting lake levels, forecasting ozone levels, etc. A frequency problem involves the determination of frequency of a specified magnitude of an event.

Indirect problems are of two types: (a) identification or modeling problems, and (b) detection, control, or instrumentation problems. Inverse problems, in general, are more difficult than direct problems. A direct problem usually has a unique solution, if it has a solution; whereas there may exist a multitude of solutions for an inverse problem. An identification problem consists in determining a mathematical characterization of the system, given a finite set of observations on input and output. The system characterization may involve either determination of system parameters or/ and system structure. Given input and pollution output for a river reach, initial and boundary conditions, and river morphology, characterizing the river system function for pollution will represent an identification problem. A detection problem is one of determining the system input, given the system output and the system function. If the pollution output is known, the initial and boundary conditions are known, and the river morphology as well as the governing equations are known, then determining the pollution input for the reach will constitute a detection problem.

In a simulation, synthesis or design problem, the nature of the expected input and the nature of the expected system response are specified. A system having this input response has to be physically realized or designed. Thus, a simulation problem involves the design of a non-existent system that is normally required to satisfy certain desirable social goals, according to certain criteria, such as cost, benefit, safety, reliability, risk, life span, etc. For example, in the case of river pollution, the identification problem will consist of determining the river system function as well as prediction of the river pollution.

Engineering decisions concerning environmental systems are frequently made with less than adequate information. Such decisions may often be based on experience, professional judgment, thumb rules, crude analyses, safety factors, or probabilistic methods. Usually, decision making under uncertainty tends to be relatively conservative. Quite often, sufficient data are not available to describe the random behavior of such systems. Although probabilistic methods allow for a more explicit and quantitative accounting of uncertainty, their major difficulty occurs due to the avail-

ability of limited or incomplete data. Small sample sizes and limited information render estimation of probability distributions of system variables with conventional methods quite difficult. Where the shortage of data is widely rampant as is normally the case in developing countries, the decisions have to be made with whatever data is available. This note revisits the techniques of statistical analyses and underscores their usefulness for decision making in environmental resources.

3. A Sample of Environmental Problems

Statistical methods are applied to a range of problems in environmental development, conservation, control, protection, and management. A short discussion of a sample of such problems is, therefore, in order. Cleaning polluted environments is a difficult task. For example, ground water, once contaminated, is difficult to clean, and its cleaning is extremely expensive and time- consuming. Cleaning of surface water systems is less difficult, but is not easy either, as vividly exemplified by the experience of oil spill in Valdes in Alaska. Most pollution-related phenomena are predominantly stochastic, periodic-stochastic, or trend-periodic-stochastic processes. To develop cleaning strategies, data is to be gained by monitoring, and information is to be extracted. There is then a whole range of statistical tools that can be employed.

There is usually a conflict between development and environmental protection. For example, when a dam is proposed for construction, there is usually an objection to dam building from fisheries biologists, environmentalists, sociologists, and so on. Because dam causes inundation, there may be relocation of people and the ecosystem of the area may be affected. When a new residential area is proposed for development, people living downstream or in the neighborhood object on the grounds of environmental degradation. Resolution of such a conflict requires information which may be meager, even though the conflict may be large and serious. What are the risks of failure in the future? What are the uncertainties in evaluating risks? What are the chances that cost will not exceed budget? These questions need to be addressed with use of statistical models.

The ongoing debate on climate change and global warming and the ensuing consequences is often based on limited information and data, or, on the models which have limited capabilities. Is climate actually changing? How large is going to be the climate change? What are its consequences? What is causing this change? Can this change be stopped and how? What is going to be its impact on human civilization? Because climatic variables are random variables, reliable statistical models are needed for change discrimination.

Environmental systems have limited physical lives. For example, urban water systems of large old cities around the world are in dilapidated state. Either they need renovation or replacement. How rapidly are these systems deteriorating? Since aging of these systems is not uniform, an assessment of the deterioration of these systems and structures is badly needed. What are the cost implications? This involves statistical sampling and the attendant statistical ramifications. Along similar lines, improvement of existing systems requires new information, new social values or demands and new paradigms. Data on performance of these systems is needed.

There is the problem of changing perception of risk. Societies are less and less tolerant of risks either because of increased stakes or increasing tendency toward intolerance. For example, people are much less willing to support development of nuclear power plants or development of chemical plants, even if they promise jobs and economic security. The consequences of failure are so frightening that people are less willing to take risks. This may be due to the lessons learnt from disastrous failures of the Three Mile Island nuclear power plant in the United States and of Chernobyl nuclear power plant in former Soviet Union, as well as of Union Carbide chemical plant in Bhopal, India, and a multitude of others. These disasters are a vivid reminder of scientific and technological limitations. Thus, risk assessment and reliability estimation are needed to justify maintenance of existing systems.

Conflicts of interests have been multiplying over the years, as evidenced by litigation in courts. These days it seems every conflict leads to litigation, and courts are becoming the ultimate bastion of conflict resolution. Why is it happening? Is that the road we want to traverse? Can there be a more efficient, just and fair way to resolve such conflicts? The precise determination of available resources (say, water), estimation of variation of resources, damages, costs, benefits, risks, etcetera usually dominate the conflict-resolution discussions.

Changes in priority of use of environmental resources are occurring rapidly these days. For example, in case of water resources, springs, aquifers, lakes, rivers, and recycled waters are the usual sources of water supply. Which sources will dominate water supply in a given area will be dictated by pollution potential, abatement of polluted waters, etc. Monitoring of water quality and the resulting observations are needed to define the priority of use of the various sources of water. Along similar lines, marketing water and water rights is going to be an important issue as the pace increasingly accelerates toward market-driven economy. It is quite possible that drinking water will eventually be an industrially produced product. This has already started occurring in the United States, and has been so in Europe for many years. Even in poor developing countries of Asia, bottled water has started taking hold.

4. Environmental Data

Before undertaking a statistical analysis, environmental data need to be carefully processed with respect to precision and accuracy, homogeneity and consistency, completeness and length of record, and statistical characteristics. Indeed many times, it is the data that dictate the type of analysis to be performed rather than the availability of technology itself.

Measurements should be sufficiently accurate or unbiased and precise or certain. The errors associated with measurements can be characterized with respect to their accuracy and precision. Precision refers to how closely individual measured values agree with each other. Thus, precision signifies (1), the number of significant figures representing a quantity or (2), the spread in repeated measurements of a particular value. Imprecision, also called uncertainty, refers to the magnitude of the scatter. Inaccuracy, also called bias, is defined as the systematic deviation from the truth. The environmental data can be (1), inaccurate and imprecise, (2), accurate and imprecise, (3), inaccurate and

precise, and (4), accurate and precise. Thus, the data error represents both the inaccuracy and imprecision.

Errors in measurements result from three sources: (a) instrumental defects, (b) improper siting or location of the measurement device, and (c) human errors. The resulting errors are of two types: systematic and random. Random errors occur when the data show no tendency toward either overestimation or underestimation of measured values for a number of successive time intervals. Random errors in the data will undoubtedly produce random errors in the output. Systematic errors occur when the error tends to persist over a series of time intervals without changing sign. Systematic errors will be reflected in the incorrect parameter values. The importance of one type of error relative to the other depends on the problem at hand. Nonlinearity of environmental processes complicates treatment of the mechanism by which errors in data are transferred to model parameters and combined with input errors in the test period to produce errors in the simulated output. All errors transmit part of their magnitudes to model parameters and then to model results, but each one does so differently.

Many environmental analyses require a long-term record of data. It is observed that measuring devices are moved from their original location for one reason or another, and this, in turn, affects the consistency of measurements. If the record is not consistent then it must be corrected for inconsistency. There are graphical and statistical techniques, including the double mass curve analysis, the von Newmann ratio test, cumulative deviation test, likelihood ratio test, run test, among others for testing the consistency of data (Singh, 1988).

It is not uncommon in environmental resources that measurement records are incomplete. Breaks may vary in length from one or two days to several years. It is often necessary to estimate the missing data in order to utilize partial records. This is especially important in data-sparse regions. Several methods are available for estimating missing data, including the arithmetic average method, the normal ratio method, the inverse distance method, modified inverse distance method, linear programming, isohyetal method, the Lagrange method, interpolation methods, maximum entropy spectral methods, finite element method, kriging, among others (Singh, 1988).

Point data are used in a variety of applications. When used in frequency analysis, the point record may not be of sufficient length. It may then be necessary to extend the point record. One simple method to accomplish the extension of the record is the station-year method. This method combines the records of several measuring sites into a single composite record such that the composite record is of the length equal to the sum of the lengths of individual records. This method assumes that the records of individual measuring sites are independent, and the area where the measuring sites are located are homogeneous, meaning that these records belong to the same population and are generated by similar physical mechanisms.

Frequently, measured data may not be used as such. Rather they are used to extract certain characteristics, such as extreme values, average values (yearly, monthly, weekly, hourly, etc.), integrated values, etc. For extreme value analyses, annual maxima or annual minima are needed. Annual maxima may be obtained in two ways. First, for each

year the instantaneous maximum value is chosen. Thus, there will be as many maximum values as the number of years of record. These annual maxima constitute the complete duration series and are employed for extreme value analyses. Second, a threshold is selected and all instantaneous values exceeding the threshold are selected. In this case, some years may contribute more than one value and some may contribute none. The maxima so chosen constitute the partial duration series and are employed for phenomenological maximum value analyses. In a similar vein, annual averages may be computed and analyzed.

Environmental data should be reduced to a common homogeneous condition. All time series modeling approaches assume that the data reflect a stationary process. This means that the historical data should be transformed to reflect natural conditions, so that the natural process can be reliably modeled. This step may encompass correction of historical data for systematic errors, filling-in of missing records, extension of data, and the reduction of data to the natural condition.

5. Characteristics of Environmental Processes

Environmental time series are often represented by such components as overyear trends and other deterministic changes, cycles or periodic changes of the day and the year, almost periodic changes, and components representing stochastic or random variations. Thus, environmental processes exhibit trends, periodicities, stochastic dependence structure components and independent stochastic components. Inconsistency (systematic errors) and nonhomogeneity (changes induced by humans or natural processes) cause overyear trends or sudden changes. These features must be identified and removed. Trends and cycles may result from sampling fluctuations. Before considering them as part of the population series, they must be tested for statistical significance. Periodicity is caused by astronomic cycles and implies that statistical characteristics change periodically within the year. Randomness in time series is caused by many processes in the earth's environment, such as turbulence, large scale vorticity, random thermodynamic processes, incoming and outgoing radiation, etc. These causes lead to variations in time series, called stochastic components. Inputs to environmental systems are mostly a combination of periodic and stochastic variations. An environmental system acts upon these inputs in three ways: (1) The input may be smoothed or magnified; (2) the periodic component may be added, attenuated, amplified or damped; and (3) randomness resulting from other factors may be added or modified.

6. Domains of Analyses

Environmental problems are analyzed in time, space, space-time, or frequency domains. In other words, they involve either (a) temporal, (b) spatial, (c) spatio- temporal, or (d) frequency analyses. In temporal analyses, spatial variability is not explicitly considered. The data correspond to either a point in space or are integrated over space. Or spatial variability is assumed not to exist and environmental systems are assumed uniform but unsteady. On the other hand, in spatial analysis temporal variability is not explicitly considered or systems are assumed to be steady but nonuniform. Both spatial and temporal variability is explicitly accounted for in spatio-temporal analysis. In this case, the systems are assumed unsteady and nonuniform. In frequency domain, either

probability distributions are derived or spectral properties are analyzed.

From an environmental perspective, the above-mentioned analyses employ either the complete record or some selected values, such as rare values, high or low extreme values, values exceeding or below a certain threshold, etc. The type of data to be used determines the choice as to the type of the methodology to be employed. Thus, statistical analyses can be classified into (a) extreme value analyses, and complete data analyses (or non-extreme value analyses).

Extreme-values may be either low values or high values. Rare values may also be included in the category of extreme values. Many environmental issues need an analysis of such values of measured environmental variables. Examples of such issues are extreme weather events, including low and high temperatures, low and high rainfalls, low and high winds, lightning, and hurricanes; floods and droughts; global warming and its impact on weather patterns; extreme levels of tropospheric ozone; to name but a few. The main objective of extreme value analyses is to derive probability distributions of the magnitude and number of extreme values, and the time interval between occurrences (called inter-arrival times) of extreme values.

The complete data analyses encompass a broad range of analyses, including (a) sampling of information, (b) probability and frequency analyses, (c) determination of functional relationships between variables, (d) forecasting and prediction of system response, (e) transfer of information between variables or systems, (f) testing hypotheses and inferences, (g) selection from amongst alternatives, (h) determination of system characteristics, and (i) risk and reliability analyses.

-
-
-

Bibliography

Akaike, H.(1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, Vol. **AS-19**, No. 6, pp. 716-723.

Box, G.E.P. and Jenkins, G.M., (1976). *Time Series analysis: Forecasting and Control*. Holden-Day, San Francisco, California.

Davison, A. C. and Smith, R. L. (1990) *Models for exceedances over high thresholds. Journal of Royal Statistical Society, Series B*, Vol. **52**, pp. 393-442.

Piegorsch, W. P., Smith, E. P., Edwards, D. and Smith, R. L.(1998) Statistical advances in environmental science. *Statistical Science*, Vol. **13**, No. 2, pp. 186-208.

Salas, J. D., Delleur, J. W., Yevjevich, V. and Lane, W. L.(1980) *Applied Modeling of Hydrologic Time Series. Water Resources Publications*, Littleton, Colorado, U.S.A.

Salas, J. D. and Yevjevich, V.(1972) Stochastic structure of water use time series. *Hydrology Papers* **52**, Colorado State University, Fort Collins, Colorado.

Singh, V. P. (1988) *Hydrologic Systems*, Vol.**1**: Rainfall-Runoff Modeling. Prentice Hall, Englewood Cliffs, New Jersey, U.S.A.

Singh, V. P. (1996) *Kinematic Wave Modeling in Water Resources: Surface Water Hydrology*. John Wiley & Sons, New York, U.S.A.

Singh, V. P. (1998) *Entropy-Based Parameter Estimation in Hydrology*. Kluwer Academic Publishers, Boston, U.S.A.

Smith, R. L.(1993) Long-range dependence and global warming. In: *Statistics for the Environment*, edited by V. Barnett and K. F. Turkman, pp. 141-161, John Wiley & Sons, New York, U.S.A.

Yevjevich, V. (1993) Water resources and statistics: past, present and future. Chapter 10 in *Statistics for the Environment*, edited by V. Barnett and K. F. Turkman, pp. 201-224, John Wiley & Sons, Chichester, England.

Biographical Sketch

Vijay P. Singh was born on July 15, 1946, in Agra, India. He obtained B. S. in Engineering and Technology in 1967 from Pant College of Technology in India; M. S. in Engineering specializing in Hydrology in 1970 from University of Guelph, Ontario, Canada; Ph. D. in Civil Engineering with an emphasis on Hydrology and Water Resources in 1974 from Colorado State University; and D. Sc. in Engineering in 1998 from the University of Witwatersrand, Johannesburg, South Africa. He is a registered professional engineer and a registered professional hydrologist.

Currently, he holds the Arthur K. Barton Endowed Professorship in Civil and Environmental Engineering at Louisiana State University. He has received more than 30 awards, including the Distinguished Service Award from the National Research Council of Italy in 1995; the Fulbright Scholar Award in 1997; International Man of the Year Award from the International Biographical Center in 1997; the Brij Mohan Distinguished Professor Award in 1999; the Distinguished Faculty Award in 1999; the Achievement in Academia Award in 1999 from Colorado State University College of Engineering; James M. Todd Technological Achievement Award in 2000 from Louisiana Engineering Society etc. He is a fellow of ASCE, AWRA, IE, IAH, ISAE, and IWRS. He has authored 9 text books, edited 25 books. He is Editor-in-Chief of Water Science and Technology Library Book Series and is a member of 9 journal editorial boards.

He serves as Senior Vice President of American Institute of Hydrology, Vice President of Indian Association of Hydrologists, and President of G. B. School. Board;

Professor Singh's research interests have encompassed a wide range of topics in both surface and subsurface water hydrology, watershed hydraulics, irrigation engineering, and water quality engineering. He has extensively worked on kinematic wave modeling; hydrodynamics of surface irrigation; erosion and sediment transport in upland watersheds; point and non-point source water quality modeling; hydrologic modeling of ungaged watersheds; flow forecasting; areal rainfall; dam break modeling; parameter estimation for frequency distributions; multivariate stochastic analysis of hydrologic extremes; entropy modeling in hydrology; network design; landfill hydrology; saltwater intrusion in coastal aquifers and ground water modeling.

Professor Singh is also actively involved in charitable activities. He founded the G.B. School in 1994 in Agra, India. The school imparts quality education to children in rural India. He recently founded the Foundation for the Aggrandizement of Rural Areas (FARA).