ENVIRONMENTAL STATISTICS and DATA ANALYSIS

WAYNE R. OTT

ENVIRONMENTAL STATISTICS and DATA ANALYSIS

WAYNE R. OTT

LEWIS PUBLISHERS
Boca Raton Ann Arbor London Tokyo

Library of Congress Cataloging-in-Publication Data

Ott, Wayne
Environmental statistics and data analysis / Wayne R. Ott
p. cm.
Includes bibliographical references and index.
1. Environmental sciences—Statistical methods. I. Title.
GE45.S73088 1995
363.7'0072—dc 20 94–15435
ISBN 0–87371–848–8 (acid-free paper)

Direct all inquiries to CRC Press, Inc., 2000 Corporate Blvd., N.W., Boca Raton, Florida 33431.

Lewis Publishers is an imprint of CRC Press

International Standard Book Number 0–87371–848–8

Library of Congress Card Number 94–15435
Printed in the United States of America
1 2 3 4 5 6 7 8 9 0

Preface

Many random processes occur in nature. To make accurate predictions about the manner in which man's activities may alter these processes and affect their outcomes, it is necessary to construct models that faithfully represent reality, including its random components. In the environmental field, the purpose of most models is to make accurate predictions about the effect of environmental pollution control activities on environmental quality or on human exposure to pollutants. Although modeling the fate and transport of pollutants through the environment is well advanced, few models have been developed adequately to include the random, or *stochastic*, nature of environmental phenomena. Stochastic models treat the phenomenon being modeled probabilistically, thus including the random components in a statistical framework. Because random phenomena abound in the environment, stochastic modeling often is more important than other kinds of modeling. Despite the importance of stochastic models in environmental analyses and decision-making, few reference works are available covering these techniques and showing how to apply them to environmental problems. Of particular importance for concentrations measured in the environment are the right-skewed distributions, such as the lognormal, and techniques to take into account source controls, such as rollback models (see Chapter 9).

To help fill the need for a reference work on environmental statistics, this book seeks to develop a comprehensive and understandable framework for applying probabilistic techniques to environmental problems of all kinds. It includes statistical models for environmental decision-making, data analysis, and field survey design, along with the theoretical basis for each model wherever possible. A model that has a sound theoretical basis is more likely to make accurate predictions than one that does not. This book also includes a considerable body of original material, not previously published, on new theories and insights to help explain observed environmental phenomena and the control of these phenomena. The new theories are included to help guide data analysts and decision makers in applying statistical models to practical problems in the environment.

The book is intended to provide students, managers, researchers, field monitoring specialists, engineers, statisticians, data analysts, and environmental decision makers with both a reference source and a body of statistical procedures for analyzing environmental data and making environmental predictions. In its structure, this book includes full documentation of each probability model presented, the theoretical basis for its origin, and examples of its application to environmental problems. In addition, BASIC computer programs, which can be programmed readily on personal computers, are included in the book wherever possible. Typical uses of these probability models include analyzing environmental monitoring data, describing the frequency distribution of exposures of the population, deciding the degree to which environmental measurements comply with health-related standards, and predicting the effect of pollutant source reductions on environmental quality. The book also introduces several new techniques, not presented elsewhere, that are important for solving practical problems.

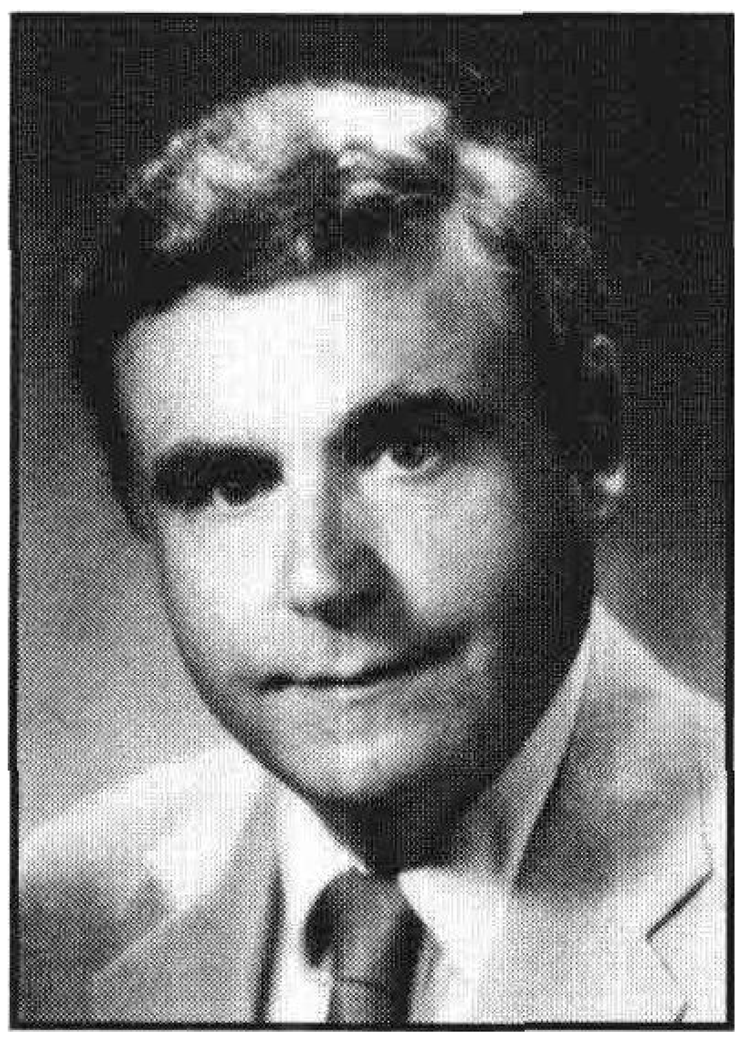
This book is dedicated to the concept that selection of a stochastic model should be based not merely on its presumed "good fit" to empirical data; rather, it should be consistent with the basic theory of the underlying physical laws responsible for generating the observed data. To assist in determining which models are theoretically appropriate for certain physical processes, the idealized physical conditions associated with each probability distribution are presented in detail. It may appear surprising to some readers, for example, to discover that commonly applied Gaussian diffusion plume models can arise naturally from probability theory and from fairly simple "random walk" Brownian motion examples.

In Chapter 8, the Theory of Successive Random Dilutions (SRD) is proposed to explain why lognormal appearing distributions occur so commonly in the diverse fields of indoor air quality, water quality, and geological systems. Hopefully, the concepts presented in this book will stimulate other investigators to consider why right-skewed distributions with a single mode occur so often in environmental pollution data. It is hoped that the SRD concepts presented in this book are fundamental to developing a general theory, and that others in the future will help extend these principles.

For more than three decades, papers have appeared in the literature on the effect of source reductions on environmental concentrations, but these predictions seem to be based more on hunches of the authors than on a proven theory or on experimental evidence. In Chapter 9, a Statistical Theory of Rollback (STR) is introduced for the first time to provide a basis for predicting the effect of changes in sources on observed environmental concentrations. The STR is derived from basic statistical concepts. Some experimental observations are included in Chapter 9 to illustrate how this theory might be applied in practice. It is hoped that this statistical theory will be applicable to a wide variety of environmental problems.

To ease presentation of the theories presented in this book, many practical examples and commonplace physical analogs are included, both to help show how a process works and to describe the physical laws responsible.

The objective of this volume is to communicate basic statistical theory to a broad environmental audience — chemists, engineers, data analysts, and managers — with as little abstract mathematical notation as possible, but without omitting important details and assumptions. It is assumed only that the reader has a basic knowledge of algebra and calculus. I hope that this book can serve as a reference source for those wishing to analyze, understand, and predict environmental quality variables in many practical settings. I also hope that this book will help provide a beginning toward establishing a comprehensive body of knowledge known as "Environmental Statistics."



Wayne R. Ott has been with the U.S. Environmental Protection Agency (EPA) and its predecessor agencies for over 28 years, conducting research on environmental statistics, environmental data analysis, environmental indices, environmental decision analysis, mathematical modeling of human exposure to environmental pollutants, Monte Carlo simulation, environmental field studies, measurement of human exposure, indoor air quality, stochastic modeling of environmental processes, human activity patterns, and quality assurance of environmental measurements.

Dr. Ott earned a B.A. degree from Claremont McKenna College in business economics; a B.S. degree in electrical engineering, an M.A. degree in communications, and an M.S. degree in engineering science from Stanford University; and received his Ph.D. in civil and environmental engineering from Stanford University. He is a member of the American Statistical Association, the International Society of Exposure Analysis, the Air and Waste Management Association, Kappa Mu Epsilon (mathematics honor society), Tau Beta Pi (engineering honor society), Sigma Xi (scientific research society), and Phi Beta Kappa (honorary member). He is a Commissioned Officer in the U.S. Public Health Service (PHS) with the rank of Captain.

In the 1970's, Dr. Ott developed mathematical techniques for analyzing water and air quality data, leading to the proposed structure of the Pollutant Standards Index (PSI), a nationally uniform air pollution index now used throughout the U.S. and in other countries. He received the Public Health Service Commendation medal in 1977 for his work in developing the PSI. In 1980, he received a competitive EPA Innovative Research Program grant award to become a visiting scholar at Stanford University's Department of Statistics in conjunction with the Societal Institute for the Mathematical Sciences (SIMS) research program. While at Stanford, he developed and wrote the computer source code for the Simulation of Human Air Pollutant Exposure (SHAPE) model, the first human activity pattern-exposure model.

He contributed to development of the Total Human Exposure concept and to EPA's Total Exposure Assessment Methodology (TEAM) field studies.

Dr. Ott has written or co-authored over 100 research reports, technical papers, and journal articles on statistical analysis of environmental data, mathematical modeling of human exposure, environmental indices, probabilistic concepts, indoor air quality, measurment methods, quality assurance, environmental data analysis, field study design, and other related topics.

Acknowledgments

I wish to acknowledge the skill, care, and tireless energy provided by the reviewers of this book, who were kind enough to examine the draft manuscript and recommend numerous changes and improvements over a period spanning a decade. Several chapters were reviewed by Steven Bayard, James Repace, Ralph Larsen, and Lynn Hildemann. All chapters were reviewed by David Mage, Lance Wallace, and William Sayers. I am extremely grateful for the important contribution they made to the completion of this book. I also appreciate the thoughtful suggestions by Paul Switzer on the concept of successive random dilutions applied to stream water quality. I am also grateful for the editorial assistance provided by Vivian Collier, Virginia Zapitz, and Le verne McClure.

NOTICE

Contents

1 Random Processes

> *Random*: A haphazard course — *at random:* without definite aim, direction, rule or method[1]

The concept of "randomness," as used in common English, is different from its meaning in statistics. To emphasize this difference, the word *stochastic* commonly is used in statistics for *random*, and a *stochastic process* is a *random process*. A stochastic process is one that includes any random components, and a process without random components is called *deterministic*. Because environmental phenomena nearly always include random components, the study of stochastic processes is essential for making valid environmental predictions.

To most of us, it is comforting to view the world we live in as consisting of many identifiable cause-effect relationships. A "cause-effect" relationship is characterized by the certain knowledge that, if a specified action takes place, a particular result always will occur, and there are no exceptions to this rule. Such a process is called deterministic, because the resulting outcome is determined completely by the specified cause, and the outcome can be predicted with certainty. Unfortunately, few components of our daily lives behave in this manner.

Consider the simple act of obtaining a glass of drinking water. Usually, one seeks a water faucet, and, after it is found, places an empty glass beneath the faucet and then turns the handle on the faucet. Turning the handle releases a piston inside the valve, allowing the water to flow. The process is a simple one: the act of turning the handle of the faucet (the "cause") brings about the desired event of water flowing (the "effect"), and soon the glass fills with water.

Like so many other events around us, this event is so familiar that we ordinarily take it for granted. If, before we operated the faucet, someone asked us, "What will happen if the handle of the faucet is turned?", we would be willing to predict, with considerable certainty, that "water will appear." If we had turned the handle and no water appeared, we probably would conclude that there is something wrong with the plumbing. Why do we feel so comfortable about making this simple cause-effect prediction? How did we arrive at this ability to predict a future event in reality?

In our mind, we possess a conceptual framework, or a "model," of this process. This model has been developed from two sources of information: (1) Our knowledge of the physical structure of faucets, valves, and water pipes and the manner in which they are assembled, and (2) Our historical experience with the behavior of other water faucets, and, perhaps, our experience with this particular faucet. The first source of knowledge comes from our understanding of the physical construction of the system and the basic principles that apply to water under pressure, valves that open and close, etc. For example, even if we had never seen a faucet or a valve before, we might be willing to predict, after the mechanism and attached pipe were described to us in detail, that turning the handle of the faucet would release the water. The second source of knowledge is derived from what we have learned from our experience with other, similar faucets. We reason thus: "Turning the faucet handle always has caused the water to flow in the past, so why shouldn't it do so in the future?" The first source of knowledge is theoretical and the second source is empirical (that is, based on actual observations). If only the second source of knowledge were available — say, 179 cases out of 179 tries in which the faucet handle is turned on and the water appears — we probably would be willing to predict (based on this information alone and with no knowledge of the internal workings of the system) that turning the handle the next time — for the 180*th* try — would allow us fill the glass with water.

These two independent sources of information — physical knowledge of the structure of a system and observational knowledge about its behavior — greatly strengthen our ability to make accurate predictions about the system's future behavior. From the first source of information, we can construct a *conceptual model* based on the internal workings of the system. From the second source of information, we can validate the conceptual model with real observations. The first source of information is theoretical in nature; the second one is empirical. A theoretical model validated by empirical observation usually provides a powerful tool for predicting future behavior of a system or process.

Unfortunately, the world about us does not always permit the luxury of obtaining both sources of information — theory and observation. Sometimes, our knowledge of the system's structure will be vague and uncertain. Sometimes, our observational information will be very limited. Despite our lack of information, it may be necessary to make a prediction about the future behavior of the system. Thus, a methodology that could help us analyze existing information about the system to improve the accuracy of our predictions about its future behavior would be extremely useful.

Consider the above example of the water faucet. Suppose little were known about its construction, attached pipes, and sources of water. Suppose that the faucet behaves erratically: when the handle is turned, sometimes the water flows and sometimes it does not, with no obvious pattern. With such uncertain behavior of the device, we probably would conclude that unknown factors (for example, clogged pipes, defective pumps, broken valves, inadequate water supplies, wells subject to rising and falling water tables) are affecting this system. The faucet may, in fact, be attached to a complex network of pipes, tanks, valves, filters, and other devices, some of which are faulty or controlled by outside forces. Because the arrival of water will depend on many unknown factors beyond the user's control, and because the outcome of each particular event is uncertain, the arrival of water from the faucet may behave as a stochastic process.

How can one make predictions about the behavior of such a process? The first step is to express the event of interest in some formal manner — such as a

"1" or "0" — denoting the presence or absence of water, or a quantitative measure (gm or m^3) denoting the amount of water arriving in a fixed time interval. Such a quantitative measure is called a *random variable*. A random variable is a function of other causative variables, some of which may or may not be known to the analyst. If all of the causative variables were known, and the cause-effect relationships were well-understood, then the process would be deterministic. In a deterministic process, there are no random variables; one can predict with certainty the rate at which the water will flow whenever the valve is opened by examining the status of all the other contributing variables.

In view of the uncertainty present in this system, how does one develop sufficient information to make a prediction? If we had no prior information at all — neither theoretical nor empirical — we might want to flip a coin, showing our total uncertainty, or lack of bias, about either possible outcome. Another approach is to conduct an *experiment*. Let K be a random variable denoting the arrival of water: if $K = 0$, water is absent, and if $K = 1$, water is present. Each turning of the faucet handle is viewed as a *trial* of the experiment, because we do not know beforehand whether or not the water will appear. Suppose that the faucet handle is tried 12 times, resulting in the following series of observations for K: $\{1,0,1,1,1,0,1,0,1,1,1,1\}$. Counting the ones indicates that the water flowed in 9 of the 12 cases, or 3/4 of the time. What should we predict for the 13*th* trial? The data we have collected suggest there may be a bias toward the presence of water, and our intuition tells us to predict a "success" on the 13*th* trial. Of course, this bias may have been merely the result of chance, and a different set of 12 trials might show a bias in the other direction. Each separate set of 12 trials is called a *realization* of this random process. If there are no dependencies between successive outcomes, and if the process does not change (i.e., remains "stationary") during the experiment, then the techniques for dealing with Bernoulli processes (Chapter 4) provide a formal methodology for modeling processes of this kind.

Suppose that we continue our experiment. The faucet is turned on 100 times, and we discover that the water appears in 75 of these trials. We wish to predict the outcome on the 101*st* trial. How much would we be willing to bet an opponent that the 101*st* trial will be successful? The information revealed from our experiment of the first 100 trials suggests a bias toward a successful outcome, and it is likely that an opponent, witnessing these outcomes, would not accept betting odds of 1:1. Rather, the bet might be set at odds of 3:1, the ratio of past successes to failures. The empirical information gained from our experiment has modified our future predictions about the behavior of this system, and we have created a model in our minds. If an additional number of trials now were undertaken, say 1,000, and if the basic process were to remain unchanged, we would not be surprised if water appeared in, say, 740 outcomes.

The problem becomes more difficult if we are asked to compare several experiments — say, two or more different groups of 1,000 observations from the faucet — to determine if a change has occurred in the basic process. Such comparison is a "trend" analysis, since it usually utilizes data from different time periods. Probabilistic concepts must be incorporated into such trend analyses to help assess whether a change is real or due to chance alone.

Our intuitive *model* of this process is derived purely from our empirical observations of its behavior. A model is an abstraction of reality allowing one to make predictions about the future behavior of reality. Obviously, our model will be more successful if it is based on a physical understanding of the characteristics of the process as well as its observed past behavior. The techniques described in this book provide formal procedures for constructing stochastic models of environmental processes, and these techniques are illustrated by applying them to examples of environmental problems.

Many of these techniques are new and have not been published elsewhere, while some rely on traditional approaches applied in the field of stochastic modeling. It is hoped that, by bridging many fields, and by presenting several new theories, each technique will present something new that will provide the reader with new insight or present a practical tool that the reader will find useful.

STOCHASTIC PROCESSES IN THE ENVIRONMENT

The process described above is an extremely simple one. The environmental variables actually observed are the consequence of thousands of events, some of which may be poorly defined or imperfectly understood. For example, the concentration of a pesticide observed in a stream results from the combined influence of many complex factors, such as the amount of pesticide applied to crops in the area, the amount of pesticide deposited on the soil, irrigation, rainfall, seepage into the soil, the contours of the surrounding terrain, porosity of the soil, mixing and dilution as the pesticide travels to the stream, flow rates of adjoining tributaries, chemical reactions of the pesticide, and many other factors. These factors will change with time, and the quantity of pesticide observed in the stream also varies with time. Similarly, the concentrations of an air pollutant observed in a city often are influenced by hundreds or thousands of sources in the area, atmospheric variables (wind speed and direction, temperature, and atmospheric stability), mechanical mixing and dilution, chemical reactions in the atmosphere, interaction with physical surfaces or biological systems, and other phenomena. Even more complex are the factors that affect pollutants as they move through the food chain — from sources to soils, to plants, to animals, and to man — ultimately becoming deposited in human tissue or in body fluids. Despite the complexity of environmental phenomena, many of these processes share certain traits in common, and it is possible to model them stochastically. There is a growing awareness within the environmental community of the stochastic nature of environmental problems.

Ward and Loftis[2] note that, with the passage of the Clean Water Act (Public Law 92-500), water quality management expanded both its programs (permits and planning) and the money devoted to wastewater treatment plants. The data collected at fixed water quality monitoring stations* assumed a new role: to identify waters in violation of standards and to evaluate the effectiveness of expenditures of the taxpayers' money. They conclude that water quality monitoring was expected to serve as a "feedback loop" by which to evaluate the effec-

*A fixed water quality monitoring station usually consists of a set of sampling points (stations) at which samples are taken (usually "grab" samples) approximately once per month.[2,3]

tiveness of regulatory programs. Unfortunately, unless the stochastic properties of these data are taken into account, the inherent randomness of these data will conceal real changes in environmental conditions:

> When data, collected to check only if a sample meets a standard, are used to evaluate management's success, the stochastic variation in the data often completely masks any improvement in controlling society's impact on water quality. Since the data cannot show management's effectiveness, the conclusion is that fixed station data are useless. However, the data are not useless: they are simply being asked to provide information they cannot show, without further statistical analysis.[2]

Standards in air and water pollution control usually are long-term goals that are well-suited to statistical formulation. Previous environmental standards often have been deterministic, however, perhaps because it was believed that probabilistic forms would complicate enforcement activities. Practically, it is impossible to design a regulatory program that can guarantee that any reasonable standard *never* will be violated, and there is a growing awareness that probabilistic concepts should be an integral part of the standard setting process. Ewing[4] states that water quality standards should be formulated in a probabilistic manner:

> The establishment of state water quality standards for both interstate and intrastate streams has recently been accomplished. In practically every case, DO [dissolved oxygen] requirements have been set without any reference to the probability of these levels being exceeded. It would seem that the state-of-the-art is rapidly advancing to the point, however, where the probabilistic concept should be recognized more specifically in the statement of the water quality standards themselves.

Drechsler and Nemetz[5] believe that incorporation of probabilistic concepts places greater demands on the design of the monitoring program but will yield a more efficient and effective water pollution control program:

> We recommend that, where appropriate, standards be altered to reflect the probability of occurrence of pollution events. This will require a greater degree of information concerning both the distribution of pollutant discharges and biological damage functions.... These measures will help overcome some of the significant weaknesses in the current regulatory system for the control of water pollution and will be more efficient and effective in the protection of both corporate and social interests.

In the air pollution field, significant progress has been made toward incorporating probabilistic concepts into the basic form of the standards. In 1979, for example, the Environmental Protection Agency (EPA) revised its National Ambient Air Quality Standard (NAAQS) for ozone from a deterministic form to a probabilistic form. The probabilistic form was based on the "expected number of days" in which the measured ozone level exceeds a certain hourly average value. The NAAQS apply uniformly on a nationwide basis, and compliance is evaluated using data collected at fixed air monitoring stations* operated in most U.S. cities. The Clean Air Act (Public Law 91-604) requires that the NAAQS be evaluated and revised from time to time, and it is anticipated that future NAAQS will be formulated on a similar statistical framework. As indicated by Curran,[6] probabilistic air quality standards have important advantages for accommodating missing data and handling rare events occurring in an unusual year:

> While the initial short-term NAAQS promulgated by EPA in 1971 were stated as levels not to be exceeded more than once per year, the revised ozone standard, which was promulgated in 1979, stated that the expected annual exceedance rate should not be greater than one. From an air quality data analysis viewpoint, this change is intuitively appealing in that it incorporates an adjustment for incomplete sampling and provides a framework for quantifying the impact of an unusual year, particularly in the development of design values. From a statistical viewpoint, this change represents a transition from the simple use of observed measurements to the use of estimates of underlying parameters.

Statistical techniques for judging compliance with these probabilistic standards are presented in Chapters 4 and 5. It appears likely that there will be steady progress toward incorporating probabilistic concepts into future environmental standards.

STRUCTURE OF BOOK

The present chapter discusses the need for probabilistic concepts in the environmental sciences and introduces common terms in statistics (e.g., deterministic and stochastic processes, models, random variables, experiments, trials, realizations). Chapter 2 briefly presents and reviews the theory of probability, including such concepts as the union and joint occurrence of events, conditional probability, and Bayes' Theorem. It includes a bibliography of books written on probability, statistics, and data analysis over four decades. Chapter 3 introduces the formal concept of probability models, both discrete and continuous, discussing measures of central tendency, dispersion, skewness, and kurtosis. It also discusses data analysis, histograms, probability plotting, and fitting probability models to observations, including goodness-of-fit tests. Chapter 4 presents Bernoulli processes, discussing the theoretical conditions required for the binomial distribution, and illustrating the binomial model by applying it to cigarette smoking activities. The binomial probability model then is applied to the interpretation of air and water quality standards, discussing how one might han-

*Fixed air monitoring stations generally consist of a group of instruments that sample air continuously, housed in a single structure or building in the city, generating 1-hour or 24-hour concentration averages. Large cities may have 10 or more of these fixed stations.

dle environmental situations that do not meet the independence and stationarity conditions of the model. In Chapter 5, the theoretical conditions required for a Poisson process are presented, and several similar environmental applications of the Poisson probability model are discussed. It is shown that some interpretations of air quality standards can be handled more effectively by modeling them as Poisson rather than Bernoulli processes. Chapter 6 is intended as a transition from discrete probability models to continuous probability models. Using fairly simple examples, it shows how the symmetrical normal distribution arises naturally from diffusion processes, satisfying the diffusion equation, and it develops the theoretical basis for the Gaussian plume model. Chapter 7 presents the normal distribution in detail, showing an important practical use of the model: calculation of confidence intervals in random sampling surveys. Another class of physical processes — those involving dilution — gives rise to right-skewed distributions. Chapter 8 introduces, in detail, the Theory of Successive Random Dilutions to help explain why distributions that are approximately lognormal (i.e., the distribution of the logarithm of concentration is approximately normal) can arise from such a great variety of environmental phenomena. Chapter 8 is intended as a transition from symmetrical distributions to right-skewed concentration distributions, and Chapter 9 formally presents the lognormal distribution. Four different methods for estimating the parameters of the lognormal distribution from data are presented. Equations are given to assist the analyst in converting from the "normal parameters" to the "geometric parameters" to the "arithmetic parameters," or to any combination of these. Chapter 9 also introduces the Statistical Theory of Rollback, which provides statistical principles for predicting pollutant concentrations after controlling the sources of pollution. The objective is to make predictions about the distribution of concentrations after a source is controlled using information about the distribution before it is controlled. Although statistical rollback theory is general and applies to any distribution, Chapter 9 illustrates statistical rollback applied to the lognormal distribution. It illustrates statistical rollback approaches using exposure field measurements on a U.S. arterial highway at two time periods separated by 11 years. Chapter 9 brings together many of the principles and concepts presented earlier in the book, introducing techniques that draw upon principles presented in earlier chapters.

REFERENCES

1. Merriam Webster's Collegiate Dictionary, Tenth Edition (Springfield, MA: Merriam-Webster, Incorporated, 1993).
2. Ward, Robert C. and Jim C. Loftis, "Incorporating the Stochastic Nature of Water Quality Into Management," *J. Water Poll. Control Feder.* 55(4):408–414 (April 1983).
3. Ott, Wayne, *Environmental Indices: Theory and Practice* (Ann Arbor, MI: Ann Arbor Science Publishers, 1978).
4. Ewing, B.B., "Probabilistic Consideration in Water Quality Management," *Water Wastes Eng.* 7:50 (1970).