Springer Series in Statistics

Advisors:
P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

Springer Series in Statistics

Alho/Spencer: Statistical Demography and Forecasting.
Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes.
Atkinson/Riani: Robust Diagnostic Regression Analysis.
Atkinson/Riani/Cerioli: Exploring Multivariate Data with the Forward Search.
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
Borg/Groenen: Modern Multidimensional Scaling: Theory and Applications, 2nd edition.
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.
Bucklew: Introduction to Rare Event Simulation.
Cappé/Moulines/Rydén: Inference in Hidden Markov Models.
Chan/Tong: Chaos: A Statistical Perspective.
Chen/Shao/Ibrahim: Monte Carlo Methods in Bayesian Computation.
Coles: An Introduction to Statistical Modeling of Extreme Values.
David/Edwards: Annotated Readings in the History of Statistics.
Devroye/Lugosi: Combinatorial Methods in Density Estimation.
Efromovich: Nonparametric Curve Estimation: Methods, Theory, and Applications.
Eggermont/LaRiccia: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation.
Fahrmeir/Tutz: Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edition.
Fan/Yao: Nonlinear Time Series: Nonparametric and Parametric Methods.
Farebrother: Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume I: Two Crops.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume II: Three or More Crops.
Ferraty/Vieu: Nonparametric Functional Data Analysis: Models, Theory, Applications, and Implementation
Ghosh/Ramamoorthi: Bayesian Nonparametrics.
Glaz/Naus/Wallenstein: Scan Statistics.
Good: Permutation Tests: Parametric and Bootstrap Tests of Hypotheses, 3rd edition.
Gouriéroux: ARCH Models and Financial Applications.
Gu: Smoothing Spline ANOVA Models.
Györfi/Kohler/Krzyżak/Walk: A Distribution-Free Theory of Nonparametric Regression.
Haberman: Advanced Statistics, Volume I: Description of Populations.
Hall: The Bootstrap and Edgeworth Expansion.
Härdle: Smoothing Techniques: With Implementation in S.
Harrell: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.
Hart: Nonparametric Smoothing and Lack-of-Fit Tests.
Hastie/Tibshirani/Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Hedayat/Sloane/Stufken: Orthogonal Arrays: Theory and Applications.
Heyde: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation.

(continued after index)

Nhu D. Le James V. Zidek

Statistical Analysis of Environmental Space-Time Processes

Springer

Nhu D. Le
British Columbia Cancer
Research Center
675 West 10th Avenue
Vancouver V5Z 1L3
Canada
nle@bccrc.ca

James V. Zidek
Department of Statistics
University of British Columbia
333-6356 Agricultural Road
Vancouver V6T 1Z2
Canada
jim@stat.ubc.ca

Library of Congress Control Number: 2005939015

ISBN-10: 0-387-26209-1
ISBN-13: 978-0387-26209-3

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springer.com

To Lynne, Hilda, Adrian, and Megan

Preface

This book presents knowledge gained by the authors along with methods they developed, over more than 30 years of experience measuring, modeling, and mapping environmental space–time fields. That experience embraces both large (continentwide) spatial domains and small. In part it comes from their research, working with students as well as coinvestigators. But much was gained from all sorts of interactions with many individuals who have had to contend with the challenges these fields present. They include statistical as well as subject area scientists, in areas as diverse as analytical chemistry, air sampling, atmospheric science, environmental epidemiology, environmental risk management, and occupational health among others. We have collaborated and consulted with government scientists as well as policy-makers, in all, a large group of individuals from whom we have learned a lot and to whom we are indebted. We hope all in these diverse groups will find something of value in this book. We believe it will also benefit graduate students, both in statistics and subject areas who must deal with the analysis of environmental fields.

In fact we have given a successful statistics graduate course based on it. The book (and course) reflect our conviction about the need for statistical scientists to learn about the phenomena they purport to explain. To the extent feasible, we have covered important nonstatistical issues involved in dealing with environmental processes. Thus in writing the book we have tried to strike a balance between important qualitative and quantitative aspects of the subject. Much of the most technical statistical-mathematical material has been placed in the starred sections, chapters, and appendices. These could well be skipped, at least on first reading. In fact the simplest path to that technical material would be through Chapter 14; it contains a more-or-less self-contained tutorial on methods developed by the authors. That tutorial relies on R software that can be downloaded by the interested reader.

When we started analyzing environmental processes, we soon came to know some of the inadequacies of geostatistical methods. These purely spatial methods had been around for a long time and proven very successful in

geostatistical application. Thanks to the SIMS group at Stanford they had even been appropriated in the 1970s for use in analyzing ozone space–time fields. However, the acid rain fields that were the initial focus of our study involved multivariate responses with up to a dozen chemical species measured at a large number of sites over a broad spatial domain. Moreover, it became clear that while these responses could be transformed to have an approximately normal distribution, their spatial covariances were far from stationary, a condition of fundamental importance in classical geostatistics. The failure of that assumption led Paul Sampson and Peter Guttorp to their discovery of an elegant route around that assumption (Chapter 6). The need to handle multivariate responses and reflect our considerable uncertainty about the spatial covariance matrix led us to our hierarchical Bayes theory, the subject of Chapters 9 and 10. Chapter 9, the simplified version, conveys the basic elements of our theory.

Chapter 10 presents the fully general (multivariate) theory. It incorporates enhancements made over time to contend with difficult situations encountered in applications. The last published extension appeared in 2002. Additional theory was developed for the book. To avoid excessive technicality, we have given much of the detail in the Appendices.

The theory in that chapter really provides the "engine" that drives our model and applications in Chapters 11–13. Chapter 11 uses that engine to drive a theory for designing networks for monitoring environmental processes, one of the most difficult challenges facing environmentricians. Other challenges are seen in Chapter 12 where the important topic of environmental process extremes is visited. In spite of their immense importance in environmental risk analysis, this topic has received relatively little emphasis in environmetrics. In contrast, the topic of Chapter 13, environmental risk, has been heavily studied. Our contributions to it, in particular, to environmental health risk analysis appear there.

The novelty of the methods emphasized in this book has necessitated the development of software for implementation. Sampson and Guttorp developed theirs for covariance modeling and we have incorporated a version of it in ours. Although our research group developed the code needed to implement our multivariate theory, that code has been greatly refined thanks to the substantial contributions of our colleague and sometime research partner, Rick White.

Although the book features a lot of our own methods and approaches, we try to give a reasonably comprehensive review of the many other, often ingenious approaches that have been developed by others over the years. In all cases we try to indicate strengths and limitations. An extensive bibliography should enable interested readers to find out more about the alternatives.

To conclude, we would like to express our deepest appreciation to all who have helped us gain the knowledge reflected in this book. Our gratitude also goes to those who helped implement that knowledge and develop the tools we needed to handle space–time fields. That includes our many co-authors,

including former students. A special thanks goes to Bill Caselton who first stimulated the second author's interest in environmental processes, and to our long time research compatriots, Peter Guttorp and Paul Sampson for a long and fruitful collaboration as well as for generously allowing us to use their software. John Kimmel, Springer's Executive Editor–Statistics, and several anonymous reviewers have provided numerous thoughtful comments and suggestions that have undoubtedly improved the book's presentation. The Copy-Editors, Valerie Greco and Natacha Menar were superb. Part of the book is based on work done while the second author was on leave at the University of Bath and later at the Statistical and Applied Mathematical Science Institute; both generously provided facilities and support. The Natural Sciences and Engineering Research Council of Canada (NSERC) has been a constant source of funding, partially supporting our research developments described in this book. Finally, we thank our wives, Hilda and Lynne for their support and patience throughout this book's long gestation period. Without that this book would certainly not have been written!

Vancouver, British Columbia — *Nhu D Le*
March 2006 — *James V Zidek*

Contents

Part I: Environmental Processes

1

First Encounters. . .

It isn't pollution that's harming the environment. It's the impurities in our air and water that are doing it.
Dan Quayle

If you visit American city,
You will find it very pretty.
Just two things of which you must beware:
Don't drink the water and don't breathe the air.
Tom Lehrer

This book concerns the "impurities" described by Dan Quayle that worry Tom Lehrer, the degree to which they are present, and the amount of harm they are causing.

1.1 Environmental Fields

On a fine summer day Vancouver's air seems clear and free of pollution. In contrast, looking east towards Abbotsford, visibility is obscured by a whitish haze that can sometimes be very thick.

That haze comes in part from Vancouver since the prevailing winds of summer transport pollution in that direction. However, at any location in an urban area, the air pollution field is a mix of "primary" and "secondary" pollutants. Local sources might include such things as automobile exhaust pipes, industrial chimneys, oil refineries, and grain storage elevators. They are commonly products of combustion. Examples would include SO_2 (sulfur dioxide) and CO (carbon monoxide). In contrast, secondary pollutants take time to form in the atmosphere and be transported to a given location, i.e., site. They come from complex photochemical processes that take place during the period of transport. Sunshine and humidity help determine the products.

These processes are not very well understood, making the forecasting of air pollution difficult. In any case, secondary pollutant fields unlike their primary cousins tend to be fairly "flat" over large urban areas. The fields also change over time.

We have introduced space–time fields with the example above because of its societal importance. Indeed, fields such as this are primary objects of study in the subject of environmental risk assessment. To quote from the Web page of the U.S. Environmental Protection Agency (http://www.epa.gov/air/concerns/):

> Breathing air pollution such as ozone (a primary ingredient in urban smog), particulate matter, carbon monoxide, nitrogen oxides, and lead can have numerous effects on human health, including respiratory problems, hospitalization for heart or lung disease, and even premature death. Some can also have effects on aquatic life, vegetation, and animals.

Indeed, the relationship between acute and chronic nonmalignant pulmonary diseases and ambient air pollution is well established. Increases in the concentration of inhalable particles (airborne particles with a diameter of no more than 10 micrograms, commonly known as PM_{10}) in the atmosphere have been associated with acute decrements in lung function and other respiratory adverse effects in children (Pope and Dockery 1992; Pope et al. 1991). There is evidence that mortality from respiratory and cardiac causes is associated with particle concentrations (Schwartz and Dockery, 1992). Increases in concentrations of ambient ozone have been associated with reduced lung function, increased symptoms, increased emergency room visits and hospitalizations for respiratory illnesses, and possibly increased mortality. This extensive literature has been reviewed by Lippman (1993) and Aunan (1996). The evidence for other chronic diseases, except lung cancer, seems far less conclusive, reflecting the limitations of most studies, particularly the inadequate characterization of air pollution exposure. Good estimates of cumulative exposure often require concentration levels at too many locations to be feasibly monitored and hence such fields need to be mapped using what little information is available.

Space–time fields such as that described above are generally viewed as "random" and described by probabilistic models, paradoxically, a view that is not inconsistent with physical laws. These laws are not fully understood. Moreover, although existing knowledge can be brought into the prediction problem through deterministic models, those models will involve a large number of constants (parameters) that need to be estimated to a high level of accuracy. Data of a requisite quality for that purpose may not be available. Finally, these models will require initial conditions specified to a level of accuracy well beyond the capabilities of science. Thus, although the outcome of say, the toss of a die is completely determined by deterministic laws of nature, these laws are of no more help now than they were, at the time of the Romans at least, for predicting that outcome. Hence, probability models are used for that purpose instead. [The interested reader should consult the entertaining book by Stewart (1989) for a discussion of such issues in a broader context.]

The reader may well wonder how the outcome of an experiment such as tossing a die can be regarded as both determined and random. Moreover, given that we are tossing that die just once, how can the probability of an "ace" be 1/6 since according to the repeated sampling school of statistics, finding it requires that we repeatedly toss the die in precisely the same manner, over

and over, while tracking the ratio of times an "ace" appears to the number of tosses. Good question!

It might be partly answered for the die experiment in that we can at least conceive of an imaginary experiment of repeated tosses. However, in our air pollution example, the thought of calculating probabilities by repeated "tosses" would strain the imagination. We would be even more challenged to provide a repeated sampling interpretation of probability for a field such as the concentration of a mineral under the earth's crust. That concentration would remain more-or-less constant over time, an important special case of the space–time model studied in the subject of geostatistics. More is said about such constant fields in Chapter 7.

A wholly different way of interpreting such probabilities underlies the theory in this book. That interpretation, found in the Bayesian paradigm, takes *probability* to represent *uncertainty*. Briefly, 1/6 would represent our fair odds of 5:1, that an ace will not occur on the toss of die.

In general, the uncertainty we have about random phenomena such as air pollution fields can be reduced through the acquisition of new information. This information can come through measurement and the analysis of the data the measurements provide. (See Chapter 11.)

However, measurement itself is subject to uncertainty. That uncertainty derives in part from inevitable error no matter how expensive the instrument. Some of it could be due to such things as misrecording or misreporting. An extreme form arises when data are missing altogether. In our air pollution example, the data can be missing because the motor in a volumetric sampler that sucks air through a filter breaks down.

A more pervasive error derives from the fact that the measurements may be mere surrogates for the real thing. For example, the concentration of SO_2 ($\mu g/m^3$) is measured through its fluorescent excitation by pulsed ultraviolet light. Measurement of O_3 (ppb) is based on the principle of the absorption of ultraviolet light by the ozone molecule. Uncertainty now resides in the exact relationship between the measurements and the thing being measured. In any case, all such uncertainty can in principle be expressed through probability models within the Bayesian framework, although finding those probabilities can involve both conceptual and technical difficulties.

The air pollution example has a number of other features commonly associated with the monitoring of space–time fields. For one thing, the random field can readily be transformed to have a joint Gaussian distribution. In fact, the logarithmic transformation often works for air pollution and there are substantive reasons for this fact.

The space–time fields seen in practice usually have regional covariates associated with them that vary with time. Time $= t$ itself may be regarded as such a covariate and in that case a simple trend line, $a + b \times t$ may be viewed as a fixed component of the responses to be measured. In fact, the coefficients a and b for this line might depend on site but since a and b will need to be estimated and the data are not usually too plentiful, a high cost

can be attached to adding so many parameters into the model. Indeed, the uncertainty added in this way may outweigh any gains in precision that accrue from making the model site-specific. The same can be said for other covariates based on time such as $\sin(t)$ and $\cos(t)$ which are commonly incorporated into the model to capture seasonality.

Quite different covariates are associated with meteorology. Temperature, humidity, as well as the easterly and northerly components of wind are examples. In the latter case, one might expect to see significant site to site variation over a region, so ideally these should be included as responses rather than as covariates to serve as predictors of the space–time fields responses. Indeed, the wind itself generates a space–time field of independent interest.

That field is the subject of the unpublished study of Nott and Dunsmuir (1998) about wind patterns over the Sydney Harbor. Their data come from 45 monitoring stations in the Sydney area and the study was undertaken in preparation for the Sydney Olympics (although the authors do not describe how their analysis was to be used).

Wind, like most commonly encountered fields, involves multivariate responses, i.e., responses (measured or not), at each location that are vectors of random variables. A lot is lost if the coordinate responses are treated separately, since the opportunity is lost to "borrow" information in one series to help make inferences about another.

Fields such as those described above have been regularly monitored in urban areas. Hourly measurements may be reported for some pollutants such as PM_{10}, Daily measurements are provided for others such as $PM_{2.5}$, a fraction of PM_{10}. There may be as many as say a dozen monitoring sites for a typical urban air basin but some pollutants may be measured at only a subset of these sites owing to technical limitations of the instruments used.

To fix ideas consider the comparatively simple network of 20 continuous ambient air quality monitoring stations maintained by the Greater Vancouver Regional District (GVRD; see the GVRD 1996 Ambient Air Quality Annual Report, http://www.gvrd.bc.ca/air/bro/aqanrep.html). Those stations transmit hourly data to an Air Quality Monitoring System computer database. Local air quality can then be compared against national and provincial guidelines. [We refer to locations (e.g., building rooftops) of ambient monitoring stations as *gauged sites*. Numerous other sites are potentially available for creating other stations. We call them *ungauged sites*.]

Each of the 20 gauged sites in the GVRD network has seven positions at which monitors or gauges could be installed, one for each of the seven fields being measured (e.g., sulphur dioxide SO_2 $\mu g/m^3$). As a purely conceptual device for explaining our theory we call the positions with monitors gauged pseudo-sites.

The data collected by the monitoring networks often have data missing for what might be termed structural reasons. In the example above, sites or quasi-sites were set up at different times and operated continuously thereafter. We see an extended analysis of monitoring data collected in just such a situation in

the next chapter. This situation leads to a monotone data pattern resembling a staircase. The top of the lowest step corresponds to the most recent start-up. The tops of the steps above, are for successively earlier starts.

Structurally missing data obtain when not all gauged sites measure the same suite of responses. In other words, not all the gauged sites have their gauges at the same quasi-sites and hence they do not collect the same data. In fact, systematically missing data of this form can emerge because monitoring networks are a synthesis of smaller networks that were originally designed for quite different purposes. Zidek et al. (2000) describe an example of such a network that provides measurements for a multivariate acid deposition field. That network in southern Ontario consists of the union of three monitoring networks established at various times for various purposes: (1) OME (Environment Air Quality Monitoring Network; (2) APIOS (Air Pollution in Ontario Study); (3) CAPMoN (Canadian Acid and Precipitation Monitoring Network described by Burnett et al. 1994).

As a brief history, both APIOS and CAPMoN were established with the initial purpose of monitoring acid precipitation, reflecting concerns of the day (see Ro et al. 1988 and Sirois and Fricke 1992 for details). In fact, CAPMoN with just three sites in remote areas began monitoring in 1978. 1983 saw an increase in its size when it merged with the APN network to serve a second purpose, that of finding source–receptor relationships. In the merged network monitoring sites could be found closer to urban areas. A third purpose for the network was then identified and it came to be used to find the relationship between air pollution and human health (Burnett et al. 1994; Zidek et al. 1998a,b).

The merged network now monitors hourly levels of nitrogen dioxide (NO_2 $\mu g/m^3$), ozone (O_3 ppb), sulphur dioxide (SO_2 $\mu g/\ m^3$) and the sulfate ion (SO_4 $\mu g/m^3$).

New features of importance continually arise and the Bayesian framework provides the flexibility needed to incorporate those features in a conceptually straightforward and coherent way. Thus, even among adherents of the repeated sampling school, the hierarchical Bayesian model has gained ground albeit disguised as something called the random effects model.

One of these new features arises when the various items in a space–time field are measured at differing or even misaligned scales. For example, some could be daily levels while others are hourly. Or some could be at the county and some at the municipal level even though say the latter were of principal interest. Fuentes and Smith refer to this feature as a change of support in an unpublished article entitled "A New Class of Nonstationary Spatial Models." That feature has become the subject of active investigation. In fact, Fuentes and Smith cite Gelfand et al. (2000) as having independently studied this feature. Much work remains to be done.

Another such feature of considerable practical importance sees both systematically missing gauges at some of the quasi-sites as well as a staircase data pattern over time. We know of no altogether satisfactory approach to

analyzing such data. In fact, it remains very much a research area at the time this book was written.

To conclude this section we describe two other examples of space–time fields in different contexts. Again the features and the problems alluded to in this section are applicable to these examples.

1.1.1 Examples

Example 1.1. Wildcat drilling in Harrison Bay
In this example, the environmental risk is ascribed to oil and gas development on the Beaufort Sea continental shelf just off the north coast of Alaska (Houghton et al. 1984). A specific response of interest was the concentrations of benthic organisms in the seabed. These "critters" form the lowest rung of the food chain ladder that eventually rises to the bowhead whale, a part of the Innuit diet. Thus, their survival was deemed vital but possibly at risk since, for example, the mud used for drilling operations, containing a number of trace metals, would be discharged into the sea.

The statistical problem addressed in this context was that of testing the hypothesis of no change in the mean levels over time of these concentrations at all sites in the seabed extending east from Point Barrow to the Canadian border. Moreover, little background data on this field were available, pointing to the need to sample the seabed before and after exploration at judiciously selected sites. Thus, the testing problem gave way to a design problem: where best to monitor the field for the intended purpose. This type of design is often referred to as the BACI (before-and-after-control-impact) design. The problem was compounded by the shortage of time before exploration was to commence, combined by the vastness of the area, the pack ice which could interfere with sampling, the high costs involved, and finally, the unpredictability of the location of the environmental impact of the drilling mud if any.

The latter depended on such things as the winds and the currents as well as the ice, all in an uncertain way. The approach proposed by the second author of this book depended on having experts from Alaska divide the area to be sampled into homogeneous blocks according to their estimates of the likelihood of an impact on the mean field if any. This could then be incorporated as a (prior) distribution in conjunction with a classical F-test of no time–space interaction, based on the before and after measurements to be taken.

This proved an effective design strategy and led to an extension by Schumacher and Zidek (1993). That paper shows among other things, that in designing such experiments, one should place the sampling points in just the regions where the likely impact is thought to be highest and lowest (to maximize the contrast in the interaction being tested). Moreover, the points should be equally divided. That seems to go against the tendency of experimenters to place their sampling points in the region of highest likely impact. The reasoning: why waste sampling points where there is little possibility of an impact? A little thought shows this reasoning to be naive, although seductive, since the

baseline levels against which impact can be measured need to be established using the *quasi-control* sites.

Example 1.2. The Rocky Mountain Arsenal
An unusual environmental field that changes little over time these days can be found at the Rocky Mountain Arsenal (RMA). This example shows the great importance that can attach to spatial mapping and large scales on which this sometimes has to be done.

A Web page maintained by the Program Manager RMA (PMRMA) and the Remediation Venture Office (RVO) of the RMA (http://www.pmrma-www.army.mil/htdocs/misc/about.html) reveals that the RMA is an 27 square mile area near Denver, Colorado. Furthermore, the pamphlet, "The Rocky Mountain Arsenal Story", published by the Public Affairs Office of Commerce City, Colorado states that starting in 1942, chemical weapons were manufactured there. After the Second World War, the need for weapons declined and some of the property was leased to the Shell Chemical Company in the 1950s whereupon the manufacture of pesticides and herbicides commenced. At the same time, the production of chemical weapons declined, ending altogether in 1969.

Throughout the site's active period, wastes were dumped in a natural basin on the site (see the PMRMA/RVO page cite above). However, those wastes leaked into the groundwater supply used for irrigating crops, leading inevitably to crop damage.

Consequently, most of the RMA was placed on the National Priorities List (NPL) in the 1987–89 period. It then became subject to the Comprehensive Environmental Response, Compensation and Liability Act of the United States This has led to a cleanup operation under the so-called Superfund program with the eventual goal of turning this area into a wildlife refuge.

According to an EPA Web page, (http://www.epa.gov/region08/superfund/sites/rmasitefs.html)

> Most of the health risks posed by the site are from: aldrin, dieldrin, dibromochloro-propane (DBCP), and arsenic. Aldrin is a pesticide that breaks down to dieldrin. Both chemicals are stored in the body and affect the central nervous system and liver. DBCP is also a pesticide, but it is not stored in the body. DBCP can affect the testes, kidneys, liver, respiratory system, central nervous system and blood cells. Arsenic is a naturally occurring element. It can cause cancer in humans.

In short, nasty stuff!

The (multivariate) response of interest in this situation would be the vector of concentrations of these hazardous agents over a variety of media such as groundwater and soil. However, a statistical question now arises. How much of the RMA was actually contaminated and in need of cleanup? Since, according to a Defense Environmental Restoration Program report cited on the EPA's

home page (http://www.epa.gov/swerffrr/ffsite/rockymnt.htm), the total cost of cleanup might come to well over 2 billion U.S. dollars, substantial savings could be realized by minimizing that estimate. Thus, in the early 1990s the second author came to serve on a tribunal convened to hear arguments from stakeholders on various sides of this question, for a variety of estimates that had been made.

While the details of this hearing are confidential, the dispute involved the spatial contamination field itself. In particular, soil samples had been taken at a number of sites and analyzed for the Chemicals of Concern (COC's) as they are called. The goal was a map of the area, giving predicted concentrations of these COCs based on the data obtained at the sampling sites. The cleanup would then be restricted to areas of highest contamination. Finally, the tribunal and no doubt many other dispute resolution mechanisms, eventually led, in 1995 as well as 1996 to the signing of two historic agreements or Records of Decision as they are called, by the Army, Shell, the Service, the Colorado Department of Public Health and Environment, and the U.S. Environmental Protection Agency. These provided a comprehensive plan for the continuation of the very expensive cleanup of the RMA. We show methods in later chapters that enable predictions such as this to be made.

Incidentally, mapping the spatial contamination field proved to be complicated by missing data, much of it being BDL (below the detection limit). These are concentrations so small they "come in under the radar" below the capacity of the measurement process to measure them to an acceptable degree of accuracy. More appallingly, a lot of the concentrations were also ADL, much to the detriment of the environment!

We begin with groundwork needed for modeling environmental space–time fields.

1.2 Modeling Foundations

Random space–time fields represent processes such as those in the examples above. Space refers generically to any continuous medium, that unlike time, is undirected. It could refer to the demarcated area of seabed in Example 1.1, for example, or to a region of the earth's surface as in Example 1.2. However, it could also refer to a lake where toxic material concentration might be the response of interest, or even to a space platform where vibration is of concern.

Subregions of the earth's surface are commonly two-dimensional domains, with points indexed by latitude and longitude, or even UTM (Universal Transverse Mercator) coordinates. (The latter, unlike the former, do not suffer the shortcoming of lines of longitude, that distances between them grow smaller near the poles.) Alternatively, they can be of higher dimensions then two as when elevation is included and we have a three-dimensional domain for our process.

1.2.1 Space–Time Domains

To describe spatial or more generally space–time processes we need a set of coordinates, say $\mathcal{I}$, to mark points in that space. In practice, $\mathcal{I}$ is taken to be finite although conceptually it is a continuum. This restriction greatly simplifies the problem from a technical perspective because then the field associated with it assumes values on a finite-dimensional rather than infinite-dimensional domain. We also avoid the need to describe small scale dependence, something that cannot be realistically done because of the complexity of most space – time processes.

1.2.2 Procedure Performance Paradigms

However, before leaving this issue, we must emphasize for completeness that one performance paradigm sometimes invoked in geostatistics for assessing procedures requires this label set to be a continuum. To expand on this point, recall that all statistical performance paradigms assume hypothetical situations, "test tracks" as it were, wherein statistical procedures must perform well to be considered acceptable. The choice of which paradigms to invoke is pretty much subjective. The repeated sampling paradigm is an example. To increase their confidence in the quality of a result, some analysts require good repeated sampling properties even when applying a procedure just once.

The large-sample paradigm is another, usually invoked in conjunction with the repeated sampling paradigm. Here, not only will sampling be repeated infinitely often but each sample will be infinitely large. How different from the situation ordinarily encountered in statistical practice!

Different situations can lead to different implementations of the large-sample paradigm. For example, time-series analysts suppose they are observing a curve (called a sample path) at timepoints separated by fixed intervals. The repeated sampling paradigm here refers to drawing curves such as the one being observed at random from a population of curves. For any fixed timepoint, say t_0, their inferential procedure might, for example, be an estimator of a population parameter such as the population average, $\mu = \mu(t_0)$, of all those curves. Such procedures of necessity rely on the measurements from just the single curve under observation; good repeated sampling properties are required under an assumption about the curves called *ergodicity* (that is of no direct concern here). The large-sample paradigm invoked in this context assumes an infinite sequence of observation times, separated by fixed intervals, that march out to infinity. The performance of procedures for inference about the population parameters such as μ can now be assessed by how well they do with this infinite sequence of observations under the repeated sampling paradigm above.

Nonparametric regression analysts invoke a different version of this paradigm. They also suppose they are observing a curve at specified sampling points, this time in a bounded range of a predictor such as time. However,

their curve is supposed to be fixed, not random, and their repeated sampling paradigm posits observation errors randomly drawn from a population of measurement errors. At the same time, the large-sample paradigm assumes measurements are made at successfully denser collections of sampling points in the range of the predictor. Thus, measurements are made at successively finer scales until, in the limit, the infinite number of points is obtained in that bounded range.

These two implementations of the repeated, large-sample paradigms differ greatly even when invoked in precisely the same context, observations of a curve measured at a collection of sampling points. So which would be appropriate, if either, for space–time processes? After all, the marker, i, could be regarded a "predictor" of the value of the field's response. Yet at the same time, our process could be considered a time-series where the curve is that traced out by an random array evolving over time.

In search of an answer, suppose that the field remains constant over time (or equivalently, that it is observed at a single timepoint). We then find ourselves in the domain of geostatistics, a much studied subject. There the field, like the curve of time-series, is considered random. Yet, a large-sample paradigm commonly used in this situation is that of nonparametric regression which assumes an ever more dense sequence of sampling points (Stein, 1999).

The reader could be forgiven for feeling somewhat confused at this point. Alas, we have no advice to offer. These different, seemingly inconsistent, choices above reflect two different statistical cultures that have evolved in different subdisciplines of statistics.

1.2.3 Bayesian Paradigm

In this book, we are not troubled by this issue, since we adopt the Bayesian paradigm. Thus, in the sequel, unknown or uncertain means random. Moreover, probabilities are subjective. In other words, the probability that an uncertain object X falls in an event set, A, $P(X \in A)$, means, roughly speaking, fair odds of $P(X \in A) \times 100$ to $[1 - P(X \in A)] \times 100$ that A occurs.

We assume a fixed index set $\mathcal{I}$ (represented by $= 1, \ldots, I$ for simplicity), while automatically acquiring performance indices for procedures that evolve out of succeeding developments. Incidentally, little attention seems to have been given to the problem of how big we can make $\mathcal{I}$ before reaching the point of diminishing returns. (We show implications of this choice in Chapter 4.) In practice, we have been guided by practical considerations. For example, in health impact analysis, the centroids of such things as census subdivisions seem appropriate since that is the level of aggregation of the health responses being measured.

Similar considerations pertain to $\mathcal{T}$ (represented as $1, \ldots, T$ for simplicity), the timepoints indexing the field. Again, this could be taken to be a continuum but is usually taken to be a finite set. Its elements may represent hours, days, weeks, or even years. It should be emphasized that, unlike space, time

is directional so cannot be regarded as another spatial coordinate (except superficially). Moreover, that special quality of time also provides a valuable structure for probabilistic modeling.

1.2.4 Space–time Fields

Finally, we are led to formulate the random space–time response series (vectors or matrices) needed for process modeling. In environmental risk assessment we may need up to three such objects, $\mathbf{X}_{it}$, $\mathbf{Y}_{it}$, and $\mathbf{Z}_t$, $t \in \mathcal{T}$, at each location $i \in \mathcal{I}$ and each time $t \in \mathcal{T}$. The $\mathbf{Y}$-process may be needed to represent the adverse environmental impact. To fix ideas, $\mathbf{Y}_{it}$ may denote the number of admissions on day t to hospital emergency wards of patients residing in region i who suffered acute asthma attacks. The $\mathbf{X}$-process can represent a real or a latent (unmeasured) process, the latter being purely contrived to facilitate modeling the $\mathbf{Y}$-process. In the example $\mathbf{X}_{it}$ might represent the ambient concentration of an air pollutant on day t in region i. Finally, the $\mathbf{Z}$-process may represent covariates that are constant over space for each timepoint; these covariates represent such things as components of time, trend, and environmental factors that affect all sites simultaneously. In the example $\mathbf{Z}_t$, $t \in \mathcal{T}$ might be the average daily temperature for the area under study on day t. A model for risk assessment might, for example, posit that the conditional average of $\mathbf{Y}_{it}$ given $\mathbf{X}_{it}$ and $\mathbf{Z}_t$, i.e., $E[\mathbf{Y}_{it} \mid \mathbf{X}_{it}, \mathbf{Z}_t]$ is given by $g(\mathbf{X}_{it}, \mathbf{Z}_t)$ for a specified function g.

1.3 Wrapup

This chapter has summarized the features of space–time response fields likely to be encountered in practice. Moreover, we have presented a number of illustrative examples of importance in their own right. Through these examples we have tried to show the great diversity and importance of the problem of mapping and measuring space–time fields. Finally, we have laid the foundations for an approach to modeling environmental space–time processes. We discuss the modeling of these processes in more detail below in Chapters 5, 9, and 10. However, modeling requires measurements, to which we turn in Chapter 4.

However, to make the ideas in this chapter more concrete, we describe in the next, a worked-out application in detail. We also demonstrate the kinds of analyses that can be done with the methods developed in this book along with the associated software.

2

Case Study

For the first time in the history of the world, every human being is now subjected to contact with dangerous chemicals, from the moment of conception until death.
Rachel Carson, Silent Spring, 1962

In this chapter, we illustrate methodology developed in this book by describing an application involving one of the chemicals Carson refers to above. Specifically, we describe a study of BC ozone data made by Le et al. (2001, hereafter LSZ). That illustration shows among other things, how to hindcast (or backcast) data from a space–time field. By this we do not mean, the opposite of forecast. Rather LSZ reconstruct unobserved historical responses through their relationship with other series that had been observed. Those are ozone levels from stations that started up at earlier times in the staircase of steps we described in Section 1.1. But they could have used any other available series such as that from temperature that might be correlated with the ozone series.

By looking ahead to Chapter 13, we can get a glimpse of the purpose of hindcasting the data, namely, environmental health impact assessment. To be more precise, LSZ require the hindcasted field for a case-control study of the possible relationship between cancer and ozone. Cancer has a long latency period and over that period the subjects would have moved occasionally from one locale to another. Their exposures to ozone would therefore have varied according to the levels prevailing in those different residential areas. However, not all those areas would have had ozone monitors, especially in the more distant past, since interest in this gas tends to be of recent vintage. The solution adopted by LSZ backcasts the missing values in historically unmonitored regions from observed values in those that were monitored. In this way, the required exposure could be predicted for the case-control study.

2.1 The Data

The monthly average ozone levels used came from 23 monitoring sites in the Province of British Columbia. These sites are listed in Figure 2.1. Averages were calculated from hourly values provided by the BC Ministry of Environment. To do so, LSZ first discarded days with fewer than 18 hourly reported

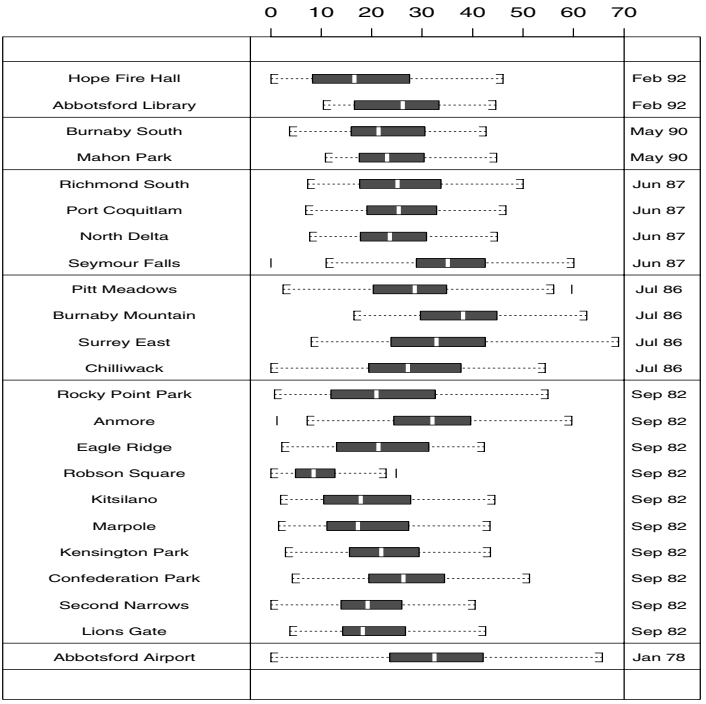

Fig. 2.1: Boxplot of monthly average ozone levels at 23 monitor sites in British Columbia and their start-up times.

values. Then daily and in turn, monthly averages were computed. That produced 204 monthly averages beginning with January, 1978 through December, 1994.

LSZ grouped stations with the same starting times beginning in 1978. The locations of these sites are shown in Figure 2.2.

2.2 Preliminaries

LSZ next transformed the data to achieve a more nearly Gaussian distribution, finding the logarithm to be suitable for this purpose.

In addition to observed responses, here log-transformed monthly values, the theory offered in Chapters 9 and 10 also allows covariates to be admitted. While these covariates may vary with time, they must be constant across space. (If they did vary across space, they could be included in the response

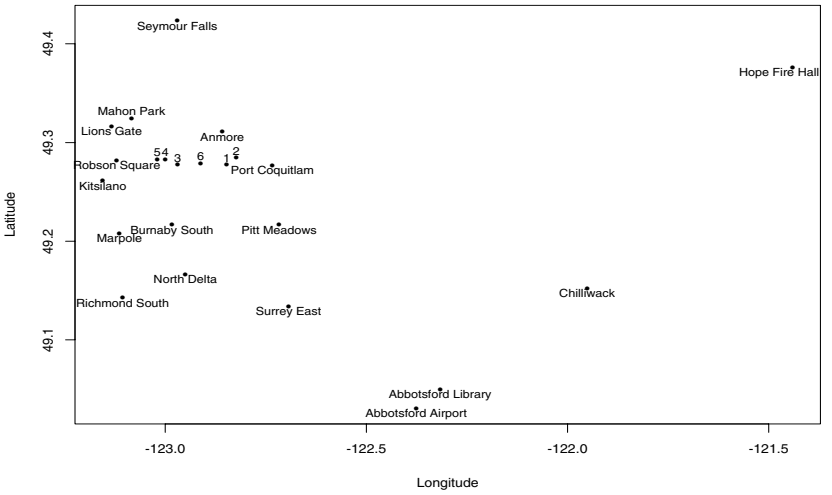

Fig. 2.2: Ozone Monitoring Sites (1 - Rocky Point Park; 2 - Eagle Ridge; 3 - Kensington Park; 4 - Confederation Park; 5 - Second Narrows; 6 - Burnaby Mountain).

vector!) LSZ adopt $Z = [1, \cos(2\pi t/12), \sin(2\pi t/12)]$ as the covariate vector. This means that

$$Y_{it} = \beta_{i0} + \beta_{i1}\cos(2\pi t/12) + \beta_{i2}\sin(2\pi t/12) + \epsilon_{it},$$

where $(\epsilon_{1,t}, \ldots, \epsilon_{23,t})$ are residuals, assumed to be independent over time and follow a Gaussian distribution with mean 0 and variance Σ (see Chapter 5 or Appendix 15.1 for a definition).

By modeling the shared effects of covariates i.e., trends in this way, LSZ are able to eliminate both temporal and spatial correlation that might be considered spurious. In other words, they remove associations over time and space that could be considered mere artifacts of confounding variables (the covariates) rather than due to intrinsic relationships. By subtracting the estimated trend from the Ys, the analysis can turn to an analysis of the residuals and a search for those associations. The trends are added back in at a later stage as necessary.

The fits of the model to the data shown in Figure 2.3 for a typical site point to a very strong yearly cycle.

That figure also depicts the partial autocorrelation function (pacf) for the series of transformed monthly averages. The pacf for lag 2, for example, shows the degree of linear correlation of current monthly values with that of two-months-ago, once the effect of last month has been factored out. In other words, if the pacf between the current month's value and its two-

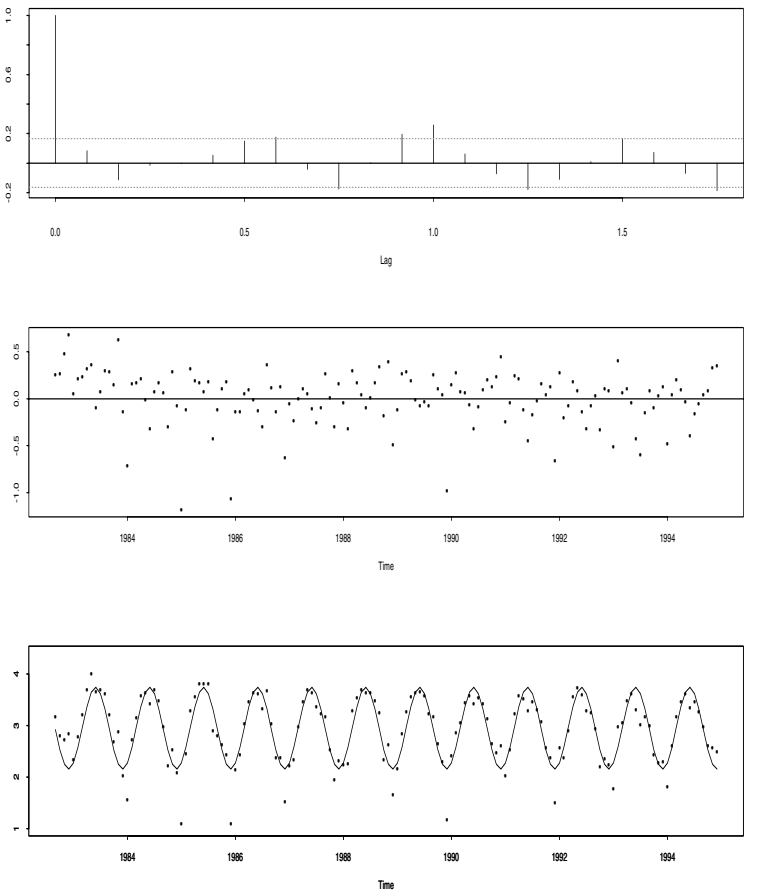

Fig. 2.3: Trend modeling: Upper: partial autocorrelation function; Middle: residual plot; Lower: fitted trend and observations.

month-old cousin were large, it could not simply be due to their both having been strongly associated with the value for one month ago (that has effectively been removed). The results suggest we may for simplicity adopt the assumption that these monthly values are independent of each other, since for the Gaussian distribution being uncorrelated means being completely independent. The analyst will not usually be in such a fortunate position as this!

2.3 Space–time Process Modeling

LSZ were now in a position to apply the theory developed in Chapters 9 and 10 of this book, using the trend model specified above. They began by grouping stations with the same starting time as follows.

Block 1: two sites, start-up time: February 1992
Block 2: two sites, start-up time: May 1990
Block 3: four sites, start-up time: June 1987
Block 4: ten sites, start-up time: July 1986
Block 5: four sites, start-up time: September 1982
Block 6: one site, start-up time: January 1978

Next comes the estimation of special parameters, called hyperparameters. These parameters, unlike say Σ above, are found not in the distribution that describes the distribution of the sample values directly, but rather they are parameters in the prior structure. The latter provides a distribution on the first-level parameters like Σ and express LSZ's uncertainty about them. (See Chapter 3. Recall, that in the Bayesian paradigm, all uncertainty can in principle be represented through a probability distribution.) It turns out these parameters can be estimated from the data. To do so, they used a standard method called the EM algorithm.

With their hyperparameters estimated, LSZ are able to turn to the development of a predictive distribution, i.e., a distribution for the unmeasured responses of interest.

LSZ require both the interpolation of the field's values at completely unmonitored sites as well as hindcasted values at those currently monitored. The predictive distribution allows for not only the imputation of these unmeasured values, but as well, the construction of say 95% prediction intervals. Figure 2.4 shows the hindcasted ozone levels and the 95% predictive intervals of the Burnaby Mountain station. To obtain the prediction intervals, LSZ simulate realizations of the field from the predictive distribution. They do this with subroutines available in standard libraries using the matric-t distributions, characterized in Appendix 15.1, that constitute the predictive distributions.

2.4 Results!

The predictive intervals between January 1978 to September 1982 proved to be large. That is hardly surprising. Only one block of stations (Block 1) was in operation. Those between September 1982 to July 1986 turned out to be smaller since by that time two blocks (Blocks 1+2) were in operation. More data were now available on which to base hindcasting.

Getting predictive distributions for ungauged sites presents a new obstacle. Whereas LSZ were able to use the EM algorithm to get estimates of hyperparameters, specifically the hypercovariance, for hindcasting, now they have to