
Ja

# Statistical

of Enviro

# 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 $\mu$ 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

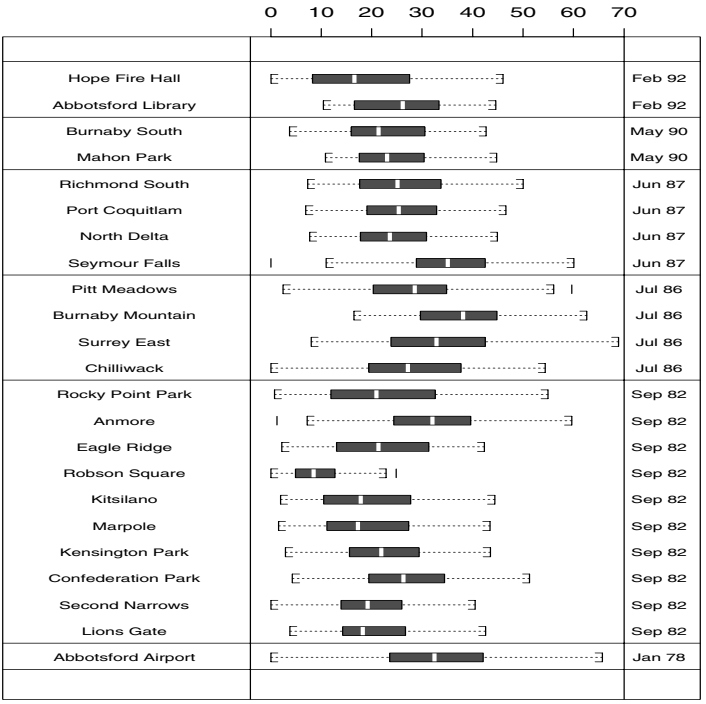


Fig. 2.1: Boxplot of monthly average ozone levels at 23 monitor sites in British Columbia and their start-up times.

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

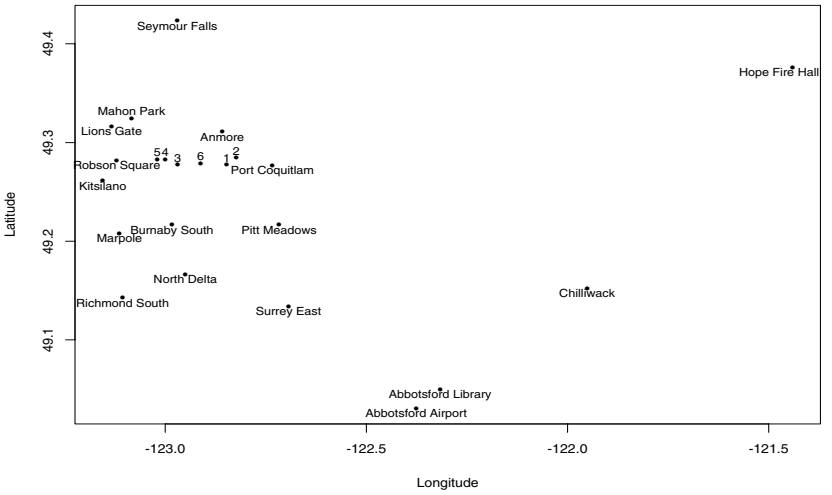


Fig. 2.2: Ozone Monitoring Sites (1 - Rocky Point Park; 2 - Eagle Ridge; 3 - Kensington Park; 4 - Confederation Park; 5 - Second Narrows; 6 - Burnaby Mountain).

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 $\Sigma$ (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 $Y$s, 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-

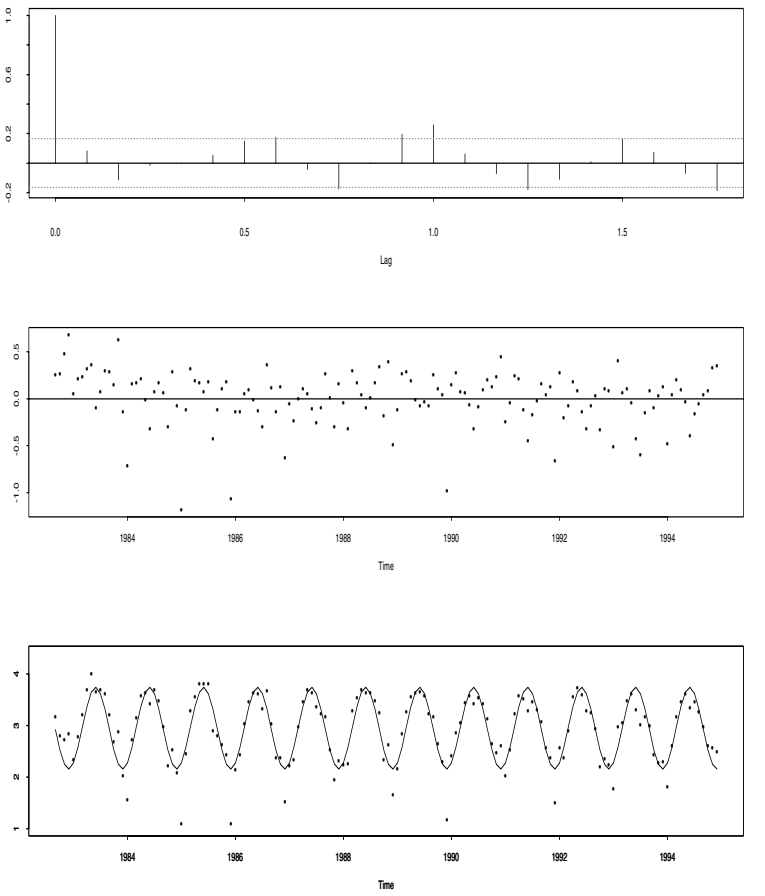


Fig. 2.3: Trend modeling: Upper: partial autocorrelation function; Middle: residual plot; Lower: fitted trend and observations.

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 $\Sigma$ 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 $\Sigma$ 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

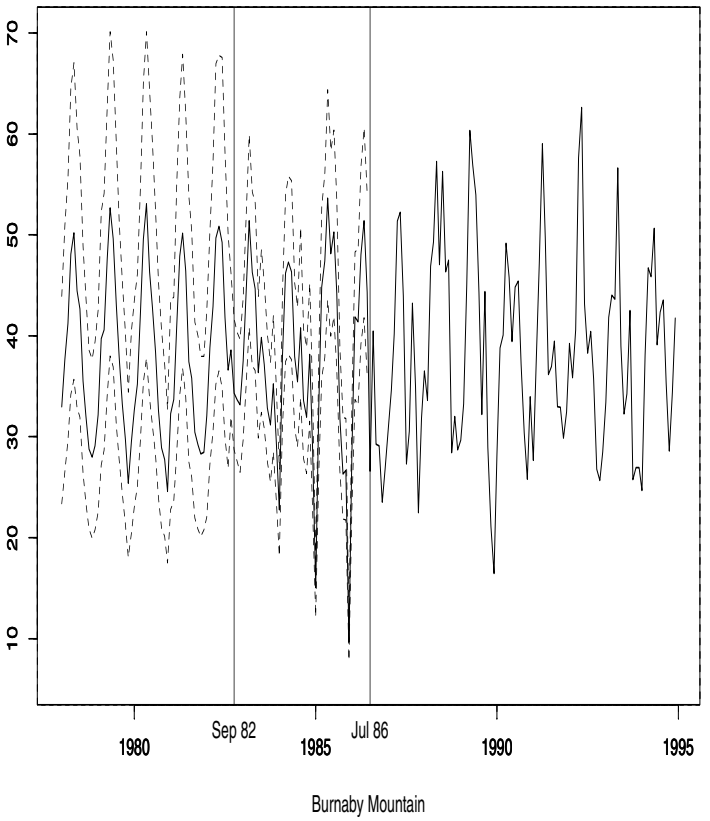


Fig. 2.4: Observed monthly average ozone levels in $\mu g/m^3$ at the Burnaby Mountain station between July 1986 and December 1994. Hindcasted values (solid) and corresponding 95% predictive intervals (dash) between January 1978 and June 1986. Vertical lines indicate when blocks are formed during this hindcasting period.

find hypercovariances between sites of interest, for which no data have ever been obtained!

LSZ find a way to do this by means of a method proposed by Sampson and Guttorp (1992) as described in Chapter 6. Briefly, the Sampson–Guttorp (SG) method provides a way of estimating spatial covariances for fields, such as those encountered in air pollution, where the degree of association (correlation) between sizes does not necessarily decline as a function of the distance between them. In fact, many other environmental factors, depending on the

context, such as elevation or salinity can have a more significant role to play in determining that association than distance. The method is something of a major breakthrough, since spatial prediction had been dominated for so many years by the methods inherited from geostatistics. There the assumption of spatial isotropy seems to be tenable even though it can fail dramatically for space–time fields.

The SG method first finds functions that connect the coordinates of locations in the *geographic plane (G-space)*, say latitude and longitude, with locations in an imaginary new, *dispersion space (D-space)*, created so that association is a decreasing function of distance. The creation of that function depends on the estimated associations between existing monitoring sites. A fitted variogram, or equivalently, the correlation function, in the D-space and the estimated mapping function are then used to obtain spatial correlations between all locations of interest.

Figure 2.5 shows the results LSZ obtain by using the method. On the right of that figure, we see the D-space coordinates obtained by applying the estimated function to a G-space grid. On the left we see a fitted variogram in the D-space. The results can be used to estimate spatial correlations between any points in the G-space. One simply identifies the D-space coordinates for any pair of sites. Then one measures the D-space distance between them. Finally, one plugs that distance into the fitted variogram to estimate their spatial correlations.

The construction of the functions that connect D- and G-spaces depends on a *smoothing parameter* which determines how much G can be changed in getting to D. At one extreme, D would be identical to G. At the other, G could be a grossly distorted version of G. Users can ensure that the G-grid is not folded in the D-space and hence maintain the spatial interpretability of the correlations. In other words, the closer the stations are the higher their correlations.

Figure 2.6 depicts 1994's predicted monthly average ozone levels (in $\mu g/m^3$). Notice the distinct annual cycles. The lowest level seems to have been in December while the highest comes in June. Predicted concentrations near the monitoring stations are strongly influenced by those stations and form the "mountain tops" (respectively, "valley floors") seen in the figure. Moving away from monitoring sites we see the surface of predicted concentrations decrease (respectively, increase) towards a regional average.

These trends would be expected and, in fact, they represent a *regression towards the mean* effect. That effect is seen even in the simplest linear predictor in ordinary regression analysis. That isn't to say the true surface trends down (up, respectively) in that way. It may go down or up in reality. However, as the information in the data becomes increasingly irrelevant, the model loses its predictive accuracy and the best prediction increasingly becomes the sample average concentration in a manner of speaking.

This phenomenon reveals the particular difficulty attached to the spatial prediction of extreme values in the field, for example, the maximum concen-

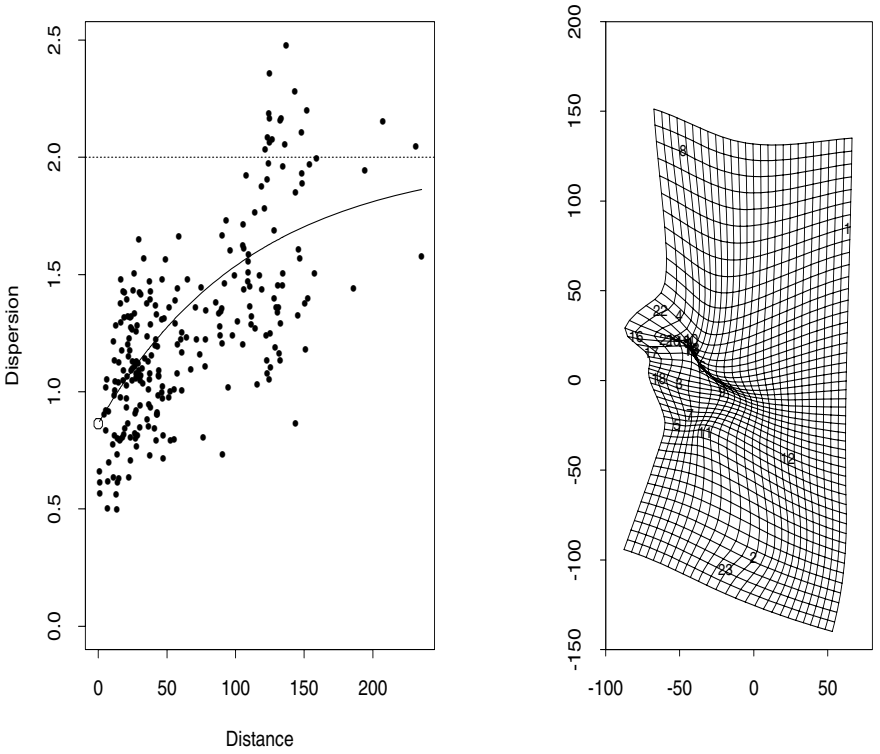


Fig. 2.5: Variogram fits and D-space coordinates with smoothing parameter = 0.3 used in the mapping function.

tration over all locations in the region. That maximum, which might well be of importance in epidemiological work or regulation, would undoubtedly occur somewhere other than the gauged stations. Yet, the concentration predictor in Figure 2.6 would be biased away from this maximum, a tail value, towards the mean, a more central value

However, the predictive distribution recognizes its own limitations in that as we move away from a gauged site the 95% prediction interval, or equivalently, its standard deviation, increases. To see this explicitly, turn to Figure 2.7 for June 1994. There we see again the features discussed above for Figure 2.6, perhaps even more clearly, this time in a contour plot. But this time LSZ also contour plot, in Figure 2.8, the standard deviations of the predictive distribution.

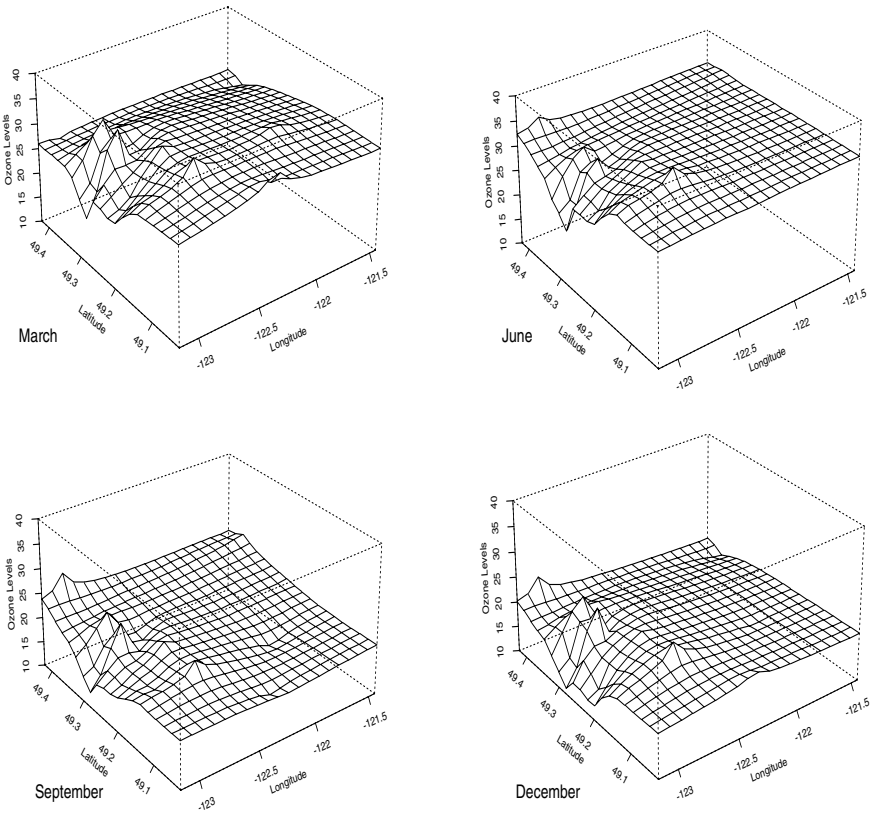


Fig. 2.6: Interpolated monthly average ozone levels ($\mu$g/m$^3$) in 1994.

Figure 2.8 reveals what we theorized above about the standard deviations. In fact as we leave the region and move northwest or southeast we see that standard deviation increasing quite dramatically. This tells us that the tails of the predictive distribution grow fatter as we move away from the gauged sites. What is not known at this time is how well the extreme quantiles predict extreme values for the field.

With the application, we have tried to introduce the reader to the subject of mapping space–time fields. We have indicated the importance of such mapping in terms of estimating human exposure to pollution fields in the setting

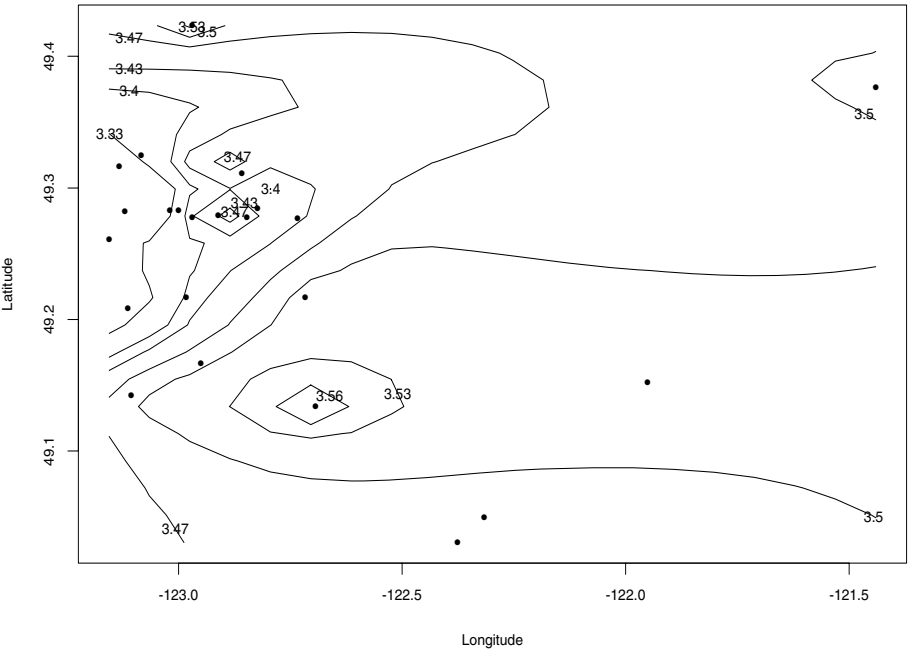


Fig. 2.7: Contour plot of the interpolated ozone field ($\mu$g/m$^3$) in June, 1994. The dots denote the stations.

of a case-control study for investigating the relationship between cancer and air pollution. Now in the next chapter, we turn to the subject of uncertainty that underlies all statistical modeling.

## 2.5 Wrapup

This chapter shows space–time modeling theory in action. With the help of the models, much can be accomplished. Unmeasured historical process concentrations can even be inferred for the purpose of assessing environmental risk. In general, uncertainty about the process can both be reduced and quantified.

Uncertainty underlies all of statistics and science itself. So it deserves special consideration and in the next chapter we investigate that elusive commodity and get a better understanding of its role in our theory.

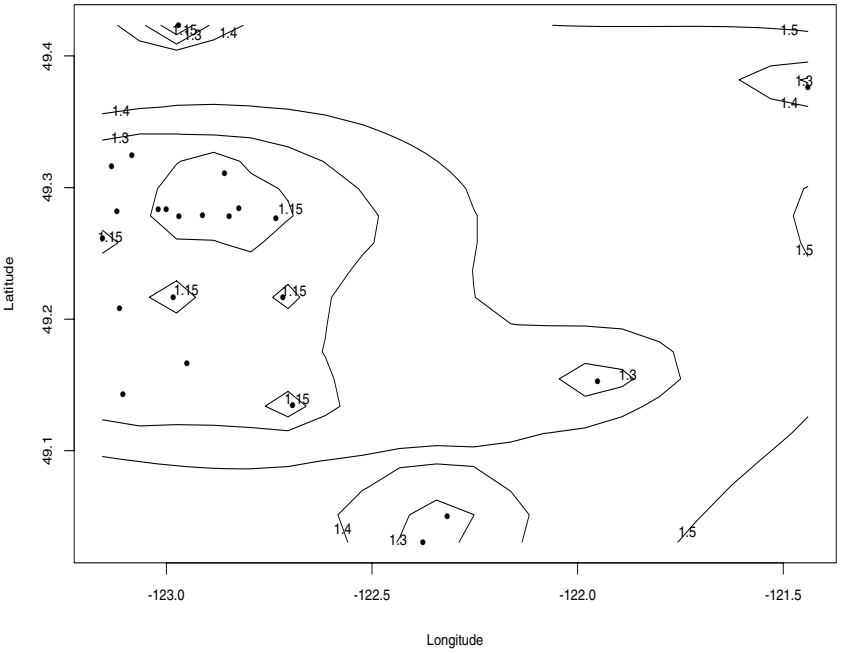


Fig. 2.8: Contour plot of the standard deviations for the interpolated values on a log-scale (ln $\mu$g/m$^3$) in June, 1994.

# 3

# Uncertainty

> *. . . there are unknowns; there are knowns that we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And. . . it is the latter category that tend to be the difficult ones.*
>
> *Donald Rumsfeld 2003.*

However elusive in its meaning, uncertainty about unknowns seems awfully important; van Eeden and Zidek (2003) found in the Science Citation Index about 30,000 articles citing it as a keyword. More importantly for us, it is the "stuff and trade" of environmental risk analysts who need to communicate it, manage it, quantify it, reduce it, and interpret it. We devote this chapter to it, describe a language for it, and present ways of measuring it. This chapter paves the way for the more focused development that follows.

## 3.1 Probability: "The Language of Uncertainty"

Bernardo and Smith (1994) describe uncertainty as "incomplete knowledge in relation to a specified objective." Frey and Rhodes (1996) say much the same: "uncertainty arises due to a lack of knowledge regarding an unknown quantity." But how should it be quantified? Lindley (2002), citing Laplace, provides an answer: "Probability is the language of uncertainty." O'Hagan (1988) more emphatically asserts that "Uncertainty is probability."

Within statistics, the Bayesian paradigm, embraced by the above quotations of Lindley as well as O'Hagan and developed below, seems ideal for discussing uncertainty (and more generally for risk assessment). There, roughly speaking, uncertainty is equivalent to randomness and the degree of uncertainty about any aspect of the world, past, present, and future, can be expressed through a probability distribution. That is the paradigm on which this book is based.

Therefore our discussion focuses on an uncertain, and hence random object of interest, $Y$, that could be a matrix, vector, or even real number. It might even involve unknown parameters in conventional modeling terminology. For example, $Y = I_{\{H\}}$ could be the indicator for a hypothesis $H$—some statement about the world—that is 1 or 0 according as $H$ is true or false, Then, although finer taxonomies are available, uncertainty can be dichotomized into either

*aleatory* (stochastic) or *epistemic* (subjective) (Helton 1997). The first obtains when $Y$'s probability distribution is known, the latter, when it is not. In fact, the latter represents the added uncertainty about $Y$ due to ignorance of $Y$'s distribution. Parameter and model-uncertainty can be considered forms of epistemic uncertainty.

Generally in this book, $Y$ represents a column vector, partitioned into $Y^{unmeas}$ and $Y^{meas}$. The first component represents things that cannot be measured. Model parameters might be included there. The realizations of a random environmental field at unmonitored locations might be put there as well. In contrast, the second component contains $Y$'s measurable but as yet unmeasured attributes. When measured, that component gets replaced by $y^{meas}$, this being a nonrandom realization of its uncertain (random) counterpart, $Y^{meas}$. The latter unlike the former has a probability distribution and is susceptible to repeated measurement in some cases.

## 3.2 Probability and Uncertainty

How can probability be used to represent an individual's uncertainty? To answer this question, we begin with a simple case, explored further in the next section, where $Y$ is again an indicator variable, say $Y = I_{\{X \in A\}}$ where $X$ is another random object, and $A$ a subset of its range of possibilities. Then uncertainty about $Y$ (and $X$) can be interpreted as the individual's willingness to give odds of $P(X \in A)/P(X \in \bar{A}) : 1$ in favor of $X$ being in $A$, $\bar{A}$ denoting the complement of $A$ (the event $X$ is not in $A$). Very large (respectively, small) odds would represent states of near certainty that $X$ is in $A$ (respectively, $X$ is not in $A$). However, 1:1 odds [when $P(X \in A) = 1/2$] represent the state of greatest possible uncertainty. This interpretation of probability conforms to its use in ordinary language.

However, representing an individual's uncertainty by means of a probability distribution in complex situations proves more challenging (O'Hagan 1998). In every case, eliciting the (joint) probability distribution is always the goal. Suppose that distribution has a joint probability density function (PDF) so that

$$P(Y \in C) = \int_C f(y^{unmeas'}, y^{meas'}) dy^{unmeas'} dy^{meas'}, \tag{3.1}$$

for every subset B where $'$s have been attached in the integrand to emphasize that the variables appearing there are merely dummy variables of integration. (The case of discrete random vectors would be handled in an analogous manner using a probability mass function instead of a PDF and summation instead of integration.)

The joint density function for $Y$ in Equation (3.1) is commonly found using the multiplication rule of probability:

$$f(y^{unmeas'}, y^{meas'}) = f(y^{meas'} \mid y^{unmeas'}) \pi(y^{unmeas'}). \tag{3.2}$$

In Equation (3.2) the density $\pi$ derives from the so-called *prior* or more formally, *a priori* ("before experience") distribution given by Equation (3.3):

$$P(Y^{unmeas} \in D) = \int_D \pi(y^{unmeas'}) dy^{unmeas'}. \tag{3.3}$$

This distribution is supposed to come purely from the individual's prior knowledge (and not from the data). In some cases, he will be an "expert" with considerable relevant experience. Then $\pi$ may be easy to elicit and welldefined. On the other hand, a novice would be forced to choose a vague prior with $\pi \approx 1$. In that case, $P$ defined by Equation (3.3), will be unbounded and not a probability distribution in which case the prior is called improper. Nevertheless, such distributions can be and are often used in Bayesian analysis (subject to certain restrictions indicated below). However, very un-Bayesian-like behavior can then ensue (Dawid et al. 1973), a fact that is generally ignored.

Once $Y^{meas}$ has been measured and $y^{meas}$ obtained, prior knowledge can be updated with the new information by means of the celebrated rule of Rev. Thomas Bayes. The result is the *posterior*, or more formally, *a posteriori* (after experience) distribution given by

$$\begin{aligned} P(Y^{unmeas} \in D \mid Y^{meas} = y^{meas}) = \int_D & f(y^{meas} \mid y^{unmeas'}) \\ & \pi(y^{unmeas'}) \\ & \times dy^{unmeas'} / f(y^{meas}), \end{aligned} \tag{3.4}$$

where

$$f(y^{meas}) = \int f(y^{meas} \mid y^{unmeas'}) \pi(y^{unmeas'}) dy^{unmeas'} \tag{3.5}$$

gives the so-called marginal density of $Y^{meas}$.

Equation (3.5) embraces an easily discovered, beautiful feature of the Bayesian approach, that the prior can be sequentially updated in a stepwise fashion as the measured data arrive, The prior for the next step is just the posterior from the previous step! Moreover, as long as the posteriors remain proper, the initial prior could well be improper. The unwary at least, would see no distinguishable difference between proper and improper.

Equation (3.5) has another distinguished feature, the so-called *likelihood function*,

$$f(y^{meas} \mid y^{unmeas'}), \tag{3.6}$$

regarded as a function of $y^{unmeas'}$. This function, one of Sir Ronald Fisher's great contributions to statistical theory (though he was not an adherent of the Bayesian school), allows the data to attach their measure of relative importance to the various hypothetical possibilities for $Y^{unmeas}$, $y^{unmeas'}$. The most plausible one would maximize the likelihood. (Where a model parameter is involved, the result is called a *maximum likelihood estimate*.)

Strangely, this measure of relative importance, unlike the prior density function, does not integrate to 1. Of course, sometimes the prior density is conceptualized as a likelihood function derived from a prior realization of $Y^{meas}$, say $y^{meas}_{prior}$. However, in this case, $\nu$ is taken to be a pristine prior representing a state of complete ignorance and the actual prior becomes on applying Equation (3.5) a purely hypothetical posterior distribution that came out of an earlier step in the analysis. (Such priors, called conjugate, are often mathematically convenient, if not always realistic choices.) Thus, the distinction made above between likelihood and prior still obtains.

That in turn implies the nonuniqueness of the likelihood. Any positive multiple would yield the same posterior. More substantively easily constructed examples show that seemingly different process models for generating $y^{meas}$ yield likelihoods that are positive multiples of one another and hence the same posterior.

Although probability gives us a language in which to discuss uncertainty, it does not directly quantify it except in simple cases. We turn to that issue next.

## 3.3 Uncertainty Versus Information

We commonly speak of "decreasing" (or "increasing") uncertainty, suggesting the existence of a quantitative ordering. Moreover, we see increasing information as a way of making that reduction. This section examines these fundamental things.

To initiate our search for such a quantitative measure of uncertainty, we return to the example introduced in the previous section where $Y = I_{\{X \in A\}}$. O'Hagan's definition above seems to point unambiguously in this simple case to $P(X \in A)$ as the appropriate measure for the individual whose uncertainty is being assessed.

Even though simple, that case allows us to see if additional information always leads to a decrease in uncertainty. The affirmative answer we might naively expect stems from our view of uncertainty and information as complementary cousins—the more of one the less of the other. Indeed, both Shannon (1948)and Renyi (1961) interpret entropy, defined below, as exactly equal to the amount of information contained in $Y$ that would be released by measuring $Y$ exactly.

However, van Eeden and Zidek (2003) show the answer can be negative. They suppose the individual learns $X \in B$. The new measure of uncertainty becomes $P(X \in A \mid X \in B)$. The result can be near 1/2, representing the state of complete uncertainty where originally that probability is near 0 (or 1), a state of near certainty about $Y$. We demonstrate this in the following example.

*Example 3.1. Information increases uncertainty*

Suppose an individual assesses her uncertainty about $X$ as represented by a uniform probability on [0,1], denoted $X \sim U[0,1]$. In particular, this individual would be quite certain (90% sure) that $X$ does not lie in $A = [0, 2/20]$. However, that individual learns that $X$ lies in $B = [1/20, 2/20]$. Using a standard formula for conditional probability the individual finds that now $P(X \in A \mid X \in C) = 1/2$ even though $P(X \in A) = 1/10$ originally. The individual who was nearly certain $X$ would not lie in $A$, is now completely uncertain about that issue.

We would emphasize that in Example 3.1 the individual's uncertainty about $X$ was aleatory so the apparent anomaly does not derive from subtleties associated with epistemic uncertainty. So why did the additional information increase rather than decrease uncertainty? The answer is that the additional information contradicted the individual's prior views about the uncertain $X$'s being in $[0, 1/10]$.

Clearly this phenomenon must be quite pervasive although it does not seem to have been much studied. In any case, the example has two important general implications: (1) some kinds of information may increase rather than decrease uncertainty; (2) any satisfactory measure of uncertainty must admit this phenomenon in similar circumstances.

### 3.3.1 Variance

So what other measures of uncertainty might be used? Two very common ones are *variance* and *entropy* (van Eeden and Zidek 2003). When $Y$ is a real-valued random variable, its variance is defined by

$$Var(Y) = E(Y - \mu_y)^2, \tag{3.7}$$

where, in general, for any function $h$ of $Y$, $E[h(Y)] = \int h(y')f_Y(y')dy'$, $f_Y$ denotes the density of $Y$'s distribution and $\mu_Y$ denotes the expectation of $Y$, $E(Y)$. $\sigma_Y = \sqrt{Var(Y)}$ represents $Y$'s standard deviation (SD), an alternative measure of $Y$'s uncertainty that has the advantage of being on the same scale as $Y$ and hence more easily interpreted than $\sigma_Y$. When $Y$ is a vector, the covariance obtains as the natural extension of the variance:

$$\Sigma_Y = E(Y - \mu_y)(Y - \mu_y)^T \tag{3.8}$$

when $Y$ is a column vector say. Here the expectation $E$ of the random matrix is computed elementwise. However, since the covariance is a matrix rather than a numerical measure, its determinant, called the generalized variance, is sometimes used as an ad hoc measure (see Example 3.2).

The variance or the conceptually equivalent SD has proven popular. Thus, for example, reporting the standard error (SE) in a reported statistical estimator to indicate uncertainty has become nearly universal in scientific reporting. Moreover, as noted by van Eeden and Zidek (2003), the United State's National Institute of Standards and Technology advocates

use of the standard deviation. To quote from the Institute's 2001 Web page (http://physics.nist.gov/cuu/Uncertainty/basic.html) "Each component of uncertainty, However, evaluated, is represented by an estimated standard deviation, termed **standard uncertainty** . . . ."

However, the variance and SD have shortcomings. These include their nonexistence for some distributions, lack of invariance under scale changes (suggesting they are not measures of intrinsic uncertainty in $Y$), undue sensitivity to the weight in the tails of the distribution, and lack of a natural multivariate extension. The lack of such an extension proves problematic when $Y$ is a vector of two independent random coordinates. One might expect their combined uncertainty to be the sum of their individual uncertainties, but the variance offers no way of expressing that fact unlike the next contender we present for the role of uncertainty measure.

### 3.3.2 Entropy

Harris (1982) suggests another generally accepted measure, the entropy of $Y$'s distribution: $H(Y) = E(-\log f_Y(Y)/m_Y(Y))$ where $f_Y$ denotes the probability density of $Y$ (with respect to counting measure in the discrete case). Here $m_Y$, with the same units as $f_Y$ [i.e., probability $\times$ ($Y$'s unit of measurement)$^{-1}$], plays the role of a reference density against which uncertainty about $Y$ is to be measured. For simplicity, we take $m = 1$ (with appropriate units) as is commonly done (Singh, 1998, p. 3).

The next example demonstrates the computation of an entropy for a distribution of central importance to this book.

*Example 3.2. Gaussian entropy*
Assume that $Y : p \times 1$ has a multivariate Gaussian distribution with mean $\mu_Y$ and covariance $\Sigma_Y$, written $Y \sim N_p(\mu_Y, \Sigma_Y)$. This means that for every row vector, $a : 1 \times p$, $aY \sim N(a\mu_Y, a\Sigma_Y a')$. When $\Sigma_Y$ has full rank $p$ this last definition is equivalent to

$$f_Y(y') = \sqrt{2\pi}^{-1}|\Sigma_Y|^{-1/2}exp[-2^{-1}(y' - \mu_Y)^T\Sigma_Y^{-1}(y' - \mu_Y)],$$

where the superscript $T$ means the transpose of the vector (or more generally, matrix) to which it is attached. The entropy is

$$\begin{aligned} E[-\log f_Y(Y)] &= E[\frac{p}{2}\log\sqrt{2\pi} + \frac{1}{2}\log|\Sigma_Y| + \\ &\qquad \frac{1}{2}(Y - \mu_Y)^T\Sigma_Y^{-1}(Y - \mu_Y)] \\ &= \frac{p}{2}\log\sqrt{2\pi} + \frac{1}{2}\log|\Sigma_Y| + \frac{p}{2}. \end{aligned}$$

The last step relies on the result,

$$E[(Y - \mu_Y)^T\Sigma_Y^{-1}(Y - \mu_Y)] = ETr[\Sigma_Y^{-1}(Y - \mu_Y)(Y - \mu_Y)^T] = Tr(I_p) = p,$$

with $Tr$ denoting the trace operator that just computes the sum of the diagonal elements of the matrix upon which it acts. We have used the identity $TrAB = TrBA$ for all matrices $A$ and $B$ of appropriate dimension.

Example 3.2 demonstrates the close connection between the generalized variance and entropy, thus endowing the latter with credentials as a measure of uncertainty, at least when $Y$ has a multivariate Gaussian distribution. However, we should emphasize that the entropy is a unitless quantity inasmuch as we have adopted a reference density of $m_Y \equiv 1$ ($f_Y$ units) before setting out on this calculation.

We can extend the definition of the entropy to the case of conditional distributions in a natural way. For example, $H(Y|A)$ could be the entropy for the conditional density $f(y'|A) = f(y')/P(A), y' \in A$ where $P(A) = P(Y \in A)$. More generally, $H(Y|X = x)$ could be used to denote the entropy in the light of the information contained in the knowledge that $X = x$. In Chapter 11, we use more elaborate versions of the entropy to help us decide where to locate new sites for monitoring an environmental space–time process.

The good news: both variance and entropy are flexible enough as to admit the phenomenon described above, as shown by van Eeden and Zidek (2003). However, these authors have also shown that even seemingly simple questions for these measures can prove quite challenging and many remain unanswered.

*Example 3.3. Behavior of entropy*
Suppose $X$ has a normal distribution with mean $\mu$ and SD $\sigma$; that is, $X \sim N(\mu, \sigma^2)$. Furthermore, suppose $B = [-b, b]$. Then it is natural to ask if the conditional variance of $X$, $Var(X \mid X \in B)$, is a monotone increasing function of $b$. Naively, one might expect an affirmative answer. However, the results above would suggest that at least if $\mid \mu \mid$ were large, the answer should be negative. The truth remains unknown.

## 3.4 Wrapup

Example 3.3 demonstrates that implications of using the variance and entropy as measures of uncertainty remain to be worked out. At the same time, we know of no realistic alternative candidates for that role.

In the next chapter we begin to address the problem of assessing environmental risk by investigating the processes used to generate the requisite data through the process of measurement.

# 4

# Measurement

*Errors using inadequate data are much less than those using no data at all.*

*Charles Babbage*

With infinite resources, we could eliminate all uncertainty about a random response field. For one thing, we could measure it completely, without error. Alternatively, we could build a perfect process model and predict the field, making measurement redundant. In the end compromise is needed. The goal is to make the data at least adequate for the task they are asked to do.

In reality, measuring environmental processes, in particular establishing a network of measurement sites, can be expensive. For example, the U.S. National Surface Water Survey was conducted to assess the degree of acidification of U.S. lakes (see for example, Eilers et al. 1987). That stratified random sample survey entailed costly visits to lakes, the collection of water samples, and their analysis in a timely fashion, all with requisite data quality assurance.

Thus a combination of deterministic and stochastic modeling seems a tempting cheap alternative to making more than just a minimal number of measurements. Uncertainty about the remaining (unmeasured) responses can then be reduced by fitting and using the model to predict them.

However, the resulting model would fall short of a complete process model. Hence its predictions would be subject to both the aleatory and epistemic uncertainties introduced in Chapter 3, even if that model were valid and perfectly fitted. In practice, additional uncertainty would derive from imperfectly fitted model parameters (and incurring uncertainty due to measurement or sampling error). Moreover, the validity of the model itself would be in doubt. The result would involve epistemic uncertainty in the terminology of Chapter 3.

The principles determining the optimal trade-off between measurement and modeling seem to be unknown. Strong adherents of the measurement school argue against models, deeming them to be subjective and hence biased. Indeed, modeling is individualistic and depends on such things as context and an investigator's background knowledge. The process may even reflect discipline bias. Thus, an atmospheric scientist might approach the problem of modeling an air pollution field quite differently than a statistician, for example. In fact, each may regard the efforts of the other as "naive." In the face of such dilemmas, proponents of measurement seem to have a case.

However, difficulties arise there as well. Measurement also proves to be highly individualistic. The choice of what to measure, say hourly or daily averages, may be dictated by such things as the supposed purposes of making the measurements or the state of an individual's knowledge about the health impact of an environmental hazard. Selection of the measuring device also entails individual choice and such devices may vary widely in their *validity* (the degree to which they measure what they are supposed to measure) and their *reliability* (the accuracy with which they measure whatever they measure). More fundamentally, Heisenberg's uncertainty principle tells us that the act of measurement may affect the process being measured and hence the outcome itself.

Fortunately, within the Bayesian paradigm we adopt, the issues raised above do not pose a problem in principle. That paradigm is founded on a subjectivist framework and so encompasses both the individualism of measurement as well as that of modeling. We discuss these things in more detail in the sequel.

This section reviews measurement issues in the sampling of space–time processes that cannot be ignored. References are given to more comprehensive and detailed sources of information. The first step in measurement involves where and when to measure an environmental space–time process. That is the subject of the next section, given in more detail in Chapter 11.

## 4.1 Spatial Sampling

Spatial sampling (or more generally environmental sampling) networks monitor space–time fields. Their designs require among other things: (1) measurement methods including associated devices, (2) data handling and quality management protocols, (3) methods for sample analysis; (4) methods for data capture as well as storage, and (5) the designation of space–time sampling points.

Their purposes may be temporary. Example 1.1 gives an example of a plan for temporarily measuring benthic sediments (biomarkers) using grab-samplers. Those measurements were made to study the effects of oil exploration in the Beaufort Sea. Another was opportunistic. A network was set up to take measurements that bracketed the closure of a smelter in the state of Washington. The inferential goal: an estimate of the difference in regional pollution levels. Many such networks measure before-and-after concentration levels in media surrounding a new industrial facility.

### 4.1.1 Acid Precipitation

However, spatial designs are commonly considered permanent as illustrated by acid deposition monitoring networks. (We see below that "permanency" can be illusive!)

Although commonly referred to as "acid rain," in fact, acid deposition can come in the form of snow or even fog. (Dry deposition of the precursors of acid rain occurs as well and networks measure them.) Complex environmental processes produce it from gaseous emissions such as sulfites $SO_2$ (smelters and power plants) and oxides of nitrogen $NO_x$ (industry, traffic, and power stations). Eventually as these gases are transported in the atmosphere they are converted to sulfuric and nitric acid, $H_2SO_4$ and $HNO_3$. In a water solution (rain), these acids generate hydrogen ions ($H^+$), the more acid molecules the more hydrogen ions.

Thus the most common measure of the acidity of a liquid is its concentration of $H^+$ ions or more precisely its "pH" value, the negative logarithm of that concentration. High acidity means low pH levels and sour tasting water. According to the United States's Environmental Protection Agency (EPA) (see http://www.epa.gov/airmarkets/acidrain/), the most acid rain falling in the UNITED STATES during 2000 had a pH of about 4.3, above the level of lemon juice (about 2.0), but below that of pure water (about 7.0).

Acidic deposition is a global environmental hazard due to the long range atmospheric transport of local emissions, sometimes thousands of kilometers. These very aggressive acids damage forests, soil, fish, materials, and human health. Thus networks for monitoring acid deposition (dry and wet) were established long ago.

In the United States, the Clean Air Status and Trends Network (CASTNET) and the National Atmospheric Deposition Program (NADP) were developed to monitor and measure dry and wet acid deposition, respectively (http://www.epa.gov/castnet/overview). Their (mainly rural) monitoring sites help determine the associations between pollution and deposition patterns. Both NADP and CASTNET yield data relevant to the health of ecosystems and they provide background pollution levels. Researchers and policy analysts use these data to investigate: (1) environmental impacts and (2) nonecological impacts of air pollution including reduced visibility as well as damage to materials, especially those of cultural and historical importance.

In the late 1970s, the NADP initiated a cooperative program among various agencies to determine geographical patterns and trends in U.S.precipitation chemistry. Under the impetus of the National Acidic Deposition Assessment Program (NAPAP) in the 1980s, the size of the monitoring network grew rapidly to eventually include more than 200 sites positioned as shown in Figure 4.1 constructed from data on the NADP's Web site.

Weekly wet deposition samples began to be collected in 1978. Eventually, the NADP network evolved into the NADP/NTN (National Trends Network). That network measures: (1) constituents of precipitation chemistry affecting rainfall acidity and (2) those that may have ecological effects. More precisely, the NADP Web site provides measurements for concentrations of calcium, magnesium, potassium, sodium, ammonia, nitrate, chlorine, sulfate, and pH.

As an illustration, Figure 4.2 compares the measured chlorine concentrations in acid deposition for Colorado and Maine. Maine is close to seawater

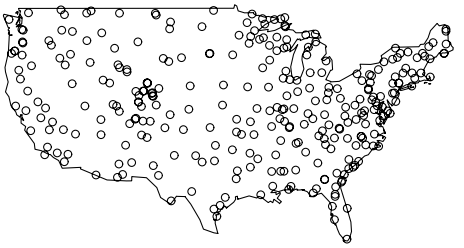

**Fig. 4.1.** Locations of NADP/NTN acid precipitation monitoring sites in 2004.

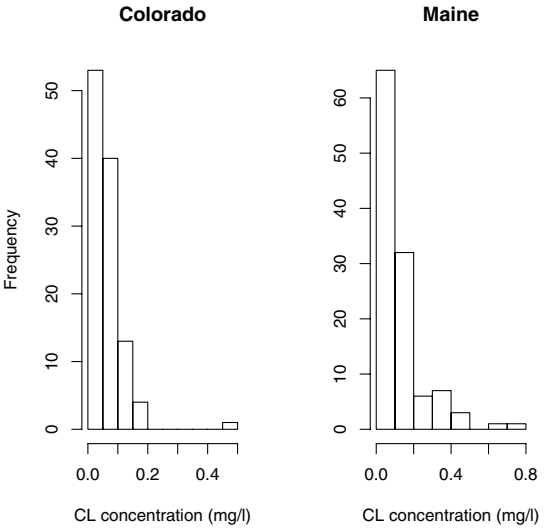


**Fig. 4.2.** Histograms of the monthly weighted averages of chlorine concentrations measured in Colorado and Maine from 1995 to 2004.

wherein salt NACL is dissolved into sodium and chlorine ions. Thus seafoam can be carried into the atmosphere by winds producing higher chlorine concentration levels. (The respective averages for these two states are 0.069 (mg/l) and 0.13 (mg/l)).

Acid precipitation travels in both directions across the Canada–United States border and not surprisingly, southern Ontario has a monitoring network. It began as an acid rain monitoring network in the 1970s. However, it has changed over time and its history is instructive as it illustrates well how such networks evolve with changing societal concerns and knowledge bases.

The current network now consists of the conjunction of three monitoring networks established at various times for various purposes: (1) Environment Air Quality Monitoring Network (OME); (2)Pollution in Ontario Study (APIOS ); (3) Canadian Acid and Precipitation Monitoring Network (CAPMoN) described by Burnett et al. 1994. (Le et al. 1997, extending the work of Brown et al. 1994a, show how to statistically integrate these three networks.)

Reflecting concerns of the day, both APIOS and CAPMoN were established to monitor acid precipitation (see Ro et al. 1988 and Sirois and Fricke 1992 for details). In 1978, CAPMoN (see Sirois and Fricke 1992) began with just three sites in remote areas, but in 1983 it grew through a merger with the APN network. The new network came to be used for a second purpose, tracing source–receptor relationships. To that end, sites could be found closer to urban areas. More recently, a third purpose for the network has been proclaimed, that of discovering relationships between air pollution and human health (Burnett et al. 1994; Zidek et al. 1998c).

Among other things, the composite network of 37 stations now monitors hourly levels of certain air pollutants including nitrogen dioxide ($NO_2$ $\mu g/m^3$), ozone ($O_3$ ppb), sulphur dioxide ($SO_2$ $\mu g/m^3$) and the sulfate ion ($SO_4$ $\mu g/m^3$). At some sites only one of the ozone or sulfate monitors had been installed. No doubt like CASTNET, the NADP/NTN, the Ontario network will continue to evolve as information needs change over time.

### 4.1.2 The Problem of Design Objectives

The lack of permanency of spatial sampling designs derives from a variety of factors. Most notably social values change. So does the state of knowledge, a close cousin of social values. The emphasis we now see on environmental protection is relatively new. (The EPA was established in the United States just over 30 years ago.) Those social values have spawned the need for regulation and control and in turn: (1) research on the nature of these processes and (2) the need to monitor environmental space–time processes.

With changing values and knowledge comes the adaption of spatial designs. The intensity of monitoring may change. For instance, TEOM monitors now make hourly measurements of particulate air pollution where just a few years ago volumetric air samplers generated their readings just once every few days. The number of spatial sampling sites may need to increase. More

strikingly the set of responses being measured may change along with the devices and methods used to make those measurements. Urban areas only began measuring particulate air pollution in the mid 1990s. Yet they do so using the same spatial sampling sites established as much as 20 to 30 years earlier, to measure pollutants such as ozone with a very different character.

Thus designers face major challenges. Traditionally, their strategies were based on carefully articulated measurement objectives as, for example, in the very elegant mathematical theory of optimal design. There the goal might be sampling points that maximally reduce the variance of the estimator of the slope of a line relating a response $Y$ to a design variable $X$, the latter constrained to lie in the interval $[a, b]$. Not surprisingly, the optimal design would put $1/2$ of the observations at $a$ and the remainder at $b$.

### 4.1.3 A Probability-Based Design Solution

But what if the designer of a permanent network knows that his original design will need to change in some as yet unknown way? He may then resort to probability-based designs such as those used to produce official statistics by government agencies. These agencies do face similar uncertainties about the use of their data.

Indeed that is the approach used in the U.S. EPA Environmental Monitoring and Assessment Program (EMAP) described next.

*Example 4.1. EMAP program*
The very ambitious EMAP program began in the United States with the goals of:

- Advancing the science of ecological monitoring and ecological risk assessment;
- Guiding national monitoring with improved scientific understanding of ecosystem integrity and dynamics;
- Demonstrating the framework of the Committee on Environmental and Natural Resources through large regional projects.

EMAP was created to produce indicators of the condition of ecological resources that could be used for monitoring. As well, multitier designs were to be "investigated" as a way of addressing multiscale data capture and analysis with the possibility of aggregating these data across tiers and natural resources.

Probability survey designs are used in EMAP for spatial site selection. These designs require the target population (sampling frame) to be spatially representative; such things as randomization, spatial balance, stratification, and equal or unequal weighting are needed.

Starting in 1989, EMAP was based on a global grid comprised of large hexagons placed on the earth's surface. These were used to generate a complex hierarchical system of hexagon grids in various size categories within

which areas were equal. Within each random sample of spatially represented hexagons, lakes, for example, could then be stratified, perhaps by size if that seemed appropriate. Then a random sample could be selected within strata in obvious ways.

However, the probability-based design approach also encounters the challenges described above. In fact, in Example 4.1 design objectives must still be specified (http://www.epa.gov/nheerl/arm/#dictionary). Moreover, modeling is implicit in stratification as it requires prior knowledge or data. At the same time, the approach ignores geographical features such as interlake distances. Yet ignoring these important features, and failing to exploit spatial associations, could lead by chance to the collection of redundant information from two adjacent lakes.

Chapter 11 presents a spatial design approach suggested by Caselton and Husain (1980). Working originally in hydrology, they later proposed it as a general approach to spatial sampling (Caselton and Zidek 1984) to circumvent the difficulties described above, while taking advantage of the designer's knowledge of spatial features. Its pros and cons are studied in that chapter.

### 4.1.4 Pervasive Principles

We leave this section by recalling important, but often ignored, principles for spatial sampling.

- **Make replicate measurements.** At least two measurements should be made at each space–time sampling point. This enables estimation of local stochastic variability (sometimes called the *nugget effect*). That variability, which may come from measurement error or interlab discrepancies, can exceed the space–time variability of central concern! The analyst cannot then draw conclusions with confidence about the latter, in the absence of estimates of the former. Yet the former cannot be easily estimated without replicate measurements. In particular, relying on modeling for inference about the former may add more uncertainty than it eliminates.
  Surprisingly often, measurements are not replicated. For example, in the Lakes Study described above just one observation was taken in each lake. Yet the cost of taking a second from some other part of the lake would have been negligible compared to the set-up costs of traveling to the lake. In another case, replicate grab-samples were taken but then blended before analysis. The explanation: to ensure that the sample was more "representative!" Much valuable information was thereby lost.
- **Use quasi-controls.** Designers commonly overlook the need for sites (*quasi-controls*) in spatial subregions unlikely to see extreme values of the process under study. For example, regulators site monitors where large values of an air criteria pollutant are expected. This seemingly natural choice stems from the need to have a good chance of detecting noncompliance with a prescribed regulatory standard .

> Yet, this approach is short-sighted. Good epidemiological analysis of the association between criteria pollutant levels and adverse health impacts needs statistical contrasts in the measurements. Good designs must ensure that observations are spread evenly over low and high response regions to maximize the chance of detecting an environmental health risk. And the point of setting regulatory standards is the minimization of environmental health risks!

For a more extensive treatment of environmental sampling, see Green's well-known book on environmental sampling (Green 1979). We now consider the actual measurement of the responses associated with a space–time field.

## 4.2 Sampling Techniques

This section gives examples of the techniques used to measure environmental processes, to show their scope and complexity as well as the perils an unwary analyst may face. The methods vary greatly. Moreover, several alternatives may be available. Although these alternatives may be similar in concept, in reality, they may measure very different characteristics of a process. Thus, observations made by one method can differ a lot from another. The choice is made by balancing their cost against their benefits such as convenience, reliability and accuracy.

Even after a method has been selected, the results seen by the analyst may change over time. This change may be due to such things as wear-out or technological upgrades of the associated devices, variations in data management protocols, and changes in the providers of analytical services. Although these changes should, in principle, be annotated in the database, in practice they often are not. Thus, changes seen in a measurement series may not come from the process itself.

### 4.2.1 Measurement: The Illusion!

Almost always, the methods will not measure the process itself but rather a surrogate strongly associated with it. The reasons for this vary. Sometimes direct measurement is just not possible. Doing so may be too expensive, dangerous, or socially unacceptable. Direct measurement, although more accurate may be too slow. Thus the surrogate may yield timely data at a lower cost but with a loss in quality.

### 4.2.2 Air Pollution

We now turn to examples.

*Example 4.2. Air pollution*
Many airborne pollutants are thought to be injurious to human health and therefore their measurement of societal importance. Carbon monoxide (CO), a product of the incomplete burning of carbon, has been called the "silent killer" since it is a clear, odorless gas that reduces the blood's oxygen capacity. That in turn can lead to a form of asphyxiation. Automobiles are an external source of CO while gas stoves provide an indoor source. It can be measured by a gas correlation method. This process is based on CO's infrared light absorption, this playing the role of the associated (surrogate) process described above.

Ozone ($O_3$) is a well-known product of atmospheric chemistry, particularly on warm summer days. It can be continuously monitored by exposing air samples to ultraviolet (UV) light. The degree of absorption of that light (the surrogate measure) is proportional to the amount of $O_3$ in the sample.

Oxides of nitrogen (NO and $NO_2$) have been associated with acute health impacts such as asthma attacks. Moreover, these gases are precursors of acid rain; through photochemistry they can be converted to $NO_3$ and, in turn, to nitric acid in wet deposition. Chemoluminescence provides a measure of the concentration of these oxides. An air sample mixed with internally generated $O_3$ emits a characteristic light whose intensity is proportional to the concentration. ($NO_2$ requires catalytic conversion to NO before being measured in this way.)

Sulphur dioxide ($SO_2$) leads to $SO_3$ and in turn to acid (sulphuric) rain in the same manner as oxides of nitrogen. It is measured by the pulsed fluorescence method. Pulses of UV light cause the $SO_2$ to release a characteristic light whose intensity is proportional to its concentration.

Finally, small, inhalable, airborne particles have been consistently associated with acute health effects, both respiratory and cardiovascular, in the form of both morbidity (disease) and mortality (death). However, these findings are of rather recent vintage, and consequently, interest in particulates, new. These particles are classified according their effective aerodynamic diameter, $PM_{10}$ and $PM_{2.5}$ referring, respectively, to those less than 10 and 2.5 microns in diameter. Once the concentrations were measured using volumetric samples and filters with long intermeasurement times (days). However, the more modern continuous fully automated method uses the *tapered element oscillating microbalance* (TEOM) monitor. A controlled volume of air is drawn into the monitor and travels into the tapered element (mass transducer). That hollow, ceramic, tapered tube resonates as air is drawn through it. The resonant frequency changes as particles pile up on a filter at the end of the tube. That frequency (the surrogate measure) gets converted into the particulate concentration measure.

### 4.2.3 Acid Precipitation Again

*Example 4.3. Acid rain revisited*

Example 10.1 describes networks set up to measure acid precipitation. We now turn to the methods used in those networks.

Acid rain derives from both wet and dry deposition. The former (both rain and snow) requires two steps. First the precipitation needs to be collected. This is done in a variety of ways. One method involves a double-sided tipping bucket. When full, a side spills its contents into a collector and opens the other side as the collector. The amount of precipitation is measured by the number of "tips." The collector must then be emptied at regular intervals, say each week, cooled to about $5^oC$, and transported that way to a lab for analysis. At the lab a large number of chemical species are measured including acidity (pH) along with the concentrations of nitrate ($NO_3^-$) ions, sulfate ($SO_4^{2-}$) ions, sodium ($Na^+$) ions, potassium (K+) ions, calcium ($Ca^{2+}$) ions, and magnesium ($Mg^{2+}$) ions. These measurements are made by such methods as chromatography and spectrophotometry.

In contrast to wet deposition, dry deposition, though important, has proven very difficult to measure. In fact, it must be inferred from concentrations in the air near ground level combined with measurements of the processes leading to exchange between the ground layer and the air above.

### 4.2.4 Toxicology and Biomarkers

In the next example, we turn to environmental toxicology and measurement through the use of biological organisms

*Example 4.4. Water pollution*
The pollution of freshwater bodies has long been a concern. (We are ignoring here their acidification from acid rain, also a serious problem.) Commonly, water samples are collected in bottles for laboratory analysis. Another method (used in a variety of other applications as well) involves a biomarker where the effect on a living organism is used as the surrogate for the response.

A predecessor can be found in the early days of coal mining when the canary was used as a biomarker of lethal gas in the mine. Another example from Example 1.1 involves the use of benthic organisms found in the mud of the seabed. Grab-samplers on ships were used to sample them. These samplers have jaws or scoops that can grab seabed sediments.

Science Daily (2000) reports that algae were used to measure the degree of improvement in the quality of the water in a polar lake (Meretta) in the Arctic (Resolute Bay, Nunavit). The latter had been badly contaminated between 1949 and 1998 by raw sewage dumped from a Canadian Department of Transport base. The sewage contained phosphorus that nurtured the algae which grew in proportion to the amount of phosphorus dumped. By measuring the concentration of the biomarker, the scientists were able to discover that the lake's water quality has been improving since the base was closed.

Incidentally, by taking core samples of that lake bed these same scientists were able to make a retrospective historical analysis of the lake's condition.

The deposition rate of diatoms manifest in the core was used as a surrogate measure of that quality over time.

That describes a few of the ways of measuring environmental fields. However, experimenters must ensure their data are good enough for their intended purposes. That brings us to the subject of data quality in the next section.

## 4.3 Data Quality

The large domains addressed in environmental science lead to a range of measurement problems. Moreover, the diversity of environmental contexts in which processes must be sampled has spawned a great variety of measurement techniques, instruments, strategies, and sampling plans. Ingenuity has been required in designing and carrying out experiments and in environmental data analysis.

Of fundamental importance has been the need for data quality assurance. Measuring devices or sites need to be appropriately located. For example, pollution monitors should not be sited near heavily used roadways. The measurements taken need to be precise. Equivalently, at each sampling point (in space and time), enough independent unbiased replicate measurements need to be taken to compensate for the noise when measurement error is large.

### 4.3.1 Cost Versus Precision

However, where cost is a binding constraint, that precision needs to be sacrificed to ensure a sample that is representative of the space–time field. Surprisingly, the theory developed for the project in Example 1.1 implied that increasing the number of grab-sampling sites was preferable to adding replicates beyond just two or three at each site. In practice, experimenters face the technically challenging design problem of formulating appropriate data quality criteria and making the optimal trade-off between precision and representativeness.

A possible compromise meriting more attention than it has received is that of using a dense set of unbiased, low-cost, temporary measuring devices/sites to generate preliminary data as a basis on which to construct a permanent system of high-quality sampling sites. Another would be a set of mobile devices/sites in conjunction with a fixed low-density set of high-quality measuring sites. Each of these alternatives could be used to increase data quality while managing sampling costs.

### 4.3.2 Interlaboratory and Measurement Issues

Subject to the considerations above, measuring devices need to be precise, well calibrated, and well maintained. To the maximum feasible extent, replicate measurements need to be taken to detect any potential drift over time in

measurement quality (in addition to compensating for any initial deficiency with measurement quality). Where analytical services (labs, for example) are used to analyze the samples, several should be employed. Each sample should be split into subsamples, one to be sent in a timely manner to each of the service providers. A statistical method must be developed to aggregate their results. The quality of the work of the competing providers needs to be assessed from time to time with calibration samples.

The problem of interlaboratory discrepancies cannot be emphasized enough. For unknown reasons, the differences between analytical service providers tends to be greater than within-service differences in analytical results even when exactly the same instruments and methods are being used.

The importance of good data cannot be overstated. In particular, it can eliminate the need for complex modeling designed to compensate for poor quality. While the latter can sometimes rescue a badly executed experiment, it only does so by replacing one source of uncertainty (the bad data) with another (model uncertainty). However, as we show in the ensuing sections, deficiencies in data quality are inevitable, because of the scale and complexity of modern environmental risk assessment. Thus, we turn to Section 4.4 and see that such error can have extremely deleterious impacts in a statistical study.

## 4.4 Measurement Error

Measurement error arises from a variety of sources and can be any one of a variety of types described in this section. (Some of the sources were indicated in Section 4.3.) However, the effects are unpredictable and can be quite deleterious. Thus experimenters must strive to reduce it and analysts, to be wary.

The most obvious source of error would be the measuring device itself. Many techniques described in Section 4.2 use surrogate measures for the true values. Concern about the accuracy of such techniques is bound to arise since their calibration is bound to depend on uncontrollable environmental factors such as outdoor temperature. Surrogates can also be indirect measurements when actual measurements are not available. For example, in spatial epidemiology, measurements made at the nearest ambient monitor often replace the true exposure of subjects to air pollution because the latter are unavailable. The inevitability of such error means it must be embraced in any statistical method used in any analysis of the resulting data.

Instrument failures provide another source of error. For instance, volumetric air samplers can fail or malfunction for periods of time until they are detected. Moreover, early gauges for measuring the acidity of precipitation could be contaminated by bird droppings with a resulting measurement bias.

The errors noted in Section 4.3 arising from analytical services are important. Good quality control programs can help to reduce them but they

cannot eliminate this source of error so these too must be recognized and accommodated in the analysis.

In practice, observations of a space–time process constitute either spatial or temporal aggregates of an underlying process. This can lead to measurement error and difficulty depending on how such data enter the analysis. As an example of such an aggregate measure, daily precipitation means the cumulative total over the day. For another, grab-sampling devices have a practical lower limit to their size so a sample of seabed mud must be an aggregate from around a specified site. As a third, continuous time air pollution monitors cannot accurately measure instantaneous pollution levels; instead they rely on averaging to reduce the noisiness of instantaneous readouts. The last example would compromise estimates of human exposure to air pollution for health risk analysis based on tracking individuals through space and time.

Data such as referred to in Example 1.2 can be censored either because the true value is below or above detection limits. Special techniques have had to be developed to handle errors of this kind since they are so common.

### 4.4.1 A Taxonomy of Types

#### Missing Data

The most extreme form of measurement error comes in the form of missing data. Detailed discussions of such data are available. For completeness, we give a brief account of the topic and suggested remedies.

Some data may be missing at random for reasons in no way associated with the process being measured. Such data pose little problem except insofar as information is lost. More problematical are data whose absence is informative though such data are uncommon in our experience. A completely contrived example would be the failure of a water toxin sampler as a result of a sudden surge in the level of the substance being measured with loss of data during and following the surge. More plausibly, the instrument would continue functioning, the offending responses would merely be censored (as above) and subsequent data would be captured. Finally, we see data missing for what might be termed structural reasons: (1) monitoring stations commence operation at different times or (2) different stations measure different subsets of a suite of environmental hazards.

We next give a taxonomy of the other types of measurement error likely to be encountered in practice. However, the vigorous and systematic study of those errors, especially with reference to study design and the development of mitigation strategies, has only begun rather recently. This may be because of complacency deriving from "... a common perception that the effect of measurement error is always to attenuate the line" (Carroll et al. 1995 p. 23). This complacency leads to the belief that the evidence against a null hypothesis is if anything reduced, that measurement error will have attenuated the slope of regression, thus reducing it towards the null value. In other words,

a belief that the correct p-value would be even smaller if it were not for the measurement error.

Recent increasing reliance on nonlinear regression models in spatial epidemiology, for example, may have helped kindle interest in the problem. That reliance can be explained by a combination of computing technology and methodological advances. The complexity of models may have challenged simplistic views born of simple linear regression models.

Those same advances may also explain why investigators have been willing to turn to the measurement or *errors in variable* (EIV) (as it is sometimes called) problem. Undoubtedly Fuller's fundamental treatise (Fuller 1987) on that problem stimulated those advances. It convincingly demonstrated the truly complex and pernicious character of measurement error. The more recent surveys of Carroll et al. (1995) and Gustafson (2004) complement and update Fuller (1987) and show the advances that have been made by the authors and others since the publication of Fuller's book.

This section taxonomizes measurement error as it has been characterized within the sampling school of statistics. Different taxa of error have seen the development of different methodological tools.

However, that taxonomy is redundant if error is treated within the ambit of the Bayesian paradigm. Its elements are then subsumed in that all uncertain quantities, including those measured with error, are treated as random variables. They can then be incorporated in any analysis through an appropriate joint distribution.

In spite of the increasing reliance upon Bayesian methods in modern statistical science, much current and recent theory for treating measurement error has been developed within the framework of the repeated sampling paradigm. For completeness we therefore describe developments from that perspective beginning with the taxonomy offered in this section. Moreover, nothing more explicit needs to be said about measurement error within a Bayesian perspective.

For continuous exposure variables, measurement error is generally characterized as either of classical or Berkson type, differential or nondifferential, structural or functional. Errors can also be of mixed type.

### Classical and Berkson Types

Classical measurement errors obtain in *analytical studies*, i.e., studies of individuals; the exposure measurement $W = X + U$ where $X$ denotes the true exposure and $U$ independent noise. The Berkson type arises, for example, when all members of a subregion are assigned a single subregional value $W$ obtained from an ambient monitor for that region and $X = W + U$, $U$ representing an independent deviation ascribable to individual differences. Carroll et al. (1995) use instead error calibration and regression calibration, respectively, to describe these two classes of error models. These two seemingly similar models are actually very different in their implications for practice.

A mixed model arises when $W = X + U$ (classical) while $X = Z + V$ (Berkson) where $Z$ is an extraneous environmental variable.

### Nondifferentiable Error

Nondifferentiable measurement error obtains when the health response $Y$ and $W$ are independent random variables while the true exposure $X$ is given. In other words, the measurement has no information about the response other than that contained in the true exposure itself. In this case, unlike that of differential error, $W$ serves merely as a surrogate for the true exposure and nothing more. Carroll et al. (1995) suggest that many situations are best described by nondifferential models. We believe in particular that they apply in the study of the acute health effects of environmental exposures as described in later sections.

### Structural Versus Functional

*Structural* measurement error refers to the case where the true exposure is random whereas *functional* means it is treated as fixed (but unknown).

### Misclassification

In the technically elementary case of binary exposure variables ($0 = low$ and $1 = high$, say) measurement error is called *misclassification*. Although conceptually relevant, the classical–Berkson dichotomy cannot be formally used. To see this note that $E(W|X)$ cannot equal $X$ (which is 0 or 1 except in degenerate cases), as it must if the classical model were to obtain. However, the concepts can be expressed through a reformulation of the measurement error model in terms of probabilities and conditional probabilities.

## 4.5 Effects

Little of a general qualitative nature is known about the effects of measurement error although a substantial methodological base for handling errors exists. By using that base, the implications of error can be assessed in particular contexts. However, some general results are known and this section describes them.

In the case of binary exposure variables, Thomas et al. (1993) show for analytical studies that quantities such as relative risk are attenuated for the nondifferential misclassification. Greenland (1982) proves analogous results for matched case-controlled studies. In fact he shows in this case the surprising result that nondifferential misclassification can have more detrimental effects than in the unmatched case, the size of the detriment growing with the closeness of the match.

However, these results reverse in cluster-based i.e., ecological studies. There populations are partitioned into groups and group attribute measures, rather than those of individuals, enter the analysis. With nondifferential misclassification, estimates of rates (slopes) for individuals based on group-level analysis will generally be inflated, rather than deflated (attenuated), towards the null as in the case of the classical error model and simple linear regression.

Thomas et al. (1993) note the complexities introduced by multilevel (discrete) exposure variables that make the effects or ecological estimates quite unpredictable.

For continuous variables the classical nondifferential measurement error model leads in simple linear regression to an attenuation towards the null of the apparent effect of exposure. This does not occur in the case of the Berkson error model, however, where the apparent effect remains unbiased.

**Crossover Designs**

The case-crossover design of MacClure (1991) seems useful for the assessment of acute health effects from exposure, in that the individual serves both as case (exposure levels at the time of failure) and control (exposure prior to failure). However, as noted by Navidi (1998) this design can be improved upon when time trends are present including, as well, exposures after the failure, provided these are not affected by the failure itself.

In general, ignoring measurement error can lead to myriad problems apart from the bias resulting from attenuation discussed above. Zidek (1997) demonstrates in the case of nondifferential structural measurement error that the curvature of nonlinear regression models can pick up the covariance of the measurement error's covariance structure.

### 4.5.1 Subtleties

**Attenuation. . . .or Not?**

To see in a simple setting some of the complexities ahead, consider just trivariate response vectors $(Y, X, X^g)$ having a joint multivariate normal distribution . Assume the commonly used impact model $E[Y \mid X] = \exp[\beta X]$. Inference concerns $\beta$ and $(X, X^g)$ has a bivariate normal distribution. Now $E[Y \mid X^g] = E[\exp[\beta X] \mid X^g]$ if $Y$ and $X^g$ are conditionally independent given $X$. Thus $E[Y \mid X^g] = \exp[\beta\beta_{XX^g}X^g + \beta^2\sigma_{X\cdot X^g}/2]$. As in the linear case, bias induced by measurement error expresses itself through $\beta_{XX^g}$. However, the "curvature" of the model now draws in a measure of how precisely the surrogate $X^g$ represents $X$ through the residual variance $\sigma_{X\cdot X^g}$. If the latter were 0, one could fit the naive model $Y = \exp bX^g$ and then correct for bias in the estimator $\hat{\beta} = b$ exactly as in the linear case. However, if $\sigma_{X\cdot X^g} \neq 0$ we see competition between the need to inflate $b$ to compensate for bias and deflate $b$ to compensate for lack of precision. To be precise, if one has a large

residual variance and fits the $Y$ on $X^g$ model above, the fitted value of $\beta$ will be close to 0.

The effect can thus dominate the attenuation that leads to bias. The effect of the error can therefore not be predicted without detailed analysis; the coefficient that transfers exposure to health impact can either be inflated or deflated by the error.

### Transfer of Causality

Zidek et al. (1996) describe a more subtle problem that can arise when both nondifferential structural measurement error and collinearity obtain. The authors assume in a hypothetical situation that a response count has a Poisson distribution with conditional mean $\exp(\alpha_0 + \alpha_1 x)$ where $x$ represents the "cause" of $Y$. A second predictor covariate $w$ has been observed but both $x$ and $w$ are measured with error according to a nondifferential classical model to yield $X$ and $W$. It is shown by means of a simulation study that if an investigator were to fit $\exp(a_0 + a_1 X + a_2 W)$ when the measurement error in $X$ is sufficiently large compared to that of $W$ while $X$ and $W$ are sufficiently strongly correlated, the analysis may well show $a_1$ and $a_2$ to be nonsignicant and significant respectively. Thus although $x$ represents the causative factor, that represented by $W$ inherits the role. Causality has thus been "transferred" through a combination of measurement error and collinearity. (This phenomenon is noted for linear regression models by Fuller 1987.)

While hypothetical, the result raises serious concerns for practice. Can any significant finding from a multivariable environmental health impact study be due to such a simple collusion among the variables?.

That concern is further reinforced by Fung and Krewski (1999) who extend and confirm the analysis of Zidek et al. (1996) by considering both Berkson and classical error models. They also investigate promising methods for mitigating the effect of measurement error in this context. We refer the reader to the comprehensive survey of Carroll et al. (1995) for a more detailed study of measurement error.

## 4.6 Wrapup

In this chapter, we have seen how the finiteness of resources means that uncertainty about an environmental space–time process can never be fully eliminated. To be sure, some uncertainty is eliminated by measuring selected variables in the space–time field. But some will remain since even with the best data quality management program, measurement error is inevitable. Moreover, all uncertainty about the unmeasured responses will remain. That suggests investing some of the available resources on models that can use the same measurements to reduce the latter sources of uncertainty. The next Chapter gives us an overview of the modeling process.

# 5

# Modeling

*When the only tool you have is a hammer, then every problem begins to look like a nail.*

*Abraham Maslow*

*All models are wrong . . .but some are useful.*

*George Box*

*A model, like a novel, may resonate with nature, but it is not a "real" thing.*

*Oreskes, Shrader-Frechette, and Belitz (1994)*

Much of this book is devoted to modeling. Building on the foundations presented in Section 1.2, we attempt in this section to put the work that follows into context by describing some of basic issues and the variety of approaches that have been taken. However, as the above quotation from Maslow suggests, our discussion is inevitably be limited by our own experience and predispositions.

## 5.1 Why Model?

Modeling can have any of a number of purposes, all seen to some extent in the analysis of data obtained from the measurement of space–time fields. We now provide a partial list of purposes on which the work presented in this book has a bearing.

### Data Summary

Models may be used to simply summarize a complex data set. For example, a large scatterplot of log ozone against daily maximum temperature may be summarized by fitting a line. The scatterplot can then be summarized simply by saying that log ozone increases by such and such an amount for a 10°C change in the maximum daily temperature. Similar summaries are very commonly used in spatial epidemiology where *relative risks* are described as the % increase in the incidence of a disease for a unit increase in the level of a pollutant.

### Knowledge Representation

Given the uncertainty that attaches to all measurements, scientists have long recognized the importance of bringing background knowledge, however crude,

into an analysis of experimental data. (To this extent at least, all statistical analysts are Bayesians)] That knowledge can come from physical theory or related studies. The knowledge may be as crude as saying that the relationship between two variables $y$ and $x$ is quadratic with a positive coefficient in the quadratic term. However, as we show, the models are often enormously complicated owing to the huge number of measured and unmeasured items connected with an environmental process.

### Covariate Adjustment

Most studies are observational rather than the randomized controlled experiments that are offered in statistics textbooks for proving causal relations. That is, the experimenter does not determine the process by which certain experimental units are treated and others not. Hence apparent associations, for example, between urban ozone and asthmatic attacks cannot prove the former causes the latter. The reason: potential confounders. In that example, both are related to the variable "population size." Areas with more people will simultaneously produce more ozone while having higher numbers of asthma attacks, even without any causal link between the two. Anticipating this kind of criticism, investigators seek to adjust their analyses for the effects of all the confounders they can think of (and measure!). To do so, they need a statistical model.

### Hypothesis Testing

Models can express a theory in such a way that its degree of accord with the data can be evaluated. When that support is small the theory can be rejected. In fact, a single observation (say a negative value) can sometimes reject a theory (say one that implies a positive response). Curiously no amount of confirmation of a theory through data collection can ever prove it!

### Prediction

Models are needed to predict or impute responses in space and time that have not been observed using those that have. In fact, spatial prediction is a major focus of this book and methods for doing so are shown in the sequel. In particular, Chapter 2 presents a case study that demonstrates the need to impute (hindcast) unmeasured historical air pollution concentrations. However, although our models can be used for temporal forecasting, limitations of space prevent us from addressing that topic in this book.

### Statistical Syntheses

Statistical models can be used to integrate data from a variety of sources. For example, in Example 10.1 we see *systematically missing* (sometimes called misaligned) data (Le et al. 1997) where some sites systemically monitor different

responses than others. A related problem derives from data with *misaligned support* (see Cressie 1993). Here we see data being collected at different levels of spatial or temporal resolution. For example, some may be from point sources while others come from cells in a grid. Some might be at the county level, others at the city level, and so on. Finally, we have seen a new direction in modeling emerging in recent years where some (simulated) data are outputs from physical models while others may come from measurement; yet all could be considered data, presenting still another situation where models would be needed for integration.

### Impact Assessment

Data from established environmental process monitors will typically give poor estimates of exposure in environmental risk analysis. To avoid the unpredictable and sometimes pernicious effects of measurement error, exposures need to be predicted (Carroll et al. 1995), ideally through a predictive distribution that allows uncertainty in the prediction to be represented.

### Data Smoothing in Disease Mapping

Things such as disease counts per interval of time can be very noisy particularly for, small geographical areas. Thus interest focuses not on the counts themselves but on the more stable, latent propensity of these areas to produce those counts. It is that propensity which is tied to intrinsic and extrinsic features of the area of interest and of concern to environmental risk managers. Such features might include income, for example, when the issue of concern is social justice with respective to environmental risks (Waller et al. 1997). Alternatively, the concern might be the effect of a hazardous waste site.

Analysts long ago recognized the benefits of borrowing strength from neighboring areas by extracting relevant information in their counts. Models are needed to help trade off the bias in these neighboring counts to improve the precision of the estimates of the propensity of real interest.

### Estimation of Trends and Gradients

In environmental risk assessment, temporal trends and spatial gradients have sometimes been of more interest than the levels of the space–time field. For example, Holland et al. (2003) find:

> Significant reductions in $SO_2$ and $SO_4^{2-}$ emissions under the Clean Air Act Amendments of 1990 have resulted in unprecedented improvements in $SO_2$ and $SO_4^{2-}$ concentrations.

Spatial gradients can point to hot-spots and possibly unrecognized environmental hazards. Estimation of the latter, a traditional topic of interest in spatial statistics, remains one of its current directions.

However, from a technical point of view finding the required estimates, i.e., estimating process derivatives, poses a more challenging problem than that of merely estimating levels. In particular, models play a fundamental role both in finding those estimates as well as in assessing their performance.

**Optimally Locating New Process Monitors**

Generally, the installation of new monitors entails a considerable start-up cost as well as substantial operating costs. The latter can include the costs of laboratory analysis, for example, in the assessment of hydrocarbon concentrations. Thus, the agencies responsible for establishing them seek to do so in an optimal way. However defined, such optimality has relied heavily on space–time modeling, another both long-standing and current direction in environmental risk assessment. In fact, the whole of Chapter 11 is given over to that very important topic.

## 5.2 What Makes a Model Good?

In Section 1.2, we describe some general performance criteria that can be used to assess models and the inferential procedures they imply. Here we present some more specific desiderata that derive from our practical experience.

*Good models. . .*

**. . . contend with time as well as space.** This point seems obvious. However, purely spatial methods, particularly those deriving from geostatistics, have been used to describe space–time processes.

**. . . come with an associated design methodology.** As we emphasize in Chapter 3, measurement and modeling are intrinsically linked. Therefore, no approach to modeling can be considered successful unless the result can pinpoint where additional measurements should be made. This feature is surely one of the great features of the geostatistical method called kriging that we describe in Chapter 7.

**. . . have predictive distributions that fully reflect uncertainty.** We emphasized, in Section 5.1, the need to make spatial and temporal predictions. However, these predictions will be of little value unless they come with realistic assessments of their own uncertainty. As well, the predictions may be used to predict exposures for input at the next level of a hierarchical model to assess impacts as in Chapter 13. For these and other reasons, predictive distributions rather than just point predictions are required.

**. . . contend with nonstationary spatial covariance.** As we see in Chapters 2 and 6, the covariance proves to be a tool of fundamental importance in analyzing space–time fields, even when they do not have a joint multivariate Gaussian distribution. Yet Chapter 2 also demonstrates the nonstationarity of spatial covariances. That is, the covariance between the responses at site pairs is not always determined by the difference between their

geographic coordinate vectors (latitude and longitude, say). Therefore, good modeling strategies need to embrace this challenging feature of environmental processes that arises in practice.

**. . . can incorporate multivariate response vectors.** Spatial and temporal methods typically borrow strength by incorporating relevant information in the responses of neighboring sample points. However, far greater strength can be found in other responses at the same sample point when the latter are highly correlated with any response of interest, as is commonly the case. Thus, space–time models should be constructed from square one to deal with multivariate response vectors.

**. . . can contend with systematically missing responses and mismatched data supports.** This property is self-evident. As noted in Section 5.1, both of these technical challenges are met commonly in practice.

**. . . can cope with very large data sets.** We are moving from the era of too little data to too much. Satellites generate vast quantities of it, for example. Analysts then face data storage and computational challenges. For example, if computing even a sample average takes a lot of time, as can sometimes be the case, more complex calculations may not even be possible. Finally, statistical challenges confront the modeler, as well. For example, to be large, a data set inevitably has to contain data from responses measured on a space–time grid of very fine resolution. To make full use of those data then requires that the fine-scale correlation be accurately modeled. That proves quite difficult, leaving a serious risk of mispecifying it. In turn this can seriously reduce the quality of statistical methods based on the model, leading to such things as bias and inefficiency. It can also yield unduly short prediction intervals.

**. . . can contend with large spatial and temporal domains.** A feature of environmental science is studies conducted on a very large domain, such as a big fraction of the Pacific Ocean or Canada's land surface. (For an example of a study involving the latter, see Chapter 12.)

Next we survey approaches to modeling space–time processes. None as far as we know, currently have demonstrated that they satisfy all the desiderata above, but then, collectively they do create a very high modeling hurdle.

## 5.3 Approaches to Modeling***

Information technology has rendered very complex modeling technically feasible, in particular, through hierarchical Bayesian models.

### Markov Chain Monte Carlo

These models have benefited tremendously from the emergence of a computational technique called the Markov chain Monte Carlo (MCMC) method. We do not have space in this book to describe that method and instead refer

the reader to Gamerman (1997) and Gilks et al. (1995). Briefly, this conceptually simple but ingenious technique allows the computer to repeatedly sample from the posterior distribution, often taking advantage of the sequence of layers that define the hierarchy. So, for example, a posterior expectation of a function of a model parameter, say $E[h(\theta) \mid Y^{meas}]$ is approximated by $(1/n)\sum_{i=1}^{n} h(\theta_i)$ when $n$ is large, where $\{\theta_1, \ldots, \theta_n\}$ are the parameter values sampled from the posterior distribution of $\theta$ given the data $Y^{meas}$.

The emergence and development of the MCMC method was stimulated by the development of high-speed processors, the very thing that makes the use of Monte Carlo methods feasible in the first place. And that development over the past ten years has been intense. Although issues still arise in its implementation, by and large it has taken the world of Bayesian statistics by storm and is now enthusiastically embraced and used, especially in complex modeling where explicit solutions are not feasible. The speaker in a recent presentation attended by the second author conceded that he had to simplify his original model which contained about 500,000 parameters because of computational difficulties associated with his implementation of MCMC. However, he beamed, he had managed to rescue his model with a simplification that brought the parameter total down to a mere 100,000!

Some general cautionary remarks seem in order before advancing to a more focused discussion of modeling strategies. As noted earlier, model complexity can make explicit model forms elusive leading to numerical methods such as MCMC. However, with MCMC methods, inevitable concerns arise about whether the chain has run long enough to burn in, and long enough to have converged. That in turn leads to a need for diagnostics with their inevitably subjective elements. Moreover, a modeling method with heavy computational requirements works against its enjoying a practical design strategy and use for modeling over big spatial–temporal domains.

### General Issues

Apart from computational issues, we find the challenge of model specification troubling. How, in fact, would one judge a good model from a poor one in this context? In principle, this should not be a issue. After all, the Bayesian paradigm tells us good models are those that correctly reflect the builder's prior knowledge. And even with a paucity of such knowledge, the data will come to the rescue by adjusting the prior model appropriately. Welcome insurance!

However, complex models with many dimensions and parameters exceed the mind's capacity for meaningful prior reasoning. Consequently, vague, even improper priors may need to be invoked (with the dangers that poses according to Dawid et al. 1973). A small amount of data relative to a surfeit of parameters, offers very little insurance via the updating mechanism. Moreover, the modeler may find it difficult to assess how much of the output comes from the data and how much from a conjunction of her prior inputs, playing out through myriad interrelated parameters. It would be hard to be sanguine

about such issues with a lot of environmental risk at stake, especially if the analyst has to sign off on the bottom line!

Recall (see Section 1.2) that our interest focuses on at most three space–time fields, $X$, $Y$, and $Z$ (not all of them being present in every application). Some of the responses (variables) comprising these fields will be measured and some not. With this notation in mind, we now list a variety of available approaches that can be taken. Note that the approaches can overlap and more than one can be used for any given problem. Finally, the list, although not exhaustive, does include the most common and powerful techniques of which we are aware.

The listed items vary greatly in their levels of generality, the first two being very general. We try to include some of their strengths and weaknesses and regret the inevitable shortcomings of our analyses that derive from a combination of space limitations as well as our wish to avoid an excess of technicality.

### 5.3.1 Modeling with Marginals

The marginal distribution functions of the individual variables are specified (see Marshall and Olkin 1988, for example). These are then combined through the use of a *copula*, a function that joins these univariate distribution functions to form multivariate distribution functions. To be more precise, in the case of two random responses, say $Y_1$ and $Y_2$, the process would begin by specifying their marginal distribution functions $F_{Y_1}$ and $F_{Y_2}$. Then a copula, $c(\cdot, \cdot) \in [0, \infty]$ would be specified. Finally, the joint distribution function would be constructed: $F_{Y_1,Y_2}(y_1', y_2') = c(F_{Y_1}(y_1'), F_{Y_2}(y_2'))$. Of course, $c$ would need to have the properties that ensure the resulting function is truly a joint distribution function. However, the real challenge lies in specifying the joint dependencies among the variables through the choice of the copula.

This simple, elegant approach does not seem to have enjoyed much success in spatial statistics, although we are not sure why. Perhaps it is because of the difficulty in specifying the dependencies. Or maybe it is because the conditional approach described next has so much more intuitive appeal to modelers.

### 5.3.2 Modeling by Conditioning

Modeling by conditioning (see Arnold et al. 1999), seems much more common than by marginalizing. To see the appeal of the approach, consider the simple identity for a joint probability density function, $f(x, y, z) = f(y \mid y,\ z)f(y \mid z)f(z)$, that we have expressed in terms of two conditional and one marginal density functions. Constructing the factors on the right-hand side would usually be simpler than the one on the left-hand side because of the small number of random variables associated with each factor. In the case of the conditional densities, that is because the conditioning variables are treated as fixed. At the

same time, we generate the necessary dependence relations as a byproduct of construction. Finally, this representation in terms of conditional distributions lends itself to the use of the MCMC method.

Since modeling by conditioning is the principal technique used in this book, we now elaborate on the approach. With the foundations of the Bayesian paradigm set out in Section 3.2, we aim for conditional predictive probability densities, conditional on the data (measured values), i.e., $f(Y^{unmeas'} \mid y^{meas})$ for any $y^{unmeas'}$ and $y^{meas}$. This conditional density can be represented in terms of the joint and marginal densities as

$$f(y^{unmeas'} \mid y^{meas}) = \frac{f(y^{unmeas'}, y^{meas})}{f(y^{meas})}. \tag{5.1}$$

In principle, only the numerator of the expression in Equation (5.1) needs to be found since the denominator can be found from it by virtue of the fact that the resulting density must integrate to 1.

**Example: Health Impact Analysis**

However, finding that numerator can be technically challenging because of the large number of variables involved as indicated in the next example.

*Example 5.1. Health impact analysis*
Zidek et al. (1998b) present an analysis wherein $Y$ and $X$ denote response arrays, $Y$ being daily counts of hospital admissions for respiratory morbidity of residents of each of I = 733 census subdivisions in southern Ontario over T = 720 summer days over the six years involved in the study, and $X$, an array of dimensions $733 \times 720 \times 5$ where 5 is the number of air pollutants involved in the study, $O_3$, $NO_2$, $NO_3$ $SO_3$, and $SO_4$. Just 31 of the 733 rows are actually measured. Finally, $Z$ is a $720 \times 3$ dimensional matrix representing fixed functions of time, and so does not contribute any random responses to the probability density in the numerator above. Nevertheless, "(unmeasured, measured)" contains $733 \times 720 + 733 \times 720 \times 5 = 3{,}166{,}560$ $X$ and $Y$ responses.

We turn now to more specialized approaches.

### 5.3.3 Single Timepoints

The approach is exemplified by the methods developed in geostatistics. That discipline was developed to enable unmeasured concentrations in a spatial field of ore to be inferred, using core samples obtained on the surface. Although one can scarcely imagine a less random field, nevertheless the discipline developed on the assumption that it was. At the same time, it was regarded as constant over time so, in effect, the selected sites were sampled at a single timepoint. *Kriging*, a method widely attributed to a South African engineer

named Krige, used a best linear unbiased predictor (BLUP) to impute the unmeasured concentrations at unsampled sites. To compute the BLUP requires the assumption of a known spatial covariance. Since it is not known, the method naively substitutes a plug-in estimate based on the sample. Getting that estimate in turn requires the additional of an isotropic spatial field.

The importance of the method leads to a more detailed discussion in Chapter 7 with variations we also describe. In the 1970s, the ozone field in a region of California was mapped anew for each timepoint for the period covered by the analysis. Over time, such fields were routinely kriged as there really was no competing methodology at the time.

The method has endured because of its many strengths and we now review these briefly along with its weaknesses.

On the positive side this mature method enjoys a very rich assortment of extensions, refinements, and software. Its great success no doubt derives from another positive feature, its extreme simplicity (that translates as flexibility, adaptability, transparency, implementability, and interpretability). Its optimality, albeit within the restricted class of linear unbiased predictors and known covariance functions, adds to its appeal. Since analysts apply the method timepoint by timepoint, no temporal covariance structure needs to be supplied, making it de facto robust against the misspecification of that structure. Finally, it comes with a very implementable design method: simply put the new monitors where the easily computed predictor variance is greatest.

However, a timepoint by timepoint approach cannot borrow strength by incorporating relevant information from adjoining timepoints. To compensate for that serious disadvantage means the method needs a large number of monitoring sites, an unrealistically large number when dealing with urban areas, for example, that may have fewer than 10. The method requires an overly simple spatial covariance model. Moreover, it ignores added uncertainty that derives from estimating that model. As a result, 95% prediction intervals may in reality be under 50% (Sun 1998).

### 5.3.4 Hierarchical Bayesian Modeling

The book's authors have developed and relied on this modeling approach, the subject of this subsection, over the past decade and we highlight it in this book (see Chapters 9 and 10). It relies on transforming the responses so their joint distribution is roughly Gaussian. Moreover, it takes advantage of the uniformity over space of parameters in trends and other systematic components. The approach also exploits that uniformity in the temporal structure. The trend and covariance without any specific structure are incorporated in the first level of the hierarchy with their uncertainty modeled in the second level.

This very general approach fully admits parameter uncertainty and yields a predictive distribution for input into impact assessment and into non-Gaussian

kriging models. Empirical assessments of performance reveal the prediction intervals to be well calibrated; e.g., 95% intervals really are (approximately) 95%. Moreover, the developed theory allows for misaligned (systematically missing) data and other structural features in the data. Developed around the Sampson–Guttorp approach (see Section 6.5.1) for estimating spatial covariance fields, it is not susceptible to the limitation of unrealistic stationarity assumption.

On the other hand, the method is considerably more complex than say, kriging. It is challenged by the nonseparability of some space–time series; there prefiltering the temporal structure, which may also remove the intersite spatial covariances, may be needed. Although the method has been much refined, elements of the theory have yet to be made fully Bayesian. Work in that direction is underway as this book is being written. However, the problem of providing user-friendly software is being solved and Chapter 14 provides a tutorial on the use of what has been developed.

### 5.3.5 Dynamic state-space Models

Process parameters do evolve over time. Building on the celebrated Kalman filter, this powerful tool incorporates that change into the process model itself. Rather than presenting an abstract description of this approach, we illustrate it with an example.

*Example 5.2. Dynamic linear model*
Huerta, et al. (2004) model the hourly sqrt($O_3$) field over Mexico City data from 19 monitors in September 1997. For time $t$ and site $i$ they assume the measurement model

$$X_{it} = \beta_t^y + S_t\alpha_{it} + Z_{it}\gamma_t + \epsilon_{it}^y.$$

Here $S_t : 2 \times 1$ has sines and cosines $\alpha$, their hourly amplitudes Z, the hourly temperatures over the city, and $\epsilon_{it}^y$, unautocorrelated errors with an isotropic exponential spatial covariance.

The parameter/process model lets the parameters change dynamically:

$$\beta_t^y = \beta_t^y + \omega_t^y.$$
$$\alpha_{it} = \alpha_{it-1} + \omega_t^{\alpha_i}.$$
$$\gamma_t^y = \gamma_t^y + \omega_t^\gamma.$$

The $Z$ (temperature) model incorporates elevation. The resulting hierarchical Bayesian model seems to model the short series quite successfully. However, for longer time periods, adaptations reflecting seasonality would need to be added, making the total number of parameters even larger.

This modeling approach enables the data to update parameters in a systematic and coherent way while minimizing the data storage requirements.

These features give it advantages over sequential model refits. Moreover, it seems intuitive, flexible, and powerful. As a Bayesian method it admits physical/prior knowledge. The approach of Chapter 11 could be adapted for use here, leading to optimal designs that change over time as knowledge about the process parameters increases. However, we doubt that such adaptable designs would be of practical value.

On the negative side, these can be very complex models with all the associated practical as well as conceptual difficulties described at the beginning of this section. Moreover, the substantial computation times lead us to wonder if we could handle a problem with a realistically large number of sites such as that in Chapter 12 with several hundred.

Finally, although a stationary spatial covariance may be appropriate in Example 5.2, that would be unrealistic in general, leading to even more prior modeling complexity, say by invoking the Bayesian approach of Damian et al. (2001). Also, we would find it hard to decide on the appropriateness of the assumption, for these hourly data, that temporal covariance components can be removed while leaving spatial covariance intact. Our experience has shown that cannot be done (Zidek et al. 2002).

### 5.3.6 Orthogonal Series

This powerful technique represents the discretized field of interest, say that associated with Y, as follows; $Y_{it} = \sum_{j=1}^{p} U_{jt}\phi_{ij}$. Here, $Y_t^T = (Y_{1t}, \ldots, Y_{pt})$ denotes the response vector over all spatial sites at time $t \in \mathcal{T}$. The $\phi_i : p \times 1$ are nonrandom orthonormal basis functions; that is, $\phi_i^T \phi_j$ is 0 or 1 accordingly as $i = j$ or $i \neq j$. This approach, one that keeps getting reused, has a very long history. To obtain one such expansion, suppose for simplicity the $Y$s have zero expectation. Let their spatial covariance matrix $\Sigma = E(Y_t Y_t^T)$ for all $t$. A well-known spectral decomposition theorem from linear algebra says that we can find an orthogonal matrix $O : p \times p$ (with $O^T O = OO^T = I_p$ such that $O\Sigma O^T = \Lambda : p \times p = diag\{\lambda_1, \ldots, \lambda_p\}$, the $\lambda$s being called $\Sigma$'s eigenvalues. If we let $U_t = OY_t$, it readily follows that $U_t$ has covariance matrix $\Lambda$, implying that the coordinates of $U_t$ are uncorrelated. Letting $Y_t = O^T U_t$ and $O^T = (O_1^T, \ldots, O_p^T)$ we obtain the famous Karhunen–Loeve orthogonal expansion $Y_{it} = \sum U_{jt}\phi_{ij}$, where $U_{t_1}$ and $U_{t_2}$ have correlation 1 or 0 accordingly as $t_1 = t_2$ or not. When the assumption of a temporally unchanging covariance matrix holds and $\Sigma$ is known, we thereby obtain an elegant regression model that for each t, represents $Y_t$ in terms of the known basis functions by means of the spatially uncorrelated random effects.

A variety of such representations has been developed. They are typically applied after systematic effects have been removed from the field and used for a variety of purposes. We now give examples. The first concerns the regression approach to design.

**Example: Designs for Monitoring**

The next two examples address modeling within a design context.

*Example 5.3. Fedorov and Mueller*
Fedorov and Mueller (1989) need such a regression model to move their optimal, regression–model based design theory into the domain of space–time processes. They comment that the ". . .most crucial assumption . . . is that the fluctuation of the observed responses is modeled by the randomness of the 'parameters'. . ." that are the $U$s, in our notation. However, they do not require the coordinates of $U_t$ be uncorrelated. They estimate that covariance, when unknown, and plug in the result as if the covariance were known, thereby underestimating the true uncertainty.

In the next example, we see the expansion used with a more comprehensive set of goals.

*Example 5.4. Mardia and Goodall*
Building on Mardia and Goodall (1993), Mardia et al. (1998; hereafter MGRA) cite the Karhunen–Loeve expansion to justify an expansion of the systematic component, not the random component, of their space–time model. However, they induce randomness in the $U$s of their model by adopting the so-called autoregressive process of order 1 [AR(1)], state space-model to describe their evolution:

$$U_t = PU_{t-1} + K\eta_t,$$

where the $\eta$ innovations process has a multivariate normal distribution with mean 0. They call the result the *Kriged Kalman Filter*. They call the $\phi$s the principal fields.

A number of discussants follow and comment on MGRA. For example, Angulo(1998) doubts the generality of the $U$-$\phi$ decomposition implied by assumptions underlying the Karhunen–Loeve expansion and argues (in line with extensions suggested in the 1994 paper) for allowing the $\phi$s also to depend on time. In fact, he goes further and offers an intermediate solution.

Stein(1998) argues that the method of MGRA is ". . . only modestly related to what would generally be called Kriging." His argument relies, in particular, on the fact that to implement their method, MGRA cannot use all the spatial data as a kriging predictor would.

Finally, Cressie and Wikle(1998) suggest that the method oversmooths. They go on to describe their own paper (Wikle and Cressie 1999) which they claim avoids oversmoothing because they admit an additional model component $V_{it}$ that represents time-varying, small-scale process variability that has spatial but not temporal structure.

However, their paper includes two other novel elements. First, it seeks to reduce the dimension of the problem by selecting just $K < p$ terms of the expansion, those corresponding to the biggest $\lambda$s. This sort of thing will

seem quite natural to anyone acquainted with principal component analysis. Second, it includes a kind of spatial–temporal AR(1) model

$$X_{it}^K = \sum_j w_{ij} X_{j;t-1}^K,$$

$X^K$ representing, in our notation, the reduced dimension systematic component that was modeled by the orthogonal expansion, and $w_{ij}$, weights that need to be selected.

Dimension reduction represents an important trend in space–time modeling, one that seeks computational efficiency when faced with increasingly large and complex data structures.

In one other notable application of this decomposition, Craigmile et al. (2003) use a wavelet decomposition of a space–time process to distinguish trend from small-scale variation. We do not go into detail about this very technical article, but note that it tackles the extremely important problem of inferring trend in the presence of long-term memory (where temporal correlations of very long lags are present). Long memory presents technically tricky issues of substantive importance since it can induce local patterns that could easily be mistaken for deterministic trend by the unwary analyst, but which in reality are merely low-frequency components of the long memory process.

Next we describe applications of orthogonal series decomposition.

### Principal Components and Empirical Orthogonal Functions

Important applications of the orthogonal series expansion above are the *principal component* (PC) approach in statistics and the *empirical orthogonal function* (EOF) analysis, which has proven very useful in physical sciences. The two approaches are complementary with each seeking structures that explain the maximum amount of variation in one of the two dimensions in a two-dimensional data set. For example, for a space–time process, the EOF approach identifies structures in the space dimension and the PC approach finds that in the time dimension. Since these are complementary and have a 1–1 relationship, they are called interchangeably depending on the tradition of a discipline.

We now describe the idea of empirical orthogonal functions and their relation to the principal components. Suppose $W : p \times 1$ denotes a random response vector across $p$ geographical sites. Furthermore, let $\mu = E(W)$ and $\Sigma = E(W - \mu)(W - \mu)'$ denote the population mean and population covariance respectively. As an aside, in physical modeling the word *population* often gets replaced by *ensemble*. Moreover, the ensemble mean is represented by $<< W >>$ instead of $E(W)$, the common statistics notation.

The positive definiteness of $\Sigma$ and the matrix diagonalization theorem imply that we can find a orthogonal matrix $Q$ such that

$$D^2 \equiv diag\{d_1^2, \cdots, d_p^2\} = Q'\Sigma Q$$

with $d_1 > \cdots > d_p > 0$ and $Q$ containing orthogonal eigenvectors of $\Sigma$. In other words, $U \equiv Q'W \equiv (U_1, \ldots, U_p)'$ has covariance matrix $\Sigma_U = D^2$, making the $\{U_i\}$ uncorrelated with decreasing variances. The ones with the largest variances are referred to as the principal components. That is because $\mathrm{tr}\Sigma = \sum_{i=1}^p \sigma_i^2 = \sum_{i=1}^p d_i^2$ so that they account for or "explain" a large fraction of $W$s total variance. At the same time,

$$W = \begin{pmatrix} W_1 \\ \vdots \\ W_p \end{pmatrix} = \begin{pmatrix} Q_1 : 1 \times p \\ \vdots \\ Q_p : 1 \times p \end{pmatrix} U = \begin{pmatrix} Q_1 U \\ \vdots \\ Q_p U \end{pmatrix}. \tag{5.2}$$

So each of $W$'s coordinate responses can be represented in terms of $\{U_i\}$, as $W_i = \sum_{j=1}^p Q_{ij} U_j$. In fact if we drop all but, say two of the principal components, we get approximately $W_i \approx Q_{i1}U_1 + Q_{i2}U_2$ and so on. This can mean a major reduction in dimension from $p$ locations to, say two, in the not unrealistic situations encountered in large-scale physical modeling where $p$ can be in the 100,000s.

Since $\Sigma$ is unknown in practice, it has to be estimated from observed realizations $W_t,\ t = 1, \ldots, T$, in the obvious way $\hat{\Sigma} = T^{-1} \sum_{t=1}^T (W_t - \hat{\mu})(W_t - \hat{\mu})'$, where $\hat{\mu} = T^{-1} \sum_{t=1}^T W_t$. Applying the diagonalization theorem above to $\hat{\Sigma}$ yields $\hat{Q}$ with orthogonal eigenvectors as columns. The eigenvectors in $\hat{Q}$ are called *empirical orthogonal functions*. The term *empirical* refers to the decomposition based on observed data.

Statisticians usually take the $\{W_t\}$ to be uncorrelated with the same covariance and mean. However, in applications they may well be autocorrelated meaning correlated over time (or space). However, as long as $t$ is moderately large, the resulting estimate $\hat{Q}$ will be quite satisfactory. The EOF analysis has been used widely in physical sciences, particularly in climatology, to identify efficient representations of data sets.

Two other particular related conditioning approaches for handling a large number of variables seem worth mentioning for completeness and we turn to them next.

### 5.3.7 Computer Graphical Models

This approach at the interface between computer and statistical science, is called *causal modeling* or sometimes *Bayesian belief modeling*. Richardson and Best (2003) discuss the use of such models in their comprehensive survey of hierarchical Bayes space–time modeling, specifically within the domain of environmental health risk analysis. This approach can help contend with complex environmental processes with a large number of random variables, including unknown parameters, where writing down a explicit expression for the joint distribution of measurable and nonmeasurable objects (such as vectors or matrices of parameters) can be impractical. The approach provides a

powerful graphical approach for organizing the web of dependence relationships among the random variables, thereby facilitating the computation of the joint distribution through a product of conditional distribution dictated by the causal model.

We illustrate the approach with a simple example.

*Example 5.5. Graphical models*
Let $(Y^{unmeas}, Y^{meas}) = (T, U, V, W, X, Y, Z)$ with the relationship among these variables depicted by the directed acyclical graph (DAG) in Figure 5.1.

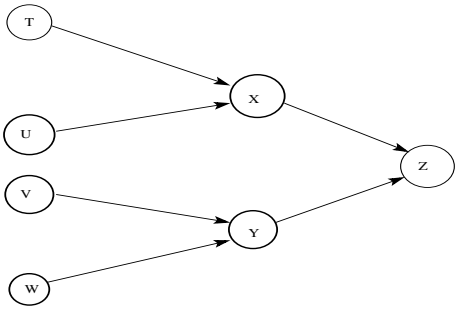


**Fig. 5.1.** This figure exemplifies a directed acyclical graph involving seven nodes and six directed edges that define the causal relationship among the variables, T,U,V,W,X,Y,Z.

The graph tells us that the joint density can be expressed as

$$f(t,u,v,w,x,y,z) = f(z|pa(z))f(x|pa(x)) \\ \times f(y|pa(y))f(pa(x))f(pa(y)) \tag{5.3}$$

with the parent sets, $pa(z) = \{x, y\}$, $pa(x) = \{t, u\}$, $pa(y) = \{v, w\}$. The arrows imply directional dependence, so that given $X$, $Z$ is independent of $T$ and $u$. (In general not all arcs need to be directed, depending on the nature of the dependence.) Here $pa(Y) = V, W$ means $V$ and $W$ constitute the parent set of $Y$ and so on. *Acyclical* means that we cannot find a sequence of arcs, including directed arcs that end trace out a path that ends up where it begins.

It can be shown that an equation like (5.3) obtains for arbitrary DAGs, making the approach broadly applicable. As well, the method has intuitive appeal. It enables easy input of conditionality relations among the variables, even for extremely large sets of variables. Moreover, the theory for DAGs enables the calculation of multivariate densities to be organized in an efficient way by judicious decomposition of the graph. Software, such as the package called HUGIN, facilitates these calculations as well as the calculation of all the appropriate conditional probabilities at the nodes when a variable at one of the nodes has been observed.

Sometimes graphs, called *chain graphs*, like that above involve some edges where the arrows are bi-directional. They arise when neither of the variables

in a pair of nodes can be said to cause the other but when they are statistically dependent for reasons other than a link through other nodes in the graph.

Decompositions such as that in Equation (5.3), can be extremely powerful tools for specifying a joint density when parent sets are not too large. Yet they have not been much used in modeling environmental space–time fields since there, the parent sets tend to be huge, involving all the remaining sites! However, it could prove valuable for modeling the spatial fields themselves, when causal relations can be identified because of well-defined spatial structures. Examples might include situations where the sites lie along the direction of a prevailing wind, or where they lie along a freshwater course.

### 5.3.8 Markov Random Fields

Another potentially powerful technique for spatial modeling at least, relies on *Markov random field* (MRF) models (see Kaiser and Cressie 2000, Lee et al. 2001 and Kaiser et al. 2002, for example). That approach has relied on the paper of Besag (1974) and the celebrated theorem of Hammersley and Clifford (1971). In a spatial process context, Besag and Higdon (1999) used this model to develop a method for analyzing the results of an agricultural field trial.

The spatial nature of this theory leads to the replacement of parent sets by neighborhoods. The approach here is quite different in character than that of causal modeling in that the goal is the local specification, for each variable (e.g., site response at a particular time) of its conditional distribution given the responses at every point in its neighborhood. Indeed, it is possible to produce all these conditional probabilities without consideration of the joint distribution itself. Now the analysis can turn to the question of whether there exists a joint distribution consistent with that specification.

Whereas the MRF methodology seems to have been an important tool in image analysis, its worth in modeling environmental space–time fields had not been convincingly established when this book was being written. In particular, we have not seen any convincing evidence of its value in spatial prediction. In any case, the majority of this book is devoted to conditionally log-Gaussian space–time fields where such methods are not needed. Convincing evidence will need to come from application to more complex stochastic structures.

The previous discussion notwithstanding, we still need to find ways of modeling the joint distribution, in particular the conditional components in its decomposition since no simple direct estimates are typically available. This is where the powerful device of *hierarchical modeling* may profitably be used. In that approach, exemplified subsequently in this book, the required joint density is found first as a conditional density given some unknown model parameters, themselves unmeasured items that in the Bayesian framework have a marginal (prior) distribution. The conditional density can be averaged with respect to the prior distribution to get the required numerator above.

That conditional density can in turn be represented by an extension of the result in Equation (5.1). In fact, we obtain precisely the same expression if we

simply extend "unmeasured" to include the model parameters as well. Now we can have a truly immense number of items in (unmeasured, measured), perhaps tens of millions. However, surprisingly, the inclusion of the parameters can actually simplify things if they are chosen judiciously because the joint density can be factored into a (large) product of joint densities of a small number of items.

### Example: Birch Tree Distribution

We illustrate a combination of the hierarchical approaches in the following example.

*Example 5.6. Crown die-back in birch trees* Based on counts obtained at single time $t$, Kaiser and Cressie (2000) develop a spatial distribution for birch trees suffering from a condition called crown die-back. These counts came from 36 sites $i$ in the northeastern United States. In a hierarchical step, these authors assume that conditional on the probability of crown die-back $X_{it}$, and total tree counts $m_{it}$ (regarded as fixed and known), the counts $Y_{it}$ are independently distributed with a binomial distribution. Modeling then turns to the joint distribution of the unknown $\{X_{it}\}$ taken to be a MRF, the neighborhood of any site $i$ being all other sites within a 48 km radius. Conditional on its neighbors, each $X_{it}$ is supposed to be a beta distribution with parameters depending on the neighboring $X$s. The authors model those parameters in terms of just three hyperparameters, assumed to be the same for all $i$ and go on to develop inferential techniques for estimating these hyperparameters. Although the authors do not consider the problem of predicting $Y_{jt}$ as some new site $j$ for which a die-back count is not available, they indicate how their method could be so used. However, empirical analysis would be needed to test and validate the approach.

The previous example merits further study. First, note that hierarchical modeling has been effectively used to separate the aleatory uncertainty (that in the measurement) and epistemic uncertainty (that in the model). The power of this decomposition would be more fully realized were some of the $\{Y_{it}\}$s found to be missing. Indeed, had the field been measured at successive times in search of trends, the pattern of missing counts could well have varied. The hierarchical decomposition would then have disentangled the complexities of describing the measurement process from those involved in modeling the latent propensity to die-back.

However, the separation of the two uncertainties might have been fully achieved by using the negative binomial rather than the positive binomial. The latter's variance unlike the former's, is too small, it being less than its mean. In practice, counts are susceptible to high levels of error, especially when based on the dichotomization of a continuous variable as in this case. Misclassification of experimental units close to the classification boundary inevitably occurs. Of course, the beta distribution in the example will help the resulting marginal

distribution of counts to better capture that measurement error but at the expense of breaking down the hierarchical demarcation between aleatory and epistemic uncertainty.

We would add that in the context of environmental statistics, counts are often spatial or temporal aggregates of nonidentically distributed random variables thereby casting additional doubt on the validity of the binomial measurement model.

Although the method for selecting neighbors may well be appropriate in the previous Example 5.6, it will not work in general. Spatial separation of sites $i$ and $j$ does not always predict well the dependence between them. That dependence may depend much more on latent environmental factors than on geography and these may not be well predicted by location. Salinity and water temperature can depend on the trajectories of ocean currents. Air pollution and acid deposition can be affected by wind directions and site elevation. Local sources can be an important determinant of environmental hazards; levels at two sites with similar sources might be more statistically associated with each other than with sites in between.

Considerations such as those above have led the authors to avoid basing the dependence between site measurements in a hierarchical model on geographical proximity. Instead, in the methods emphasized in this book, arbitrary dependence structures are permitted. We have been able to gain that generally, by concentrating on fields where at least conditional on model parameters in the first level of a hierarchical model, the responses (possibly after a suitable transformation) are supposed to have a multivariate Gaussian distribution. In the next Section 5.4, we describe such fields and their properties. Chapters 8–10 present Bayesian methods for making spatial predictions for them.

### 5.3.9 Latent Variable Methods

Here we use the representation $Y_{it} = \sum_j a_{ijt} W_{jt}$ where the $W_{jt}$s are uncorrelated for every $t$. This approach has been used quite a lot in modeling space–time fields (see Higdon 1998, Higdon et al. 1999 as well as Fuentes and Raftery 2005 for some recent applications). The variation of Gelfand et al. (2004) uses the so-called co-regionalization approach, an important idea in geostatistics (Wackernagel 2003). In fact, that idea is used by Schmidt and Gelfand (2003) to account for the dependence in a multivariate response vector of pollutants.

The method gives a powerful intuitive representation for finding reasonable covariance structures. However, unlike the basis functions in orthogonal expansions, for example, the latent variables themselves may not derive from physical or mathematical considerations. Instead, and this is one of their advantages, they may be purely intuitive, conceptual devices. This can make them difficult to implement and make the results seem personalistic or even arbitrary.

### 5.3.10 Physical–Statistical Models

Environmental processes are generally distributed over very large space–time domains. That, their complexity, and the amount of data commonly available from sources such as monitors and satellites make modeling them challenging from concept to implementation.

Commonly, deterministic models have been used to model those processes. However, such usage has generated debate at a fundamental level as the quote from Oreskes et al. (1994) at the beginning of this chapter illustrates.

Moreover, such models may prove unsatisfactory where large domains are involved. The so-called *butterfly effect* tells us that tiny perturbations in initial conditions can propagate into gross changes in model outputs as they dynamically evolve over time. Thus, forecasting weather more than two or three days ahead is difficult. In fact, dynamic nonlinear models can be susceptible to chaotic behavior, small shifts in conditions leading to abrupt and unpredictable changes in model outputs. Such behavior is the subject of *chaos theory*.

Problems associated with deterministic models have made statistics increasingly valuable. An early and somewhat ad hoc use is called *data assimilation*. Here deterministic model parameters are adjusted to get their outputs to agree with field measurements.

Outputs from deterministic models can be quite complex and difficult to interpret. So statistical models can help summarize, understand, and exploit their outputs. Example 5.7 illustrates such an application.

*Example 5.7. Coupled Global Climate Model*
Fu (2002) and Fu et al. (2003) are concerned with maximum annual precipitation over many years and more than 300 grid cells covering all of Canada. Since much of Canada is uninhabited, most precipitation goes unmonitored. Nevertheless it is of great importance since it tracks climate change and its implications for the nation. So how can the missing precipitation measurements be inferred?

The answer: It can be simulated by the Coupled Global Coupled Climate Model (CGCM1). That model with a surface resolution of $3.7^{o} \times 3.7^{o}$, runs uncoupled ocean and atmospheric models separately for the time period of interest. These outputs are adaptively integrated in 14-year blocks. Eventually a variety of climate response variables such as temperature and rainfall can be deterministically generated over long time periods and all grid cells. Apart from simulating data for nonmonitored grid cells, the model can also be used for scenario analysis to determine the effect in centuries to come of various levels of greenhouse gas emissions.

Fu and her coinvestigators use a hierarchical Bayes model for the distribution of the logarithm of simulated data. They fit a joint distribution over the grid cells for the simulated data, treated as random. That adds stochastic uncertainty to that deterministic output. Moreover, the distribution can answer questions involving a number of grid cells simultaneously. For example,

they can find the probability that the maximum annual precipitation will fall below a critical level (drought) in every one of a combination of grid cells representing an agriculture area. Thus, the model summarizes the simulated data in a very usable way.

Statistical models can also be used to analyze the outputs of those models. Fuentes et al. (2003) demonstrate such an analysis. However, unless it is sensibly done, the comparison of measurements with simulated data will not be meaningful. Outputs from the latter are typically on the mesoscale representing spatial scales of 100–100,000m and temporal scales of 100–10,000s. In contrast, measurements are made on the microscale. A direct comparison is as meaningful as comparing as apples and oranges! It is meaningless at an even deeper level, according to Oreskes et al. (1994), who state:

> To claim that a proposition (or model) is verified because empirical data match a predicted outcome is to commit the fallacy of affirming the consequent.

Simply replacing deterministic with statistical models also proves unsatisfactory. Although the latter admit input and output uncertainty, they lack the backbone needed for large domains. Hence the idea of merging physical with statistical models was born. That has led in recent years to the dramatic convergence now underway, of two very distinct modeling cultures of physical and statistical modeling.

Broadly speaking, two sorts of deterministic models have been addressed. The first involves just a few differential space–time model equations. Two approaches have been suggested for building a superstructure upon them (Wikle et al. 2001). These are illustrated in the following example, chosen for its simplicity.

*Example 5.8. Making deterministic models statistical.*
An environmental growth process $\{X(t),\ t > 0\}$ at a single site is governed by the following dynamic equation,

$$\frac{dX(t)}{dt} = \lambda X(t), \tag{5.4}$$

$\lambda$ being the growth parameter. Equation (5.4) is easily solved with the result $X(t) = \lambda t$, if constants are ignored. However, the growth parameter would generally be uncertain and hence random within a Bayesian framework. Thus $X(t)$ becomes a stochastic process with a lot of backbone. Equation (5.4) becomes a stochastic differential equation. So much for the first approach.

The second approach reconstructs the equation using a finite difference approximation as

$$\frac{X(t+\delta) - X(t)}{\delta} = \lambda X(t), \text{ or}$$
$$X(t+\delta) = K_\delta X(t), \tag{5.5}$$

for some $\delta > 0$ (the smaller, the more accurate) where $K = 1 + \delta\lambda$. To allow for uncertainty in the approximation error, a random evolution error is added:

$$X(t + \delta) = K_\delta X(t) + \epsilon(t).$$

The result is the state equation of a state-space model, i.e., Kalman filter that governs the dynamic growth in $X(t)$.

Of course, the first approach will commonly not work because the explicit equation solutions are not available. The second may fail for processes over large spatial domains because of prohibitively expensive computational burdens. However, the ensemble Kalman filter has been developed to yield a practical approximation in such situations (Bengtsson et al. 2003).

The second sort of deterministic model that can be merged with a statistical one involves a very large number of differential equations, making deconstruction of the equations impractical. Computers solve the equations, albeit slowly. Their outputs represent grid cells on the earth's surface at successive times. The finer the grid cells, the more burdensome is the computation. Worse still, in some applications, meteorology, for example, an ensemble of such models must be used. (Their results must then be amalgamated in some way.)

The following example is about a *chemical transport model* (CTM) used to forecast air pollution levels up to an elevation of several hundred kilometers. Our context is the troposphere , the approximately 20 km layer of atmosphere closest to the earth. (The *stratosphere* lies above it.)

*Example 5.9. Multiscale Air Quality Simulation Platform (MAQSIP)*
The deterministic MAQSIP model forecasts (among other things) hourly ozone concentrations at a grid cell resolution of 6 km $\times$ 6 km. It relies on two inputs. The first comes from another model, the NCAR/Penn State Mesoscale Model (MM5) computer model that provides the required meteorological inputs. The second inputs estimates of the precursor emissions that get turned into ozone through photochemical processes in the atmosphere. These processes along with the transportation of the products are simulated by differential equations, solved by difference methods such as the one in Equation (5.5) although more complex. In fact, numerous linear differential equations are needed to describe the processes of atmospheric chemistry alone.

How can computer model outputs for grid cells and point measurements at monitoring stations be meaningfully merged? Fuentes and Raftery (2005) answer that question with a technique called *Bayesian melding*. However, it applies only in a purely spatial context. Time is not allowed!

The key component of this Bayesian method is something called the truth, a latent process $\{Z(s) : s \in D\}$ over the domain of interest $D$. The monitors in that domain yield measurements from another process $\{\hat{Z}(s) : s \in D\}$ made at a finite discrete subset of $D$. The simulated data from the deterministic model output is again represented by a process $\{\tilde{Z}(s) : s \in D\}$ although, in

fact, only values of a finite collection of grid cells are available. How are the measurements and simulated data tied together?

The answer: through the truth. Modeling output can be regarded as an integral of that process, measurements, as noisy observations of it, all within a Bayesian framework, hence the name, Bayesian melding.

The model:

$$\begin{aligned}
&\hat{Z}(s) = Z(s) + e(s);\\
&Z(s) = \mu(s) + \epsilon(s);\\
&\tilde{Z}(s) = a(s) + b(s)Z(s) + \delta(s);\\
&\tilde{Z}(B) = \frac{1}{B}\int_B a(s)ds + \frac{1}{B}\int_B b(s)Z(s)ds + \frac{1}{B}\int_B \delta(s)ds;\\
&e(s) \sim N(0, \sigma_e^2 I) \text{ independent of Z(s)};\\
&\mu(s) = X(s)\beta;\\
&\epsilon \sim N(0, \Sigma(\theta));\\
&\delta(s) \sim N(0, \sigma_\delta^2 I) \text{ independent of Z(s) and e(s)},
\end{aligned} \tag{5.6}$$

where $s$ is a vector of a site's geographic coordinates and $B$, a grid cell for which model output has been provided. Thus, the model output process is a biased noisy variant of the truth.

The truth's mean function $\mu(s)$ is a polynomial in $s$s coordinates. In other words, $\mu(s) = X(s)\beta$, $X(s)$ being a polynomial function of the coordinates of $s$.

The arbitrary covariance matrix of the true underlying process $\Sigma(\theta)$ could well be, for example, a member of the Whittle–Matern class introduced in Chapter 6. In that case the covariance will have uncertain parameters $\theta$ about which inference is necessary, namely $\theta = (\sigma, \rho)$, where $\sigma$ is the variance and $\rho$ is the range.

Notice that in Equation (5.6), the simulated data process is a noisy, multiplicatively, and additively biased version of the truth. In general, these biases may depend on site location. However, taking the multiplicative bias as constant greatly simplifies the model.

This complicated model cannot be implemented without resorting to numerical techniques. Interested readers should consult Fuentes and Raftery (2005). In particular, integrals are approximated in MCMC runs by averaging over a sample of integrand values taken at a random subset of sites.

We turn now to a case where more tractable solutions are more readily available.

## 5.4 Gaussian Fields

In this section, we describe response fields whose joint distributions possess the most famous distribution in probability and statistics, named after its creator

(also commonly referred to as the *normal distribution*. That distribution, although easily described in the elementary case when responses are real-valued (see Appendix 15.1), requires more technical apparatus when responses constitute a multidimensional array. We provide that apparatus below.

### Multivariate–Normal Distribution

Consider to begin with, a random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ (T denoting transpose). Define the inner product between $\mathbf{a} = (a_1, \ldots, a_p)^T$ and $\mathbf{X}$ to be $(\mathbf{a}, \mathbf{X})_v = \mathbf{a}^T\mathbf{X}$. Then $\mathbf{X}$ has a p-dimensional multivariate Gaussian distribution, denoted $N_p(\boldsymbol{\mu}, \Sigma)$, if for every fixed vector $\mathbf{a}$,

$$(\mathbf{a}, \mathbf{X})_v \sim N[(\mathbf{a}, \mu)_v, (\mathbf{a}, \Sigma\mathbf{a})_v], \tag{5.7}$$

where $\sim$ means is distributed as, $\mu^T = (\mu_1, \ldots, \mu_p)$ and $\Sigma : p \times p = (\Sigma_{ij})$. This definition implies in particular that:

1. For each i, $X_i \sim N(\mu_i, \sigma_{ii})$ so that in particular, $\mu_i = E(X_i)$ and $\sigma_{ii} = E(X_i - \mu_i)^2$;
2. $\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j)$.

We write $E(\mathbf{X}) = \mu$ and $\Sigma = \mathrm{Cov}(\mathbf{X}) = E(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T$. Thus, in particular, $\Sigma$ is a symmetric matrix so that $\Sigma^T = \Sigma$. Hence, $(\Sigma^T a, a)_v = (\Sigma a, a)_v = (a, \Sigma a)_v$. More generally,

$$\mathrm{Cov}(\mathbf{U}, \mathbf{V}) \stackrel{\triangle}{=} E(\mathbf{U} - E\mathbf{U})(\mathbf{V} - E\mathbf{V})^T$$

denotes the covariance between any two random vectors $\mathbf{U}$ and $\mathbf{V}$, whatever be their dimensions where $\stackrel{\triangle}{=}$ means defined by.

### Matric-Normal Distribution

We get the definition of the matric-Gaussian distribution for a random matrix $\mathbf{X} : p \times q$ by extending the one above for random vectors. Here $\mathbf{X}$ could represent, for example, the daily responses corresponding to $q$ chemical species for each of $p = 24$ hours. To define it, we need an inner product for matrices, $\mathbf{A}$ and $\mathbf{X}$, namely, $(\mathbf{A}, \mathbf{X})_m \stackrel{\triangle}{=} tr\ \mathbf{A}\mathbf{X}^T$ where $\mathbf{A}$ has the same dimensions as $\mathbf{X}$ and $tr$ denotes the trace operator (the sum of diagonal elements of any square matrix upon which it operates). Then $\mathbf{X} \sim N_{p\times q}(\mu, \Sigma)$ will mean that for every constant matrix $\mathbf{A}$, Equation (5.7) is satisfied, the inner product now being that for matrices, $(\cdot, \cdot)_m$.

Observe that for the inner product defined in the last paragraph, $(\mathbf{A}, \mathbf{X})_m = \sum_{i=i}^{p} \sum_{j=1}^{q} a_{ij} X_{ij}$. Curiously we could obtain this last result by vectorizing both $\mathbf{A}$ and $\mathbf{X}$ through the operator vec that stacks the rows of the matrix $\mathbf{X}$ into a tall vector of dimension $pq \times 1$ and using the vector inner product. While using this common device conveniently turns the problem into an

analysis of vectors, it destroys the intrinsic structure of the response and can make certain trivial results extremely difficult to prove.

But what are $\mu$ and $\Sigma$? The answers follow directly from the definition: (1) $\mu$ is the (unique) $p \times q$ matrix satisfying $(\mathbf{A}, \mu)_m = E(\mathbf{A}, \mathbf{X})_m$; (2) the (unique) linear operator mapping the space of $p \times q$ matrices into itself satisfying $V(\mathbf{A}, \mathbf{X})_m = (\mathbf{A}, \Sigma\mathbf{A})_m$ for all $p \times q$ matrices $\mathbf{X}$. By expressing these inner products explicitly, we find: (1) $\mu_{ij} = E(X_{ij})$ for all i and j; (2) $\Sigma = (\Sigma_{(ij),(i'j')})$ where the operation of $\Sigma$ is defined by $\Sigma\mathbf{A} = ((\Sigma\mathbf{A})_{ij})$ where

$$(\Sigma\mathbf{A})_{ij} = \sum_{i'=1}^{p} \sum_{j'=1}^{q} \Sigma_{(ij),(i'j')} a_{(i'j')} \tag{5.8}$$

for all i and j. By vectorizing the matrices we can express $\Sigma$ as a $pq \times pg$.

We can carry our analysis further by choosing $\mathbf{A}$ to be the matrix whose rows consist of 0s except for the $k$th element which is $\mathbf{a}_k$. Then $(\mathbf{A}, \mathbf{X})_m = (\mathbf{a}_{k\cdot}, \mathbf{X}_{k\cdot})_v$, implying $\mathbf{X}_{k\cdot} \sim N_q(\mu_{k\cdot}, \Sigma_{k\cdot})$, where $\mu_{k\cdot} = E(\mathbf{X}_{k\cdot})$ and $\Sigma_{k\cdot} = \mathrm{Cov}(\mathbf{X}_{k\cdot})$. More explicitly, $\Sigma_{k\cdot} : q \times q = (\Sigma_{(kj)(kj')})$. In a similar way, we can show the columns of $\mathbf{X}$ have a marginal normal distribution, column l having the covariance matrix, $\Sigma_{\cdot l} : p \times p = (\Sigma_{(il)(i'l)})$. Finally, we find $\mathrm{Cov}(\mathbf{X}_{k\cdot}, \mathbf{X}_{k'\cdot}) : q \times q = (\Sigma_{(kj)(k'j')})$ and $\mathrm{Cov}(\mathbf{X}_{\cdot l}, \mathbf{X}_{\cdot l'}) : p \times p = (\Sigma_{(il)(i'l')})$. Thus, $\Sigma$ provides a very rich description of the covariances among the matrix of responses, albeit at the cost of a very large number of uncertain parameters.

An important special case obtains when $\mathrm{Cov}(\mathbf{X}_{k\cdot}, \mathbf{X}_{k'\cdot}) = \Lambda_{kk'}\Omega$ for all k and $k'$. This implies a strong form of separability that gives rise to something referred to as Kronecker structure: $\mathrm{Cov}(X_{kj}, X_{k'j'}) = \Lambda_{kk'}\Omega_{jj'}$. From the previous expression we deduce that $\mathrm{Cov}(\mathbf{X}_{\cdot l}, \mathbf{X}_{\cdot l'}) = \Omega_{ll'}\Lambda$ for all $l$ and $l'$. In this case the covariance structure is said to have Kronecker structure.

That structure is commonly expressed by writing $\Sigma = \Lambda \otimes \Omega$. Equation (5.8) implies that in this case $\Sigma$'s operation can be described by $\Sigma a = \Lambda\ a\ \Omega$. It follows immediately that $\Sigma$'s inverse operator $\Sigma^{-1}$ for which $\Sigma^{-1}\Sigma = I$, the identity operator, is given by $\Sigma^{-1} = \Lambda^{-1} \otimes \Omega^{-1}$. This result would be difficult to establish by the vectorized matrices approach described above. [There, we would have $\Sigma = \Lambda \otimes \Omega = (\Lambda_{ij}\ \Omega)$, a common expression for the Kronecker product, one that we in fact use in the sequel.]

The transpose of $\Sigma$, $\Sigma^T$ also proves easy to find. The transpose is defined as the (unique) operator on the space of all $p \times q$ matrices for which $(\Sigma^T a, a)_m = (a, \Sigma\ a)_m$ for all a. It readily follows that $\Sigma^T = \Lambda^T \otimes \Omega^T = \Lambda \otimes \Omega$, since $\Lambda$ and $\Omega$ are symmetric. This result in turn implies that $\Sigma$ is symmetric.

In general, $\mathbf{X}$ is a multidimensional array, for example, a $365 \times 24 \times 5$ dimensional array when responses obtain for every one of, say 5 species, for every one of the 24 hours for every one of the 365 days in a year. However, we leave it to the reader to consider the development of a definition of a Gaussian distribution for such response arrays.

In practice, responses typically do not themselves have a joint multivariate Gaussian distribution. However, in many cases their logarithmically trans-

formed values do (see Ott 1995). We turn in Section 5.5 to this very important distribution in environmental risk analysis.

## 5.5 Log Gaussian Processes

In this section, we describe the log Gaussian (normal) random response model in some detail to emphasize how different it is from the normal. In particular, its responses are positive, unlike those of the normal and its distribution has a much heavier right tail.

We say that a positive random variable $Y$ has a log Gaussian or log normal distribution with mean $\mu$ and standard deviation $\sigma$ and write $Y \sim LN(\mu, \sigma)$ if $\log\ Y \sim N(\mu, \sigma)$. A number of properties of the log Gaussian process are easily found from those of its Gaussian partner. First

$$\mu_X = E(X) = E(\exp[\log\ X]) = E(\exp Y). \tag{5.9}$$

The latter is just the moment generating function $M(t)$ of the normally distributed Y evaluated at $t = 1$. Hence $\mu_X = exp(\mu + \sigma^2/2)$.

One point bears emphasis, namely parameters such as means of the normal and associated log normal, unlike the responses, are not connected in a simple way. Practitioners do not always recognize this fact. For instance, given data, $x_1, \ldots, x_n$ from a log Gaussian population, an obvious estimate of $\mu$ would be $\hat{\mu} = \bar{y}$, where $\bar{y} = \sum_{i=1}^{n}\ y_i$ and $y_i = \log x_i$. In other words, $\hat{\mu}$ is the log of the geometric mean of the $x$s. However, one cannot simply invert the log transformation, that is take $\hat{\mu}_X^{naive} = \exp \hat{\mu}$, to estimate $X$s expected value, in spite of the latter's obvious appeal. Equation (5.9) makes the naiveté of this estimator clear and suggests an alternative, namely $\hat{\mu}_X = exp(\hat{\mu} + \hat{\sigma}^2/2)$ where $\hat{\sigma}^2$ is say, the sample variance of the $y$s. Since $\hat{\mu}_X^{naive} < \hat{\mu}_X$, the naive estimate of $\mu_X$ seriously underestimates $X$s mean when $X$ is quite uncertain. This bias could in turn seriously understate an environmental risk associated with the response $X$.

To further emphasize the differences between the log normal and normal distributions, consider $X$s variance $\sigma_X^2$ that can readily be found in terms of $\mu$ and $\sigma^2$. To do so, we need to recognize that $E(X^2)$ is just the Gaussian moment generating function above $M(t)$ evaluated at $t = 2$; that is, $E(X^2) = M(2) = exp(2\mu + 2\sigma^2)$. It follows immediately that $\sigma_X^2 = (\mu_X)^2(exp\sigma^2 - 1)$, a quantity very different from the square of the antilogarithm of $\sigma$ unless the latter is large and $\mu = 0$.

Although simple attributes of the log Gaussian model can be found as illustrated above, in general, it does not offer anything like the tractability of the Gaussian model. For example, it is not clear how to develop a multivariate log Gaussian model nor what its properties would be. Hence, in practice, statistical analysts will (often without much consideration for substantive issues) log transform the measured responses and model the log Gaussian field. Indeed, we do that in this book.

However, measurements are recorded and generally understood in their original scales. In particular, subject area investigators will see the transformed measurements in terms of the knowledge and language of their subject. Consider, for example, the findings of Zidek et al. (1998c) that show a statistical association between log transformed daily average ozone concentrations (among other things) and daily counts of adverse health outcomes. In fact, a unit increase (appropriately defined) in log (ozone) concentrations would result in a 5% increase in these adverse outcomes. Subject area workers would need to consider these findings by taking the antilog of that unit increase. They would read that finding as saying a $100 * (\exp -1)\%$ increase in ozone would lead to a 5% increase in adverse outcomes. Imagine the skepticism with which the findings expressed this way would be greeted! (As an aside, analysts will also face technical hurdles in going back to the original measurement scale when reporting their findings. Fortunately, estimates of marginal means and variances of the log normal distribution can readily be found using the formulas above.)

In rebuttal, the statistician might well point out that the scales of measurement are arbitrary. Why not use the log scale in the first place? There are substantive reasons why the logarithmic transformation makes substantive sense because of the cascade effect in the formation of some secondary pollutants. Acidity (pH) is measured on a log scale. The Richter scale measures the logarithm of the amplitude of waves recorded by seismographs. (In a different context, decibels are measured on a log scale. So are stock indices typically, even though the latter are nominally in dollars, because of the divisibility of money; gains of the whole must be the same as the gains on any part.) Furthermore, as we have shown by examples in Section 4.2, conceptual scales such as ppb for ozone are, in reality, transformed measurements of a surrogate such as UV light absorption. In summary, the appropriate scale for response measurement is itself a subjective choice and the logarithmic scale in the case of environmental response fields has much to recommend it.

## 5.6 Wrapup

That completes our general discussion of modeling with an all too brief survey of the variety of methods that have been developed. The importance of such modeling in the assessment and management of environmental risk has led to a great deal of often ingenious work on this broad topic. However, the subject is undergoing a tremendous amount of current development, making it difficult to be completely comprehensive in our treatment. The authors hope they have covered at least the most important of those approaches. Moreover, they have tried to give some sense of their strengths and weaknesses. More can be found in later chapters.

However, the book now turns to much more specialized topics with an emphasis on Gaussian fields. Hence, the next chapter introduces the spatial covariance, a topic of great importance in that context.

# Part II: Space–Time Modeling

# 6

# Covariances

*Science is nothing but developed perception, interpreted intent, common sense rounded out and minutely articulated.*

*George Santayana*

The covariance between any two random variables measures the strength of their relationship. The covariance structure of a spatial random field indicates the strength of the relationships between variables representing its levels at different domain locations. For some, such as those in geological applications, that structure may well be homogeneous, meaning these relationships are similar over the entire geographical domain; their strength is similar in all directions. For others, particularly in environmental contexts, that structure can be highly nonhomogeneous, the strengths of relationships depending strongly on location and direction.

Since the covariance structure reflects the strengths of relationship between random variables within the field, it plays an important role in the spatial prediction problem. However, modeling such structures is not a simple task because covariances, besides capturing the features of the random fields such as those mentioned above, must satisfy certain mathematical properties. The modeling problem proves much more complicated for nonhomogeneous random fields than homogeneous ones. For the latter, one simple mathematical expression may adequately capture key features of the covariance field because of its similarity over locations and directions. On the other hand, simple mathematical expressions will usually be inadequate for nonhomogeneous random fields because of their varying behavior over locations and directions.

The covariance modeling problem can be further complicated by a lack of data. For example, in geological applications the cost of obtaining a measurement is generally quite substantial, so often only single measurements are made at a small number of locations scattered over the geographical field. For environmental applications, data are generally from networks of monitoring stations and although the cost of collecting repeated measurements at any one station may not be very high, the operational cost for networks with large numbers of stations is prohibitively expensive. Thus in such applications, repeated measurements, sometimes for multiple pollutants, are often available but only for a limited number of monitoring locations.

This chapter presents a variety of approaches to modeling spatial covariance structures. First, we discuss basic statistical concepts for characterizing covariance structures of spatial processes, including definitions of moments and variograms. The latter are mathematically related to the covariance function but more convenient for certain purposes in spatial prediction.

Second, we introduce the concept of stationarity that describes characteristics of certain (homogeneous) random fields . As noted above, spatial covariance modeling for nonhomogeneous fields can be very complicated and generally impractical or impossible without imposing some sort of restrictions on the random field. Stationarity allows us to impose specific restrictions at various levels of a random field model. For example, a random field that is second-order stationary would have its mean and variance not depend on location and hence be much easier to estimate.

We go on to describe characteristics of suitable covariance models for stationary processes. Suitability criteria are required since covariance models must meet certain technical conditions such as *nonnegative definiteness*. Several commonly used models are given, notably for processes with isotropic stationary covariance structure; for these, the correlation between any two locations in the field depends solely on the distance between them.

Finally, we discuss methods for modeling the spatial covariance of nonstationary processes. Since their intersite covariances depend on location as well as direction, any naive estimation procedure would require data from all locations of interest. The use of such a procedure would usually present the modeler with an insurmountable challenge since data in such abundance are almost never available. Hence he must rely on more sophisticated approaches. Various such approaches have been proposed in recent years. We find that of Sampson and Guttorp (1992) particularly appealing. They create an imaginary (pseudo) region and map the locations of interest in the geographical region onto it by using a complicated mathematical function. They constrain that mapping so that the covariance structure of the random field can be expressed as a function depending only on the distance between the locations in the new region. Fuentes (2001) and Higdon et al. (1999) propose methods primarily based on the need for mathematical tractability and computational convenience. They represent the random process as a weighted combination of local processes, each stationary and weighted in accordance with location and direction. A process with such a representation can have a nonstationary covariance structure. However, the local processes would usually be virtual or latent with no physical interpretation.

## 6.1 Moments and Variograms

### 6.1.1 Finite-Dimensional Distributions

Let $Y$ denote an environmental random field over a geographical region, for example, monthly average levels of ozone concentrations over a city. The random

field $Y$ is said to have the finite-dimensional cumulative distribution (or CDF) $F$ if for any finite set of locations in the geographical region, $\{s_1, \ldots, s_n\}$, and any positive integer $n$,

$$F_{s_1,\ldots,s_n}(x_1, \ldots, x_n) \equiv P\{Y(s_1) \leq x_1, \ldots, Y(s_n) \leq x_n\},$$

where $P$ denotes probability.

### Moments

Define the $k$th-order moment of the random field $Y$ at any location $s$ as

$$E[Y(s)]^k \equiv \int x^k dF_s(x)$$

provided this integral exists, where $dF_s(x)$ denotes the differential element of probability allocated to $x$ by the distribution $F_s$. ($E|Y(s)|^k < \infty$ ensures the existence of the $k$th-order moment and all moments of order less than $k$.) For some random fields such as those with Gaussian finite-dimensional distributions, all moments exist. For others such as those with Cauchy finite-dimensional distributions, few or even no moments exist.

### Expectation

The expectation of a random field $Y$ is defined to be its first-order moment

$$\mu(s) \equiv E[Y(s)]$$

for any location $s$ (provided it exists). The expectation is generally (but not always) allowed to depend on $s$. It is also called the *mean* or *expected value*.

### Variance and Covariance

The variance of a random field $Y$ is defined as the second-order moment about the expectation $\mu(s)$,

$$var[Y(s)] \equiv E[Y(s) - \mu(s)]^2$$

for any location $s$ (provided it exists). Like the expectation $\mu(s)$ the variance generally depends on $s$.

An important variant of the second-order moment, the *covariance* is defined as

$$C(s_1, s_2) \equiv E[(Y(s_1) - \mu(s_1))(Y(s_2) - \mu(s_2))]$$

for any two locations $s_1$ and $s_2$. The covariance is generally allowed to depend on the locations of the associated variables and its existence is ensured by that of the variance (Schwarz's inequality). Note that when $s_1 = s_2 = s$, the covariance becomes the variance, i.e.,

$C(s_1, s_1) \equiv var[Y(s)]$.

The expectation and covariance completely determine the distribution of a Gaussian random field. That is, although these fields possess moments of all orders, only the first two moments are needed to completely specify the distribution.

### Variogram

The *variogram* between any two locations, $s_1$ and $s_2$, on the plane supporting a random field, is defined as the variance of the difference between $Y(s_1)$ and $Y(s_2)$,

$$\begin{aligned} 2\gamma(s_1, s_2) &\equiv var[Y(s_1) - Y(s_2)] \\ &= E[(Y(s_1) - Y(s_2)) - (\mu(s_1) - \mu(s_2))]^2. \end{aligned}$$

The function $\gamma(s_1, s_2)$, called a *semi-variogram*, is closely related to the covariance for random fields having special features that we describe below. Matheron (1962) introduced the terms *variogram* and *semi-variogram*, although the concept had been used in earlier scientific publications (Kolmogorov 1941, De Wijs 1951, Jowett 1952, and Matern 1960, among others). More detail can be found in Cressie (1991).

## 6.2 Stationarity

Stationarity is a concept describing how some random fields $Y$ behave across the geographical region over which they obtain. For example, the probability distribution at any specific location is the same for all locations. However, that property is so strong that few processes will possess it. Thus, several weaker versions of stationarity have been defined to enable a finer characterization of the stochastic nature of random fields. We now describe them.

### Strict Stationarity

A random field $Y$ is said to be *strictly stationary* if for any vector $h$ the finite-dimensional distributions of $\{Y(s_1), \ldots, Y(s_n)\}$ and $\{Y(s_1+h), \ldots, Y(s_n+h)\}$ are identical for an arbitrary $n$. That is, the random field is invariant under translation.

Strict stationarity implies that moments of any order, if they exist, will not depend on location. Thus this condition imposes a very strong requirement that few random fields will meet, making it of little use in applications. However, in environmental as well as geostatistical applications weaker versions of stationarity such as those limited to the first two moments may be sufficient to provide a foundation for modeling and analysis. In particular, random fields with Gaussian finite-dimensional distributions are fully characterized by the first two moments.

### Second-Order Stationarity

A random field is said to be *second-order stationary* if: (a) the expectation exists and is not a function of the location, and (b) the covariance exists and depends only on the vector $h$ separating the two locations. That is, a second-order stationary random field would have, for all locations $s$ and any $h$,

$$\begin{aligned} \mu(s) &= E[Y(s)] = \mu \\ C(s+h,s) &= C(s+h-s) = C(h). \end{aligned}$$

For $h = 0$, the covariance becomes the variance that second-order stationarity implies must equal $C(0)$. That is,

$$var[Y(s)] = C(s,s) = C(0).$$

Thus second-order stationarity implies the existence of the variance that does not depend on the location $s$.

### Remarks

- The $C(h)$ function is sometimes called a *covariogram*. In the field of time-series, it is called the autocovariance function.
- When $C(0) > 0$, the *correlogram* $\rho(h)$, the correlation between two points separated by a vector $h$, is defined as

  $$\rho(h) = C(h)/C(0).$$

  In time-series, it is called the *autocorrelation function*.

Second-order stationarity implies the variogram, i.e., $var[Y(s) - Y(s+h)]$ can be written as

$$\begin{aligned} var[Y(s) - Y(s+h)] &= var[Y(s)] + var[Y(s+h)] - 2\ cov[Y(s), Y(s+h)] \\ &= C(0) + C(0) - 2C(h) \\ &= 2[C(0) - C(h)]. \end{aligned}$$

The first equality is true for the variance of the difference of any two random variables while the second is implied by second-order stationarity. Thus the semi-variogram, a function of only the separating vector $h$, can be expressed as

$$\gamma(h) = C(0) - C(h).$$

It is worth noticing that second-order stationarity requires the existence of the covariance that is also a function depending only on $h$. This requirement in turn ensures not only the existence of the variance but also that the variogram, or equivalently the semi-variogram, depends only on the vector separating the locations. However, this last property is implied by a slightly weaker property of a random field known as intrinsic stationarity.

Specifically, a random field is said to be intrinsically stationary if (a) its expectation exists and is not a function of location, and (b) for any two locations separated by a vector $h$, the variance of the difference $[Y(s) - Y(s + h)]$ exists and is a function of $h$,

$var[Y(s) - Y(s + h)] = 2\gamma(h)$.

This property of intrinsic stationarity is slightly weaker than its second-order cousin since it assumes only the existence of the variance of the difference. That does not imply the variance or covariance exists as required by second-order stationarity. The reverse is always true as described above.

## 6.3 Variogram Models for Stationary Processes

In this section, we describe models that can play the role of variograms or semi-variograms for the second-order stationary processes. Their suitability derives from their possession of certain mathematical conditions and properties consistent with those required of the covariance $C(h)$ they induce, as described in the previous section.

### 6.3.1 Characteristics of Covariance Functions

#### Nonnegative and Positive Definiteness

The covariance as defined above must have the following properties:

- $C(0)$ must be greater than or equal to zero since $C(0) = Var[Y(s)] \geq 0$ for any $s$;
- $C(h) = C(-h)$ for any vector $h$ since the covariance is an even function;
- $|C(h)| \leq C(0)$ where $|.|$ denotes the absolute value, this inequality being derived by applying Schwarz's inequality (Shorack and Wellner 1986).

Furthermore, the covariance matrix for $Y(s_1), \ldots, Y(s_n)$, an $n \times n$ matrix denoted by $\Sigma = (\Sigma_{ij})$, must also be nonnegative definite where $\Sigma_{ij} = C(s_i - s_j)$ equals the covariance between the responses at the two corresponding locations. In other words, for any nonzero vector $a$, the quadratic form $a^T \Sigma a$ must be greater than or equal to 0. Explicitly, the nonnegative definiteness condition can be written as

$$\begin{aligned} a^T \Sigma a &= \sum_i \sum_j a_i a_j C(s_i - s_j) \geq 0 \\ &= \sum_i \sum_j a_i a_j C(h_{ij}) \geq 0, \end{aligned}$$

where $a_i$ and $a_j$ are elements of $a$ and $h_{ij}$ denotes the vector separating $s_i$ and $s_j$. The function $C(h)$ satisfying this condition is also said to be *nonnegative definite*.

Hence under second-order stationarity, suitable models for the variogram, or semi-variogram, defined through $\gamma(h) = C(0) - C(h)$ in this case, must allow C to satisfy the above properties, as well as the nonnegative definiteness condition. In particular they must be nonnegative even functions of $h$, and ensure that the nonnegative definiteness condition for the covariance function is satisfied. Obviously not just any mathematical function would provide a suitable variogram model. Conditions that do ensure suitability can be found, for example, in Doob (1953), Journel and Huijbregts (1978), Cressie (1991, 1993), and Wackernagel (2003).

**Remarks**

- If one restricts the variance to be greater than 0, then the condition on the quadratic form above will become $\sum_i \sum_j a_i a_j C(h_{ij}) > 0$. Such a $C(h)$ function is said to be *positive definite*.
- In environmental and geographical practice, it is often sensible, but not always necessary, to assume that the covariance decreases as the distance between the two locations $h$ increases; i.e., $\gamma(h)$ is an increasing function of $h$.

**Anisotropy and Isotropy**

The covariance $C(h)$ is a function of vector $h$, specified by its length and direction. In environmental and geostatistical applications, covariance functions often exhibit different behavior in different directions. Random fields with such covariances are called *anisotropic*. On the other hand, when $C(h)$ depends only on the length of $h$, denoted by $|h|$, the field is said to be *isotropic*. In that case the strength of association within the field is the same in every direction. This kind of association has been widely assumed in geostatistical applications.

In cases where the anisotropic covariance function $C(h)$ can be represented as an isotropic covariance function $f(|h_1|)$ by linearly transforming the vector $h$ to $h_1$, the anisotropy is called *geometric anisotropy*. Here the geometric anisotropy can be reduced to isotropy by a linear transformation of the coordinators of $h$. More details on geometric anisotropy can be found in Journel and Huijbregts (1978), Cressie (1991), and Wackernagel (2001, 2003).

## 6.4 Isotropic Semi-Variogram Models

We now describe various models commonly seen in the literature for an isotropic variogram $2\gamma(|h|)$ . Because of isotropy, we have for simplicity used $|h|$, the length of $h$, as the argument for the semi-variogram function $\gamma(\cdot)$.

The definition $\gamma(h) = C(0) - C(h)$ implies $\gamma(0) = 0$. However, $\gamma(h)$ need not approach 0 as $h$ tends towards zero. That is, the semi-variogram function

can be discontinuous at the origin. That mysterious discontinuity, called the nugget effect (Matheron 1962), actually reflects local variability in the random field. At the other extreme, when the semi-variogram approaches a limiting value as the separation $h$ tends towards infinity, the limit is called a sill (Journel and Huijbregts 1978). More discussion on the behavior of semi-variogram functions can be found in Journel and Huijbregts (1978), Cressie (1991), and Wackernagel (2003).

All models described here can validly be extended to a domain of at least three-dimensions unless otherwise stated. In other words, they can be used even when say depth or elevation are added as coordinates to the location vector. As well, they all have a common component $\gamma(0) = 0$ and so, for brevity, it is not listed for each model.

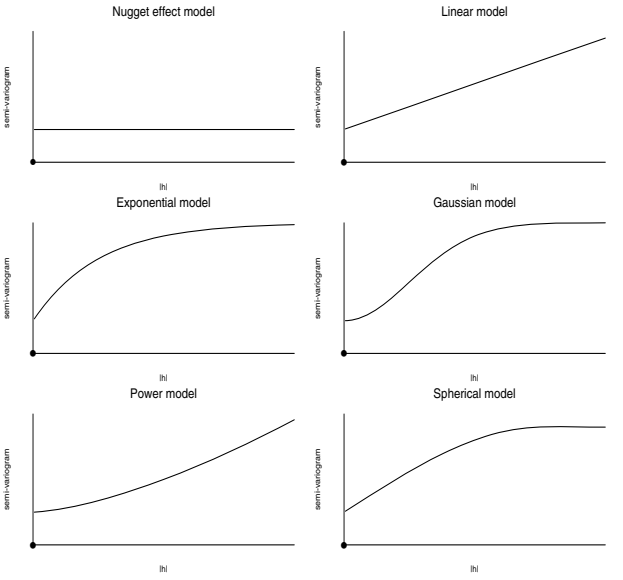


**Fig. 6.1.** Sketches of various semi-variogram models.

### Nugget Effect Model

$$\gamma(h) \equiv C_0 \geq 0$$

for $|h| > 0$. This semi-variogram model, displayed in Figure 6.1, reveals a constant nugget effect at all distances. Equivalently the spatial correlation is constant for any distance in the random field.

### Exponential Model

$$\gamma(h) = C_0 + b\left(1 - exp\left\{-\frac{|h|}{a}\right\}\right)$$

for $|h| > 0$, where $C_0 \geq 0$, $b \geq 0$ and $a \geq 0$. This semi-variogram model, displayed in Figure 6.1, increases exponentially as the distance $|h|$ increases, the sill being $C_0 + b$. Equivalently, the covariance decreases exponentially as $|h|$ increases. The parameter $a$ determines how sharply the covariance drops off. When $|h| = 3a$, the covariance has dropped to about 95% of its maximal value at the origin. The distance corresponding to such a 95% drop has been termed the practical range (Journel and Huijbregts 1978; Wackernagel 2003). This semi-variogram model behaves linearly near the origin.

**Gaussian Model**

$$\gamma(h) = C_0 + b\left(1 - exp\left\{-\frac{|h|^2}{a}\right\}\right)$$

for $|h| > 0$, where $C_0 \geq 0$, $b \geq 0$ and $a \geq 0$. The Gaussian semi-variogram again increases exponentially as $|h|$ increases (see Figure 6.1). This semi-variogram model has a practical range of $\sqrt{3}a$ with a sill of $C_0 + b$ and behaves parabolically at the origin (Matheron 1972; Journel and Huijbregts 1978).

**Stable Model**

$$\gamma(h) = C_0 + b\left(1 - exp\left\{-\frac{|h|^\lambda}{a}\right\}\right)$$

for $|h| > 0$, where $C_0 \geq 0$, $b \leq 0$, and $0 < \lambda \leq 2$. The Gaussian and the exponential models are special cases of this class of semi-variogram models, studied by Schoenberg (1938).

**Whittle–Matern Model**

$$\gamma(h) = C_0 + b(1 - (|h|/a)^\nu K_\nu(|h|/a)$$

for $|h| > 0$ where $C_0 \geq 0$, $\nu > 0$ , $a \geq 0$, and $K_\nu$ is a modified Bessel function of order $\nu$ (see Abramowitz and Stegun 1970, for details). This semi-variogram model with order 1 was originally suggested by Whittle (1954). This is an intermediate choice between the exponential and the Gaussian ones. For example, the exponential semi-variogram model is a special case with $\nu = .5$ (Fuentes 2001). This model, subsequently generalized by Matern (1960), has a sill of $C_0 + b$.

**Rational Quadratic Model**

$$\gamma(h) = C_0 + b\left\{\frac{|h|^2}{1 + |h|^2}\right\}$$

for $|h| > 0$ where $C_0 \geq 0$ and $b \geq 0$. Schoenberg (1938) showed this to be a valid semi-variogram model with a sill of $C_0 + b$, attained as $|h|$ approaches infinity.

**Spherical Model**

$$\gamma(h) = \left\{ \begin{array}{ll} C_0 + b\left(\frac{3|h|}{2a} + \frac{|h|^3}{2a^3}\right) & 0 < |h| \leq a \\ C_0 + b & |h| > a \end{array} \right\},$$

where $C_0 \geq 0$, $b \geq 0$, and $a \geq 0$. This semi-variogram model, studied by Matheron (1965, 1970), steadily increases from the nugget effect of $C_0$ to the sill of $C_0 + b$ when $h \geq a$, as displayed in Figure 6.1. Equivalently, the correlation steadily decreases from its highest value $b$ near the origin to zero as the distance increases and reaches the range $a$. This type of semi-variogram is widely used in mining applications (Journel and Huijbregts 1978; Wackernagel 2001).

**Cauchy Model**

$$\gamma(h) = C_0 + b\left(1 - 1/\left[1 + (|h|/a)^2\right]^{\lambda}\right)$$

for $|h| > 0$ where $C_0 \geq 0$ and $\lambda \geq 0$. This valid semi-variogram (Yaglom 1986), has a sill of $C_0 + b$ and behaves linearly near the origin.

**Triangular Model**

$$\gamma(h) = \left\{ \begin{array}{ll} C_0 + b\frac{|h|}{a} & 0 \leq |h| \leq a, \\ C_0 + b & |h| > a \end{array} \right.$$

where $C_0 \geq 0$, $b \geq 0$, and $a \geq 0$. This semi-variogram model, valid for one-dimensional space (Yaglom 1986), behaves linearly near the origin.

**Hole-Effect Model**

$$\gamma(h) = C_0 + b\left[1 - a\frac{\sin(|h|/a)}{|h|}\right]$$

for $|h| > 0$ where $b \geq 0$ and $a \geq 0$. This semi-variogram model reveals a *hole effect* in that its growth is not monotonic with respect to $|h|$. This semi-variogram, also called a *wave* model, behaves parabolically near the origin.

**Linear Model**

$$\gamma(h) = C_0 + b|h|$$

for $|h| > 0$ where $C_0 \geq 0$, and $b \geq 0$. Here the semi-variogram increases linearly from its nugget as the distance $|h|$ increases (see Figure 6.1). Equivalently, the spatial correlation drops off linearly as the distance increases. The parameter $b$ determines how fast the correlation drops off, $b = 0$ corresponding to the nugget effect model. When $b > 0$, this semi-variogram is unbounded.

**Power Model**

$$\gamma(h) = C_0 + b|h|^\lambda$$

for $|h| > 0$ where $C_0 \geq 0$, $b \geq 0$, and $0 \leq \lambda < 2$. Yaglom (1957), Whittle (1962), and Christakos (1984) studied this general model for unbounded semi-variograms. A specific example with $\lambda = 1.3$ and $b = .5$ is seen in Figure 6.1. Its unbounded character implies possibly negative correlations because $C(h) = C(0) - \gamma(h)$ where $C(0)$ is fixed and $\gamma(h)$ can be arbitrarily large.

**De Wijsian Model**

$$\gamma(h) = \frac{3}{2} b \log(|h|^2 + a)$$

for $|h| > 0$ where $a \geq 0$ and $b \geq 0$ (De Wijs 1951). This semi-variogram behaves linearly near the origin. Such logarithmic semi-variogram models have been extensively studied in the early stages of geostatistical research (e.g., Krige 1951, and Matheron 1955, 1962). More discussion can be found in Journel and Huijbregts (1978).

## 6.5 Correlation Models for Nonstationary Processes

Spatially stationary processes have been successfully used in geostatistical applications for several decades dating back to Krige (1951). However, in some situations the assumption of homogeneous covariance behavior across the entire domain of the field proves untenable, particularly in environmental applications (see Escoufier et al. 1984, Chami and Gonzalez 1984, Haas 1990, Sampson and Guttorp 1992, Brown et al. 1994a, Higdon et al. 1999, Fuentes 2001, Le et al. 2001, and Kibria et al. 2002 among others). Recognizing this problem, several authors in recent years have developed new approaches for dealing with nonstationary processes. We sketch these approaches below.

### 6.5.1 The Sampson–Guttorp Method

Sampson and Guttorp (1992) propose a highly original, nonparametric approach to estimate the spatial covariance structure for the entire random field without assuming stationarity. Let's call it the SG method for short. Briefly, their method first constructs a smooth mapping function between locations in the geographic space, where stationarity of the random field is not assumed, to locations in a (virtual) new space where isotropy is assumed. Sampson and Guttorp call geographic space *G-space,* and the new one *dispersion space* or *D-space.* By means of this construction, an isotropic variogram model can then be fitted using the observed correlations and distances in D-space. The smooth

mapping function, estimated from the observed correlations between monitoring stations in conjunction with the estimated isotropic variogram model, then estimates spatial correlations between any two locations of interest.

The action of the SG method can be demonstrated using Figure 6.2 based on the monthly average levels of multiple pollutants ($O_3$, $SO_4$, $NO_3$) obtained at seven stations in Southern Ontario where the spatial covariance component among seven stations is estimated as described in Brown et al.(1994a). The right panel shows the corresponding D-space coordinates, the result of applying the mapping function to a rectangular grid in G-space. The left panel shows the fitted variogram in D-space. The results in the panels can be used to estimate spatial correlations between any two points in the G-space. This could be done, for example, by first identifying the corresponding points in D-space using the grid. Their interdistance in D-space can then be calculated. Finally, the fitted variogram at that distance can be evaluated. The result: an estimate of their spatial correlations.

Note that the SG method does not require the units of the D-plane coordinates to be explicitly specified. Furthermore, there is a built-in smoothing parameter in the mapping function to control the distortion between the G- and D-spaces. This feature allows users to ensure that the grid is not folded in the D-space and hence maintain the spatial interpretability of the correlations; that is, locally, the closer the stations are together the higher their between-response correlations.

To give a more precise description of the method, let $f : R^2 \to R^2$ be a 1-1 smooth nonlinear mapping from a G-space (including all locations of interest in the geographic space) to a D-space. For a location $s_i$ in G-space, the corresponding location $z_i$ in D-space is obtained as $z_i = f(s_i)$; equivalently $s_i = f^{-1}(z_i)$ where $f^{-1}$ denotes the inverse mapping of $f$.

The variogram of the random field $Y$ between locations $s_i$ and $s_j$ can be expressed in terms of D-space locations

$$\begin{aligned} 2\gamma(s_i, s_j) &\equiv var[Y(s_i) - Y(s_j)] \\ &= var[Y(z_i) - Y(z_j)] \\ &= 2g(|h_{ij}^D|), \end{aligned}$$

where $z_i$ and $z_j$ are the corresponding locations in D-space, $|h_{ij}^D| = |z_i - z_j|$ is the distance in D-space, and $g$ denotes the semi-variogram in D-space. Since isotropic stationarity on D-space is assumed, the semi-variogram $g$ is just a function of $|h_{ij}^D|$. Suitable semi-variogram models such as those described in the previous section could be used for $g$. Note that Sampson and Guttorp refer to the variogram between locations in G-space, $2\gamma(s_i, s_j)$, as *dispersion* to emphasize that the random field is nonstationary in the G-space.

With this framework, Sampson and Guttorp (1992) propose a two-step approach for estimating $g$ and $f$ using sample dispersions between locations $s_1, \ldots, s_n$ in G-space denoted by $d_{ij}^2$. First a multidimensional scaling approach (Mardia et al. 1979) is used to form a new two-dimensional repre-

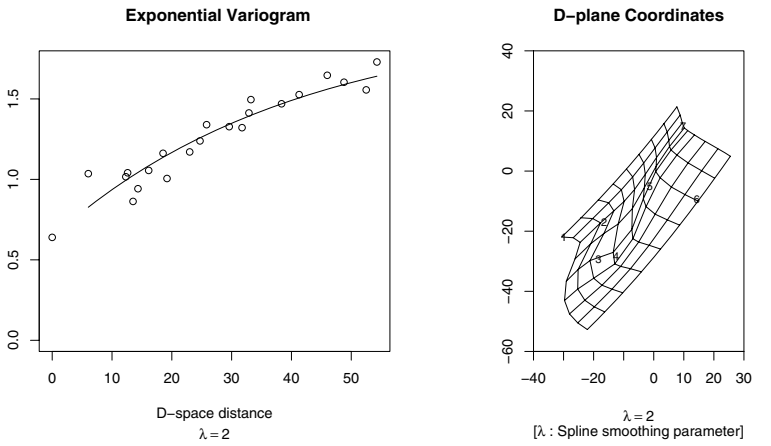


Fig. 6.2: Fitted variogram using D-space interstation distance (left panel) and D-space coordinates (right panel) with smoothing parameter.

sentation $(z_1, \ldots, z_n)$ of the G-space locations $(s_1, \ldots, s_n)$, the isotropy assumption being appropriate for the new representation. A semi-variogram $g$ is estimated using the intersite distances between the new locations in D-space, $(z_1, \ldots, z_n)$. The multidimensional scaling algorithm determines a new representation of $z_i$ so that

$$\delta(d_{ij}) \equiv \delta_{ij} \approx |z_i - z_j|,$$

where $\delta$ is a monotone function. Solving this relationship yields an estimate for $g$ since

$$d_{ij}^2 \equiv (\delta^{-1}(\delta_{ij}))^2 \approx g(|z_i - z_j|).$$

The new representation of $z_i$ is selected so that the intersite distances in D-space $|h_{ij}^D|$ minimize the following stress criterion

$$\min_\delta \left[ \sum_{i<j} \frac{(\delta(d_{ij}) - h_{ij}^D)^2}{\sum_{i<j} (h_{ij}^D)^2} \right],$$

that minimum being taken over all monotone functions.

Second, the thin-plate spline approach (Wabba and Wendelberger 1980) is used to estimate the smooth mapping $f$ between the original locations $s_i$ and the new ones $z_i$. Specifically,

$$f(s) = \alpha_0 + \alpha_1 s^{(1)} + \alpha_2 s^{(2)} + \sum_{i=1}^{n} \beta_i u_i(s),$$

where $u_i(s) = |s - s_i|^2 \log|s - s_i|$ and $s^{(j)}$ indicates the $j$th coordinate of the location $s$. The parameters to be fitted are $\alpha$s and $\beta$'s. For the bivariate problem, Sampson and Guttorp (1992) compute the function $f$ as two thin-plate splines, $f_1$ and $f_2$ for the two coordinates of $z_i$. A smoothing parameter is incorporated into this second step, allowing users to choose a desirable level of smoothness.

For any specified value smoothing parameter $\lambda$, the parameters, $\alpha$s and $\beta$s are chosen to minimize

$$\sum_{i=1}^{n}\sum_{x=1}^{2}(z_{ix} - f_j(s_i))^2 + \lambda[J_2(f_1) + J_2(f_2)],$$

where $z_{i1}$ and $z_{i2}$ denote the first and second coordinate of $z_i$, while $J_2$ measures the smoothness of the functions defined as

$$J_2(f) = \int_R \left[ \left( \frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left( \frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left( \frac{\partial^2 f}{\partial x_2^2} \right)^2 \right] dx_1 dx_2, \ j = 1, 2.$$

This construction ensures that $\beta \to 0$ when $\lambda \to \infty$; that is, in this case, it is a simple linear mapping. See Sampson and Guttorp (1992) for more details.

With the estimated $\hat{f}$ and $\hat{g}$, the variogram between any two locations $s_1$ and $s_2$ in G-space can be estimated by:

- Obtaining the corresponding locations in D-space through

  $$z_j = f(s_j) \qquad \text{for } \ j = 1, 2;$$

- Then calculating the D-space distance $|h_{12}^D|$ between $z_1$ and $z_2$;
- And finally, evaluating $\quad 2\gamma(h) = 2\hat{g}(|h_{12}^D|)$.
  Equivalently the covariance between the locations can be estimated by

  $$C(h) = C(0) - \hat{g}(|h_{12}^D|),$$

  where $C(0)$ is the variance.

To allow for the nonconstant variance field, the above approach can be first applied with the correlation matrix and then any estimate of the variance field can be simply incorporated. The resulting covariance matrix is ensured to be nonnegative definite through this construction.

The SG method has been successfully used in a wide range of environmental applications due to its flexibility in modeling nonstationary features (see, for example, Monestiez et al. 1993, Guttorp et al. 1992, 1993, 1994, Brown et al. 1994a, Le et al. 1997, 2001, Meiring et al. 1998, Kibria et al. 2002). Variants of this approach based on the maximum likelihood principle have been studied by Mardia and Goodall (1993) and Smith (1996). The SG approach has recently been enhanced by putting it into a Bayesian framework that accounts for model uncertainty (see Damian et al. 2001 for details).

R codes for this method are currently available online and free of charge. Instructions for downloading and R tutorials for using the software in real applications are given in Chapter 14.

### 6.5.2 The Higdon, Swall, and Kern Method

Higdon et al. (1999) propose a process convolution approach for constructing two-dimensional, nonstationary Gaussian processes, thus allowing the spatial dependence structure to vary as a function of location. Their approach is developed by first representing a stationary Gaussian process as a moving average of a Gaussian white noise process with a normal convolution kernel and second, generalizing the kernel to allow for nonstationarity. This relatively simple construction results in valid nonstationary Gaussian processes for a wide range of kernel functions, at least in principle.

Specifically, a process $x(s)$ is said to be a white noise process if

$$\int_A x(u)du \sim N(0, b^2 area(A))$$

where $N(\mu, \sigma^2)$ denotes a Gaussian distribution with mean $\mu$ and variance $\sigma^2$; here $area(A)$ denotes the area of region $A$ while $b$ is a constant. Let $Y(s)$ be any stationary Gaussian process at location $s$ over a two-dimensional spatial domain $R^2$ having a correlogram $\rho(h)$ given by

$$\rho(h) = \int_{R^2} k(s)k(s-h)ds.$$

The process $Y(s)$ can then be expressed as the convolution of a Gaussian white noise process $x(s)$ with a kernel $k(s)$ through

$$Y(s) = \int_{R^2} k(s-u)x(u)du.$$

For instance, if $k(s)$ were the two-dimensional standard normal kernel defined as

$$k(s) = \frac{1}{2\pi} exp\left\{-\frac{1}{2}s^T s\right\}$$

then the process $Y(s)$ would have the usual isotropic Gaussian correlation function,

$$\rho(h) = exp\{-h^T h\}.$$

The above representation can be generalized by using a smoothing kernel, denoted by $k_s$, which depends on spatial location $s$. The process resulting from this process convolution approach defined as

$$Y(s) = \int_{R^2} k_s(u)x(u)du,$$

is a nonstationary Gaussian process with correlation between two locations $s$ and $s_1$ given by

$$\rho(s, s_1) \propto \int_{R^2} k_s(u)k_{s_1}(u)du.$$

This construction would work for any kernel $k(s)$ satisfying

$$\sup \int_{R^2} k_s^2(u)du < \infty.$$

Higdon et al. (1999) choose a bivariate Gaussian kernel for their application:

$$k_s(s) = \frac{1}{2\pi}|\Sigma|^{-1/2} exp\left\{-\frac{1}{2}s^T \Sigma_s^{-1} s\right\},$$

where $\Sigma_s$ is a function of location $s$. Since there is a 1-1 mapping from a bivariate Gaussian distribution to its one standard deviation ellipse, $\Sigma_s$ is obtained by varying the ellipses spatially. The two foci of each ellipse at location $s$ are randomly drawn from two independent Gaussian distributions that in turn define $\Sigma_s$ (see Higdon et al. 1999 for more details). The construction of $k_s$ is motivated mostly by the pragmatic considerations of mathematical tractability and computability. Further work is needed to assess the applicability of the process convolution method.

### 6.5.3 The Fuentes Method

Fuentes (2001) proposes another approach for constructing nonstationary processes. Her method assumes that a nonstationary process representing the random field is a weighted average of local isotropic stationary processes that are uncorrelated with each other. In particular, the geographical region is divided into well-defined subregions, each having a local isotropic stationary process. The local stationary processes, assumed to be independent, have spatial covariance representing locally the spatial structure of the nonstationary process. With weights appropriately chosen, for example, to be positive, this construction yields valid nonstationary processes with local isotropic stationarity and parameters of the local processes allowed to vary across the region. The parameters are estimated by means of a method based on the spectral density.

More precisely, Fuentes represents the nonstationary process $Y(s)$ as a linear combination of local, orthogonal stationary processes $Y_i(s)$

$$Y(s) = \sum_{i=1}^{k} Y_i(s) w_i(s),$$

for $i = 1, \cdots, k$, i.e., where $cov(Y_i(s), Y_j(s)) = 0$ for $i \neq 0$. The geographical region is divided into $k$ well-defined subregions $S_1, \ldots, S_k$ and $Y_i(s)$ is a local isotropic stationary process in a subregion $S_i$. The weights, $\{w_i(s)\}$ come from a positive kernel function centered at the centroid of $S_i$.

The covariance between any two locations $s_1$ and $s_2$ in the geographical region can be written as

$$\begin{aligned}\mathrm{Cov}(Y(s_1), Y(s_2)) &= \sum_{i=1}^{k} w_i(s_1) w_i(s_2) cov(Y_i(s_1), Y_i(s_2)) \\ &= \sum_{i=1}^{k} w_i(s_1) w_i(s_2) C_{\theta_i}(|h|),\end{aligned}$$

where $C_{\theta_i}(|h|)$, having parameter $\theta_i$ and representing the covariance between two locations $s_1$ and $s_2$ with respect to the local process $Y_i(.)$, is a function of only the distance $|h|$ due to the isotropic stationarity. Since the parameter $\theta_i$ could change from subregion to subregion, $\mathrm{Cov}(Y(s_1), Y(s_2))$ is generally a function of not only $|h|$ but also the locations of $s_1$ and $s_2$ and hence the process $Y(s)$ is nonstationary.

For parameter estimation, Fuentes (2001) proposes a spectral density approach that begins by transforming the covariance functions to their spectral densities (i.e., the Fourier transforms of the covariance functions). The approach then estimates the corresponding parameters from the data. The covariance can then be obtained by inverting the Fourier transformations with their estimated parameters. For instance, Fuentes (2001) assumes the local spatial covariance has the Whittle–Matern isotropic form; i.e.,

$$C_{\theta_i}(|h|) = b_i(|h|/a_i)^{\nu_i} K_{\nu_i}(|h|/a_i),$$

where $\theta_i = (b_i, \nu_i, a_i)$ is a vector of parameters with $\nu_i \geq 0$ , $a_i \geq 0$, and $K_{\nu_i}$ is a modified Bessel function with order $\nu_i$ (see Abramowitz and Stegun 1970 for details). The corresponding spectral density has the form

$$f_i = g(a_i, \nu_i, b_i)(a_i^{-2} + |\omega|^2)^{(-\nu_i - 1)},$$

where $\omega$ denotes frequency in the spectral domain and $g$ is a known function of $a_i$, $\nu_i$, and $b_i$. The parameters of each local process are then estimated by fitting the spectral density with its observed *periodogram* (a nonparametric estimate of the spectral density). See Fuentes (2001) for more detail.

It is worth noticing that the observed periodogram in each subregion is obtained by using the measurements in that region, implicitly assuming that the measurements in the $i$th subregion are solely from the local process $Y_i(s)$. This implicit assumption presents a conceptual challenge since for any other locations in the $i$th subregion, the random field is a weighted average of all the local processes, $Y_1, \ldots, Y_k(s)$, and hence does not solely depend on $Y_i(s)$.

## 6.6 Wrapup

That completes our tour of spatial covariance structures. We have seen the important role these structures play in spatial statistics. Moreover, we surveyed the great variety of models that have been suggested in the case of stationary fields where intersite correlations depend only on the difference between the geographical locations of these sites.

However, this last assumption generally proves untenable in modeling environmental processes over broad geographical domains. We saw ways that have been proposed to cope with such processes. In particular, the so-called SG approach enables the theory for the stationary case to be adapted to the nonstationary one.

Our tour has laid the groundwork for spatial modeling, at least for the case where the Gaussian distribution is an adequate representation for the distribution of the environmental field's random response (suitably transformed). It is to that topic we now turn in the following chapter.

# 7

# Spatial Prediction: Classical Approaches

*The only useful function of a statistician is to make predictions, and thus to provide a basis for action.*
*William Edwards Deming*

Predicting an unmeasured (hence uncertain) response is a central problem of statistics. In simple linear regression, the prediction problem involves a response $Y$ and a covariate $x$ thought to be of some assistance in predicting $Y$. That $x$ could be time, for example. Or it could be a physical measurement such as *height* as a predictor of *weight*.

### Calibration

Along with the predictor $\hat{Y}$, a prediction interval $(\hat{Y}_L, \hat{Y}_U)$ is required to indicate the level of confidence that may be placed on $\hat{Y}$. To be valid, that prediction interval must be calibrated in some appropriate sense. In other words, if a prediction interval is said to be a 95% interval, 95% must be interpretable in some quantitatively meaningful sense. For example, under the resampling paradigm it might be taken to mean that the interval would contain the observed value of $Y$, say $y$ "19 times out of 20" if $Y$ were sampled repeatedly from its distribution conditional on the value, say $x = x_f$ at which the prediction were being made. However, this interpretation would make little sense when the experiment is to be performed just once.

On the other hand, the Bayesian paradigm offers a valid interpretation in any case: the interval contains the unmeasured response $Y$ with fair odds 19:20 in the view of the individual making the prediction. Presumably, to be fair under repeated sampling, 95% of the prediction intervals computed by that individual would need to include the observed value of $Y$, $y$, under repeated sampling. In other words, he would want his personal prediction interval to be well calibrated.

### Spatial Prediction

Predicting unmeasured responses at locations of interest, using observations made at sites scattered over the field's domain, is commonly called *spatial*

*interpolation* or *spatial prediction*. The need to make such predictions arises in diverse fields such as mining, geology (geostatistics), environmental health, engineering, soil science, and hydrology. For example, studies of environmental health often require estimates of pollution levels at locations where monitoring data are not available, using observed levels at monitoring stations. These estimated concentration levels over spatial fields are needed to assess the health impacts of environmental pollution. In mining applications, ore concentration levels at unsampled deposits are needed after observations are made at selected locations in the geographical region.

Spatial prediction is complicated by the sparseness of the monitoring sites in the geographical domain of the field. Further complications can arise from nonhomogeneity in that field. For example, although those in mining applications are generally homogeneous, the cost of obtaining an observation can be substantial. Thus, only single measurements are made at a few sites scattered over the relevant region. In environmental applications, data come from networks of widely dispersed monitoring stations. Although the cost of collecting repeated measurements at any one station may not be high, the operational cost for networks with large numbers of stations is prohibitively expensive. Thus, in such applications, repeated measurements for multiple pollutants are often available for only a small number of monitoring locations. Moreover, these environmental fields are usually nonhomogeneous.

## Kriging

In this chapter, we focus our discussion on classical approaches to spatial interpolation, notably kriging. As with any other form of prediction, prediction intervals (ellipsoids in the case of multivariate responses) are needed to gauge the confidence that can be placed on an interpolated value. However, the high cost of measurements means that typically few observations are available for predicting the random fields encountered in geostatistics, one of the earliest domains in which the need for spatial interpolation emerged. Yet, by exploiting the homogeneity of fields encountered in that domain, Krige, a South African mining engineer developed, in 1951, a method that was able to use the relatively small amount of data for spatial prediction. That method, formalized later by Matheron (1962) and now commonly known as kriging, demonstrated the great power of spatial statistics.

Let us consider a particularly simple setting for the spatial interpolation problem. There, values of the concentration levels of the random field are measured at locations $s_1, \ldots, s_n$ to yield $Y(s_i)$ for all $i$. The objective is to estimate the concentration level $Y(s_0)$ for the location $s_0$ seen in Figure 7.1.

The kriging interpolator, a weighted linear combination of the observed values $\{Y(s_i)\}$, has coefficients chosen to make it unbiased and have minimal prediction error, commonly known as the *kriging variance* in geostatistics. The kriging interpolator is hence called the Best Linear Unbiased Predictor (BLUP). Correspondingly, the kriging variance can be expressed as a specific

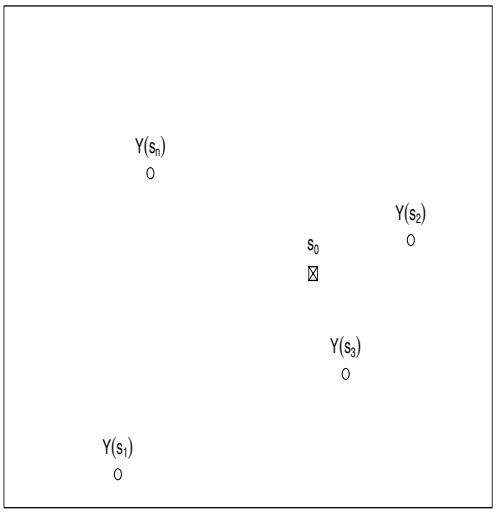


**Fig. 7.1.** The spatial interpolation problem is to estimate the concentration $Y(s_0)$ at location $s_0$ given the observed concentrations $Y(s_i)$.

function of the weights and observations. Thus, application of the method reduces mainly to estimating the optimal weights.

Those weights turn out to be specific functions of the covariances of the random field between the locations, $s_0, s_1, \ldots, s_n$. When the covariance structure of the random field is isotropic, the weights reduce further to functions of the distances between locations. In fact, they are proportionally inverse to a monotonic transformation of the distances from $s_0$ to $s_i$ for $i = 1, \ldots, n$. That is, observations closer to a location of interest would be given correspondingly greater weight in the kriging interpolator, an intuitively appealingly characteristic of the method.

Furthermore, isotropy of a covariance field implies that same monotonic transformation is assumed for any locations of interest in the domain of the random field. Thus, application of the method becomes quite straightforward under the assumption of isotropy; the objective simply reduces to identifying an appropriate monotonic transformation for the intersite distances. Generally, that transformation would be specified through a mathematical expression with a number of unknown parameters estimated from the observed data. The isotropic variogram models described in Chapter 6 are candidates for the role of such a transformation. Because of its simplicity in applications,

when the covariance structure is isotropic, use of kriging has been commonly justified by such an assumption. A particularly important aspect of its implementation has been another assumption, that the estimated weights are in fact known. Emphatically, they are not! However, this second assumption makes simple, the calculation of both the optimal kriging interpolator and its corresponding variance. The method's derivation does not assume specific distributions for the random field, giving it a superficially appealing robustness against misspecification of that distribution. In fact, that robustness is illusory. To be both linear and optimal would essentially require a joint Gaussian distribution for that field. Anyway, that distribution is assumed in practice for the derivation of prediction intervals.

Since its invention, the kriging methodology has been the most widely used approach for spatial interpolation, particularly in geostatistical problems. It has been so popular and successful that in recent years it has also been adopted for use in other fields including environmetrics. Several enhancements of the method have been developed to deal with specific characteristics of particular applications. Below, we describe the basic mathematical setting for kriging and some of its enhancements.

This chapter, which addresses fields with only a spatial and no temporal component, is included for a number of reasons. First, the methods presented here have played an important historical role in the evolution of spatial statistics. Moreover, they have played a key role in modeling environmental processes, as noted in Section 5.3. Indeed in the 1970s, the SIMS Group, led by Paul Switzer at Stanford, pioneered the use of the geostatistical approach in that role for air pollution analysis. Finally, the primary (detrending/prefiltering) strategy could be used to reduce a spatial–temporal field to an independent sequence of spatial fields, each susceptible to treatment by the methods offered in this chapter.

## 7.1 Ordinary Kriging

Suppose the random field $Y(.)$ is intrinsically stationary; that is, for any location $s$

$$E[Y(s)] = \mu$$
$$Var[Y(s) - Y(s+h)] = 2\gamma(|h|),$$

where $\gamma(|h|)$ is the semi-variogram and a function of the distance $|h|$ separating two locations. Let $Y(s_1), \ldots, Y(s_n)$ be the random variables representing the field at locations $s_1, \ldots, s_n$ with realizations $y_1, \ldots, y_n$. The problem is to predict the random field at location $s_0$, $Y(s_0)$, from the observed data as displayed in Figure 7.1.

Kriging, a local interpolation method for such spatial interpolation problems, is a weighted average of levels in the surrounding area. It is an optimal linear estimator of the form

$$Y^*(s_0) = \sum_{i=1}^{n} \alpha_i Y(s_i), \tag{7.1}$$

where the weights $\alpha_i$ are chosen to make the estimator unbiased and of minimal prediction error.

To achieve unbiasedness, the expectation of the kriging estimator must be identical to the expectation of the random field at location $s_0$. In other words,

$$\begin{aligned} E[Y^*(s_0)] &= E\left[\sum_{i=1}^{n} \alpha_i Y(s_i)\right] \\ &= \sum_{i=1}^{n} \alpha_i E[Y(s_i)] \\ &= \sum_{i=1}^{n} \alpha_i \mu. \end{aligned}$$

The last equality derives from the stationarity of the mean. Hence the estimator is unbiased; i.e., $E[Y^*(s_0)] = E[Y(s_0)]$, if the weights sum to unity. That is,

$$\sum_{i=1}^{n} \alpha_i = 1. \tag{7.2}$$

**Kriging Variance**

The prediction error is quantified by the mean-squared prediction error $\sigma^2_{s_0} \equiv E[Y^*(s_0) - Y(s_0)]^2$. In geostatistical terminology, it is called the estimation variance or kriging variance. It can be expanded in terms of the semi-variogram, using the unbiasedness condition (7.2), as

$$\begin{aligned} \sigma^2_{s_0} &\equiv E\left[Y^*(s_0) - Y(s_0)\right]^2 \\ &= E\left[\sum_{i=1}^{n} \alpha_i (Y(s_i) - Y(s_0))\right]^2 \\ &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j (Y(s_i) - Y(s_j))^2/2 - \sum_{i=1}^{n} \alpha_i (Y(s_i) - Y(s_0))^2\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j E[Y(s_i) - Y(s_j)]^2/2 - \sum_{i=1}^{n} \alpha_i E[Y(s_i) - Y(s_0)]^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \gamma(|h_{ij}|) - \sum_{i=1}^{n} \alpha_i \gamma(|h_{i0}|), \end{aligned} \tag{7.3}$$

where $\gamma(|h_{ij}|)$ is the semi-variogram and function of the distance $|h_{ij}|$ between locations $s_i$ and $s_j$. The last equality comes from the intrinsic stationarity of the random field.

### Kriging Weights

The mean-squared prediction error given by (7.3) can be minimized by choosing the weight $\alpha_i$ appropriately, in conjunction with the unbiasedness condition (7.2). To do so, we can employ an optimization technique commonly known as the Lagrange multiplier method. With that method, the optimal weights are found by setting to zero, the partial derivative, with respect to each $\alpha_i$ of the objective function defined as

$$f(\alpha_1, \ldots, \alpha_n, \lambda) = \sigma^2_{s_0} + 2\lambda \left( \sum_{i=1}^{n} \alpha_i - 1 \right).$$

The Lagrange multiplier $\lambda$, properly chosen, ensures achievement of the unbiasedness condition in the minimization process. The partial derivative $\partial f / \partial \lambda$ set to zero yields that unbiasedness condition.

Setting the partial derivatives to zero yields a system of $n+1$ linear equations to be solved for the $n$ optimal weights $\alpha_1, \ldots, \alpha_n$. The set of these linear equations is called the ordinary kriging system (Matheron 1971; Journel and Huijbregts 1978; Cressie 1991). It can be written as

$$\begin{cases} \sum_{j=1}^{n} \alpha_j \gamma(|h_{ij}|) + \lambda = \gamma(|h_{i0}|) \quad i = 1, \ldots, n \\ \qquad\qquad \sum_{j=1}^{n} \alpha_j = 1. \end{cases} \tag{7.4}$$

The optimal weights, $\alpha_1, \ldots, \alpha_n$, satisfying the kriging system (7.4) are used in Equations (7.1) and (7.3) to yield the theoretical kriging estimator and its corresponding kriging variance, respectively. Hence, given the semi-variogram, kriging gives a best linear unbiased predictor. Prediction intervals can be constructed accordingly. For instance, when the random field $Y(\cdot)$ is Gaussian, the interval

$$[Y^*(s_0) - 1.96\sigma_{s_0}, Y^*(s_0) + 1.96\sigma_{s_0}] \tag{7.5}$$

has a 95% nominal prediction level.

### Estimated Optimal Weights

In applications, the semi-variogram is generally unknown. The realizations, $y_1, \ldots, y_n$, are used to identify a suitable semi-variogram model and to obtain the estimated values $\hat{\gamma}(.)$. Solving the kriging system (7.4) using $\hat{\gamma}(|h_{ij}|)$ yields the estimated optimal weights, i.e., the $\hat{\alpha}$s. General solutions for the kriging systems are given in the next section.

The kriging interpolator for a location $s_0$ and its corresponding estimated kriging variance are then given by

$$\hat{Y}^*(s_0) = \sum_{i=1}^{n} \hat{\alpha}_i y_i \tag{7.6}$$

$$\hat{\sigma}^2_{s_0} = \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\alpha}_i \hat{\alpha}_j \hat{\gamma}(|h_{ij}|) - \sum_{i=1}^{n} \hat{\alpha}_i \hat{\gamma}(|h_{i0}|). \tag{7.7}$$

When the random field follows a Gaussian distribution, the estimated 95% prediction interval can be obtained from (7.5) by replacing the unknown roots of the kriging variances by their estimated values, i.e.,

$$[\hat{Y}^*(s_0) - 1.96\hat{\sigma}_{s_0}, \hat{Y}^*(s_0) + 1.96\hat{\sigma}_{s_0}]. \tag{7.8}$$

**Remarks**

- Although the ordinary kriging method described above is for spatial prediction of one location at a time, the method can be simply extended for the simultaneous prediction of multiple locations. The kriging solution in such a setting is given in the next section.
- When the random field is not Gaussian, the prediction interval given in (7.5) may not yield the correct 95% nominal level.
- The kriging predictor is an *exact interpolator* (Journel and Huijbregts 1978). That is, the kriging interpolator coincides with the measurement at any location where one has been made; the corresponding kriging variance is then zero at such a location. This can easily be seen by solving the above kriging system with the weight corresponding to that location set to 1 while the remaining weights are set to zero.
- The above kriging system is derived without assuming the existence of a spatial covariance. When that covariance does exist, the kriging variance $\sigma^2_{s_0}$ can be expressed as a function of the covariance between locations $C(s_i, s_j)$ as

$$\sigma^2_{s_0} = 2\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j C(s_i, s_j) - 2\sum_{i=1}^{n}\alpha_i C(s_i, s_0) + var(Y(s_0)).$$

  The kriging system resulting from minimizing this function has the same form as the system (7.4) where $\gamma(|h_{ij}|)$ and $\gamma(|h_{i0}|)$ are replaced by $C(s_i, s_j)$ and $C(s_i, s_0)$. One can then derive the theoretical kriging estimator and the kriging variance as functions of $C(.,.)$.
  Hence it is not necessary to assume the stationarity of the spatial covariance in the derivation of the kriging estimation method. However, in practice, it is generally difficult to estimate the spatial covariance which is nonstationary, particularly in geological applications where the amount of observed data is limited.

## 7.2 Universal Kriging

Ordinary kriging as described above is developed to deal with interpolation problems where the random field has a constant mean. However, in practice, environmental and geological fields often exhibit nonconstant mean values. That fact has led to the development of the *universal kriging* method (Matheron 1969) which interpolates random fields whose mean function does depend

on location through specific structural forms. More precisely, the universal kriging method assumes that a random field with nonconstant expectation is a linear combination of two components. The first is deterministic, a function that depends on location. The second is probabilistic, i.e., second-order stationarity. Specifically, let $Y(s_1), \ldots, Y(s_n)$ be the random variables representing the field at locations $s_1, \ldots, s_n$. Assume

$$Y(s) = \mu(s) + Z(s), \tag{7.9}$$

where $\mu(s)$, a function of location $s$, is the deterministic component and $Z(s)$ is a second-order stationary process with a constant mean, assumed to be zero (without loss of generality). Suppose the drift $\mu(s)$ can be represented as a linear combination of known functions $\{f_l(s), l = 1, \ldots, k\}$, with unknown coefficients $\{a_l\}$,

$$\mu(s) = \sum_{l=1}^{k} a_l f_l(s).$$

Thus, the mean and the covariance of the random field can be expressed as

$$E[Y(s)] = \sum_{l=1}^{k} a_l f_l(s)$$

$$\begin{aligned} E[(Y(s_1) - \mu(s_1))(Y(s_2) - \mu(s_2))] &\equiv E[Z(s_1)Z(s_2)] \\ &= C(s_1 - s_2). \end{aligned}$$

**Universal Kriging Predictor**

Like ordinary kriging, universal kriging yields a predictor that is again a weighted average of the levels in the surrounding region,

$$Y^*(s_0) = \sum_{i=1}^{n} \alpha_i Y(s_i), \tag{7.10}$$

where the weights $\alpha_i$ are chosen to make the estimator unbiased with minimal prediction error.

The unbiasedness condition is achieved by letting $E[Y^*(s_0)] = E[Y(s_0)]$, or

$$\mu(s_0) - \sum_{i=1}^{n} \alpha_i \mu(s_i) = 0.$$

Equivalently, that condition can be written

$$\sum_{l=1}^{k} a_l (f_l(s_0) - \sum_{i=1}^{n} \alpha_i f_l(s_i)) = 0.$$

Since the $a_l$s are generally nonzero, the universal condition becomes

$$f_l(s_0) = \sum_{i=1}^{n} \alpha_i f_l(s_i) \; l = 1, \ldots, k. \tag{7.11}$$

The term *universal* used by Matheron (1969) refers to the unbiasedness of the kriging estimator when the random field has nonconstant mean.

### Universal Kriging Variance

Using the unbiasedness condition (7.11), the kriging variance can be expressed as

$$\begin{aligned}
\sigma^2_{s_0} &\equiv E[Y^*(s_0) - Y(s_0)]^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j C(s_i, s_j) - 2\sum_{i=1}^{n} \alpha_i C(s_i, s_0) \\
&\quad + var(Y(s_0)) \qquad (7.12) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j C(s_i - s_j) - 2\sum_{i=1}^{n} \alpha_i C(s_i - s_0) \\
&\quad + var(Y(s_0)), \qquad (7.13)
\end{aligned}$$

where $C(s_i - s_j)$ denotes the covariance between locations $s_i$ and $s_j$ for the second-order stationary process $Z(\cdot)$. The kriging variance (7.13) can be minimized by choosing the weight $\alpha$ appropriately by the Lagrange multiplier method. The resulting $n + k$ linear equations, called the *universal kriging system*, are

$$\begin{cases} \sum_{j=1}^{n} \alpha_j C(s_i - s_j) + \lambda_l f_l(s_i) = C(s_i - s_0) & i = 1, \ldots, n \\ \sum_{j=1}^{n} \alpha_j f_l(s_j) = f_l(s_0) & l = 1, \ldots, k. \end{cases} \tag{7.14}$$

The optimal weights $\alpha_1, \ldots, \alpha_n$ satisfying the universal kriging system (7.14) are used in equations (7.10) and (7.13) to yield the theoretical universal kriging estimator and its corresponding kriging variance, respectively. Prediction intervals can be constructed accordingly. For instance, when the random field $Y(.)$ is Gaussian, the interval

$$[Y^*(s_0) - 1.96\sigma_{s_0}, Y^*(s_0) + 1.96\sigma_{s_0}] \tag{7.15}$$

has the 95% nominal prediction level.

### Universal Kriging Solution

We now present specific solutions for the optimal weights in kriging. Consider a setting as in (7.9) for universal kriging and more generally, $m$ locations of interest for spatial prediction. In vector notation, the relationship (7.9) can be simply expressed as

$$\begin{aligned}
Y &= X\beta + Z \\
Y_0 &= X_0\beta + Z_0,
\end{aligned}$$

where $Y$ denotes the levels of the random field at the $n$ locations with measurements and $Y_0$, the level at the $m$ locations of interest to be predicted. Here

$$Y = (Y(s_1), \ldots, Y(s_n))^T$$
$$Y_0 = (Y(s_{0_1}), \ldots, Y(s_{0_m}))^T$$

$$Z = (Z(s_1), \ldots, Z(s_n))^T$$
$$Z_0 = (Z(s_{o_1}), \ldots, Z(s_{o_m}))^T$$

$$\beta = (a_1, \ldots, a_k)^T$$

and

$$X = \begin{pmatrix} f_1(s_1) & \ldots & f_k(s_1) \\ \vdots & & \vdots \\ f_1(s_n) & \ldots & f_k(s_n) \end{pmatrix}$$

$$X_0 = \begin{pmatrix} f_1(s_{0_1}) & \ldots & f_k(s_{0_1}) \\ \vdots & & \vdots \\ f_1(s_{0_m}) & \ldots & f_k(s_{0_m}) \end{pmatrix}.$$

The random components $Z$ and $Z_0$ have covariance matrices $\Sigma_{yy}$ and $\Sigma_{00}$, respectively, as well as a cross-covariance matrix $\Sigma_{y0}$. These matrices are known with elements being the covariance between the corresponding locations. For example, $\Sigma_{y0}$ is a matrix with $n$ rows and $m$ columns given by

$$\Sigma_{y0} = \left\{C(s_i - s_{0_j})\right\}_{n \times m}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

In this setting, the kriging interpolator is defined as

$$Y_0^* = \Theta Y$$

where $\Theta$ is the $m \times n$ matrix of weights. The problem is then to determine the optimal weights, $\Theta$. Let $\Lambda$ be the $n \times m$ matrix of the Lagrange multipliers.

**Universal Kriging System**

The universal kriging system, which is an extension of (7.14) to multiple locations, can be expressed in terms of $\Theta$ and $\Lambda$ as

$$\begin{cases} \Sigma_{yy}\Theta^T + X\Lambda = \Sigma_{y0} \\ \Theta X = X_0. \end{cases} \tag{7.16}$$

Solving the system (7.16) yields

$$\Lambda = \left(X^T \Sigma_{yy}^{-1} X\right)^{-1} \left(X^T \Sigma_{yy}^{-1} \Sigma_{y0} - X_0\right)$$

and

$$\Theta = (\Sigma_{0y} - M^T X^T)\Sigma_{yy}^{-1}.$$

Hence the kriging interpolator and the corresponding kriging variance are

$$\begin{aligned} Y^*(s_0) &= (\Sigma_{0y} - M^T X^T)\Sigma_{yy}^{-1} Y \\ var(Y^*(s_0)) &= E(\Theta Y - Y_0)^2 \\ &= \sigma_{00} - \Sigma_{0y}\Sigma_{yy}^{-1}\Sigma_{y0} + M^T \left(X^T \Sigma_{yy}^{-1} X\right) M. \end{aligned}$$

**Remarks**

- For the universal kriging system to have a solution, it is necessary that the deterministic functions $f_l$s be linearly independent; that is, the matrix having $(f_l(s_1), \ldots, f_l(s_n))^T$, $l = 1, \ldots, k$ as columns, must have full rank.
- The deterministic functions could always be chosen to ensure that the probabilistic component has mean zero. For example, this can be achieved by including a function with constant value.
- As with ordinary kriging, it is not necessary to assume stationarity of the spatial covariance in the derivation of the universal kriging predictor. However, the stationarity assumption is often made in practice so that the spatial covariance structure can be estimated using available data.
- Ordinary kriging is a special case that corresponds to the situation where

  $$f_1 = 1 \quad \text{and} \quad f_2 = \cdots = f_k = 0.$$

- Several authors have applied kriging to a transformation of the random field (see, for example, Howarth and Earle 1979 and Verly 1983). The main idea is to identify a transformation for $Y(\cdot)$ so that the resulting process is Gaussian and to apply the kriging method to the transformed random field. This approach is called *trans-Gaussian kriging* (see Cressie 1991 for more information).

More detailed discussion of kriging methods can be found, for example, in Journel and Huijbregts (1978), Cressie (1991), Kitanidis (1997), and Wackernagel (2003).

## 7.3 Cokriging

In some applications, multiple measurements obtain at each location for different random fields. For example, environmental monitoring stations often simultaneously measure several air pollutants such as $O_3$ and $PM_{10}$. In mining applications, the silver concentration at a location could be observed together with the lead and zinc concentration levels and other minerals. These multiple processes may be correlated and so using all observed data may improve the

prediction for any specific random field. The cokriging method, a generalization of the kriging method, has been developed to deal with such multivariate interpolation problems (Journel and Huijbregts 1978; Cressie 1991).

Let $Y_k(s)$, $k = 1, \ldots, K$, denote the correlated random fields, representing the concentration levels for $K$ different minerals, assumed to be second-order stationary. That is, for $k, k_1, k_2 \in \{1, \ldots, K\}$,

$$E[Y_k(s)] = \mu_k \qquad \forall s$$
$$E[(Y_{k_1}(s_1) - \mu_{k_1})(Y_{k_2}(s_2) - \mu_{k_2})] = C_{k_1 k_2}(s_1 - s_2),$$

where $C_{k_1 k_2}(h)$ denotes the cross-covariance between the $k_1$th and $k_2$th random fields. When $k_1 = k_2$, the cross-covariance reduces to the usual univariate covariance. For each random field $k$, let $Y_k(s_1), \ldots, Y_k(s_{n_k})$ be the random variables representing the $k$th random field at locations $s_1$ to $s_{n_k}$.

### Cokriging Estimator

The cokriging estimator for a specific random field, say $k_0 \in \{1, \ldots, K\}$, at location $s_0$ is a linear combination of the form

$$Y^*_{k_0}(s_0) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \alpha_{ki} Y_k(s_i). \tag{7.17}$$

The weights, $\alpha_{ki}$, are chosen to ensure the estimator is unbiased and has minimal prediction error.

The unbiasedness condition is achieved by letting $E[Y^*_{k_0}(s_0)] = E[Y_{k_0}(s_0)]$, or

$$\mu_{k_0} \left(1 - \sum_{i=1}^{n_{k_0}} \alpha_{ik_0}\right) - \sum_{k \neq k_0}^{K} \mu_k \sum_{i=1}^{n_k} \alpha_{ki} = 0.$$

Hence the unbiasedness condition can be written in terms of $K$ constraints

$$\begin{cases} \sum_{i=1}^{n_{k_0}} \alpha_{ik_0} = 1 \text{ and} \\ \sum_{i=1}^{n_k} \alpha_{ki} = 0 \ \forall k \neq k_0. \end{cases}$$

The first constraint implies $n_{k_0}$ must be different from zero; that is, the cokriging estimator for the $k_0$th random field must depend on at least one random variable of that field. Hence, in applications, at least one observation from the random field of interest must be available for this method to work.

### Cokriging Variance

The prediction error is quantified by the *cokriging variance*,

$$\sigma^2_{k_0, s_0} = E[\sum_{k=1}^{K} \sum_{i=1}^{n_k} \alpha_{ki} Y_k(s_i) - Y_{k_0}(s_0)]^2$$

expressible as

$$\sigma^2_{k_0,s_0} = C_{k_0k_0}(0) - \sum_{k=1}^{K}\sum_{i=1}^{n_k} \alpha_{ki} C_{k_0k}(s_i - s_0). \tag{7.18}$$

The cokriging variance can be minimized subject to the $K$ unbiasedness constraints above using the Lagrange multiplier method. The minimization leads to a system of $(\sum_{k=1}^{K} n_k + K)$ linear equations to be solved for $\alpha$s, known as the *cokriging system*:

$$\begin{cases} \sum_{k_1=1}^{K}\sum_{i_1=1}^{n_k} \alpha_{k_1i_1} C_{k_1k_2}(s_{i_1} - s_{i_2}) - \lambda_{k_2} = C_{k_0k_2}(s_{i_2} - s_0) \\ \qquad \text{for} \quad k_2 = 1, \ldots, K, \quad i_2 = 1, \ldots, n_{k_2} \\ \qquad\qquad \sum_{i=1}^{n_{k_0}} \alpha_{ik_0} = 1 \\ \qquad\qquad \sum_{i=1}^{n_k} \alpha_{ki} = 0 \quad \forall k \neq k_0. \end{cases} \tag{7.19}$$

The $\alpha$s satisfying the cokriging system (7.19) are used in Equations (7.17) and (7.18) to obtain the theoretical cokriging predictor and its corresponding variance. The prediction interval can be then constructed accordingly. For instance, when the multivariate random field $Y(\cdot)$ is Gaussian, the interval

$$[Y^*_{k_0}(s_0) - 1.96\sigma_{k_0,s_0}, Y^*_{k_0}(s_0) + 1.96\sigma_{k_0,s_0}]$$

has a 95% nominal prediction level.

**Remarks**

- The cokriging method described here is quite similar to the *ordinary kriging* method, except that different univariate random fields must be considered simultaneously. This is sometimes called ordinary cokriging (Wackernagel 2001) to distinguish it from *universal cokriging* where the universal kriging method is similarly extended to deal with multivariate responses (Cressie 1991).
- Unlike the kriging method, the general formulation of the cokriging system in terms of the cross semi-variogram, defined as

  $$E[(Y_{k_1}(s_1) - Y_{k_1}(s_2))(Y_{k_2}(s_1) - Y_{k_2}(s_2))] = 2\gamma_{k_1k_2}(s_1 - s_2),$$

  is available only if the cross-covariance is symmetric (Journel and Huijbregts 1978); that is, $C_{k_1k_2}(h) = C_{k_2k_1}(h)$ for any vector $h$. In this case, the cross semi-variogram system is obtained by simply replacing the $C_{k_1k_2}(s_{i_1} - s_{i_2})$ by $-\gamma_{k_1k_2}(s_{i_1} - s_{i_2})$ in (7.19).

## 7.4 Disjunctive Kriging

The kriging and cokriging methods described above restrict the predictors to be linear; that is, if we let $Y(s_1), \ldots, Y(s_n)$ be the random variables representing the field at locations $s_1, \ldots, s_n$, the co/kriging predictors for $Y(s_0)$

are linear combinations of the $\{Y(s_i)\}$. The estimators are then optimized by choosing weights to minimize mean-squared prediction errors. However, such linear predictors would generally be suboptimal and nonlinear ones more appropriate.

More precisely, it can be shown that the conditional expectation $E[Y(s_0) \mid Y(s_1), \ldots, Y(s_n)]$ minimizes the mean-squared prediction error $E[Y(s_0) - Y^*(s_0)]^2$ among all possible predictors and hence is optimal. That conditional expectation is, by definition, an orthogonal projection of $Y(s_0)$ onto the vector space spanned by all functions $f(Y(s_1), \ldots, Y(s_n))$ (see Journel and Huijbregts 1978 for more details). Technically speaking, these functions must be measurable but for simplicity, we take that requirement as understood, without repetitively stating it, here and below.

That conditional expectation would generally be nonlinear with only a few important exceptions. For example, when the random field is Gaussian, the conditional expectation $E[Y(s_0) \mid Y(s_1), \ldots, Y(s_n)]$ has a linear structure and coincides with the kriging predictor, making the latter optimal in this special case.

The *disjunctive kriging method* (Matheron 1976) produces a more general than linear predictor that is closer to the conditional expectation $E[Y(s_0) \mid Y(s_1), \ldots, Y(s_n)]$. The disjunctive kriging predictor is defined as

$$Y^*_{DK}(s_0) = \sum_{i=1}^{n} f_i(Y(s_i)), \tag{7.20}$$

where the $f_i$s are specified functions. The $\{f_i\}$ are chosen to make $Y^*_{DK}(s_0)$ an orthogonal projection of $Y(s_0)$ onto the vector space spanned by $\{g_i(Y(s_i))\}$ where the $g_i$ are any (measurable) functions; that is,

$$E\left\{\left[Y(s_0) - \sum_{i=1}^{n} f_i(Y(s_i))\right] h_j(Y(s_j))\right\} = 0,\ j = 1, \ldots, n \tag{7.21}$$

for any functions, $\{h_i(Y(s_i))\}$.

Journel and Huijbregts (1978) show that the orthogonality condition (7.21) leads to the following disjunctive kriging system to be solved for $\{f_i\}$,

$$E\left[Y(s_0) \mid Y(s_j)\right] = \sum_{i=1}^{n} E\left[f_i(Y(s_i)) \mid Y(s_j)\right],\ j = 1, \ldots, n. \tag{7.22}$$

The system (7.22) reveals that disjunctive kriging requires knowledge of the bivariate distributions of $\{Y(s_i), Y(s_0)\}$ and $\{Y(s_i), Y(s_j)\}$ for $i, j \in \{1, \ldots, n\}$. In contrast, the best estimator, the conditional expectation $E[Y(s_0) \mid Y(s_1), \ldots, Y(s_n)]$, would require the knowledge of the $(n+1)$-dimensional distribution of $\{Y(s_0), Y(s_1), \ldots, Y(s_n)\}$.

Although disjunctive kriging reduces complexity, finding solutions of the system (7.22) would not generally be straightforward. Solutions to (7.22) are obtainable when the random field $Y(s)$ follows a class of models, called *isofactorial* (Matheron 1976, 1984) with bivariate Gaussian distributions as special

cases (Cressie 1991). For bivariate Gaussian distributions having marginal standard normal distributions, the disjunctive kriging estimator can be expressed in terms of the $K$th-order Hermite polynomial expansion (see, for examples, Beckmann 1973 and Journel and Huijbregts 1978 for more details)

$$Y^*_{DK}(s_0) = \sum_{i=1}^{n} \sum_{k=0}^{K} \eta_k(Y(s_i)) f_{ik},$$

where $\eta_k$ is the $k$th Hermite polynomial defined as

$$\eta_k(x) = (k!)^{-1/2} exp\left(\frac{x^2}{2}\right) \frac{\partial^k}{\partial y^k} \left\{ exp\left(-\frac{x^2}{2}\right) \right\}.$$

The $(K+1)n$ unknown parameters $\{f_{ik}\}$ satisfy the corresponding disjunctive kriging system

$$\sum_{i=1}^{n} \rho^k_{ij} f_{ik} = b_k \rho^k_{0j},\ k = 1, \ldots, K,\ j = 1, \ldots, n, \tag{7.23}$$

where for all $i$ and $j$, $\rho_{ij}$ is the correlation between $Y(s_i)$ and $Y(s_j)$, while for all k, $b_k$ is the $k$th coefficient of the Hermite expansion of $Y(s_0)$ given by

$$b_k = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} \frac{\partial^k}{\partial y^k} \left\{ exp\left(-\frac{x^2}{2}\right) \right\}.$$

The mean-squared prediction error, i.e., *disjunctive kriging variance* can be written as (Journel and Huijbregts 1978)

$$E[Y(s_0) - Y^*_{DK}(s_0)]^2 = \sum_{k=0}^{K} b_k \left( b_k - \sum_{i=1}^{n} f_{ik} \rho^k_{0i} \right).$$

More details concerning disjunctive kriging can be found in Journel and Huijbregts (1978) and Cressie (1991).

**Remarks**

- The disjunctive kriging method for optimally predicting $g(Y(s_0))$, $g$ being any function of $Y(s_0)$, can be analogously derived. In fact, the resulting expressions would be very similar. The disjunctive kriging system would have (7.22) as before, with $E[Y(s_0) \mid Y(s_j)]$ replaced by $E[g(Y(s_0)) \mid Y(s_j)]$. For the bivariate Gaussian case, the system for the parameters $\{f_{ik}\}$ is the same as (7.23), where $b_k$ is the $k$th coefficient of the Hermite expansion of $g(Y(s_0))$. Flexibility in choosing $g$ would allow the method to be used in a wide range of applications where nonlinear estimates are required. For instance, selecting $g(x) = xI_{x \geq x_o}$, where $I_{x \geq x_o}$ is an indicator of whether $x \geq x_o$, would deal with the optimal prediction of mineral level above $x_0$ in geological applications.

- Other nonlinear kriging methods have been studied including indicator and probability kriging by several authors (see, for example, Journel 1983, 1988, and Bilonick 1988). Basically when the interest is in some transformation of the random field, these methods first transform the random field and then apply linear kriging methods on the transformed variables. In principle, the approximation to the conditional expectation by these methods would be less accurate than disjunctive kriging because the latter allows for a richer class of transformations. On the other hand, disjunctive kriging solutions are generally difficult to obtain, requiring specific assumptions on the bivariate distributions. Hence, in practice it is not clear which if any is superior to the others.

## 7.5 Wrapup

As discussed above, kriging and its variants produce unbiased predictors that are also optimal in the sense of minimizing the prediction error for spatial interpolation problems. Optimal linear predictors generally rely on features of the random field such as the semi-variogram, that are related to moments up to second-order. They are completely determined when these moments are known and specified. However, in applications semi-variograms are unknown and must be estimated from the data. The interpolators are then obtained by simply plugging in the estimated semi-variograms, as if these estimates were known and specified without error. This common practice underestimates the imprecision of the spatial interpolators since uncertainty associated with the estimation of the semi-variograms are ignored (Hughes and Lettenmeier 1981; Kitanidis 1986; Le and Zidek 1992; Handcock and Stein 1993). This deficiency can be serious in commonly encountered situations where data are in limited supply, for example, in hydrological applications (Kitanidis 1986), and uncertainty will be large. Prediction intervals derived from interpolation methodology that fails to incorporate all relevant uncertainty represents unwarranted confidence in the interpolated values (Sun 1998). This deficiency can potentially lead to seemingly valid decisions or regulatory actions that are in fact unjustified (Le and Zidek 1992).

A related shortcoming of kriging stems from its reliance on parametric (isotropic) models for the semi-variogram. Although such isotropic models may work for some geological applications, they would generally be unrealistic, particularly for environmental problems. It is worth noticing that when models are inappropriate, no matter how many additional data are obtained, nothing in the methodology enables it to update the initial, overly simplistic parametric semi-variogram models (Yakowitz and Szidarovszky 1985). To partially overcome this deficiency, Haas (1990, 1992) proposes a modification called moving-window regression residual kriging that applies the (isotropic) kriging method locally to regions defined by a moving window across the geographical field. That is, for each location of interest, Haas's approach first

identifies a surrounding local region based on a fixed number of available data points in the area and then obtains the kriging interpolator using the usual isotropic models for the semi-variogram. The method is later extended to include temporal components (Haas, 1995) as well as multivariate responses (Haas 1996). Since isotropy is only required in local regions, the modified method can generally cope with nonstationary random fields. However, use of the method entails a trade-off between the number of monitoring sites (observations) needed for reasonable estimates of the semi-variograms and the size of the local regions. The larger the number of observations, the more precise are the estimates but less tenable the isotropy assumption. Furthermore, as with kriging in general, uncertainty associated with estimating the semi-variogram cannot be accommodated in calculating such things as prediction intervals, a fact whose significance increases as the amount of data dwindles.

In recent years, Bayesian approaches have been developed to overcome these deficiencies. The uncertainty associated with covariance modeling is directly incorporated through prior distributions, in the derivation of the predictive distribution of concentration levels at several locations of interest. Some authors have developed Bayesian versions of kriging where uncertainty associated with parameter estimation is accounted for, but generally the random field is still assumed to be isotropic stationary (Kitanidis 1986; Handcock and Stein 1993; Hjort and Omre 1994; De Oliveira et al. 1997; and Gaudard et al. 1999).

We, on the other hand, have adopted an alternative approach to kriging where uncertainty about the covariance field, not assumed to be stationary, is taken into account in the derivation of the predictive distribution (Le and Zidek 1992). Our integrated framework has since been extended to embrace multivariate responses with specific kinds of missing data patterns (Brown et al. 1994a; Le et al. 1997, 2001; and Kibria et al. 2002). The covariance field is assumed to have one of a rich class of multivariate conjugate priors (Brown et al. 1994b) with the associated hypercovariance matrix estimated by the Sampson and Guttorp method (see Chapter 6). We describe Bayesian approaches in the next chapter.

# 8

# Bayesian Kriging

*In these matters the only certainty is that nothing is certain.*

*Pliny the Elder*

Bayesian approaches to spatial interpolation, which tries to capture the uncertainty referred to by Pliny the Elder, have been developed in recent years to overcome limitations associated with their classical counterparts, especially kriging and its variants. Specifically, classical methods lack the ability to incorporate model uncertainty, notably that associated with model parameters. Moreover, they rely on the assumption of an isotropic covariance field that is often unrealistic in environmental applications.

In general, the flexible Bayesian framework allows model parameters to be treated as uncertain, that is, "random" in Bayesian parlance. Thus, in deriving spatial predictive distributions the parameters are endowed with so-called *prior distributions* representing the investigator's personal knowledge (or beliefs) about the parameters prior to observing the data.

*Conjugate prior* distributions, ones having the same functional form as the associated likelihood and indexed by hyperparameters, prove a particularly convenient choice. By varying their hyperparameters, a wide range of personal beliefs can be represented. Moreover, their mathematical tractability simplifies the derivation of the predictive distribution. Finally, although the advent of Markov chain Monte Carlo (MCMC) methods have made feasible the use on nonconjugate priors, using them comes with a heavy computational price. The latter can prove prohibitive in some instances, for example, in the design of monitoring networks or in the analysis of fields with large spatial domains. Thus, conjugate priors retain an important role in modeling spatial fields.

The incorporation of prior distributions allows uncertainty associated with model parameters to be accounted for in a natural way. Whereas classical methods generate a point prediction (unless supplementary distributional assumptions are tacked on), the Bayesian approach yields the full spatial distribution as an intrinsic outcome. Thus, one gets not only the expectation of the unmeasured response as a point prediction, but such things as its various quantiles as well. This sort of bonus feature is particularly useful in environmental health impact studies where adverse health effects can be caused by exposure levels above certain thresholds or associated with different exposure indices

such as daily one-hour maximum or eight-hour moving average. Indeed, the Bayesian framework can handle arbitrarily complex exposure metrics since their distribution can be simulated from the joint predictive distribution of unmeasured responses.

Thus, the general solution to the spatial interpolation problem under the Bayesian paradigm is the joint predictive distribution of the levels of the random field at the locations of interest, given the data measured at the monitoring sites, $p(Y(s) \mid D)$. Here $Y(s)$, a vector, represents the concentration levels of the random field at a vector of locations $s$ in the domain of the field. At the same time, $D$ represents the data obtained from the monitoring sites. Assuming the concentration levels follow a joint distribution indexed by a set of parameters $\theta$, the predictive distribution is given by

$$\begin{aligned} p[Y(s) \mid D] &= \int p[Y(s), \theta \mid D] d\theta \\ &= \int p[Y(s) \mid \theta, \mid D] p[\theta \mid D] d\theta. \end{aligned}$$

The last equation reveals the (posterior) predictive distribution to be a weighted average of the predictions conditional on various $\theta$s. The predictive distribution can be decomposed into several components. First, future responses can be predicted. Or spatial predictions at locations other than where the gauged stations are situated can be obtained. Finally, unmeasured historical values based on those that were measured can be predicted. This latter application is called *hindcasting* particularly in oceanology for wind and wave predictions. In all cases, prediction intervals (or ellipsoids) can be derived. For example, a 95% simultaneous predictive interval $(a, b)$, can be obtained from the equation $p(a < Y^{[u]} < b \mid D) = .95$. Although these predictive intervals can be derived in closed form in some cases, generally numerical approaches are needed.

Kitanidis (1986) published one of the first articles to use the Bayesian paradigm in spatial interpolation. In a hydrological context, the author develops a theoretical framework for deriving the spatial predictive distribution with a covariance field assumed to be known up to a scale parameter. Specifically, assuming a Gaussian field in conjunction with a conjugate prior distribution for the scale parameter, the author obtains a multivariate Student $t$ distribution in closed form as the resulting predictive distribution. Kitanidis (1986) derives the kriging estimator and its corresponding variance as special cases of the (Bayesian) posterior mean and covariance matrix, thereby creating implicitly the concept of *Bayesian kriging*. That is, if no prior information on the parameter is available, then these Bayesian predictors become identical to kriging predictors. More generally, Bayesian predictors have the advantage that they can admit partial knowledge about the parameters.

Although a conceptually important development, the Kitanidis approach lacks applicability due to the strong assumption of a covariance known up to a scale parameter. Later, Handcock and Stein (1993) take a similar approach

in an application to topographical data. They advance the Kitanidis theory by assuming instead, that the covariance field can be represented by a specific parametric functional form. More specifically, these authors use the isotropic Whittle–Matern class of functions with unknown parameters to model the covariance structure. In this framework, numerical integration is required to obtain the posterior mean and covariance matrix. This development allows the Bayesian kriging approach to have broader applicability, albeit with the restriction to stationary Gaussian fields with isotropic spatial correlation.

The Handcock and Stein approach is subsequently extended by De Oliveira et al. (1997) where the random fields need to be nonlinearly transformed to have Gaussian finite distributions and uncertainty associated with such transformations taken into account. Gaudard et al. (1999) develop computational tools for obtaining spatial predictive distributions with arbitrary prior distributions on the model parameters.

## 8.1 The Kitanidis Framework***

Bayesian kriging theory can be formulated starting from the framework studied by Kitanidis (1986). Consider a setting similar to that of universal kriging described in Chapter 7, with $Y(s_1), \ldots, Y(s_g)$ being the random variables representing the field at $g$ locations $s_1, \ldots, s_g$, having measurements $y(s_1), \ldots, y(s_g)$. Let $Y(s_{o_1}), \ldots, Y(s_{o_u})$ denote the random fields at $u$ locations to be predicted.

### 8.1.1 Model Specification

Assume that for any location $s$ in the field,

$$Y(s) = \mu(s) + Z(s),$$

where $\mu(s)$ is a function of location, $s$ is the deterministic component, and $Z(s)$ follows a Gaussian distribution with mean zero. Suppose the *drift* $\mu(s)$ can be represented as a linear combination of known functions $\{f_l(s), l = 1, \ldots, k\}$ with unknown coefficient $a_l$,

$$\mu(s) = \sum_{l=1}^{k} a_l f_l(s).$$

In vector notation, the relationship can be expressed as

$$\begin{aligned} Y^{[g]} &= X^{[g]}\beta + Z^{[g]} \\ Y^{[u]} &= X^{[u]}\beta + Z^{[u]}, \end{aligned} \tag{8.1}$$

where the superscripts $g$ and $u$ denote the monitored (gauged) sites and unmonitored (ungauged) sites, respectively, with

$$Y^{[g]} = (Y(s_1), \ldots, Y(s_g))^T$$
$$Y^{[u]} = (Y(s_{o_1}), \ldots, Y(s_{o_u}))^T$$

$$Z^{[g]} = (Z(s_1), \ldots, Z(s_g))^T$$
$$Z^{[u]} = (Z(s_{o_1}), \ldots, Z(s_{o_u}))^T$$

$$\beta = (a_1, \ldots, a_k)^T$$

and

$$X^{[g]} = \begin{pmatrix} f_1(s_1) \ldots f_k(s_1) \\ \vdots \qquad\quad \vdots \\ f_1(s_g) \ldots f_k(s_g) \end{pmatrix}$$

$$X^{[u]} = \begin{pmatrix} f_1(s_{o_1}) \ldots f_k(s_{o_1}) \\ \vdots \qquad\quad \vdots \\ f_1(s_{o_u}) \ldots f_k(s_{o_u}). \end{pmatrix}.$$

The random components $Z^{[g]}$ and $Z^{[u]}$ have covariance matrices $Q_{gg}$ and $Q_{uu}$, respectively, as well as a cross-covariance matrix $Q_{ug}$. These matrices are assumed to be known up to a scale parameter. That is,

$$Q_{gg} = \frac{1}{\theta} S_{gg} \; Q_{uu} = \frac{1}{\theta} S_{uu} \; Q_{ug} = \frac{1}{\theta} S_{ug}, \tag{8.2}$$

where the $S$ matrices are known and $\theta$ is an unknown parameter vector.

### 8.1.2 Prior Distribution

The parameters $\beta$ and $\theta$ are assumed to have conjugate prior distributions, namely, the normal and gamma distributions. Specifically,

$$\begin{aligned} \beta \mid \theta &\sim N_k\left(\beta_0, (\theta F)^{-1}\right) \\ \theta &\sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu q}{2}\right), \end{aligned} \tag{8.3}$$

where $N_k(\mu, \Sigma)$ denotes the $k$-variate Gaussian distribution with density

$$f(x, \mu, \Sigma) \propto |\Sigma|^{-k/2} exp\left\{\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

and Gamma$(r, \lambda)$ denotes the gamma distribution with density

$$f(x, r, \lambda) \propto x^{r-1} \exp\{\lambda x\}.$$

The parameters indexing the prior distributions, $\{F, \nu, q\}$, are the hyperparameters. Here $\nu, q$ represent the shape of the prior distribution for $\theta$ and $F$, a $k \times k$ matrix, representing the correlation among elements of $\beta$.

### 8.1.3 Predictive Distribution

The spatial predictive distribution for $Y^{[u]}$ based on (8.1)–(8.3), given the observed measurements $y^{[g]} \equiv (y(s_1), \ldots, y(s_g))^T$, the hyperparameters, and the known $S$ matrices, is a $u$-variate Student $t$ distribution with $\nu + g$ degrees of freedom having posterior mean

$$E\left(Y^{[u]} \mid y^{[g]}\right) = \left(X^{[u]} - \tau X^{[g]}\right) H^{-1} F\beta_0 + \left(\tau + \left(X^{[u]} - \tau X^{[g]}\right) H^{-1} \left(X^{[g]}\right)^T S_{gg}^{-1}\right) y^{[g]} \tag{8.4}$$

and posterior covariance matrix

$$V\left(Y^{[u]} \mid y^{[g]}\right) = l\left(S_{u|g} + \left(X^{[u]} - \tau X^{[g]}\right) H^{-1} \left(X^{[u]} - \tau X^{[g]}\right)^T\right) \times \frac{\nu + g}{\nu + g - 2}, \tag{8.5}$$

where

$$\begin{aligned}
S_{u|g} &= S_{uu} - S_{ug} S_{gg}^{-1} S_{gu} \\
\tau &= S_{ug} S_{gg}^{-1} \\
H &= F + E \\
E &= \left(X^{[g]}\right)^T S_{gg}^{-1} X^{[g]} \\
l &= \frac{\nu q + \hat{\nu}\hat{q} + \left(\hat{b} - \tilde{b}\right) E \left(\hat{b} - \tilde{b}\right)^T}{g + \nu} \\
\hat{b} &= E^{-1} \left(X^{[g]}\right)^T S_{gg}^{-1} y^{[g]} \\
\tilde{b} &= (F + E)^{-1} \left(F\beta_0 + E\hat{b}\right) \\
\hat{\nu} &= g - k \\
\hat{q} &= \frac{\left(y^{[g]}\right)^T S_{gg}^{-1} y^{[g]} - \hat{b}^T X^{[g]} S_{gg}^{-1} y^{[g]}}{g - k}.
\end{aligned}$$

A $u$-variate Student $t$ distribution with mean $\mu$, covariance matrix $gV/(g-2)$, and $g$ degrees of freedom, denoted as $t_u(\mu, V, g)$ has a density function given by

$$f(x) \propto |V|^{-1/2} \left[g + (x - \mu)^T V^{-1} (x - u)\right]^{-(g+u)/2}.$$

### 8.1.4 Remarks

1. Through the Bayesian framework, the posterior mean and covariance matrix given in (8.4) and (8.5) take into account prior knowledge associated

with model parameters expressed via hyperparameters $F$ and $\nu$. The non-informative prior knowledge corresponds to the limiting case, $F = 0$ and $\nu = 0$. In this case, the posterior mean (8.4) reduces to the kriging interpolator as presented in Chapter 7. The posterior covariance matrix (8.5) is then given by the kriging covariance increased by a scale factor $g/(g-2)$ with $\theta$ replaced by $1/l$ (see Kitanidis 1986 for details). Thus, kriging can be considered as a special case of this Bayesian estimation framework.

2. Omre (1987) as well as Omre and Halvorsen (1989) take a slightly different approach where the uncertainty associated with model parameters of the covariance field is directly accounted for in the derivation of the kriging variance. The approach is specifically termed *Bayesian kriging* by Omre (1987). Hjort and Omre (1994) subsequently show that similar results can be obtained by starting with Gaussian random fields and then deriving the corresponding posterior moments.

## 8.2 The Handcock and Stein Method***

Handcock and Stein (1993) use a similar framework as that developed by Kitanidis (1986) to examine the effect of uncertainty in the covariance function on the prediction. Unlike the Kitanidis framework where the covariance field is assumed known (up to a scale parameter), the key feature of the Handcock and Stein approach is that the covariance field can be parametrically represented by specific functional forms with unknown parameters. These hyperparameters can be estimated based on the available data or assumed to follow specific prior distributions. Their derivation is for one unobserved location although it should be easy to extend to multiple locations. The authors illustrate the approach using a topographical data set.

Specifically, as in the Kitanidis framework and the universal kriging, it is assumed the random field follows the model specification (8.1). That is,

$$Y^{[g]} = X^{[g]}\beta + Z^{[g]}$$
$$Y^{[u]} = X^{[u]}\beta + Z^{[u]},$$

where the superscripts $g$ and $u$ denote the monitored (gauged) sites and unmonitored (ungauged) site $s_o$, respectively, with

$$Y^{[g]} = (Y(s_1), \ldots, Y(s_g))^T$$
$$Y^{[u]} = Y(s_o)$$

$$Z^{[g]} = (Z(s_1), \ldots, Z(s_g))^T$$
$$Z^{[u]} = (Z(s_o)$$

$$\beta = (a_1, \ldots, a_k)^T$$

and

$$X^{[g]} = \begin{pmatrix} f_1(s_1) \ldots f_k(s_1) \\ \vdots \qquad\quad \vdots \\ f_1(s_g) \ldots f_k(s_g) \end{pmatrix}$$

$$X^{[u]} = \begin{pmatrix} f_1(s_o) \ldots f_k(s_o) \end{pmatrix}.$$

The random components, $Z^{[g]}$ and $Z^{[u]}$, are a zero-mean Gaussian process with covariance matrix $Q_{gg}$ among the monitored sites and variance at the un-monitored locations $Q_{uu}$, respectively, as well as a cross-covariance matrix $Q_{ug}$. The covariance matrix of the combined ungauged and gauged sites can be written as

$$Q = \begin{pmatrix} Q_{uu}\ Q_{ug} \\ Q_{gu}\ Q_{gg} \end{pmatrix}.$$

Handcock and Stein (1993) then use the isotropic Whittle–Matern class, with unknown parameters, to model the covariance structure as a parametric function of interdistances between locations. That is, if $q_{ij}$ denotes the $(i, j)$th element of $Q$ and covariance between locations $s_i$ and $s_j$, $q_{ij}$ can be written as

$$q_{ij} = \frac{1}{\theta} K_\eta(|s_i - s_j|).$$

Here $|s_i - s_j|$ denotes the distance between the locations and $K_\eta(x)$ has the general form

$$K_\eta(x) = \frac{1}{2^{\eta_2 - 1}\Gamma(\eta_2)} \left(\frac{x}{\eta_1}\right)^{\eta_2} \kappa_{\eta_2}\left(\frac{x}{\eta_1}\right) \tag{8.6}$$

where $\kappa_{\eta_2}$ is a modified Bessel function of order $\eta_2$ (Abramowitz and Stegun 1970). The two parameters $\eta_1$ and $\eta_2$ control the range of correlation and the smoothness of the random field.

In terms of spatial correlations, denote

$$Q_{gg} = \frac{1}{\theta} S_{gg}\ Q_{uu} = \frac{1}{\theta} S_{uu}\ Q_{ug} = \frac{1}{\theta} S_{ug}.$$

The elements of $S$s represent the spatial correlations between the locations; that is, the $(i, j)$th element of $S$ is $K_\eta(|s_i - s_j|)$.

Under the assumption that the prior distribution, $pr(\beta \mid \eta, \theta)$, is locally uniform and $\theta$ has a Jeffrey's invariant prior, Handcock and Stein (1993) show that the predictive distribution of $Y^{[u]}$ given the observed data at the monitored locations is a Student $t$ distribution for a given value of hyperparameter $\eta$.

The distribution is specified by

$$Y^{[u]} \mid Y^{[g]}, \eta \sim t_{g-k}\left(\hat{Y}^{[u]}, \frac{g}{g-k}\hat{\theta} V_\eta, \nu\right), \tag{8.7}$$

where

$$\begin{aligned}
\hat{Y}^{[u]} &= \tau Y^{[g]} + \left(X^{[u]} - \tau X^{[g]}\right)\hat{\beta} \\
\tau &= S_{ug}S_{gg}^{-1} \\
\hat{\beta} &= \left(\left(X^{[g]}\right)^T S_{gg}^{-1} X^{[g]}\right)^{-1}\left(X^{[g]}\right)^T S_{gg}^{-1} Y^{[g]} \\
V_\eta &= \left(S_{u|g} + \left(X^{[u]} - \tau X^{[g]}\right)\left(\left(X^{[g]}\right)^T S_{gg}^{-1} X^{[g]}\right)^{-1}\left(X^{[u]} - \tau X^{[g]}\right)^T\right) \\
S_{u|g} &= S_{uu} - S_{ug}S_{gg}^{-1}S_{gu} \\
\hat{\theta} &= \left((Y^{[g]} - X^{[g]}\hat{\beta})S_{gg}^{-1}(Y^{[g]} - X^{[g]}\hat{\beta})'\right)^{-1}.
\end{aligned}$$

When the hyperparameter $\eta$ for the Matern covariance function is known, the predictive distribution (8.7) is completely specified. In the case where $\eta$ is unknown, a prior distribution can be imposed. The predictive distribution is then generally not simplified and numerical methods are required for evaluation. Handcock and Stein (1993) illustrate the method through an application of topographical data and perform a sensitivity analysis with respect to the choice of priors. Their results suggest that the approach is reasonably robust. More details can be found in Handcock and Stein (1993).

Recently the Handcock and Stein approach has been extended to more general settings by Gaudard et al. (1999). The authors develop a Bayesian kriging framework for stationary processes that allows for arbitrary prior distributions and use the recent advents in Markov chain Monte Carlo methodologies for analysis. This MCMC-based approach for dealing with stationary spatial processes is well covered in the recent book by Banerjee et al. (2004) and is not covered here to due to space availability. The Handcock and Stein approach has also been extended to deal with non-Gaussian random fields by De Oliveira et al. (1997) which is described in the next section.

## 8.3 The Bayesian Transformed Gaussian Approach

In environmental applications, the random fields are often non-Gaussian and it is hence necessary to transform the variables to satisfy the normality assumption required in the formulation of the Bayesian kriging framework described above. De Oliveira et al. (1997) extend the Handcock and Stein approach to incorporate such a transformation into the Bayesian kriging framework to cope with the non-Gaussian random fields. Their approach, termed *Bayesian Transformed Gaussian* (BTG), starts by assuming the transformed response variable follows a Gaussian distribution, then deriving its predictive distribution similar to the Hancock and Stein results and ends by converting the derived predictive distribution to that of the untransformed response vari-

able. Uncertainty about the transformation can be incorporated in the last step although numerical methods are required for implementation.

The approach requires the transformation to belong to a parametric family of monotone transformations. Thus this approach, unlike trans-Gaussian kriging (Cressie 1993), avoids the selection of one specific transformation. It is well-known that using a single transformation may lead to bias in prediction since the selection could be substantially affected by a few influential data points (see Atkinson and Shepard 1996). The BTG approach takes into account the uncertainty associated with the transformation operation. The approach is hence partially robust against model misspecification.

### 8.3.1 The BTG Model

Denote the random field by

$$Y^{[g]} = (Y(s_1), \ldots, Y(s_g))^T$$
$$Y^{[u]} = (Y(s_{o_1}), \ldots, Y(s_{o_u}))^T,$$

where the superscripts $g$ and $u$ denote the $g$ monitored sites $(s_1, \ldots, s_g)$ and the $u$ unmonitored site $(s_{o_1}, \ldots, s_{o_u})$, respectively.

Let $\mathcal{G} = \{h_\lambda(.) : \lambda \in R\}$ be a parametric family of transformations where $h_\lambda(.)$ is a monotone transformation for a given $\lambda$ and the first derivative $g'_\lambda(x) = (\partial/\partial x)h_\lambda(x)$ exists and is continuous. The Box–Cox power transformation is an example of such families (Box and Cox 1964) defined as

$$h_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0. \end{cases}$$

Another example is that proposed by Aranda-Ordaz (1981) to transform proportions defined as

$$h_\lambda(x) = \begin{cases} \log((1-x)^{-\lambda} - 1) - \log\lambda & \text{if } \lambda \neq 0 \\ \log(-\log(1-x)) & \text{if } \lambda = 0 \end{cases}$$

for $0 < x < 1$.

De Oliveira et al. (1997) assume that the transformed response follows a Gaussian distribution given by

$$\begin{pmatrix} h_\lambda(Y^{[u]}) \\ h_\lambda(Y^{[g]}) \end{pmatrix} \mid \beta, \theta, \eta, \lambda \Bigg) \sim N_{u+g} \left( \begin{pmatrix} X^{[u]}\beta \\ X^{[g]}\beta \end{pmatrix}, \frac{1}{\theta} S_\eta \right), \tag{8.8}$$

where

$$\beta = (a_1, \ldots, a_k)^T$$

$$X^{[g]} = \begin{pmatrix} f_1(s_1) & \ldots & f_k(s_1) \\ \vdots & & \vdots \\ f_1(s_g) & \ldots & f_k(s_g) \end{pmatrix}$$

$$X^{[u]} = \begin{pmatrix} f_1(s_{o_1}) & \dots & f_k(s_{o_1}) \\ \vdots & & \vdots \\ f_1(s_{o_u}) & \dots & f_k(s_{o_u}). \end{pmatrix}.$$

Here $S_\eta$ denotes the correlation matrix; that is, its elements represent pairwise correlations between corresponding locations. The spatial correlation between two locations $s_i$ and $s_j$ is assumed to be isotropic and follow a known function of the distance between the locations, i.e., $K_\eta(|s_i - s_j|)$ in (8.6).

### 8.3.2 Prior Distribution

De Oliveira et al. (1997) caution that the choice of the prior distribution should be cautiously chosen since the interpretation of $\beta$, $\theta$, and $\eta$ depend on the specific value of $\lambda$. For example, assuming the prior distribution for these parameters to be independent could lead to nonsensical results since the location and scale of the transformed response variable and the spatial correlation are functions of $\lambda$ (Box and Cox 1964).

Following suggestions by Box and Cox (1964), the authors assume the prior distribution to be

$$p(\beta, \theta, \eta, \lambda) \propto \frac{p(\eta)p(\lambda)}{\theta J_\lambda^{k/g}}, \tag{8.9}$$

where $p(\eta)$ and $p(\lambda)$ are assumed to be continuous and $J_\lambda = \prod_{i=1}^{g} |h'_\lambda(y(s_i))|$ is the Jacobian of the transformation for a given $\lambda$.

Another family of prior distributions proposed by Perrichi (1981) could also be used where

$$p(\beta, \theta, \eta, \lambda) \propto p(\eta)p(\lambda)\theta^{k/2-1}.$$

De Oliveira et al. (1997) use these two prior distributions in the data analysis that yields essentially identical results in spatial prediction.

### 8.3.3 Predictive Distribution

Under the model (8.8)-(8.9), the predictive distribution for the transformed response variable at the ungauged locations given the observed data at the gauged sites and the hyperparameters $(\lambda, \eta)$, is a $k$-variate Student $t$ distribution (De Oliveira et al. 1997)

$$h_\lambda(Y^{[u]}) \mid \eta, \lambda, Y^{[g]} \sim t_{g-k}\left(m_{\lambda,\eta}, \frac{g}{g-k}\tilde{q}_{\lambda,\eta}C_\eta, \nu\right), \tag{8.10}$$

where

$$
\begin{aligned}
m_{\lambda,\eta} &= \tau h_\lambda(Y^{[g]}) + \left(X^{[u]} - \tau X^{[g]}\right)\hat{\beta} \\
\tau &= S_{ug}S_{gg}^{-1} \\
\hat{\beta} &= \left(\left(X^{[g]}\right)^T S_{gg}^{-1} X^{[g]}\right)^{-1}\left(X^{[g]}\right)^T S_{gg}^{-1} h_\lambda(Y^{[g]}) \\
C_\eta &= \left(S_{u|g} + \left(X^{[u]} - \tau X^{[g]}\right)\left(\left(X^{[g]}\right)^T S_{gg}^{-1} X^{[g]}\right)^{-1}\left(X^{[u]} - \tau X^{[g]}\right)^T\right) \\
S_{u|g} &= S_{uu} - S_{ug}S_{gg}^{-1}S_{gu} \\
\tilde{q}_{\lambda,\eta} &= \left((Y^{[g]} - X^{[g]}\hat{\beta})S_{gg}^{-1}(Y^{[g]} - X^{[g]}\hat{\beta})'\right)^{-1}.
\end{aligned}
$$

It is easy to see that this predictive distribution is identical to that given in (8.7) derived by Handcock and Stein (1993) when no transformation is used; i.e., $h_\lambda(x) = x$.

For a general transformation, the predictive distribution of the untransformed response variable can be derived using the distribution (8.10) and the prior distribution of $\eta$ and $\lambda$. Specifically, the predictive distribution can be expressed as

$$
p(Y^{[u]} \mid Y^{[g]}) = \int\int p(Y^{[u]} \mid \lambda, \eta, Y^{[g]}) p(\lambda, \eta \mid Y^{[g]}) \partial\lambda\partial\eta. \tag{8.11}
$$

Generally the closed form for this predictive distribution is not available since the integration involved is intractable and has to be done numerically. De Oliveira et al. (1997) propose a numerical integration algorithm for this evaluation which is described next.

### 8.3.4 Numerical Integration Algorithm

The predictive distribution (8.11) can be rewritten using Bayes' rules as

$$
p(Y^{[u]} \mid Y^{[g]}) = \frac{\int\int p(Y^{[u]} \mid \lambda, \eta, Y^{[g]}) p(Y^{[g]} \mid \lambda, \eta) p(\lambda) p(\eta) \partial\lambda\partial\eta}{\int\int p(Y^{[g]} \mid \lambda, \eta) p(\lambda) p(\eta) \partial\lambda\partial\eta}. \tag{8.12}
$$

Assuming the prior distributions $p(\lambda)$ and $p(\eta)$ to be proper, De Oliveira et al. (1997) propose the following Monte Carlo approach to evaluate the expression:

1. Partition the effective range of $Y^{[u]}$ into a set $S$.
2. Generate $m$ realizations from the prior distributions $p(\lambda)$ and $p(\eta)$; that is,

   $$\text{simulate } \eta_1, \ldots, \eta_m \sim_{iid} p(\eta) \text{ and } \lambda_1, \ldots, \lambda_m \sim_{iid} p(\lambda).$$

3. For each value of $z_0 \in S$, the approximation to $p(Y^{[u]} = z_0 \mid Y^{[g]})$ is given by

   $$\hat{p}_m(z_0) = \sum_{i=1}^{m} p(Y^{[u]} = z_0 \mid \lambda_i, \eta_i, Y^{[g]}) w(\lambda_i, \eta_i),$$

where

$$w(\lambda_i, \eta_i) = \frac{p(Y^{[g]} \mid \lambda_i, \eta_i)}{\sum_{i=1}^{m} p(Y^{[g]} \mid \lambda_i, \eta_i)}.$$

Here $p(Y^{[u]} = z_0 \mid \lambda_i, \eta_i, Y^{[g]})$ is the density of the predictive distribution (8.10) evaluated at $z_0$ and

$$p(Y^{[g]} \mid \lambda_i, \eta_i) = c|S_{gg}|^{-1/2}|(X^{[g]})' S_{gg}^{-1} X^{[g]}|^{-1/2} \tilde{q}_{\lambda_i,\eta_i}^{-(g-k)/2} J_{\lambda_i}^{1-(k/n)}, \tag{8.13}$$

where $c$ is the proportionality constant and not relevant in the calculation of $w(\lambda_i, \eta_i)$.

It can be proved that as $m \longrightarrow \infty$, the approximation $\hat{p}_m(z_0)$ converges to $p(Y^{[u]} = z_0 \mid Y^{[g]})$ (Geweke 1989).

In the case that the prior distributions $p(\lambda)$ and/or $p(\eta)$ are improper, a Markov chain Monte Carlo approach may be used (Tanner 1996). More details can be found in De Oliveira et al. (1997).

## 8.4 Remarks

The Bayesian framework for interpolation described above overcomes several deficiencies associated with classical kriging. It allows for uncertainty associated with the parameters to be taken into account in the derivation of the predictive distribution. It can be used to simultaneously predict random field values at multiple locations, along with their corresponding error bands. However, these developments still generally assume the field to be stationary and/or isotropic. Such assumptions are mostly unrealistic for environmental factors (Guttorp et al. 1993; Brown et al. 1994a; Le et al. 2001). On the other hand, the recently proposed integrated framework for Bayesian spatial and temporal interpolation starting with our work (Le and Zidek 1992) overcomes this deficiency. In this integrated framework, unlike that of Kitanidis, the spatial covariance field is left completely unspecified and hence stationarity is not required. We originally derived our method for a univariate response (Le and Zidek 1992) but have with coinvestigators subsequently extended it to multivariate responses with various patterns of missing data (Brown et al. (1994a); Le and Zidek 1994a; Le et al. 1997, 2001; and Kibria et al. 2002). Empirical comparisons with real data using cross-validation suggests the method works quite well (Sun et al. 1998). In particular, Sun (1998) demonstrates that the multivariate Bayesian spatial predictors outperform cokriging, a variant of kriging in a multivariate setting. The integrated frameworks for various settings are described in the next chapter.

# 9

# Hierarchical Bayesian Kriging

*The only relevant thing is uncertainty—the extent of our own knowledge and ignorance.*

*Bruno de Finetti*

In accord with de Finetti's comment uncertainty (and ignorance!) abound in environmental science. So analysis must fully embrace it. The approach featured in this chapter seeks to do just that in the analysis of environmental processes.

The kriging approach (Krige 1951, see Chapter 7) has been very successful in spatial interpolation (or prediction) for geological applications. There random fields are quite homogeneous with respect to spatial correlation. That fact has made success possible in spite of a paucity of data, there being just one single realization typically available at each of the limited number of locations due to the substantial cost of making measurements.

However, in environmental application spatial interpolation problems are quite different. There the random fields are generally quite heterogeneous. They often change over time adding temporal components to the interpolation problem. Moreover, they are influenced by meteorology, making prediction challenging. Finally concentration levels at each monitored location are often measured sequentially over time, yielding more data but additional complexity. Historically due to a lack of practical alternatives, kriging has been used. However, its success has been limited by the heterogeneity of these random fields. The assumption of an isotropic covariance structure is simply not tenable as demonstrated in several applications presented in the literature (Guttorp et al. 1993; Brown et al. 1994a; Le et al. 1997, 2001; Kibria et al. 2002). This deficiency also presents itself in the Bayesian kriging approaches described in Chapter 8 since they too rely on that crucial assumption.

In recent years a Bayesian alternative to kriging has been developed for use in environmental settings where the assumption of isotropy is not realistic. These developments (see Le and Zidek 1992, hereafter LZ), based on an integrated Bayesian hierarchical framework, overcome deficiencies associated with kriging and its Bayesian variants.

### Homogeneous Subdomains

A key initial step is the subdivision, if necessary, of the region of interest into relatively homogeneous subregions. These subdivisions can flow from structural factors. For example, if a study concerns a number of distinct communities, community could be the partitioning factor. Alternatively, that factor could be topographical zone if a region were divided by, say ridges of higher elevation. Other such factors include *meteorological regime* and *catchment area*. Alternatively partitioning can be done on a statistical basis as in Wu and Zidek (1992) who cluster a large group of sites (and in turn the region) on the basis of data they generated in the past.

Ideally the random field should be pretty well described by a common temporal model. For example, at all sites within a subregion the field should have similar time trends. They should be predicted by a common covariate model, temperature being an example of a covariate for ozone concentrations over an urban area. Moreover, after subtracting the common trend and covariate model, the resulting field of residuals should be approximately second-order stationary at all sites and be approximately described by a single time-series model. For example, the residuals at each site could be a zero mean *autoregressive process of order 1 (AR(1))*.

Provided the subregions are not too small, enough data will be available to accurately estimate the common coefficients in the shared models. In fact Savage's *principle of precise measurement* (see Savage 1971) can be used to justify doing so even within a Bayesian framework. Prior modeling can thus be avoided at this stage. Moreover, if the partitioning strategy has been effective, removal of all the shared components from the original field, as demonstrated in Chapter 14, will leave a much simpler residual random field to model. The approach avoids the necessarily high cost of a complex approach such as dynamic linear modeling described in Section 5.3 (West and Harrison 1999), with its heavy computational burden.

The subdivisions can be surprisingly large in some cases. For example, the $PM_{10}$ field over the Northeastern United States is extremely flat; a single subdivision will cover a very large area. In his study of acid deposition fields, Sun's single subdivision is even larger (Sun 1998).

At the same time, as demonstrated in ensuing developments, the theory does not need the strategy above to completely achieve its ideal result. For one thing it has the capacity to admit trend and covariates on a site by site basis. As well, the flexibility of the multivariate spatial predictor enables us to avoid modeling short-term (and hence complex) autocorrelation structures, as might be seen in hourly average concentrations, for example. Anyway we now assume this preliminary analysis has been completed.

### Modeling Trend and Covariance

Specifically our approach assumes, in the first level of the Bayesian hierarchy, that the (suitably transformed) random field follows a finite-dimensional

Gaussian distribution, its mean function depending on an unknown parameter or parameter matrix $B$. The corresponding covariance field $\Sigma$ is given no specific structure. At the second level of the hierarchy, conjugate prior distributions are adopted for the parameters $B$ and $\Sigma$. The predictive distribution, conditional on hyperparameters, is then derived from the prior distribution. The method turns out to be very general, accommodating such things as trends and seasonality in the mean function while allowing model misspecification to be corrected as data become available. Model uncertainty has been incorporated in the posterior predictive distribution which is thereby made robust against model misspecification.

The LZ method can deal with random fields having nonstationarity spatial features since the predictive (posterior) distribution is derived without any specific structures necessarily required for the hyperparameters. Such features can be captured in the estimation of the hypercovariance matrix using, for example, the Sampson and Guttorp method (see Section 6.5.1) for estimating nonstationary spatial covariance structure as demonstrated in Chapter 2. Given the hyperparameters, the resulting predictive distribution is a product of multivariate $t$-distributions and hence is completely specified (Le and Zidek 1992). The LZ approach has since been extended to multivariate responses. At each monitoring station, a set of environmental responses is measured, yet not all stations need to measure the same set. In fact the methodology can cope with a variety of missing data patterns (Brown et al. 1994a; Le et al. 1997, 2001; Kibria et al. 2002).

**Hierarchical Approach**

The value of the hierarchical approach embodied in this integrated framework, unlike that of Kitanidis (1986) among others, lies in its lack of restriction on the form of the $\Sigma$. In particular, the random field need not be stationary. The covariance structure can be modeled at the second level through the hyperparameters, i.e., $B$ and $\Sigma$. Past and future data will then update these prior models through Bayes rule to achieve increasingly more realistic versions of their level-one counterparts $B$ and $\Sigma$. In any case, the uncertainty about the level-one parameters $B$ and $\Sigma$ is reflected in the predictive distribution given above. Consequently, prediction intervals will realistically account for that uncertainty through the heavy-tailed Student-$t$ distributions.

At level two of the hierarchy, parametric models may be specified to accommodate the remaining prior knowledge about the random field under consideration. This will leave some additional (third-stage) hyperparameters unspecified.

If a strictly Bayesian approach were taken, another distribution would be added to the hierarchy to represent the uncertainty about these hyperparameters. One such approach would be a robust Bayesian approach such as that of Pilz (1991).

**Why Empirical Bayes?**

Another would be the use of a diffuse prior on the hyperparameters. However, the latter would also be un-Bayesian and could lead to un-Bayesian characteristics of the posterior distribution (Dawid et al. 1973). But the simplest option is to estimate the hyperparameters from the marginal distribution of the data conditional on them—the empirical Bayes solution. For one thing, the predictive distribution will not be especially sensitive to the choices made of these hyperparameters (Haitovsky and Zidek 1986). In addition the approach would allow for the priors to be matching priors in the sense that the posterior predictive intervals would be well calibrated. That is, 95% intervals would contain the true response about 95% of the time. Meeting that requirement is absolutely crucial and there seems to be no other way given the complexity of the models involved. Finally this empirical Bayes solution would help in designing environmental monitoring networks where computational simplicity is essential.

This chapter offers a mathematical derivation of the integrated framework. That derivation starts with a relatively simple setting where at each station in the monitoring network, univariate measurements (for example, for a specific pollutant) are collected at regular intervals over the same period. A more general setting that reflects practical problems encountered in environmetrics is given, specifically where the observed data follow a monotone pattern. This pattern can arise when data from different monitoring networks have been combined in a single network to improve the performance of spatial interpolators. The method has also been extended to cope with multivariate responses where multiple responses are measured at each monitoring station with different starting times of operations. The simple case with only two blocks for univariate and multivariate responses is presented in this chapter. The notationally challenging general case, with more than two blocks is described in the next chapter along with the corresponding parameter estimation methods.

## 9.1 Univariate Setting

Suppose $g$ locations where levels of a univariate random field are completely observed over time and a further $u$ locations where prediction is required as in Figure 9.1 are considered. Let $Y_t$ be a $p$-dimensional (i.e., strung out) random row vector denoting the random field at time $t$. The first $u$ coordinates are those with no data available (ungauged locations) and the remaining $g$ coordinates are those yielding the data (the gauged stations), $y_t(s_1), \ldots, y_t(s_g)$ for $t = 1, \ldots, n$. The vector $Y_t$ can be partitioned accordingly as $Y_t \equiv (Y_t^{(u)}, Y_t^{(g)})$, where $Y_t^{(u)}$ corresponds to the responses at the $u$ locations without observations and $Y_t^{(g)}$ corresponds to those at the $g$ monitoring locations.

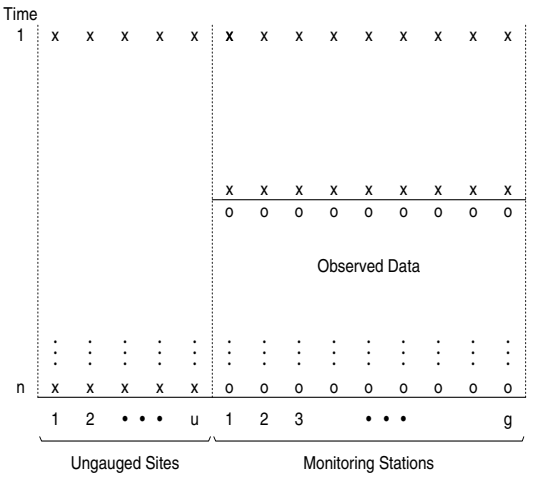


**Fig. 9.1.** Diagram for the observed data (o) at monitoring stations and the unobserved data (×) at locations of interest.

### 9.1.1 Model Specification

Suppose the random variables $\{Y_t\}$, assumed to be independent over time, follow a joint Gaussian distribution; i.e.,

$$Y_t \mid z_t, B, \Sigma \overset{\text{independent}}{\sim} N_p(z_t B, \Sigma), \tag{9.1}$$

where $N_p(\mu, \Sigma)$ denotes the $p$-dimensional Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$. Here $z_t \equiv (z_{t1}, \ldots, z_{tk})$ denotes a $k$-dimensional row vector of covariates and $B$, a $(k \times p)$ matrix of regression coefficients with $p = u + g$,

$$B = \begin{pmatrix} \beta_{1,1} \cdots \beta_{p,1} \\ \vdots \\ \beta_{1,k} \cdots \beta_{p,k} \end{pmatrix} \equiv \begin{pmatrix} B^{(u)} & B^{(g)} \end{pmatrix},$$

partitioned in accord with the partitioning of $Y_t$. The covariates are allowed to vary with time but they must be constant across sites. In contrast, the corresponding regression coefficients may vary over sites.

The covariance matrix $\Sigma$ is partitioned accordingly as

$$\Sigma = \begin{pmatrix} \Sigma_{uu} & \Sigma_{ug} \\ \Sigma_{gu} & \Sigma_{gg} \end{pmatrix}.$$

Here the matrices $\Sigma_{gg}$ and $\Sigma_{uu}$ denote the covariances of $Y_t^{(g)}$ and $Y_t^{(u)}$, respectively. The matrix $\Sigma_{ug}$ represents the corresponding cross-covariance.

### Prior Distribution

Assume $B$ and $\Sigma$ have conjugate prior distributions (see, for example, Anderson 2003, page 272),

$$B \mid B_o, \Sigma, F \sim N_{kp}\left(B_o, F^{-1} \otimes \Sigma\right) \tag{9.2}$$

$$\Sigma \mid \Psi, \delta \sim W_p^{-1}(\Psi, \delta), \tag{9.3}$$

where $W_p^{-1}(\Psi, \delta)$ denotes the $p$-dimensional inverted Wishart distribution with scale matrix $\Psi$ and $m$ degrees of freedom; it is proper if $p < m$.

With this prior model, $F$, $B_o$, and $\Psi$ are hyperparameter matrices, respectively, of dimensions $k \times k$, $k \times 1$, and $p \times p$.

It is convenient to reparameterize $\Sigma$ as $(\Sigma_{gg}, \Sigma_{u|g}, \tau)$ where $\Sigma_{u|g}$ is a $(u \times u)$ matrix denoting the residual covariance of $Y_t^{(u)}$-residuals after optimal linear prediction based on $Y_t^{(g)}$; it is given by

$$\Sigma_{u|g} \equiv \Sigma_{uu} - \Sigma_{ug}\Sigma_{gg}^{-1}\Sigma_{gu}.$$

The $(g \times g)$ matrix $\Sigma_{gg}$ is the covariance matrix of $Y_t^{(g)}$ and $\tau$, a $(u \times g)$ matrix representing the slope of the optimal linear predictor of $Y_t^{(u)}$ based on $Y_t^{(g)}$ given by

$$\tau \equiv \Sigma_{ug}\Sigma_{gg}^{-1}.$$

This 1–1 transformation is achieved through the well-known Bartlett decomposition (Bartlett 1933) given in Appendix 15.2.

Using these new parameters, the conjugate prior distribution (9.3) for $\Sigma$ can be equivalently presented as

$$\begin{aligned} \Sigma_{gg} \mid \Psi, \delta &\sim W_g^{-1}(\Psi_{gg}, \delta - u) \\ \Sigma_{u|g} \mid \Psi, \delta &\sim W_u^{-1}(\Psi_{u|g}, \delta) \\ \tau \mid \Sigma_{u|g}, \Psi &\sim N_{ug}(\tau_o, \Sigma_{u|g} \otimes \Psi_{gg}^{-1}). \end{aligned} \tag{9.4}$$

Here $(\Psi_{gg}, \Psi_{u|g}, \tau_o)$ denotes the decomposition of the prior parameter matrix $\Psi$ analogous to that of $\Sigma$; that is, $\tau_o = \Psi_{ug}\Psi_{gg}^{-1}$ and $\Psi_{u|g} = \Psi_{uu} - \Psi_{ug}\Psi_{gg}^{-1}\Psi_{gu}$. Moreover, $\Sigma_{gg}$ is independent of $(\Sigma_{u|g}, \tau)$ when the prior distribution is proper. See, for example, Caselton et al. (1992) for a more detailed derivation.

### 9.1.2 Predictive Distribution

Let $D = \left\{(y_1^{(g)}, z_1), \ldots, (y_n^{(g)}, z_n)\right\}$ be the observed responses, in other words, the data. That is, given the covariate $z_t$s, the $\{y_t^{(g)}\}$s are independent realizations of

$$Y_t^{(g)} \mid B, \Sigma, z_t \sim N_g(z_t B^{(g)}, \Sigma_{gg}) \tag{9.5}$$

corresponding to the second component of the $Y_t$ in model (9.1). That is, $D$ represents the partially observed data with no observations for the first $u$-coordinates.

We now describe the joint spatial predictive distribution of the random vector $Y_f$, representing the spatial field at a future time $f$. That distribution must be conditional on all the available data $D$ along with the new covariate vector $z_f$. It is described by (9.1) and given in the next theorem (see Le and Zidek 1992).

First, define $\hat{B}^{(g)}$ and $S$ as follows.

$$\hat{B}^{(g)} = A^{-1}C \tag{9.6}$$

$$S = \sum_{t=1}^{n} (y_t^{(g)} - z_t \hat{B}^{(g)})^T (y_t^{(g)} - z_t \hat{B}^{(g)});$$

here

$$C = \sum_{t=1}^{n} z_t^T y_t^{(g)} \tag{9.7}$$

and

$$A = \sum_{t=1}^{n} z_t' z_t,$$

$\hat{B}^{(g)}$ and $S$ being, respectively, the usual least-squares estimates of $B^{(g)}$ and the residual sum of squares in the regression setting.

Partition the prior matrix $B_o$ in a manner analogous to $B$, as $\left(B_o^{(u)}, B_o^{(g)}\right)$. Let the prior distributions of $B$ and $(\Sigma_{gg}, \Sigma_{u|g}, \tau)$ be defined as in (9.2) and (9.3) or equivalently (9.4); $t_r$ denotes the $r$-variate $t$-distribution as described in Appendix 15.1.

The predictive distribution can now be stated. However, before doing so, some additional notation is helpful and it is introduced next.

### Degrees of Freedom

First, we need degrees of freedom, $l = \delta + n - u - g + 1$ and $q = \delta - u + 1$. Here $l$ represents those of the marginal distribution of the observable responses, and $q$, those of the conditional distribution of the unobservable responses conditional on the observable ones. Both share $\delta$, an integer that comes from the priors. That integer represents how much hypothetical prior data went into constructing them. To see this, note that $n$, the number of independent sample records, and $\delta$ have symmetrical roles in the calculation of $l$.

Specifying $\delta$ is part of the difficult general problem of eliciting meaningful priors in complicated situations, here modeling space–time processes. Although, progress is being made on that front, we have chosen instead to let the data help with its specification. In any case, $l$ and $q$ must be positive to ensure that nondegenerate predictive distributions obtain "at the end of the day."

Notice that $l$ and $q$ decrease as $u$ increases. This important observation points anew to the well-known adage, "There is no such thing as a free lunch."

This means that with a fixed amount of data, the number of locations at which unmeasured responses can be predicted is bounded. Although this point seems completely obvious, it has gone largely unrecognized and generally spatial predictions have been made seemingly without limit. How has this been possible?

The answer can be found in the prediction intervals. These have usually been computed as if the prediction were being made at just a single point, i.e., as if $u = 1$. In fact, simultaneous confidence intervals are needed to achieve a simultaneous coverage probability of, say 95%. (As we show in Section 10.8, these can be easily found within the ambit of our theory via confidence ellipsoids.) These simultaneous intervals increase in width as $u$ increases to the point where the (simultaneous) predictions will be seen as completely unreliable, even though at each point they may be sensible.

So how can this conundrum be resolved? The answer lies in the purpose to which these predictions are put. If the decision-maker really needs these predictions to be simultaneously valid, as say when computing some quantity based on an aggregate of them, then he will only be able to make a small number of them reliably.

That point must be borne in mind in implementing our theory with a fitted $\delta$. In fact, we have deliberately included a third-level prior for this hyperparameter to give the investigator some control over its size. However, $q$ will always come out to be positive so the naive user can artificially push $\delta$ up by merely picking a bigger $u$, and seemingly, enjoying that free lunch!

Notice that in $l$, $g$ of the $\delta + n$ degrees of freedom are paid for including that many gauged sites. This loss results from the growth in model uncertainty as the number of responses (sites) owing to the increase in the number of parameters lurking in the distribution model. That uncertainty drains some of the gains in certainty that accrue from the data, a seemingly natural result. However, we were greatly surprised by the simplicity and elegance of this little formula for $l$ that captures the subtle trade-off being made.

The reader might well be surprised that $n$ does not appear in $q$ as it does in $l$. Why not? The important answer to this question bears emphasizing and stems from the nature of the inverted Wishart conjugate prior we have adopted. It makes that posterior distribution for the unobserved responses independent of the data! Of course, if we had added another layer of prior modeling to our hierarchy, this apparent deficiency would have been rectified. However, our more pragmatic strategy accomplishes the same thing, albeit implicitly, since we have chosen to estimate the hyperparameters, including $\delta$. Hence, $n$ does implicitly come into the formula for $q$ and the other elements of the posterior predictive distribution that rely on the hyperparameters.

### Weights

Another in the cast of notational characters that play fundamental roles is $W = (A + F)^{-1}F^{-1}$. Recall that $Z = \sum_{t=1}^{n} z_t' z_t$ plays the same role as the famous $X'X$ matrix in ordinary linear regression. In particular, $A^{-1}$ would

determine the covariance of the vector of coefficients in the linear model relating the observations at the monitoring sites to their covariates, if ordinary least-squares were employed to fit that model.

However, in our Bayesian framework, we must also incorporate the prior counterpart of the $A^{-1}$, namely, $F^{-1}$. That need leads us to the matrix of weights $W$ given above. It tells us how much weight, roughly speaking, should go on the prior mean $B_o^{(g)}$ matrix of coefficients of the linear model for the gauged, i.e., monitoring, sites. In contrast, the counterpart of $W$, $I - W$, gives the weights to be put on the least-squares estimate of those coefficients, namely, $\hat{B}^{(g)}$. In fact, we might anticipate the result below that the posterior mean is a combination of the prior mean matrix and the least-squares estimated matrix.

That concludes our presentation of basic notation. Now we are ready for the predictive distribution itself.

**Predictive Distribution**

The predictive distribution of $Y_f = \left(Y_f^{(u)'}, Y_f^{(g)'}\right)$ conditional on the covariate vector $z_f$ and prior parameters $B_o$ and $(\Psi_{gg}, \Psi_{u|g}, \tau_o)$, is

$$Y_f^{(g)} \mid D \sim t_g\left(\mu^{(g)}, \frac{c}{l}\hat{\Psi}_{gg}, l\right) \tag{9.8}$$

$$Y_f^{(u)} \mid Y_f^{(g)} = y_f^{(g)}, D \sim t_u\left(\mu^{(u)}, \frac{d}{q}\Psi_{u|g}, q\right), \tag{9.9}$$

where, with the notation discussed above, we have the constants,

$$c = 1 + z(A+F)^{-1}z^T$$
$$d = 1 + zF^{-1}z^T + (y_f^{(g)} - z_f B_o^{(g)})\Psi_{gg}^{-1}(y_f^{(g)} - z_f B_o^{(g)})^T$$
$$q = \delta - u + 1$$

and

$$\hat{\Psi}_{gg} = \Psi_{gg} + S + (\hat{B}^{(g)} - B_o^{(g)})'(A^{-1} + F^{-1})^{-1}(\hat{B}^{(g)} - B_o^{(g)}).$$
$$\mu^{(g)} = (I - W)\hat{B}^{(g)} + WB_o^{(g)}$$
$$\mu^{(u)} = z_f B_o^{(u)} + \tau_o\left(y_f^{(g)} - z_f B_o^{(g)}\right). \tag{9.10}$$

Le and Zidek (1992) derive this result and an alternative proof can be obtained as a special case of that given in Appendix 15.4.

Here the posterior covariance matrix, $\hat{\Psi}_{gg}$ up to a scale factor, includes contributions from the prior distribution $(\Psi_{gg})$, the observations $(S)$, and the model. The posterior means $\mu^{(g)}$ and $\mu^{(u)}$ include contributions from the prior distribution and the observations.

**Remarks**

- $\hat{B}^{(g)}, S, C$, and $A$ are given in Equations (9.6) and (9.7). Thus, given the hyperparameters $\{B_o, F, \Psi, \delta\}$, the predictive distribution is completely characterized as a product of two Student's $t$-distributions. For a specified time $f$, where $1 \leq f \leq n$ (i.e., when the measurements are available at the $g$ gauged stations), the predictive distribution for the ungauged locations is given by (9.9). The resulting predictive distribution has a heavier tail than that of the Gaussian distribution, reflecting the incorporation of the uncertainty associated with the model parameters through the prior distributions.
- The incorporation of the observed data, the model, and the prior knowledge is demonstrated in the predictive distribution. In particular, the contribution of the data to the predictive distribution can be seen in its parameters. The mean of the predictive distribution at the gauged sites is a weighted average of the best linear estimate based on the observed data and the prior mean. On the other hand, the mean associated with the ungauged stations is the best linear predictor based on the observed data and the prior knowledge. The matrix $\hat{\Psi}_{gg}$ reflecting the covariances between the gauged sites, is the sum of the corresponding prior matrix, the sample residual sum of squares, and that from the prior model.
- It is important to note that we impose no stationarity restriction on the form of $\Sigma$ in deriving the predictive distribution. Hence the covariance structure of the random field can be incorporated in the modeling of $\Psi$. Methods for estimating the hyperparameters $\{B_o, F, \Psi, \delta\}$, in particular, for incorporating nonstationarity features in the estimation of $\Psi$, are given in Chapter 10.
- The regression model specification (9.1) allows time-varying covariates. This enables trend and seasonality to be incorporated directly into the derivation of the predictive distribution; there is no need to remove them in an ad hoc preliminary analysis. Furthermore, the regression coefficients are allowed to vary from location to location to account for location-specific strengths of the trend and seasonality.
- The integrated framework can be modified to incorporate in $z_t$, spatial modeling coordinates $f_l(s_i), i = 1, \cdots p$, and $l = 1, \cdots L$, where the $f_l$s are specified functions and $s_i$ represents the location of station $i$. These spatial modeling coordinates are used in the universal kriging approach described in Section 7.2. Such spatial models may well be important in applications such as those of geostatistics. These models may also be useful in some kinds of environmental interpolation which are intrinsically local in nature. However, they do not seem useful for environmental problems generally where the measured responses are produced by macroscale space–time processes. In these situations local coordinates have little explanatory value (for a discussion, see Wu and Zidek 1992).

## 9.2 Missing Data

Missing data present an extreme form of measurement error but one that inevitably presents itself in any scientific investigation. They go missing for a variety of reasons. One project involving the authors saw hourly measurements systematically missing at 2am each day. Further investigation revealed that to be the time when the measuring device was shut down for recalibration! High marks for instrument accuracy, but that missing hour proved an obstacle to the analyst since statistical software commonly requires complete data sets. What should the analyst do about them?

Various answers can be given and we touch on a few in this section. However, much has been written on this topic (see Little and Rubin 1987).

Obviously, filling in the missing values with real measurements would be ideal. But that has never been possible in our experience. If computation time is not an issue, then hierarchical Bayesian models embracing the Kalman filter (dynamic linear model) can admit any pattern of missing measurements in a partially observed space–time process, simply as unknown quantities. The analysis can proceed without them.

In fact, this approach also provides a predictive distribution for the missing values that can be used to impute them if necessary, as the mean of that distribution along with, say 95% prediction intervals. If the latter cannot be computed due to the intractability of that distribution, multiple imputation can be used to repeatedly impute those values and hence characterize their uncertainty. That can be done as part of an MCMC run, for example.

In general this very appealing approach does not work in problems even of moderate size where perhaps dozens of sites yield hourly measurements over an entire summer. The computation becomes too demanding even on very high-speed processors. Therefore, other methods are commonly used to fill them in. One approach uses the EM algorithm. However, that may not work because of the analytical difficulties involved. In fact, the technical analysis can be very involved; some models are just not very tractable.

Ad hoc methods are most commonly used. One of these uses linear spatial regression. Suppose the measurement at hour $t$ is missing at site $i$. Assume data are available at that hour for a nonempty set of other sites $S$ (which varies with site and time). Now find the times other than $t$ when all those sites including $i$ had nonmissing measurements. Using those data, fit a regression model with site $i$'s response as the predictand. Use the fitted model to predict the missing value for site $i$ at time $t$. In fact, that response can be imputed at random from a normal with mean equal to the fitted value and variance, the estimated residual regression variance.

Of course, for some $(t, i)$ pairs the data will be missing at all other sites. In that case it may be possible to fit a time-series model for that site and fill in the data using that model as the predictor. When all else fails, it may be necessary to fill in the missing value with an appropriate average of values

that "neighbor" the missing one in some appropriate sense, again at random to maintain the variance in the observed values.

Fortunately two commonly seen patterns of missing data can be conveniently addressed by the hierarchical Bayes approach developed in this book. (Some data may have to be selectively imputed to achieve one of them.) The first pattern, addressed in the next section, is generally called *monotone missing*. We refer to this as the *staircase pattern*. Within the staircase pattern, a second pattern of missing data may arise in the case of multiple responses. That pattern, called *systemically missing* data (Le et al. 1997) arises when the different responses are measured by design at the difference sites. The resulting data are sometimes referred to as misaligned. For any given step, each site has a multiplicity of responses, some of which are unmeasured throughout the time period spanned by that step. We emphasize just the first of these patterns in this chapter and leave the second to Chapter 10 (Section 10.7).

## 9.3 Staircase Pattern of Missing Data

In many applications, data from different networks with stations having different operational periods must be combined in spatial prediction. Even in single networks, stations may be added over time, resulting in their having different starting times of operation. After appropriately reordering the stations in this situation, the data matrix will have a staircase structure. That is, when the data are reassembled in an increasing order of operational periods, the data matrix appears to be an ascending staircase as displayed in Figure 9.2. Each step of the staircase consists of stations with the same start-time. The spatial interpolation objective remains as it was, to obtain the predictive distribution of the random field at locations of interest given the observed responses (data).

Solutions for spatial–temporal interpolation problems having this staircase data structure have been derived by Le et al. (2001). As in the simple setting of no missing data, the authors assume the response vector follows a Gaussian distribution (perhaps after transformation of the raw data). The novelty of this work lies in its incorporation of a general conjugate prior distribution for the covariance matrix, namely a generalized inverted Wishart distribution (GIW). As its name suggests, this distribution, discovered by Brown et al. (1994b) generalizes the well-known inverted Wishart distribution. It overcomes one of the latter's main limitations, its single degrees of freedom parameter. The inverted Wishart works that single parameter pretty hard, making it represent all the uncertainty associated with a positive random matrix (the covariance matrix in this case). In contrast, the GIW distribution has a multiplicity of such parameters to represent that uncertainty. The extension proves useful in environmental applications; one may have different levels of prior knowledge about the covariance structures in different geographical subregions. Moreover, for the staircase, the GIW distribution allows the modeler to express

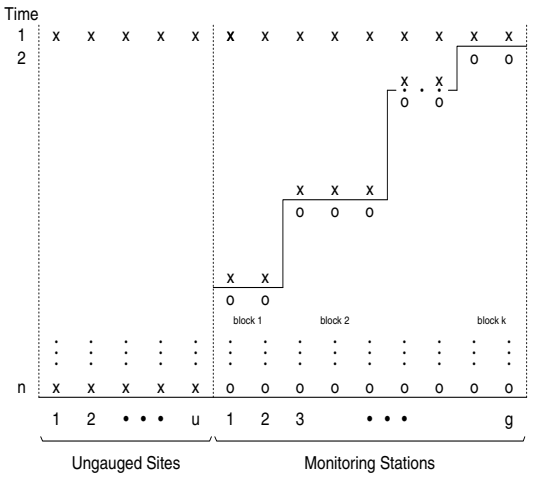


**Fig. 9.2.** Diagram for observed data (o) at monitoring stations having a monotone pattern and unobserved data (×) at locations of interest.

different levels of uncertainty for its steps. That feature accords with common sense. Less would be known about regions where stations have only recently been established. Under the Gaussian-generalized inverted Wishart model for this staircase structure, the resulting predictive distribution is a product of matric-$t$ distributions (Le et al. 2001). However, the mathematical work in the general case of several blocks entails a heavy burden of notation that obscures the main ideas (see Chapter 10). Thus, we present the simpler case of two blocks where we hope the key ideas are clearer.

### 9.3.1 Notation

Consider a special case where $k = 2$ as illustrated in Figure 9.2, i.e., a two-block setting for the data. Let $g_1$ and $g_2$ denote the numbers of monitoring stations in Blocks 1 and 2, respectively, with $g = g_1 + g_2$. In turn, $m_1$ and $m_2$ are the numbers of missing responses in each of the two blocks.

Denote the response variables at the gauged and ungauged sites by

$$Y \equiv \left[Y^{[u]}, Y^{[g]}\right] \equiv \left[Y^{[u]}, Y^{[g_1]}, Y^{[g_2]}\right],$$

where:

- $Y^{[u]} : n \times u$ denotes the matrix of responses at ungauged sites;
- $Y^{[g]} : n \times g$ denotes the matrix of responses at all gauged sites;
- $Y^{[g_j]} : n \times g_j$ denotes the matrix of responses for the $j$th block, $j = 1, 2$.

Partition the responses $Y^{[g_j]}$ into missing and observed components as

$$Y = \left[ Y^{[u]}, \begin{pmatrix} Y^{[g_1^m]} \\ Y^{[g_1^o]} \end{pmatrix}, \begin{pmatrix} Y^{[g_2^m]} \\ Y^{[g_k^o]} \end{pmatrix} \right],$$

where:

- $Y^{[g_j^m]} : m_j \times g_j$ denotes the matrix of missing responses at the gauged sites in the $j$th block, for $j = 1, 2$;
- $Y^{[g_j^o]} : (n - m_j) \times g_j$ denotes the matrix of observed responses at the gauged sites in the $j$th block.

In other words, each row of $Y$ represents the observed ($o$) and unobserved, i.e., missing ($m$) responses at a specific time; here the superscript $b$ in $[g_a^b]$ is either $o$ or $m$ accordingly as the responses are observed or missing. At the same time, the subscript, $a = 1, 2$, designates the block number.

Suppose $l$ time-varying covariate responses $Z_t = (Z_{t1}, \ldots, Z_{tl})^T$ obtain at each timepoint $t$. Temperature measured at a central site would be an example. Assume they are constant across all sites. Let

$$Z = \begin{pmatrix} Z_1^T \\ \vdots \\ Z_n^T \end{pmatrix}.$$

Partition the $l \times (u + g)$ coefficient matrix $\boldsymbol{\beta}$ corresponding to the $l$ covariates and covariance matrix $\Sigma$ of dimension $(u + g) \times (u + g)$ over gauged and ungauged sites as

$$\boldsymbol{\beta} = (\boldsymbol{\beta}^{[u]}, \boldsymbol{\beta}^{[g]}) \text{ and } \Sigma = \begin{pmatrix} \Sigma^{[u,u]} & \Sigma^{[u,g]} \\ \Sigma^{[g,u]} & \Sigma^{[g,g]} \end{pmatrix}.$$

The coefficient matrix $\boldsymbol{\beta}^{[g]}$ corresponding to the gauged sites is further partitioned in conformance with the block structure as

$$\boldsymbol{\beta}^{[g]} = (\boldsymbol{\beta}^{[g_1]}, \boldsymbol{\beta}^{[g_2]}).$$

Likewise partition the covariance matrix $\Sigma^{[g,g]}$ as

$$\Sigma^{[g]} = \begin{pmatrix} \Sigma^{[g_1,g_1]} & \Sigma^{[g_1,g_2]} \\ \Sigma^{[g_2,g_1]} & \Sigma^{[g_2,g_2]} \end{pmatrix}.$$

The following 1–1 transformation (Bartlett 1933) of the matrix $\Sigma$ simplifies the derivation.

$$\Sigma_{22} = \Sigma^{[g_2,g_2]},$$

$$\Gamma_1 = \Sigma^{[g_1,g_1]} - \Sigma^{[g_1,g_2)]}(\Sigma^{[g_2,g_2]})^{-1}\Sigma^{[g_2,g_1]},$$

$$\tau_1 = (\Sigma^{[g_2,g_2]})^{-1}\Sigma^{[(g_2,g_2),g_1]}.$$

### 9.3.2 Staircase Model Specification

The response matrix $Y$ is assumed to follow a Gaussian-generalized inverted Wishart model. Specifically, using the notation described above,

$$\begin{cases} Y \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, I_n \otimes \Sigma), \\ \boldsymbol{\beta} \mid \Sigma, \boldsymbol{\beta}_0, F \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma), \\ \Sigma \sim GIW(\Psi, \delta), \end{cases} \tag{9.11}$$

where $N(\cdot, \cdot)$ denotes the Gaussian distribution, $\boldsymbol{\beta}_0$ is the $l \times (g+u)$ hyperparameter mean matrix of $\boldsymbol{\beta}$, $F^{-1}$ is an $l \times l$ positive definite matrix representing the variance component of $\boldsymbol{\beta}$ between its $l$ rows, and $Z$ is the matrix of covariates. GIW represents the generalized inverted Wishart distribution with $\Psi$ being a collection of hyperparameters and $\delta = (\delta_0, \delta_1, \delta_2)^T$ representing degrees of freedom, as described in Appendix 15.1.

That is, the above GIW distribution is defined, through the Bartlett decomposition, in a stepwise fashion starting with

$$\begin{cases} \Sigma^{[g,g]} \sim GIW(\Psi^{[g]}, \delta^{[g]}), \\ \Gamma^{[u]} \sim IW(\Psi_0, \delta_0), \\ \tau^{[u]} \mid \Gamma^{[u]} \sim N\left(\tau_{00}, H_0 \otimes \Gamma^{[u]}\right), \end{cases} \tag{9.12}$$

where $\Gamma^{[u]} = \Sigma^{[u|g]} = \Sigma^{[u,u]} - \Sigma^{[u,g]}(\Sigma^{[g,g]})^{-1}\Sigma^{[g,u]}$; $\tau^{[u]} = (\Sigma^{[g]})^{-1}\Sigma^{[gu]}$. IW denotes the inverted Wishart with hyperparameters $(\Psi_0, \delta_0)$; the matrix $\tau_{00}$ is the hyperparameter of $\tau^{[u]}$; and the matrix $H_0$ is the variance component of $\tau_u$ between its rows.

The stepwise definition then continues with $\Sigma^{[g,g]}$ through the Bartlett decomposition of $\Sigma^{[g,g]}$, conformably with the block structure, into a new set of variables $\{\Sigma_{22}, \Gamma_1, \tau_1\}$ as described above. The distribution of $\{\Sigma_{22}, \Gamma_1, \tau_1\}$ is given as

$$\begin{cases} \Sigma_{22} \sim IW(\Psi_2, \delta_2), \\ \tau_1 \mid \Gamma_1 \sim N(\tau_{01}, H_1 \otimes \Gamma_1), \\ \Gamma_1 \sim IW(\Psi_1, \delta_1). \end{cases} \tag{9.13}$$

The hyperparameters involved in this two-block Gaussian-GIW model can be written as

$$\mathcal{H} = \{\boldsymbol{\beta}_0, F, \Psi, \delta\}, \tag{9.14}$$

where

$$\Psi = \{\Psi_0, \tau_{00}, H_0, \Psi_1, H_1, \tau_{01}, \Psi_2\},$$

and

$$\delta = (\delta_0, \delta_1, \delta_2)^T.$$

The dimensions of $\Psi$'s components are as follows.

$$\Psi_0 : u \times u, \quad \tau_{00} : g \times u, \quad H_0 : g \times g, \quad \Psi_k : g_k \times g_k$$

$$\Psi_1 : g_1 \times g_1, \quad \Psi_2 : g_2 \times g_2, \quad \tau_1 : g_2 \times g_1, \quad H_1 : g_2 \times g_2.$$

### 9.3.3 The GIW Distribution

1. The GIW distribution, discovered by Brown et al.(1994b), generalizes the IW distribution by allowing different degrees of freedom for a random positive definite matrix.
2. The GIW distribution is a conjugate prior for a Gaussian distribution. This prior is very flexible and quite natural to deal with the staircase structure of the observed data. For example, different degrees of freedom for the $k$ blocks can be expressed through the hyperparameter vector $\delta$.
3. The GIW modeling method also allows considerable latitude in selecting the numbers of blocks in the GIW structure. For example, one could group all sites that started operation at the same time in one block or one could select each site as a block in the stair-case structure.

### 9.3.4 Predictive Distributions

Let $Y_{unob}$ denote the list all the unobserved responses at all locations, i.e.,

$$Y_{unob} = \left\{ Y^{[u]}, Y^{[g_1^m]}, Y^{[g_2^m]} \right\}.$$

Furthermore, the list D includes all the data, i.e., measurements made at the gauged sites:

$$D = \left\{ Y^{[g_1^o]}, \ldots, Y^{[g_k^o]} \right\}.$$

Here we have used the notation $\{\cdot\}$ to emphasize that individual components in the list may have different dimensions since each block can have different numbers of missing observations $m_j$.

With this notation we are in a position to state precisely a key result. Under the model (9.11), the predictive distribution of the unobserved responses conditional on the observed data $D$ and the hyperparameter set $\mathcal{H}$ is given by

$$(Y_{unob} \mid D, \mathcal{H}) \sim \left( Y^{[u]} \mid Y^{[g_1^m]}, Y^{[g_2^m]}, D, \mathcal{H} \right) \left( Y^{[g_1^m]} \mid Y^{[g_2^m]}, D, \mathcal{H} \right) \times \left( Y^{[g_2^m]} \mid D, \mathcal{H} \right), \tag{9.15}$$

where the three components of the conditional distributions are specified as follows.

$$\left(Y^{[g_2^m]} \mid D, \mathcal{H}\right) \sim t_{m_2 \times g_2}\left(\mu^{[2]}_{(u|g)}, \Phi^{[2]}_{(u|g)} \otimes \Psi^{[2]}_{(u|g)}, \delta^{[2]}_{(u|g)}\right); \tag{9.16}$$

$$\left(Y^{[g_1^m]} \mid Y^{[g_2^m]}, D, \mathcal{H}\right) \sim t_{m_1 \times g_1}\left(\mu^{[1]}_{(u|g)}, \Phi^{[1]}_{(u|g)} \otimes \Psi^{[1]}_{(u|g)}, \delta^{[1]}_{(u|g)}\right); \tag{9.17}$$

$$\left(Y^{[u]} \mid Y^{[g_1^m]}, Y^{[g_2^m]}, D, \mathcal{H}\right) \sim t_{n \times u}\left(\mu^{[u|g]}, (\delta_0 - u + 1)^{-1}\Phi^{[u|g]} \otimes \Psi_0, \delta_0 - u + 1\right). \tag{9.18}$$

Here $t_{m_j \times g_j}$ denotes a matric $t$-distribution as described in Appendix 15.1 and for $j = 1, 2$

$$\mu^{[j]}_{(u|g)} = \mu^{[j]}_{(1)} + A^{[j]}_{12}(A^{[j]}_{22})^{-1}(Y^{[g_j^o]} - \mu^{[j]}_{(2)}),$$

$$\Phi^{[j]}_{(u|g)} = \tfrac{\delta_j - g_j + 1}{\delta_j - g_j + n - m_j + 1}\left[A^{[j]}_{11} - A^{[j]}_{12}(A^{[j]}_{22})^{-1}A^{[j]}_{21}\right],$$

$$\Psi^{[j]}_{(u|g)} = \tfrac{1}{\delta_j - g_j + 1}\left[\Psi_j + (Y^{[g_j^o]} - \mu^{[j]}_{(2)})'(A^{[j]}_{22})^{-1}(Y^{[g_j^o]} - \mu^{[j]}_{(2)})\right],$$

$$\delta^{[j]}_{(u|g)} = \delta_j - g_j + n - m_j + 1.$$

At the same time,

$$\mu^{[u|g]} = Z\boldsymbol{\beta}_0^{[u]} + (Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]})\tau_{00},$$

$$\Phi^{[u|g]} = I_n + ZF^{-1}Z' + (Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]})\tilde{\epsilon}^{[g]}H_0(Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]})',$$

with

$$\begin{pmatrix} \mu^{[j]}_{(1)} \\ \mu^{[j]}_{(2)} \end{pmatrix} : \begin{pmatrix} m_j \times g_j \\ (n - m_j) \times g_j \end{pmatrix} = Z\boldsymbol{\beta}_0^{[g_j]} + \tilde{\epsilon}^{[g_{j+1}, \ldots, g_k]}\tau_{0j},$$

$$\begin{pmatrix} A^{[j]}_{11} & A^{[j]}_{12} \\ A^{[j]}_{21} & A^{[j]}_{22} \end{pmatrix} : \begin{pmatrix} m_j \times m_j & m_j \times (n - m_j) \\ (n - m_j) \times m_j & (n - m_j) \times (n - m_j) \end{pmatrix}$$

$$= I_n + ZF^{-1}Z' + \tilde{\epsilon}^{[g_{j+1}, \ldots, g_k]}H_j(\tilde{\epsilon}^{[g_{j+1}, \ldots, g_k]})',$$

where

$$\tilde{\epsilon}^{[g_{j+1}, \ldots, g_k]} = \begin{cases} Y^{[g_{j+1}, \ldots, g_k]} - Z\boldsymbol{\beta}_0^{[g_{j+1}, \ldots, g_k]}, \text{ for } j = 1, \ldots, k-1, \\ 0, \text{ for } j = k. \end{cases}$$

Le et al. (2001) derive that distribution. An alternative derivation obtains as a special case of the general derivation given in Appendix 15.4.

The posterior means, $\mu^{[j]}_{(u|g)}$ for $j = 1, 2$, that combine contributions from data and prior knowledge, represent the best linear predictor for the missing observations at the gauged sites in blocks 1 and 2. Similarly, $\mu^{[u|g]}$ represents the best linear predictors for the contaminant levels at the ungauged sites. The matrices $\Phi^{[j]}_{(u|g)}$ $\Psi^{[j]}_{(u|g)}$ for $j = 1, 2$, and $\Phi^{[u|g]}$ represent the covariance structure of the predictive distribution. For each block of the gauged stations, the observed data from stations within the block contribute to the covariance structure through $\Psi^{[j]}_{(u|g)}$. Moreover, the data from other blocks contribute through $\Phi^{[j]}_{(u|g)}$, the usual form of residual covariance. For the ungauged stations, the observed data at the gauged ones contribute though $\Phi^{[u|g]}$.

**Remarks**

1. Le et al. (2001) refer to (9.16) and (9.17) as hindcasting since they give the joint predictive distribution of the response variables at the gauged sites during their ungauged time period. More precisely,
$$(Y_{hindcasting} \mid D, \mathcal{H}) = \left(Y^{[g_1^m]} \mid Y^{[g_2^m]}, D, \mathcal{H}\right) \times \left(Y^{[g_2^m]} \mid D, \mathcal{H}\right).$$
2. They also refer to (9.18) as *spatial interpolation* since it is the predictive distribution of the response variables at the ungauged sites during the time period under consideration. More precisely
$$(Y_{interpolation} \mid Y_{hindcasting}, D, \mathcal{H}) = \left(Y^{[u]} \mid Y^{[g_1^m, \ldots, g_k^m]}, D, \mathcal{H}\right).$$
3. The result (9.15) can be used to obtain predictive distributions for forecasting. This can be achieved by appropriately choosing $Z$ corresponding to the first $m_k$ components. For example, to forecast the $(n+1)$st month in the illustrative application in Chapter 2, we let the first component of $Z$ be $[1, \cos(2\pi(n+1)/12), \sin(2\pi(n+1)/12)]$
4. In the case $m_j = m$ and $\delta_j = \delta \ \forall j$ (i.e., no staircase), the result (9.15) reduces to the predictive distribution for the simple univariate setting as given in (9.8)–(9.9).

## 9.4 Wrapup

This chapter presents a hierarchical Bayesian framework for environmental space–time fields . It precedes the general framework in the next chapter. By presenting a special case, where the burden of notation is modest, the authors hope the nature of that framework will become clear. Moreover, the material in the next chapter is just a formalistic extension of the material in Section 9.3.

The one substantive distribution lies in Section 10.6 where we find estimators of the hyperparameters for the general model. Anyway, the next chapter

completes our general framework, prior to turning to design issues and applications.

# Part III: Design and Risk Assessment

# 10

# Multivariate Modeling***

*One cannot escape the feeling that these mathematical formulas have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverers, that we get more out of them than was originally put into them.*

*Heinrich Hertz*

Methods for analyzing environmental processes must be able to contend with multivariate response data. Most networks for monitoring such processes collect measurements for a multiplicity of responses (e.g., pollutants or hours) at each station to reduce costs, among other things. For example, the spatial–temporal predictions for Philadelphia by Kibria et al. (2002) concern simultaneous measurements of $PM_{2.5}$ and $PM_{10}$ at each station in the network. Even if a single response is of primary concern, the multivariate approach allows information from the others to be incorporated in inferences about it, with resulting gains in precision.

*Example 10.1. Acid deposition*
Section 4.1.1 describes the acid deposition field. A network of nearly 300 sites monitor its concentrations of nine chemical species yielding a staircase data pattern since they came online at varying times. Tables 10.1 and 10.2 give the interresponse correlations for the two sites singled out for study in that section.

| | Ca | Mg | K | Na | $NH_4$ | $NO_3$ | Cl | $SO_4$ | pH |
|---|---|---|---|---|---|---|---|---|---|
| Ca | 100 | 99 | 96 | 87 | 70 | 62 | 82 | 81 | 59 |
| Mg | 99 | 100 | 95 | 89 | 70 | 64 | 84 | 82 | 62 |
| K | 96 | 95 | 100 | 79 | 80 | 59 | 76 | 83 | 46 |
| Na | 87 | 89 | 79 | 100 | 60 | 64 | 95 | 77 | 58 |
| $NH_4$ | 70 | 70 | 80 | 60 | 100 | 65 | 66 | 84 | 21 |
| $NO_3$ | 62 | 64 | 59 | 64 | 65 | 100 | 75 | 73 | 11 |
| Cl | 82 | 84 | 76 | 95 | 66 | 75 | 100 | 82 | 47 |
| $SO_4$ | 81 | 82 | 83 | 77 | 84 | 73 | 82 | 100 | 26 |
| pH | 59 | 62 | 46 | 58 | 21 | 11 | 47 | 26 | 100 |

**Table 10.1.** Interresponse correlations $\times 100$ for the nine pollutants measured at a site in Colorado as part of the NADP/NTN acid deposition monitoring program.

| | Ca | Mg | K | Na | $NH_4$ | $NO_3$ | Cl | $SO_4$ | pH |
|---|---|---|---|---|---|---|---|---|---|
| Ca | 100 | 59 | 31 | -7 | 79 | 64 | -10 | 73 | -51 |
| Mg | 59 | 100 | 23 | 68 | 38 | 38 | 55 | 39 | -23 |
| K | 31 | 23 | 100 | 4 | 42 | 12 | 7 | 27 | -3 |
| Na | -7 | 68 | 4 | 100 | -24 | 7 | 84 | -18 | 7 |
| $NH_4$ | 79 | 38 | 42 | -24 | 100 | 56 | -24 | 88 | -57 |
| $NO_3$ | 64 | 38 | 12 | 7 | 56 | 100 | 1 | 61 | -81 |
| Cl | -10 | 55 | 7 | 84 | -24 | 1 | 100 | -19 | 14 |
| $SO_4$ | 73 | 39 | 27 | -18 | 88 | 61 | -19 | 100 | -79 |
| pH | -51 | -23 | -3 | 7 | -57 | -81 | 14 | -79 | 100 |

**Table 10.2.** Interresponse correlations like those in Table 10.1 but here for the state of Maine.

These tables reveal some consistently strong associations. Some seem surprising, for instance between $NO_3$ and Ca concentrations in both states. While that between Na and Na in Maine is expected, why Colorado?

In any case, this example illustrates the importance of the topic of this chapter, multivariate response models. Coupled with strong associations such as those above, they enable the species of interest (say $NO_3$) to borrow strength from another (Ca) in tasks such as spatial prediction and design (see Chapter 11).

The tables also reveal some interesting association "flips" like Colorado's large positive association between Cl and Ca going to negligible in Maine. The latter demonstrates the possible failure of the Kronecker product covariance structure assumed below, when dealing with continentwide processes. On that large scale, analyses of the type described in this chapter need to be done locally. Fortunately the common availability of large amounts of temporal data makes that approach very practical.

This chapter extends the preceding one's integrated framework to admit multivariate responses and the general $k$-step staircase pattern of missing data. Although the developments take us through some very complex notation, all follows from some simple basic ideas as the Hertz quote above suggests.

Like its univariate cousin, the multivariate extension assumes the random field's first hierarchical level has a joint Gaussian distribution. Its mean function depends on an unknown parameter (matrix) $B$ while the corresponding covariance matrix $\Sigma$ has no specific structure. Conjugate prior distributions are assumed for $B$ and $\Sigma$ at the hierarchy's second level, the generalized inverted Wishart (GIW) (see Appendix 15.1) being a good choice for the latter in any Gaussian random field such as that assumed here. A Kronecker structure imposed on the hyperscale matrix sidesteps the estimation of the otherwise prohibitively large number of parameters. In other words, assume equality across sites of the hyperscale matrix for the covariance of the multivariate response.

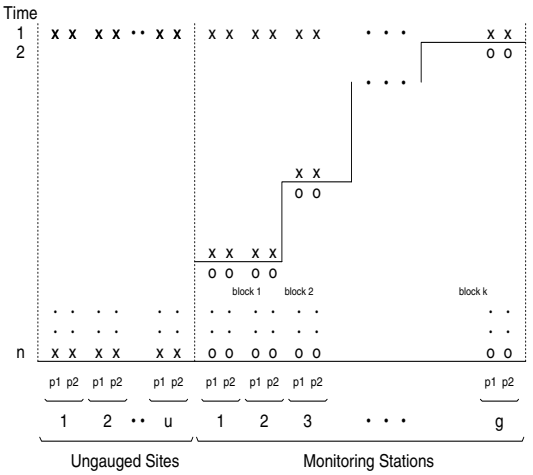


**Fig. 10.1.** Diagram for data (o) at monitoring stations having a monotone pattern and unobserved responses (×) at locations of interest. In this case the response at each station has two responses: p1 and p2.

## 10.1 General Staircase

Responses are organized in a staircase structure; the steps, each consisting of sites with the same start-up time, are arranged in increasing order as in Figure 10.1. As in the univariate case the resulting predictive distribution can be expressed as a product of conditional matric-$t$ distributions. The unobserved responses in each block conditional on the data and the responses at the higher block(s) then follow a matric-$t$ distribution.

This chapter presents the derivation of the predictive distribution while proofs are deferred to Appendix 15.4. The posterior distributions for $B$ and $\Sigma$ given the data, along with their posterior expectations are also derived.

Given the hyperparameters, the predictive distributions are completely determined. As noted in earlier chapters, a strictly Bayesian approach would entail the addition of another distribution layer to the hierarchy to accommodate the uncertainty about these hyperparameters. However, for reasons given in the introduction of Chapter 9, we opt instead for an empirical Bayes solution described in detail in Section 10.6. The mathematical derivations are given later in the chapter. We begin with a glossary of the notation used in those derivations.

### 10.1.1 Notation

Let

$p$ = number of responses considered at each station;

$n$ = number of timepoints (e.g., number of months);

$u$ = number of locations with no monitors (i.e., ungauged sites);

$g$ = number of locations with monitors (i.e., gauged sites).

### Staircase Notation

As in the univariate case, the stations are organized into $k$ blocks where the $g_j$ $(j = 1, 2, \ldots, k)$ sites in the $j$th block have the same number of timepoints $m_j$ at which, by design, no measurements are taken. These blocks are numbered so that the measurements correspond to a monotone data pattern or a staircase structure as depicted in Figure 10.1, that is,

$$m_1 \geq m_2 \geq \cdots \geq m_k \geq 0.$$

If the responses prior to the first monitor in operation are of interest, then $m_k$ is set to be bigger than 0.

### Response Variables

The response variables can accordingly be organized as

$$Y = \left[ Y^{[u]}, Y^{[g]} \right].$$

Here $Y^{[u]} : n \times up$ denotes the unobserved responses at ungauged sites while $Y^{[g]} : n \times gp$ is given by

$$Y^{[g]} = \left[ Y^{[g_1]}, \ldots, Y^{[g_k]} \right] = \left[ \begin{pmatrix} Y^{[g_1^m]} \\ Y^{[g_1^o]} \end{pmatrix}, \ldots, \begin{pmatrix} Y^{[g_k^m]} \\ Y^{[g_k^o]} \end{pmatrix} \right],$$

the missing (unmeasured) and measured responses at gauged sites. Thus:

- $Y^{[g_j^m]} : m_j \times g_j p$ is the matrix of missing responses at the $g_j$ gauged sites for the $m_j$ timepoints;
- $Y^{[g_j^o]} : (n - m_j) \times g_j p$ is the matrix of measurements at the $g_j$ gauged sites for timepoint $(n - m_j)$ .

That is, each row of the matrix $Y$ represents the multivariate responses, measured and unmeasured, for all locations at a given time. As in Chapter 9, here the superscripts $m$ and $o$ denote the missing and observed responses, respectively, and the subscript indicates a particular block from 1 to $k$.

### Covariates

Suppose $l$ time-varying covariate responses $Z_t = (Z_{t1}, \ldots, Z_{tl})^T$ are obtained at each timepoint $t$ and assumed constant across all sites. Let

$$Z = \begin{pmatrix} Z_1^T \\ \vdots \\ Z_n^T \end{pmatrix}.$$

**Parameter Partitioning**

Partition the coefficient matrix $\boldsymbol{\beta} : l \times (u + g)p$ corresponding to the $l$ covariates in conformance with the block structure as

$$\boldsymbol{\beta} = (\boldsymbol{\beta}^{[u]}, \boldsymbol{\beta}^{[g_1]}, \ldots, \boldsymbol{\beta}^{[g_k]}).$$

Likewise, partition the covariance matrix $\Sigma : (u+g)p \times (u+g)p$ over gauged and ungauged sites conformably as

$$\Sigma = \begin{pmatrix} \Sigma^{[u,u]} & \Sigma^{[u,g]} \\ \Sigma^{[g,u]} & \Sigma^{[g,g]} \end{pmatrix},$$

$\Sigma^{[u,u]} : up \times up$ being for the ungauged sites. Further partition the covariance matrix $\Sigma^{[g,g]} : gp \times gp$ for the gauged site blocks:

$$\Sigma^{[g,g]} = \begin{pmatrix} \Sigma^{[g_1,g_1]} & \ldots & \Sigma^{[g_1,g_k]} \\ \vdots & \ldots & \vdots \\ \Sigma^{[g_k,g_1]} & \ldots & \Sigma^{[g_k,g_k]} \end{pmatrix}.$$

As well, for $j = 1, \ldots, k$ let

$$\Sigma^{[g_j,\ldots,g_k]} = \begin{pmatrix} \Sigma^{[g_j,g_j]} & \ldots & \Sigma^{[g_j,g_k]} \\ \vdots & \ldots & \vdots \\ \Sigma^{[g_k,g_j]} & \ldots & \Sigma^{[g_k,g_k]} \end{pmatrix}.$$

**The Bartlett Transformation**

Deriving the predictive distribution is facilitated by reparameterizing the matrix $\Sigma$ through the recursive 1–1 Bartlett transformation for the $k$ blocks described in Appendix 15.2. Specifically, define new parameters as

$$\Gamma^{[u]} = \Sigma^{[u,u]} - \Sigma^{[u,g]}(\Sigma^{[g,g]})^{-1}\Sigma^{[g,u]},$$

$$\tau^{[u]} = (\Sigma^{[g,g]})^{-1}\Sigma^{[g,u]},$$

$$\Gamma_k = \Sigma^{[g_k,g_k]}, \quad \text{and for } j = 1, \ldots, k-1$$

$$\Gamma_j : g_j p \times g_j p = \Sigma^{[g_j,g_j]} - \Sigma^{[g_j,(g_{j+1},\ldots,g_k)]}(\Sigma^{[g_{j+1},\ldots,g_k]})^{-1}\Sigma^{[(g_{j+1},\ldots,g_k),g_j]},$$

$$\tau_j : (g_{j+1} + \cdots + g_k)p \times g_j p = (\Sigma^{[g_{j+1},\ldots,g_k]})^{-1}\Sigma^{[(g_{j+1},\ldots,g_k),g_j]},$$

where

$$\Sigma^{[(g_{j+1},\ldots,g_k),g_j]} = \begin{pmatrix} \Sigma^{[g_{j+1},g_j]} \\ \vdots \\ \Sigma^{[g_k,\ g_j]} \end{pmatrix},$$

for $j = 1, \ldots, k-1$.

## 10.2 Model Specification

Assume the response matrix $Y$ follows the Gaussian and generalized inverted Wishart model specified by

$$\begin{cases} Y \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, I_n \otimes \Sigma), \\ \boldsymbol{\beta} \mid \Sigma, \boldsymbol{\beta}_0, \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma), \\ \Sigma \sim GIW(\Theta, \delta). \end{cases} \tag{10.1}$$

Here: $N(\cdot, \cdot)$ denotes the multivariate Gaussian distribution of appropriate dimension, $\boldsymbol{\beta}_0 : l \times (g+u)p$, the hyperparameter mean matrix of $\boldsymbol{\beta}$, $F^{-1} : l \times l > 0$, the variance component of $\boldsymbol{\beta}$ between its $l$ rows, and $Z$ the matrix of covariates. 1–1 denotes the generalized inverted Wishart distribution (Appendix 15.1), where $\{\Theta, \delta\}$ is a set of model parameters specified below and $\otimes$ represents the Kronecker product between two matrices defined as

$$A_{p\times q} \otimes B_{m\times n} = \begin{bmatrix} a_{11}B & \cdots & a_{1q}B \\ \vdots & & \vdots \\ a_{p1}B & \cdots & a_{pq}B \end{bmatrix}_{pm\times qn}.$$

### The GIW Prior

The GIW prior distribution for $\Sigma$ in (10.1) is equivalently defined in terms of $(\Gamma^{[u]}, \tau^{[u]})$ and $\{(\Gamma_1, \tau_1), \ldots, (\Gamma_{k-1}, \tau_{k-1}), \Sigma_k\}$ as follows.

$$\begin{cases} \tau^{[u]} \mid \Gamma^{[u]} \sim N\left(\tau_{00}, H_0 \otimes \Gamma^{[u]}\right) \\ \Gamma^{[u]} \sim IW(\Lambda_0 \otimes \Omega, \delta_0) \\ \tau_j \mid \Gamma_j \sim N(\tau_{0j}, H_j \otimes \Gamma_j),\ j = 1, \ldots, k-1 \\ \Gamma_j \sim IW(\Lambda_j \otimes \Omega, \delta_j),\ j = 1, \ldots, k, \end{cases} \tag{10.2}$$

where IW denotes the inverted Wishart distribution.

In this model $\tau^{[u]}$ is the slope of the optimal linear predictor of $Y^{[u]}$ based on $Y^{[g]}$ and $\Gamma^{[u]}$, the residual covariance of the optimal linear predictor. Similar interpretations apply to $\tau_j$ and $\Gamma_j$, for $j = 1, \ldots, k-1$.

Let $\mathcal{H}$ be the set of the hyperparameters in (10.1)–(10.2); i.e., $\mathcal{H} = \{\Theta, \delta, F, \beta_0\}$ where $\Theta$ labels the set of hyperparameters:

$$\Theta = \{(\tau_{00}, H_0, \Lambda_0), \Omega, (\tau_{01}, H_1, \Lambda_1), \ldots, (\tau_{0,k-1}, H_{k-1}, \Lambda_{k-1}), \Lambda_k\}$$

with degrees of freedom parameters $\delta = (\delta_0, \delta_1, \ldots, \delta_k)$. The dimensions of $H_j$ and $\Lambda_j$ are $(g_{j+1} + \cdots + g_k)p \times (g_{j+1} + \cdots + g_k)p$ and $g_j \times g_j$, respectively. $\Omega$

represents the hyperscale matrix between responses assumed to be common across all stations. The spatial hyperscale matrix itself is represented by the set of $\{\Lambda\}$s, for $j = 1, \ldots, k$, the spatial hyperscale matrix being readily reconstituted through the inverse of the Bartlett transformation as described in Appendix 15.2.

Note that through this specification, the Kronecker structure imposed in the prior model implies that the covariance field can be considered as two separable components: $\Omega$ denotes the hyperscale matrix between responses assumed to be common across all sites; $\Lambda_j$ and $\Lambda_0$ represent the conditional spatial component covariance between the sites within the blocks. The formulation thus reduces the number of parameters in the model and greatly simplifies their estimation. Furthermore it allows for the separation of the covariance field's spatial component and hence facilitates use of the nonparametric spatial covariance interpolator (Sampson and Guttorp 1992) to estimate the spatial hyperscale matrix among all the gauged and ungauged sites. Section 10.6 gives the details.

## 10.3 Predictive Distributions

Let $D$ and $Y_{unob}$ denote the data set and unobserved responses, respectively. That is,

$$D = \left\{ Y^{[g_1^o]}, \ldots, Y^{[g_k^o]} \right\}$$
$$Y_{unob} = \left\{ Y^{[u]}, Y^{[g_1^m]}, \ldots, Y^{[g_k^m]} \right\}.$$

Denote the multivariate responses at stations from the $j$th to $k$th blocks and the corresponding coefficient matrix as

$$Y^{[g_j, \ldots, g_k]} = \left[ \begin{pmatrix} Y^{[g_j^m]} \\ Y^{[g_j^o]} \end{pmatrix}, \ldots, \begin{pmatrix} Y^{[g_k^m]} \\ Y^{[g_k^o]} \end{pmatrix} \right]$$

and

$$\boldsymbol{\beta}^{[g_j, \ldots, g_k]} = (\boldsymbol{\beta}^{[g_j]}, \ldots, \boldsymbol{\beta}^{[g_k]}).$$

Represent the residuals between the responses and the prior means for the $(j+1)$th to $k$th blocks as

$$\tilde{\epsilon}^{[g_{j+1}, \ldots, g_k]} = \begin{cases} Y^{[g_{j+1}, \ldots, g_k]} - Z\boldsymbol{\beta}_0^{[g_{j+1}, \ldots, g_k]}, \text{ for } j = 1, \ldots, k-1, \\ 0, \text{ for } j = k. \end{cases}$$

For the gauged stations, each block has a best linear predictor based on its data and the residual covariance matrix under model (10.1). They are

$$\begin{pmatrix} \mu_{(1)}^{[j]} \\ \mu_{(2)}^{[j]} \end{pmatrix} : \begin{pmatrix} m_j \times g_j p \\ (n-m_j) \times g_j p \end{pmatrix} = Z\boldsymbol{\beta}_0^{[g_j]} + \tilde{\epsilon}^{[g_{j+1},\ldots,g_k]}\tau_{0j},$$

$$\begin{pmatrix} A_{11}^{[j]} & A_{12}^{[j]} \\ A_{21}^{[j]} & A_{22}^{[j]} \end{pmatrix} : \begin{pmatrix} m_j \times m_j & m_j \times (n-m_j) \\ (n-m_j) \times m_j & (n-m_j) \times (n-m_j) \end{pmatrix}$$

$$= I_n + ZF^{-1}Z' + \tilde{\epsilon}^{[g_{j+1},\ldots,g_k]} H_j (\tilde{\epsilon}^{[g_{j+1},\ldots,g_k]})'.$$

In the same way the best linear predictor for the gauged sites conditional on the complete responses at all gauged sites ( i.e., from the 1st to $k$th blocks) and residual covariance matrix, respectively, can be expressed as

$$\mu^{[u|g]} = Z\boldsymbol{\beta}_0^{[u]} + (Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]})\tau_{00},$$

$$\boldsymbol{\Phi}^{[u|g]} = I_n + ZF^{-1}Z' + (Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]}) H_0 (Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]})'.$$

### Predicting Unobserved Responses

**Theorem 10.1.** *Under Model (10.1) the predictive distribution of the unobserved responses conditional on the data $D$ and the hyperparameter set $\mathcal{H}$ is given by*

$$(Y_{unob} \mid D, \mathcal{H}) \sim \left(Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \prod_{j=1}^{k-1} \left(Y^{[g_j^m]} \mid Y^{[g_{j+1}^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \times \left(Y^{[g_k^m]} \mid D, \mathcal{H}\right), \tag{10.3}$$

*where*

$$\left(Y^{[g_k^m]} \mid D, \mathcal{H}\right) \sim t_{m_k \times g_k p}\left(\mu_{(u|g)}^{[k]}, \boldsymbol{\Phi}_{(u|g)}^{[k]} \otimes \boldsymbol{\Psi}_{(u|g)}^{[k]}, \delta_{(u|g)}^{[k]}\right) \tag{10.4}$$

$$\left(Y^{[g_j^m]} \mid Y^{[g_{j+1}^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \sim t_{m_j \times g_j p}\left(\mu_{(u|g)}^{[j]}, \boldsymbol{\Phi}_{(u|g)}^{[j]} \otimes \boldsymbol{\Psi}_{(u|g)}^{[j]} \delta_{(u|g)}^{[j]}\right) \tag{10.5}$$

$$\left(Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \sim t_{n \times up}\left(\mu^{[u|g]}, \frac{\boldsymbol{\Phi}^{[u|g]} \otimes \Lambda_0 \otimes \Omega}{\delta_0^*}, \delta_0^*\right). \tag{10.6}$$

*Here $t_{a\times b}$ denotes a matric-t distribution as described in Appendix 15.1 and for $j = 1, \ldots, k$,*

$$\mu^{[j]}_{(u|g)} = \mu^{[j]}_{(1)} + A^{[j]}_{12}(A^{[j]}_{22})^{-1}(Y^{[g^o_j]} - \mu^{[j]}_{(2)}),$$

$$\delta^*_0 = \delta_0 - up + 1$$

$$\delta^{[j]}_{(u|g)} = \delta_j - g_j p + n - m_j + 1,$$

$$\Phi^{[j]}_{(u|g)} = \frac{\delta_j - g_j p + 1}{\delta_j - g_j p + n - m_j + 1}\left[A^{[j]}_{11} - A^{[j]}_{12}(A^{[j]}_{22})^{-1}A^{[j]}_{21}\right],$$

$$\Psi^{[j]}_{(u|g)} = \frac{1}{\delta_j - g_j p + 1}\left[\Lambda_j \otimes \Omega + (Y^{[g^o_j]} - \mu^{[j]}_{(2)})'(A^{[j]}_{22})^{-1}(Y^{[g^o_j]} - \mu^{[j]}_{(2)})\right].$$

Proof: Given in Appendix 15.4.

To interpret this predictive distribution's parameters, notice that $(Y^{[g^o_j]} - \mu^{[j]}_{(2)})$ is the residual vector obtained from using $\mu^{[j]}_{(2)}$ to predict $Y^{[g^o_j]}$, the observed responses at the gauged stations in the $j$th block. Then $\mu^{[j]}_{(u|g)}$ represents the best predictor for the unobserved responses at gauged stations in the $j$th block based on the observed responses at the same stations as well as those in the $(j+1)$th to $k$th blocks. The $(n - m_j)$ component in the degrees of freedom $\delta^{[j]}_{(u|g)}$ reflects the contribution of the observed responses $Y^{[g^o_j]}$ in the prediction.

The contributions of the data are also reflected in the predictive covariance structure expressed through $\Phi^{[j]}_{(u|g)}$ and $\Psi^{[j]}_{(u|g)}$. Here the measurements from the gauged stations in the $j$th block contribute through $\Psi^{[j]}_{(u|g)}$ while the responses at other stations from the $(j+1)$th to $k$th blocks adjust through $\Phi^{[j]}_{(u|g)}$.

**Remarks**

- When a single response is of interest, it might seem worthwhile to use a spatial predictor based on univariate theory. That turns out to be naive. In fact, much predictive strength can be borrowed from the remaining responses through their correlations with the one of interest. In fact Sun et al. (1998) compare the accuracy of the purely univariate approach against the marginalized multivariate approach and find substantial improvements are possible in their application. For the four air pollutants in that application they found the following mean-squared error values in a cross-validatory assessment whose results are shown in Table 10.3. The improvement for $SO_4$ in particular, is dramatic.
- Brown et al. (1994a) derive the predictive distribution for multivariate responses using the Gaussian-inverted Wishart distribution. Their derivation assumes that the data at all stations are complete (i.e., multivariate response but no staircase pattern). That situation is a special case of this

| Pollutant | log $NO_2$ | log $SO_4$ | log $O_3$ | log $SO_2$ |
|---|---|---|---|---|
| Multivariate*100 | 19 | 14 | 5 | 62 |
| Univariate*100 | 28 | 127 | 13 | 76 |

Table 10.3: Mean-squared prediction error for a univariate and marginalized multivariate Bayesian posterior spatial predictor. The units are $100 \times \log^2 \mu g\ m^-3$.

predictive distribution presented here with $m_j = m$ and all $\delta_j = \delta$, an unknown constant.

- When the mean of $Y$ is assumed to be zero, the above predictive distribution reduces to that of Kibria et al. (2002).

## 10.4 Posterior Distributions

In some applications, the model parameters themselves will be of inferential interest. For example, the model transfer coefficients in $\boldsymbol{\beta}$ give insight into the role of the covariates in shaping the joint response surface. The spatial covariance may reveal the influence of latent factors such as wind speed or direction. Thus the posterior distributions for $\boldsymbol{\beta}$ and $\Sigma$ are of interest and given here.

Since observation numbers may differ from block to block, the posterior distributions need to reflect such unbalanced data appropriately. As a direct result, $\boldsymbol{\beta}$'s posterior distribution is a product of distributions; each corresponds to the regression coefficients in individual blocks. Similarly $\Sigma$'s posterior distribution is a product of distributions conveniently presented through the recursive Bartlett transformation.

First some notation. Let

$$K_j : (n - m_j) \times n = (0, I_{n-m_j}),$$

$$Z_{(j)} = K_j Z,$$

$$Y_{(j)}^{[g_{j+1},\ldots,g_k]} = K_j Y^{[g_{j+1},\ldots,g_k]}.$$

Here $Y_{(j)}^{[g_{j+1},\ldots,g_k]}$ denotes the matrix of the last $(n - m_j)$ responses from all stations in blocks $(j+1)$ to $k$ and $Z_{(j)}$, the covariates from time $m_j + 1$ to $n$.

After removing their prior means, let the observation residuals from times $(m_j + 1)$ to $n$ for all stations in blocks $(j + 1)$ to $k$ be

$$\tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]} = Y_{(j)}^{[g_{j+1},\ldots,g_k]} - Z_{(j)}\boldsymbol{\beta}_0^{[g_{j+1},\ldots,g_k]};$$

and after subtracting their prior means, let the residuals of the $(n - m_j)$ observations from stations in the $j$th block

$$\tilde{\epsilon}^{[g_j^o]} = Y^{[g_j^o]} - Z_{(j)}\boldsymbol{\beta}_0^{[g_j]}.$$

Let

$$\hat{\boldsymbol{\beta}}^{[g_j]} = (Z_{(j)}^T Z_{(j)})^{-1} Z_{(j)}^T Y^{[g_j^o]},$$

$$\hat{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]} = (Z_{(j)} Z_{(j)}^T)^{-1} Z_{(j)}^T Y_{(j)}^{[g_{j+1},\ldots,g_k]}.$$

Here $\hat{\boldsymbol{\beta}}^{[g_j]}$ and $\hat{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]}$ are the usual maximum likelihood estimates for $\boldsymbol{\beta}^{[g_j]}$ and $\boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}$, respectively, given the observations in the $j$th block and the last $(n - m_j)$ observations from stations in blocks $j+1$ to $k$.

The corresponding best linear estimates are denoted by

$$\tilde{F}_j = Z_{(j)} Z_{(j)}^T + F,$$

$$W_j = \tilde{F}_j^{-1} Z_{(j)}^T Z_{(j)}$$

$$\tilde{\boldsymbol{\beta}}^{[g_j]} = W_j \hat{\boldsymbol{\beta}}^{[g_j]} + (I - W_j)\boldsymbol{\beta}_0^{[g_j]},$$

$$\tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]} = W_j \hat{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]} + (I - W_j)\boldsymbol{\beta}_0^{[g_{j+1},\ldots,g_k]}.$$

**The Posterior**

**Theorem 10.2.** *Under the model (10.1), the joint posterior density for $\boldsymbol{\beta}$ and $\Sigma$ given the data $D$ and the hyperparameters $\mathcal{H}$ can be presented as*

$$f(\boldsymbol{\beta}, \Sigma \mid D, \mathcal{H}) = f(\boldsymbol{\beta} \mid \Sigma, D, \mathcal{H}) f(\Sigma \mid D, \mathcal{H}),$$

*where*

$$(i) \quad f(\boldsymbol{\beta} \mid \Sigma, D, \mathcal{H}) = f(\boldsymbol{\beta}^{[u]} \mid D, \Gamma^{[u]}, \tau^{[u]}, \mathcal{H}) f(\boldsymbol{\beta}^{[g_k]} \mid D, \Gamma_k, \mathcal{H}) \times \prod_{j=0}^{k-1} f(\boldsymbol{\beta}^{[g_j]} \mid D, \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}, \tau_j, \Gamma_j, \mathcal{H}) \tag{10.7}$$

*with*

$$\boldsymbol{\beta}^{[g_k]} \mid D, \Sigma_{kk}, \mathcal{H} \sim N_{l \times g_k p}\left(\tilde{\boldsymbol{\beta}}^{[g_k]}, \tilde{F}_k^{-1} \otimes \Gamma_k\right),$$

$$\boldsymbol{\beta}^{[g_j]} \mid D, \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}, \tau_j, \Gamma_j, \mathcal{H} \sim N_{l \times g_j p}\left(\tilde{\boldsymbol{\beta}}^{[g_j]} + (\boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]} - \tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]})\tau_j, \tilde{F}_j^{-1} \otimes \Gamma_j\right), \tag{10.8}$$

*and*

$$\boldsymbol{\beta}^{[u]} \mid D, \boldsymbol{\beta}^{[g_1,\ldots,g_k]}, \Gamma^{[u]}, \tau^{[u]}, \mathcal{H}$$
$$\sim N_{l\times up}\left(\boldsymbol{\beta}_0^{[u]} + (\boldsymbol{\beta}^{[g_1,\ldots,g_k]} - \tilde{\boldsymbol{\beta}}^{[g_1,\ldots,g_k]})\tau^{[u]}, F_j^{-1} \otimes \Gamma^{[u]}\right),$$

$$(ii) \quad f(\Sigma \mid D, \mathcal{H}) = f(\tau^{[u]} \mid D, \Gamma^{[u]}, \mathcal{H}) f(\Gamma^{[u]} \mid D, \mathcal{H}) \prod_{j=1}^{k-1} f(\tau_j \mid D, \Gamma_j, \mathcal{H})$$
$$\times \prod_{j=1}^{k} f(\Gamma_j \mid D, \mathcal{H}) \tag{10.9}$$

*with*

$$\tau^{[u]} \mid \Gamma^{[u]}, D, \mathcal{H} \sim N\left(\tau_{00}, H_0 \otimes \Gamma^{[u]}\right)$$
$$\Gamma^{[u]} \mid D, \mathcal{H} \sim IW(\Lambda_0 \otimes \Omega, \delta_0)$$
$$\tau_j \mid D, \Gamma_j, \mathcal{H} \sim N\left(\tilde{\tau}_{0j}, \tilde{H}_j \otimes \Gamma_j\right)$$
$$\Gamma_j \mid D, \mathcal{H} \sim IW(\tilde{\Psi}_j, \tilde{\delta}_j).$$

*Here*

$$\tilde{\Psi}_k = \Lambda_k \otimes \Omega + (\tilde{\epsilon}^{[g_k^o]})^T [I_{n-m_k} + Z_{(k)} F^{-1} Z_{(k)}^T]^{-1} \tilde{\epsilon}^{[g_k^o]}$$

*and for* $j = 1, \ldots, k-1$

$$\tilde{\Psi}_j = \Lambda_j \otimes \Omega + (\tilde{\epsilon}^{[g_j^o]} - \tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]} \tau_{0j})^T \Big[ I_{n-m_j} + Z_{(j)} F^{-1} Z_{(j)}^T$$
$$+ (\tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]})^T H_j \tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]} \Big]^{-1} (\tilde{\epsilon}^{[g_j^o]} - \tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]} \tau_{0j}),$$
$$\tilde{H}_j^{-1} = H_j^{-1} + (\tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]})^T [I_{n-m_j} + Z_{(j)} F^{-1} Z_{(j)}^T]^{-1} \tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]},$$

$$\tilde{\tau}_{0j} = \tilde{H}_j \left[ H_j^{-1} \tau_{0j} + (\tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]})^T [I_{n-m_j} + Z_{(j)} F^{-1} Z_{(j)}^T]^{-1} \tilde{\epsilon}^{[g_j^o]} \right],$$
$$\tilde{\delta}_j = \delta_j + n - m_j.$$

Proof: Given in Appendix 15.4.

Note that the posterior distributions corresponding to the ungauged sites $\tau^{[u]}$ and $\Gamma^{[u]}$ remain the same as the prior distributions. This is a direct result of the selected parameterization of the parameter space.

## 10.5 Posterior Expectations

Relevant posterior means can be derived from the above posterior distributions, (10.7)–(10.9). Some posterior moments are derived below, conditional on data $D$ and hyperparameter set $\mathcal{H}$. The notation in previous sections is used here (see also Le et al. 2001).

- **The posterior mean of $\boldsymbol{\beta}$.** The usual conditional argument leads to that mean:

$$E(\boldsymbol{\beta}^{[g_k]} \mid D, \mathcal{H}) = \tilde{\boldsymbol{\beta}}^{[g_k]},$$

$$E(\boldsymbol{\beta}^{[g_j]} \mid D, \mathcal{H}) = \tilde{\boldsymbol{\beta}}^{[g_j]} + \left[E(\boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]} \mid D, \mathcal{H}) - \tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]}\right] \tilde{\tau}_{0j},$$

where

$$E(\boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]} \mid D, \mathcal{H}) = [E(\boldsymbol{\beta}^{[g_{j+1}]} \mid D, \mathcal{H}), \ldots, E(\boldsymbol{\beta}^{[g_k]} \mid D, \mathcal{H})]$$

is computed recursively for $j = 0, \ldots, k-1$.
Notice that $E(\boldsymbol{\beta}^{[g_j]} \mid D, \mathcal{H}) = \tilde{\boldsymbol{\beta}}^{g_j}$ for $j = 1, \ldots, k-1$, when the data are complete, i.e., when $D$ contains no missing blocks.

- **The posterior mean of $\Sigma^{-1}$** is obtained recursively as follows.

$$E\left[\Sigma^{-1} \mid D, \mathcal{H}\right] =$$

$$\begin{pmatrix} \delta_0(\Lambda_0 \otimes \Omega)^{-1} & -\delta_0(\Lambda_0 \otimes \Omega)^{-1}\tau_{00}^T \\ -\delta_0\tau_{00}(\Lambda_0 \otimes \Omega)^{-1} & \delta_0\tau_{00}(\Lambda_0 \otimes \Omega)^{-1}\tau_{00}^T \\ & \quad + upH_0 + E\left[\Sigma_{(11)}^{-1} \mid D, \mathcal{H}\right] \end{pmatrix},$$

where for $j = 1, \ldots, k-1$

$$E\left[\Sigma_{(jj)}^{-1} \mid D, \mathcal{H}\right] =$$

$$\begin{pmatrix} \tilde{\delta}_j\tilde{\Psi}_j^{-1} & -\tilde{\delta}_j\tilde{\Psi}_j^{-1}\tilde{\tau}_{0j}^T \\ -\tilde{\delta}_j\tilde{\tau}_{0j}\tilde{\Psi}_j^{-1} & \tilde{\delta}_j\tilde{\tau}_{0j}\tilde{\Psi}_j^{-1}\tilde{\tau}_{0j}^T + g_jp\tilde{H}_j + E\left[\Sigma_{(j+1,j+1)}^{-1} \mid D, \mathcal{H}\right] \end{pmatrix},$$

and

$$E\left[\Sigma_{(kk)}^{-1} \mid D, \mathcal{H}\right] \equiv E\left[\Gamma_k^{-1}\right] = \tilde{\delta}_k\tilde{\Psi}_k^{-1}.$$

To obtain these results notice that

$$\Sigma_{(jj)}^{-1} = \begin{pmatrix} \Gamma_j^{-1} & -\Gamma_j^{-1}\tau_j^T \\ -\tau_j\Gamma_j^{-1} & \tau_j\Gamma_j^{-1}\tau_j^T + \Sigma_{(j+1,j+1)}^{-1} \end{pmatrix}, \text{ for } j = 1, \ldots, k-1,$$

and then get the corresponding expectations from the posterior distributions.

- **The posterior expectation of** $\mathbf{E}\{\log|\Gamma_\mathbf{j}| \mid \mathbf{D}, \mathcal{H}\}$ for $j = 1, \ldots, k$, depends on the digamma function (the derivative of the Gamma function) denoted by $\psi$. Then

$$E\{\log|\Gamma_j| \mid D, \mathcal{H}\} = -g_j p \; log \; 2 - \sum_{i=1}^{g_j p} \psi\left(\frac{\tilde{\delta}_j - i + 1}{2}\right) + \log|\tilde{\Psi}_j|.$$

These results are direct applications of those in Chen (1979).

- **The posterior expectation** $\mathbf{E}\left\{\boldsymbol{\beta}\Sigma^{-\mathbf{1}} \mid \mathbf{D}, \mathcal{H}\right\}$ corresponding to the gauged sites, is obtained by recursion as follows.

$$E\left\{\boldsymbol{\beta}\Sigma^{-1} \mid D, \mathcal{H}\right\} = E\left\{\boldsymbol{\beta}^{[g_1,\ldots,g_k]}\Sigma^{-1}_{(11)} \mid D, \mathcal{H}\right\},$$

$$E\left\{\boldsymbol{\beta}^{[g_j,\ldots,g_k]}\Sigma^{-1}_{(jj)} \mid D, \mathcal{H}\right\} = \left(\tilde{\delta}_j(\tilde{\boldsymbol{\beta}}^{[g_j]} - \tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]}\tilde{\tau}_{0j})\tilde{\Psi}_j^{-1},\right.$$

$$E\left\{\boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}\Sigma^{-1}_{(j+1,j+1)} \mid D, \mathcal{H}\right\} - \tilde{\delta}_j(\tilde{\boldsymbol{\beta}}^{[g_j]} - \tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]}\tilde{\tau}_{0j})\tilde{\Psi}_j^{-1}\tilde{\tau}_{0j}^T$$

$$\left. + \; g_j p\tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]}\tilde{H}_j\right),$$

and finally

$$E\left\{\boldsymbol{\beta}^{[g_k]}\Sigma_{kk}^{-1} \mid D, \mathcal{H}\right\} \equiv E\left\{\boldsymbol{\beta}^{[g_k]}\Gamma_k^{-1} \mid D, \mathcal{H}\right\} = \tilde{\delta}_k\tilde{\boldsymbol{\beta}}^{[g_k]}\tilde{\Psi}_k^{-1}.$$

These results are obtained by writing

$$\boldsymbol{\beta}^{[g_j,\ldots,g_k]}\Sigma^{-1}_{(jj)} = \left((\boldsymbol{\beta}^{[g_j]} - \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}\tau_j)\Gamma_j^{-1}, \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}\Sigma^{-1}_{(j+1,j+1)}\right.$$
$$\left. -(\boldsymbol{\beta}^{[g_j]} - \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}\tau_j)\Gamma_j^{-1}\tau_j^T\right),$$

and then taking expectations recursively.

- **The posterior expectation** $\boldsymbol{\beta}\Sigma^{-\mathbf{1}}\boldsymbol{\beta}^{\mathbf{T}}$ corresponding to the gauged sites, is as follows:

$$E\left\{\boldsymbol{\beta}\Sigma^{-1}\boldsymbol{\beta}^T \mid D, \mathcal{H}\right\} = \tilde{\delta}_k\tilde{\boldsymbol{\beta}}^{[g_k]}\tilde{\Psi}_k^{-1}\tilde{\boldsymbol{\beta}}^{[g_k]^T}$$

$$+ \textstyle\sum_{j=1}^{k-1}\tilde{\delta}_j\left[\tilde{\boldsymbol{\beta}}^{[g_j]} - \tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]}\tilde{\tau}_{0j}\right]\tilde{\Psi}_j^{-1}\left[\tilde{\boldsymbol{\beta}}^{[g_j]} - \tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]}\tilde{\tau}_{0j}\right]^T$$

$$+ \textstyle\sum_{j=1}^{k-1} g_j p\tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]}\tilde{H}_j\tilde{\boldsymbol{\beta}}^{[g_{j+1},\ldots,g_k]^T} + \sum_{j=1}^{k} g_j p\tilde{F}_j^{-1}.$$

To obtain this result observe that

$$\boldsymbol{\beta}\Sigma^{-1}\boldsymbol{\beta}^T = \boldsymbol{\beta}^{[g_k]}\Sigma_{kk}^{-1}\boldsymbol{\beta}^{[g_k]^T} + \sum_{j=1}^{k-1}(\boldsymbol{\beta}^{[g_j]} - \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}\tau_j)\Gamma_j^{-1}(\boldsymbol{\beta}^{[g_j]} - \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}, \tau_j)^T.$$

and then take expectation on both sides of the equation.

## 10.6 Hyperparameter Estimation

The predictive distributions derived through the integrated framework developed above are completely characterized by their hyperparameters. However, these hyperparameters are themselves uncertain, calling in principle for an additional layer of prior modeling.

However, as noted in the introduction to Chapter 9, they may instead be estimated. In fact advantages accrue in this context from using an empirical Bayes approach. Here this means estimating them by maximizing the marginal likelihood, i.e., the marginal joint density function of all the measured responses (conditional on those hyperparameters) evaluated at their observed values. This procedure is referred to as type-II maximum likelihood estimation (type-II MLE). Besides simplicity, this approach offers the important advantage of helping ensure the predictive distributions are well calibrated: 95% prediction intervals will cover unmeasured responses about that percentage of the time.

However, in the formulation of the prediction problem we have imposed no restriction on the forms of $\Sigma$ and its hyperparameters. Thus the required joint marginal depends only on the gauged site hyperparameters associated with the monitoring stations. They can be estimated by type-II MLE in the first of a two step procedure. The remainder associated with the ungauged sites are estimated in the second step by the nonparametric approach in Sampson and Guttorp (1992) (see Chapter 6). Several papers have used this two-step procedure (Brown et al. 1994a; Sun et al. 1998; Sun 1998; Le et al. 2001; Kibria et al. 2002).

Yet even that approach proves challenging due to the large number of hyperparameters involved, especially in the covariance model. An additional difficulty arises because our responses here are multivariate. Finally those difficulties are compounded by the lack of assumed stationarity or isotropy for the random field.

### 10.6.1 Two-Step Estimation Procedure

The first step of the Type-II maximum likelihood computes the hyperparameter values that maximize the marginal distribution $f(D \mid \mathcal{H}_g)$ where

$$D = \left\{Y^{[g_1^o]}, \ldots, Y^{[g_k^o]}\right\} \tag{10.10}$$

denotes the data and

$$\mathcal{H}_g = \{F, \boldsymbol{\beta}_0, \Omega, (\tau_{01}, H_1, \Lambda_1, \delta_1), \ldots, (\tau_{0,k-1}, H_{k-1}, \Lambda_{k-1}, \delta_{k-1}), (\Lambda_k, \delta_k)\}, \tag{10.11}$$

are the hyperparameters of interest. The subscript $g$ indicates that not all the hyperparameters are involved in this marginal distribution. Although $f(D \mid \mathcal{H}_g)$ can be written as a product of matric-$t$ distributions as in (10.35), direct maximization of this marginal density presents a challenge. Using the EM algorithm helps circumvent it. Recall that while $\Omega$, reflects the covariance within sites (for example, between responses), the $\Lambda$s represent the residual spatial covariance (i.e., between sites).

The second step yields estimates of the remaining hyperparameters $\Lambda_0$, $\tau_{00}$, $H_0$, and $\delta_0$ representing the spatial covariance between the ungauged sites. Since the spatial dependence structure can be nonstationary, these remaining hyperparameters are related to those corresponding to the gauged sites estimated in the first step. This step uses the SG method to extend the spatial covariance estimates when stationarity is not assumed. First compute the spatial covariance matrix between gauged sites from the estimated residual spatial covariances (i.e., the estimated $\Lambda_k$) through the Bartlett transformation (Appendix 15.2). Next extend the spatial covariance matrix to ungauged sites using the SG method to obtain estimates for the covariance matrix between the ungauged sites and the cross-covariance. Finally estimate $\Lambda_0$, $\tau_{00}$, and $H_0$ to complete the estimation process.

### 10.6.2 Spatial Covariance Separability

By design the SG method applies to spatial covariance matrices between the gauged sites, not between responses. That needs an assumption that the overall covariance separates into spatial and within-site components. That need is partially met by the choice of the prior model for the $\Gamma_k$; it implies the spatial hypercovariance matrix $\Lambda_k$ between stations in the $k$th block is separable from the covariance between responses. However, the prior models for $\Gamma^{[u]}, \Gamma_1, \ldots, \Gamma_{k-1}$ specify separation only in terms of the residual hypercovariance matrices $\Lambda_j$ for $j = 1, \ldots, k-1$ and $\Lambda^{[u]}$, not in the unconditional ones. Thus other conditions are needed to ensure the separability of the covariances between and within sites. To find them below we start with the unconditional covariance matrix for the multivariate responses and derive the required conditions.

**Covariance Matrix**

Let $Y_t$ be the response vector at time $t$ and $z_t$ a vector of covariates. Assume $Y_t$ has the Gaussian-generalized-inverted-Wishart model specified by (10.1) and (10.2). The variance–covariance matrix of $Y_t$ can be written as follows:

$$V(Y_t) = E[V(Y_t \mid \boldsymbol{\beta})] + V(E[Y_t \mid \boldsymbol{\beta}])$$

$$= E(\Sigma) + E[V(z_t\boldsymbol{\beta} \mid \Sigma)] + V(E[z_t\boldsymbol{\beta} \mid \Sigma])$$

$$= (1 + z_t F^{-1} z_t')E(\Sigma). \tag{10.12}$$

Note that $\Sigma$ can be expressed in terms of the hyperparameters through the Bartlett decomposition as $(\Gamma^{[u]}, \tau^{[u]})$ and $\{(\Gamma_1, \tau_1), \ldots, (\Gamma_{k-1}, \tau_{k-1}), \Sigma_k\}$ as follows.

$$\Sigma = \begin{pmatrix} \Gamma^{[u]} + (\tau^{[u]})^T \Sigma^{[g_1,\cdots,g_k]} \tau^{[u]} & (\tau^{[u]})^T \Sigma^{[g_1,\cdots,g_k]} \\ \Sigma^{[g_1,\cdots,g_k]} \tau^{[u]} & \Sigma^{[g_1,\cdots,g_k]} \end{pmatrix}, \tag{10.13}$$

where $\Sigma^{[g_1,\cdots,g_k]}$ is recursively defined as

$$\Sigma^{[g_j,\cdots,g_k]} = \begin{pmatrix} \Gamma_j + \tau_j^T \Sigma^{[g_{j+1},\cdots,g_k]} \tau_j & \tau_j^T \Sigma^{[g_{j+1},\cdots,g_k]} \\ \Sigma^{[g_{j+1},\cdots,g_k]} \tau_j & \Sigma^{[g_{j+1},\cdots,g_k]} \end{pmatrix}, \tag{10.14}$$

and

$$\Sigma^{[g_k,g_k]} = \Gamma_k.$$

Details can be found in Appendix 15.4.

Hence, $E(\Sigma)$ can be obtained (componentwise) using the prior distribution (10.2) and through the relationships given in (10.13) and (10.14). That is,

$$E[\Gamma^{[u]}] = \frac{1}{\delta_0 - up - 1} \Lambda_0 \otimes \Omega \tag{10.15}$$

$$E[(\tau^{[u]})^T \Sigma^{[g_1,\ldots,g_k]}] = \tau_{00}^T E[\Sigma^{[g_1,\ldots,g_k]}] = \tau_{00}^T \eta^{[1,\ldots,k]} \tag{10.16}$$

$$\begin{aligned} E[(\tau^{[u]})^T \Sigma^{[g_1,\ldots,g_k]} (\tau^{[u]})] &= E[E(\tau^{[u]})^T \Sigma^{[g_1,\ldots,g_k]} (\tau^{[u]}) \mid \tau^{[u]}] \\ &= E[(\tau^{[u]})^T (E\Sigma^{[g_1,\ldots,g_k]})(\tau^{[u]})] \\ &= E[(\tau^{[u]})^T \eta^{[1,\ldots,k]} (\tau^{[u]})] \\ &= \tau_{00}^T \eta^{[1,\ldots,k]} \tau_{00} + \frac{tr(\eta^{[1,\ldots,k]} H_0)}{\delta_0 - up - 1} \Lambda_0 \otimes \Omega. \end{aligned} \tag{10.17}$$

Similarly the remaining components can be obtained by recursion. More precisely let

$$\eta^{[j,\ldots,k]} = E[\Sigma^{[g_j,\ldots,g_k]}] \tag{10.18}$$

represent the covariance matrix corresponding to the response from blocks $j$ to $k$. Thus for $j = 1, \ldots, k-1$, $\eta^{[j,\ldots,k]}$ can be recursively evaluated as

$$\eta^{[j,\ldots,k]} = \begin{pmatrix} a_j(\Lambda_j \otimes \Omega) + \tau_{0j}^T \eta^{[j+1,\ldots,k]} \tau_{0j} & \tau_{0j}^T \eta^{[j+1,\ldots,k]} \\ \eta^{[j+1,\ldots,k]} \tau_{0j} & \eta^{[j+1,\ldots,k]} \end{pmatrix} \tag{10.19}$$

and

$$\eta^{[k]} = \Lambda_k \otimes \Omega/(\delta_k - g_k p - 1) \tag{10.20}$$

where $a_j = (1 + tr(\eta^{[j+1,\ldots,k]} H_j))/(\delta_j - g_j p - 1)$.

Substituting (10.15) to (10.20) in (10.12) yields the covariance matrix

$$V(Y_t) = (1+z_t F^{-1} z_t^T) \begin{pmatrix} a_0(\Lambda_0 \otimes \Omega) + \tau_{00}^T \eta^{[1,\ldots,k]} \tau_{00} & \tau_{00}^T \eta^{[1,\ldots,k]} \\ \eta^{[1,\ldots,k]} \tau_{00} & \eta^{[1,\ldots,k]}, \end{pmatrix} \tag{10.21}$$

where $a_0 = (1 + tr(\eta^{[1,\ldots,k]} H_0))/(\delta_0 - up - 1)$.

**Separability Conditions**

The spatial covariance $V(Y_t)$ can be separated into two components $\Psi$ and $\Omega$ and expressed as $\Psi \otimes \Omega$, if for $j = 0, \ldots, k-1$,

$$\tau_{0j} = \xi_{0j} \otimes I_p, \tag{10.22}$$

where $\Psi$ denotes the spatial covariance component.

The separability conditions [Equation (10.22)] can be derived as follows. First substitute the separability condition for $\tau_{0j}$ [Equation (10.22)] into Equation (10.19) to get

$$\eta^{[j,\ldots,k]} = \Psi^{[j,\ldots,k]} \otimes \Omega, \ j = k-1, \ldots, 1, \tag{10.23}$$

where

$$\Psi^{[j,\ldots,k]} = \begin{pmatrix} a_j \Lambda_j + \xi_{0j}^T \Psi^{[j+1,\ldots,k]} \xi_{0j} & \xi_{0j}^T \Psi^{[j+1,\ldots,k]} \\ \Psi^{[j+1,\ldots,k]} \xi_{0j} & \Psi^{[j+1,\ldots,k]} \end{pmatrix} \tag{10.24}$$

and

$$\Psi^{[k]} = \Lambda_k/(\delta_k - g_k p - 1). \tag{10.25}$$

Then substitute the separability condition for $\tau_{00}$ in (10.22) into the covariance matrix (10.21) to obtain the desired separability conclusion

$$V(Y_t) = (1 + z_t F^{-1} z_t^T) \Psi \otimes \Omega,$$

where

$$\Psi = \begin{pmatrix} a_0 \Lambda_0 + \xi_{00}^T \Psi^{[1,\ldots,k]} \xi_{00} & \xi_{00}^T \Psi^{[1,\ldots,k]} \\ \Psi^{[1,\ldots,k]} \xi_{00} & \Psi^{[1,\ldots,k]} \end{pmatrix}. \tag{10.26}$$

The above separability conditions are used in the estimation of hyperparameters, those for gauged sites described next.

### 10.6.3 Estimating Gauged Site Hyperparameters

The EM algorithm (Dempster et al. 1977; Chen 1979) facilitates the computation of the type-II maximum likelihood estimates for $\mathcal{H}_g$. Le and Zidek (1994) and Brown et al. (1994a) first employed this approach for application to environmental problems.

First, additional assumptions are made to further simplify the estimation problem. Specifically, to reduce the number of hyperparameters to be estimated, assume in accordance with the inverted Wishart distribution, that

$$H_j = (\Lambda^{[j+1,\ldots,k]} \otimes \Omega)^{-1}, \tag{10.27}$$

where

$$\Lambda^{[j,\ldots,k]} = \begin{pmatrix} \Lambda_j + \xi'_{0j}\Lambda^{[j+1,\ldots,k]}\xi_{0j} & \xi'_{0j}\Lambda^{[j+1,\ldots,k]} \\ \Lambda^{[j+1,\ldots,k]}\xi_{0j} & \Lambda^{[j+1,\ldots,k]} \end{pmatrix} \tag{10.28}$$

and

$$\Lambda^{[k]} = \Lambda_k.$$

With this assumption, the EM iterative approach for estimating the hyperparameters is described next, starting with the essentials of the EM algorithm.

*EM Algorithm*

Let $U^*$,$V^*$, and $\varrho^*$ be random objects such as vectors or matrices where $U^*$ is measured to yield data $u_o$ while $V^*$ cannot be so it represents missing data. At the same time, $\varrho^*$ represents a collection of nuisance parameters endowed with randomness by the prior distribution. Finally, given another collection of parameters $\theta'$, the random objects have a joint conditional probability density function $f(u^*, v^*, \varrho^*|\theta')$ where $f$ is known. The problem: maximize $f(u_o, |\theta')$, or equivalently when $f$ is strictly positive, $\ln f(u_o, |\theta')$ with respect to $\theta'$.
The famous EM algorithm relies on a simple but ingenious trick that uses an identity that is trivial to obtain:

$$\begin{aligned} lnf(u_o, |\theta') &= [\ln f(u_o, V^*, \varrho^*|\theta')] + [-\ln f(V^*, \varrho^*|u_o, \theta')] \\ &= E_{\theta^{old}}[\ln f(u_o, V^*, \varrho^*|\theta')] + \\ &\qquad E_{\theta^{old}}[-\ln f(V^*, \varrho^*|u_o, \theta')], \end{aligned}$$

where $E_{\theta^{old}}$ means take expectations with respect to the conditional distribution obtained if $\theta'$ were set equal to $\theta^{old}$, the "expectation" step in the EM algorithm.
The trick relies on the celebrated information inequality that implies

$$E_{\theta^{old}}[-\ln f(V^*, \varrho^*|u_o, \theta')] > E_{\theta^{old}}[-\ln f(V^*, \varrho^*|u_o, \theta^{old})]$$

as long as $\theta' \neq \theta^{old}$. Thus as long as we choose $\theta^{old}$ to make

$$E_{\theta^{old}}[\ln f(u_o, V^*, \varrho^*|\theta')] > E_{\theta^{old}}[\ln f(u_o, V^*, \varrho^*|\theta^{old})]$$

the algorithm ensures $\ln f(u_o, |\theta^{new}) > \ln f(u_o, |\theta^{old})$. Of course ideally $\theta^{new}$ should be taken to maximize $E_{\theta^{old}}[-\ln f(V^*, \varrho^*|u_o, \theta')]$, the "maximization" step in the EM algorithm.
If we now replace $\theta^{old}$ by $\theta^{new}$ and continue the cycle repeatedly, we will under very general conditions converge to a point that maximizes $\ln f(u_o, |\theta')$. More details can be found in Dempster et al. (1977) and Wu (1982).

To apply the EM algorithm to our estimation problem, let $U^*$ and $V^*$ together represent the measured and unmeasured components of $Y^{[g]}$ while $\varrho^* = \{\boldsymbol{\beta}, \Sigma\}$. The algorithm would then require at iteration $p+1$, in the $E$-step, the computation of

$$\begin{aligned} Q(\mathcal{H}_g \mid \mathcal{H}_g^{(p)}) &= E\left(\log[f(Y^{[g]}, \boldsymbol{\beta}, \Sigma \mid \mathcal{H}_g) \mid D, \mathcal{H}_g^{(p)}\right) \\ &= E\left[\log f(Y^{[g]} \mid \boldsymbol{\beta}, \Sigma) \mid D, \mathcal{H}_g^{(p)}\right] + \\ &\qquad E\left[\log f(\boldsymbol{\beta}, \Sigma \mid \mathcal{H}_g) \mid D, \mathcal{H}_g^{(p)}\right] \end{aligned} \tag{10.29}$$

given the previous parameter estimates $\mathcal{H}^{(p)}$ from iteration $p$. Then it would require at the $M$-step maximization of $Q(\mathcal{H}_g \mid \mathcal{H}_g^{(p)})$ over $\mathcal{H}_g$ to get $\mathcal{H}_g^{(p+1)}$. Here, the expectation is taken over $\boldsymbol{\beta}$ and $\Sigma$ with respect to the posterior distribution $\boldsymbol{\beta}, \Sigma \mid D, \mathcal{H}_g^{(p)}$.

Notice that $E\left[\log f(Y^{[g]} \mid \boldsymbol{\beta}, \Sigma) \mid D, \mathcal{H}_g^{(p)}\right]$ does not depend on $\mathcal{H}_g$. Thus the algorithm requires only that we compute

$$Q^*(\mathcal{H}_g \mid \mathcal{H}_g^{(p)}) = E\left[\log f(\boldsymbol{\beta}, \Sigma \mid \mathcal{H}_g) \mid D, \mathcal{H}_g^{(p)}\right] \tag{10.30}$$

at the $E$-step and maximize $Q^*$ over $\mathcal{H}_g$ at the $M$-step.

With the parameterization introduced above and the prior distributions specified in (10.1)–(10.2), we have

$$\begin{aligned} f(\boldsymbol{\beta}, \Sigma \mid \mathcal{H}_g) &\propto f(\boldsymbol{\beta} \mid \Sigma, \mathcal{H}_g) \prod_{j=1}^{k-1} f(\tau_j \mid \Gamma_j, \mathcal{H}_g) \prod_{j=1}^{k} f(\Gamma_j \mid \mathcal{H}_g) \\ &\propto |F|^{gp/2} |\Sigma|^{-1/2} \text{etr}\left\{-\frac{1}{2} F(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\right\} \\ &\quad \times \prod_{j=1}^{k-1} |\Gamma_j|^{-(g_{j+1}p + \cdots + g_k p)/2} |H_j|^{g_j p/2} \\ &\quad \times \prod_{j=1}^{k-1} \text{etr}\left\{-\frac{1}{2} H_j^{-1}(\tau_j - \tau_{0j})^T \Gamma_j (\tau_j - \tau_{0j})\right\} \end{aligned}$$

$$\times \prod_{j=1}^{k} c(g_j p, \delta_j) |\Gamma_j|^{-(\delta_j + g_j p + 1)/2} |\Lambda_j \otimes \Omega|^{\delta_j/2}$$
$$\times \prod_{j=1}^{k-1} \text{etr}\left\{ -\frac{1}{2} \Gamma_j^{-1} (\Lambda_j \otimes \Omega) \right\}, \tag{10.31}$$

where

$$c(p, \delta) = \left[ 2^{\delta p/2} \pi^{p(p+1)/4} \prod_{i=1}^{p} \Gamma\left( \frac{\delta - i + 1}{2} \right) \right]^{-1}.$$

Simplifying (10.31) yields

$$\begin{aligned} f(\boldsymbol{\beta}, \Sigma \mid \mathcal{H}_g) \propto\ & |F|^{gp/2} \text{etr}\left\{ -\frac{1}{2} F(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \right\} \\ \times\ & \prod_{j=1}^{k-1} |H_j|^{g_j p/2} \text{etr}\left\{ -\frac{1}{2} H_j^{-1} (\tau_j - \tau_{0j}) \Gamma_j (\tau_j - \tau_{0j})^T \right\} \\ \times\ & c(g_k p, \delta_k) |\Gamma_k|^{-(\delta_k + g_k p + 1)/2} |\Lambda_k \otimes \Omega|^{\delta_k/2} \\ \times\ & \text{etr}\left\{ -\frac{1}{2} \Sigma_k^{-1} (\Lambda_k \otimes \Omega) \right\} \\ \times\ & \prod_{j=1}^{k-1} c(g_j p, \delta_j) |\Gamma_j|^{-(l + \delta_j + g_j p + \cdots + g_k p + 1)/2} |\Lambda_j \otimes \Omega|^{\delta_j/2} \\ \times\ & \prod_{j=1}^{k-1} \text{etr}\left\{ -\frac{1}{2} \Gamma_j^{-1} (\Lambda_j \otimes \Omega) \right\}. \end{aligned} \tag{10.32}$$

The separability condition (10.22) in conjunction with (10.32) implies $Q^*(\mathcal{H}_g \mid \mathcal{H}_g^{(p)})$ given in (10.29) can be expressed as

$$\begin{aligned} Q(\mathcal{H}_g \mid \mathcal{H}_g^{(p)}) = & \ CONST + \frac{gp}{2} \log |F| \\ & - \frac{1}{2} tr\left\{ FE[(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \mid D, \mathcal{H}_g^{(p)}] \right\} \\ & + \sum_{j=1}^{k-1} tr\left\{ -\frac{1}{2} H_j^{-1} E[(\tau_j - \xi_{0j} \otimes I_p) \Gamma_j (\tau_j - \xi_{0j} \otimes I_p)^T \mid D, \mathcal{H}_g^{(p)}] \right\} \\ & + \sum_{j=1}^{k} \log c(g_j p, \delta_j) - \frac{l + \delta_k + g_k p + 1}{2} E[\log |\Gamma_k| \mid D, \mathcal{H}_g^{(p)}] + \frac{\delta_k p}{2} \log |\Lambda_k| \\ & + \frac{\delta_k g_k}{2} \log |\Omega| - \frac{1}{2} tr\left\{ (\Lambda_k \otimes \Omega) E[\Gamma_k^{-1} \mid D, \mathcal{H}_g^{(p)}] \right\} \\ & + \sum_{j=1}^{k-1} \frac{l + \delta_j + g_j p + \cdots + g_k p + 1}{2} E[\log |\Gamma_j| \mid D, \mathcal{H}_g^{(p)}] + \sum_{j=1}^{k-1} \frac{\delta_j p}{2} \log |\Lambda_j| \end{aligned}$$

$$+ \sum_{j=1}^{k-1} \frac{\delta_j g_j}{2} \log|\Omega| - \sum_{j=1}^{k-1} \frac{1}{2} tr\left\{(\Lambda_j \otimes \Omega) E[\Gamma_j^{-1} \mid D, \mathcal{H}_g^{(p)}]\right\}, \tag{10.33}$$

where $CONST$ denotes a constant not depending on hyperparameters to be estimated.

The objective function in (10.33) involves the hyperparameters and the expectations for the unobserved random variables. Maximizing this function can be done in an iterative two-step procedure: (**E-step**) evaluating the posterior expectations given the data and the current estimates of hyperparameters and (**M-step**) maximizing the resulting objective function. That is, suppose at the $p$th iteration, the current estimate is

$$\mathcal{H}_g^{(p)} = \left(\boldsymbol{\beta}_0^{(p)}, F^{(p)}, \Omega^{(p)}, \Lambda_k^{(p)}, \delta_k^{(p)}, [\Lambda_j^{(p)}, \delta_j^{(p)}, \xi_{0j}^{(p)}], j = 1, \ldots, k-1\right). \tag{10.34}$$

The two steps in the EM algorithm at the $(p+1)$st iteration become

- ***E*-step:** Substitute the posterior expectations involved in (10.33) conditional on $D$ and $\mathcal{H}_g^{(p)}$ using the results given in Section 10.5.
- ***M*-step:** Maximize the resulting $Q^*(\mathcal{H}_g \mid \mathcal{H}_g^{(p)})$ over $\mathcal{H}_g$ to obtain the updated estimate $\mathcal{H}_g^{(p+1)}$ of $\mathcal{H}_g$ at step $(p+1)$. Specific maximization equations for this M-step are given below.

**Maximization Equations:**
The new estimate of $\mathcal{H}_g$ at the $(p+1)$th iteration is

$$\mathcal{H}_g^{(p+1)} = \left(\boldsymbol{\beta}_0^{(p+1)}, F^{(p+1)}, \Omega^{(p+1)}, \Lambda_k^{(p+1)}, \delta_k^{(p+1)}, [\Lambda_j^{(p+1)}, \delta_j^{(p+1)}, \xi_{0j}^{(p+1)}]\right., \\ \left. j = 1, \ldots, k-1\right),$$

the quantities that maximize the objective function given below from the $E$-step. The following result is useful for the derivation.

*Lemma 1: Let $D$ and $G$ be positive definite matrices, the maximum of*

$$f(G) = N \log|G| - trG^{-1}D$$

*occurs at $G = D/N$* (Anderson 2003-Lemma 3.2.2).

- By first rearranging

$$tr\left\{(\Lambda_j \otimes \Omega) E[\Gamma_j^{-1} \mid D, \mathcal{H}_g^{(p)}]\right\} = tr\Omega C_j^{(2)}$$

as described in Appendix 15.3 and then applying Lemma 1 to $Q^*$ in (10.33), the new estimate of $\Omega$ is given by

$$\Omega^{(p+1)} = \sum_{j=1}^{k} \delta_j^{(p+1)} g_j \left[\sum_{j=1}^{k} C_j^{(2)}\right]^{-1}.$$

- Similarly by rearranging

$$tr\left\{(\Lambda_j \otimes \Omega)E[\Gamma_j^{-1} \mid D, \mathcal{H}_g^{(p)}]\right\} = tr\Lambda_j C_j^{(1)}$$

as described in Appendix 15.3 and applying Lemma 1 to $Q^*$ in (10.33), the new estimates $\Lambda_j^{(p+1)}$, for $j = 1, \ldots, k$, satisfy

$$\Lambda_j^{(p+1)} = \delta_j^{(p+1)} p (C_j^{(1)})^{-1}.$$

- Taking the partial derivative of $Q^*$ in (10.33) with respect to $\delta_j$ and setting it to zero yields the following equation for the new estimate $\delta_j^{(p+1)}$,

$$-\tfrac{1}{2} g_j p \log 2 - \tfrac{1}{2} \sum_{i=1}^{g_j p} \psi\left(\frac{\delta_j^{(p+1)} - i + 1}{2}\right) - \tfrac{1}{2} E[\log|\Gamma_j| \mid D, \mathcal{H}_g^{(p)}] + \tfrac{1}{2} p \log|\Lambda_j^{(p+1)}| + \tfrac{1}{2} g_j \log|\Omega^{(p+1)}| = 0,$$

where $\psi(x) = d[\log \Gamma(x)]/dx$ denotes the digamma function.

**Note:** Le et al. (1998) show that in the univariate case (where $p = 1$), the degrees of freedoms are not identifiable and propose the use of a prior gamma distribution for it to bypass the problem. That approach is used here. In other words the unspecified degrees of freedom, $\delta_1, \ldots, \delta_k$, are assumed to have a gamma distribution

$$\pi(\delta) \propto (\delta_1 \cdots \delta_k)^{\alpha - 1} \exp\{-r(\delta_1 + \cdots + \delta_k)\}$$

with specified hyperparameters $\alpha$ and r. Hence, the estimation equation when $p = 1$ becomes

$$-\tfrac{1}{2} g_j p \log 2 - \tfrac{1}{2} \sum_{i=1}^{g_j p} \psi\left(\frac{\delta_j^{(p+1)} - i + 1}{2}\right) - \tfrac{1}{2} E[\log|\Gamma_j| \mid D, \mathcal{H}_g^{(p)}] + \tfrac{1}{2} p \log|\Lambda_j^{(p+1)}| + \tfrac{1}{2} g_j \log|\Omega^{(p+1)}| + \tfrac{\alpha - 1}{\delta_j} - r = 0.$$

- Maximizing $Q^*$ in (10.33) with respect to $F$ and $\boldsymbol{\beta}$ using Lemma 1 yields the following equations that $F^{(p+1)}$ and $\boldsymbol{\beta}_0^{(p+1)}$ satisfy

$$F^{(p+1)} = (gp)\left(E[(\boldsymbol{\beta} - \boldsymbol{\beta}_0^{(p+1)})\Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^{(p+1)})^T \mid D, \mathcal{H}^{(p)}]\right)^{-1}$$

$$\boldsymbol{\beta}_0^{(p+1)} = \left(E[\Sigma^{-1} \mid D, \mathcal{H}^{(p)}]\right)^{-1} E[\Sigma^{-1}\boldsymbol{\beta}^T \mid D, \mathcal{H}^{(p)}].$$

It is possible to impose additional structures on $\boldsymbol{\beta}_0$ without much difficulty. For example, to impose an exchangeable structure between stations but allowing the coefficients to be different within the multivariate response, simply express $\boldsymbol{\beta}_0$ as

$$\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^* R,$$

where $\boldsymbol{\beta}_0^*$ is a $(l \times p)$ matrix of coefficients and

$$R_{p \times gp} = [I_p, \cdots, I_p].$$

The estimate of $\boldsymbol{\beta}_0^*$ is then

$$(\boldsymbol{\beta}_0^{*T})^{(p+1)} = \left(RE[\Sigma^{-1} \mid D, \mathcal{H}^{(p)}]R'\right)^{-1} RE[\Sigma^{-1}\boldsymbol{\beta}^T \mid D, \mathcal{H}^{(p)}].$$

- The new estimate $\xi_{0j}^{(p+1)}$ (or equivalently $\tau_{0j}^{(p+1)}$) is obtained by maximizing $Q^*$ in (10.33) with respect to $\xi_{0j}$. That is, the maximization problem is equivalent to

$$\begin{aligned}
&\max_{\xi_{0j}} tr\left\{-\frac{1}{2}H_j^{-1}E[(\tau_j - \xi_{0j}\otimes I_p)\Gamma_j^{-1}(\tau_j - \xi_{0j}\otimes I_p)^T \mid D, \mathcal{H}_g^{(p)}]\right\} \\
&\text{or } \max_{\xi_{0j}} tr\left\{H_j^{-1}(\xi_{0j}\otimes I_p)E[\Gamma_j^{-1}\tau_j^T \mid D, \mathcal{H}_g^{(p)}]\right. \\
&\quad \left.-\frac{1}{2}H_j^{-1}(\xi_{0j}\otimes I_p)E[\Gamma_j^{-1} \mid D, \mathcal{H}_g^{(p)})(\xi_{0j}\otimes I_p]^T\right\} \\
&\text{or } \max_{\xi_{0j}} tr\left\{(\xi_{0j}\otimes I_p)E[\Gamma_j^{-1}\tau_j^T \mid D, \mathcal{H}_g^{(p)}]H_j^{-1}\right. \\
&\quad \left.-\frac{1}{2}(\xi_{0j}\otimes I_p)E[\Gamma_j^{-1} \mid D, \mathcal{H}_g^{(p)})(\xi_{0j}\otimes I_p]^T H_j^{-1}\right\}.
\end{aligned}$$

The last expression can be written as

$$\max_{\xi_{0j}}\left\{\text{vec}(\xi_{0j})\text{vec}(D) - \frac{1}{2}\text{vec}(\xi_{0j})\ G\ \text{vec}(\xi_{0j})\right\},$$

where $G$ and $D$ are specific functions of $H_j^{-1}$, $E[\Gamma_j \mid D, \mathcal{H}_g^{(p)}]$, and $E[\Gamma_j\tau_j^T \mid D, \mathcal{H}_g^{(p)}]$ as described in Appendix 15.3.
Thus, the optimal choice $\xi_{0j}^{(p+1)}$ satisfies

$$\text{vec}(\xi_{0j}^{(p+1)}) = G^{-1}\text{vec}(D).$$

The estimates for the hyperparameters can be obtained by iterating these *EM* steps until the marginal density (10.35) given below converges. When $p = 1$, the contribution of the prior distribution for $\delta$s needs to be incorporated in the marginal density for the optimization.

**Marginal Distribution** $f\left(\left\{Y^{[g_1^o]}, \ldots, Y^{[g_k^o]}\right\} \mid \mathcal{H}_g\right)$
Assume the response matrix $Y$ has the Gaussian-generalized inverted Wishart model specified by (10.1) and (10.2). Then in the notation of Section 10.2, the marginal distribution can be written as

$$f\left(\left\{Y^{[g_1^o]}, \cdots, Y^{[g_k^o]}\right\} \mid \mathcal{H}_g\right) = \prod_{j=1}^{k} t_{(n-m_j)\times g_j p}\left(\mu_o^{[j]}, \Phi_o^{[j]}\otimes\Psi_o^{[j]}, \delta_o^{[j]}\right), \tag{10.35}$$

where $t_{a\times b}$ denotes a matric-$t$ distribution (see Appendix 15.1) and

$$\mu_o^{[j]} = \mu_{(2)}^{[j]}$$

$$\Phi_o^{[j]} = A_{22}$$

$$\Psi_o^{[j]} = \frac{1}{\delta_j - g_j p + 1} [\Lambda_j \otimes \Omega]$$

$$\delta_o^{[j]} = \delta_j - g_j p + 1$$

with

$$\begin{pmatrix} \mu_{(1)}^{[j]} \\ \mu_{(2)}^{[j]} \end{pmatrix} : \begin{pmatrix} m_j \times g_j p \\ (n - m_j) \times g_j p \end{pmatrix} = Z\boldsymbol{\beta}_0^{[g_j]} + \tilde{\epsilon}^{[g_{j+1},\ldots,g_k]}\tau_{0j},$$

$$\begin{pmatrix} A_{11}^{[j]} & A_{12}^{[j]} \\ A_{21}^{[j]} & A_{22}^{[j]} \end{pmatrix} : \begin{pmatrix} m_j \times m_j & m_j \times (n - m_j) \\ (n - m_j) \times m_j & (n - m_j) \times (n - m_j) \end{pmatrix}$$

$$= I_n + ZF^{-1}Z^T + \tilde{\epsilon}^{[g_{j+1},\ldots,g_k]} H_j (\tilde{\epsilon}^{[g_{j+1},\ldots,g_k]})^T.$$

### 10.6.4 Estimating Ungauged Site Hyperparameters

Given the Type-II maximum likelihood estimates for hyperparameters associated with gauged sites as described above, the hyperparameters $\Lambda_0$, $\tau_{00}$, and $H_0$, associated with the ungauged sites can be estimated through the Sampson–Guttorp method (Sampson and Guttorp 1992). That nonparametric method extends the spatial hypercovariance associated with the gauged sites to include that of the ungauged sites. The extension is carried out by

- First estimating the covariance matrix associated with the gauged sites using the type-II maximum likelihood estimates as described above;
- Then applying the Sampson–Guttorp method to get estimates of the covariance matrix associated with the ungauged sites and the corresponding cross-covariance matrix; these are called SG estimates;
- Finally obtaining estimates for $\Lambda_0$ and $\tau_{00}$ using the resulting SG estimates and the Bartlett transformation.

Details are as follows. First write $\Psi$, the spatial component covariance matrix for all the sites given in (10.26) as

$$\begin{bmatrix} \mathcal{M}^{[u,u]} & \mathcal{M}^{[u,g]} \\ \mathcal{M}^{[g,u]} & \mathcal{M}^{[g,g]} \end{bmatrix} \tag{10.36}$$

with $\mathcal{M}^{[u,u]}$ representing the spatial covariance between the ungauged sites and $\mathcal{M}^{[u,g]}$ corresponding to the cross-covariance. $\mathcal{M}^{[g,g]}$ denoting the spatial covariance between the gauged sites is in fact $\Psi^{[1,\ldots,k]}$ that can be written in terms of the hyperparameters as given in (10.24) and (10.25).

As in the case of the gauged sites, assume in accordance with the inverted Wishart distribution, that $H_0 = \left(\Lambda^{[1,\ldots,k]} \otimes \Omega\right)^{-1}$ where $\Lambda^{[1,\ldots,k]}$ is defined in (10.28).

Let $\hat{\Lambda}_j$, $\hat{\tau}_{oj}$, and $\hat{\delta}_j$ denote the type-II MLE estimated hyperparameters associated with the gauged sites. $\mathcal{M}^{[g,g]}$ can be estimated by substituting $\hat{\Lambda}_j$, $\hat{\tau}_{oj}$, and $\hat{\delta}_j$ into expressions (10.24) and (10.25), yielding $\hat{\mathcal{M}}^{[g,g]}$. Similarly $H_0$ can be estimated yielding $\hat{H}_0$.

The SG method (see Chapter 6) estimates $\mathcal{M}^{[u,u]}$ and $\mathcal{M}^{[u,g]}$ based on $\hat{\mathcal{M}}^{[g,g]}$, yielding $\tilde{\mathcal{M}}^{[u,u]}$ and $\tilde{\mathcal{M}}^{[u,g]}$. An example of how the SG method works is demonstrated in Chapter 14 using R codes.

Equating (10.26) with (10.36) yields

$$\mathcal{M}^{[g,u]} = \mathcal{M}^{[g,g]}\xi_0^{[u]}.$$

Hence $\xi_{00}$ can be estimated by

$$\tilde{\xi}_{00} = (\hat{\mathcal{M}}[g,g])^{-1}\tilde{\mathcal{M}}^{[g,u]}.$$

Thus $\tau_{00} = \xi_{00} \otimes I_p$ can be estimated by

$$\tilde{\tau}_{00} = (\hat{\mathcal{M}}^{[g,g]})^{-1}\tilde{\mathcal{M}}^{[g,u]} \otimes I_p.$$

Similarly, $\Lambda_0$ can be estimated by

$$\tilde{\Lambda}_0 = \frac{\delta_0 - up - 1}{1 + tr((\Psi^{[1,\ldots,k]} \otimes \Omega)H^{[u]})}\left(\tilde{\mathcal{M}}^{[u,u]} - \tilde{\xi}_{00}^T\hat{\mathcal{M}}^{[g,g]}\tilde{\xi}_{00}\right).$$

Given the lack of a spatial model for interpolating degrees of freedom over space, $\delta_0$ has to be selected before interpolating the data. Kibria et al. (2002) propose a couple of potential estimates including

$$\tilde{\delta}_0 = \min(\hat{\delta}_1, \cdots \hat{\delta}_k) \text{ or } \frac{\hat{\delta}_1 + \ldots + \hat{\delta}_k}{k},$$

subject to the condition that $\delta_0 \geq up$.

We turn now to a missing data pattern introduced in Section 9.2.

## 10.7 Systematically Missing Data

To this point we have supposed that for any given step in the staircase all responses at all the sites there are measured. In practice, each site may have some responses that are never measured, by design. This happens, for example, in composite networks constructed by merging a number of subnetworks set up for different purposes. The set of species each measures may overlap but vary over the subnetworks. Thus at each step of the staircase empty vertical columns of missing values stand atop the missing gauges at each site. The missing measurements are called *systematically missing* and the data, *misaligned*.

**Open Problem** The double staircase of missing data presents an unsolved problem. Within each of the steps in the main staircase, small secondary staircases are formed because the gauges at the sites there, for measuring different responses, were installed at different times.

Dealing with these data requires the concept of the *quasi-site*. Each existing monitoring site has purely imaginary positions corresponding to individual monitors i.e., gauges and so has any geographical location that is not yet monitored. These imaginary locations are called quasi-sites. With this concept we can dichotomize all quasi-sites gauged and ungauged. The ungauged ones include those at unmonitored locations and the missing components at the monitoring sites. Moreover, the ungauged quasi-sites at the monitoring sites can be permuted to the bottom of the staircase to join with the unmonitored sites (and their quasi-sites). As a result the empty vertical columns in the staircase disappear and we can more or less proceed as we did earlier in this chapter.

Things are not quite that simple and some technical obstacles must be overcome (see Le et al. 1997). For simplicity, suppose the data staircase has just one step. Then $Y^{[g]}$ has empty columns corresponding to the total of $h$ ungauged quasi-sites, $j_1, \ldots, j_h$ located on monitoring sites that only partially measure the vector-valued responses. The gauged quasi-sites correspond to the remaining columns $j_{h+1}, \ldots, j_{gp}$. Let $r_i$, $i = 1, \ldots, gp$ be an $gp \times 1$-dimensional vector with $i$th element 1, the rest 0. Let $R_1$ and $R_2$ be indicator matrices that mark the missing and present columns: $R_1 = (r_{j_1}, \ldots, r_{j_h})$ and $R_2 = (r_{j_{h+1}}, \ldots, r_{j_{gp}})$. Finally let $R \equiv (R_1, R_2)$, an orthogonal matrix.

Observe that $Y^{[g]}$ can be rearranged through permutation as

$$Y^{[g]} = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \end{bmatrix},$$

where $Y^{(1)} \equiv Y^{[g]}R_1$ and $Y^{(2)} \equiv Y^{[g]}R_2$ consisting of just the missing and present columns, respectively, for the monitoring sites. To find the distribution of these new matrices, recall [see Equation (10.1)] that the prepermuted $Y$ follows the Gaussian-generalized-inverted-Wishart model specified by

$$\begin{cases} Y \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, I_n \otimes \Sigma), \\ \\ \boldsymbol{\beta} \mid \Sigma, \boldsymbol{\beta}_0, \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma), \\ \\ \Sigma \sim IW(\Phi, \delta), \end{cases}$$

Notice that for the purpose of explicating our results we have replaced $\Sigma$'s GIW prior by the simpler IW.

Because, after the permutation, the response vector $Y$ has effectively been partitioned into three parts, we need to partition $\beta$, $\Sigma$, $\beta^o$, and $\Phi$ accordingly. For example, we first partition $\Sigma$ as

$$\Sigma = \begin{pmatrix} \Sigma_{uu} & \Sigma_{ug} \\ \Sigma_{gu} & \Sigma_{gg} \end{pmatrix},$$

where $\Sigma_{uu}$ and $\Sigma_{gg}$ are $up \times up$, $gp \times gp$ matrices, respectively. The covariance matrix $R^{'}\Sigma_{gg}R$ corresponding to $Y^{(1)}$ and $Y^{(2)}$, can be further partitioned as

$$R^{'}\Sigma_{gg}R = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \equiv \begin{pmatrix} R_1^{'}\Sigma_{gg}R_1 & R_1^{'}\Sigma_{gg}R_2 \\ R_2^{'}\Sigma_{gg}R_1 & R_2^{'}\Sigma_{gg}R_2 \end{pmatrix},$$

where $\Sigma_{11}$ and $\Sigma_{22}$ are $h \times h$, $(gp-h) \times (gp-l)$ matrices, respectively. Furthermore, let

$$\Psi_{gg} = R'\Phi_{gg}R = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix} = \begin{pmatrix} R_1'\Phi_{gg}R_1 & R_1'\Phi_{gg}R_2 \\ R_2'\Phi_{gg}R_1 & R_2'\Phi_{gg}R_2 \end{pmatrix},$$

with $\Psi_{11}$ being $h \times h$ and $\Psi_{22}$ being $(gp-h) \times (gp-h)$. Finally, let $\Psi_{1|2} = \Psi_{11} - \Psi_{12}\Psi_{22}^{-1}\Psi_{21}$, and

$$(\beta_0^{(1)}, \beta_0^{(2)}) = \beta_0^{[g]}R = (\beta_0^{[g]}R_1, \beta_0^{[g]}R_2).$$

We obtain Bayesian spatial predictors as a special case of the analysis earlier in this chapter. More specifically, conditional on $Y^{[2]} = y^{[2)]}$ and the hyperparameters, the predictive distribution of unmeasured responses at the gauged sites is

$$Y^{(1)} \sim t_{n\times h}\left(\mu_0^{(1)}, \frac{1}{\delta^*}\ P_{(1|2)} \otimes \Psi_{(1|2)}, \delta^*\right),$$

where

$$\begin{aligned} \mu_0^{(1)} &= Z\beta_0^{(1)} + (y^{(2)} - Z\beta_0^{(2)})\Psi_{22}^{-1}\Psi_{21} \\ P_{(1|2)} &= I_n + ZF^{-1}Z' + (y^{(2)} - Z\beta_0^{(2)})\Psi_{22}^{-1}(y^{(2)} - Z\beta_0^{(2)})' \\ \delta^* &= \delta - up - h + 1. \end{aligned}$$

Similarly the predictive distribution at the ungauged locations is

$$Y^{[u]} \sim t_{n\times up}\left(\mu_0^{[u]}, \frac{1}{\delta^*}\ P_{(1|2)} \otimes \Phi_{(u|2)}, \delta^*\right),$$

where

$$\begin{aligned} \mu_0^{[u]} &= Z\beta_0^{[u]} + (y^{(2)} - Z\beta_0^{(2)})\Psi_{22}^{-1}\Psi_{21} \\ \Phi_{u|2} &= \Phi_{uu} - \Phi_{ug}(R_2\Psi_{gg}R_2')^{-1}\Phi_{gu}. \end{aligned}$$

The joint conditional predictive distribution of $\left(Y^{[u]}, Y^{(1)}\right) \mid Y^{(2)} = y^{(2)}$ can be derived in the same way, but details are omitted.

Applying these results to a specific time $t$ yields the important special case $Y_t^{[u]} \mid Y^{(2)} = y^{(2)}$ to be a multivariate-$t$ distribution, a special case of a matric-$t$ distribution:

$$Y_t^{[u]} \sim t_{1\times up}\left(Z_t\beta_0^{[u]} + (y_t^{(2)} - Z_t\beta_0^{(2)})\Psi_{22}^{-1}R_2'\Phi_{gu}, \frac{1}{\delta^*}P_{t|2} \otimes \Phi_{u|2}, \delta^*\right), \quad (10.37)$$

where

$$P_{t|2} = 1 + Z_t F^{-1} Z_t' + (y_t^{(2)} - Z_t \beta_0^{(2)}) \Psi_{22}^{-1} (y_t^{(2)} - Z_t \beta_0^{(2)})'$$

and $\Phi_{u|2}$ is defined above. A similar result can be obtained for $Y_t^{(1)} \mid Y^{(2)} = y^{(2)}$.

**Remarks**

- These results can readily be extended to the case of more than one step in the staircase but that extension is straightforward.
- Usually interest focuses on the specific case, $t = n+1$, i.e., the future. However, in some applications such as that described in Chapter 2 unmeasured past responses need to be imputed. In such cases, the entire field of missing measurements may have to be constructed and the uncertainties in doing so correctly disclosed.
- The hyperparameters can be estimated with the approach discussed earlier in the chapter. The methods are incorporated in the software that is posted online as a companion to this book. See Chapter 14 for more details.

We turn in the next section to the problem of characterizing those uncertainties.

## 10.8 Credibility Ellipsoids

This section covers an important task, the specification of predictive sets that correctly reflect the uncertainty in a mapped environmental field. Such maps are usually drawn by spatially predicting the field at a grid of geographical locations and then applying a contouring program of some kind to those predictions. How dense should that grid of points be and how uncertain is the resulting map?

This is an important question since such mapping is so common and often done for regulation, control, and abatement programs. In such cases, stating unrealistically small levels of uncertainty can lead to false confidence in the predictions. That in turn can lead to unfair actions of great severity against producers of an environmental hazard, at least when the predicted levels are high. When they are low, they can also lead to bad decisions to ignore negative impacts on human health and welfare. In short, stating uncertainty accurately is important.

Commonly, spatial mappers construct their maps one grid point at a time and calculate, say 95% prediction intervals at the same time, point by point. There would seem to be no limit to the number of grid points that could be done in this fashion.

In fact, our spatial predictor allows us to do that by applying the theory to the special case, $up = 1$. That is to say, predict unmeasured responses, response by response, site by site for all responses and all sites. The multivariate-

$t$ distribution in (10.37) then reduces to the univariate-$t$. Moreover, a standard variance formula for the multivariate-$t$ distribution of $Y_t^{[u]} \mid Y^{(2)} = y^{(2)}$ yields the variance of the predictor:

$$Var(Y_t^{[u]} \mid Y^{(2)} = y^{(2)}) = (\delta^* - 2)^{-1} P_{t|2} \Phi_{u|2}.$$

This variance yields a pointwise posterior credibility interval for the predictand at each of the sites in our grid.

As ever, there is "no such thing as a free lunch" and in this case, a cost to pay for this site by site approach. In fact that approach ignores the multiplicity of the imputations made simultaneously over this (correlated) spatial field. Overall, the chances are much less than 95% that all the imputed measurements will simultaneously lie within their intervals. Hence the point by point approach renders a poor characterization of the uncertainty in the mapped field.

However, our theory yields a much more satisfactory characterization of that uncertainty. In fact, the multivariate-$t$ predictive distribution such as that of $Y_t^{[u]} \mid Y^{(2)} = y^{(2)}$ in (10.37) allows us to derive a simultaneous credibility region. More precisely for the unmonitored sites, let $\hat{y}_t^{[u]} = Z_t \beta_0^{[u]} + \left(y^{(2)} - Z\beta_0^{(2)}\right) \Psi_{22}^{-1} \Psi_{21}$ be the $Y_t^{[u]}$s predictor in the systematically missing case above. Then conditional on $Y^{(2)} = y^{(2)}$, the $1 - \alpha$ level $(0 < \alpha < 1)$ simultaneous posterior credibility region is

$$\{Y_t^{[u]} \ : \ (Y_t^{[u]} - \hat{y}_t^{[u]}) \Phi_{u|2}^{-1} (Y_t^{[u]} - \hat{y}_t^{[u]})' < b\},$$

where

$$b = [up * P_{t|2} * F_{1-\alpha, up, \delta^*}] * (\delta^*)^{-1}.$$

This characterization shows that indeed, the "lunch is not free" and it is paid in degrees of freedom as $u$ and $p$ increase. That price increases particularly rapidly since these numbers enter through their product and soon use up all the prior information expressed through $\delta$, especially if $h$, the number of ungauged quasi-sites, is large. Of course, the analyst can artificially inflate $\delta$ to accommodate a large value of $u$ but in doing so must recognize that false certainty is being injected into the problem.

How well do these credibility ellipsoids represent uncertainty? Sun et al. (1998) provide an answer to that question through a cross-validatory assessment in the application discussed in Section 13.5. In that application, which concerned logarithmically transformed daily concentrations of $NO_2$, $SO_2$, $O_3$, and $SO_4$ over southern Ontario, the authors systematically removed and then predicted the measurements at each monitoring site, over a 24-week summer period. The univariate nominal 95% prediction intervals, fitted site by site, included the removed measurements 90%, 98%, 99%, and 100% of the time, respectively, quite a deviation from 95%. The authors noted that this may have been due to fitting a single degrees of freedom parameter for all four pollutants. In fact, $NO_2$ has quite a heavy tail and a smaller $\delta$ would have been

appropriate. The authors ran the same experiment to assess the credibility ellipsoids for the four air pollutants. The results are in the Table 10.4. On the

| Nominal coverage (%) | 50 | 80 | 90 | 95 | 99 |
|---|---|---|---|---|---|
| Empirical coverage (%) | 57 | 82 | 90 | 94 | 98 |

Table 10.4: Empirical coverage probabilities for credibility ellipsoids of various nominal levels and four air pollutants.

whole, the credibility ellipsoids seem to be quite well calibrated.

## 10.9 Wrapup

That completes a very technical chapter that presents a theory for the spatial prediction of multivariate responses among other things. Computer codes in R for implementing the approach are available free of charge. Download instructions and an R tutorial on how to fit the models and estimate the hyperparameters are given in Chapter 14.

All of the formulas have been included in the software and so the details could be skipped. However, this material is needed by critical readers who demand an understanding of the models and methods presented there before being willing to accept their validity. Moreover, it is needed by anyone contemplating their extension to even more general cases.

The framework set out in Chapters 9 and 10 lays the foundation for a theory of design that is the subject of the next chapter. Other applications are seen in Chapters 12 and 13.

# 11

# Environmental Network Design

*It is often said that experiments must be made without preconceived ideas. That is impossible. Not only would it make all experiments barren, but that would be attempted which could not be done.*

*Henri Poincaré*

*If you need statistics, you did the wrong experiment.*

*Ernest Rutherford*

Yet, in contradiction to Rutherford's famous remark, you need statistics to do the right experiment! And that is not all. For us, the experimental design is a network of sites at which gauges, i.e., monitors, are placed to measure the environmental field of concern. As Poincare's remark suggests, choosing those sites requires prior knowledge. Moreover, making that choice optimal, requires a design objective.

### But What's the Objective?

However, defining that objective can be difficult since usually a number of reasonable, often conflicting objectives can be discerned. For example, regulators may wish to build a network that detects noncompliance with proclaimed quality standards. They would prefer to gauge sites near anticipated hot-spots. In contrast, epidemiologists concerned with the health effect of a perceived hazard would want to split those sites equally between areas of high risk and areas of low risk to maximize contrast and the power of their health effects analyses. Some investigators might be interested in measuring extremes, others trends. Each of these can be measured in a variety of different ways, and those different metrics may well imply different optimal designs. Designing to monitor a multivariate response field leads to even greater challenges since now different levels of importance can attach to the different coordinates or to some index computed from them. Also involved are cost as well as levels of temporal and spatial aggregation. In combination, these identified goals and associated factors can lead to myriad possible objectives. Although the multiattribute decision paradigm has a useful role to play here, clearly the combination of terms in the resulting objective may be very large indeed. (For a discussion of the multiattribute approach in the context of network design, see the "*sampsn2.pdf*" document of PD Sampson, P Guttorp, and DM Holland at http://www.epa.gov/ttn/amtic/files/ambient/ pm25/workshop/spatial/.

Specifying that objective may even seem impossible since many of the future uses of the network simply cannot be foreseen. Zidek et al. (2000) give an example of a network comprised of several networks established at different times for different purposes. The original network, established to measure acidic deposition, tended to be located in rural areas. Later, as the state of knowledge of environmental risk evolved, air pollution came to dominate acid as a societal concern and the (by now composite) network tended to be located in urban areas.

### Network Costs Are Large

Yet the high cost of network construction and maintenance leads to persistent demand for rational designs that, in practice, cannot be ignored. This led to a solution that seems to embrace the spirit of all the objectives while not emphasizing any one of them. That solution for network design, which uses entropy to define an objective function, was proposed by Caselton and Husain (1980), Caselton and Zidek (1984, hereafter CZ) and again by Shewry and Wynn (1987), and Sebastiani and Wynn (2000). It has also been embraced in the work of Bueso et al. (1998, 1999b), Angulo et al. (2000), and Angulo and Bueso (2001). In fact, the idea of using entropy in experimental design goes back at least to Lindley (1956). There is a substantial body of work on optimal design in the Bayesian context although none covers environmental applications as discussed here; for a review see Verdinelli (1991).

### Entropy Approach

The entropy approach to design was implemented by Caselton et al. (1992, hereafter, CKZ) to obtain a method of ranking stations for possible elimination from an existing network; refinements were added by Wu and Zidek (1992). Guttorp et al. (1993) tackled the complementary problem of extending an existing network. Le and Zidek (1994) extended the approach to a multivariate setting. Zidek et al. (2000) proposed a method for incorporating costs. Le et al. (2004) address the design problem for multivariate responses where the existing monitoring network has stations with different operational periods, resulting in a monotone (staircase) data pattern.

The basic idea underlying the just cited work is that all data have the fundamental purpose of reducing uncertainty about some aspect of the world. As discussed in Chapter 3, the postulates of Bayesian theory imply uncertainty can be quantified in terms of probability distributions. And the postulates of entropy theory, in turn, imply that the uncertainty in any distribution is indexed by its entropy. Ineluctably, an optimal design must minimize residual entropy after data have been collected.

We develop our entropy approach within a hierarchical Bayes framework with some estimated components. That framework is natural; designers invariably need prior information. For example, Linthurst and his coinvestigators

(Linthurst et al., 1986, p. 4) relied on their expectations of low alkalinity to help select their sample of surface water bodies in the United States. After all, the real information only comes from the experiment being designed to produce it! We now turn to a description of the approach, the subject of this chapter.

### Joint Predictive Distribution***

We first need the joint predictive distribution of concentration levels at locations of interest. The Bayesian hierarchical models described in Chapter 9 and generalized in Chapter 10 provide the distributions we use to demonstrate the entropy design approach. We derive specific optimal design criteria and discuss computational and other related design issues. But before any of that, we review, in the next section, some of the basic approaches that have been taken to network design.

## 11.1 Design Strategies

A sampling domain may seem to offer a continuum of possible monitoring locations (sites). However, in practice only a small discrete set of possibilities are usually be available due to such things as accessibility. That is the set-up addressed in this chapter.

### Probability or Model-Based?

Generally designs may be *probability-based* or *model-based*. The former includes simple random sampling: sites are sampled at random with equal probability (usually without replacement). The measured responses, which may even be a time series of values, would then be (approximately) independent and their associated inferential theory quite simple. As well, such designs prove quite robust since nothing is assumed about the population of possible responses.

However, these designs can also be very inefficient under the simplest of assumptions about the population. Moreover, sampling sites could end up adjacent to each other by chance, thereby making one of them redundant except in exceptional cases. Thus, samplers commonly rely on population models and sometimes achieve dramatic increases in efficiency under these models. For example, they may postulate a population that consists of a union of homogeneous geographical strata. Under that model, only a small number of sites would need to be selected from each stratum. Because of their appeal, such designs have been used in a survey of U.S. lakes (Eilers et al. 1987) and in EMAP (see, for example, http://www.epa.gov/emap).

While stratification diversifies sampling, adjacent pairs of sites could still obtain, either within strata or on opposite sites of a common boundary. Moreover, knowledge about environmental fields can well exceed what can be accommodated by the models of probability-based theory. That knowledge can lead to greater gains in efficiency than achievable through probability-based designs and hence model-based designs are commonly used in practice to achieve design optimality.

### Regression or Random Field Approach?

Broadly speaking, two distinct approaches have emerged for selecting model-based (or optimal) designs (Federov and Müller, 1988, 1989), based either on *regression models* or *random field models*. The latter are emphasized in this chapter. However, the former are reviewed as they has been offered as an approach to network design. Their advantages and disadvantages in that role are described.

### Regression Model-Based Approach

Regression model (optimal design) theory originally had nothing to do with monitoring networks. Originating with Smith (1918), it was refined by Elfving (1952), Keifer (1959), and others (see Silvey 1980, Fedorov and Hackl 1997, and Müller 2001 for reviews). The theory addresses continuous sampling domains, $\mathcal{X}$. However, optimal designs there, $\xi$, have finite support, $x_1, \ldots, x_m \in \mathcal{X}$ with $\sum_{i=1}^{m} \xi(x_i) = 1$. In all, $n \times \xi(x_i)$ (suitably rounded) responses would then be measured at $x_i$ for all $i = 1, \ldots, m$ to obtain $y_1, \ldots, y_n$. Underlying the method is a regression model, $y(x) = \eta(x, \beta) + \varepsilon(x)$ relating the $y$s to the selected (and fixed) $x$s. Another key assumption: the $\varepsilon$s are independent from one sample point $x$ to another. Optimality was then defined in terms of the efficiency of estimators of $\beta$, thus yielding an objective function $\Phi(M(\xi))$ to be optimized, where $M(\xi)$ denotes the information matrix and $\Phi$ a positive function that depends on the criterion adopted. For example, in ordinary linear regression, $M(\xi) = \sigma^2[\mathbf{X}'\mathbf{X}]^{-1}$. $\Phi$ could be any of a number of possibilities including $\Phi(A) = -\log|A|$ (D-optimality) or $\Phi(A) = \text{Trace}(A)$ (A-Optimality). An elegant mathematical theory emerged together with numerical algorithms for computing the optimum design approximately.

To illustrate, suppose that conditional on $x \in [a, b]$, $y(x) = \alpha + \beta x + \varepsilon(x)$ and the $\epsilon$s are independent of the $x$s as well as each other. Then to minimize the variance of the least-squares estimator of $\beta$, the optimal design would have $x_1 = a, x_2 = b$ while $\xi(x_1) = \xi(x_2) = 1/2$.

Regression-based optimal design theory as described above encounters difficulties in application to network design. There monitors must be located at a subset of available sites and then simultaneously measure the field of interest regularly for an indefinite period. For example, every TEOM particulate air pollution monitor at an urban sampling site yields hourly observations. To

measure $n$ responses each time would entail gauging $n$ sites, forcing $\xi \equiv 1/n$. That in turn would completely determine the design once its support were specified, making the classical theory of design irrelevant.

Nevertheless, a sustained effort has been made to adapt the regression model paradigm to encompass network design. Federov and Müller (1989) cite Gribik et al. (1976) as an early attempt. However, the major push came later (Fedorov and Müller 1989). The motive may have been a unified optimal design theory. However, Fedorov and Müller (1989) give a more pragmatic reason. They argue in their paper that hitherto, only suboptimal designs could be found, feasible algorithms being limited to adding just one station at a time, albeit optimally. However, algorithms from the regression model theory offered promise (and algorithms!) by which genuinely optimal designs could be computed. (This reason may not be quite as compelling for the maximum entropy designs proposed in the next section where quick algorithms are now available for finding the optimum designs, at least for networks of moderate size.)

To that end, Fedorov and Müller (1988), assume that at time $t = 1, \ldots, T$, $y_t(x_i) = \eta(x_i, \beta_t) + \varepsilon_t(x_i)$. Once again, the $\varepsilon$s are all independent of each other. However, the $\beta_t$s are random and autocorrelated. Moreover, $\eta(x_i, \beta_t) = g^T(x_i)\beta_t$ for a known vector-valued $g$. Thus, this ingenious model captures both temporal and spatial covariance. By the way, the latter is not as restricted as it might seem at first glance, since the coordinates of $g$ can be eigenfunctions of the spatial covariance kernel when it is known. That covariance can thus be approximated well if the dimension of g is sufficiently large. But this comes at the expense of fixing the variances of these random effects to be eigenvalues of that kernel. The design objectives embrace the performance of either a linear predictor of a single $\beta_t$, say at time $t =$ "now" or of the mean of the common $\beta_t$ distribution. These objectives would not seem compelling when the coefficients are merely artifacts of the eigenvector expansion associated with the covariance kernel rather than quantities of substantive interest such as the slope of a genuine regression model.

The authors recognize the limitation mentioned above, that in this context the optimum design must be a subset of the available design set. However, to bring in the classical theory and associated algorithms, they relax that restriction and admit general $\xi$s, albeit subject to a boundedness requirement, so that established numerical search solutions now obtain. They call this substitution a "continuous approximation" and solve that problem instead of the original. The result will not usually be a feasible solution to the original problem and Fedorov and Müller (1988, 1989) note the challenge of interpreting it, seen variously as a local density, an indicator of a hot-spot, or a design with more than one monitor at some sites. Further work in this direction described in Müller (2001) may help clarify the nature of this approximation. However, it is unclear about the value of substituting the approximate problem (and big associated toolbox of computational algorithms) for the hard to solve exact

discrete design problem (and inevitable feasible to compute approximations), an issue that seems to need further investigation.

Apart from the problem of interpreting the optimum, issues of a more technical nature arise. First, suppose a genuine regression model (as opposed to eigenfunction expansion) is used above so that the objective function is substantively meaningful. Then the range of spatial covariance kernels will be restricted unless the $\varepsilon$s are allowed to be spatially correlated. That need is met in the extensions of the above model above described in the reviews of Fedorov (1996) and Müller (2001). However, the resulting design objective function "does not have much in common with [the original] besides notation" in the words of Fedorov (1996, page #524). A new toolbox will have to be created except in simple cases where an exhaustive search is needed. Back to square one!

While the regression model above does have substantive appeal, its value is uncertain. Environmental space–time fields tend to be so complex that their random response fields are only crudely related to spatial site coordinates. Moreover, the shape of that field can vary dramatically over time and season. In other words, finding a meaningful, known vector-valued function $g$ above would generally be difficult or impossible.

The alternative, the eigenfunction expansion, also presents difficulties according to Fedorov (1996), relating to the problem of accurately approximating the spatial covariance kernel. Complications can arise in particular when the size of the proposed network is large. Moreover, while the eigenfunctions are known to exist under very general conditions, it is not clear that actually finding them in usable form will be possible in problems of realistic size.

To summarize, the regression model approach does offer a very highly evolved theory for design, along with a substantial toolbox of algorithms for computing optimal designs, at least approximately. It also offers a broad range of objective functions which formally embraces that which comes out of the maximum entropy approach in the Gaussian case we introduce in the next section. However, forcing the network design problem into the regression model mold proves challenging both in terms of interpretation of the resulting optima as well as satisfying the assumptions underlying that approach.

### Link to Geostatistics

Perhaps the strongest link between the regression modeling and random field approaches can be found in geostatistics. Wackernagel (2003) gives a very readable recent account of that subject while Myers (2002) addresses space–time processes from the perspective of geostatistical modeling. Because until very recently, that subject has concerned itself with spatial fields while we focus on space–time fields, this approach is not described in detail.

Unlike the regression modeling approach (above) that emphasizes parameter estimation, geostatistics has tended to focus on the prediction of unmeasured values in a spatial field that, paradoxically, is regarded as random even

though it is fixed. Two methods are commonly employed, cokriging and universal kriging. The first concerns the prediction of an unmeasured coordinate of the response vector, say $y_1(x_0)$ using an optimal linear predictor based on the observed response vectors at all the sampling sites. The coefficients of that optimal predictor are found by requiring it to be unbiased and to minimize the mean-square prediction error. They depend on the covariances between responses and between the sites, covariances that are unrealistically assumed to be known and later estimated from the data usually without adequately accounting for the additional uncertainty thereby introduced. In contrast to the first, the second relies on a regression model precisely of the form given above, $y(x) = g^T(x)\beta + \varepsilon_t(x)$, where the $\varepsilon$s are assumed to have a covariance structure of known form. However, unlike the regression modeling approach above, the goal is prediction of the random response (possibly a vector) at a point where it has not been measured. Moreover, $g$ (that may be a matrix in the multivariate case) can represent an observable covariate process. Optimization again relies on selecting coefficients by minimizing mean-squared prediction error subject to the requirement of unbiasedness. Designs are commonly found iteratively one future site at a time, by choosing the site $x_0$ where the prediction error of the optimum predictor proves to be greatest. Other approaches to model-based designs have been proposed. For example, Bueso et al. (1999a), offer one based on stochastic complexity.

That completes the survey of regression-based approaches. The maximum entropy approach is described in the next section.

## 11.2 Entropy-Based Designs

Entropy reflects the reduction in uncertainty when a random variable is observed. Thus, in an optimal environmental design context, the objective is to maximize the uncertainty reduction when selecting a number of stations where their responses are measured. This idea, first proposed by Caselton and Zidek (1984), can be formalized. However, before doing that we give an introduction to entropy. For a more detailed description, see Theil and Fiebig (1984).

## 11.3 Entropy

Suppose $X$, a discrete random variable, takes a finite number of possible outcomes $E_1, \ldots, E_n$ with probabilities $p_1, \ldots, p_n$, respectively. Let $\phi$ be a function defined on the interval $(0, 1]$, $\phi(p_i)$ representing the uncertainty associated with the event $X = E_i$, for $i = 1, \ldots, n$. Before observing $X$, we would expect the reduction in our uncertainty, $H(X)$, to be the weighted average,

$$H(X) = \sum_{i=1}^{n} p_i \phi(p_i).$$

Now add an additional axiom, that the expected reduction in uncertainty from jointly observing two independent random variables be the sum of the reductions from observing each random variable separately, that is

$$H(X,Y) = H(X) + H(Y) \text{ whenever} X, Y \text{ are independent.}$$

This implies that $\phi(p_i) = -\log(p_i)$. Thus, the uncertainty in the discrete random variable $X$ becomes what is known as its entropy,

$$H(X) = -\sum_{i=1}^{p} p_i \log(p_i).$$

However, extending that definition to cover continuous random variables $Y$ proves problematical. Simply taking the limit through a series of progressively finer discrete approximations does not work. In fact, the outcome would not be finite. Alternatively, one might suppose entropy could be defined by analogy. Sums and probabilities could be replaced by integrals and density functions, respectively. In other words, why not take

$$H(Y) = E[-\log f(Y)],$$

where $f$ is the probability density function (PDF) of the continuous random variable $Y$? The answer is that the result would not be invariant under transformations of $Y$. That is, merely changing the scale of measurement of $Y$ from Celsius to Fahrenheit, for example, would lead to a different index of our state of uncertainty, a nonsensical result! That deficiency derives from the fact that $f$, unlike $p$ in the discrete case, is not a probability. Instead it is a rate, the rate of change in probability per unit change in $Y$.

Jaynes (1963) proposes instead, that

$$H(Y) \equiv -E\left[\log \frac{f(Y)}{h(Y)}\right],$$

where $h$ is a reference measure representing complete ignorance. Although defining $h$ unambiguously remains an unresolved issue, the Jaynes entropy has come to be widely used in the continuous case and, in any case, seems a good index of uncertainty. After all, it has many natural properties such as invariance and additivity for independent random responses. In any case, this is the one we use in our approach to designing environmental networks in this chapter. In the example below and throughout the chapter, the reference measure $h$ is chosen to ensure invariance of the entropy under affine transformations of $X$.

### Example: Entropy of Multivariate *t* Distribution

*Example 11.1. Normal—inverted Wishart*
Assume $Y$, a $g$-dimensional random vector, has a multivariate $t$ distribution $t_g(\mu, s^{-1}\Psi, \delta)$. That distribution can be considered to be a marginal distribution deriving from a conjunction of a Gaussian and an inverted Wishart distribution defined as

$$Y \mid \Sigma \sim N_g(\mu, \Sigma);$$
$$\Sigma \mid \Psi, \delta \sim IW(\Psi, \delta)$$

with $s = \delta - g + 1$. Conditional on the hyperparameters $\Psi, \delta$, the joint entropy $H(Y, \Sigma)$ can be decomposed in two ways:

$$H(Y, \Sigma) = H(Y \mid \Sigma) + H(\Sigma) \tag{11.1}$$
$$H(Y, \Sigma) = H(\Sigma \mid Y) + H(Y).$$

The equivalence of the left-hand sides of (11.1) implies

$$H(Y) = H(Y \mid \Sigma) + H(\Sigma) - H(\Sigma \mid Y). \tag{11.2}$$

The components of (11.2) can be computed using the reference measure

$$h(Y, \Sigma) = h(Y)h(\Sigma) = |\Sigma|^{-(g+1)/2}.$$

The first component $H(Y \mid \Sigma)$ is then

$$\begin{aligned} H(Y \mid \Sigma) &= \frac{1}{2}E(\log|\Sigma| \mid \Psi) + \frac{g}{2}(\log(2\pi) + 1) \\ &= \frac{1}{2}E(\log|\Psi|) + \frac{1}{2}E(\log|\Sigma\Psi^{-1}|) + \frac{g}{2}(\log(2\pi) + 1) \\ &= \frac{1}{2}E(\log|\Psi|) + c_1(g, \delta). \end{aligned}$$

The constant, $c_1(g, \delta)$ for given $g$ and $\delta$, like its cousins below, $c_2, \ldots, c_5$, plays no role in our theory and hence is not specified. The last equality obtains since $\Psi\Sigma^{-1} \sim W(I_g, \delta)$.

Similarly the second component $H(\Sigma)$ can be expressed as

$$\begin{aligned} H(\Sigma) &= E[\log f(\Sigma)/h(\Sigma)] \\ &= \frac{1}{2}\delta\log|\Psi| - \frac{1}{2}\delta E(\log|\Sigma|) - \frac{1}{2}E(tr\Psi\Sigma^{-1}) + c_2(g, \delta) \\ &= -\frac{1}{2}\delta E(\log|\Sigma\Psi^{-1}|) - \frac{1}{2}E(tr\Psi\Sigma^{-1}) + c_2(g, \delta) \\ &= \frac{1}{2}\delta E(\log|\Sigma^{-1}\Psi|) - \frac{1}{2}E(tr\Psi\Sigma^{-1}) + c_2(g, \delta) \\ &= c_3(g, \delta), \text{ since again } \Psi\Sigma^{-1} \sim W(I_g, \delta). \end{aligned}$$

Similarly, the last component is

$$\begin{aligned} H(\Sigma \mid Y) &= \frac{1}{2}(\delta + 1)log|\Psi| - \frac{1}{2}(\delta + 1)E(\log|\Psi + YY'|) + c_4(g, \delta) \\ &= -\frac{1}{2}(\delta + 1)E(\log|1 + Y'\Psi^{-1}Y|) + c_4(g, \delta) \\ &= c_5(g, \delta). \end{aligned}$$

The first equality obtains because $\Sigma$ has the following posterior distribution (see Anderson 2003, for example)

$$\Sigma \mid Y, \Psi, \delta \sim IW(\Psi + YY', \delta + 1).$$

Furthermore, note that

$$|\Psi + YY'| = |\Psi|(1 + Y'\Psi^{-1}Y)$$

and for $Y \sim t_g(\mu, s^{-1}\Sigma, s)$ the quadratic term $Y'\Psi^{-1}Y$ has an $F$ distribution with degrees of freedom depending on $g$ and $\delta$.

Substituting these last results back into (11.2) yields, for the total entropy of $Y$,

$$H(Y) = \frac{1}{2} log|\Psi| + c(g, \delta), \tag{11.3}$$

where $c$ is a constant depending on $g$ and $\delta$.

## 11.4 Entropy in Environmental Network Design

As noted earlier, entropy can be an appealing design criterion because it sidesteps the problem of specifying a particular design objective. Moreover, that criterion fits well into the Bayesian framework adopted for the spatial–temporal stochastic models discussed in the book.

To describe the approach more precisely, we associate a random variable with every site in a spatial random field representing concentration levels, for example. The variables corresponding to the sites in the discrete random field may be stacked to obtain a random vector. The random vector field is observed at $g$ discrete gauged sites at sampling times $j = 1, \cdots, n$, yielding a $g \times 1$ data vector, $X_j^{(2)} = \left(X_j^{(21)}, \cdots, X_j^{(2g)}\right)'$ at time $j$. Of interest is a $u \times 1$ vector, $X_{n+1}^{(1)} = \left(X_{n+1}^{(11)}, \ldots, X_{n+1}^{(1u)}\right)'$, of unmeasured future values at $u$ ungauged sites at time $n + 1$. The spatial field is over the domain of $u + g$ discrete sites. Let $X_j$ denote the gauged and ungauged responses combined at time $j$; i.e., $X_j' \equiv (X_j^{(1)'}, X_j^{(2)'})$.

Suppose $X_j$ has the joint probability density function $f_j$ for all $j$. The total uncertainty about $X_j$ may be expressed by the entropy of its distribution; i.e., $H_j(X_j) = E[-\log f_j(X_j)/h(X_j)]$, where $h(\cdot)$ is a not necessarily integrable reference density (see Jaynes 1963). Note that the distributions involved in $H_j$ may be conditional on certain covariate vectors $\{z_j\}$ regarded as fixed.

Given the network's mission to monitor the environment, we regard the next value $X_{n+1}$ as being of primary interest. However, $X_{n+1}$'s probability density function $f_{(n+1)}(\cdot) = f_{(n+1)}(\cdot \mid \theta)$ depends on a vector of unspecified model parameters, say $\theta$, so it cannot be used directly in computing its uncertainty $H(X_{n+1})$. Uncertainty about $\theta$ could be absorbed by averaging $f_{(n+1)}(\cdot \mid \theta)$ with respect to $\theta$'s distribution to obtain $X_{n+1}$'s marginal distribution and hence its entropy. However, $\theta$ is of interest in its own right. For example, $\theta$ may include the (spatial) covariance matrix of $X_{n+1}$, $\Sigma$, which

has potential use in spatial interpolation. Therefore it is important in its own right. Thus we, like Caselton et al. (1992), include reducing the uncertainty about $\theta$ among the network's objectives. As a result, the total entropy becomes $H_{n+1}(X, \theta)$ conditional on the data, $D \stackrel{defn}{=} \left\{X_j^{(2)}, j = 1, \ldots, n\right\}$.

For purposes of optimizing environmental design, partition $X_{n+1} = (U, G)$ into two subvectors, $U \equiv X_{n+1}^{(rem)'}$ representing stations not selected for the network at time $n + 1$, and $G \equiv X_{n+1}^{(sel)'}$ those selected. The objective? Find $G$, of preset dimension, that maximizes the corresponding entropy.

### Fundamental Identity

Towards our achievement of that objective, note that the total a priori uncertainty $H(X_{n+1}, \theta)$, denoted by $TOT$ and conditional on $D$, is reduced by observing $X_{n+1}$. Now in terms of the prospective gauged and ungauged sites, that total can be decomposed as

$$TOT = PRED + MODEL + MEAS,$$

where, assuming $h(X_{n+1}, \theta) = h_1(X_{n+1})h_2(\theta)$ and $h_1(X_{n+1}) = h_{11}(U)h_{12}(G)$,

$$PRED = E[-\log(f(U \mid G, \theta, D)/h_{11}(U)) \mid D],$$
$$MODEL = E[-\log(f(\theta \mid G, D)/h_2(\theta)) \mid D],$$

and

$$MEAS = E[-\log(f(G \mid D)/h_{12}(G)) \mid D].$$

Assuming negligible measurement error, eliminating all uncertainty about $G$ by observing it would lead to an expected reduction in uncertainty given by $MEAS$. Thus, it is optimal to select the gauged stations so as to maximize $MEAS$.

Since $TOT$ is fixed, it follows that the same selection of those sites meets another design objective, that of minimizing $PRED + MODEL$. The latter represents the residual uncertainty about the model parameters and the values of the random field at the ungauged sites, after observing $G$. Incidentally, it is easily seen that had $H(X_{n+1})$ been decomposed analogously instead of $H(X_{n+1}, \theta)$, the same optimization criterion, maximization of $MEAS$, would have been achieved.

### Extension or Reduction?

In practice, the redesign of an environmental network can involve either an extension or a reduction of an existing one. Bueso et al. (1998) describe well, these two broad design objectives. Earlier, CKZ considered the problem of reducing the number of sites in a network that has been providing data for some time. In this framework, the problem would be to optimally partition

$X_{n+1}^{(2)}$ so that after appropriately relabeling the coordinates of $X_{n+1}^{(2)}$, it can be written as $\left(X^{(rem)'}, X^{(sel)'}\right)$ where $X^{(rem)}$ and $X^{(sel)}$ are $u_1$ and $g_1$ dimensional vectors respectively, $u_1 + g_1 = g$, corresponding to the sites that will be ungauged and gauged in the future. Using 48 months of available data, Wu and Zidek (1992) also implement this approach in an analysis of 81 selected sites from the NADP/NTN network, an existing network of wet deposition monitoring stations in the United States.

For extending an environmental network, this framework can be used by gauging a specified number $u_2$ of sites corresponding to coordinates of $X_{n+1}^{(1)}$. That is, the new gauged sites are selected by optimal partitioning of $X^{(1)}$ which, after reordering its coordinates, yields $X^{(1)} = \left(X^{(rem)'}, X^{(add)'}\right)$ where $X^{(rem)'}$ is a $u_1$-dimensional vector representing the future ungauged sites and $X^{(add)'}$ is a $u_2$-dimensional vector representing the future gauged sites. The resulting network will consist of the sites corresponding to the coordinates of $\left(X^{(add)'}, X^{(2)'}\right) \equiv G$, which is of dimension $(g + u_2)$.

Next, the criteria for redesigning an environmental network including extension and reduction are provided. First the simple univariate setting is described and then generalization to more complex settings is discussed.

## 11.5 Entropy Criteria

Consider the simple univariate setting of Chapter 9 where concentration levels are observed at times 1 to $n$ for $g$ locations. Assume no measurements are available at $u$ other specified locations. To obtain the entropy criterion, the predictive distribution for all locations at a future time is required. For completeness, we briefly review that distribution which is derived in Chapter 9. Note that we impose no condition of isotropy in its derivation, thereby endowing our design criterion with an advantageous feature.

## 11.6 Predictive Distribution

Let $Y_t$ be a $p$-dimensional (i.e., strung out) random row vector denoting the random field at time $t$. The first $u$ coordinates are those with no data available (ungauged sites) and the remaining $g$ coordinates are those with observed data (gauged locations), $y_t(s_1), \ldots, y_t(s_g)$ for $t = 1, \ldots, n$. The vector $Y_t$ can be partitioned as $Y_t = (Y_t^{(u)}, Y_t^{(g)})$ corresponding to the locations without observations $(u)$ and those with measurements $(g)$.

The random variable $Y_t$ is assumed to be independent and follow a Gaussian distribution

$$Y_t \mid z_t, B, \Sigma \overset{\text{independent}}{\sim} N_p(z_t B, \Sigma), \tag{11.4}$$

where $z_t \equiv (z_{t1}, \ldots, z_{tk})$ is a $k$-dimensional row vector of covariates and $B$ denotes a $(k \times p)$ matrix of regression coefficients with $p = u + g$,

$$B \equiv \left( B^{(u)} \;\; B^{(g)} \right),$$

partitioned in accord with the partitioning of $Y_t$. The covariance matrix $\Sigma$ is partitioned accordingly as

$$\Sigma = \begin{pmatrix} \Sigma_{uu} & \Sigma_{ug} \\ \Sigma_{gu} & \Sigma_{gg} \end{pmatrix}.$$

Assume $B$ and $\Sigma$ follow conjugate prior distribution (c.f. Anderson 2003),

$$B \mid B_o, \Sigma, F \sim N_{kp}\left(B_o, F^{-1} \otimes \Sigma\right) \tag{11.5}$$

$$\Sigma \mid \Psi, \delta \sim W_p^{-1}(\Psi, \delta), \tag{11.6}$$

where $N_p(\mu, \Sigma)$ denotes the $p$-dimensional Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$. $W_p^{-1}(\Psi, \delta)$ denotes the $p$-dimensional inverted Wishart distribution with scale matrix $\Psi$ and $m$ degrees of freedom.

Let $D = \left\{ (y_1^{(g)}, z_1), \ldots, (y_n^{(g)}, z_n) \right\}$ be the data. The spatial predictive distribution is given below. Before presenting it, we give required notation. First, we need the least-squares estimator of $B^{(g)}$, namely, $\hat{B}^{(g)} = (\sum_{t=1}^n z_t z_t')^{-1} \sum_{t=1}^n z_t^T y_t^{(g)}$. Second comes the residual sum of squares: $S = \sum_{t=1}^n (y_t^{(g)} - z_t \hat{B}^{(g)})^T (y_t^{(g)} - z_t \hat{B}^{(g)})$. Third are the weights we need to combine the prior and data-based versions of $B^{(g)}$; they appear in the weights matrix $W = (A + F)^{-1} F^{-1} = A^{-1}(F^{-1} + A^{-1})^{-1}$. Note that the inverses of $A$ and $F$ represent *precision* and determine whether the data based or prior should get the most weight, the latter when $W$ is large, that is, when $F^{-1}$ is small. Finally, we need a couple of rescaling values that derive from the covariates and from model fit residuals obtained when using the prior version of $B^{(g)}$:

$$c = 1 + z(A + F)^{-1} z^T$$

$$d = 1 + z F^{-1} z^T + \left( y_f^{(g)} - z_f B_o^{(g)} \right) \Psi_{gg}^{-1} \left( y_f^{(g)} - z_f B_o^{(g)} \right)^T.$$

With the notation we can state the following result.

*The predictive distribution of* $Y_f = \left( Y_f^{(u)'}, Y_f^{(g)'} \right)$ *given covariate vector* $z_f$ *and the prior hyperparameters* $B_o$ *and* $(\Psi_{gg}, \Psi_{u|g}, \tau_o)$*, is*

$$Y_f^{(g)} \mid D \sim t_g\left( \mu^{(g)}, \frac{c}{l} \hat{\Psi}_{gg}, l \right) \tag{11.7}$$

$$Y_f^{(u)} \mid Y_f^{(g)} = y_f^{(g)}, D \sim t_u\left( \mu^{(u)}, \frac{d}{q} \Psi_{u|g}, q \right), \tag{11.8}$$

*where* $t_r$ *denotes the* $r$*-variate Student's t-distribution,* $l = \delta + n - u - g + 1$*,* $q = \delta - u + 1$*, and*

$$\hat{\Psi}_{gg} = \Psi_{gg} + S + (\hat{B}^{(g)} - B_o^{(g)})'(A^{-1} + F^{-1})^{-1}(\hat{B}^{(g)} - B_o^{(g)})$$
$$\mu^{(g)} = (1 - W)\hat{B}^{(g)} + WB_o^{(g)}$$
$$\mu^{(u)} = z_f B_o^{(u)} + \tau_o \left( y_f^{(g)} - z_f B_o^{(g)} \right).$$

.

Here $\Psi_{gg}$ and $\Psi_{u|g}$ represent the hypercovariance matrix between gauged sites and residual hypercovariance matrix between ungauged sites, respectively.

## 11.7 Criteria

The total entropy $H(Y_f)$ can be expressed, using the predictive distribution (11.7)–(11.8), as

$$\begin{aligned} H(Y_f \mid D) &= H(Y_f^{(u)} \mid Y_f^{(g)}, D) + H(Y_f^{(g)} \mid D) \\ &= \frac{1}{2} \log |\Psi_{u|g}| + c_u(u, q) + \frac{1}{2} \log |\hat{\Psi}_{gg}| + c_g(g, l), \end{aligned} \tag{11.9}$$

where $c_u(u, q)$ and $c_g(g, l)$ are constants depending on the degrees of freedom and the dimensions of the ungauged and gauged sites, respectively. The last equality is obtained by applying the entropy of the multivariate $t$ distribution given in (11.3). The resulting decomposition can be used to establish the entropy criterion by maximizing $MEAS$ as described in Section 11.2. Specific criteria for redesigning, reducing, and extending a network are given below.

**Criterion for Redesign**:
The purpose of redesigning a network is to select a new set of locations for a given dimension (say $g_1$) among the $(u + g)$ locations to maximize the corresponding entropy. The new locations are selected from both the stations in the existing network and the new potential sites. Specifically, partition $Y_f$ into two components denoted by $(Y_f^{(rem)'}, Y_f^{(sel)'})$, where $Y_f^{(sel)}$ corresponds to a set of $g_1$ locations. Partition $sel$ further to $sel_g$ and $sel_u$ corresponding to the gauged and ungauged sites, respectively. The optimal entropy criterion, using (11.9), becomes

$$\max_{sel_u, sel_g} \left[ \left( \frac{1}{2} log|\Psi_{u|g}| + c_u(u, q) \right)^{sel_u} + \left( \frac{1}{2} log|\hat{\Psi}_{gg}| + c_g(g, l) \right)^{sel_g} \right]. \tag{11.10}$$

**Criterion for Reduction**:
We must find the $g_1$ gauged sites we wish to keep from an existing network. To this end, partition $Y_f^{[g]}$ into two components, $(Y_f^{[g]})^{(rem)'}, (Y_f^{[g]})^{(sel)'})$, $(Y_f^{[g]})^{(sel)}$ corresponding to a subset of $g_1$ locations. The optimal criterion becomes

$$\max_{sel_g} \left( \frac{1}{2} \log |\hat{\Psi}_{gg}| \right)^{sel_g} . \tag{11.11}$$

**Criterion for Extension**:
For augmenting the network, we have to find $u_1$ sites, among the $u$ available, to add to the existing network. Partition $Y_f^{[u]}$ into two components, $(Y_f^{[u]})^{(rem)'}, (Y_f^{[u]})^{(add)'})$, *add* corresponding to the chosen locations. The *add* sites, a vector of dimension $u_1$, are selected to maximize the corresponding entropy in (11.9). The optimality criterion becomes

$$\max_{add} \left( \frac{1}{2} \log |\Psi_{u|g}| \right)^{add} . \tag{11.12}$$

The entropy approach has been extended to more complex situations, including those with multivariate responses (Le and Zidek 1994), systematically missing data (where not all stations in the existing network measure the same set of pollutants Zidek et al. 2000) and staircase patterns of missing data (Le et al. 2001). These extensions are specifically for augmenting a network and the Kronecker structure is imposed on the hypercovariance matrix to reduce the number of parameters that need to be estimated. In all cases, the entropy criterion is also to maximize $\log |\Lambda|^{add}$ where $\Lambda$ is the hypercovariance matrix among the ungauged sites. The entropy approach to design is illustrated in the example below.

## 11.8 Incorporating Cost

Discussion so far has been about optimal designs based purely on the reduction of uncertainty as reflected by entropy. However, in applications, cost is also an important consideration. Costs accrue from the initial preparation of a site as well as from its ongoing operations. Hence, the costs may well vary from site to site. For example, in network redesign, using an existing site avoids the sometimes substantial preparation costs, depending on such things as the ease of access and the cost of new equipment.

Furthermore, entropy theory suggests uncertainty reduction will increase monotonically as the number of selected sites increases. However, the growth rate will begin to tail off at some point where a marginal gain in entropy will be seen (Caselton et al. 1992). Meanwhile, the costs will also increase in a monotone fashion until eventually cost will outweigh benefit. At that point a practically optimal design will obtain.

Zidek et al. (2000) propose a direct approach to incorporating costs in a composite objective criterion that requires a cost to entropy conversion factor. Denote by $E(s)$ the reduction in entropy per period, assumed constant over time. Let $C_{op}(s)$ be the cost of operating the network over a single time period. Then the total cost of running the network for that period is $C(s) = C_{op}(s) + C_{init}(s)$ where $C_{init}(s)$ denotes the per-time-period cost for the

initial preparation of the network. Define the composite objective function as $O(s) = E(s) - DE \times C(s)$, $DE$ being the cost to entropy conversion factor. This factor gives the number of entropy units that one would trade for a unit of cost. In practice, this factor would have to be elicited from those charged with redesigning the network. When $DE = 0$ the objective function becomes the entropy approach described above.

**More "Bang for the Buck!"**

An appealing alternative, suggested by Dr. Larry Phillips (personal communication), would simply maximize the "bang for the buck," i.e., maximize $E(s)/C(s)$. This approach enjoys the advantage of bypassing the need to specify $DE$. However, as an ad hoc method, it lacks the normative credentials of the multiattribute approach we propose above.

## 11.9 Computation***

The exact optimal design in Equations (11.10)–(11.12) cannot generally be found in reasonable time since finding it is an NP-hard problem (Ko et al. 1995). That makes suboptimal designs necessary in problems of large or moderate size. Among the alternatives are exchange algorithms, in particular, the (DETMAX) procedure of Mitchell (1974a,b) cited by Ko et al.(1995). They also cite the greedy algorithm of Guttorp et al. (1993). At each step, the latter adds (or subtracts if the network is being reduced) the station that maximally improves the design's objective criterion. Ko et al. (1995) introduce a greedy plus exchange algorithm. The former starts with the complete set of all sites $K$, and first reduces it to the required number by the greedy algorithm. It then applies an exchange algorithm to the resulting greedy network $S$. Specifically, while possible, it successively exchanges site pairs $i \in S$ and $j \in K \backslash S$ so that the objective function at $(S \backslash i) \cup \{j\}$ exceeds its values at $S$. Finally, Wu and Zidek (1992) propose the idea of clustering the prospective sites into suitably small subgroups before applying an exact or inexact algorithm so as to get suboptimal designs that are optimal, at least within clusters.

**Exact Algorithms**

Exact algorithms for moderate-sized problems are available. The obvious one, complete enumeration, is used in this chapter and in Guttorp et al. (1993) in cases where $K$ is not too large. Ko et al. (1995) offer a more sophisticated branch-and-bound technique that we now describe. Using their notation, let $F$ denote a subcollection of sites that must be added to the network, $K$ being the collection of all sites. They seek to extend $F$ to some $S \supset F$ of sites that are to be added. Finally, if certain sites $K \backslash (E \cup F)$ are ineligible, their goal would entail finding

$$\nu(\Lambda_0, F, E, s) := \max_{S:\#(S)=s, F\subseteq S\subseteq E\cup F} |\Lambda_0[S,S]| \tag{11.13}$$

and the associated $S = S^{optimal}$ where "#(S)" stands for the number of sites in $S$ and in general, $\Lambda_0[E', F']$ refers to the submatrix of $\Lambda_0$ with rows $E'$ and columns $F'$. The algorithm requires a good initial design $S^*$ obtained by the greedy algorithm, for example. This design yields as a target to beat, the initial lower bound $LB := |\Lambda_0[S^*, S^*]|$. As well, it provides an initial active subproblems set $\mathcal{L} = \{\mathrm{L}\}$ consisting of just one element $L := (\Lambda_0, F, E, s)$ as well as a global upper bound $UB := b(\Lambda_0, F, E, s)$. For $b$, Ko et al. (1995) find

$$b(L) := |\Lambda_0[F,F]| \Pi_{i=1}^{s-f} \lambda_i(\Lambda_{0[E\cdot F]}), \tag{11.14}$$

where $\Lambda_{0[E\cdot F]} = \Lambda_0[E] - \Lambda_0[E,F]\Lambda_0[F,F]^{-1}\Lambda_0[F,E]$ and the $\{\lambda_i\}$ are the ordered eigenvalues of its matrix argument in decreasing order, $\lambda_1 \geq \cdots \lambda_{s-f}$ with $f = \#(F)$.

At a general step in the execution of the algorithm, the LB would correspond to the best design $S^*$ obtained to that step. At the same time, $\mathcal{L}$ would have a multiplicity of elements and the required global upper bound would be $UB := \max_{\mathrm{L}\in\mathcal{L}} b(\mathrm{L})$. $UB > LB$ suggests that $S^{optimal}$ has not been reached and new branches need to be explored in search of the optimum, i.e., new active subproblems need to be added to the $\mathcal{L}$. We do this by first deleting an active subprogram $(\Lambda_0, F', E', s)$ from $\mathcal{L}$ and then selecting a branching index $i \in E'$. Four (nondistinct) cases obtain and determine which subproblems to add. First, one of (i) $\#(F') + \#(E') - 1 > s$ or (ii) $\#(F') + \#(E') - 1 = s$ obtains. If (i), add $(\Lambda_0, F', E'\backslash i, s)$ to $\mathcal{L}$ and compute $b(\Lambda_0, F', E'\backslash i, s)$ (needed to find the new UB). If (ii), $S := F' \cup E'\backslash i$ is the only feasible solution. If $\Lambda_0[S,S] < LB$, S supplants the current $S^*$ and LB moves up to $LB := \Lambda_0[S,S]$. Next, one of (iii) $F' + 1 < s$ or (iv) $F' + 1 = s$ prevails. If (iii), add $(\Lambda_0, F' \cup \{i\}, E'\backslash i, s)$ to $\mathcal{L}$ and compute $b((\Lambda_0, C(\Lambda_0, F' \cup \{i\}, E'\backslash i, s), s)$. If (iv), $S = F' \cup \{i\}$; the only available feasible solution can supplant the current $S^*$ and move LB (computed as above) even higher. Finally, recompute the UB and determine whether the program has terminated with $UB \leq LB$. If not, delete another active subproblem, create new branches, and carry on as long as possible.

Ko et al. (1995) show their algorithm to be much quicker than complete enumeration. Jon Lee (personal communication) suggests that problems with site totals of about 80 can be routinely tackled. No doubt by improving UBs and methods of selecting the active problems for deletion, further increases in the algorithm's domain are possible. Nevertheless, for realistic continentwide redesign problems having hundreds or even thousands of prospective sites, exact optimization seems out of the question. Therefore, the finding of Ko et al. (1995) is encouraging in that the greedy/swap algorithm described above often produced the exact optimum, where the latter is computable.

The branch-and-bound algorithm can be extended in various ways. Bueso et al. (1998) extend it to the case where observations are made with error and the goal is the prediction not only of responses at ungauged sites but

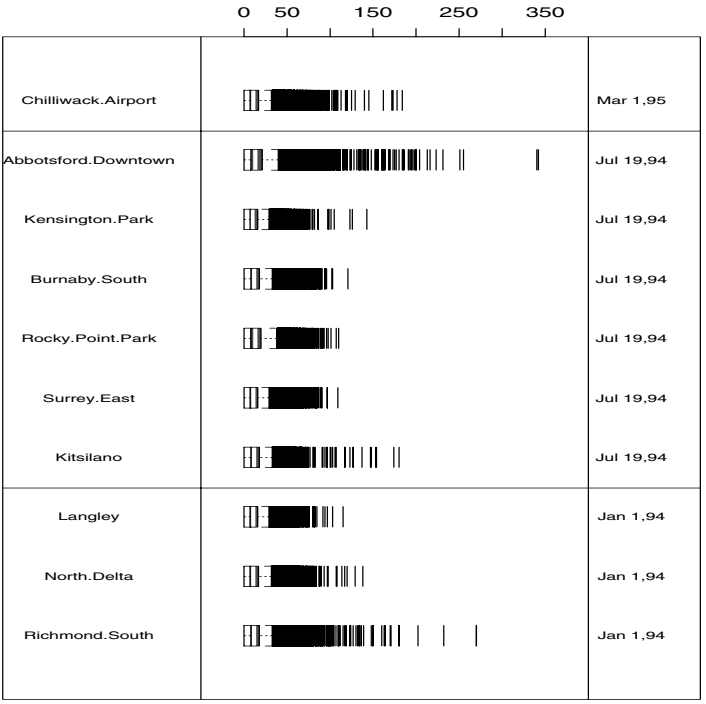


Fig. 11.1: Boxplot of hourly $PM_{10}$ levels ($\mu g/m^3$) at ten monitoring sites in Greater Vancouver and their start-up times.

those at the gauged sites as well. Lee (1998) extends it to incorporate linear constraints (e.g., limiting cost). His approach differs from the approach of Zidek et al. (2000) where cost is also incorporated.

## 11.10 Case Study

This section illustrates the use of the above entropy design theory by redesigning GVRD's $PM_{10}$ network. That network had ten stations measuring hourly $PM_{10}$ levels with different start dates, resulting in a staircase data pattern. Each step of the staircase consists of stations having the same starting time. Figure 11.1 shows the names of the stations along with their start dates and the boxplots of the hourly $PM_{10}$ measurements.

**Add Six New Sites!**

Our objective: augment the existing network with an optimal subset of six stations among 20 potential sites. Locations of the existing stations and potential sites are displayed in Figure 11.2. Since the hour to hour $PM_{10}$ levels are highly dependent, a 24-dimensional vector representing hourly $PM_{10}$ measurements for each day obtained from the detrended series is considered. The staircase pattern and the multivariate responses fit in with the general Bayesian hierarchical model setting in Chapter 10.

**Preliminaries**

The trend for the log-transformed hourly $PM_{10}$ levels is modeled with seasonal components, hourly, and daily effects, and meteorological covariates. We do not go into detail, but briefly, the seasonal components are captured by sine and cosine functions for monthly, semi-annual, and annual cycles. Since the trend model is linear in its coefficients, they like the other coefficients are fitted by regression analysis. In a similar way, we incorporate meteorological data, including "visibility index," "sealevel pressure," "dewpoint temperature," "wind speed," "rain," and "relative humidity."

After removing the trend, we need to filter out day to day autocorrelation. This we do with the help of a standard (multivariate) autogressive time-series model of order one, i.e., MAR(1) model. Standard statistical software packages include routines for fitting such models. After fitting that model we are left with approximately *whitened*, i.e., unautocorrelated, residuals.

**On to Residuals and Design**

Those residuals are multivariate responses having a monotone missing data pattern. The predictive distribution for the multivariate residuals at all locations of interest, 10 existing and 20 new potential sites, is a product of conditional matric-$t$ distributions as given by the results in Chapter 10.

The hyperparameters are estimated using the moment method proposed by Kibria et al. (2002). Table 11.1 shows the estimated hypercovariance matrix between gauged stations.

The residual hypercovariance matrix between potential (ungauged) sites conditional on the existing sites $\Lambda_0$ is estimated using the Sampson and Guttorp (SG) method (Sampson and Guttorp 1992) described in Chapter 6 and based on the estimated hypercovariance matrix among the gauged sites.

**Results!**

Figure 11.3 demonstrates the actions of the SG method in this application. The right panel shows the corresponding D-space coordinates resulting from applying the mapping function to a biorthogonal grid in G-space. The left

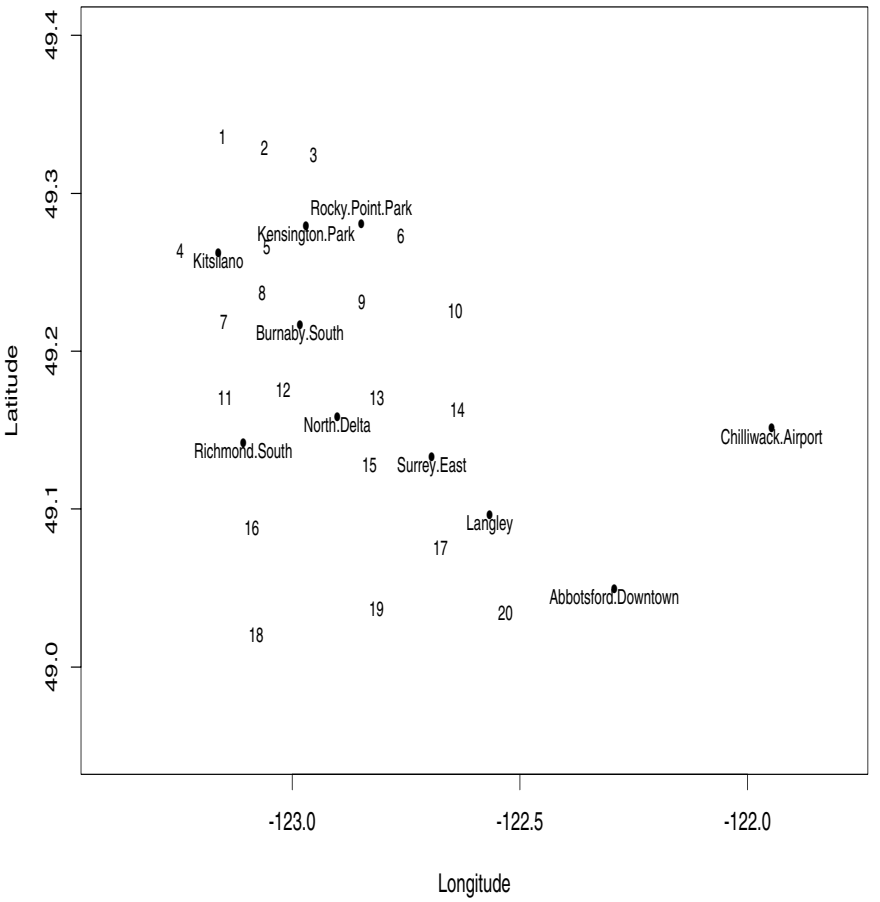


**Fig. 11.2.** $PM_{10}$ Monitoring stations and potential sites.

panel shows the fitted variogram in D-space. The figure shows a good fit for the variogram model using this mapping function with spline smoothing parameter of 2. Users specify this built-in map smoothing parameter that controls the distortion between the G-space and the D-space. This feature ensures that the grid is not folded in the D-space and hence maintains the spatial interpretability of the correlations; that is, correlations are reflected in intersite distances in dispersion space. The deformation on the right panel

| | |
|---|---|
| Chilliwack airport | 75 46 38 40 34 38 27 40 36 29 |
| Abbotford downtown | 46 79 36 46 33 41 29 40 38 31 |
| Kensington Park | 38 36 70 55 47 42 38 38 46 34 |
| Burnaby South | 40 41 55 77 48 49 43 43 55 43 |
| Rocky Point Park | 34 33 47 48 64 39 36 34 41 33 |
| Surrey East | 38 41 42 49 39 64 34 45 47 36 |
| Kitsilano | 27 29 38 43 36 34 59 30 37 41 |
| Langley | 40 40 38 43 34 45 30 64 41 31 |
| North Delta | 36 38 46 55 41 47 37 41 72 39 |
| Richmond South | 29 31 34 43 33 36 41 31 39 65 |

Table 11.1: Estimated hypercovariance matrix at existing stations after multiplying entries by 100.

indicates the nonstationarity of the field. Failure to capture nonstationarity results in a suboptimal design as illustrated in the analysis below.

The panels in Figure 11.3 can be used to estimate spatial correlations between any points in the G-space, e.g., by first identifying the points in D-space using the grid, then measuring the distance in D-space between them, and finally applying the fitted variogram to the distance to estimate their spatial correlations. The residual hypercovariance matrix among the ungauged site conditional on existing stations $\Lambda^{[u]}$ is estimated accordingly.

The predictive distribution of the responses at ungauged locations follows a matric $t$ distribution and the entropy criterion is to find an *add* subset of six locations that maximizes $\log|\Lambda_0|^{add}$. Applying the entropy criterion yields the optimal subset of six sites {Sites: 10, 12, 16, 18, 19, 20} among the 20 potential sites, to augment the existing network. The locations of the selected sites are depicted in Figure 11.4 along with the locations of existing stations and potential sites, the latter accompanied by their ranking based on their estimated hypervariances (i.e., the diagonal element of $\Lambda_0$).

**Optimal Design!**

The optimum solution seems sensible in that five of the six sites {Sites: 10, 16, 18, 19, 20} have the five largest variances and are generally far away from existing stations. However, note that the sixth selected site, Site 12, has a smaller estimated hypervariance than two unselected ones (Sites: 14 and 17). The trade-off between variance and correlation with nearby stations is demonstrated here. Site 14 is not selected in spite of its having a large estimated variance because it is closer to existing stations than Site 12. Furthermore, Site 14 is located in a region of stronger spatial correlation than that of Site 12, as indicated by the stretching in the region containing Site 12 in Figure 11.3's right panel. The nonstationarity of this field also plays an important role in the selection of Site 12 over Site 17. The two sites are roughly the same distance from existing stations, Site 17 having larger estimated hypervariance;

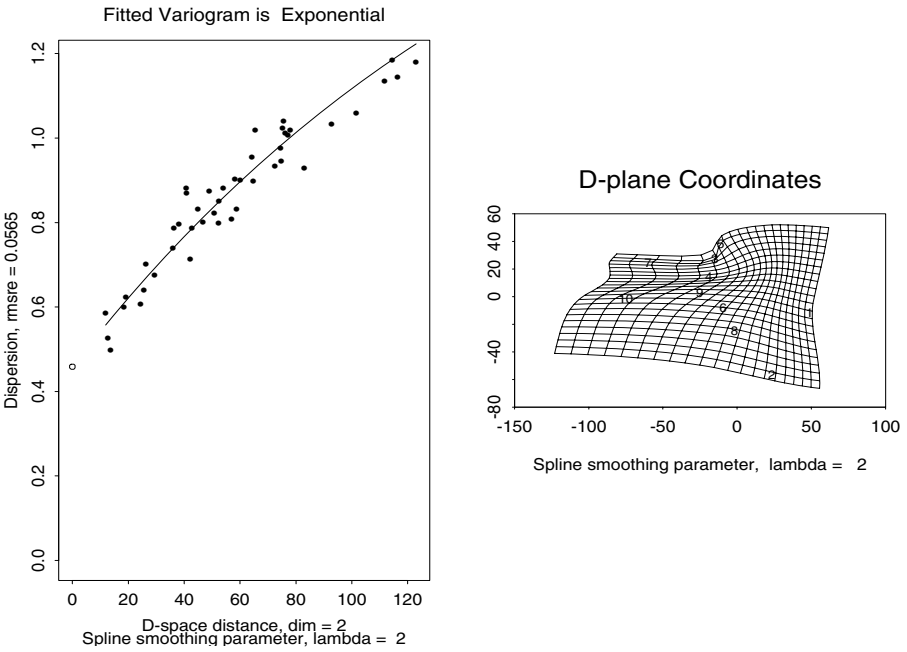


Fig. 11.3: Transforming the geographic plane to dispersion space (right panel) and fitting a variogram to the empirical variogram over dispersion space for Vancouver's hourly $PM_{10}$ field.

however, the spatial correlation is weaker at Site 12 than at Site 17. This fact can be clearly seen in the right panel of Figure 11.3 where the region containing Site 17 does not show any stretching, in comparison with the region containing Site 12 showing more stretching and hence less spatial correlation for the same distance in G-space.

## 11.11 Pervasive Issues***

We now need to step back a bit and view the design problem from a more general perspective in order to see some important issues not revealed hitherto.

### State-Space Model Framework

To do this, we formulate the problem in terms of a general state-space model, a very important tool in the space–time process modeling. That model has three components: (1) the measurement model, (2) the process model, and (3)

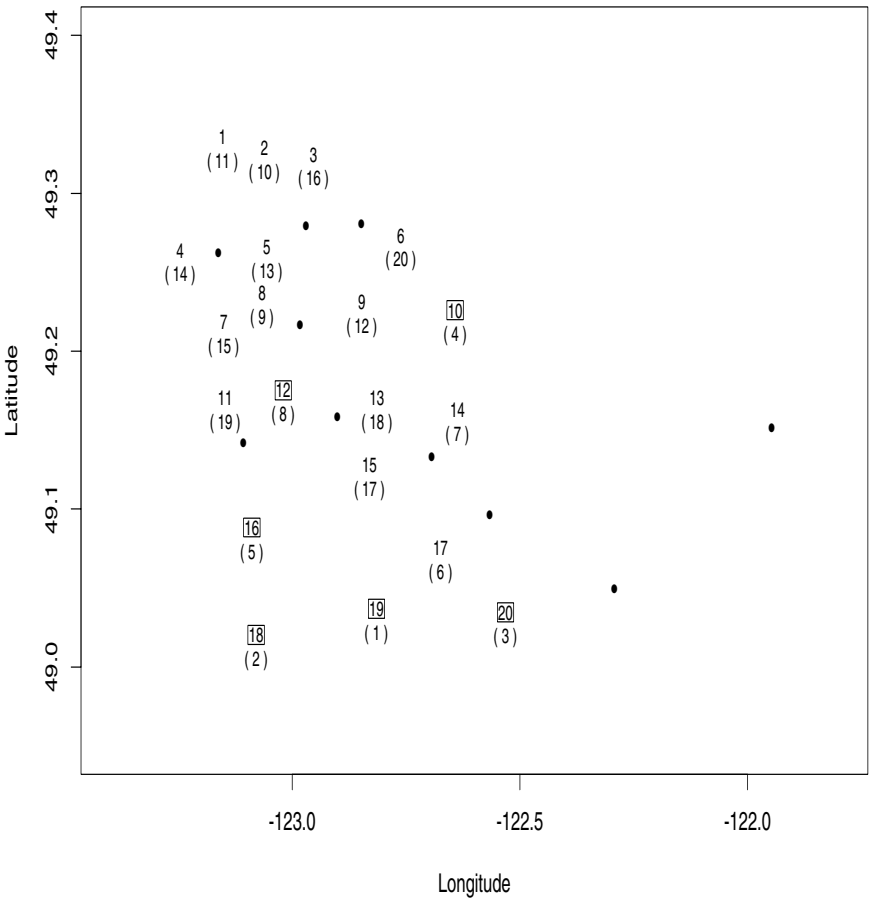


Fig. 11.4: Locations of existing network (point) and potential new sites (number) with their rank based on estimated variance (in brackets); selected sites are marked with squares.

the parameter model. Let $\mathbf{X}_t$ be the $(1 \times n_t)$-dimensional vector of responses observed at time t. These measured responses are related to $(1 \times (u+g)q)$-dimensional state vectors $\mathbf{S}_t$ by the *measurement model*:

$$\mathbf{X}_t = \mathbf{S}_t\mathbf{H}_t + \varepsilon_t, \quad t = 1, \ldots, (n+1). \tag{11.15}$$

This model is a composition of two others. The first is

$$\mathbf{X}_t = (\mathbf{Y}_t + \varepsilon_t^1)\mathbf{F}_t^1, \quad t = 1, \ldots, (n+1), \tag{11.16}$$

where $\mathbf{Y}_t$ is of dimension $1 \times (u+g)r$ while $\mathbf{F}_t^1$ is a $((u+g)r \times n_t)$-dimensional *design vector* of ones and zeros that determines which of the responses is measured. In fact, $\mathbf{F}_t^1$ will generally be random, designating the data missing both at random as well as by design. However, it is assumed that the former are missing for reasons ancillary to the process of their generation and measurement. Thus, they can be treated as fixed.

### Design Problem

The design problem discussed at the beginning of this section is that of selecting $\mathbf{F}_{n+1}^1$. In other words, optimally partition the vector of all measurements of pollutants and sites that could be taken at time $n+1$, into those that are actually taken and those not. The latter therefore remain uncertain along with all the parameters and latent variables in the process at that time. Following the earlier reasoning in this section, the objective is to minimize the residual uncertainty about all these uncertain quantities by selecting $\mathbf{F}_{n+1}^1 = \mathbf{F}_{n+1}^{1opt}$ so as to maximize MEAS; that is,

$$\mathbf{F}_{n+1}^{1opt} = \arg\min_{\mathbf{F}_{n+1}^1} H(\mathbf{X}_{n+1}|\mathbf{X}^n), \tag{11.17}$$

superscripts like $n$ denoting here and in the sequel all items up to and including those up to that specified time. The resulting design will change dynamically as $n$ increases since other practical considerations are ignored including cost in this section.

The second model needed to reach (11.15) is

$$\mathbf{Y}_t = \mathbf{S}_t\mathbf{F}_t^2 + \varepsilon_t^2, \quad t = 1, \ldots, n,$$

$\mathbf{F}_t^2$ relating responses measured and unmeasured to the state-space vectors $\mathbf{S}_t$. Generally $F_t^2$, unlike the design matrices, will involve unknowns. Finally, the so-called $((u+g)q \times n_t)$ *output matrix* $\mathbf{H}_t$ is just the composition of the $((u+g)q \times (u+g)r)$ state transition matrix $\mathbf{F}_t^2$ with the $((u+g)r \times n_t)$ measured response output matrix $\mathbf{F}_t^1$. The measurement error vectors $\varepsilon_t = (\varepsilon_t^2 + \varepsilon_t^1)\mathbf{F}_t^1$ resulting from the combination of these two models are assumed to have zero mean, to have covariance matrix $\mathbf{F}_t^{1\prime}\Sigma\mathbf{F}_t^1$ (assumed known for the purposes of this section), and to be independent of each other as well as other uncertain elements of the process and measurement models. Note that $\Sigma = \Sigma_{\varepsilon_t^2} + \Sigma_{\varepsilon_t^1}$ combines the spatial covariance of the responses with measurement noise.

### State Evolution Equation

We adopt the following class of process models,

$$\mathbf{S}_t = \mathbf{S}_{t-1}\theta_t + \nu_t, \quad t = 1, \ldots, (n+1) \tag{11.18}$$

where the process noise variables $\nu_t$ have zero means covariances $\Sigma_\nu$ and are independent of each other as well as of the other random process vectors. Returning to the general case, for the purposes of this section both $\theta$ as well as the covariances $\Sigma_\nu$ and $\Sigma$ are assumed known. However, a more realistic approach such as that in the next section would add a parameter model that specifies prior distributions for these components.

**Entropy Approach**

Now assume all measurement and state-space processes above have a multivariate Gaussian distribution (possibly after an appropriate transformation). With these assumptions the entropy in Equation (11.17) can be explicitly evaluated. To that end let

$$\begin{aligned} \hat{\mathbf{S}}_t &= E(\mathbf{S}_t|\mathbf{X}^t), \\ \hat{\mathbf{P}}_t &= Cov(\mathbf{S}_t|\mathbf{X}^t) \text{ so that,} \\ \mathbf{S}_t|\mathbf{X}^t &\sim N_{(u+g)q}(\hat{\mathbf{S}}_t,\ \hat{\mathbf{P}}_t) \end{aligned} \tag{11.19}$$

for all $t = 1, \ldots,\ (n+1)$. We now find the conditional distribution,

$$\mathbf{X}_{t+1}|\mathbf{X}^t \sim N_{n_{t+1}}[E(\mathbf{X}_{t+1}|\mathbf{X}^t), Cov(\mathbf{X}_{t+1}|\mathbf{X}^t)]$$

needed to compute the entropy. As a first step we find the conditional distribution $\mathbf{S}_{t+1}|\mathbf{X}^t \sim N_{pq}[E(\mathbf{S}_{t+1}|\mathbf{X}^t), Cov(\mathbf{S}_{t+1}|\mathbf{X}^t)]$. First,

$$\begin{aligned} E(\mathbf{S}_{t+1}|\mathbf{X}^t) &= E(E[\mathbf{S}_{t+1}|\mathbf{S}_t]|\mathbf{X}^t) \\ &= E(\mathbf{S}_t\theta|\mathbf{X}^t) \\ &= \hat{\mathbf{S}}_t\theta. \end{aligned} \tag{11.20}$$

Similarly,

$$\begin{aligned} \mathrm{Cov}(\mathbf{S}_{t+1}|\mathbf{X}^t) &= \mathrm{Cov}(E[\mathbf{S}_{t+1}|\mathbf{S}_t]|\mathbf{X}^t) \\ &\quad + E(\mathrm{Cov}[\mathbf{S}_{t+1}|\mathbf{S}_t]|\mathbf{X}^t) \\ &= \mathrm{Cov}(E[\mathbf{S}_t\theta_t|\mathbf{X}^t]) \\ &\quad + E(\Sigma_\nu) \\ &= \theta_t'\hat{\mathbf{P}}_t\theta_t + \Sigma_\nu. \end{aligned} \tag{11.21}$$

Then

$$\begin{aligned} Cov(\mathbf{X}_{t+1}|\mathbf{X}^t) &= Cov([\mathbf{S}_{t+1}\mathbf{F}^2_{t+1} + \varepsilon^2_{t+1} + \varepsilon^1_{t+1}]\mathbf{F}^1_{t+1}|\mathbf{X}^t) \\ &= \mathbf{H}'_{t+1}Cov(\mathbf{S}_{t+1}|\mathbf{X}^t)\mathbf{H}_{t+1} + \mathbf{F}^{1\prime}_{t+1}\Sigma\mathbf{F}^1_{t+1} \\ &= \mathbf{H}'_{t+1}[\theta_t'\hat{\mathbf{P}}_t\theta_t + \Sigma_\nu]\mathbf{H}_{t+1} + \mathbf{F}^{1\prime}_{t+1}\Sigma\mathbf{F}^1_{t+1}, \end{aligned} \tag{11.22}$$

by Equation (11.21). Finally, from standard theory for the Gaussian distribution, it follows that the entropy to be maximized, $H(\mathbf{X}_{n+1}|\mathbf{X}^n)$, is, apart from irrelevant constants, the logarithm of the determinant of the (conditional) covariance matrix:

$$|\mathbf{H}'_{n+1}[\theta'_{n+1}\hat{\mathbf{P}}_n\theta_{n+1} + \Sigma_\nu]\mathbf{H}_{n+1} + \mathbf{F}^{1\prime}_{n+1}\Sigma\mathbf{F}^1_{n+1}|. \tag{11.23}$$

The logarithm ensures that the optimal design will remain invariant under rescaling; multiplication of the normal density by the Jacobean of the transformation simply becomes an additive shift. Recalling that $\mathbf{H}_{n+1} = \mathbf{F}^2_{n+1}\mathbf{F}^1_{n+1}$, the optimal design matrix is found by finding the maximal $n_{n+1} \times n_{n+1}$ subdeterminant, that is, *generalized subvariance* of the covariance

$$\mathbf{F}^{2\prime}_{n+1}[\theta'_{n+1}\hat{\mathbf{P}}_n\theta_{n+1} + \Sigma_\nu]\mathbf{F}^2_{n+1} + \Sigma. \tag{11.24}$$

The apparent simplicity of the state-space model above disguises the difficulty of its formulation in specific cases as seen in the following example.

### Example: AR(3) Model

*Example 11.2. The AR(3) model*
Consider the case of Li et al. (1999) where an autoregressive model of order three obtains at every site $j = 1, \ldots, (u + g)$. The number of response species is $r = 1$. There,

$$\begin{aligned}\mathbf{Y}_{tj} - \beta_j\mathbf{Z}_t = {} & [\mathbf{Y}_{(t-1)j} - \beta_j\mathbf{Z}_{t-1}]\rho_{1j} + [\mathbf{Y}_{(t-2)j} - \beta_j\mathbf{Z}_{t-2}]\rho_{2j} \\ & + [\mathbf{Y}_{(t-3)j} - \beta_j\mathbf{Z}_{t-3}]\rho_{3j} + \varepsilon^2_{tj},\end{aligned}$$

where $r = 1$, $\beta_j : 1 \times l$ is a vector of response-dependent trend coefficients, the $\mathbf{Z}_t : \ l \times r$ are ancillary (and hence fixed) covariates with the same value at all sites $j$, and $\rho_{ij} : r \times r, \ i = 1, 2, 3$ are the autoregressive coefficient matrices. Equivalently, with an abuse of notation,

$$\mathbf{Y}_{tj} = \mathbf{Y}_{(t-1)j}\rho_{1j} + \mathbf{Y}_{(t-2)j}\rho_{2j} + \mathbf{Y}_{(t-3)j}\rho_{3j} + \beta_j\mathbf{Z}_t + \varepsilon^2_{tj}, \tag{11.25}$$

where $\mathbf{Z}_t$ now stands for $\mathbf{Z}_t - \mathbf{Z}_{t-1}\rho_{1j} - \mathbf{Z}_{t-2}\rho_{2j} - \mathbf{Z}_{t-3}\rho_{3j}$. A standard reformulation of this model would have $\mathbf{Y}_{tj} = \mathbf{S}_{tj}\mathbf{F}^2_{tj} + \varepsilon^2_{tj}$, where

$$\mathbf{S}_{tj} : 1 \times (l + 3r) = [\beta_j, \mathbf{Y}_{(t-1)j}, \mathbf{Y}_{(t-2)j}, \mathbf{Y}_{(t-3)j}]$$

and

$$\mathbf{F}^2_{tj} : (l + 3r) \times r = \begin{pmatrix} \mathbf{Z}_t \\ \rho_{1j} \\ \rho_{2j} \\ \rho_{3j} \end{pmatrix}. \tag{11.26}$$

However, this dynamic state-space model fails since space and time are inseparable while the residual independence assumption in Equation (11.18) proves invalid. That problem is observed by Zidek et al. (2002) who adopt a different approach, one that does not model fine-scale autocovariance structures (and, as a bonus, avoids the risk of misspecifying them). Instead, they adopt the 24-hour site response vector as a basic building block, i.e., day as a temporal unit. Since $r$ species are responding each hour at each site ($r = 1$ in

the specific case under consideration here), that response vector is $(24r \times 1)$-dimensional. The resulting vector series at each site could then be modeled by a multivariate AR model (MAR; used below in Section 11.10). Moreover, in the case of Li et al. (1999), the $(24r \times 24r)$-dimensional autoregression coefficient matrices depend little on $j$ so that $\rho_{ij} = \rho_i$ for all $j$ proves a tenable assumption. A MAR of order 1 would yield

$$\mathbf{Y}_{tj} = \mathbf{Y}_{(t-1)j}\rho_1 + \beta_j \mathbf{Z}_t + \varepsilon^2_{tj}. \tag{11.27}$$

With an appropriate change in dimensions it can be written

$$\mathbf{S}_t : 1 \times (u+g)(l+24r) = [\mathbf{S}_{t1}, \ldots, \mathbf{S}_{tp}];$$

$$\mathbf{F}^2_t = \begin{pmatrix} \mathbf{F}^2_{t1} & 0 & \cdots & 0 \\ 0 & \mathbf{F}^2_{t2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \mathbf{F}^2_{tp} \end{pmatrix}.$$

For the autoregressive model in Equation (11.27), obtain

$$\mathbf{S}_{tj} = \mathbf{S}_{(t-1)j}\theta_t + \nu_{tj}, \quad j = 1, \ldots, (u+g), \tag{11.28}$$

where

$$\theta_{tj} = \begin{pmatrix} I_l & \mathbf{Z}_{t-1} & 0 & 0 \\ 0 & \rho_1 & I_{24r} & 0 \end{pmatrix}. \tag{11.29}$$

The matrices in Equation (11.29) can be combined to determine $\theta_t : (u + g)(l + 24r) \times (u + g)(l + 24r)$, namely,

$$\theta_t = \begin{pmatrix} \theta_t & 0 & \cdots & 0 \\ 0 & \theta_t & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \theta_t \end{pmatrix}.$$

**Remarks**

**3.a** A generalized subvariance obtained from Equation (11.24) will tend to be small when either its columns or rows are nearly collinear; i.e., the associated responses are highly associated. That can occur because of strong spatial association between sites, as expressed through the intrinsic component of variation $\Sigma$ in that equation. Or it may derive from strong temporal association as expressed through the remaining terms, i.e., extrinsic component. In any case, sites will tend to be omitted from the network, either because they are predictable from other sites in the present, or from measurements made in the past.

**3.b** Typically in applications, data to time $n$ will derive from a permanent set of monitoring sites and the design goal will be judicious augmentation of that network. To address this situation, let us represent the (symmetric) matrix of Equation (11.24) more simply as

$$\Xi = \begin{pmatrix} \Xi_{uu} & \Xi_{ug} \\ \Xi_{gu} & \Xi_{gg} \end{pmatrix}, \tag{11.30}$$

where $\Xi_{gg}$ represents covariance at the permanent sites. Then using a familiar identity for matrix determinants, the design optimization problem associated with Equation (11.24) becomes that of adding a fixed number of sites to the network with maximal generalized sub-variances of $\Xi_{u \cdot g} = \Xi_{uu} - \Xi_{ug}\Xi_{gg}^{-1}\Xi_{gu}$ of appropriate dimension.

**3.c** Our formulation of the design problem through $F_t^1$ allows us to dynamically expand or contract the monitoring network at each successive time, the optimal basis for making alterations being expressed in Remark 3.a. One can conceive of hypothetical cases where dynamically changing networks might be desirable. For example, mobile monitors might be used to track the radiation plume generated by the failure of a nuclear power generator. Or in military operations, they might be used to follow hazardous agents released in the battlefield. However, such designs would generally be impractical because of such things as their high operating or administrative costs.

Through $\hat{\mathbf{P}}_n$, the extrinsic component of the design criterion above is a function of past data. Moreover, that component rather than the intrinsic component may point to the deletion of sites at time $t = n + 1$ whose responses are well predicted from past data including those which they produced. Their deletion will eliminate the very source of information that justified their removal in the first place. Thus in time, the quality of the network insofar as it provides information about nonmonitored sites (including some of those that were removed from the network), could degrade.

This suggests a need for a practical compromise and acceptance of a suboptimal permanent design after time $t = n$. That compromise may be achievable by filtering the data and relying primarily on the intrinsic component of covariance.

*Example 11.3. Example 11.2 continued*

To arrive at an appropriate compromise design criterion, transform the responses as $\mathbf{Y}_{tj}^* = \mathbf{Y}_{tj} - \mathbf{Y}_{(t-1)j}\rho_1 = \beta_j \mathbf{Z}_t + \varepsilon_{tj}^2$. Then the design at time $t = n + 1$ will not depend on past measurements as predictors of current responses. These transforms have an added benefit in that they eliminate the autoregression matrices from the design criterion, simplifying technical analysis when they are unknown. However, in practice this would require that they be well estimated and not subject to much uncertainty.

**3.d** Another issue that must be confronted arises from the uncertainty about parameters such as the $\theta$s and the covariances that were assumed known. The result of incorporating that additional uncertainty makes the conditional distribution of $\mathbf{X}_{n+1}|\mathbf{X}^n$ non-Gaussian. In fact, that distribution will typically not have a tractable form, making a convenient analytical representation of the entropy impossible. Evaluating that entropy numerically is not a practical option since the combinatorial design optimization problem is computationally intensive. Finding it is generally very difficult for realistically large values of $u + g$. Adding the additional burden of numerically evaluating the entropy at each iteration can make the burden prohibitively large.

**3.e** The measurement noise represented by $\Sigma_{\varepsilon_t^2}$ could conceivably vary in magnitude from response to response in extreme cases. In fact, it could dominate the selection of an optimum design. For this and other reasons, it might be argued that the optimum design should not be selected such as that above to include its capacity to reduce uncertainty about measurements that could have but have not been taken. Instead the goal could be to maximally reduce uncertainty about $\mathbf{Y}_{n+1}$ rather than $\mathbf{X}_{n+1}$ given measurements to time $n$. These objectives would be essentially equal when measurement noise is negligible. The design objective criterion in that case would obtain from that in Equation (11.24) after subtracting the measurement noise covariance.

## 11.12 Wrapup

Other approaches to spatial sampling design have been developed that do not fit neatly into the design taxonomy used in this chapter. Richard Smith in his 2004 Hunter lecture describes two approaches that seek to compromise between the prediction of random fields and the estimation of their model coefficients. One of these is due to Zhu (2002) who also presents an annealing optimization algorithm. (See also Zhu and Stein (2005) as well as Zhu and Stein (2006).) The other is due to Zimmerman (2004). He shows, in particular, that optimal designs for prediction (with known covariances) and designs for estimating covariance parameters are antithetical, pointing anew to the problem posed by a multiple objectives.

In this chapter we have developed a hierarchical Bayesian framework for redesigning an existing monitoring network, stimulated in part by the multiple objectives problem. We realistically allow for the possibility of staggered start-up times for current stations, i.e., staircase data patterns. As well, we take the lack of a well-defined design objective as a given. That leads us to adopt the generic objective of minimizing the entropy of the posterior probability distribution of the quantities of interest. Roughly speaking, the new network stations would be those with highly unpredictable response vectors, either because of their lack of dependence on the other stations, or because of their

high intrinsic variability. Our results indicate that the proposed approach yields sensible designs that capture the nonstationarity of the spatial field.

A number of practical issues need to be addressed in implementing the entropy method. First it must be possible to transform and prefilter the data to validate the distributional assumptions we make. Computation is a major practical consideration. We deliberately chose a relatively small number of potential sites in our case study to make complete combinatorial optimization feasible. But the numerical problem rapidly becomes overwhelming as the existing network increases in size.

This relates directly to an issue raised in Section 11.2. There it was noted that the uncertainty about ungauged sites can be reduced not only by borrowing information from current measurements at gauged sites but, to a lesser extent, from previous such measurements as well, at least when the autocorrelation in the individual series is sufficiently strong. However, incorporating that component of the model seems to lead to an intractable entropy calculation and in turn to the computational problem indicated in the last paragraph. Further work is needed to address this issue.

Finally, it should be recognized that in practice "good" rather than "optimal" designs are needed and optimal designs such as those in this chapter must be considered as tentative proposals susceptible to modification depending on the circumstances prevailing in the context of their implementation. These optimal designs may well be valuable starting points, however, since they can be explicated in terms of their axiomatic underpinnings and proposed changes to these optimal designs can be interpreted in terms of the axioms. This can provide a degree of confidence and clarity in the typically complex situation confronting a designer. However, the entropy approach, founded on a coherent normative theory, should make the resulting designs defensible in an operational context.

We should emphasize that as a compromise, an entropy optimal design will not yield optimal designs for specific objectives. In fact, it would be of limited use when interest lies in monitoring the extreme values of the time-series of responses at the spatial sites. (The reasons become apparent in the following chapter.) That interest stems from the fact that risk can often be the result of environmental space–time processes generating an extreme value such as a 100-year flood, for example. However, monitoring and modeling fields of extreme values presents very challenging problems without an entirely satisfactory solution as this book is being written. We describe some of those problems in the next chapter.

# 12

# Extremes

*We have been fortunate so far. But we have seen that when winds fail to blow, the concentrations of poisonous clouds over our cities can become perilous.*

*Lyndon B Johnson, "To Renew a Nation": Special address to Congress.*

This chapter describes some of the challenges presented by fields of extremes such as that described by Johnson. No book on environmental processes would be complete without some discussion of this difficult subject. After all, much environmental risk, the topic of Chapter 13, derives from extremely large (or small) values generated by such things as wind, rain, and air pollution. The latter motivates much of the discussion in this chapter. However, most of the material applies to environmental space–time fields in general.

Air pollution monitoring seems to have begun as a result of the worryingly high association between air pollution and human morbidity (or mortality) established by a great many studies. Regulation and control policies were instituted in conjunction with these monitoring programs.

### Extremes and Air Quality Standards

Consider, for example, the AQS (air quality standards) set for criteria responses in the United States, specifically $PM_{2.5}$, i.e., fine airborne particulates. They require that 3 year averages of annual 98th percentiles of daily means of hourly concentrations must lie below 65 $\mu$g m$^{-3}$ at each of an urban area's monitoring network. In fact, this criterion will be exceeded with some regularity, so it is not as extreme, relatively speaking, as a 1000-year flood level. But it does lie well above the mean level of the $PM_{2.5}$ field and it does exemplify the complexity of the kinds of extreme responses that might be encountered in practice.

### Adequacy of Monitoring Programs?

How well are such extremes actually monitored? Generally resource limitations mean few sites are monitored in most urban areas, just ten sites for $PM_{10}$ and only two for $PM_{2.5}$, for example, in the Greater Vancouver Regional District (GVRD) although it covers a large geographical area. Thus,

in terms of their intended function of protecting human health by detecting noncompliance, it is by no means certain these networks have enough monitors.

Incidentally, a low density of monitors does not in itself demonstrate the inadequacy of those networks. Indeed, if unmeasured extremes are well predicted by data obtained from monitoring, our uncertainty about the field would be eliminated or greatly reduced. In fact, our studies of particulate fields over Philadelphia and London indicate that for those cities, at least, the entire field can be predicted pretty well from just a few well placed monitors.

However, that is far from true in Seattle and Vancouver, BC (as we show farther along in this chapter). Moreover, we should not complacently expect urban networks to be adequate for monitoring extremes. After all, classical design approaches such as those described in Chapter 11 emphasize inference about the response field, not its extremes. Similarly, even though the entropy approach described in that chapter was intended to get around the need to specify particular design objectives, that approach can lead to unsatisfactory designs for monitoring extremes (as an analysis later in this chapter shows).

Our discussion of the complex topic of this chapter begins by reviewing approaches to modeling environmental space–time fields of extremes. We begin with extreme value theory itself and some difficulties that theory encounters in processes that may involve hundreds of sites. These deficiencies point to the need for the alternative approach we describe along with its strengths and weaknesses.

## 12.1 Fields of Extremes

Since extremes, rare by definition, are seldom measured, generally acceptable models are needed instead. The search for them begins with plausible judicious assumptions that point to a fairly specific process model. The ultimate goal: an extreme value distribution with just a few parameters that can be estimated from the available (nonextreme) data. However, the ultimate test of its success is user acceptance for such things as structural design against extreme winds, for example. For single-site models, that test has certainly been passed.

### 12.1.1 Theory of Extremes

This section begins with the classical theory that concerns processes at a single spatial site.

#### Single Sites

The systematic study of extreme values began with the seminal paper of Fisher and Tippett (1928).

**The Fisher–Tippett "Trinity"**

Assuming $X_1, X_2, \ldots, X_n$ are independently and identically distributed (iid) random variables with distribution $F$, they proved the distribution of the maximum $M_n = \max\{X_1, X_2, \ldots, X_n\}$ converges to exactly one of three distributions as $n \to \infty$. More precisely,

$$P(\frac{M_n - b_n}{a_n} \leq x) \to H(x), \quad \text{as n} \to \infty, \tag{12.1}$$

where $a_n$ and $b_n$ are normalizing constants that keep *H(x)* from being degenerate. Their celebrated result tells us $H$ must be one of three types, each involving a parameter $\alpha > 0$:

1. (Gumbel):

$$H(x) = \exp\{-\exp(-x)\}, \quad -\infty < x < \infty;$$

2. (Fréchet):

$$H(x) = \begin{cases} 0 & \text{if x} < 0 \\ \exp(-x^{-\alpha}) & \text{if } 0 < \text{x} < \infty; \end{cases}$$

3. (Weibull):

$$H(x) = \begin{cases} \exp\{-(-x)^{\alpha}\} & \text{if } -\infty < \text{x} < 0, \\ 1 & \text{if x} > 0. \end{cases}$$

**Ensuing Development**

Ensuing development produced a very general theory (Gumbel 1958; Leadbetter et al. 1983; Coles 2001; Embrechts et al. 1997). Moreover, the Fisher–Tippett result assumed the role of a paradigm in that development. Their three types came to be combined in the *Generalized Extreme Value* (GEV) distribution, with cumulative distribution function:

$$H(x) = \begin{cases} exp\left[-\left\{1 + \xi\left(\frac{(x-\mu)}{\sigma}\right)\right\}^{-1/\xi}\right], & 1 + \xi(x-\mu)/\sigma > 0,\ \xi \neq 0 \\ exp\left\{-exp\left[-\frac{(x-\mu)}{\sigma}\right]\right\} & \xi = 0. \end{cases}$$

Alternatives emerged. One of these was the class of *peak over threshold (POT)* models that avoid fixed intervals and instead, look at exceedances over high thresholds. The number of such exceedances comprises the response of interest and it may be modeled by a nonhomogeneous Poisson process. Another approach models the response conditional on its exceeding a specified threshold. Its conditional distribution is the generalized Pareto distribution (Pickands 1975; Davison and Smith 1990) with right-hand tail,

$$H(x) = 1 - \lambda\left\{1 + \frac{\xi(x-u)}{\sigma}\right\}_{+}^{-1/\xi}, \quad x > u$$

for parameters $\lambda > 0$, $\sigma > 0$, and $\xi \in (-\infty, \infty)$. Finally, there is a class of models developed by hydrologists using the probability weighted moments

(PWM) approach. However, the latter seems too complex for our purposes (Katz et al. 2002). More details about these approaches can be found in Smith (2001) and Fu et al. (2003).

**The Adequacy of Models**

Turning to more general issues, the use of the extreme value distribution models will always be questioned when grave risks are involved. An obvious concern would be the accuracy of any asymptotic approximations, for example, in the Fisher–Tippett results. Even Tippett seemed concerned on this point. Gumbel (1958) quotes him as saying that even $n = 1000$ is insufficiently large to guarantee reasonable accuracy when modeling distribution tail areas smaller than 0.05. In fact, Zidek et al. (1979), citing this and other concerns, devise an alternative approach for finding structural design criteria (that they successfully used to produce codes for long-span highway bridges).

Another source of concern stems from the (unrealistic) assumption of independence in the sequence of process responses. At issue is the robustness of the limit distributions when the condition fails. That issue arises, for example, when extremes are used for detecting a trend in atmospheric temperature. There we would anticipate underlying trends expressing themselves through clusters of nonindependent process responses exceeding a threshold or even by a sequence of such clusters separated by at least a fixed number of consecutive nonexceedances. (Ledford and Tawn 2003 present diagnostic methods for assessing dependence within or between temporal clusters of extremes.) How best to handle such autocorrelated series seems uncertain.

Clearly much remains to be done for single-site series. However, our primary interest lies elsewhere, in the multiplicity of series encountered in studying space–time fields. We go to that topic next.

## Multiple Sites

Multivariate extreme value theory has been an important research direction in recent years (Joe 1994). A natural starting point would be the obvious extension of Equation (12.1). That idea doesn't work very well.

**Extending Fisher–Tippett**

The class of limit distributions in the multivariate case, unlike its univariate cousin, turns out to be very large, limited only by a property called multivariate regular variation (Smith 2004). Moreover, the coordinate random variables in this multivariate limit must be asymptotically dependent except when the variables are actually preasymptotically independent. That property will not always be desirable. [It means $\lim_{q\to 1} Pr\{F_2(X_2) > q | F_1(X_1) > q\} \neq 0$.] In fact, empirical evidence suggests responses may well be asymptotically independent. Bortot et al. (2000) provide some evidence; so does Fu (2002). In particular, the latter indicates that responses at some pairs of sites can

be asymptotically dependent (or at least highly correlated) while some are independent, at least where particulate air pollution fields are concerned.

Bortot et al. (2000) present a model for the case of asymptotic independence, while allowing for arbitrarily strong preasymptotic dependence. Coles and Pauli (2002) extend earlier work to obtain models that embrace both pairwise dependence and independence.

Finally, Hefferman and Tawn (2004), going after the same objective, offer a promising new approach involving conditional distributions. For the $d$-dimensional case, they model the joint distribution of $(d-1)$ subvectors assuming the remaining one tends to its upper end-point. Their approach allows some components not to become extreme, while encompassing previous models when all are. We now turn to alternative approaches.

Smith (2004) describes one he calls a "radically new direction" and ascribes it to Ledford and Tawn (1996, 1997). Their extension of multivariate extreme value theory for the two-dimensional case models just the upper tail:

$$P(\min\{X_1, X_2\} > x) = L(x)x^{-1/\eta} \quad \text{and}$$
$$P(X_1 > x_1, X_2 > x_2) = L(x)x_1^{-c_1}x_2^{-c_2},$$

for indices $c_1$, $c_2$, $\eta$, and a slowly varying function $L$. These distributions overcome the deficiency mentioned above; asymptotic independence is allowed. However, they are bivariate. Moreover, the bivariate normal (and other) distributions are excluded, at least when the correlation is strictly less than one, a serious limitation.

**A Different Approach**

An entirely different approach starts by taking the marginal distributions to be members of one of the univariate families discussed above. Then separately, a multivariate dependence structure is specified (in some cases on a restricted support above marginal thresholds). Reiss and Thomas (1997) present a number of such univariate to multivariate approaches.

Distributions obtained this way tend to be very complex as, for example, that of Coles and Tawn (1996). Although their paper focuses on a univariate case, the extremes of an areal average of rainfall over a region, they base their process models on joint distributions of extremes over sites in a region. The required spatial dependence models form the centerpiece of their extensive investigation.

That complexity can make simpler approaches such as that of Kharin and Zwiers (2000) appealing. Their approach, like the former, addresses precipitation extremes. It uses the marginal GEV distribution and incorporates spatial correlation rather through the distribution of marginal parameters over space. First, estimate the parameters of the marginal GEV distributions separately. Regarding these parameter estimates as indexed by their associated spatial site coordinates makes them a spatial field in their own right. Then predict that field at a particular site, by averaging all the parameter estimates in a neighborhood of it.

Approaches such as the latter can be criticized for the ad hoc way in which spatial dependence is superimposed through the smoothing of the parameter estimates from univariate marginals. For example, we see no compelling basis for selecting one smoothing method over another. Generally, joint distributions obtained by the methods described above need not yield tractable expressions for the conditional and marginal distributions needed for simulating the distributions of complex metrics calculated from these extremes.

In a different direction, Bortot et al. (2000) restrict the support of their distribution to regions above specified thresholds and allow the margins to be selected from the class of GPD distributions. The joint dependence structure is then provided by the Gaussian distribution. They see their approach as addressing a subtle issue revolving around asymptotic independence of the extremes of coordinate responses.

Difficulties associated with the use of extreme value theory, in particular, in dealing with complex metrics and large domains, lead us in the next section to a different approach to modeling extreme fields.

## 12.2 Hierarchical Bayesian Model

The complex models surveyed above are applied in cases with just a few coordinate responses (five in the case of Hefferman and Tawn 2004). We need methods for large numbers of sites such as the 312 seen in Fu et al. (2003). Furthermore, we believe parameter uncertainty must be reflected in a model's specification. That points to the need for a Bayesian approach. Finally, any successful joint distribution model must embrace the (combinatorial optimization) problem of selecting optimal network designs for monitoring extremes. The work of Fu (2002) and Fu et al. (2003) points to an approach to modeling space–time extreme processes that solves some of the problems suggested in the previous section.

That approach links to Bortot et al. (2000) in that our process distribution, conditional on the spatial covariance model, is a special case of theirs (with their thresholds set to $-\infty$). More specifically, we use a log-Gaussian distribution. According to Coles and Tawn (1994), that choice would be in line with a recommendation of the World Meteorological Organization. However, they criticize that approach on the grounds that it induces asymptotic pairwise independence between sites (Reiss and Thomas 1997).

However, our unconditional distribution or process model, differs from that of Bortot et al. (2000). To get it, we follow the line of development in Chapter 10 and specify a prior distribution on the conditional distribution's parameters. That leads to a log multivariate-$t$ process distribution model. By varying the number of degrees of freedom in that process model, we can move between the log-Gaussian process at one extreme to very heavy-tailed distributions at the other. We spare the reader the details since the development is a straightforward application of our theory and its software implementation is demon-

strated in Chapter 14. Instead we give an example, one of several we have explored that suggest the proposed theory may have a practical role to play. In particular, these examples demonstrate the suitability of the log-matric-$t$ distribution as an approximation to the joint distribution of an extremes field in these specific cases.

Although more complex than the log-Gaussian distribution, that distribution has two of the latter's important advantages: namely, it has conditional distributions in the same family and it has an explicitly computable entropy. The latter proves of immense value in the combinatorial optimization problem faced in environmental design considered in Section 12.4 where computing times are immense, even without introducing MCMC computation or numerical integration. (These difficulties will not easily be overcome since that problem is known to be $NP$-hard.)

We turn now to our assessment and demonstrate the applicability of the our approach.

### 12.2.1 Empirical Assessment

This section describes an analysis reported in Fu (2002) and Fu et al. (2003) of extremes in the data introduced in Section 11.10, that is, hourly log $PM_{10}$ concentrations collected at stations in the GVRD, restricted to 1996 to make computation feasible.

That analysis included the development of a log-multivariate Gaussian-inverted-Wishart model. Using the approach in Sun et al. (2000), spatial correlation due to shared temporal patterns in the data series were removed by subtracting a solitary trend model (fitted for all ten stations) from the original log-transformed series. That first term of the (additive) trend model is just the "overall effect" found by averaging the (transformed) data across all sites and hours. Next comes the "hour effect" term that has 24 values each, the value for that hour across all sites. The next three terms for "day effect," "linear time trend," and"seasonal effect" are found in a similar way. Finally, comes the "meteorological effect." To find the last effect due to meteorology subtract the sum of the effects just described from the data and regress the results (residuals) on various meteorological variables (Sun et al. 2000). Last, autocorrelation was removed to make the resulting series consistent with the modeling assumption that they are uncorrelated. Finally, for this investigation, the extreme values are taken to be the weekly maxima of these detrended whitened residuals.

Standard diagnostic checks suggest the marginal (sitewise) distributions of the weekly maxima may be approximated by a Gaussian distribution.

To estimate hyperparameters of the prior distribution, an isotropic spatial correlation structure was adopted and an exponential semivariogram fitted with the result:

$$\gamma(h) = \begin{cases} 0.2 + 0.1(1 - exp(-h/0.2)), & \text{if } \mathrm{h} > 0 \\ 0, & \text{if } \mathrm{h} = 0. \end{cases}$$

(As a technical point, $\hat{F}^{-1} = 0$, $m = 34$, and $c = 13$ were chosen in the notation of Fu et al. 2003.)

To assess the model as a multivariate distribution, a two-deep cross-validation was used, two out of ten sites being removed repeatedly at random to play the role of the ungauged sites. Their values were then predicted and it was determined how frequently the predictive distribution's credibility intervals actually contained the observed values at the two omitted sites. Of course, the coverage fractions do vary over the weeks of 1996. In fact, for 50% credibility intervals individual site coverage fractions ranged from 30 to 75%. However, on average the observed data lay inside the prediction interval 54% of the time. At the other extreme, 95% credibility interval coverages ranged from 80% but most fell in the 95–100% range with an average of 94%. Finally, coverage for the 80% intervals coverage averaged out to 81%.

Overall, we conclude that the log matric-$t$ distribution provides a reasonable model for the extremes seen here. This finding (and others like it not reported here) in turn, suggests the design of networks for monitoring extremes might well be based on that approximation. However, as we show in the next section, other issues present themselves in that context.

## 12.3 Designer Challenges

This section concerns issues that arise in designing networks for monitoring fields of extremes.

### 12.3.1 Loss of Spatial Dependence

Any site pair can be asymptotically dependent or independent. If they lay along the prevailing wind direction they may have very similar extreme values and thus be asymptotically dependent. Alternatively, they may be asymptotically independent if they lie in directions orthogonal to that path.

Some models may be seen as deficient because they make all site pairs the former, others because of the latter. In particular, neither type may seem acceptable when both sorts of asymptotic behavior are seen in the field being modeled.

Yet in reality only the preasymptotic case ever obtains, making the relevance of asymptotic cases doubtful. Indeed, what may be more important is a model's capacity to flexibly admit both types of dependence in the preasymptotic case, as that in Section 12.2 will do.

This section concerns process maxima over successive time periods of length $n$; that is, $Y_{i(r+1)} = \max_{j=k}^{k+n-1} X_{ij}$ for $k = 1 + rn$, $r = 0, \ldots,$ $i = 1, \ldots, I$, where $X_{ij}$ denotes site $i$'s response at time j, $i = 1, \ldots, I$, $j = 1, \ldots, n$. In particular, it focuses on $Cov(Y_{i(r+1)}, Y_{i'(r+1)})$ for $i \neq i'$ as $r$ grows large.

The covariance may not, in fact, be a good measure of dependence except for models deriving from Gaussian distributions. However, the latter are important, a subclass that potentially offers good approximations to suitably transformed extreme value distributions.

We now give examples to illustrate the complexity of fields of extremes. The first concerns the eastern United States.

*Example 12.1. Hourly ozone concentrations over the eastern United States*
The data for this example come from the AIRS database, maintained by the United State's Environmental Protection Agency. The analysis concerns hourly ozone concentrations over 120 days in the summer of 1997, more specifically, the 16 sites with the lowest fraction of missing hourly values (less than 18 over the 2880 hours). Prior to the analysis, the data were square-root transformed to make them have a more Gaussian distribution, but no other processing was done.

Intersite correlations were computed for these 16 sites and hourly concentration maxima over varying time intervals ("spans"). Figure 12.1 depicts the results for site #1 and each one of the remainder.

Notice the wide range of correlations for the span of 144 hours (6 day intervals). In part, this derives from their large standard errors. After all, they are based on merely 20 six-day maxima at each site. However, it also reflects the great diversity of associations among the sites due to the latent factors that determine ozone levels.

In descending order, site-pairs (1,8),(1,6), and (1,2) have the three smallest correlations (0.14, 0.21, and 0.27, respectively) at that span. Both #6 and #8 are well far away from site #1 so those two small correlations at least would be expected. In contrast, Site #2 is actually quite close by #1. A surprising result.

At the other extreme, the three largest span 144 correlations, 0.81, 0.66, and 0.64 come from site pairs (1,14), (1,11), and (1,16), respectively. Once again we are surprised since although #14 is right next door to #1, both #11 and #16 are quite some distance away. However, the latter two are collinear with #1. Could this reflect the prevailing wind direction so that these three sites share the same extremes? Curiosities such as this abound in the study of extreme fields.

Similarly nonintuitive features obtain in our second example that comes from Fu (2002) and Chang et al.(2006) and a very different part of North America.

*Example 12.2. Vancouver's 1996 $PM_{10}$ field*
In this example, we emulate an analysis reported by Chang et al. (2006). The data are measured hourly ambient log $PM_{10}$ concentrations recorded at nine monitoring sites during the 240 week period up to the end of 2001. The site locations in the Greater Vancouver Regional District (GVRD) are portrayed in Figure 12.2. In this example, temporal effects were removed so that spatial

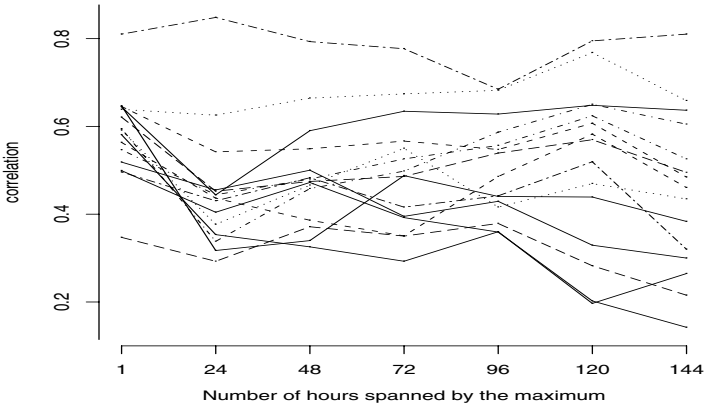


Fig. 12.1: Correlations between site #1 and 15 other sites.

structure alone could be expressed through the maxima computed for the series over varying time spans. These spans stretched from one hour at one extreme to 84 days at the other.

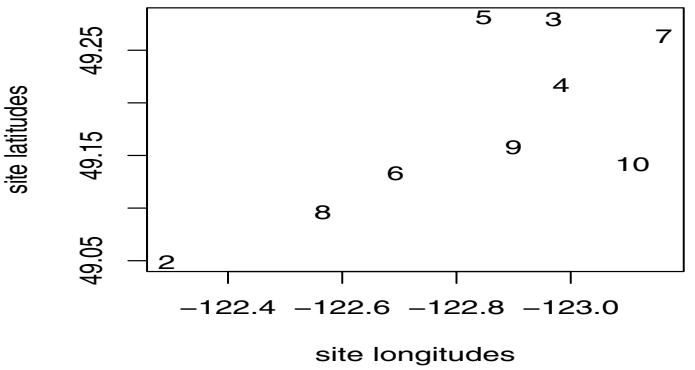


Fig. 12.2: Location of nine $PM_{10}$ monitoring sites in the Greater Vancouver Regional District.

The intersite correlations are depicted in Figure 12.3. Notice how the correlation declines for most site pairs, but persists for a few in agreement with the previous example.

To examine this phenomenon more closely, we list in Table 12.1 those correlations for the maxima at the nine sites computed over a very long-time span of 84 days. The table presents them in increasing order from left to right and down its rows. A glance at the table shows site pair $(3, 6)$ to be the winner. That pair's correlation actually seems to increase as the time span increases (although this could be a spurious product of sampling error). However, $(8, 9)$ is a close competitor even though that pair has a shorter interdistance compared to that of the former pair as shown by Figure 12.2 reveals.

At the other extreme, site #2 seems to be uncorrelated with most of the other sites $2, 5, 8, 6, 3, 9,$ and 10. That does not seem surprising. Site #2 is situated near a major roadway with fairly heavy traffic volumes during rush hours that can generate relatively large particulate concentrations. Mysteriously, site #2 does have a positive, albeit small, association with site #4.

Roughly speaking, we see evidence of asymptotic dependence in the following pairs: $(3, 6)$; $(8, 9)$. Weaker evidence of such dependence is found for: $(4, 9)$; $(3, 5)$; $(7, 10)$;$(5, 6)$; $(3, 8)$. However, we know of no substantive evidence that would lend support for those findings.

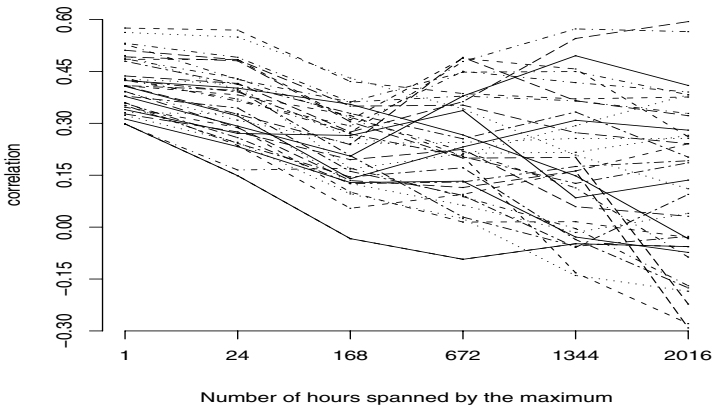


Fig. 12.3: Intersite correlations for the maxima over time spans of between 1 and 2016 hours of log $PM_{10}$ concentrations. These were obtained from nine monitoring sites in the Greater Vancouver Regional District between 1996 and 2001.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (2,7) | (7,9) | (4,6) | (2,5) | (2,8) | (2,6) | (4,5) | (2,3) |
| -0.29 | -0.28 | -0.22 | -0.19 | -0.18 | -0.17 | -0.09 | -0.07 |
| (2,9) | (2,10) | (5,9) | (5,10) | (9,10) | (8,10) | (3,10) | (6,10) |
| -0.06 | -0.06 | -0.03 | -0.03 | -0.03 | 0.03 | 0.04 | 0.09 |
| (6,10) | (3,4) | (7,8) | (3,7) | (5,8) | (4,7) | (2,4) | (6,9) |
| 0.09 | 0.11 | 0.14 | 0.19 | 0.19 | 0.20 | 0.24 | 0.24 |
| (6,9) | (6,7) | (3,9) | (4,10) | (4,8) | (5,7) | (6,8) | (4,9) |
| 0.24 | 0.26 | 0.26 | 0.27 | 0.28 | 0.32 | 0.33 | 0.38 |
| (3,5) | (7,10) | (5,6) | (3,8) | (8,9) | (3,6) | | |
| 0.38 | 0.38 | 0.39 | 0.41 | 0.57 | 0.59 | | |

Table 12.1: In ascending order by site pair, intersite correlations of successive maxima over 2016 hours for detrended, whitened log $PM_{10}$ concentrations. In all nine Vancouver sites were studied during a subperiod of 1997–2001.

In any case the results emphasize the need for models, for example, of Coles and Pauli (2002), that allow flexibility in specifying the between-site dependence.

This problem of decreasing between-site field dependence has important implications for prediction and design. In particular, it seems likely that in some urban areas, a fairly dense grid of monitoring stations will be needed to ensure that extreme values over the region are reliably detected.

**Decline of Dependence**

To seek a better understanding of the phenomenon revealed in the last section, Chang et al. [(2006); hereafter referred to as CFLZ] conducted a simulation study. Following Fu (2002) and Fu et al. (2003), they adopted a log matric-$t$ process model.

More precisely, they chose ten aligned monitoring sites, labeled $i = 1, \ldots, 10$, with site responses having mean 0 and variance 1. They varied the number of measurements $n$ from which the maximum is found in each replicate while fixing the number of replicates (extreme values) at $N = 5000$. Using an algorithm from Kennedy and Gentle (1980, pages 231–232) they sampled the covariance matrices $\Sigma$ from an inverted Wishart distribution (see Appendix 14.1), i.e., $\Sigma \sim IW(\Psi, m)$ with varying degrees of freedom $m$ and isotropic (hyper-) covariance kernel $\Psi$, $\Psi_{ij} = exp[-\alpha|i-j|]$, $i,\ j = 1, \ldots, 10$.

That algorithm proceeds as follows: (1) generate $\Sigma \sim IW(\Psi, m)$; (2) generate $n$ random vectors $X_k = (X_{k1}, \ldots, X_{k,10},\ k = 1, \ldots, n$; (3) find $Y_i = (Y_{i1}, \ldots, Y_{i,10})$ where $Y_{ij} = \max_{k=1}^{n} X_{kj},\ j = 1, \ldots, 10$; (4) repeat this process $N$ times to get replicates of the vector of extremes.

As in the previous subsection, interest focuses on between-site correlations for extreme values with varying $n$ and degrees of freedom. In general, they found the between-site correlation declined as the number of degrees of freedom of the inverted Wishart increased (making the tails on the resulting $t$

distribution heavier). The same result obtained as $n$ increased. The between-site correlation tended to decline more as $n$ increased when the number of degrees of freedom was large rather than small.

The relationship between the between-site correlations for extremes and that of the original response field was also explored. CFLZ found an empirical power law characterized this relationship quite well:

$$Cor_{ext} = \beta(Cor_{raw})^{\gamma} + \delta,\ \gamma > 0.$$

Here $Cor_{ext}$ denotes the between-site correlation for extremes $Cov_{raw}$ that for the original response field. CFLZ fitted their power law for varying $n$ and degrees of freedom to obtain the results summarized in Table 12.2.

Note that for large values of $\gamma$, the "power" in the power law means more rapidly declining between-site correlations for extremes. That is, the extremes at any pair of sites tend to have a weaker linear association than when the power is small. Indeed, when $\gamma = 1$ there is no loss of correlation in going from the original response field to the extremes field.

Thus smaller powers are associated with heavier tails (that is, smaller degrees of freedom). Since empirical studies have shown that log matric-$t$ air pollution predictive distributions fit observed fields well, this finding, if validated by more rigorous analysis, would constitute good news: unmeasured extremes can be better predicted than if, for example, the field were log-Gaussian.

Another conclusion suggested by these findings: the power will also decline with $n$ (the number of data points, for example, hour or days on which replicate extremes are calculated). This finding, if confirmed, would have implications for developing realistic compromise air quality criteria and designing the associated monitoring networks.

Finally, note that the range of powers in the power law is greater when the degrees of freedom is large relative to the range when it is small.

### 12.3.2 Uncertain Design Objectives

As noted in Chapter 11, model–based design strategies have generally focused on predicting unobserved values or on estimating parameters of a regression function. But what if, instead, extremes are of interest, say for regulation or assessing the benefits of an abatement strategy? More specifically, how well does the maximum entropy design work?

To answer such questions, performance criteria must be specified. Not surprisingly the answers turn out to be mixed depending on the choice. That raises anew the issue of which to use and hence, whether confronting that choice can be avoided in this context.

#### Entropy-Based Design

Consider anew the hourly $PM_{10}$ ($\mu g\ m^{-3}$) concentration field over the Greater Vancouver Regional District (GVRD). The ten sites in Figure 12.4 monitor

| DF | n | Power Law |
|---|---|---|
| 20 | 24 | $Cor_{ext} = 0.89 \times Cor_{raw}^{2.28} + 0.10$ |
| | 100 | $Cor_{ext} = 0.94 \times Cor_{raw}^{2.83} + 0.05$ |
| | 500 | $Cor_{ext} = 0.93 \times Cor_{raw}^{2.98} + 0.06$ |
| 30 | 24 | $Cor_{ext} = 0.94 \times Cor_{raw}^{2.49} + 0.05$ |
| | 100 | $Cor_{ext} = 0.94 \times Cor_{raw}^{3.05} + 0.06$ |
| | 500 | $Cor_{ext} = 0.96 \times Cor_{raw}^{3.38} + 0.03$ |
| 40 | 24 | $Cor_{ext} = 0.94 \times Cor_{raw}^{3.04} + 0.05$ |
| | 100 | $Cor_{ext} = 0.96 \times Cor_{raw}^{3.06} + 0.01$ |
| | 500 | $Cor_{ext} = 0.97 \times Cor_{raw}^{3.85} + 0.02$ |
| $\infty$ | 24 | $Cor_{ext} = 0.93 \times Cor_{raw}^{2.80} + 0.07$ |
| | 100 | $Cor_{ext} = 0.97 \times Cor_{raw}^{3.51} + 0.02$ |
| | 500 | $Cor_{ext} = 0.98 \times Cor_{raw}^{4.40} + 0.01$ |

Table 12.2: Empirical power laws relating intersite correlations for extremes (denoted $Cor_{ext}$) against original responses ($Cor_{raw}$) for varying degrees of freedom (DF).

that field (see Zidek et al. 2002). Note that its stations started operation at different times (an issue addressed by Le et al. 2001 and Kibria et al. 2002).

Suppose the hypothetical designer seeks to add to the network six sites from among the 20 candidates in Figure 12.4, these being centered in Census Tracts (regions with reasonably large populations). Calculation of the entropy requires the hypercovariance (hereafter called "covariance") matrix of the predictive log $\mathrm{PM}_{10}$ ($\mu$g m$^{-3}$) distribution, supposed to have the Kronecker product form $\Lambda \otimes \Omega$. Here $\Lambda: \ 30 \times 30$ represents the spatial covariance between the $10 + 20 = 30$ sites $\Omega: \ 24 \times 24$ the within-site correlations between hours. Let

$$\Lambda = \begin{pmatrix} \Lambda_{uu} & \Lambda_{ug} \\ \Lambda_{gu} & \Lambda_{gg} \end{pmatrix}.$$

Here $u$ refers to unmonitored sites (ungauged) and $g$ to monitored (gauged) sites regardless of when they began operation. Thus, $\Lambda_{uu}: \ 20 \times 20$ is

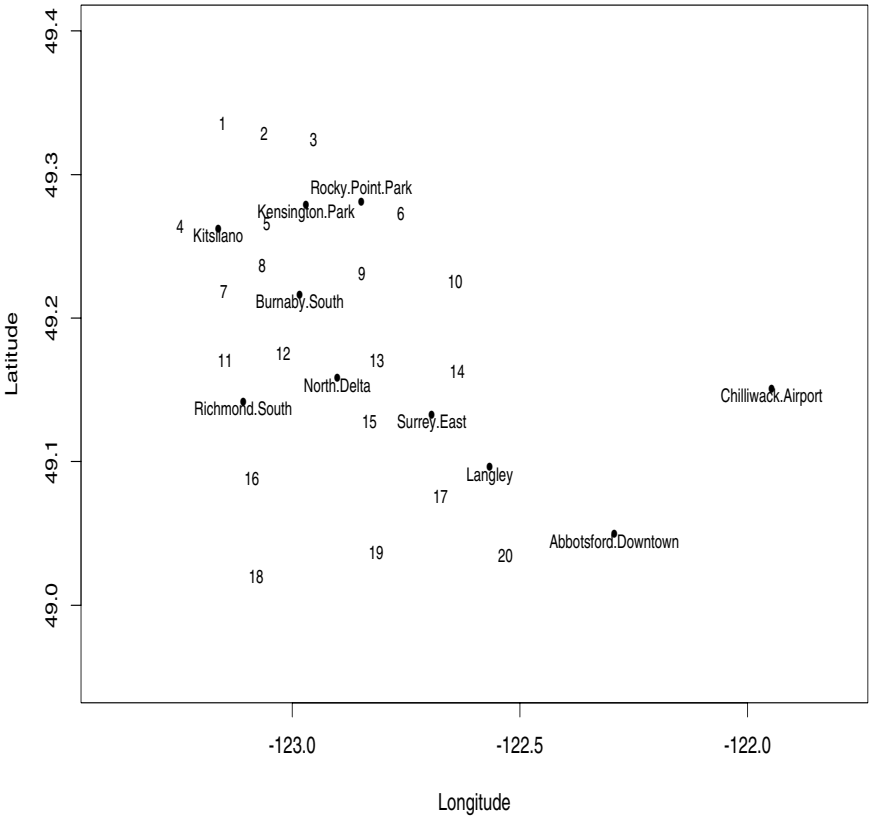


Fig. 12.4: Hourly $PM_{10}$ ($\mu$g $m^{-3}$) concentration monitoring sites in the Greater Vancouver Regional District.

the covariance matrix for candidate sites, six to be chosen. Next, $\Lambda_{u \cdot g} = \Lambda_{uu} - \Lambda_{ug}\Lambda_{gg}^{-1}\Lambda_{gu}$ is the conditional spatial covariance of the ungauged sites given data from the gauged sites.

All the components of the covariance structure can be estimated (see Kibria et al. 2002). Chapter 14 describes software for doing so. Finally, ignoring irrelevant terms and factors, the entropy of any proposed $6 + 10 = 16$ station network of gauged sites including the original 10 is the logarithm of the determinant of $\Lambda_{a \cdot g}$, the $16 \times 16$ submatrix of the estimated $\Lambda_{u \cdot g}$ corresponding to that proposed network. Here $a$ denotes the added stations. So the six new

sites $a$ must be chosen to maximize this determinant and thereby find the maximum entropy optimal design.

Figure 12.5 shows all the sites with their ranks based on the size of the (conditional) variances of their predictive distributions. The existing sites are dots and the potential sites are numbered.

The designer might now wonder which potential site has the (conditional, log-transformed concentration, predictive) distribution with the largest hypervariance. He finds the answer in the figure: #19. That site lies well outside the cluster of existing sites so its large conditional hypervariance is not surprising. Thus, it seems a likely candidate for membership in the set of six new stations to be selected for the network.

Le et al. (2005) confirm that choice- Site #19 does lie in the optimal set of six new stations. The remaining four of the top six hypervariance-ranked sites also make it in: #s 10, 16, 18, 20. However, the remaining selected site #12 is a surprise, coming in ahead of the sixth and seventh ranked sites #14 and #17. Presumably, the entropy criterion has shrewdly recognized that responses at the latter will be predictable from those at sites #19 and #20, once these are added.

### Does It Work for Extremes?

We can now ask how well the network of the last section would work for monitoring extremes? The answer depends on how you look at things. A regulator might view it in terms of her need to enforce compliance with national or local standards. More specifically, she might prefer that the six new sites, *add* for short, be chosen from among the 20 as those most likely to be noncompliant. As a byproduct, the remainder *rem* would be more likely than those in the *add* subgroup to be compliant.

Suppose the regulator adopts an ad hoc criterion of 50 (ppb) for $PM_{10}$ concentrations. More formally,

$$add = \arg\max_{add'} \text{Prob}[\max_{t=\text{hours}, j \in add'} Y_{tj} > 50 \text{ (ppb)}]. \tag{12.2}$$

Here *Prob* means with respect to the joint conditional predictive probability for unmonitored responses given those at the gauged sites. That probability cannot be found explicitly. Hence finding *add* entails repeatedly simulating the field of 20 unmeasured values hour by hour over the entire day. Then for any proposed subset of six sites $add'$ the fraction of times $\max_{t=hours, j \in add'} Y_{tj} > 50$ (ppb) estimates that probability. By systematically varying $add'$ over all possible subsets and selecting the maximum among these estimated probabilities, *add* is obtained.

This analysis reveals the advantage of having a predictive distribution over a mere predictive point estimate. For any proposed regulatory criterion, the

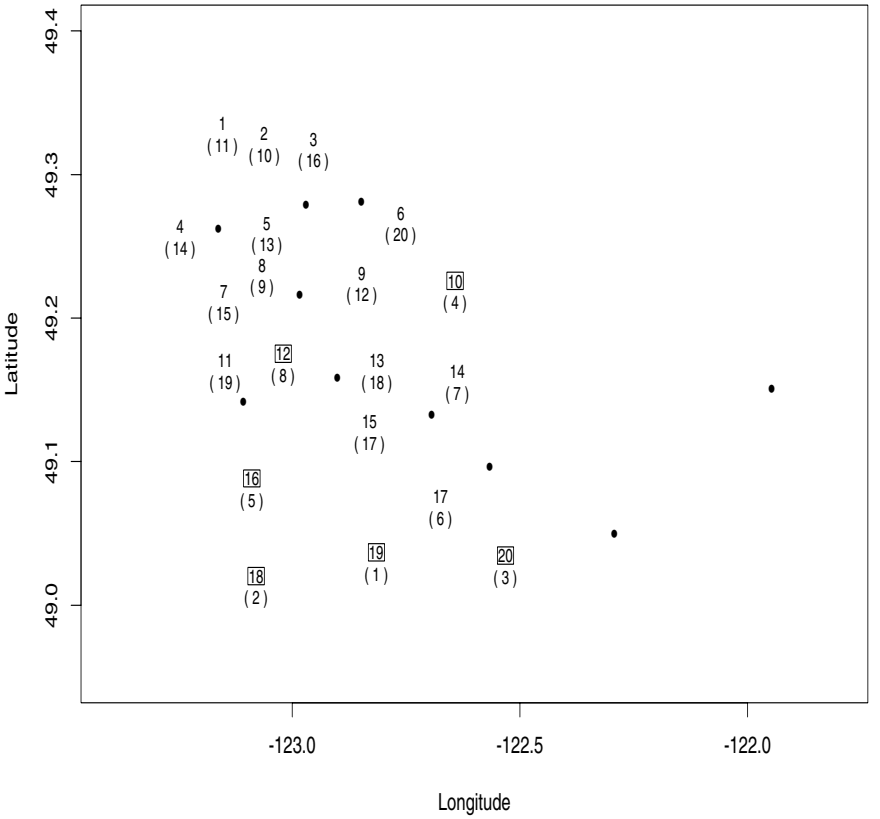


Fig. 12.5: The numbered locations represent sites that might potentially be gauged when adding six new monitoring stations. The numbers appearing in brackets indicate their ranking by the size of their conditional predictive hypervariances in $\Lambda_{u \cdot g}$.

required probability of noncompliance can be found as above. Indeed, much more complicated metrics such as the largest eight hour moving average over the day could have been handled.

However, the choice of a metric is only one of the designer's issues. Her approach also requires that she specify the day for which this probability is calculated, for it, unlike the entropy, depends on the conditional mean. That, in turn, depends on the values at the gauged sites which change from day

to day. Hence, this seemingly simple criterion has led the designer to a day-dependent design. How stable would that optimum design be over days?

To answer that question, suppose she picked February 28, 1999 when Vancouver's particulate air pollution levels can be high. Table 12.3 shows in ranked order the top ten choices of the six *add* sites based on the entropy criterion. For each such choice, the table gives the probability defined above that they will be in noncompliance that day. More to the point their order is reported there, it being calculated with respect to our noncompliance criterion, among all the 38,760 candidate subsets of six possible *add* subsets.

| MaxEnt Order | Selected Sites | 100*Prob | Order |
|---|---|---|---|
| 1 | 10 12 16 18 19 20 | 45.9 | 64 |
| 2 | 10 14 16 18 19 20 | 46.1 | 55 |
| 3 | 10 12 14 18 19 20 | 44.6 | 145 |
| 4 | 8 10 16 18 19 20 | 45.0 | 123 |
| 5 | 2 10 16 18 19 20 | 45.1 | 109 |
| 6 | 2 10 12 18 19 20 | 43.6 | 222 |
| 7 | 8 10 14 18 19 20 | 43.7 | 212 |
| 6 | 2 10 14 18 19 20 | 43.9 | 200 |
| 9 | 10 16 17 18 19 20 | 49.5 | 1 |
| 10 | 1 10 16 18 19 20 | 45.1 | 109 |

Table 12.3: The top ten choices among the 38,760 available, of subsets of six new sites to be added to Vancouver's existed set of ten sites gauged to measure $PM_{10}$ concentrations. Here *Prob* means the probability of noncompliance while *Order* means order with respect to the noncompliance probability criterion.

A number of observations can be made about the results in that table with the help of Figure 12.6. Boxplots show the sitewise distribution of simulated daily maxima obtained from the predictive response distribution for that day.

- The entropy criterion does reasonably well. Its ninth ranking candidate turns out to be the best with respect to noncompliance.
- The noncompliance probabilities for the top ten choices by the entropy criterion are quite similar. However, the importance of these seemingly small differences should not be minimized in a regulatory environment, where the financial and other penalties for noncompliance can be large.
- Sites #10, 18, 19, and 20 are always selected in the top ten, because of their large conditional response variances as manifest in the boxplots of Figure 12.6. Between-site correlations are essentially ignored, no surprise in the light of the results of the previous section showing that for extreme values, between site correlations tend to be small.

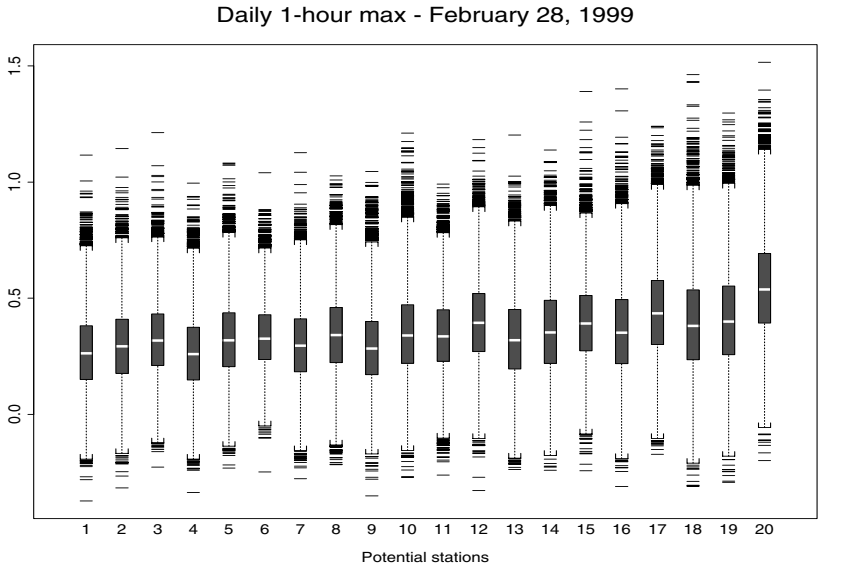


Fig. 12.6: Through boxplots this figure depicts the sitewise distribution of simulated daily maxima log ozone (log $\mu$g m$^{-3}$) values for February 28, 1999.

- The ninth ranking choice in the table is best by the noncompliance criterion, that choice substituting Site #17 for Site #12, since the former's daily ozone maxima tend to be larger while their variances are similar and large (see Figure 12.6).

However, the entropy criterion does not do so well on other days such as August 1, 1998, with respect to noncompliance criteria. Figure 12.7, like Figure 12.6, indicates the distribution of daily log ozone distributions for the potential new sites. Notice that the level of log ozone at Site #19 is much lower than it was above, and hence it is not a contender by the noncompliance criterion. Sites #16 and # 18 have also dropped out. Site #7 looks likely to be in noncompliance that day.

We see that the optimum noncompliance design is not stable; it can vary from day to day. This last result implies that any fixed design, no matter how chosen, would be less than optimal on lots of days.

In any case, the method above for finding that fixed design seems impractical since it would require the designer to select and use a specific day. A possibly preferable alternative would use a weighted average over days of the criterion *prob*s in Equation (12.2) as the design objective function based on multicriteria optimization considerations. Equal weights would have some appeal since the objective function could then be interpreted as the expected

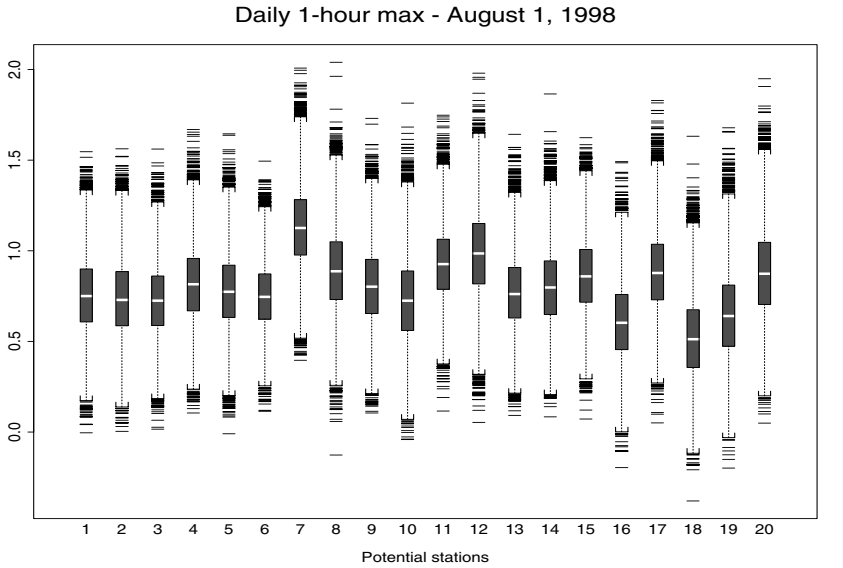


Fig. 12.7: Boxplots show the sitewise distribution of simulated daily maxima log ozone (log $\mu$g m$^{-3}$) values for August 1, 1998.

*prob* based on selecting the day at random. However, this new criterion would weight equally days when noncompliance was unlikely. Why should such days be accorded any role in picking the winning design?

The approach we prefer would not be based on the concentrations observed at the gauged sites at all. Instead, we would take full advantage of the theory in Kibria et al. (2002) and use the predictive distribution for the gauged as well as that for the ungauged sites conditional on the values at the gauged sites (see Figure 12.8).

The simulations represented in Figure 12.8 are obtained in the following steps.

1. For the first day of 1998, generate a random vector from the marginal predictive distribution for the gauged sites on that day.
2. Conditional on that vector, generate a random vector from the predictive distribution for ungauged sites.
3. Compute the daily average log ozone concentration for that day at every site.
4. Repeat steps 1-3 for each day of 1998;
5. At each site find the 99th percentile of daily averages and take its antilogarithm.

This figure suggests that if the annual 99th percentile of daily average ozone levels were used as the compliance metric, stations #5, 7, 8, 9, 12, and 13

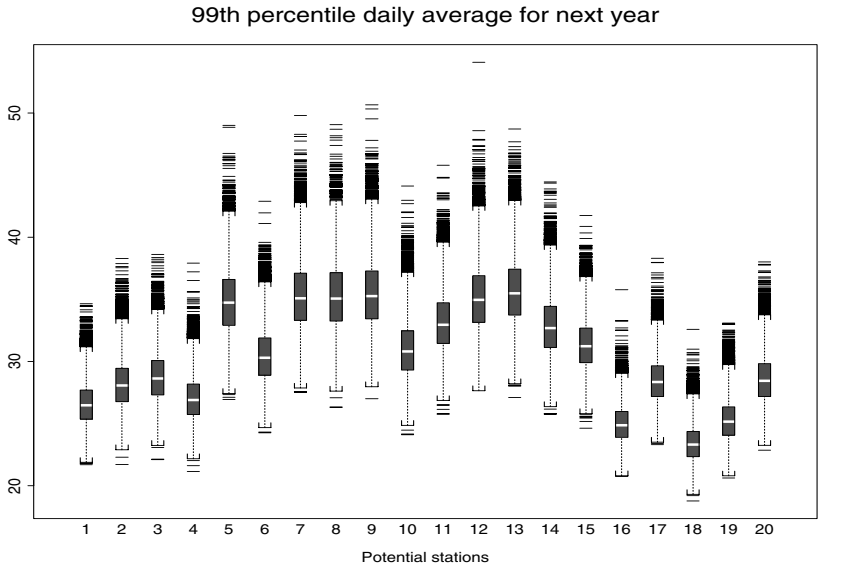


Fig. 12.8: Boxplots show the sitewise distribution of simulated annual 99th percentiles of daily average ozone ($\mu$g m$^{-3}$) concentrations.

would be the most likely to noncomply. In any case, in this figure we see major differences among the potential sites with respect to the distribution of this metric.

### Other Regulatory Criteria

Stepping back a bit, we might suppose that the regulator would like a design that gives her the greatest chance of detecting a noncompliance event itself without regard to the specific concentration of the response that is noncompliant. From that perspective she might want to add sites with a high collective risk of noncompliance. Or she might seek the sites with a minimum risk of noncomplying and leave these out of the extended network.

To explore these ideas, define three distinct compliance events $R$, $A$, and $G$. Their associated conditional and marginal probabilities would be estimated through simulation using our predictive distribution. Once a particular set of, say six candidate sites among 20 have been selected, these events would be, respectively, compliance of the remaining sites, of the added sites, and of the gauged sites. Whether each of these events occurred could then be determined for any series of simulated responses at both gauged and ungauged sites, over hours and even days (if a given design criterion's metric were based on a multiplicity of days). After sufficiently many replicate simulation runs, the probabilities for these events, as described below, could be found as the

relative outcome frequencies. We illustrate these calculations below, albeit using a computationally simpler model.

Even with this new approach a number of possible design objectives present themselves. For example, the designer could choose to select the six *add* sites to maximize $P(\bar{A}|G)$. Alternatively, she could maximize $P(\bar{A}|\bar{G})$ where $\bar{A}$ means A is not in compliance. Still a third would be to maximize $P(\bar{A})$. Since $P(\bar{A}) = P(\bar{A}|G)P(G)+P(\bar{A}|\bar{G})P(\bar{G})$, we see that the first two are related to but different from the third. In fact, the third could be interpreted as a multicriteria optimization problem that combines the objective functions from the first two.

She could alternatively express the regulators' goals through minimization of $P(R|G)$, $P(R|\bar{G})$, or $P(R) = P(R|G)P(G) + P(\bar{R}|\bar{G})P(\bar{G})$. Still other credible choices exist, the maximization of $P(\bar{A}|G, R)$, for example. Numerous other options exist.

However, we do not see what principles she could invoke to compel a choice of a single criterion for designing a network that best meets the regulator's goals of detecting noncompliance, knowing that the different choices yield different designs. We see this as a major challenge facing the designer of a network to monitor extremes, one that does not seem to have been very much explored.

To get a better understanding of this issue Chang et al. [(2006); or CFLZ for short] did a simulation study resembling the one in Section 12.3.1. To make their study realistic they used a total of 30 sites with locations depicted in Figure 12.4. Sites #21–#30 were taken as gauged. Of the remaining sites #1–#20, six were to be gauged. CFLZ assumed the joint 30-dimensional response matrix has a multivariate-$t$ distribution with 35 degrees of freedom and they generated their responses using the algorithm of Kennedy and Gentle (1980, pages 231-232). All marginal site response distributions had mean zero. The between-site covariances were those estimated by Le et al. (2001) and used in Section 12.4. The variances themselves are depicted in Figure 12.9.

Notice that of the nongauged sites #9, #10, #12, #13, #10, and #20 are the ones having the largest variances, making them obvious candidates for network membership.

Their study proceeded by generating a large number of replicates of the 30-dimensional response vector representing the values at all the sites, gauged and ungauged. They found the subset of those vectors for which the ten gauged sites were compliant, i.e., had a maximum less than a specified threshold (the compliance event $G$). Finally, within that subset they looked at all the possible choices of the subset of *add* sites among the 20 ungauged sites.

CFLZ considered two possible compliance criteria, denoted $A$ and $R$. For the first they estimated, by a relative frequency, the probability (conditional on $G$) that the six selected sites would be noncompliant. They then selected the six sites that maximized that conditional probability. For the second they looked at the subsets of 14 unselected sites and their estimated (conditional) probability of noncompliance. Finding the subset of 14 with the minimum

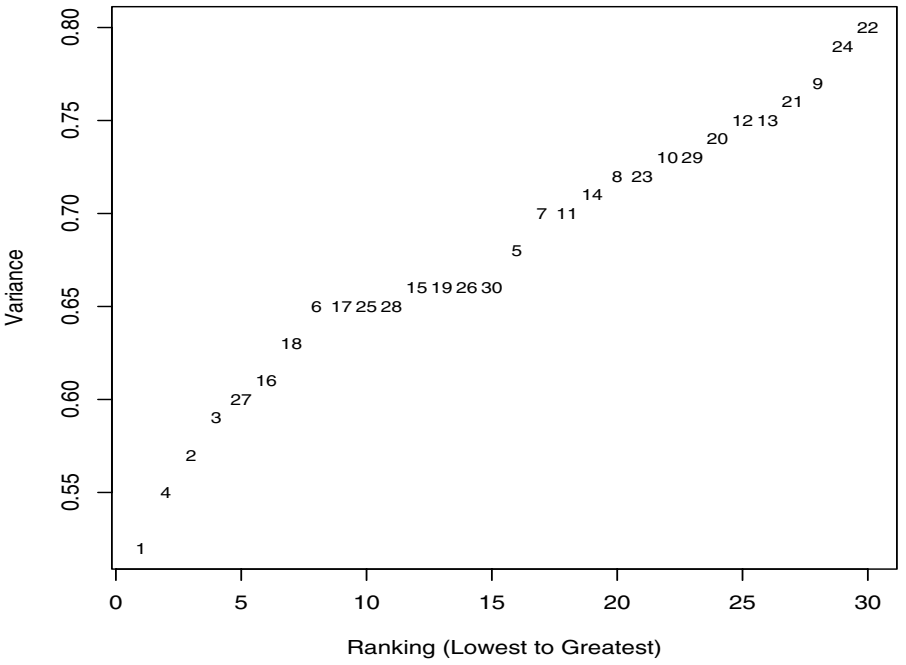


Fig. 12.9: The ranked site variances based on estimated variances.

conditional probability, they got their complement as the six sites (the *add* set) to be added to the network of ten gauged sites. Note that these two criteria do not yield the same *add* set.

After experimenting with a variety of options, CFLZ settled on three possible compliance thresholds for their study, $\log 2$, $\log 4$, and $\log 7$, the first being the most stringent. Any site exceeding the threshold would be in noncompliance. $G$ is the event that all of the ten gauged sites are in compliance. When the threshold is set to $\log 2$, $G$'s occurrence would be quite informative and alter appreciably, the probability distribution of the responses at the 20 ungauged sites.

The effect of learning $G$ occurred can be seen in Figures 12.10 and 12.11. They depict the variances of sites conditional on $G$ for the most and least stringent thresholds. (We should emphasize that CFLZ did not condition on the responses actually measured at those gauged sites!) Not only does the variance of the responses at the ungauged sites change, but so does their variance order due to the complex interaction of between-site covariances and

their marginal variances. To interpret them these figures need to be compared with each other and with Figure 12.9.

First observe that the response variance of Site #20, located at the extreme southeast corner of Figure 12.4 retains the top variance rank among the 20 ungauged sites in all three figures. Clearly the level of uncertainty about its response remains unchanged by knowledge that $G$ has occurred.

On the other hand, that knowledge has a substantial impact on Sites #18 and #19; their variance ranking moves close to the top, although below that of Site #20. We found this result surprising since they, like Site #20, are on the southern boundary of Figure 12.4 and well away from the "pack." Why the knowledge should have increased their uncertainty is a mystery. However, the result indicates if new sites are added to the network on the basis of compliance probabilities conditional on $G$, Sites #18, #19, and #20 will be very strong candidates.

The other site worth remarking on is Site #9. In that case response uncertainty has gone down thanks to knowledge that $G$ has occurred. That is because unlike the other sites, #9 is in the middle of the collection of gauged sites (that are known to be in compliance given the conditioning event). In fact, the most stringent threshold moves its variance rank lower than the least stringent, as intuition would suggest.

In line with the previous observation, note that with the most stringent threshold $\log 2$ $G$'s occurrence imposes a substantial constraint on the simulated field, its dispersion. Hence its marginal variances are smaller than for the least stringent.

The CFLZ study reveals that Sites #18 and #9 must be included in the optimal group of six no matter what threshold is used ($\log 2$, $\log 4$, $\log 7$) or which criterion [maximize compliance over subset of six sites, $P(A \mid G)$ or minimize noncompliance over subsets of 14 sites, $P(R \mid G)$] is used. Only the other 3 selected sites vary somewhat. In fact, for the most stringent threshold, both criteria $A$ and $G$ select site #16 as well. However, whereas $A$ chooses #1 and #10 for the remaining two slots $G$ picks #4 and #17. Whichever criterion is selected, the conditional probability of the six selected sites being in compliance is about 0.54 for the most stringent threshold, one that is hard to meet.

Note: CFLZ used very large-sample sizes in their simulation, and the estimated standard deviation of the simulated estimates of the respective probabilities show them to be significantly different, not just different due to chance.

At the other extreme when the least stringent threshold is used, both criteria select #9 and #10 to be added. For the sixth and last slot, $A$ picks #12 while $G$ goes for #14. Now the chance of the added six being compliant comes in at around 0.95.

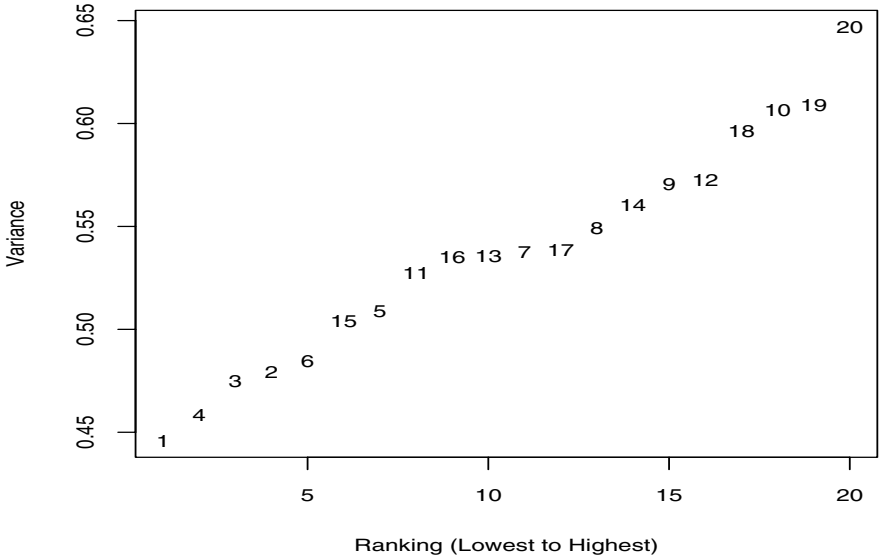


Fig. 12.10: The ranked variances of sites for the response distribution, conditional on the ten gauged sites being in compliance. Here the threshold is log 2.

Generally the results of CFLZ show convincingly that the optimal design depends on the criterion selected, forcing a designer to select from among myriad seemingly plausible compliance criteria. However, in this case at least only small differences are seen in the resulting criterion probabilities among the top five designs, thus offering some hope that at least at the practical level, the eventual choice of a winner among the leading contenders will not be critical. Clearly, more analysis will be required to determine if this hope is realistic. For the time being, how might she escape from this morass?

Her dilemma makes us reconsider the possibility of using an information-based (entropy) approach. However, we know from the results in the previous section that criterion cannot be applied directly to the response field. Therefore a new approach is needed and one is described in the next section.

## 12.4 Entropy Designs for Monitoring Extremes

Section 12.3.2 has shown us the desirability of bypassing specific design objectives in favor of a more generic choice such as the entropy. However, Section 12.3.2 has shown the futility of applying the entropy criterion directly to

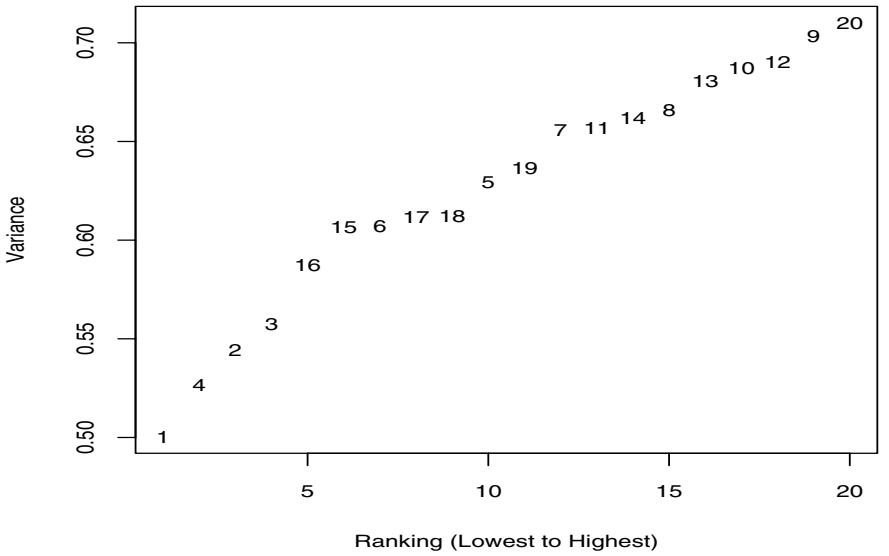


Fig. 12.11: The ranked variances of sites for the response distribution, conditional on the ten gauged sites being in compliance. Here the threshold is log 7.

the response field itself. So in this section we explore its use on the field of extremes; that is, after all, the posited object of inferential interest in this chapter.

Various approaches would lead to an entropy-based design criterion. The most obvious would be use of a multivariate joint distribution for the extremes field. However, we know of no distribution that would yield the conditional predictive distribution we need. (See Fu et al. 2003 for a recent review.) Another possibility would use a predictive distribution for the response field to estimate the required conditional probability density functions by a Monte Carlo approach. That of Kibria et al. (2002), as developed within a design framework by Le et al. (2005), offers one such possibility. However, that approach also fails due to the curse of dimensionality; reliably estimating a joint density function over a high-dimensional domain proves impossible.

In this section we propose a third approach that derives from the framework set out in Section 12.

## 12.5 Wrapup

Environmental monitoring criteria have been based on extremes because of their perceived association with risk. Consequently, networks have been established to monitor such things as air pollution and acidic deposition and in their associated extremes. Surprisingly no attempt seems to have been made to define explicit technical design objective functions that express societal concern about the risk associated with extremes. These functions would accommodate and embrace knowledge, concern, risk perceptions, and political demands.

As a consequence, the specification of such criteria has been left to the designers. This chapter has pointed to some of the technical challenges they face. Among other things, we have delineated a large number of seemingly credible alternatives, all of which are consistent with the goal of detecting noncompliance and thereby enforcing standards. Yet they yield a variety of different designs!

This complacency may derive from a belief that the layout of a network is not critical. That would certainly be true, for example, of designs for monitoring London's $PM_{10}$ field since it is quite flat (Zidek et al. 2003). Even a single station would characterize that field quite well, no matter where it was placed.

However, such complacency may not be warranted where extremes are concerned. An important issue presented in this chapter is the diminished intersite correlations among extremes compared to the response fields from which they derive. This discovery leads us to wonder how well current monitoring programs work, especially in guarding the susceptible and sensitive, such as the old and young in urban areas, if indeed extremes are important determinants of risk. In fact, the near independence of extremes between sites in certain areas would suggest the need for a dense network of monitors to adequately protect the associated population. In fact, the high cost of setting up and maintaining monitors has severely restricted their numbers.

In any case this chapter also offers a practical way of addressing some of the challenges in design and spatial prediction, in particular the multiplicity of objectives confronting the designer. That approach is based on an entropy criterion and the use of a joint matric-$t$ distribution as an approximation to the actual joint distribution of spatial extremes. However, the support for our approach is both limited and empirical. More validation and testing would clearly be desirable.

# Part IV: Implementation

# 13

# Risk Assessment

> *More than any other time in history, mankind faces a crossroads. One path leads to despair and utter hopelessness. The other, to total extinction. Let us pray we have the wisdom to choose correctly.*
>
> *Woody Allen*

The prospects facing risk managers may not be as bleak as those confronting Woody Allen. Nevertheless, without proper management, the consequences can be unnecessarily severe. Such management begins with an assessment of those risks, the topic of this chapter.

## 13.1 Environmental Risk Model

London's fog of 1952 ranks among the most famous space–time processes in history, even though it lasted only a few days (see Bates and Caton 2002 for a description). The BBC Web page quotes one observer, Barbara Fewster, on recalling her 16-mile walk home, in heels, guiding her fianceé's car:

> It was the worst fog that I'd ever encountered. It had a yellow tinge and a strong, strong smell strongly of sulphur, because it was really pollution from coal fires that had built up. Even in daylight, it was a ghastly yellow color.

The ensuing sharp rise in mortality could be attributed to it unambiguously. In time the public's interest in environmental risk (ER) grew along with measures in the United fKingdom to reduce it.

The decades since have seen a sharp rise in societal concern about ER and demands for its reduction. The United States passed its Clean Air Act in 1970 and created the Environmental Protection Agency (EPA) in 1971. The NAPAP program (see Example 10.1) was launched in 1980 to:

- Specify the cause and origin of acid deposition.
- Assess the impact on environment, society, and economy caused by acid deposition.
- Remove or weaken the harmful impacts of acid deposition by the regulation or elimination of the discharge of original materials, on the basis of the research results.

That program produced a lot of statistical work. This may have been due to the subtlety of the environmental impacts of concern; though small, they can be pervasive and hence their overall effect enormous. Anyway that work offered a lot of analysis about such things as gradients and trends on the one hand, as well as new theory for risk assessment on the other. In particular, it gave new approaches to modeling space–time fields (Le and Zidek 1992; Sampson and Guttorp 1992).

Recent years have seen a surge in ER studies, so far beyond the scope of a single chapter that we limit ourselves to a brief overview. As well this chapter presents methodology for ER assessment (ERA) tailor-made for what is called *longitudinal* or *time-series* analysis, one that minimizes computational complexity while tying in with predictive exposure distributions (like that in Chapter 9).

We begin with an overview.

## 13.2 Environmental Risk

Informally, quantitative risk is sometimes defined as "consequence $\times$ probability" (personal communication, John Lockwood). That definition reflects the need to average the product of a measure of the loss produced by a risky outcome and its likelihood.

Though simple in concept, this formula proves difficult to apply. For one thing, numerical losses are hard to estimate. (How much is the cost of a life?) Uncertainties about the outcomes of the judicial processes needed to settle claims add to the challenge. Finally, losses are borne, not by one individual, but by a group of "stakeholders," such as insurance companies and the bereaved.

The probabilities are also hard to specify since by nature, risky outcomes are rare. They may represent extremes (Chapter 12) and include such things as a one hundred year flood, an earthquake, or the failure of a nuclear power plant. Consequently, they generate so little data for estimating those probabilities that lots of uncertainty remains, even when statistical models such as those in Chapter 12 are used.

ERA, that is, monitoring, data capture, data analysis, interpretation, conclusions, and implications, concerns the impacts of an environmental hazard (EH). These impacts can be on living things (see Examples 1.1 and 1.2), the subject of environment health risk. But they can also be on nonliving things such as building materials that are eroded by acid precipitation (see Examples 4.1 and 4.4). Graphically, ERA may be portrayed as $EH \longrightarrow exposure \longrightarrow dose \longrightarrow response \longrightarrow impact$. This diagram covers two major components of ERA, namely, environmental exposure assessment (EAA) and environmental toxicology assessment (ETA) (Bailer et al. 2003). Very briefly, the former concerns the level of exposure, the latter, the dose–response, i.e., how much of that exposure will be converted into a response.

As noted by Bailer et al. (2003):

> . . . the definition of impact may not always be obvious.

In fact, the EHs may not even be known (the known unknowns or even unknown unknowns in Rumsfield's terminology (Chapter 3). For instance, the health impacts of particulate air pollution became a concern only in the 1990s whereas in contrast, the impact of ozone was studied in the 1970s.

EHs include such things airborne particles, gases as well as fumes, contaminants of water as well as subsoils, pesticides, metals and solvents as well as vapors, radiation, and radioactive materials (Sen 2003). Quite a list!

Membership on this list is controversial. Should electromagnetic radiation be on it? What about cell-phone radiation? Critics still question the presumption that airborne particles cause illhealth. Controversy about membership on the list has stimulated a whole new field of inquiry called "environmental epidemiology," the concern of much of this chapter because of its close ties to environmental space–time processes analysis.

The following example illustrates ERA.

*Example 13.1. Pollution versus mental development*
Budtz-Jørgensen et al. (2003) study the consumption of pilot whale meat in the Faroe Islands because it exposes the population to methylmercury. They focus on a 1986-87 birth cohort of 1022 children in a prospective study of possible adverse effects of prenatal exposure, that being determined from both umbilical cord blood and maternal hair.

At age 7 (years), 90% of the cohort members were tested in a variety of ways centering on nervous system function, in particular, by using the Boston Naming Test (BNT). There each child is presented with a sequence of drawings of objects and asked to name each one. Those who do not respond correctly at each stage are given two successively more helpful cues. Finally, the investigators compute two scores, the totals correct with and without cues.

The study shows a strong association between ethylmercury concentrations and the scores, especially the cued scores. Specifically, a tenfold increase in that environmental hazard leads to a predicted drop of 1.6 in that score.

### Confounders

Example 13.1 proves instructive. First the study described there like most, is observational: treatment levels are decided by nature, not the experimenter through randomization. That means the association can come through one or more *confounders*, the latter being in the words of Budtz-Jørgensen et al. (2003) quoting Miettenin,

> an extraneous determinant of the response which has imbalanced distributions between the categories of the exposure.

As an example, these authors note that socioeconomic status can generally be a confounder since a high status tends to imply good health, a low status, greater exposure to the hazard. Thus the confounder can skew the results in favor of a positive association by amplifying the exposure contrasts.

Population size can be a confounder in ecological studies that use regional aggregates, in other words cluster statistics. Populous regions generate both larger aggregate adverse health counts and higher aggregates of the EH being studied for its association with those counts. Symbolically, more people $\longrightarrow$ more hazard $\longrightarrow$ more bad health. Naive analyses do not adjust for confounders and discover totally spurious but often strong associations between the two.

Ideally studies should be conducted separately within each confounder category. However, in practice that will almost never be possible, there being too many confounders (known and unknown) and thus a huge number of categories created through cross-classification. Instead regression modeling strategies are used, all identified potential confounders being measured and included in the model. That way results can be adjusted for them.

However, not all potential confounders are "known unknowns." Long ago randomized experiments were proposed to deal with that problem and ensure that the two populations being compared are identical with respect to all potential confounders. Yet in practice, randomization is seldom possible.

### Longitudinal Studies

To deal with this difficulty an ingenious design was developed to enable within-cluster comparisons. More precisely, response and hazard levels are tracked over time within each cluster. An association can claim causality credentials since within each cluster the confounders, such as aggregate socioeconomic status, will be relatively stable over time.

The inferential approach associated with such a design, *longitudinal data analysis*, is an important tool for environmental epidemiology. In contrast, the classical approach called *cross-sectional data analysis*, has lost much of its popularity.

### Risk Analysis

The topic of *risk analysis* completes our review and includes both ERA as well as risk management. It entails such things as intervention or abatement, regulation or control, and penalizing offenders. It may also include the tricky tasks of risk communication.

How the latter should best be done remains a mystery especially since the probabilities are small and the consequences ill-defined. Fascinating studies (Kahneman et al. 1982) have shown how badly experts and nonexperts can be at reasoning about uncertainty and making decisions involving it. Moreover, the perception of risk may deviate markedly from reality.

We illustrate the difficulty with an amusing example included not because it concerns environmental risk but because its simplicity makes the point so eloquently.

*Example 13.2. Prisoner's paradox*
This famous example, originally phrased as the three prisoners paradox, comes from an American TV game show involving three unmarked doors. Behind one is hidden a valuable prize. The contestant selects a door and the show's host opens one of the other two, revealing no prize. Should the contestant then switch his or her guess to the other of the unopened pair of doors as the host invites him or her to do?

In fact, the contestant should switch.Yet most do not. Moreover, most are unable to figure out why when told they should have done so to increase their chances of winning.

## 13.3 Risk in Postnormal Science

In the latter part of the twentieth century, growing concern about the risks associated with environmental processes coupled with the technology needed to address it led to a transformation of normal science into postnormal science [Funtowicz and Ravetz(undated)]. Stereotypically normal science, driven by curiosity, the need for reproducibility, and the quest for truth, led to carefully controlled laboratory experiments. In contrast postnormal science is associated with complexity such as that seen in environmental processes. More than just complicated, the latter has great uncertainty and a multitude of perspectives associated with it. In Funtowicz and Ravetz(undated) we find it described as follows.

> For policy purposes, a very basic property of observed and analyzed complex systems might be called "feeling the elephant," after the Indian fable of the five blind men trying to guess the object they were touching by feeling a part of an elephant. Each conceived the object after his own partial imaging process (the leg indicated a tree, the side a wall, the trunk a snake, etc.); it is left to an outsider to visualize the whole elephant.

Such is the situation confronting a crew of environmental scientists analyzing an environmental process over a very large space–time domain. Uncertainty abounds while the level of risk is uncertain. Decisions are urgent. Postnormal science must bridge between the process and the policy-maker. It is driven by stakeholders with different values, who may see their levels of risk differently, along with the funding envelopes that increasingly drive the science and decide the measurements to be taken. Whole new approaches to science are developing.

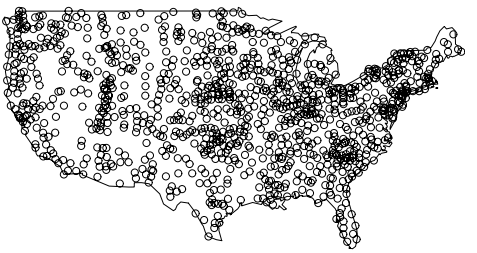

**Fig. 13.1.** Location of 1221 historical meteorological monitoring stations.

The following example concerns one of the processes about which so much controversy has arisen, global climate change. While limitations of space force its superficiality, it does illustrate some of the complexity described above.

*Example 13.3. Global climate change*
This examples concerns the maximum monthly temperatures since 1884 over the coterminus United States. Historical data on this and other environmental processes can be found on the U.S. Historical and Climatological Network (HCN). The data from that source we look at have been adjusted for such things as the effects of urbanization to reveal, ignoring model uncertainty, the intrinsic changes in those maxima over that time period. Our focus on these maxima derives from the argument that extreme values more sensitively indicate change than, say the mean temperature.

The HCN network involves 1221 individual monitoring stations. Figure 13.1 shows the overall network to be quite dense.

Just ten of those stations were chosen for our purposes, one from each state that monitored climate since at least 1890. Their locations in Figure 13.2 reveal they are widely distributed.

Turning to the issue of climate change, we plot in Figure 13.3 the time-series of the annual averages of the monthly maxima for the ten sites in our investigation. A lot of different pictures emerge from that plot. Some temperatures seem to have been trending upwards, others down. In fact, the two hottest states among the ten, Texas and Mississippi seem to be trending in opposite directions. We take a closer look at them in the next two figures. Someone in Mississippi looking at this plot in 1980 might be forgiven for believing an ice age was well on its way. In contrast, except for a brief 20-year

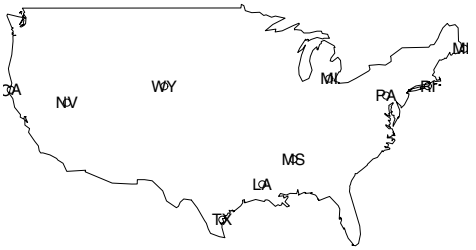


**Fig. 13.2.** Location of the ten historical meteorological monitoring stations selected for analysis because of their long historical records.

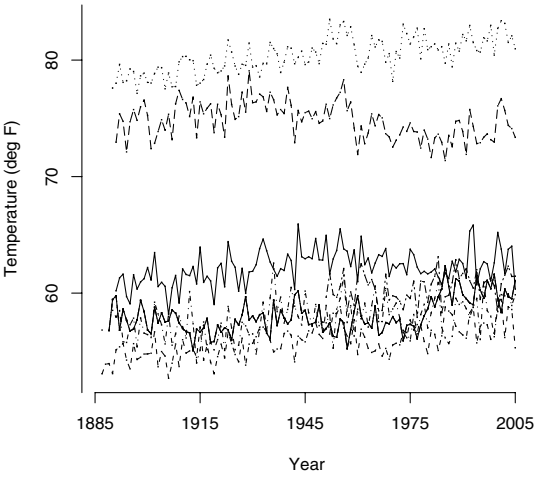


**Fig. 13.3.** Time-series plots of the annual average of monthly maximum temperatures for ten selected sites in the United States, 1884–2003.

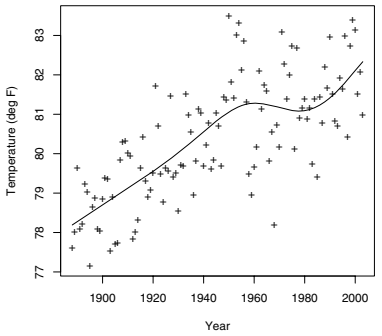


Fig. 13.4: Time-series of annual averages of monthly maximum temperatures (deg F) for a site in Texas.

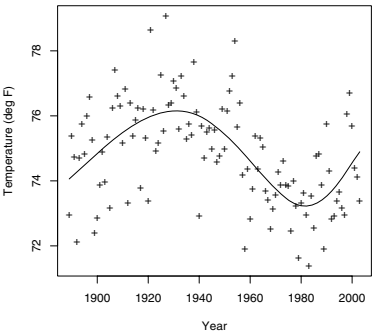


Fig. 13.5: A plot for Mississippi like that on the left.

period, at any time during the last century, a Texan might have concluded global warming was well underway.

We now turn to the topic of environmental epidemiology, a particularly important one in postnormal science.

## 13.4 Environmental Epidemiology***

Environmental epidemiology concerns the relationship between human health and environmental hazards. Increasing knowledge and societal concern have led to a substantial increase in the support for such studies. Moreover, regulatory agencies have needed and used the results they report to shape abatement programs as well as for regulatory and control policies.

However, the health effects involved are generally so subtle that increasingly sophisticated statistical methods have had to be developed. These methods are needed to detect those effects as well as quantify uncertainties in any inferences that might be made about them using available data. In practice policy-makers need accurate estimates of uncertainty to support action when the negative estimates warrant it. Moreover, such action will be warranted when large populations are at risk even when the effects are subtle.

Perversely, many departures from the assumptions underlying the methods used to quantify uncertainty (and risk) tend to produce misleadingly low estimates. Those estimates may then lead policy-makers to be overconfident about the issue of concern. For example, dependent data yield higher confidence or credibility regions than independent data. If the latter is assumed, uncertainty about the associated estimates will be misleadingly small. (Dependent data

contain less information than uncorrelated data. In the extreme case of perfect dependence the data actually consist of a single datum!) Consequently recent years have seen increased emphasis on those departures, diagnostic tools for detecting them, and ways for getting around them.

This chapter presents an environmental health risk assessment method for longitudinal data (Zidek et al. 1998a) designed to deal with one of those departures, namely, measurement error. Such error can have unpredictable and potentially pernicious effects on the estimates and their associated levels of uncertainty (Chapter 4).

**Measurement Error and Estimating Equations**

The method, developed in Duddek et al. (1995) and refined by Zidek et al. (1998a), extends one of Burnett and Krewski (1994) as well as Lindstrom and Bates (1990). It addresses that error by imputing exposures at unmonitored sites using a predictive distribution conditional on the observed levels of the hazardous substance of concern. In the terminology of Carroll et al. (1995) we are using regression calibration.

Another feature of the method is its use of the well-known generalized estimating equations (GEE) approach for fitting a health effect (impacts) model. Although other methods such as Poisson regression are commonly used instead, the GEE enjoys robustness against model misspecification (with large samples) and simplicity. Thus only first and second moments of the predictive distribution are needed. The resulting computational simplicity enables the method to handle large problems such as that in Section 13.5.

**Data Clusters**

The method assumes the data are clustered. A *cluster* can consist of data from a single subject such as a time-series of measurements made on him. Alternatively, it could be from a group of individuals defined by a characteristic, such as having a home address in a specified geographical area.

The method assumes *random effects* represent cluster contributions to the measured responses. The Bayesian framework provides one justification for that assumption. Then randomness represents uncertainty about the cluster parameters. However, non-Bayesians have also embraced this approach, allowing them to assume these cluster effects have a joint distribution, and creating a soft but powerful link between clusters. In particular, these effects can be marginalized out across clusters and strength can be borrowed. Thus small, even insignificant individual cluster effects can attain significance if they point consistently in the same direction. That provides a powerful tool for ERA.

### 13.4.1 Impact Assessment***

To make things more precise, suppose a study involves $K$ clusters and $T$ times. For a given cluster–time pair $i$ and $t$, $i \in \mathcal{I}, t \in \mathcal{T}$, $Y_{it}$ denotes a measurable

response. That response could be the number of hospital admissions for respiratory morbidity, the number of school absences, or the number of organisms that die. At the same time, $\mathbf{X}_{it}^{(1)}$ represents an associated vector of covariates, some measured with error and some not. Set $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})^T$, the vector of all cluster $i$ responses over time and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_I)$, the combination of all cluster responses.

### Regression Model

We begin with the regression model

$$E(Y_{it} \mid \mathbf{a}_i, \mathbf{X}_{it}) = \zeta_{it} = m_{it}\zeta(\mathbf{a}_i^T \mathbf{X}_{it}) \tag{13.1}$$

that links the measurable response to the covariates in cluster $i$ at time $t$. The constant $m_{it}$ can account for such things as the population size of cluster $i$ and slowly varying seasonal patterns in health outcomes, when the study concerns only acute effects. Notice the simplification imposed, of making $\zeta_{it}$ depend on the covariate vector $\mathbf{X}_{it}$ only through the function of one variable $\zeta$ and the linear form $\mathbf{a}_i^T \mathbf{X}_{it}$. Much more complicated relationships could be handled at the cost of model complexity. However, even the simple model captures many commonly used impact functions such as the one in Section 13.5. For simplicity, divide both sides of Equation (13.1) by $m_{it}$ to get

$$E(Y_{it} \mid \mathbf{a}_i, \mathbf{X}_{it}) = \zeta(\mathbf{a}_i^T \mathbf{X}_{it}), \tag{13.2}$$

where with an abuse of notation now the new $Y_{it}$ is the old one divided by $m_{it}$.

The subtlety of the effects generally makes $\mathbf{a}_i^T \mathbf{X}_{it}$ small. Thus, if we let $\mathbf{X}_{it}^o$ represent a vector of baseline values for the coordinates of $\mathbf{X}_{it}$ (say the "typical" values in cluster $i$ at time $t$), we might well approximate $\zeta$ with the first two terms in its Taylor expansion. To get that expansion, let $\zeta'(s) = d\zeta(s)/ds$, $s \in (-\infty, \infty)$ denote $\zeta$'s first derivative. Then approximately, the expected number of adverse outcomes is given by $\zeta_{it} \approx \zeta(\mathbf{a}_i^T \mathbf{X}_{it}^o) + \zeta'(\mathbf{a}_i^T \mathbf{X}_{it}^o)\mathbf{a}_i^T(\mathbf{X}_{it} - \mathbf{X}_{it}^o) = \zeta(\mathbf{a}_i^T \mathbf{X}_{it}^o)[1 + \mathbf{R}_{it}(\mathbf{X}_{it} - \mathbf{X}_{it}^o)]$, where $\mathbf{R}_{it} = [\zeta'(\mathbf{a}_i^T \mathbf{X}_{it}^o)/\zeta(\mathbf{a}_i^T \mathbf{X}_{it}^o)]\mathbf{a}_i^T$. In particular, the $j$th coordinate of $R_{it}$; i.e., $R_{itj}$ is called the relative risk of $\mathbf{X}_{itj}$. If we interpret $\zeta(\mathbf{a}_i^T \mathbf{X}_{it}^o)$ as the baseline number of expected outcomes at time $t$ (normalized by $m_{it}$) then $100 \times R_{itj}$ would be the % change in that number as a result of a unit change in the level of $\mathbf{X}_{itj}$ above its baseline.

Relative risks are standard indices of risk in environmental epidemiology. Often they are stated not for just a unit change in the environmental hazard but some other change, for example, 10 $\mu$g m$^{-3}$ in the case of $PM_{10}$. The conversion is easy; just multiply by 10 in the latter case. If this were, say 2%, it would represent a seemingly small relative risk of mortality. Yet in a cluster with a population size of 100,000 about 2,000 excess deaths would occur. However subtle, this effect could hardly be called negligible! Of course,

accurate estimates of the uncertainty in that estimate would be needed to support regulatory action and intervention.

Finally, note that $\mathbf{X}_{it}$ can be replaced by $\mathbf{X}_{it} - \mathbf{X}_{it}^o$ in the regression model when acute effects are of concern. (That change was made in the Case Study reported in Section 13.5.) There it's the deviation from baseline that would more likely be associated with abrupt changes in impact levels. As a desirable byproduct, the degree of collinearity between $\mathbf{X}$'s coordinate responses would thereby be reduced. After all, much of that problem derives from the cross-correlation induced by long term trend patterns.

*Example 13.4. Logistic model*
Let $\zeta(s) = \exp(s)/[1 + \exp(s)]$, the so-called logistic model. Then using the multiplication rule for differentiation,

$$\begin{aligned}\zeta'(s) &= \exp(s)/[1 + \exp(s)] - \exp(s)/[1 + \exp(s)]^2 \times \exp(s) \\ &= \zeta(s)[1 - \zeta(s)].\end{aligned}$$

Then

$$R_{it} = [1 - \zeta(\mathbf{a}_i^T \mathbf{X}_{it}^o)]\mathbf{a}_i^T .$$

Notice that $\zeta(s) \to 1$ as $s \to \infty$. Thus in agreement with intuition the relative risk tends to zero as $\mathbf{a}_i^T \mathbf{X}_{it}^o \to \infty$; a small change in a high baseline level would produce a smaller effect than a similar change in a low level. Clearly the baseline needs to be selected carefully. Yet that important issue receives little attention.

This model makes the effect of the baseline's deflation factor $1 - \zeta(\mathbf{a}_i^T \mathbf{X}_{it}^o)$ common to all elements of $\mathbf{X}$. This property could be unrealistic in some situations, a cost of the model's simplicity. On the other hand, the relative risk would be monotonically increasing from approximately zero for very small values of $a_{kj} X_{itj}^o$, and bounded (if $a_{ij} > 0$), seemingly natural properties.

A much more common regression model, the one we use in Section (13.5), appears in the next example. It like that in the previous example which ensures the fitted means are positive (unlike the linear model, say).

*Example 13.5. Exponential model*
Let $\zeta(s) = \exp(s)$. Now $R_{it} = a_k^T$ so the effect of hazards in $X$ does not depend on their size relative to the baseline. Moreover, that risk does not depend on time. We wonder if these properties are biologically justified. In particular, would the impact of a unit's increase in ozone above its springtime baseline have the same relative risk as in summer say, another issue that has received little attention in environmental epidemiology. The conclusion implied by this model, that the risk is unbounded, is not justified. In fact, the model should only be regarded as an approximation to that in Example (13.4) when $\mathbf{a}_i^T \mathbf{X}_{it}$ is small.

### Random Effects

Equation (13.2) involves a vector of regression coefficients $\mathbf{a}_i$ for cluster $i$. As noted above, advantages accrue from taking as random at least some of the coordinates in that vector. Therefore assume $\mathbf{a}_i = \beta + \mathbf{b}_i$, $\{\mathbf{b}_i\}$ being the random cluster $i$ effects vector with vector mean zero and covariance matrix $\mathbf{D}$. (To make a coordinate of $\mathbf{a}_i$ nonrandom, simply set to zero the variance in $\mathbf{D}$ for the associated coordinate of $\{\mathbf{b}_i\}$. This convention considerably simplifies the model, bypassing the need to split the linear form $\mathbf{a}_i^T \mathbf{X}_{it}$ into two pieces, one fixed and one random as is commonly done.)

The risk assessment of an environmental hazard commonly entails a test of the hypothesis of no association with the measured health response. More precisely, for the hazard's coefficient in the regression model $\mathbf{a}_{ij}$ that (null) hypothesis $H_o$ would state either $\mathbf{a}_{ij} = 0$ or $\mathrm{Var}(\mathbf{a}_{ij}) = 0$, accordingly as that coefficient represents a fixed or random effect.

### "Working" Error Covariance

However, the need to borrow strength across clusters requires in addition to the model in (13.2), one for its covariance. For that model we make the "working assumption" that the outcome responses are not correlated across time. That assumption, although unrealistic can be made since the GEE approach features a robust covariance estimate that overcomes that deficiency (Liang and Zeger 1986). To express that assumption, let $\delta$ be the Dirac delta function. In other words, $\delta_{uv} = 0$ unless $u = v$ when it is 1. Now assume

$$Cov\,(Y_{it_1}, Y_{it_2} \mid \mathbf{a}_i, \mathbf{X}_{it_1}, \mathbf{X}_{it_2}) = \phi \zeta_{it_1} \delta_{t_1 t_2},$$

$\phi$ being an unknown real number called the *overdispersion* parameter.

We may not have observed all elements of the covariate vector $\mathbf{X}_{it}$. That is after all the point of our analysis. However, assume we can find with an approach like that in Chapters 9 and 10,

$$\begin{aligned} E\,(\mathbf{X}_{it}) &= \mathbf{z}_{it} \\ Cov\,(\mathbf{X}_{it_1}, \mathbf{X}_{it_2}) &= \mathbf{G}_{it_1 t_2}. \end{aligned} \tag{13.3}$$

Set to 0 elements of the latter for covariates measured without error. Finally let $\mathbf{G}_{it_1 t_2} = \mathbf{0}$ when $t_1 \neq t_2$.

### Practical Approximations

Nonlinear models generally prove quite intractable in statistical analysis. In particular, their parameters cannot be estimated without resorting to numerical algorithms, the subject to which we now turn. These algorithms find optimal estimates through iterative approximation, each iteration requiring the local linearization of any nonlinear functions involved. Developing fast accurate algorithms has proven quite challenging. Nonetheless, functions with a

large number of parameters may defeat the best such algorithms because of insufficient computer memory or of the existence of excessively many local optima, making the resulting estimates depend critically on the initial estimates supplied.

We begin with some assumptions needed to ensure the positive definiteness of the approximate covariance matrix we develop below. Assume $\zeta$ is: (1) positive; (2) three times differentiable; (3) strictly log convex. The latter makes the second derivative of its logarithm is positive; i.e., $\zeta''(u)/\zeta(u) - (\zeta'(u)/\zeta(u))^2 > 0$. This in turn makes $0 \leq \zeta(u) + 2v\zeta'(u) + v^2\zeta''(u)$, for all $u$ and $v$ since this quadratic cannot then have any real roots when set to zero. In other words it can never be zero and must therefore always be positive, the positive definiteness condition required below.

Suppose also the mean in Equation (13.3) approximates the unmeasured $X$s reasonably well. Then we can make a Taylor expansion and compute the expectation of the result, dropping terms of order higher than two. The resulting approximation is denoted $\approx$ meaning, *approximately equal to.*

To begin, recall the general identity $E[U] = E[E(U|V)]$ a result that holds for any pair of random vectors. Then

$$\begin{aligned} E\left(Y_{it} \mid \mathbf{a}_i\right) &= EE[(Y_{it} \mid \mathbf{X}_{it}, \mathbf{a}_i] \mid \mathbf{a}_i) \\ &= E\left(\zeta(\mathbf{a}_i^T \mathbf{X}_{it}) \mid \mathbf{a}_i\right) \\ &\approx E\left(\zeta(\mathbf{a}_i^T \mathbf{z}_{it}) + \zeta'(\mathbf{a}_i^T \mathbf{z}_{it})\mathbf{a}_i^T(\mathbf{X}_{it} - \mathbf{z}_{it}) + \right. \\ &\quad \left. \zeta''(\mathbf{a}_i^T \mathbf{z}_{it})\mathbf{a}_i^T(\mathbf{X}_{it} - \mathbf{z}_{it})(\mathbf{X}_{it} - \mathbf{z}_{it})^T \mathbf{a}_i \mid \mathbf{a}_i\right) \\ &= \zeta(\mathbf{a}_i^T \mathbf{z}_{it}) + \zeta''(\mathbf{a}_i^T \mathbf{z}_{it})\mathbf{a}_i^T \mathbf{G}_{itt}\mathbf{a}_i. \end{aligned}$$

Using the approximation above, we circumvent the problem created by $\zeta$'s nonlinearity in $\mathbf{X}$. However, that in $\mathbf{a}$ remains. To deal with it we use an idea of Lindstrom and Bates (1990) and let $\beta_{oi} = \beta + \mathbf{b}_i^o$, $\mathbf{b}_i^o$ representing a current estimator of cluster $i$'s random effect. This estimator would presumably be fairly close to the optimum at least after a number of iterations. This gives a basis for a further Taylor expansion approximation. The result (with details omitted):

$$\begin{aligned} E\left(Y_{it} \mid \mathbf{a}_i\right) &\approx \eta_{it}(\mathbf{a}_i) \quad \text{where} \qquad (13.4) \\ \eta_{it}(\mathbf{a}_i) &= \zeta\left(\beta_{oi}^T \mathbf{z}_{it}\right) + \hat{\mathbf{Z}}_{it}\left(\mathbf{a}_i - \beta_{oi}\right) \\ &\quad + \frac{1}{2}\zeta''\left(\beta_{oi}^T \mathbf{z}_{it}\right)\left[\mathbf{z}_{it}^T\left(\mathbf{a}_i - \beta_{oi}\right)\left(\mathbf{a}_i - \beta_{oi}\right)^T \mathbf{z}_{it} + \beta_{oi}^T \mathbf{G}_{itt}\beta_{oi}\right] \\ \hat{\mathbf{Z}}_{it} &= \zeta'\left(\beta_{oi}^T \mathbf{z}_{it}\right)\mathbf{z}_{it}^T. \end{aligned}$$

While these approximations take care of the mean, finding a covariance approximation remains. The next result uses another well-known identity $Cov[U] = E[Cov(U|V)] + Cov[E(U|V)]$ for any pair of random vectors for which the requisite covariances exist. The result:

$$
\begin{aligned}
Cov\left(Y_{it_1}, Y_{it_2} \mid \mathbf{a_i}\right) &\approx \Lambda_{it_1t_2}(\mathbf{a}_i) \quad \text{where} \\
\Lambda_{it_1t_2}(a_i) &= \delta_{it_1t_2}\phi E\left(Y_{it_1} \mid \mathbf{a}_i\right) \\
&\quad + \zeta'\left(\mathbf{a}_i^T\mathbf{z}_{it_1}\right)\zeta'\left(\mathbf{a}_i^T\mathbf{z}_{it_2}\right)\mathbf{a}_i^T\mathbf{G}_{it_1t_2}\mathbf{a}_i.
\end{aligned} \tag{13.5}
$$

To estimate the fixed effect parameters, we need the corresponding approximations after the random effects have been eliminated by averaging them out of Equations (13.4) and (13.5). The result:

$$
\begin{aligned}
E\left(Y_{it}\right) &\approx \mu_i(\beta_{oi}); \\
\mu_i(\beta_{oi}) &= \zeta\left(\beta_{oi}^T\mathbf{z}_{it}\right) + \hat{\mathbf{Z}}_{it}\left(\beta - \beta_{oi}\right) + \frac{1}{2}\zeta''\left(\beta_{oi}^T\mathbf{z}_{it}\right) \\
&\quad \times \left\{\mathbf{z}_{it}^T[\mathbf{D} + (\beta - \beta_{oi})]\left(\beta - \beta_{oi}\right)^T\right]\mathbf{z}_{it} + \beta_{oi}^T\mathbf{G}_{itt}\beta_{oi}\}; \\
Cov\left(Y_1, Y_2\right) &\approx \Sigma_{12}(\alpha_o) \quad \text{where} \\
\Sigma_{it_1t_2}(\beta_{oi}) &= \Lambda_{it_1t_2}(\beta_{oi}) + \hat{\mathbf{Z}}_{it_1}\mathbf{D}\hat{\mathbf{Z}}_{it_2}^T.
\end{aligned} \tag{13.6}
$$

The regularity conditions ensure the positive definiteness of $\Lambda_{itt}$ and $\Sigma_{it_1t_2}(\beta_{oi})$.

The approximations above can be compactly summarized using vector-matrix:

$$
\begin{aligned}
E\left(\mathbf{Y}_i \mid \mathbf{a_i}\right) &\approx \left(\eta_{it_1}(\mathbf{a_i}), \ldots, \eta_{\mathbf{it_T}}(\mathbf{a_i})\right)^T \\
&\equiv \eta_i(\mathbf{a_i}); \\
Cov\left(\mathbf{Y}_i \mid \mathbf{a_i}\right) &\approx diag\left(\Lambda_{it_1t_1}(\mathbf{a_i}), \ldots, \Lambda_{it_Tt_T}(\mathbf{a_i})\right) \\
&\equiv \Lambda_i(\mathbf{a_i}); \\
E\left(\mathbf{Y}_i\right) &\approx \left(\mu_{it_1}(\beta_{oi}), \ldots, \mu_{it_T}(\beta_{oi})\right)^T \\
&\equiv \mu_i(\beta_{oi}); \\
Cov\left(\mathbf{Y}_i\right) &\approx \Lambda_i(\beta_{oi}) + \hat{\mathbf{Z}}_i\mathbf{D}\hat{\mathbf{Z}}_i^T \\
&\equiv \mathbf{\Sigma}_i(\beta_{oi}); \\
\hat{\mathbf{Z}}_i^T &= \left(\hat{\mathbf{Z}}_{it_1}^T, \ldots, \hat{\mathbf{Z}}_{it_T}^T\right).
\end{aligned}
$$

This summary simplifies programming in object-oriented programming languages such as R that can handle vector and matrix objects.

### The Burnett–Krewski Approach

This section presents an adaptation of methods of Burnett and Krewski (1994) for nonlinear regression. The results yield two kinds of analysis. The first is *cluster-specific* (Zeger et al.1988) where $\beta$ reflects the response's change at a typical cluster due to a covariate's change while the $\{\mathbf{b}_i\}$ model response rates among different clusters. The second is called *population average*, to which we turn presently.

To implement the GEE approach we (unrealistically) assume that conditional on the random effects, the $\{Y_{it}\}$ have joint normal probability density and use the approximations in the previous section for the required means

and covariances. That joint distribution (with clusters) is assumed to be stochastically independent) provides a *quasi-likelihood* for the parameters. We may combine it with the prior distribution for the random effects to obtain a *quasi-posterior* distribution for those parameters conditional on the data and all parameters/hyperparameters but those in the $\{\mathbf{b}_i\}$:

$$\Pi_i \pi(\mathbf{b}_i \mid \mathbf{y}_i, \beta, \ldots) \propto \Pi_i exp\left\{ \frac{[\mathbf{y}_i - \eta_i(\beta + \mathbf{b}_i)]^T \Lambda_i^{-1} [\mathbf{y}_i - \eta_i(\beta + \mathbf{b}_i)]}{2} - \mathbf{b}_i^T D^{-1} \mathbf{b}_i \right\}. \tag{13.7}$$

$D$ has been augmented as necessary to make it nonsingular and exposition easier.

Their estimates may be found by solving the estimating equations obtained by setting equal to zero, the derivative (or more properly, *column gradient*) of the posterior's logarithm with respect to the random effects vector. In deriving those equations, we fix $\Lambda_i \equiv \Lambda_i(\mathbf{a}_i)$ at $\mathbf{a}_i = \beta_{oi}$. Additional simplification obtains from evaluating the gradient of $\eta_{it}$ at $\mathbf{a}_i = \beta_{oi}$ to get $\hat{Z}_{it}$ ($\hat{Z}_i$ the corresponding vector). Thus for the random effects vector $\mathbf{b}_i$ we obtain the estimating equations:

$$\mathcal{W}_i \equiv \hat{\mathbf{Z}}_i^T \Lambda_i^{-1} (\mathbf{y}_i - \eta_i(\beta + \mathbf{b_i})) - \mathbf{D}^{-1} \mathbf{b}_i = 0. \tag{13.8}$$

To solve these equations Fisher's scoring algorithm is used. That first entails computing a matrix, the *row* gradient of $\mathcal{W}_i$ and taking the expectation of the result with respect to $\mathbf{b}_i$. That matrix is given by:

$$\mathbf{A} = -\hat{Z}_i^T \Lambda_i \hat{Z}_i - D^{-1}.$$

The iterative solution of Equation (13.8) proceeds at the next step by finding $\hat{\mathbf{b}}_i^*$ as the solution of

$$A\hat{\mathbf{b}}_i^* = A\hat{\mathbf{b}}_i - \mathcal{W}_i.$$

To put this last equation into a more explicit form uses a matrix identity:

$$(P + A^T R Q)^{-1} = P^{-1} - P^{-1} Q^T (R^{-1} + Q P^{-1} Q^T)^{-1} Q P^{-1}.$$

Applying that identity gives

$$\hat{\mathbf{b}}_i^* = D \hat{Z}_i^T \Sigma_i^{-1} \tilde{r}_i, \tag{13.9}$$

where $\tilde{r}_i = y_i - \eta_i(\beta + \hat{\mathbf{b}}_i) + \hat{Z}\mathbf{b}_i$, a surprisingly simple result. Recall that we had augmented $D$ to make it nonsingular. Now reset to zero the affected elements. This last equation then returns a 0 for each of the fixed effects.

To obtain estimating equations for $\beta$ we proceed in a similar fashion, this time with the approximations obtained in the previous section after eliminating the random effects. A uniform prior distribution for $\beta$ is adopted for convenience (Zidek et al. 1998b). That is equivalent to finding the maximum quasi-likelihood estimator, to which we now turn. After taking a logarithm

and multiplying the result by $-2$, the quasi-likelihood for $\beta$ and the hyperparameters becomes

$$\sum_i \log|\Sigma_i| + \mathbf{r}_i^T \Sigma_i^{-1} \mathbf{r}_i, \tag{13.10}$$

where $\mathbf{r}_i = \mathbf{y}_i - \mu_i(\beta + \hat{\mathbf{b}}_i)$. On setting to zero this expression's (column) gradient with respect to $\beta$ we get the estimating equation

$$\sum_i \hat{X}_i^T \Sigma_i^{-1} \mathbf{r}_i = 0,$$

$\hat{X}_i^T : T \times I$ being the matrix obtained by computing $\mu$'s gradient with respect to $\beta$. (Its leading term is just $\hat{Z}_i$.) Again we can appeal to Fisher's scoring algorithm and get updated estimate

$$\beta^* = \mathbf{H} \sum_i \hat{X}_i \Sigma_i^{-1} \mathbf{r}_i, \tag{13.11}$$

where $\mathbf{H} = \sum_i \hat{X}_i^T \Sigma_i^{-1} \hat{X}_i$.

It only remains to estimate $D$ and $\phi$ using the quasi-likelihood. The approach of Laird and Ware (1982) yields the updating equations for this purpose:

$$\hat{D}^* = \hat{D} + \hat{D} \left( I^{-1} \sum_i \hat{Z}_i^T \Sigma_i^{-1} (\mathbf{r}_i \mathbf{r}_i^T - \Sigma_i) \Sigma_i^{-1} \hat{Z}_i \right) \hat{D} \tag{13.12}$$

$$\hat{\phi}^* = \hat{\phi}(IT)^{-1} \sum_i \mathbf{r}_i^T \Sigma_i^{-1} \mathbf{r}_i. \tag{13.13}$$

Now if the $\{\mathbf{b}_i\}$ had been observed $D$ could simply be estimated by $I^{-1} \sum_i \mathbf{b}_i \mathbf{b}_i^T$. Since they are not, use of the EM algorithm (see Chapter 10) is suggested. Thus in simplified notation we seek $D = \arg\max_D E_{old}[\log f(\mathbf{y}, \mathbf{b})]$, where $E_{old}$ denotes the conditional expectation given $\mathbf{Y}$ and the hyperparameter estimates from the previous iteration including $D$. However, the new $D$ appears in $\log f(\mathbf{y}, \mathbf{b})$ only in the prior density for $\mathbf{b}$. Thus

$$\begin{aligned}
\hat{D}^* &= \arg\max_D E_{old} \log f(\mathbf{y}, \mathbf{b}) \\
&= \arg\max_D E_{old}[\log|D| + tr D^{-1} \sum_i \mathbf{b}_i \mathbf{b}_i^T] \\
&= E_{old}[I^{-1} \sum_i \mathbf{b}_i \mathbf{b}_i^T] \\
&= I^{-1} \sum_i \mathrm{Cov}_{old}(\mathbf{b}_i) + E_{old}[\mathbf{b}_i] E_{old}[\mathbf{b}_i]^T \\
&= I^{-1} \sum_i \hat{D} - \hat{D} \hat{Z}_i \Sigma_i^{-1} \hat{Z}_i^T \hat{D} + \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^T.
\end{aligned}$$

The last result gives us Equation (13.12).

The same approach yields for $\hat{\phi}$,

$$\hat{\phi}^* = \hat{\phi}(IT)^{-1} \sum_i \mathbf{r}_i^T \Sigma_i^{-1} \mathbf{r}_i + \hat{\phi}(IT)^{-1} \sum_i (I - \Sigma_i^{-1} \mathbf{r}_i \mathbf{r}_i^T) \mathbf{M}_i^*,$$

where $\mathbf{M}^* = I - \hat{\phi}\Sigma_i^{-1}\ diag\{\mu_{i1} \dots \mu_{iT}\}$. However, the second term in the last equation has expectation zero, making the first term an unbiased estimator. Thus it is omitted in Equation (13.13) for simplicity. This concludes our derivation of estimates for the first approach (cluster-specific) and turn in the next section to the second.

### Inference

To make inferences about estimated parameters requires the specification of their distributions, in particular, standard errors. In practice, asymptotic results including normality are often used because finite sample results are intractable. Here the GEE approach has much to offer since in particular, its asymptotic theory circumvents the potential difficulties arising from the use of unrealistic working covariances. It turns out that as long as the mean function has been correctly specified a robust covariance estimate called the sandwich estimate is available.

The idea goes back to Liang and Zeger (1986). The asymptotic robust covariance estimator for $\hat{\beta}$ suggested by Equation (13.11) is given by

$$Cov(\hat{\beta}) = \mathbf{H}[\sum_i \hat{Z}_i^T \Sigma_i^{-1} \mathbf{r}_i \mathbf{r}_i^T \Sigma_i^{-1} \hat{Z}_i]\mathbf{H}. \tag{13.14}$$

This covariance estimator can be used to construct confidence ellipsoids as well as to test hypotheses about the $\beta$ the parameter vector of central interest in environmental epidemiology.

### Population Average Versus Cluster-Specific Models

A second approach to the analysis of longitudinal health effects data (Zidek et al. 1998a) focuses on the average effect over all clusters, obtained from marginalizing the cluster-specific model or more precisely its expectation

$$E\left(Y_{it} \mid \mathbf{X}_{it}\right) = \zeta_{it}^* = \zeta_{it}^*(\beta, \mathbf{X}_{it}). \tag{13.15}$$

Now $\beta$ is the *population-average regression parameter* where before it was the regression parameter for a typical cluster. It represents the expected population response due to changes in the mean levels of the covariates. Hence Zeger et al. (1988) call this a *population-average* model. If the covariates in $X_{it}$ were at the population-level, Equation (13.15) would give their hypothetical effect.

These dual roles for $\beta$ can lead to confusion and point to the kinds of subtleties that arise when nonlinear models are used (Zidek et al. 1996). Care must be taken interpreting the results. To emphasize the point, suppose $\beta^*$ satisfies $\zeta_{it}^*(\beta^*, \mathbf{X}_{it}) = \zeta_{it} = \zeta_{it}([\beta + \mathbf{b}_i]^T \mathbf{X}_{it})$ with $\mathbf{b}_i = 0$. Then it has the population-level coefficients leading to the same impact in cluster $i$ at time $t$

as if $i$ had the regression parameters of a typical cluster. Yet generally $\beta^* \neq \beta$. In other words, the population average impact of $\mathbf{X}_{it}$ need not be the same as at a typical cluster (where $\mathbf{b}_i = 0$)! Consider the following example.

*Example 13.6. Cluster-specific versus population average*
Here

$$\zeta_{it} = exp(\beta + b_i)X_{it},$$

$X_{it}$ being one-dimensional. Furthermore $b_i \sim N(0, D)$, D being a scalar variance. Then

$$\zeta^*_{it}(\beta, X_{it}) = E[exp(\beta + b_i)X_{it}] = exp(\beta X_{it} + DX^2_{it}/2),$$

$\exp(\mu t + \sigma^2 t^2/2)$ being the moment-generating function for a Gaussian random variable with mean 0 and standard deviation $\sigma$. In contrast, at a typical cluster defined by $b_i = 0$, $\zeta_{it} = exp(\beta X_{it})$. Thus, $\beta^*$ defined above solves $\beta^* X_{it} + DX^2_{it}/2 = \beta X_{it}$. In other words, $\beta^* = \beta - DX_{it}/2$. Thus the fitted population-level parameter would have to be substantially different (smaller if the covariate were positive) from the typical cluster parameter if it were to produce the same results in that cluster at that time.

How different can these two vectors $\beta^*$ and $\beta$ be? Example 13.6 makes clear that the answer to that question depends on the between-cluster response variability since we are integrating over $\mathbf{b}_i$ to get the population-level model. They can be very different.

In practice, both methods would not be applied in the same context. Environmental health risk assessors would use the cluster-specific approach whereas environmental health risk managers would fit a population-level model.

### Cluster-Specific Models in Population Average Analysis

Cluster-specific modeling helps even when interest focuses on the population (as noted by Burnett and Krewski 1994). In particular it yields a working covariance for the population model. To get that model, evaluate the covariance matrix of the cluster-specific model at $\mathbf{b}_i = 0$ for all $i$. Replacing $\mathbf{D}$ by $\Gamma$ highlights the distinction. However, maximizing the quasi-log-likelihood remains the objective. Expanding $\zeta_{it}[(\beta^T + \mathbf{b}_i)\mathbf{X}_{it}]$ around $\mathbf{b}_i = 0$ and discarding all but the leading term yields

$$E\left(Y_{it} \mid \mathbf{X}_{it}\right) = \zeta^*\left(\beta^{*T}\mathbf{X}_{it}\right) \tag{13.16}$$

for the population regression function.

This simplification lets us borrow results from the cluster-specific analysis (13.16) as a random effects model with $\mathbf{b}_i \equiv 0$ and hence $\mathbf{D} = \mathbf{0}$. Thus

$$E[Y_{it}] \approx \nu_{it} \equiv \mu_{it}(\beta) = \zeta(\beta^T z_{kt}) + \frac{1}{2}\zeta''(\beta^T z_{it})\beta^T \mathbf{G}_{itt}\beta.$$

The equations leading to the working covariance follow.

$$\begin{aligned}
\nu_i &= (\nu_{it_1}, \ldots, \nu_{it_T})^T; \\
\mathbf{s}_i &= \mathbf{y}_i - \nu_i; \\
M_i &= \left[\zeta'\left(\beta^T \mathbf{z}_{it_1}\right)\mathbf{z}_{it_1}, \ldots, \zeta'\left(\beta^T \mathbf{z}_{it_T}\right)\mathbf{z}_{it_T}\right]^T; \\
\mathbf{V}_{it} &= \tau\nu_{it} + [\zeta'\left(\beta^T \mathbf{z}_{it}\right)]^2 \mathbf{z}_{it}^T G_{ktt} \mathbf{z}_{it}; \\
\mathbf{V}_i &= diag\left\{\nu_{kt_1}, \ldots, \nu_{kt_T}\right\}; \\
\mathbf{W}_i &= \mathbf{V}_i + M_i \Gamma M_i^T.
\end{aligned}$$

Formally, we can represent the responses as

$$Y_{it} = \nu_{it} + \mathbf{M}_{it}\mathbf{b}_i + \mathbf{U}_{it} + \epsilon_{it}, \tag{13.17}$$

where $\mathbf{M}_{it}$ denotes the $t$th row of $\mathbf{M}_i$ while $\{\mathbf{b}_i\}$, $\{\mathbf{U}_{it}\}$, and $\{\epsilon_{it}\}$ are mutually independent and normally distributed with means 0 and variances/covariances $\Gamma = Cov(\mathbf{b}_i)$,

$$Var(\mathbf{U}_{it}) = [\zeta'(\beta^T z_{it})]^2 \beta^T \mathbf{G}_{itt} \beta$$

while $Var(\epsilon_{it}) = \tau\nu_{it}$, respectively.

To maximize the quasi-log-likelihood,

$$\mathcal{P} = \sum_{i=1}^{n} \left(\ln \mid \mathbf{W}_i \mid + \mathbf{s}_i' \mathbf{W}_i^{-1} \mathbf{s}_i\right), \tag{13.18}$$

invoke the representation of the responses given in Equation (13.17) and formally appeal to results for the cluster-specific model. The recursive equations we need for fitting the working covariance matrix are obtained in that way.

## 13.5 Case Study

The study of Burnett et al. (1994) served as the genesis for Duddek et al. (1995) and Zidek et al. (1998a) (Section 13.4). That study focuses on daily hospital admission counts in Ontario due to respiratory problems for the years 1983 through 1988, classified by the hospital of admission. Strong relationships between these counts and daily concentrations of certain airborne pollutants were found.

### Rationale for the Study

However, for nonlinear models, measurement error (ME) can have quite unpredictable consequences (Chapter 4). Moreover, Burnett et al. (1994) rely on ambient monitors, in many cases far away from the hospital catchment areas involved. Thus some of the significant associations could have been an artifact of ME. That concern warranted a follow-up study by Duddek et al. (1995), successively refined in Zidek et al. (1998b) and Le et al. (1999).

The follow-up studies reanalyzed the data used by Burnett et al. (1994) and for that purpose developed a new methodology (a forerunner of that in the previous section) to run in conjunction with a predictive exposure distribution such as those in Chapters 9 and 10.

### Clusters

The follow-up studies classified each hospital admission as being in one of the 733 Census Subdivisions (CSDs) depending on the patient's place of residence (cluster;) urban clusters having populations of about 100,000). They used only daily cluster admission totals for each of the six years in the study. [In fact only those for May to August were used since Burnett and Krewski (1994) found the strongest association between pollution and admissions during those months.] They assessed the average effect of changes in the mean pollution level over all CSDs using a population-average approach (Section 13.4). [In contrast, that of Le et al. (1999) was cluster-specific.]

### Impact Model

The health impact model needed to adjust for both seasonal variation as well as day-of-the-week effects, both of which could affect the daily hospital admissions counts. Given those factors plus the goal of estimating population-average effects, the expected admission count $Y_{it}$ given the pollution levels $\mathbf{X}_{it}$ was modeled as

$$\begin{aligned}\mathbf{E}\left(Y_{it} \mid \mathbf{b}_i, \mathbf{X}_{it}\right) &= \zeta(\beta^T \mathbf{X}_{it}) \\ &\equiv m_{it} exp\left(\beta^T \mathbf{X}_{it}\right).\end{aligned}$$

The multiplier ($m_{it}$) accounts for seasonality (trend), the day-of-the-week effect, and CSD population size. The first two were estimated prior to fitting the model (Burnett and Krewski 1994). The 1986 Census provided population counts that, together with the large quantity of data used in model fitting, meant the $\{m_{it}\}$ could be treated as known. As this was a population-level analysis the conditional covariance of $Y_{it}$ defined in Section 13.4 was

$$\mathbf{Cov}\left(Y_{it_1}, Y_{it_2} \mid \mathbf{X}_{it_1}, \mathbf{X}_{it_2}\right) = \delta_{it_1 t_2} \phi \zeta(\beta^T \mathbf{X}_{it}).$$

### Air Pollutants

Burnett et al. (1994) study four pollutants $SO_4$($\mu$g m$^{-3}$), $O_3$ (ppb), $SO_2$ ($\mu$g m$^{-3}$), and $NO_2$ ($\mu$g m$^{-3}$). However Zidek et al. (1998b) consider only $SO_4$, and $O_3$. These were of primary concern thanks to the results of Burnett et al. (1994). More importantly, Duddek et al. (1995) showed $SO_4$ and $NO_2$ to be much less strongly associated with the admission counts. These analyses also included maximum daily temperature and average daily humidity as climatic variables. For computational simplicity pollution variables as well as climate variables were taken to be uncorrelated in constructing the working covariance matrix.

### Model Uncertainty Versus Standard Errors

The follow-ups were carried out one summer at a time. That simplified calculation and more importantly enabled an assessment of temporal changes in impact as well as model uncertainty. Standard errors (SEs) do not measure the latter, reflecting only parameter estimate uncertainty. Worse still they are often asymptotic, leaving uncertainty about their legitimacy in finite samples. Model deficiencies might well be seen in the warp of year to year parameter estimates subject to model structural limitations. In other words they may well make the parameter estimators wobble over cases and times as they attempt to compensate for the model's inadequacies.

In contrast small SEs might lead an analyst relying on just a single aggregate analysis to unwarranted complacency. Separating the analysis into six summers, Duddek et al. (1995) and Zidek et al. (1998b) seek to discover that wobble and therefore some indication of model reliability, and true parameter uncertainty.

Of course their approach costs significance, the separate analyses being based on just 1/6 of the data. However, all is not lost. If they demonstrate model reliability the separate estimates can be averaged to get an even better overall estimate. Moreover, since separate data sets are approximately independent, its standard error can readily be computed as the square root of the average of the six squared SEs. While the result will not be quite as efficient as that which would have been obtained from an aggregate analysis, the investigators have bought some insurance against the possibility of a misspecified model.

In fact, initially Duddek et al. (1995) did not incorporate the uncertainty in the pollutant levels, setting $\mathbf{G}_{it}$ equal to zero. Their results are shown in Table 13.5, giving fitted ozone model coefficients $\beta_1^*$ with ozone-lagged 0, 1, and 2 days. (Earlier studies had shown the irrelevance of longer lags.) Table 13.5 gives the corresponding results for nitrogen dioxide.

### Results for Population Averages

Table 13.5 suggests log $O_3$–lag 1 day and log $O_3$–lag 2 days compete for association with daily admission counts. For all years both are significant save 1987 when only log $O_3$–lag 2 is. Duddek et al. (1995) therefore chose log$O_3$–lag 2 days as their covariate. Overall their results seem stable over time and they pooled their six estimates (and SEs) to obtain a strongly significant combined $\beta_1^*$ estimate of 0.053 (SE = 0.0082).

Duddek et al. (1995) find log$NO_2$–lag 2 to be a substantially better predictor overall so choose lag 2 in this case as well. However that predictor proved insignificant in 1983, 1984, and 1987 [even though the combined estimate in this case 0.039 (SE = 0.0081) was quite significant]. These results lead Duddek et al. (1995) to incorporate uncertainty in the interpolated values of log $O_3$ to assess the robustness of earlier results against ME. By their two-stage

analysis the authors aimed to see if the added uncertainty due to error would significantly affect the findings. In fact the new analysis for the two pollutants lag-2 days shows no change to two significant digits (see Table 13.5). The explanation lies in the relative lack of uncertainty relative to all other sources and the low level of pollutant–admissions association (i.e., the fitted model coefficients). In other words the spatial field is interpolated with sufficient precision that only bias due to the ME affects the analysis.

Table 13.1: Estimated log ozone transfer coefficients regarding interpolation error as negligible.

| $\beta_1^* \times 1000$ (robust SE$\times$ 1000 in parentheses) | | | |
|---|---|---|---|
| | Lag | | |
| Summer | 0 | 1 | 2 |
| 1983 | -13 (16) | 40 (14) | 33 (17) |
| 1984 | 14 (23) | 75 (26) | 59 (22) |
| 1985 | 29 (20) | 47 (21) | 57 (24) |
| 1986 | 25 (19) | 79 (18) | 64 (20) |
| 1987 | 2 (20) | 21 (19) | 33 (19) |
| 1988 | 31 (16) | 59 (15) | 73 (17) |

Table 13.2: Estimated nitrogen dioxide transfer coefficients regarding interpolation error as negligible.

| $\beta_1^* \times 1000$ (robust SE$\times$ in parentheses) | | | |
|---|---|---|---|
| | Lag | | |
| Summer | 0 | 1 | 2 |
| 1983 | -11 (14) | 37 (14) | 22 (18) |
| 1984 | 54 (0.021) | 39 (0.018) | 31 (19) |
| 1985 | 2 (22) | 41 (20) | 70 (21) |
| 1986 | -9 (20) | 0.004 (18) | 61 (18) |
| 1987 | 40 (21) | -32 (19) | 2 (20) |
| 1988 | 7 (20) | 48 (19) | 47 (22) |

### Cluster-Specific Results

For simplicity Zidek et al. (1998c) restrict their cluster specific analysis to just 1988 and the 100 CSDs with the largest average daily hospital admission counts. Moreover, they consider only the three variables found to be important

Table 13.3: Estimated lag-2 pollutant transfer coefficients incorporating interpolation error.

| $\beta_1^* \times 1000$ (robust SE$\times$ in parentheses) | | |
|---|---|---|
| Summer | $O_3$ | $NO_2$ |
| 1983 | 33 (17) | 22 (18) |
| 1984 | 59 (22) | 31 (19) |
| 1985 | 57 (24) | 70 (21) |
| 1986 | 0.064 (20) | 61 (18) |
| 1987 | 33 (19) | 2 (20) |
| 1988 | 0.073 (17) | 47 (22) |

by Zidek et al. (1998b), log $SO_4$–lag 1 and log $O_3$–lags 2,3. In fact they concentrate mainly on the first of these, the single largest fraction of airborne particulate pollution.

They start with just the top 10 CSDs ranked in descending order by their average 1988 daily hospital admission numbers. Their results appear in Table 13.5, the estimated random effects (i.e., $\{\hat{b}_k\}$s) for the three pollution variables as well their typical effects (the $\hat{\beta}$s). Their relative risks obtain from adding the random effects to the typical effects to get $\hat{\beta} + \hat{b}_k$.

These random effects vary a lot over these ten subdivisions indicating substantial variation in impact among them. Thus their inclusion seems important. Some CSDs seem to be well above the typical effect level for one of the two log–ozone variables, yet below that for the other. The authors offer no explanation for this curious inconsistency.

By extending their study to include the rest of the 100 CSDs they find the typical effects of the logged and lagged pollution impact coefficients drop to 103, 37, and 169 from 123, 39, and 209, respectively, the values in Table 13.5 Nevertheless, their $z$-scores increase to 2.88, 2.24, and 4.25 from 1.95, 1.43, and 3.31. That is a direct consequence of having the additional information in the larger data set and the consequent reduction in the standard errors of 35, 17, and 39 compared with 63, 39, and 63 originally. This change demonstrates a benefit of the hierarchical Bayes approach.

For a more detailed analysis Le et al. (1999) focus on just log $SO_4$–lag 1. Their results including typical effects for each group of ten CSDs appear in Table 13.5. The impact typically drops, going from the top to bottom 10, not surprising when you consider that the latter has a smaller number of daily hospital admissions (42 versus 60). More surprising is the greater variability in the impact coefficients, something Le et al. (1999) are not able to explain.

### Value of Including Random Effects

Le et al. (1999) explore the benefit of including random effects when dealing with a larger number of clusters, namely, the top 100 CSDs involved in their case study. More precisely, they compare the fits with and without including the random effects. It turns out that including them yields a typical effect of 55 (with SE = 17) against 51 (SE = 17) without. Echoing this result they find the respective quasi-likelihoods to be 7060 and 7045. In this respect, their results agree with those of Burnett and Krewski (1994), although the latter investigate ozone rather than sulfate. They had concluded that fitting random effects was of negligible benefit in terms of model fit at least.

However, Le et al. (1999) argue for including those effects nonetheless. For one thing, they may point to potential hot-spots. These can be due to unknown environmental hazards or known ones that have simply not been detected. Just such spots were seen around the Rocky Mountain Arsenal described in Example 1.2. Of course apparent hot-spots can be artifacts of chance variation so confirmation is vital, a topic beyond the scope of this book.

The estimated random effects prove to be far from normally distributed as their prior distribution assumes. Instead they appear to be bimodal pointing to CSDs with extremely small sulfate impacts, a finding of some interest in its own right. Another inconsistency lies in their negative average value, disagreeing with the zero mean assumption built into the quasi-likelihood.

As another curious feature of the analysis, effects for big Census Subdivisions have big random effects. Yet these values decline sharply to values systematically below zero, the values for the bottom ten Census Subdivisions.

## 13.6 Wrapup

In this chapter, we presented a practical approach to environmental risk assessment, one that fits well into the hierarchical Bayesian framework that provides the technical framework for much of the material in this book. Moreover, we illustrated its use with a case study.

However, the extensive field of risk assessment now offers many alternatives to the approach described above and illustrated again in Chapter 14. One popular approach uses Poisson regression methods for impact data expressed as counts in place of the one we use that devolves from a normal quasi-likelihood function. A good discussion of this and other topics within the context of longitudinal data analysis can be found in Diggle et al. (1994).

We turn in the next chapter to a tutorial that illustrates software that implements many of the methods described in this book.

| Census Sub-division | Random Effects $\times 10^3$ | | |
|---|---|---|---|
| | Ozone Lag 2 | Sulfate Lag 1 | Ozone Lag 3 |
| 1 | 18 | 37 | -50 |
| 2 | 16 | 16 | 5 |
| 3 | -34 | -16 | 36 |
| 4 | 29 | -4 | 18 |
| 5 | 57 | 77 | -103 |
| 6 | -19 | -13 | 3 |
| 7 | -19 | -15 | 6 |
| 8 | 23 | 30 | -27 |
| 9 | -12 | -38 | 25 |
| 10 | 32 | 60 | -75 |
| Typical Effects (SE) | 123 (63) | 39 (27) | 209 (63) |

Table 13.4: Random effects[a] $\times$ 1000 for selected log transformed pollutant concentrations for the top ten census-subdivisions in southern Ontario, 1988, ranked by average daily hospital admission totals.
[a] Units admissions/CSD/day/unit of explanatory variables (log $O_3$ ($\mu$g/m$^3$) and log $SO_4$ (ppb)).

| | Random Effects $\times 10^3$ | | | | | | | | | | Typical Effect $\times$1000 (SE) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Top 10 | 20 | 12 | -13 | 1 | 50 | -12 | -11 | 22 | 24 | 41 | 60 (24) |
| Bottom 10 | 73 | 91 | -52 | -8 | -21 | 40 | 33 | 99 | 109 | -36 | 42 (15) |

Table 13.5: Random effects[a] $\times$ 1000 of log $SO_4$–lag 1 concentrations for the top and bottom ten southern Ontario Census Subdivisions (among the top 100), 1988, ranked by average daily admissions.
[a] Units admissions/CSD/day/unit of log $SO_4$ ($\mu$g/m$^3$).

# 14

# A Tutorial in R

In this chapter, an example is used to illustrate the spatial interpolation approach presented in Chapters 9 and 10 using R on a Windows platform. The example also demonstrates the environmental network extension in Chapter 11. The complete software package is available for downloading free of charge from http://enviRo.stat.ubc.ca. The package contains relevant R functions and instructions for implementation, as well as illustrating examples including the one presented below.

The data set (data) used in this example consists of hourly $O_3$ concentration levels (ppb) from nine stations (S1–S9) in New York State. Other information includes month(mm from 4 to 9), day within month (dd from 1 to 31), hour within day (hr from 0 to 23), weekday (wkday from 2–8), sequential number of week (wk from 1 to 27). Each row of the data set represents an hourly record starting at April 1, 1995, hour 0, and ending at September 30, 1995 hour 23; i.e., there are 4392 records (24 hours $\times$ 183 days). The last six stations have no missing observations; stations 1, 2, 3 have 2616, 2016, 72 missing hourly observations, respectively, at the starting time.

```
> data
 mm dd hr wkday wk S1 S2 S3 S4 S5 S6 S7 S8 S9
  4  1  0     7  1 NA NA NA 22 34 38 30 33 31
  4  1  1     7  1 NA NA NA 19 33 37 29 32 35
  4  1  2     7  1 NA NA NA  9 34 36 21 27 34
  4  1  3     7  1 NA NA NA  8 34 32 15 27 34
  4  1  4     7  1 NA NA NA 10 34 26 21 30 33
----- records deleted ------
  9 30 19     7 27 34 38 24 11 33 27 13 41 24
  9 30 20     7 27 32 31 15 17 27 23 13 38 28
  9 30 21     7 27 28 29 14 16 27 16  8 35 21
  9 30 22     7 27 27 28 11 20 34 11  8 32 16
  9 30 23     7 27 29 25  9 28 37  7  3 29 14
> missing.num = apply(is.na(data[,6:14]),2,sum)
> missing.num
```

```
  S1   S2   S3   S4   S5   S6   S7   S8   S9
2616 2016   72    0    0    0    0    0    0
```

Locations of the stations are given in Figure 14.1 with lat–long coordinates specified in location.

```
> round(location,2)
        lat   long
 [1,] 42.64 -73.32
 [2,] 42.14 -74.66
 [3,] 42.72 -73.58
 [4,] 43.30 -75.87
 [5,] 42.73 -75.94
 [6,] 43.46 -74.67
 [7,] 42.68 -73.91
 [8,] 43.01 -73.80
 [9,] 42.90 -73.40
```

## 14.1 Exploratory Analysis of the Data

First load the relevant R functions with the corresponding dll files. From the download mentioned above, all functions are stored within subdirectories under one large directory called LZ-Rcodes . Assume that the directory is copied to the C drive. The current version of the R function is denoted ver0.1 . This should be changed accordingly with newer versions in future updates.

```
> dyn.load("C:/LZ-Rcodes/SG-method/SG.dll")
> dyn.load("C:/LZ-Rcodes/LZ-design/LZ.design.dll")
> source("C:/LZ-Rcodes/SG-method/SG.ver0.1.r")
> source("C:/LZ-Rcodes/LZ-EM.staircase/LZ-EM.staircase.ver0.1.r")
> source("C:/LZ-Rcodes/LZ-design/LZ.design.ver0.1.r")
> source("C:/LZ-Rcodes/LZ-pred.dist/predict.ver0.1.r")
```

The square root of $O_3$ levels for each station are plotted in Figure 14.2 indicating stations starting operation at different times (i.e., staircase pattern of missing data).

The deterministic trend of the 03 levels is examined by fitting a linear model with hour, weekday, and week as factors for each station separately. The corresponding design matrix is obtained using the model.matrix() function and the linear fit is obtained using the lm() function.

```
> hr = as.factor(data[,3])
> wkday = as.factor(data[,4])
> week = as.factor(data[,5])
> y = sqrt(data[,6:14])
> x = model.matrix(~ hr + wkday + week,
              contrasts=list(hr= "contr.helmert",
```


Fig. 14.1: Locations of monitoring stations.

```
              wkday = "contr.helmert", week ="contr.helmert"))
> fit = list()
> for (i in 1:9)
     fit[[i]] = lm(y[,i] ~ x -1, singular.ok =T , na.action=na.omit)
```

The estimated effects are plotted in Figure 14.3. The results show consistent patterns of hourly and weekday effects for all stations. Except for the first few weeks in April the weekly effects show little temporal trend.

```
> par(mfrow=c(2,2))
> plot(fit[[1]]$coef[2:24],ylim=c(-.5,.5),type="n",xlab="Hour",
                                ylab="Hourly Effects")
> for (i in 1:9) points(fit[[i]]$coef[2:24])
> plot(fit[[1]]$coef[25:30],ylim=c(-.5,.5),type="n",xlab="Weekday",
                                ylab="Weekday Effects")
> for (i in 1:9) points(fit[[i]]$coef[25:30])
```

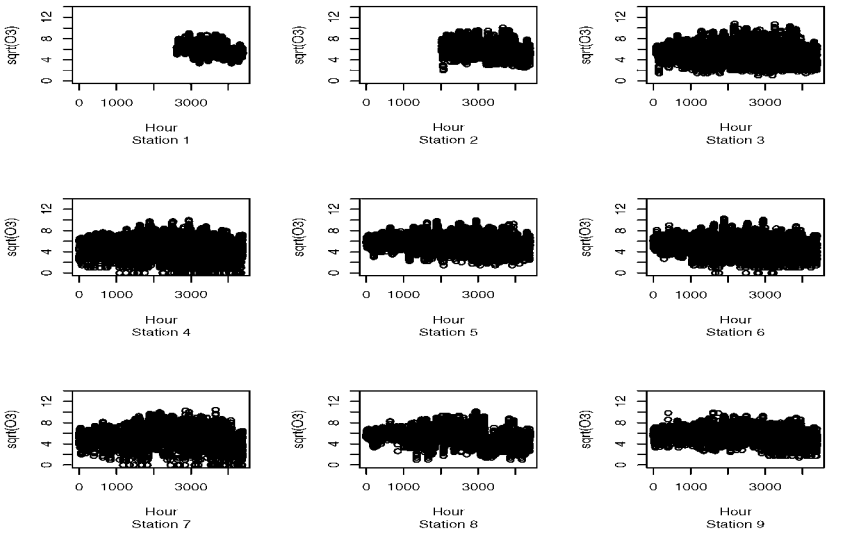


Fig. 14.2: Observed data at the monitoring stations.

```
> plot(fit[[1]]$coef[31:56],ylim=c(-.5,.5),type="n",xlab="Week",
```

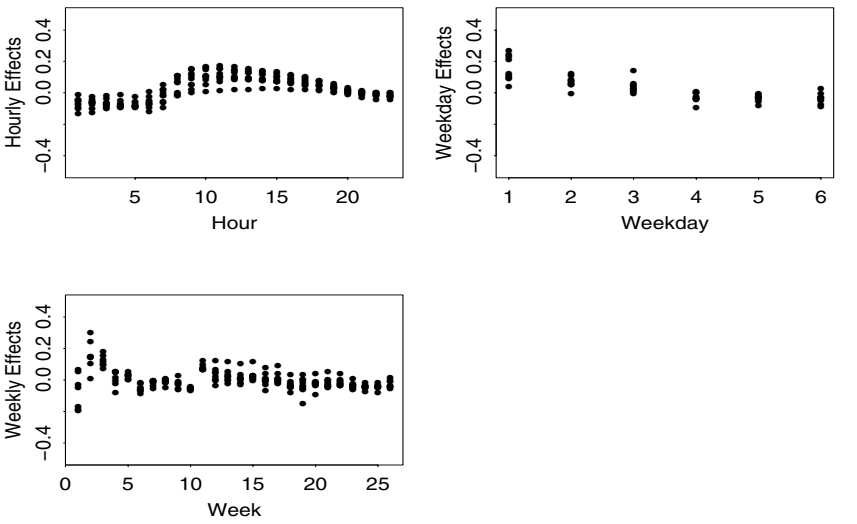


Fig. 14.3: Estimated effects.

```
                                        ylab="Weekly Effects")
> for (i in 1:9) points(fit[[i]]$coef[31:56])
```

The QQ-plots are obtained for the fitted residuals using the qqnorm() command and displayed in Figure 14.4 which indicates that normality assumption is reasonable.

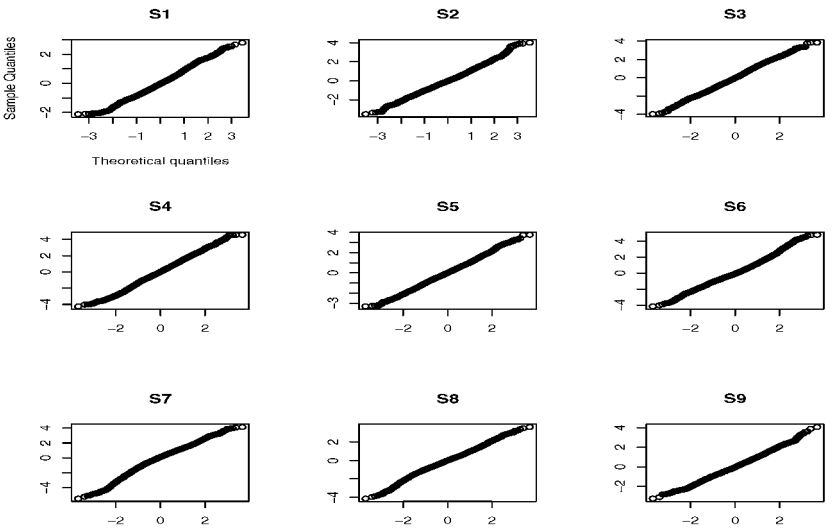


Fig. 14.4: QQ-plots for the fitted residuals.

```
> par(mfrow=c(3,3))
> for (i in 1:9) qqnorm(fit[[i]]$resid)
```

The temporal autocorrelations are plotted (Figure 14.5) using the acf() function. The results seem to indicate an AR(2) autocorrelation structure for each of the stations with a very strong lag-1 correlation of $\geq .9$ consistently.

```
> par(mfrow=c(3,3))
> for (i in 1:9) acf(fit[[i]]$resid,type="partial")
```

The spatial correlations between the stations are obtained

```
> lmfit.resid = NULL
> for (i in 1:9) lmfit.resid = cbind(lmfit.resid,c(rep("NA",
                  missing.num[i]),fit[[i]]$resid))
> lmfit.resid.corr = cor(lmfit.resid,na.method="available")
```

Similarly the spatial correlations after taking out the AR(2) structure are computed.

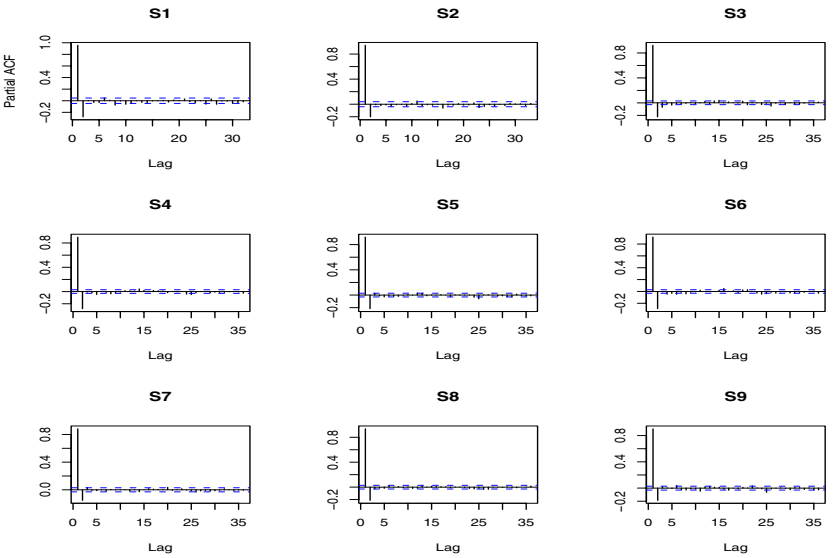


Fig. 14.5: Partial autocorrelation functions.

```
> arfit = list()
> for (i in 1:9) arfit[[i]] = ar(fit[[i]]$resid,aic=F,order=2)
> ar.resid = NULL
> for (i in 1:9) ar.resid = cbind(ar.resid,
           c(rep("NA",missing.num[i]),arfit[[i]]$resid) )
> ar.resid.corr = cor(ar.resid,use = "pairwise")
```

The estimated spatial correlations before and after taking the AR(2) structure are plotted side by side versus the interdistances between the stations in Figure 14.6. First the lat–long coordinates are transformed to a rectangular coordinate system through a Lambert projection using the Flamb2() function. The output of this function includes the new coordinates of the stations as well as the reference coordinates.

```
> coords = Flamb2(location)
> coords
$xy:
               x         y
[1,]  106.949788 -16.94391
[2,]   -3.076874 -73.19558
[3,]   21.416492 -55.14407
[4,] -100.584494  57.17452
[5,] -106.787916  -6.51416
[6,]   -3.036232  73.19694
[7,]   58.952310 -13.05582
```

```
[8,]   67.267631  24.26941
[9,]   99.811622  11.92334

$latrf1:
[1] 42.40136

$latrf2:
[1] 43.19204

$latref:
[1] 42.7967

$lngref
[1] 74.62735
```

The intersite distances between the locations are calculated using the Fdist() function.

```
> dist = Fdist(coords$xy)
> par(mfrow=c(1,2))
> plot(-.2,0,xlim=c(0,300),ylim=c(-.2,1),xlab="Dist",
          ylab="Spatial correlation (detrended sqrtO3)",type="n")
> for (i in 1:8) for (j in (i+1):9)
          points(dist[i,j],lmfit.resid.corr[i,j])
> plot(-.2,0,xlim=c(0,300),ylim=c(-.2,1),xlab="Dist",
          ylab="Spatial correlation (AR(2) resid)",type="n")
> for (i in 1:8) for (j in (i+1):9)
          points(dist[i,j],ar.resid.corr[i,j])
```

The results show a substantially reduced spatial correlation when an AR(2) process is taken out. The pre-AR(2) spatial correlations are mostly between 0.4 to 0.6 but the post-AR(2) ones are reduced to around 0.1. This phenomenon has been observed and studied by Zidek et al. (2002) who term it a *spatial correlation leakage* problem. The authors show that in an AR process with spatially correlated residuals, removing the AR structure by first fitting the corresponding coefficients and then subtracting them from the original process could reduce the spatial correlation substantially. The reduction depends on the strength of the autocorrelation which is quite strong in this example.

This spatial correlation leakage presents a special challenge for the spatial interpolation problem. The usual approach where the temporal component is first taken out then the residuals are interpolated would not work in this case. Generally speaking there are no observations at the new locations to take advantage of the strong temporal correlation yet the residuals from monitoring stations are not helpful due to the reduced spatial correlation. Thus an approach where both temporal and spatial components are modeled simultaneously is needed. The Bayesian hierarchical approach presented in Chapters 9 and 10 is an option for this problem as illustrated in the next section.

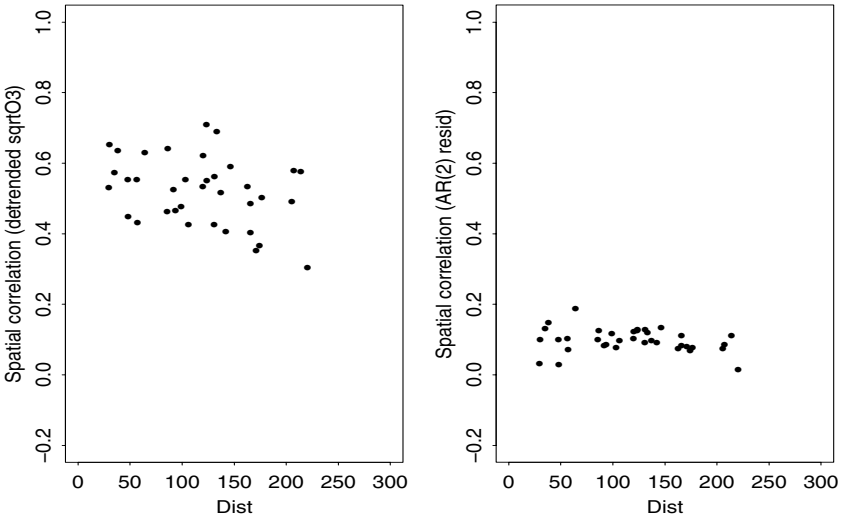


Fig. 14.6: Estimated spatial correlations before and after taking out the AR(2) structure.

## 14.2 Spatial Predictive Distribution and Parameter Estimation

Suppose one is interested in interpolating the hourly $O_3$ levels at unobserved locations for a specific hour of the day. The method described in Chapter 10 can be used where one could consider a multivariate response consisting of the hourly $O_3$ levels at that hour as well as several preceding hours. The approach would then allow for the use of consecutive hourly levels at monitoring stations in the interpolation and thus take advantage of the strong temporal pattern.

In this illustrative example a 4-consecutive hour multivariate response for each day is considered where the last element is the hour of interest (11AM–noon) and the preceding 3 hours (8–10AM) are used to capture the temporal pattern. The choice of 4 hour sis based on the available data (183 days) for the estimation of the hyperparameters and the independence requirement of the model for the multivariate response. Here with an AR(2) structure observed, the 18-hour gap from one daily response to the next day would reduce the temporal correlation substantially.

The daily four-hour response is then assumed to follow a model specified by Equations (10.1–10.2). The predictive distribution for the unobserved locations and times, conditional on the observed data and the hyperparameters, is given by Equations (10.3–10.6). That is, the predictive distribution is completely characterized given the hyperparameters. Specifically the hyper-

parameters include those associated with the gauged sites

$$\mathcal{H}_g = \{F, \boldsymbol{\beta}_0, \Omega, (\tau_{01}, H_1, \Lambda_1, \delta_1), \ldots, (\tau_{0,k-1}, H_{k-1}, \Lambda_{k-1}, \delta_{k-1}), (\Lambda_k, \delta_k)\}, \quad (14.1)$$

and those associated with ungauged sites $\Lambda^{[u]}$, $\tau_0^{[u]}$, $H^{[u]}$, and $\delta^{[u]}$. As described in Sections 10.6.3 and 10.6.4, the hyperparameters associated with the gauged sites can be estimated through an EM algorithm approach and those associated with the new locations (i.e., ungauged sites) can be estimated via the Sampson–Guttorp method. The R session on how the multivariate response is constructed and the hyperparameters are estimated is presented next.

### 14.2.1 Parameter Estimation: Gauged Sites Through the EM-algorithm

First the data from each station are organized into a 24-hour (0–23) matrix (183 days × 24) and then these matrices are combined side by side into a larger matrix "series24hr" (183 days × 216).

```
> series24hr = NULL
> for (i in 6:14) { x = sqrt(data[,i])
    temp = t(matrix(x,nrow=24))
    series24hr = cbind(series24hr,temp) }
```

The multivariate response consisting of 4 consecutive hours from 8AM to 12 noon is extracted and denoted by "hr8.11":

```
> n = 4
> tt = c(1:n)
> for (i in 2:9) tt = c(tt,c(1:n)+24*(i-1))
> hr8.11 = series24hr[,tt+2*n]
```

The month and weekday factors corresponding to the rows are obtained for trend fitting.

```
> month  = as.factor((matrix(data[,1],byrow=T,ncol=24))[,1] )
> weekday  = as.factor((matrix(data[,4],byrow=T,ncol=24))[,1] )
```

The $\mathcal{H}_g$ hyperparameters are estimated by the EM algorithm, described in Section 10.6.3 Chapter 10, using the staircase.EM() function below. In this R function call, the covariates "month" and "weekday" are used as categorical factors. The current version assumes an exchangeable structure between stations for $\boldsymbol{\beta}_0$ but allows for different coefficients from each element of the multivariate response (i.e., hours in this case). The default block structure is based on the staircase of the missing data; i.e., the stations having the same number of observations are grouped together as a block. In this example the default option is used and thus there are four blocks associated with the observed data. The first three blocks have 1 station each from S1 to S3 and the last block have six stations from S4 to S9. Note that data must be ordered in decreasing number of missing observations for this function.

```
> Z  =model.matrix(~month+weekday, contrasts =
       list(month= "contr.helmert",weekday = "contr.helmert"))
> emfit.hr8.11 =  staircase.EM(hr8.11,p=4,covariate=Z)
```

The estimated $\Omega$ (*Omega*), $\Lambda_j$ (*Lambda*), $\delta_j$ (*Delta*), and $\boldsymbol{\beta}_0$ (*Beta0*) are the output of this staircase.EM() function:

```
> round(emfit.hr8.11$Omega,2)
     [,1] [,2] [,3] [,4]
[1,] 0.60 0.41 0.28 0.21
[2,] 0.41 0.44 0.34 0.26
[3,] 0.28 0.34 0.37 0.31
[4,] 0.21 0.26 0.31 0.37

> emfit.hr8.11$Lambda
[[1]]
         [,1]
[1,] 5.215931

[[2]]
         [,1]
[1,] 14.89936

[[3]]
         [,1]
[1,] 31.43736

[[4]]
          [,1]      [,2]     [,3]       [,4]      [,5]      [,6]
[1,] 80.998313 19.524571 18.14417   4.874237  9.303233  4.330475
[2,] 19.524571 45.605702 16.97445   8.322499  6.512192  8.185238
[3,] 18.144173 16.974450 68.79851  10.461332 12.383675 12.341112
[4,]  4.874237  8.322499 10.46133 152.889043 22.222718 15.451960
[5,]  9.303233  6.512192 12.38368  22.222718 53.193556 13.383386
[6,]  4.330475  8.185238 12.34111  15.451960 13.383386 44.501323

> emfit.hr8.11$Delta
[[1]]
[1] 23.59728

[[2]]
[1] 22.49019

[[3]]
[1] 30.23065

[[4]]
[1] 76.23103
```

```
> round(emfit.hr8.11$Beta0[,1:4],2)
       [,1]  [,2]  [,3]  [,4]
 [1,]  5.08  5.19  5.35  5.46
 [2,] -0.08 -0.12 -0.13 -0.11
 [3,] -0.07 -0.05 -0.02 -0.01
 [4,] -0.03 -0.04 -0.01  0.00
 [5,] -0.20 -0.18 -0.17 -0.17
 [6,] -0.21 -0.19 -0.17 -0.16
 [7,]  0.14  0.14  0.17  0.17
 [8,]  0.10  0.08  0.07  0.04
 [9,]  0.08  0.06  0.05  0.04
[10,]  0.01  0.00  0.00  0.00
[11,] -0.03 -0.02 -0.03 -0.03
[12,] -0.02 -0.02 -0.02 -0.03
```

Here only estimated $\beta_0$s corresponding to four different hours at the first station are displayed. Other stations have the same estimates since an exchangeable structure is used across sites. Recall that $\Omega$ represents the covariances between hours (up to a scale) and the estimated one seems able to capture the AR structure seen in the exploratory data analysis. $\Lambda$s represent the residual covariances between gauged stations within each block (i.e., conditional on observed data in the preceding blocks). The staircase.EM() function also gives the estimate of the unconditional covariance matrix between all monitoring stations $\Psi$ (*Psi*) as given in (10.26).

```
> round(emfit.hr8.11$Psi[[1]],2)
       [,1] [,2]  [,3] [,4] [,5] [,6] [,7] [,8] [,9]
 [1,]  0.64 0.21 -0.02 0.08 0.05 0.14 0.13 0.08 0.15
 [2,]  0.21 1.78  0.57 0.16 0.28 0.41 0.36 0.17 0.36
 [3,] -0.02 0.57  2.09 0.11 0.18 0.23 0.76 0.22 0.29
 [4,]  0.08 0.16  0.11 1.58 0.38 0.35 0.10 0.18 0.08
 [5,]  0.05 0.28  0.18 0.38 0.89 0.33 0.16 0.13 0.16
 [6,]  0.14 0.41  0.23 0.35 0.33 1.34 0.20 0.24 0.24
 [7,]  0.13 0.36  0.76 0.10 0.16 0.20 2.98 0.43 0.30
 [8,]  0.08 0.17  0.22 0.18 0.13 0.24 0.43 1.04 0.26
 [9,]  0.15 0.36  0.29 0.08 0.16 0.24 0.30 0.26 0.87
```

The estimated (unconditional) spatial correlations between all monitoring stations are obtained and then displayed in Figure 14.7. The results indicate that the spatial correlations are much higher than those displayed in Figure 14.6 and so the effect of the correlation leakage problem has been reduced through this multivariate modeling.

```
> {\rm Cov} = emfit.hr8.11$Psi[[1]]
> em.corr.hr8.11 = {\rm Cov} / sqrt(matrix(diag(cov),9,9)*
                                        t(matrix(diag(cov),9,9)))
> plot(-.2,0,xlim=c(0,300),ylim=c(-.2,1),xlab="Dist",
              ylab="Spatial correlation (Hours 8-11)", type="n")
> for (i in 1:8) for (j in (i+1):9)
```

```
points(dist[i,j],em.corr.hr8.11[i,j])
```

### 14.2.2 Parameter Estimation: The Sampson–Guttorp Method

The estimated unconditional spatial covariance matrix Psi" among the monitoring stations is now nonparametrically extended to the ungauged locations of interest using the Sampson–Guttorp method as described in Section 5.3 of Chapter 10. The estimation procedure provides estimates for $\Lambda^{[u]}$, $\tau_0^{[u]}$, and $H^{[u]}$. The SG-estimation procedure is currently not fully automated and has to be done in several sequential steps, denoted by **Step**, as described below. The SG method does not assume a constant variance is not required for the SG method. Hence the approach starts with the estimation of the correlation and then any estimate of the variance field can be incorporated as seen in Step 5.

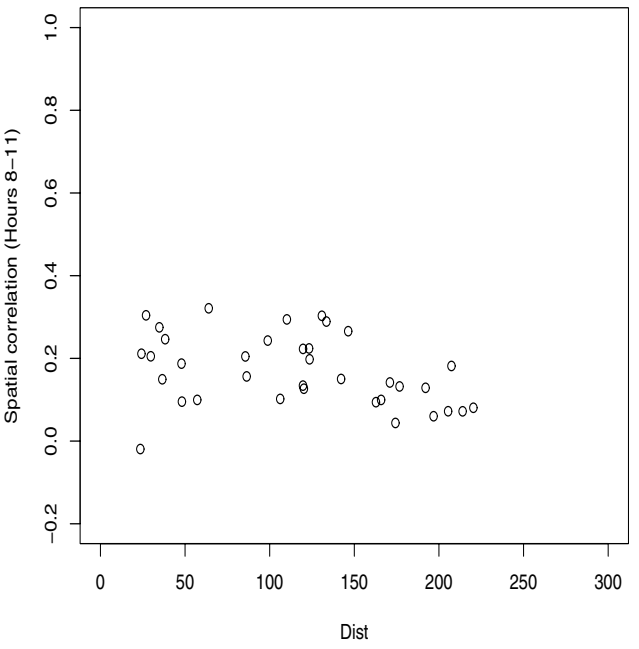


Fig. 14.7: Spatial correlations based on the hierarchical model.

- **Step 1**: The objective of this first step is to identify, with a dispersion matrix and coordinates of the stations, a new configuration of the original coordinates (locations in geographical space) where the estimated correlation would follow an isotropic model. The newly configured locations are in a space called *D-space* or dispersion space. Sampson and Guttorp (1992) use the "dispersion" term, instead of the usual "variogram," to emphasize that the spatial correlation structure in the geographical space may not be isotropic.

  The Falternate3() function is written for that purpose. This function uses an alternating iterative algorithm trying to optimally relocate the stations in $D$-space using the multidimensional scaling method and then fitting the variogram. The exponential variogram is used as a default option for this function. The exponential semi-variogram is defined as

  $$\gamma(h) = a0 + (2 - a0)(1 - \exp(-t0 \times h)),$$

  where $a0$ and $t0$ are parameters to be estimated and $h$ is the distance between locations. The other option is the Gaussian semi-variogram defined as

  $$\gamma(h) = a0 + (2 - a0)(1 - \exp(-t0 \times h^2)).$$

  In this example, the exponential variogram is used. The function seems to work better with small distances and so the coordinates are scaled down by a factor 1/10.

```
> coords.lamb  = coords$xy/10
> disp = 2-2*em.corr.hr8.11
> sg.hr8.11 = Falternate3(disp,coords.lamb,alter.lim=100)

VAR-convergence:  TRUE
  nlm code =   1
    criterion:  0.9104484
MDS-convergence:  FALSE
  nlm code =   3
   criterion:  0.5627452

...... deleted output .....

MDS-convergence:  TRUE
  nlm code =   2
   criterion:  0.1850346
VAR-convergence:  TRUE
  nlm code =   2
    criterion:  0.1850346
There were 50 or more warnings (use
warnings()  to see the first 50)
```

  At each iteration, the results indicating the movements of the locations from the original locations and the fitted variograms are displayed in a

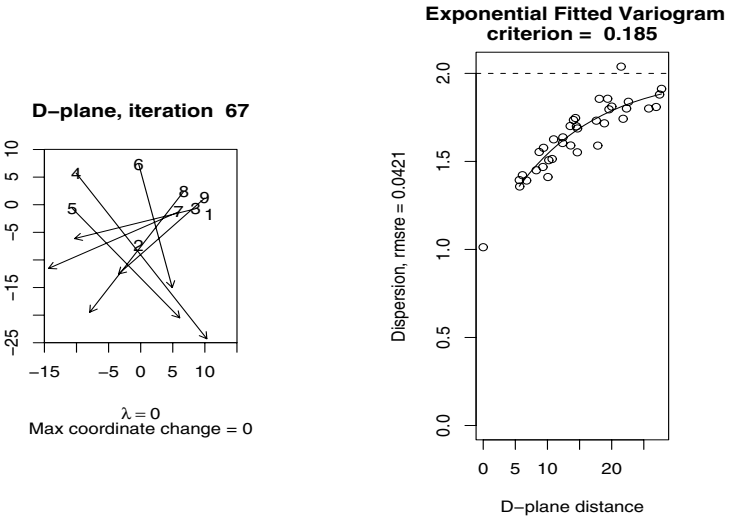


Fig. 14.8: Movements of stations and the dispersion fit at the last iteration.

graphical window. The iterative procedure converges after 67 iterations. Here no smoothing is imposed on the fit. Figure 14.8 shows the results for the last iteration. Note that the warnings seen in the output are only related to the setplot() function for displaying the results.
The results show a very good fit for the dispersion using the exponential variogram with the interdistances in D-space. The SG fitted values are given below where $a0 = variogfit\$a[1]$ and $variogfit\$a[2] = 2 - variogfit\$a[1]$.

```
> sg.hr8.11
$variogfit
$variogfit$objf
[1] 0.1850346

$variogfit$t0
[1] 0.07678341

$variogfit$a
[1] 1.0124313 0.9875687

$variogfit$fit
 [1] 1.617620 1.810046 1.544093 1.825860 1.787610 1.880034
 [7] 1.776782 1.679263 1.814506 1.361316 1.676004 1.516186
[13]1.744501 1.566878 1.358375 1.875078 1.680045 1.412109
[19]1.882803 1.821930 1.780427 1.863412 1.672172 1.650873
```

```
[25]1.767989 1.664286 1.652708 1.548914 1.748855 1.382073
[31]1.519645 1.753477 1.618884 1.494055 1.574384 1.476157


$ncoords
             [,1]       [,2]
 [1,]  10.6949788  -1.694391
 [2,]  -0.3076874  -7.319558
 [3,] -10.3063760  -6.151099
 [4,]  10.3006164 -24.292021
 [5,]   6.0745665 -20.502635
 [6,]   4.9124224 -15.007853
 [7,] -14.3623410 -11.553414
 [8,]  -7.9408930 -19.485040
 [9,]  -3.4463130 -12.557688
```

- **Step 2**: The SG method next fits a thin-plate smoothing spline between the original locations and the D-space locations identified in Step 1. This step allows the user to view the deformation of the geographical space in the D-space and to select a suitable value for the smoothing parameter of the thin-plate spline. This can be achieved by the Ftransdraw() function as demonstrated below. The function is an interactive one showing the fitted variogram and the mapping transformation from Step 1 from the geographical space into D-space.
  First a grid of points over the range of stations is created with the Fmgrid() function. The Ftransdraw() function fits thin-plate splines between the G-space locations and the D-space location and applies the fitted spline to G-space grid points. It then draws the corresponding grid points in D-space allowing the users to interactively choose a suitable value for the smoothing parameter ("lambda").

```
> apply(coords.lamb,2,range)
             x         y
[1,] -10.67879 -7.319558
[2,]  10.69498  7.319694
> coords.grid = Fmgrid(c(-11,11),c(-7.5,7.5),xn=10,yn=10)
> deform  <- Ftransdraw(disp=disp, Gcrds=coords.lamb,MDScrds=
                        sg.hr8.11$ncoords, gridstr=coords.grid)

Click anywhere on plot to continue (Left button)
VAR-convergence: TRUE
  nlm code =   1
    criterion:  0.1850346
Enter value for new lambda (Hit return to stop)
1:
```

The interactive Ftransdraw() function first displays the deformation of the G-space rectangular grid when no smoothing is imposed (i.e., $\lambda = 0$).

Typically the D-space image would have some folding as seen in Figure 14.9. Generally speaking a folded D-space is not desirable since it implies that two locations farther apart could have higher correlation than that corresponding to those located between them. A smoothing parameter that smooths out any folds in the D-space would avoid that problem. The function interactively prompts the user for a new value. Here the value 50 is provided as input to the function.

```
Enter value for new lambda (Hit return to stop)
1: 50
Read 1 item
Click anywhere on plot to continue (left button)
VAR-convergence:
TRUE
  nlm code =   1
    criterion:  0.7648719
Enter value for new lambda (Hit return to stop)
```

The results for choosing "lambda = 50" are displayed in Figure 14.10, showing a reasonable choice for smoothing. Hitting a return key without providing a new value for lambda will terminate the function. Notice that there is a trade-off between the variogram fit and the smoothness of the deformation. Although this selection is somewhat ad hoc, the general goal is to find a small value of $\lambda$ that yields an unfolded transformation.

- **Step 3:** This step combines the results in Step 1 and the smoothing parameter identified in Step 2 to create a thin-plate smoothing spline for mapping coordinates from the geographical space to the D-space. The sinterp() function fits the thin-plate spline with the selected smoothing parameter. The estimated coefficients $\alpha$s and$\beta$s for this example are store in *sol*, one column for each coordinate.

```
> Tspline = sinterp( coords.lamb, sg.hr8.11$ncoords, lAM = 50 )
> Tspline$sol
                [,1]          [,2]
 [1,]  1.249889e-02  3.917129e-03
 [2,]  1.057970e-03  6.469526e-04
 [3,] -4.843946e-03  2.046898e-03
 [4,]  1.880209e-03 -2.047277e-04
 [5,] -6.478255e-05 -8.801582e-04
 [6,]  1.715288e-03  3.824566e-03
 [7,] -8.671632e-03 -1.411987e-03
 [8,] -4.148977e-03 -5.562235e-03
 [9,]  5.769824e-04 -2.376438e-03
[10,] -2.764702e+00 -1.438088e+01
[11,] -4.899633e-01 5.622610e-01
[12,]  2.929522e-01 -8.106002e-01
```

The bgrid() function evaluates the so-called biorthogonal grid depicting the contraction and expansion of the thin-plate spline (see Sampson and Gut-

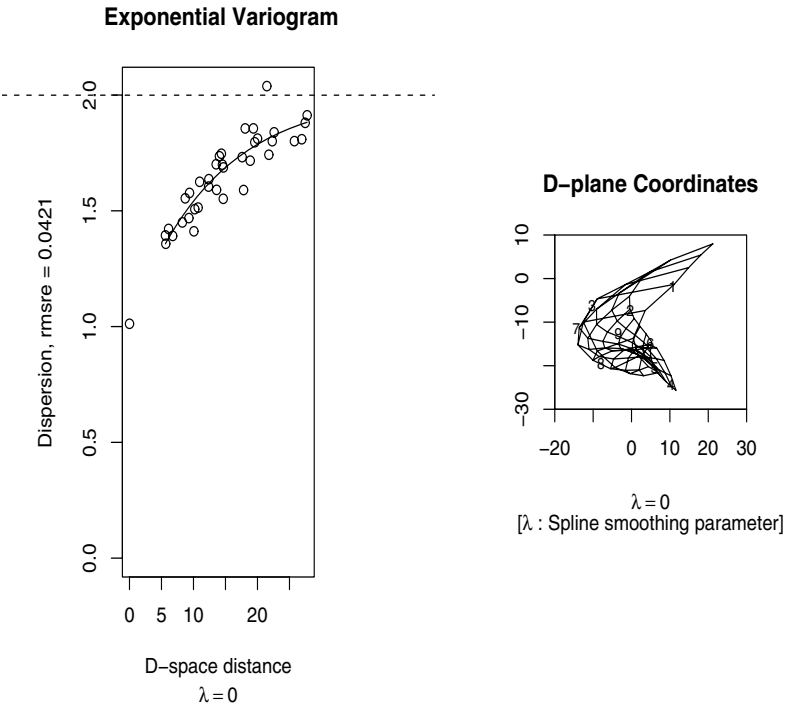


Fig. 14.9: Variogram fit and deformation with no smoothing.

torp 1992 for more details). Solid lines indicate contraction while dashed lines indicate expansion. The results in this example are displayed in Figure 14.11 showing the contraction along the Southeast to Northwest direction.

```
> Tgrid  = bgrid(start=c(0,0), xmat=coords.lamb,
                                          coef=Tspline$sol)
> tempplot = setplot(coords.lamb, ax=T)
> text (coords.lamb)
> draw(Tgrid, fs=T)
```

- **Step 4:** This step uses the thin-plate spline in Step 3 and the corresponding variogram fitted in Step 1 to estimate the dispersions between the stations and the new locations of interest. The results are obtained by first converting the new locations to the Lambert coordinates using the same reference point as before, then evaluating their corresponding locations in the D-space using the selected thin-plate spline, and finally calculating the correlations using the fitted variogram parameters and the interdistances in the D-space. In this example, a grid with $10 \times 10$ points covering sta-

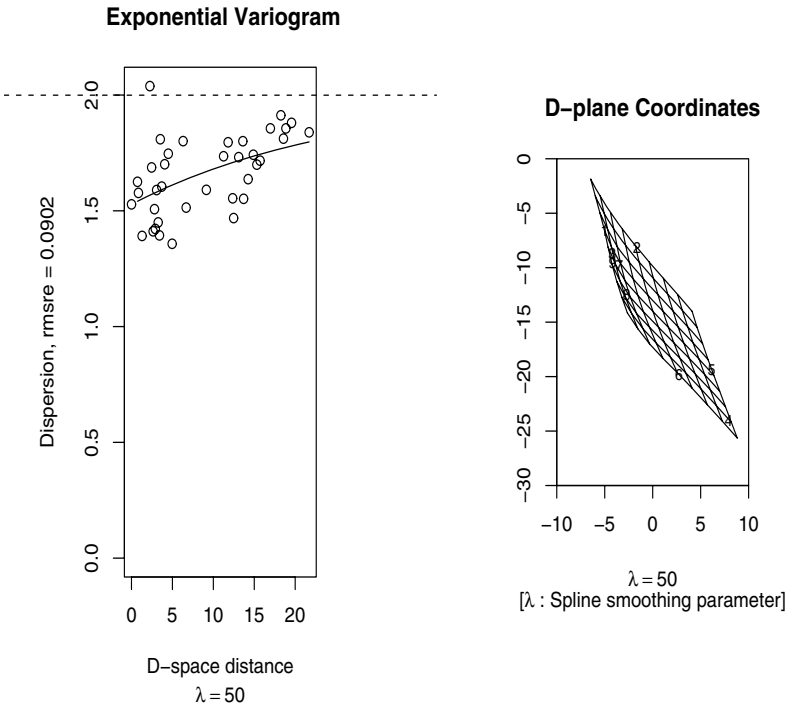


Fig. 14.10: Variogram fit and deformation with lambda = 50.

tions is used, resulting in 100 coordinates. First the coordinates of the new locations in geographical space are obtained.

```
> lat10 <- seq(min(location[,1]),max(location[,1]),length=10)
> long10 <- seq(max(abs(location[,2])),
                          min(abs(location[,2])),length=10)
> llgrid <-  cbind(rep(lat10,10),c(outer(rep(1,10),long10)))
```

The station coordinates are also attached at the end. All locations are converted to Lambert coordinates using the reference points used earlier. Note that the Lambert projected coordinates must be scaled by 1/10 as before to ensure the same unit for distance.

```
> newcrds <- rbind(llgrid,abs(location))
> z <- coords
> newcrds.lamb <- Flamb2(newcrds,latrf1=z$latrf1,
        latrf2= z$latrf2,latref=z$latref,lngref=z$lngref)$xy/10
```

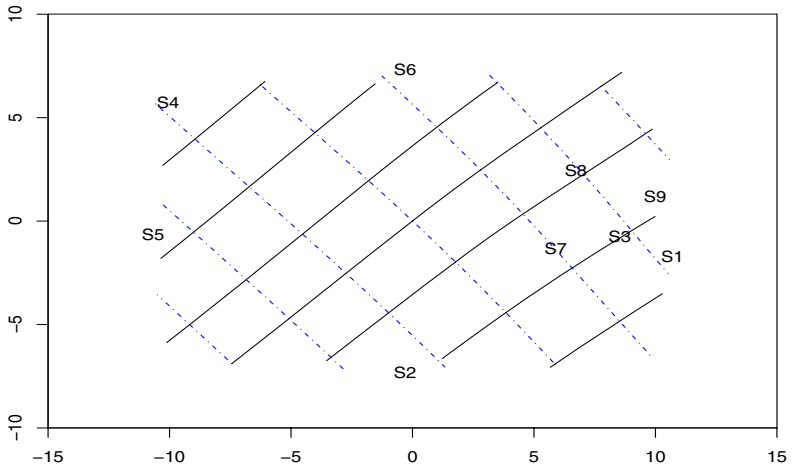


Fig. 14.11: Biorthogonal grid for the thin-plate spline.

Next the estimated correlations between the locations are obtained using the corrfit() function. The estimated correlations for the first ten locations are given below.

```
> corr.fit = corrfit(newcrds.lamb, Tspline, sg.hr8.11, model = 1)
> round(corr.fit$cor[1:10,1:10],2)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] 1.00 0.44 0.40 0.36 0.32 0.29 0.26 0.24 0.21  0.19
 [2,] 0.44 1.00 0.44 0.40 0.36 0.32 0.29 0.26 0.24  0.21
 [3,] 0.40 0.44 1.00 0.44 0.40 0.36 0.32 0.29 0.26  0.24
 [4,] 0.36 0.40 0.44 1.00 0.44 0.40 0.36 0.32 0.29  0.26
 [5,] 0.32 0.36 0.40 0.44 1.00 0.44 0.40 0.36 0.32  0.29
 [6,] 0.29 0.32 0.36 0.40 0.44 1.00 0.44 0.40 0.36  0.32
 [7,] 0.26 0.29 0.32 0.36 0.40 0.44 1.00 0.44 0.40  0.36
 [8,] 0.24 0.26 0.29 0.32 0.36 0.40 0.44 1.00 0.44  0.40
 [9,] 0.21 0.24 0.26 0.29 0.32 0.36 0.40 0.44 1.00  0.44
[10,] 0.19 0.21 0.24 0.26 0.29 0.32 0.36 0.40 0.44  1.00
```

- **Step 5:** This step estimates variances at all locations and then combines with the estimated correlation matrix in Step 4 to get an estimated covariance matrix.

```
> psi = emfit.hr8.11$Psi[[1]]
> round(diag(psi),2)
[1] 0.64 1.78 2.09 1.58 0.89 1.34 2.98 1.04 0.87
```

In this example, the variance field appears to be heterogeneous. One option is to use the same thin-plate spline above to get smoothed estimates of site variances. The variance estimates are then combined with the estimated correlation matrix to get the covariance matrix estimate ("covfit").

```
> Tspline.var = sinterp(coords.lamb,matrix(diag(psi),ncol=1),lam=50)
> varfit = seval(newcrds.lamb,Tspline.var)$y
> temp = matrix(varfit,length(varfit),length(varfit))
> covfit = corr.fit$cor * sqrt(temp * t(temp))
```

The SG-method for extending the spatial covariance matrix from the gauged sites to the ungauged ones is now completed.

### 14.2.3 Parameter Estimation: Ungauged Sites

The SG results are now used to estimate the hyperparameters associated with the ungauged sites $\Lambda_0$ ( "Lambda.0"), $\tau_{00} \equiv \xi.0 \otimes I_p$ ("Xi0.0"), $H_0$ ("H.0"), and $\delta_0$ ("Delta.0"). The staircase.hyper.est() function achieves this objective. It combines the results from the staircase.EM() function (estimating the hyperparameters at gauged sites) and the SG results (extending the spatial covariance matrix to ungauged sites) to obtain estimates for hyperparameters associated with ungauged sites. In this example there are 100 new locations (u = 100), nine monitoring stations each having 4-dimensional response (p = 4).

```
> u = 100 # number of new locations
> p = 4  # dimension of the multivariate response
> hyper.est = staircase.hyper.est(emfit= emfit.hr8.11,
                            covfit=covfit,u =u, p=p)
```

All hyperparameters associated with the predictive distribution as given in Equations (10.3)–(10.6) are now estimated and stored in "hyper.est". Thus, the predictive distribution for the $O_3$ concentration levels at the new 100 locations, from 8–12AM between April 1, 1995, and September 30, 1995, is completely characterized. Besides the estimated hyperparameters, the output from the staircase.hyper.est() function includes all the results from the staircase.EM() fit which can be used for generating realizations from the predictive distribution as demonstrated below.

## 14.3 Spatial Interpolation

With the availability of the predictive distribution, spatial interpolation can be obtained relatively easily. Although the predictive distribution is nonstandard, the mean and the covariance matrix can be analytically derived through conditional reasoning. For other quantiles, the derivation could be tedious and numerical methods may be required.

Alternatively, it is relatively simple to generate realizations from the predictive distribution and spatial interpolation can be easily done using these simulated samples. The pred.dist.simul() function generates realizations for a given timepoint ("tpt"). Here a sample of N = 1000 replicates is generated for all new 100 locations, each with four hours, on September 30, 1995 (i.e., tpt = 183).

```
> simu = pred.dist.simul(hyper.est,tpt = 183, N=1000)
> dim(simu)
[1] 1000  400
```

The sample mean and variance for hourly $O_3$ can be computed from this simulated sample. The contours of the mean and variance surfaces are displayed in Figures 14.12 and 14.13.

```
> x = apply(simu,2,mean)
> X11()
> par(mfrow=c(2,2))
> # Plot the contours
> for (i in 1:4) {
    tt = i+ 4*c(0:99)
    x1 = x[tt]
    hr = matrix(x1 ,byrow=T, ncol=10)
    contour(-long10,lat10, hr, xlab="Long", ylab="Lat",
      main=paste("Mean: Day 183; ", 7+i,"-",8+i,"am"))
 }
> # Plot the corresponding variance field
> x = simu
> X11()
> par(mfrow=c(2,2))
> # Plot the contours
> for (i in 1:4) {
    tt = i+ 4*c(0:99)
    x1 = x[,tt]
    x2 = diag(var(x1))
    vv = matrix(x2 ,byrow=T, ncol=10)
    contour(-long10,lat10, vv, xlab="Long", ylab="Lat",
      main=paste("Var: Day 183; ", 7+i,"-",8+i,"am"))
```

## 14.4 Monitoring Network Extension

The resulting predictive distribution (Section 14.2) can be used to redesign an environmental network as described in Chapter 11. In this illustrative example, suppose that 36 among the 100 locations used in Section 14.2 are considered as potential new sites as displayed in Figure 14.14. The objective is to select an optimal set of 3 locations among these 36 to add to the current network. This can be achieved as follows.

First the coordinates of the potential sites are extracted and plotted.

```
> # Identify potential sites
> # Load mapping library
> library(maps)
> library(mapproj)
> library(mapdata)
>
> par(mfrow=c(1,1))
> map('state', region = c('new york','vermont','mass'))
> text(c(-74.5,-72.8,-71.8),c(44.5,44.5,42.5),
                              c("NY","VT","MA"), cex=.7)
> text(location[,2],location[,1], cex=1)
>
> tt = c(3:8)
> potential.site = NULL
> for (i in 2:7) potential.site = c(potential.site, tt + 10*(i-1))
> potential.coord = llgrid[potential.site,]
> points(-potential.coord[,2],potential.coord[,1])
```

The ldet.eval() function evaluates the log determinants for all combinations of three potential sites as given in the optimality criterion for extension [Equation (11.12)] $\max_{add} \left(\frac{1}{2}\log|\Lambda_0|\right)^{add}$. The function returns the combination with the largest log determinant.

```
> # Extracting the subset of Lambda.0
> hyper.cov = hyper.est$Lambda.0[potential.site,potential.site]
> nsel = 3
> sel = ldet.eval( (hyper.cov+ t(hyper.cov))/2,nsel,all =F)
> text(-potential.coord[sel$coord.sel,2],
                        potential.coord[sel$coord.sel,1], "X")
```

The selected sites are displayed in Figure 14.14. It should be noticed that the ldat.eval() function uses the symmetry of the covariance matrix to reduce computing time and so a symmetric matrix must be provided. This completes the illustrative example of the software. The instructions used in this example are available at the Web site mentioned at the beginning of the chapter.

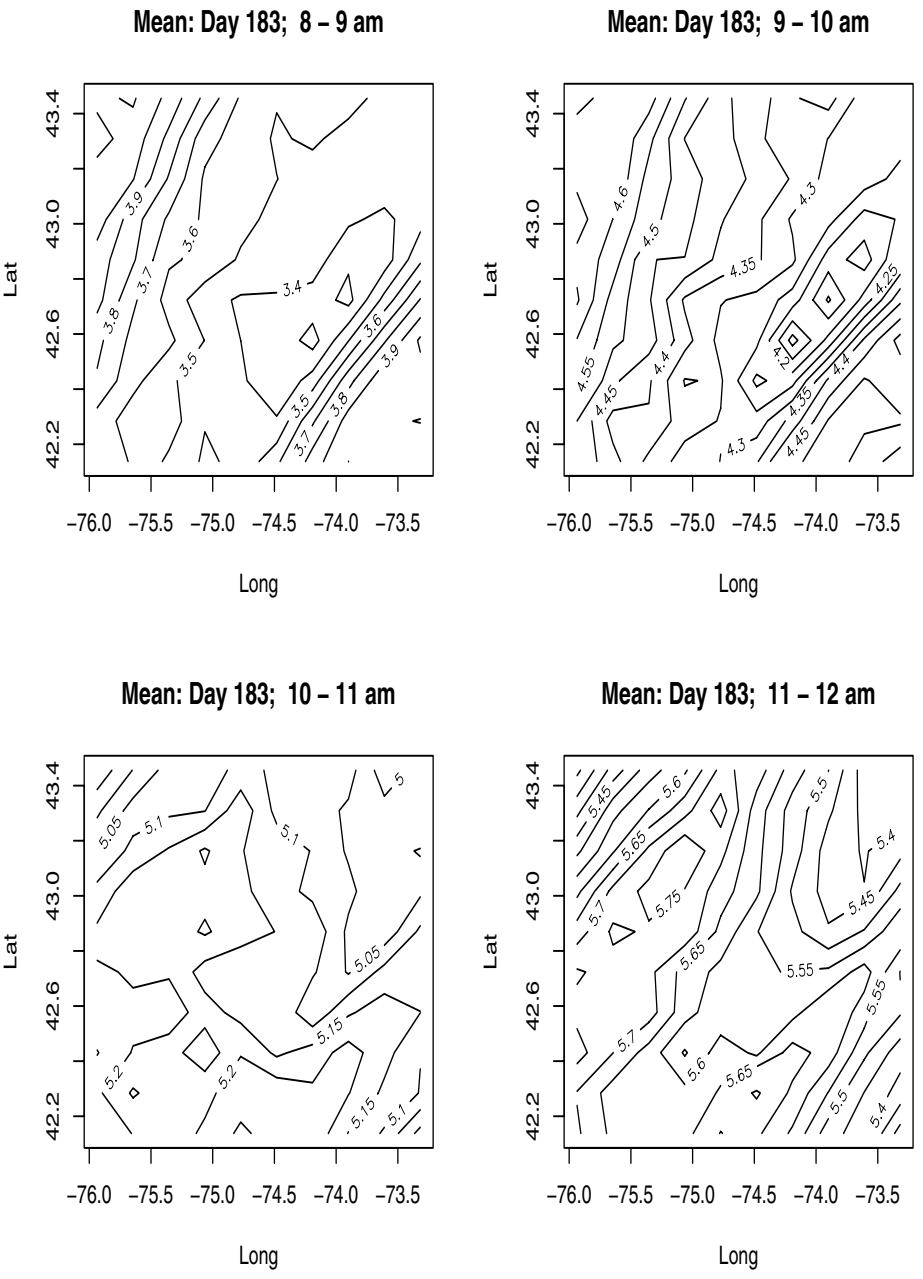


Fig. 14.12: Contour plots of sample mean.

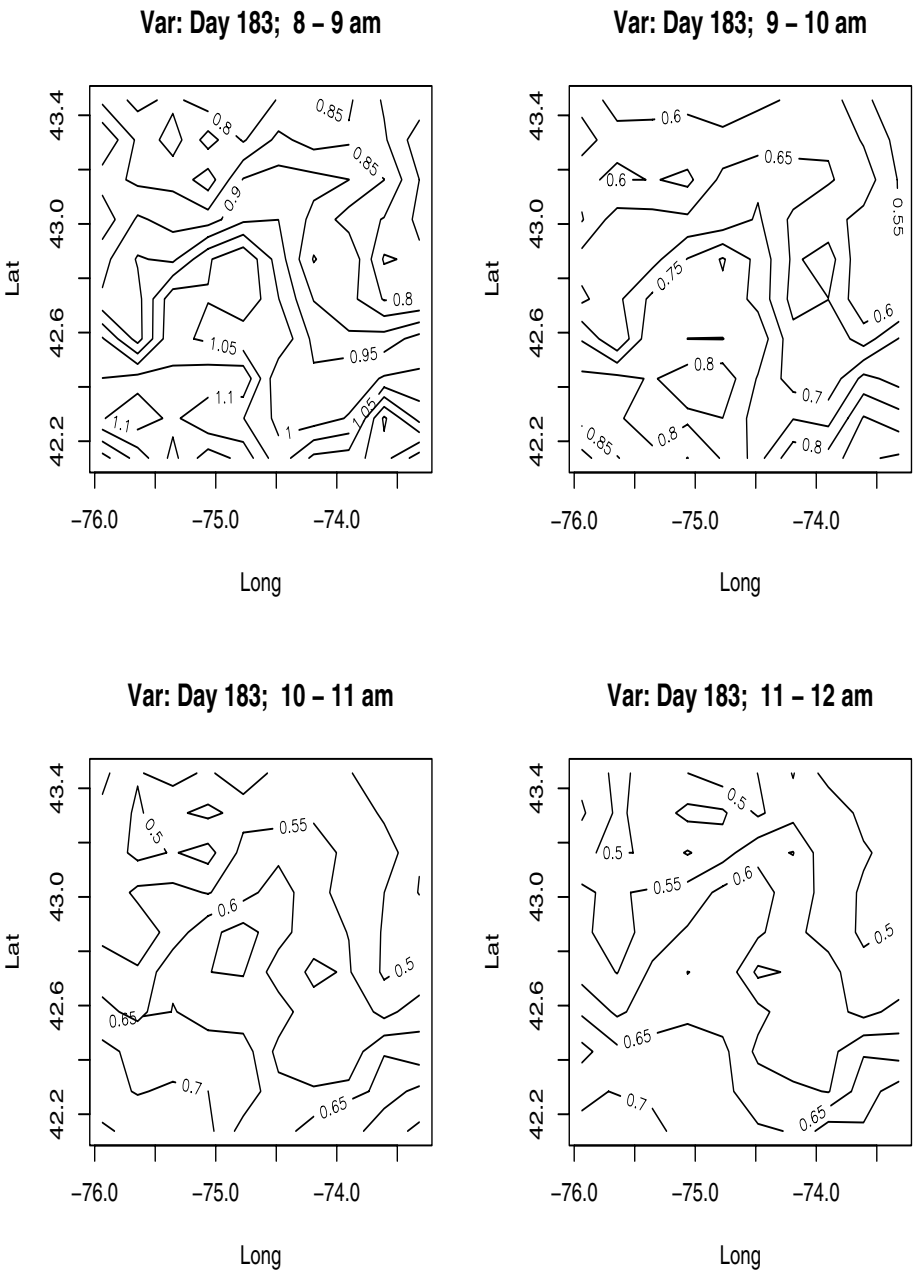


Fig. 14.13: Contour plots of sample variance.

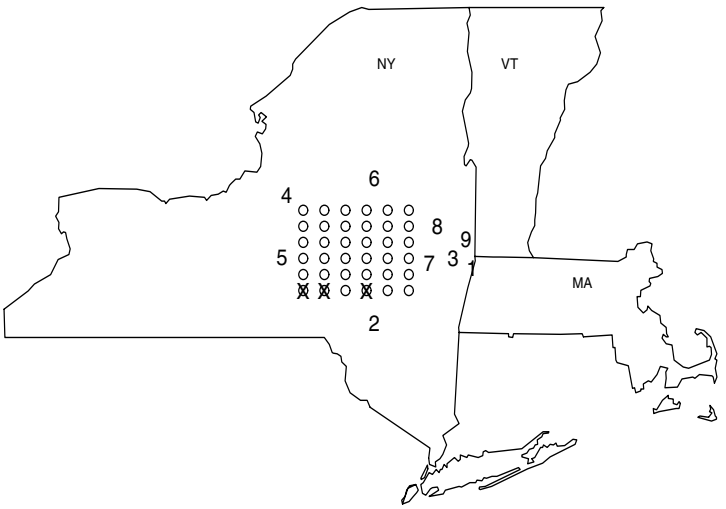


Fig. 14.14: Monitoring locations and potential new sites; the three selected sites are marked by X.

# 15

# Appendices

## 15.1 Probabilistic Distributions

### 15.1.1 Multivariate and Matrix Normal Distribution

- *Multivariate normal distribution:* A $p$-dimensional vector-valued random variable $X$ is said to have a multivariate normal distribution if its density has the form

  $$f(X) = (2\pi)^{-1/2}|\Sigma|^{-1/2}\left\{-(X-\mu)^T\Sigma^{-1}(X-\mu)/2\right\}$$

  for any vector $\mu$ and a positive definite matrix $\Sigma$. The distribution is denoted by $X \sim N_p(\mu, \Sigma)$ with

  $$\begin{aligned} E(X) &= \mu \\ Cov(X) &= \Sigma. \end{aligned}$$

- *Matrix normal distribution:* A $n \times m$ matrix valued random variable $X$ is called a matrix normal distribution if its density function has the form

  $$f(X) = (2\pi)^{\frac{-nm}{2}}|A|^{-\frac{m}{2}}|B|^{-\frac{n}{2}}\text{etr}\left\{-\frac{1}{2}[A^{-1}(X-\mu)][(X-\mu)B^{-1}]'\right\}$$

  for any $n \times m$ matrix $\mu$ and positive definite matrices $A$ and $B$ specified by

  $$A = (a_{ij})_{n\times n}, \quad B = (b_{ij})_{m\times m}.$$

  Let

  $$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = (x^{(1)}, \ldots, x^{(m)}).$$

  The distribution is denoted by $X \sim N(\mu, A \otimes B)$. The distribution has the following properties.

- $E(X) = \mu$
- $\text{var}[vec(X)] = A \otimes B$ and $\text{var}[vec(X')] = B \otimes A$.
- $X \sim N(\mu, A \otimes B)$ if and only if $X' \sim N(\mu', B \otimes A)$.
- $\text{cov}(x_i, x_j) = a_{ij}B, \quad \text{cov}(x^{(i)}, x^{(j)}) = b_{ij}A.$
- For any matrix $C_{c\times n}$ and matrix $D_{m\times d}$
  $CXD \sim N(C\mu D, CAC' \otimes D'BD)$.
- For any matrix $F_{m\times m}$
  $EXFX' = \mu F\mu' + A\text{tr}(FB)$,
  and for any matrix $G_{n\times n}$
  $EX'GX = \mu G\mu' + \text{tr}(AG)B$.
  Thus,
  $EXB^{-1}X' = \mu B^{-1}\mu' + mA$
  and
  $EX'A^{-1}X = \mu' A^{-1}\mu + nB$.

### 15.1.2 Multivariate and Matric-*t* Distribution

- *Multivariate-t distribution:* A $p$-dimensional vector-valued random variable $X$ is said to have a multivariate-$t$ distribution with $\nu$ degrees of freedom, if its density is of the form

$$f(X) = \frac{\Gamma\left(\frac{p+\nu}{2}\right)\sqrt{|A|}}{\Gamma(\nu/2)\sqrt{2\pi p}} \times \left[1 + \frac{1}{\nu}(X-\mu)^T A(X-\mu)\right]^{-(p+\nu)/2}$$

  for any vector $\mu$ and a positive definite matrix $A$. The distribution is denoted by

$$X \sim t_p(\mu, A, \nu)$$

  with $A$ called the *precision matrix* and

$$E(X) = \mu$$
$$Cov(X) = \frac{\nu}{\nu-2}A.$$

- *Matric-t distribution:* A $n \times m$ matrix-valued random variable $X$ is said to have a matric-$t$ distribution with $\delta$ degrees of freedom, if its density function has the form

$$f(X) \propto |A|^{-\frac{m}{2}}|B|^{-\frac{n}{2}}|I_n + \delta^{-1}[A^{-1}(X-\mu)][(X-\mu)B^{-1}]^T|^{-\frac{\delta+n+m-1}{2}},$$

  for positive definite matrices $A$ and $B$ of dimensions $n \times n$ and $m \times m$, respectively, and any $n \times m$ matrix $\mu$. The normalizing constant of the density is given by

$$K = (\delta\pi^2)^{-(mn/2)}\frac{\Gamma_{n+m}[(\delta+n+m-1)/2]}{\Gamma_n[(\delta+n-1)/2]\Gamma_m[(\delta+m-1)/2]},$$

  where

$$\Gamma_p(t) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma[t - (i-1)/2] \tag{15.1}$$

denotes the multivariate gamma function.
The distribution is denoted by $X \sim t_{n\times m}(\mu,\ A \otimes B,\ \delta)$, and has the following properties; see Sun (2001) for more details:

- $E(X) = \mu$
- When $\delta > 2$,
  $\text{var}[vec(X)] = \delta(\delta-2)^{-1} A \otimes B.$
  and
  $\text{cov}(x_i, x_j) = \delta(\delta-2)^{-1} a_{ij} B, \quad \text{cov}(x^{(i)}, x^{(j)}) = \delta(\delta-2)^{-1} b_{ij} A.$
- $X \sim t_{n\times m}(\mu,\ A \otimes B,\ \delta)$, if and only if $X' \sim t_{m\times n}(\mu',\ B \otimes A,\ \delta)$.
- If $n = 1$ and $A = 1$, $X$ has an multivariate $t$-distribution; i.e.,
  $X \sim t_m(\mu,\ B,\ \delta).$
- If $m = 1$ and $B = 1$, $X$ has an $n$-variate $t$-distribution; i.e.,
  $X \sim t_n(\mu,\ A,\ \delta).$
- If $X \sim t_{n\times m}(\mu,\ A \otimes B,\ \delta)$, and $C_{c\times n}$ and $D_{m\times d}$ are of full rank (i.e., rank $c$ and $d$, respectively), then
  $Y = CXD \sim t_{c\times d}(C\mu D, CAC' \otimes D'BD,\ \delta).$

### 15.1.3 Wishart and Inverted Wishart Distribution

- *Wishart distribution:* A $p \times p$ positive definite matrix $S$ is said to have a Wishart distribution with $m$ degrees of freedom, if its density is of the form

$$f(S) = \left[2^{mp/2}\Gamma_p(m/2)\right]^{-1} |A|^{-m/2} |S|^{(m-p-1)/2} e^{-tr(A^{-1}S)/2}$$

  for any positive definite matrix $A$ where $\Gamma_p$ is the multivariate gamma function defined by (15.1). The distribution is denoted by $S \sim W_p(A, m)$.
- *Inverted Wishart distribution:* A $p \times p$ positive definite matrix $\Sigma$ has an inverted Wishart distribution with $\delta$ degrees of freedom if its density function is of the form

$$f(\Sigma) = [2^{mp/2}\Gamma_p(m/2)]^{-1} |\Psi|^{\delta/2} |\Sigma|^{-(\delta+p+1)/2} exp\{-tr\Sigma^{-1}\Psi/2\}$$

  for any positive definite matrix $\Psi$. The distribution is denoted by $\Sigma \sim W_p^{-1}(\Psi, \delta)$.

**Properties of the Wishart and Inverted Wishart Distribution**

- $Y \sim W_p^{-1}(\Psi, \delta)$ if and only if $Z = Y^{-1} \sim W_p(\Psi^{-1}, \delta)$.
- If $Z \sim W_p(\Sigma, \delta)$ then $E(Z) = \delta\Sigma$ and $E(Z^{-1}) = \Sigma^{-1}/(\delta - p - 1)$ provided $\delta - p - 1 > 0$.
- If $Y \sim W_p^{-1}(\Psi, \delta)$, then $E(Y) = \Psi/(\delta - p - 1)$ and $E(Y^{-1}) = \delta\Psi^{-1}$.

– If $Y \sim W_p^{-1}(\Psi, \delta)$, then

$$E \log |Y| = -p \log 2 - \sum_{i=1}^{p} \eta \left[ \frac{1}{2}(\delta - i + 1) \right] + \log |\Psi|,$$

where $\eta$ is the digamma function; that is, $\eta(x) = d[\log \Gamma(x)]/dx$ (Chen 1979).

See Anderson (2003) for more details.

### 15.1.4 Generalized Inverted Wishart Distribution

Let a $g \times g$ positive matrix $\Sigma$ having a $k$-block structure be written as

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,k} \\ \vdots & \cdots & \vdots \\ \Sigma_{k,1} & \cdots & \Sigma_{k,k} \end{pmatrix},$$

with $\Sigma_{i,j}$ having dimensions $g_i \times g_j$. That is, $g = g_1 + \cdots + g_k$.

Denote the submatrix corresponding to the $j$th to $k$th blocks as $\Sigma^{[j,\cdots,k]}$ where

$$\Sigma^{[j,\ldots,k]} = \begin{pmatrix} \Sigma_{j,j} & \ldots & \Sigma_{j,k} \\ \vdots & \ldots & \vdots \\ \Sigma_{k,j} & \ldots & \Sigma_{k,k} \end{pmatrix}.$$

Let

$$\Sigma^{[j(j+1)]} = (\Sigma_{j,j+1}, \ldots, \Sigma_{j,k})$$

and

$$\Sigma^{[(j+1)j]} = (\Sigma_{j+1,j}, \cdots, \Sigma_{k,j}).$$

Let $\Psi$ be a $g \times g$ positive definite having the same k-block structure denoted by $\{\Psi_{i,j}\}$. Similarly let $\Psi^{[j,\cdots,k]}$ denote the submatrix corresponding to the $j$th to $k$th blocks. Let $\delta = (\delta_1, \cdots, \delta_k)$ be $k$-dimensional positive vectors and denote $\delta^{[j,\cdots,k]} = (\delta_j, \cdots, \delta_k)$.

$\Sigma$ is said to have a generalized inverted Wishart distribution, denoted by $GIW(\Psi, \delta)$, if

$$\begin{cases} \Sigma^{[2,\cdots,k]} \sim GIW(\Psi^{[2,\cdots,k]}, \delta^{[2,\cdots,k]}), \\ \\ \Gamma_1 \sim IW(\Psi_1, \delta_1), \\ \\ \tau_1 \mid \Gamma_1 \sim N\left(\tau_{01}, H_1 \otimes \Gamma_1\right), \end{cases}$$

where

$$\Gamma_1 = \Sigma_{1,1} - \Sigma^{[1(2)]}(\Sigma^{[2,\cdots,k]})^{-1}\Sigma^{[(2)1]}$$

$$\tau_1 = \left(\Sigma^{[2,\cdots,k]}\right)^{-1}\Sigma^{[(2)1]},$$

$$\Psi_1 = \Psi_{1,1} - \Psi^{[1(2)]}(\Psi^{[2,\cdots,k]})^{-1}\Psi^{[(2)1]},$$

and IW denotes the inverted Wishart distribution with $\delta_1$ degrees of freedom and the hyperparameter matrix $\Psi_1$; the matrix $\tau_{01}$ is the hyperparameter of $\tau_1$; and the matrix $H_1$ is the variance component of $\tau_1$ between its rows.

Note that the GIW distribution is defined recursively. At the first step, the matrix $\Sigma$ is partitioned into two components: the first block and the 2nd to $k$th blocks with distribution specified above. The next step is to partition the 2nd to $k$th blocks into two components: the second block and the 3rd to $k$th blocks where

$$\begin{cases} \Sigma^{[3,\cdots,k]} \sim GIW(\Psi^{[3,\cdots,k]}, \delta^{[3,\cdots,k]}), \\ \Gamma_2 \sim IW(\Psi_2, \delta_2), \\ \tau_2 \mid \Gamma_2 \sim N\left(\tau_{02}, H_2 \otimes \Gamma_2\right), \end{cases}$$

with

$$\begin{aligned} \Gamma_2 &= \Sigma_{2,2} - \Sigma^{[2(3)]}\left(\Sigma^{[3,\cdots,k]}\right)^{-1}\Sigma^{[(3)2]} \\ \tau_2 &= \left(\Sigma^{[3,\cdots,k]}\right)^{-1}\Sigma^{[(3)2]}, \\ \Psi_2 &= \Psi_{2,2} - \Psi^{[2(3)]}\left(\Psi^{[3,\cdots,k]}\right)^{-1}\Psi^{[(3)2]}. \end{aligned}$$

Similarly at the $j$th step for $j < k$, the submatrix $\Sigma^{[j,\cdots,k]}$ is partitioned into two components: the $j$th block and the remaining blocks with

$$\begin{cases} \Sigma^{[j+1,\cdots,k]} \sim GIW(\Psi^{[j+1,\cdots,k]}, \delta^{[j+1,\cdots,k]}, \\ \Gamma_j \sim IW(\Psi_j, \delta_j), \\ \tau_j \mid \Gamma_j \sim N\left(\tau_{0j}, H_j \otimes \Gamma_j\right) \end{cases}$$

with

$$\begin{aligned} \Gamma_j &= \Sigma_{j,j} - \Sigma^{[j(j+1)]}\left(\Sigma^{[j+1,\cdots,k]}\right)^{-1}\Sigma^{[(j+1)j]} \\ \tau_j &= \left(\Sigma^{[j+1,\cdots,k]}\right)^{-1}\Sigma^{[(j+1)j]}, \\ \Psi_j &= \Psi_{j,j} - \Psi^{[j(j+1)]}\left(\Psi^{[j+1,\cdots,k]}\right)^{-1}\Psi^{[(j+1)j]}. \end{aligned}$$

At the last step,

$$\Gamma_k \sim IW(\Psi_k, \delta_k)$$

with

$$\begin{aligned} \Gamma_k &= \Sigma_{k,k} \\ \Psi_k &= \Psi_{k,k}. \end{aligned}$$

See Brown et al. (1994b) for more details.

## 15.2 Bartlett Decomposition

### 15.2.1 Two-Block Decomposition

Let a covariance matrix $\Sigma$ be represented as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

The matrix $\Sigma$ could be decomposed as

$$\Sigma = T \Delta T^T,$$

where

$$\Delta = \begin{pmatrix} \Sigma_{1|2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} I & \tau \\ 0 & I \end{pmatrix},$$

with

$$\Sigma_{1|2} \equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

and

$$\tau \equiv \Sigma_{12}\Sigma_{22}^{-1}.$$

Hence

$$\Sigma = \begin{pmatrix} \Sigma_{1|2} + \tau\Sigma_{22}\tau^T & \tau\Sigma_{22} \\ \Sigma_{22}\tau^T & \Sigma_{22} \end{pmatrix}.$$

This one–to–one transformation is commonly known as the Bartlett decomposition (Bartlett 1933).

### 15.2.2 Recursive Bartlett Decomposition for Multiple Blocks

Let a covariance matrix $\Sigma$ be represented as having $k$ blocks

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,k} \\ \vdots & \cdots & \vdots \\ \Sigma_{k,1} & \cdots & \Sigma_{k,k} \end{pmatrix},$$

with $\Sigma_{i,j}$ having dimensions $g_i \times g_j$. That is, $g = g_1 + \cdots + g_k$.

Denote the submatrix corresponding to the $j$th to $k$th blocks as $\Sigma^{[j,\cdots,k]}$ where

$$\Sigma^{[j,\ldots,k]} = \begin{pmatrix} \Sigma_{j,j} & \cdots & \Sigma_{j,k} \\ \vdots & \cdots & \vdots \\ \Sigma_{k,j} & \cdots & \Sigma_{k,k} \end{pmatrix}.$$

Let

$$\Sigma^{[j(j+1)]} = (\Sigma_{j,j+1}, \cdots, \Sigma_{j,k})$$

and

$$\Sigma^{[(j+1)j]} = (\Sigma_{j+1,j}, \cdots, \Sigma_{k,j}).$$

Applying the Bartlett decomposition for two-blocks above recursively, one can represent the (sub-)matrix $\Sigma^{[j,\ldots,k]}$ , for $j < k$, as

$$\Sigma^{[j,\cdots,k]} = \begin{pmatrix} \Gamma_j + \tau_j^T \Sigma^{[j+1,\cdots,k]} \tau_j & \tau_j^T \Sigma^{[j+1,\cdots,k]} \\ \Sigma^{[j+1,\cdots,k]} \tau_j & \Sigma^{[j+1,\cdots,k]} \end{pmatrix},$$

$$\Gamma_j = \Sigma_{j,j} - \Sigma^{[j,(j+1)]} (\Sigma^{[j+1,\ldots,k]})^{-1} \Sigma^{[(j+1),j]},$$

$$\tau_j = (\Sigma^{[j+1,\ldots,k]})^{-1} \Sigma^{[(j+1),j]},$$

and

$$\Sigma_{kk} = \Sigma^{[k,k]}.$$

## 15.3 Useful Matrix Properties

1. Let $\mathbf{A}_{a\times a}, \mathbf{B}_{b\times b}, \mathbf{C}$ be matrices having elements $a_{ij}, b_{ij}$, and $c_{ij}$, respectively, then

$$tr(\mathbf{A} \otimes \mathbf{B})\mathbf{C} = tr(\mathbf{BM}), \tag{15.2}$$

where $\mathbf{M}$ is given below. Take a simple setting for example,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

Partition $\mathbf{C}$ as

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix},$$

where each $\mathbf{C}_{ij}$ is a $b \times b$ matrix. Then in this case

$$\begin{aligned} tr(\mathbf{A} \otimes \mathbf{B})\mathbf{C} &= tr[a_{11}\mathbf{BC}_{11} + a_{12}\mathbf{BC}_{21} + a_{21}\mathbf{BC}_{12} + a_{22}\mathbf{BC}_{22}] \\ &= tr(\mathbf{BM}), \end{aligned}$$

where

$$\begin{aligned} \mathbf{M} &= a_{11}\mathbf{I}_b\mathbf{C}_{11} + a_{12}\mathbf{I}_b\mathbf{C}_{21} + a_{21}\mathbf{I}_b\mathbf{C}_{12} + a_{12}\mathbf{I}_b\mathbf{C}_{22}] \\ &= [\, a_{11}\mathbf{I}_b \; a_{12}\mathbf{I}_b \,] \begin{bmatrix} \mathbf{C}_{11} \\ \mathbf{C}_{21} \end{bmatrix} \\ &\quad + [\, a_{21}\mathbf{I}_b \; a_{22}\mathbf{I}_b \,] \begin{bmatrix} \mathbf{C}_{12} \\ \mathbf{C}_{22} \end{bmatrix} \\ &= (a_1 \otimes \mathbf{I}_b)\mathbf{C}[\mathbf{1},] + (a_2 \otimes \mathbf{I}_b)\mathbf{C}[,\mathbf{2}]], \end{aligned}$$

$a_i$ denoting the ith row of $\mathbf{A}$, $\mathbf{C}$**[1,]**, the first column of $\mathbf{C}$ and so on. Similarly a general expression for $\mathbf{M}$ can be obtained.
Note that $M = (m_{rc})$ can be explicitly expressed in terms of individual elements of $\mathbf{A}$ and $\mathbf{C}$ where

$$m_{rc} = \sum_{i=1}^{a} \sum_{j=1}^{a} a_{ij} c_{(c-1)b+j,(r-1)b+i}.$$

2. Let $\mathbf{A}_{a\times a}, \mathbf{B}_{b\times b}, \mathbf{C}$ be matrices having elements $a_{ij}, b_{ij}$, and $c_{ij}$ respectively; then

$$tr(\mathbf{A} \otimes \mathbf{B})\mathbf{C} = tr(A\tilde{\mathbf{M}}) \tag{15.3}$$

when $\tilde{\mathbf{M}}$ is given below.
Let a so-called permutation matrix $\mathbf{P} = \mathbf{P}^T = \mathbf{P}^{-1}$ such that $P(\mathbf{A} \otimes \mathbf{B})P^T = \mathbf{B} \otimes \mathbf{A}$. In fact, $\mathbf{P}$ turns out to have a simple form when expressed in terms of the so-called basis vectors $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$. More explicitly (with b = 2 in this case),

$$\begin{aligned}\mathbf{P} &= \begin{bmatrix} \mathbf{e}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{e}_1 \\ \mathbf{e}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{e}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_b \otimes \mathbf{e}_1 \\ \mathbf{I}_b \otimes \mathbf{e}_2 \end{bmatrix}.\end{aligned}$$

Thus,

$$tr(\mathbf{A} \otimes \mathbf{B})\mathbf{C} = tr(\mathbf{B} \otimes \mathbf{A})\mathbf{PCP}$$

and adopting the above result yields (with a = 2 in this case)

$$\tilde{\mathbf{M}} = (\mathbf{b}_1 \otimes \mathbf{I}_a)\mathbf{[PCP][1,]} + (\mathbf{b}_2 \otimes \mathbf{I}_a)\mathbf{[PCP][,2]}.$$

This result can be extended in the obvious way to the general case where, for example,

$$\mathbf{P} = \begin{bmatrix} \mathbf{I}_b \otimes \mathbf{e}_1 \\ \vdots \\ \mathbf{I}_b \otimes \mathbf{e}_a \end{bmatrix}.$$

Note that $\tilde{M} = \tilde{m}_{rc}$ can be explicitly expressed in terms of individual elements of $\mathbf{B}$ and $\mathbf{C}$ where

$$\tilde{m}_{rc} = \sum_{i=1}^{b} \sum_{j=1}^{b} b_{ij} c_{(j-1)a+c,(i-1)a+r}.$$

3. Let $A_{qp\times hp}, B_{qp\times qp}, C_{hp\times hp}$, and $\xi_{h\times q}$ be matrices with dimensions given; then

$$tr(I_p \otimes \xi)A \quad = \quad tr\{\xi D\} \quad = \quad vec(\xi)' vec(D) \tag{15.4}$$

and

$$tr(I_p \otimes \xi)B(I_p \otimes \xi')C \quad = \quad vec(\xi)'Gvec(\xi), \tag{15.5}$$

where $D$ and $G$ are defined below.
Specifically, let

$$P_{p,q} = \begin{bmatrix} I_p \otimes e_{q1} \\ \vdots \\ I_p \otimes e_{q1} \end{bmatrix}$$

be a general permutation matrix where $e_{qj}$ denotes the q-dimensional row vector all of whose elements are 0 save for the $j$th which is 1. Then for any matrices, $\alpha : k \times r$ and $\beta : l \times s$,

$$P_{k,l}(\alpha \otimes \beta)P_{r,s}^T = \beta \otimes \alpha.$$

Hence,

$$\begin{aligned} tr(\xi^{h\times q} \otimes I_p)A &= tr[P_{p,h}(I_p \otimes \xi)P_{p,q}^T A] \\ &= tr[\{(I_p \otimes \xi)P_{p,q}^T AP_{p,h}\}] \\ &= tr\{(I_p \otimes \xi)\tilde{\mathbf{A}}\}, \end{aligned}$$

where $\tilde{\mathbf{A}} = P_{p,q}^T AP_{p,h}$.
Similarly

$$\begin{aligned} tr\{(\xi \otimes I_p)B(I_p \otimes \xi)C\} &= tr\{(I_p \otimes \xi)P_{p,q}^T BP_{p,h}(I_p \otimes \xi')P_{p,p}^T CP_{q,h}\} \\ &= tr\{(I_p \otimes \xi)\tilde{\mathbf{B}}(I_p \otimes \xi)\tilde{\mathbf{C}}\}, \end{aligned}$$

where $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{C}}$ have definitions analogous to $\tilde{\mathbf{A}}$ above.
Partition the matrices involved as

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{A}_{11}^{q\times h} & \cdots & \tilde{A}_{1p} \\ \vdots & \vdots & \vdots \\ \tilde{A}_{p1} & \cdots & \tilde{A}_{pp} \end{bmatrix}$$

and similarly

$$\tilde{\mathbf{B}} = \begin{bmatrix} \tilde{B}_{11}^{h\times q} & \cdots & \tilde{B}_{1p} \\ \vdots & \vdots & \vdots \\ \tilde{B}_{p1} & \cdots & \tilde{B}_{pp} \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{C}} = \begin{bmatrix} \tilde{C}_{11}^{h\times h} & \cdots & \tilde{C}_{1p} \\ \vdots & \vdots & \vdots \\ \tilde{C}_{21} & \cdots & \tilde{C}_{pp} \end{bmatrix}.$$

Observe that:

$$(I_p \otimes \xi) = \begin{bmatrix} \xi & \ldots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \xi \end{bmatrix}.$$

Hence

$$tr(I_p \otimes \xi)A = tr\left\{ \begin{bmatrix} \xi & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \xi \end{bmatrix} \begin{bmatrix} \tilde{A}_{11}^{q\times h} & \cdots & \tilde{A}_{12} \\ \vdots & \vdots & \vdots \\ \tilde{A}_{p1} & \cdots & \tilde{A}_{pp} \end{bmatrix} \right\}$$
$$= tr\left\{\xi D\right\},$$

where $D = \tilde{A}_{11} + \cdots + \tilde{A}_{pp}$. Similarly

$$(I_p \otimes \xi)B = \begin{bmatrix} \xi & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \xi \end{bmatrix} \begin{bmatrix} \tilde{B}_{11}^{h\times q} & \cdots & \tilde{B}_{1p} \\ \vdots & \vdots & \vdots \\ \tilde{B}_{p1} & \cdots & \tilde{B}_{pp} \end{bmatrix}$$
$$= \begin{bmatrix} \xi\tilde{B}_{11} & \cdots & \xi\tilde{B}_{1p} \\ \vdots & \vdots & \vdots \\ \xi\tilde{B}_{p1} & \cdots & \xi\tilde{B}_{pp} \end{bmatrix}.$$

As well

$$(I_p \otimes \xi')C = \begin{bmatrix} \xi' & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \xi' \end{bmatrix} \begin{bmatrix} \tilde{C}_{11}^{h\times q} & \cdots & \tilde{C}_{1p} \\ \vdots & \vdots & \vdots \\ \tilde{C}_{p1} & \cdots & \tilde{C}_{pp} \end{bmatrix}$$
$$= \begin{bmatrix} \xi\prime\tilde{C}_{11}^{h\times q} & \cdots & \xi\prime\tilde{C}_{1p} \\ \vdots & \vdots & \vdots \\ \xi\prime\tilde{C}_{p1} & \cdots & \xi\prime\tilde{C}_{pp} \end{bmatrix}.$$

Thus,

$$tr(I_p \otimes \xi)B(I_p \otimes \xi')C = tr\left\{ \begin{bmatrix} \xi\tilde{B}_{11} & \cdots & \xi\tilde{B}_{1p} \\ \vdots & \vdots & \vdots \\ \xi\tilde{B}_{p1} & \cdots & \xi\tilde{B}_{pp} \end{bmatrix} \begin{bmatrix} \xi\prime\tilde{C}_{11}^{h\times q} & \cdots & \xi\prime\tilde{C}_{1p} \\ \vdots & \vdots & \vdots \\ \xi\prime\tilde{C}_{p1} & \cdots & \xi\prime\tilde{C}_{pp} \end{bmatrix} \right\}$$
$$= \sum_{r=1}^{p}\sum_{s=1}^{p} tr\xi\tilde{B}_{rs}\xi'\tilde{C}_{sr}$$
$$= \sum_{r=1}^{p}\sum_{s=1}^{p} tr\xi\tilde{B}_{rs}(\tilde{C}'_{sr}\xi)'.$$

However, recall that for matrices, $U$ and $V$ with conformable dimensions,

$$(U, V) = tr(UV') = \sum U_{ij}V_{ij} = vec(U)'vec(V).$$

Thus

$$tr(I_p \otimes \xi)B(I_p \otimes \xi')C = \sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}(\xi\tilde{B}_{rs})_{kl}(\tilde{C}'_{sr}\xi)_{kl}$$

$$= \sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{k=1}^{h}\sum_{l=1}^{q}\sum_{u=1}^{q}\sum_{v=1}^{h} \xi_{ku}\tilde{B}_{rsul}\tilde{C}_{rskv}\xi_{vl}$$
$$= vec(\xi)'Gvec(\xi),$$

where using a double index notation, $G = (Gku, vl)$, with $Gku, vl = \sum_{rs}\tilde{B}_{rsul}\tilde{C}_{srkv}$ and $vec(\xi) = (\xi_{ku}) : hq \times 1$ with double subscripts ordered in accord with those of $G$. Similarly $tr\{\xi D\} = vec(\xi)'vec(D)$.

## 15.4 Proofs for Chapter 10

LEMMA 1. *Define matrices* $Y : n \times g$, $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0 : l \times g$, $\Sigma > 0, \Psi > 0 : g \times g$, $Z : n \times l$, $F > 0 : l \times l$, *and* $A > 0 : n \times n$. *The Gaussian inverted Wishart model*

$$\begin{cases} Y \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, A \otimes \Sigma) \\ \boldsymbol{\beta} \mid \Sigma \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma) \\ \Sigma \sim IW(\Psi, \delta) \end{cases}$$

*implies the following predictive distribution*

$$Y \sim t_{n\times g}\left[Z\boldsymbol{\beta}_0, (\delta - g + 1)^{-1}(A + ZF^{-1}Z^T) \otimes \Psi, \delta - g + 1\right]$$

*and the posterior distributions*

$$\boldsymbol{\beta} \mid \Sigma, Y \sim N(W\hat{\boldsymbol{\beta}} + (I - W)\boldsymbol{\beta}_0, \tilde{F}^{-1} \otimes \Sigma),$$

$$\Sigma \mid Y \sim IW(\Psi + (Y - Z\boldsymbol{\beta}_0)^T(A + ZF^{-1}Z^T)^{-1}(Y - Z\boldsymbol{\beta}_0), \delta + n),$$

where

$$W = (Z^TA^{-1}Z + F)^{-1}Z^TZ$$
$$\hat{\boldsymbol{\beta}} = (Z^TA^{-1}Z)^{-1}Z^TA^{-1}Y,$$
$$\tilde{F} = Z^TA^{-1}Z + F.$$

*Proof of Lemma 1.* The proof follows arguments of Anderson (2003); see also Brown (1993).

*Note.* Using the identity

$$(A + ZF^{-1}Z^T)^{-1} = A^{-1} - A^{-1}Z(F^{-1} + Z^TA^{-1}Z)^{-1}Z^TA^{-1}$$

one can show, on setting $(A + ZF^{-1}Z^T)^{-1} = W$

$$(Y - Z\boldsymbol{\beta}_0)^T[A + ZF^{-1}Z^T]^{-1}(Y - Z\boldsymbol{\beta}_0) =$$
$$(Y - Z\hat{\boldsymbol{\beta}}_0)^TA^{-1}(Y - Z\hat{\boldsymbol{\beta}}_0) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^TW(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

reflecting the contributions from the likelihood and the prior distribution.

To state the next lemma, let $Y = (Y^{[u]}, Y^{[g]})$, $Y^{[u]}$ and $Y^{[g]}$ having $n \times u$ and $n \times g$ dimensions, respectively. We adopt the following transformation of the partitioned covariance matrix $\Sigma$ of $Y$:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \to (\Sigma_{22}, \tau, \Gamma)$$

for matrices $\Sigma_{11} : u \times u$, $\Sigma_{21} : g \times u$, $\Sigma_{22} : g \times g$, and

$$\tau = \Sigma_{22}^{-1}\Sigma_{21}, \quad \Gamma = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

LEMMA 2. *Adopt the Gaussian and generalized inverted Wishart model specified by:*

$$\begin{cases} Y \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, A \otimes \Sigma), \\ \boldsymbol{\beta} \mid \Sigma \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma), \\ \tau \mid \Gamma \sim N(\tau_0, H \otimes \Gamma), \\ \Gamma \sim IW(\Psi_1, \delta_1), \\ \Sigma_{22} \sim IW(\Psi_2, \delta_2), \end{cases} \tag{15.6}$$

*where*

$$Z : n \times l,$$

$$\boldsymbol{\beta} = (\boldsymbol{\beta}^{[u]}, \boldsymbol{\beta}^{[g]}) : (l \times u, l \times g) \quad \textit{and}$$

$$\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0^{[u]}, \boldsymbol{\beta}_0^{[g]}) : (l \times u, l \times g).$$

*Then the predictive distribution of* $(Y^{[u]} \mid Y^{[g]})$ *is*

$$Y^{[u]} \mid Y^{[g]} \sim t_{n \times u}\left(\mu^{[u|g]}, \Phi^{[u|g]} \otimes \Psi^{[u|g]}, \delta_1 - u + 1\right),$$

*where*

$$\mu^{[u|g]} = Z\boldsymbol{\beta}_0^{[u]} + (Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]})\tau_0,$$

$$\Phi^{[u|g]} = A + ZF^{-1}Z^T + (Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]})H(Y^{[g]} - Z\boldsymbol{\beta}_0^{[g]})^T,$$

$$\Psi^{[u|g]} = (\delta_1 - u + 1)^{-1}\Psi_1.$$

*Proof of Lemma 2.* (i) Suppose $\boldsymbol{\beta} = 0$. Then by standard results for the multivariate normal distribution, the conditional distribution of $(Y^{[u]} \mid Y^{[g]}, \Sigma)$, which does not depend on $\Sigma_{22}$, can be expressed as

$$(Y^{[u]} \mid Y^{[g]}, \tau, \Gamma) \sim N(Y^{[g]}\tau, I_n \otimes \Gamma).$$

Applying Lemma 1 to this distribution with the prior distributions of $\tau$ and $\Gamma$ in (15.6) yields

$$(Y^{[u]} \mid Y^{[g]}) \sim t_{n\times u}\left(Y^{[g]}\tau_0, (\delta_1 - u + 1)^{-1}(I_n + Y^{[g]}HY^{[g]^T}) \otimes \Psi_1, \delta_1 - u + 1\right). \quad (15.7)$$

(ii) Now suppose $\boldsymbol{\beta}$ follows the distribution in (15.6). Notice that

$(A + ZF^{-1}Z^T)^{-1/2}(Y - Z\boldsymbol{\beta}_0) \mid \Sigma \sim N(0, I_n \otimes \Sigma)$.

The lemma follows immediately from the result in (i).

Lemma 3. *In the setting of Lemma 2, assume further*

$$Y^{[g]} = \begin{pmatrix} Y^{[g]}_{(1)} \\ Y^{[g]}_{(2)} \end{pmatrix},$$

*where the matrix* $Y^{[g]}_{(1)} : m \times g,\ m < n$, *holds the unobserved responses at the gauged sites. Let*

$$\begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix} : \begin{pmatrix} m \times g \\ (n-m) \times g \end{pmatrix} = Z\boldsymbol{\beta}^{[g]}_0,$$

$$\begin{pmatrix} A_{11}\ A_{12} \\ A_{21}\ A_{22} \end{pmatrix} : \begin{pmatrix} m \times m & m \times (n-m) \\ (n-m) \times m & (n-m) \times (n-m) \end{pmatrix} = A + ZF^{-1}Z^T,$$

$$\mu_{(u|g)} = \mu_{(1)} + A_{12}(A_{22})^{-1}(Y^{[g]}_{(2)} - \mu_{(2)}),$$

$$\Phi_{(u|g)} = \frac{\delta_2 - g + 1}{\delta_2 - g + n - m + 1}\left[A_{11} - A_{12}(A_{22})^{-1}A_{21}\right],$$

$$\Psi_{(u|g)} = \frac{1}{\delta_2 - g + 1}\left[\Psi_2 + (Y^{[g]}_{(2)} - \mu_{(2)})^T(A_{22})^{-1}(Y^{[g]}_{(2)} - \mu_{(2)})\right],$$

$$\delta_{(u|g)} = \delta_2 - g + n - m + 1.$$

*Then the predictive distribution of* $Y^{[g]}_{(1)}$ *given data* $Y^{[g]}_{(2)}$ *is*

$$\left(Y^{[g]}_{(1)} \mid Y^{[g]}_{(2)}, \mathcal{H}\right) \sim t_{m\times g}\left(\mu_{(u|g)}, \Phi_{(u|g)} \otimes \Psi_{(u|g)}, \delta_{(u|g)}\right). \quad (15.8)$$

*Proof of Lemma 3.* The GIW model (15.6) implies the GIW submodel

$$\begin{cases} Y^{[g]} \mid \boldsymbol{\beta}^{[g]}, \Sigma_{22} \sim N(Z\boldsymbol{\beta}^{[g]}, A \otimes \Sigma_{22}), \\ \boldsymbol{\beta}^{[g]} \mid \Sigma_{22} \sim N(\boldsymbol{\beta}_0^{[g]}, F^{-1} \otimes \Sigma_{22}), \\ \Sigma_{22} \sim IW(\Psi_2, \delta_2). \end{cases}$$

Therefore, by Lemma 1

$$Y^{[g]} \sim t_{n\times g}\left[Z\boldsymbol{\beta}_0^{[g]}, (\delta_2 - g + 1)^{-1}(A + ZF^{-1}Z^T) \otimes \Psi_2, \delta_2 - g + 1\right]. \quad (15.9)$$

Conditioning the distribution in (15.9) upon $Y_{(2)}^{[g]}$ yields the predictive distribution (15.8).

*Proof of (9.15)–(9.18)* Part (i) of the proof is a straightforward application of Lemma 3. For part (ii), the distribution is obtained by first applying Lemma 2 to $Y^{[g_j]}$ conditional on $Y^{[g_{j+1},\ldots,g_k]}$ and then applying Lemma 3 to $Y^{[g_j^m]}$ conditional on $Y^{[g_j^o]}$ and $Y^{[g_{j+1},\ldots,g_k]}$. Part (iii) is an immediate result of Lemma 2.

*Proof of (10.7) and (10.9)* . (i) Model (10.1) implies

$$\begin{cases} \boldsymbol{\beta}^{[g_k]} \mid \Sigma_{kk}, \boldsymbol{\beta}_0, F \sim N(\boldsymbol{\beta}_0^{[k]}, F^{-1} \otimes \Sigma_{kk}), \\ Y^{[g_k^o]} \mid \boldsymbol{\beta}^{[g_k]}, \Sigma_{kk}, \mathcal{H} \sim N(Z_{(k)}\boldsymbol{\beta}^{[g_k]}, I_{n-m_k} \otimes \Sigma_{kk}), \end{cases} \quad (15.10)$$

which gives the posterior distribution of $(\boldsymbol{\beta}^{[g_k]} \mid D, \Sigma_{kk}, \mathcal{H})$, as in (10.7), by means of Lemma 1.

Similarly, model (10.1) implies

$$\begin{cases} \boldsymbol{\beta}^{[g_j]} \mid \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}, \tau_j, \Gamma_j, \mathcal{H} \sim N\left(\boldsymbol{\beta}_0^{[g_j]} + (\boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]} - \boldsymbol{\beta}_0^{[g_{j+1},\ldots,g_k]})\tau_j, F^{-1} \otimes \Gamma_j\right), \\ Y^{[g_j^o]} \mid Y_{(j)}^{[g_{j+1},\ldots,g_k]}, \boldsymbol{\beta}^{[g_j,g_{j+1},\ldots,g_k]}, \tau_j, \Gamma_j, \mathcal{H} \sim N(Z_{(j)}\boldsymbol{\beta}^{[g_j]} + \tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]}\tau_j, F^{-1} \otimes \Gamma_j), \end{cases} \quad (15.11)$$

where

$$\tilde{\epsilon}_{(j)}^{[g_{j+1},\ldots,g_k]} = Y_{(j)}^{[g_{j+1},\ldots,g_k]} - Z_{(j)}\boldsymbol{\beta}_0^{[g_{j+1},\ldots,g_k]}.$$

Applying Lemma 1 again gives the posterior of $(\boldsymbol{\beta}^{[g_j]} \mid D, \boldsymbol{\beta}^{[g_{j+1},\ldots,g_k]}, \tau_j, \Gamma_j, \mathcal{H})$ as in (10.7).

(ii) Combining (15.10), (15.11), as well as (10.2) and using the result of Lemma 1 yields the distribution (10.9)

*Proof of Corollary 1.* Taking conditional expectations of the $\boldsymbol{\beta}$s given $(D, \mathcal{H})$ in Theorem 2 yields the result.

# 16

# References.


Abramowitz M, Stegun IE (1970). *Handbook of Mathematical Functions.* Washington, DC: National Bureau of Stndards.

Anderson TW (2003). *An Introduction to Multivariate Statistical Analysis.* New York: Wiley.

Angulo JM, Bueso MC (2001). Random perturbation methods applied to multivariate spatial sampling design. *Environmetrics*, 12, 631–646.

Angulo JM, Bueso MC, Alonso FJ (2000). A study on sampling design for optimal prediction of space–time stochastic processes. *Stochastic Environ Res Risk Assess*, 14, 412–427.

Angula JM (1998). The kriged Kalman filter. Discussant. *Test*, 7, 217–285.

Aranda-Ordaz FJ (1981). On two families of transformations to additivity for binary response data. *Biometrika*, 68, 357–363.

Arnold BC, Castillo E, Sarabia JM (1999). *Conditional Specification of Statistical Models.* New York: Springer.

Atkinson AC, Shepard N (1996). Deletion diagnostics for transformations of time-series. *J Forecasting,* 15, 1–17.

Aunan K (1996). Exposure-response Functions for health effects of air pollutants based on epidemiological findings. *Risk Anal*, 16, 693–709.

Bailer AJ, Oris JT, See K, Hughes MR, Schaefer R (2003). Defining and evaluating impact in environmental toxicology. *Environmetrics*, 14, 235–243.

Banerjee S, Carlin BP, Gelfand AE (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall, CRC.

Bartlett MS (1933). On the theory of statistical regression. *Proc Roy Soc Edinburgh*, 53, 260–283.

Bates DV, Caton RB (2002). *A Citizen's Guide to Air Pollution. Second Edition.* Vancouver: David Suzuki Foundation.

Beckmann P (1973). *Orthogonal Polynomials for Engineers and Physicists.* Boulder, CO: Golem Press.

Bengtsson T, Snyder C, Nychka D (2003). Toward a nonlinear ensemble filter for high- dimensional systems. *J Geophys Res Atmosphere*, 108, No. D24, 8775.

Bernardo JM, Smith AFM (1994). *Bayesian Statistics.* Chichester: Wiley.

Besag J (1974). Spatial interaction and the statistical analysis of lattice systems. *J Roy stat Soc, Ser B*, 36, 192–236.

Besag J, Higdon D (1999). Bayesian analysis of agriculture field experiments. *J of Roy stat Soc, Ser B*, 61, 691–746.

Bilonick RA (1988). Monthly hydrogen ion deposition maps for the Northeastern U.S. from July 1982 to September 1984. *Atmos Environ* 22(9), 1909–1924.

Bortot P, Coles SG, Tawn J (2000). The multivariate Gaussian tail model. *Appl Statist*, 49, 31–49.

Box GEP, Cox DR (1964). An analysis of transformations (with discussion). *J Roy Stat Soc, Ser B*, 26, 211–252.

Brown PJ (1993). *Measurement, Regression, and Calibration.* Oxford: Clarendon.

Brown PJ, Le ND, Zidek JV (1994a). Multivariate spatial interpolation and exposure to air pollutants. *Can J Statist*, 22, 489–510.

Brown PJ, Le ND, Zidek JV (1994b). Inference for a covariance matrix. In *Aspects of Uncertainty: A Tribute to DV Lindley*, (Eds) PR Freeman and AFM Smith. Chichester: Wiley, 77–92.

Budtz-Jørgensen E, Keiding N, Grandjean P, Weihe P, White RF (2003). Statistical methods for the evaluation of health effects of prenatal mercury exposure. *Environmetrics*, 14, 105–120.

Bueso MC, Angulo JM, Alonso FJ (1998). A state-space model approach to optimum spatial sampling design based on entropy. *Environ Ecol Statist*, 5, 29–44.

Bueso MC, Angulo JM, Curz-Sanjulià n J, García-Aróstegui JL (1999b). Optimal spatial sampling design in a multivariate framework. *Math Geol*, 31, 507–525.

Bueso MC, Angulo JM, Qian G, Alonso FJ (1999a). Spatial sampling design based on stochastic complexity. *J Mult Anal*, 71, 94–110.

Burnett D, Krewski D (1994). Air pollution effects on hospital admission rates: a random effects modeling approach, *Can J Statist*, 22, 441–458.

Burnett RT, Dales RE, Raizenne MR, Krewski D, Summers PW, Roberts GR, Raad-Young M, Dann T, Brook J (1994). Effects of low ambient levels of ozone and sulphates on the frequency of respiratory admissions to Ontario hospitals. *Environ Res*, 65, 172–94.

Carroll RJ, Ruppert D, Stefanski LA (1995). *Measurement Error in Nonlinear Models.* London: Chapman and Hall.

Caselton WF, Husain T (1980). Hydrologic networks: Information transmission. *J Water Resources Plan Manage Div, A.S.C.E.*, 106, 503–520.

Caselton WF, Zidek JV (1984). Optimal monitoring network designs. *stat Prob Letters*, 2, 223–227.

Caselton WF, Kan L, Zidek JV (1992). Quality data networks that minimize entropy. In *Statistics in the Environtal and Earth Sciences*, (Eds) P Guttorp and A Walden. London: Griffin.

Chami H, Gonzalez PL (1984). Amelioration d'un Reseau de Surveillance de la Pollution Atmospherique. *Montpellier Cedex, Unite de Biometrie.*

Chang H, Fu AQ, Le ND, Zidek JV (2006). Designing environmental monitoring networks to measure extremes. *Environmetrics.* To appear.

Chen CF (1979). Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *J Roy stat Soc, Ser B* 41, 235–48.

Christakos G (1984). On the problem of permissible covariance and variogram models. *Water Resources Res*, 20, 251–265.

Coles SG (2001). *An Introduction to Statistical Modelling of Extreme Values.* UK:Springer.

Coles SG, Pauli F (2002). Models and inference for uncertainty in extremal dependence. *Biometrika*, 89, 183–196.

Coles SG, Tawn JA (1996) Modelling extremes of the areal rainfall process. *J Roy Statist Soc, Ser B* 58, 329–347.

Coles SG, Tawn JA (1994). Statistical methods for multivariate extremes: an application to structural design. *Appl Statist*, 43, 1–48.

Craigmile PF, Guttorp P, Percival DB (2003). Trend assessment in a long memory dependence model using a discrete wavelet transformation. *Environmetrics*, 14, 1–23.

Cressie N (1991). The origins of kriging. *Math Geol*, 22, 239–252.

Cressie N (1993). *Statistics for Spatial Data.* New York: Wiley.

Cressie N and Wikle CK (1998) The kriged Kalman filter. Discussants. *Test*, 7, 217–285.

Damian D, Sampson PD, Guttorp P (2001). Bayesian estimation of semi–parametric non-stationary spatial covariance structure. *Environmetrics*, 12, 161–176.

Davison AC, Smith RL (1990). Models for exceedances over high thresholds (with discussion). *J Roy stat Soc, Ser B*, 52, 393–442.

Dawid AP, Stone M, Zidek JV (1973). Marginalization paradoxes in Bayesian and structural inference. *J Roy Statist Soc, Ser B* 35, 189–233.

Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J Roy stat Soc, Ser B*, 39, 1–38.

De Oliveira V, Kedem B, Short DA (1997). Bayesian prediction of transformed Gaussian random fields. *J Amer stat Assoc*, 92, 1422–1433.

De Wijs HJ (1951). Statistics of ore distribution, part I: Frequency distribution of assay values. *Geologie en Mijnbouw*, 13, 365–375.

Diggle PJ, Liang KY, Zeger SL (1994). *Analysis of Longitudinal Data.* Oxford: Clarendon.

Doob JL (1953). *Stochastic Processes.* New York: Wiley.

Duddek C, Le ND, Zidek JV, Burnett RT (1995). Multivariate imputation in cross sectional analysis of health effects associated with air pollution (with discussion). *J Environ Ecol Statist*, 2, 191–212.

Eilers JM, Kanciruk P, McCord RA, Overton WS, Hook L, Blick DJ, Brakke DF, Lellar PE, DeHan MS, Silverstein ME, Landers DH (1987). Characteristics of lakes in the western United States. *Data Compendium for Selected Physical and Chemical Variables, Vol 2.* Washington, DC: Environmental Protection Agency. EP/600/3-86-054b.

Elfving G (1952). Optimum allocation in linear regression theory. *Ann Math Statist*, 23, 255–262.

Embrechts P, Klüppelberg C, Mikosch T (1997). *Modelling Extremal Events for Insurance and Finance.* Berlin: Springer-Verlag.

Escoufier Y, Camps R, Gonzalez PL (1984). Bilan et Perspectives dans l'Approache Statistique de la Constitution d'un Reseau d'Alerte a la Pollution Atmospherique. *Matapli 3, Montpellier Cedex, Unite de Biometrie.*

Federov VV (1996). Design for spatial experiments: Model fitting and prediction. In *Handbook of Statistics, Vol 13*, (Eds) S Ghosh and CR Rao. Amsterdam: Elsevier, 515–553.

Federov VV, Hackl P (1997). *Model-Oriented Design of Experiments, Volume 125* of *Lecture Notes in Statistics.* New York: Springer-Verlag.

Federov VV, Müller W (1988). Two approaches in optimization of observing networks. In *Optimal Design and Analysis of Experiments*, (Eds) Y Dodge, VV Fedorov, and HP Wynn. New York: North Holland.

Federov VV, Müller W (1989). Comparison of two approaches in the optimal design of an observation network. *Statist*, 3, 339–351.

Fisher RA, Tippett LHC (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc Cambridge Phil Soc*, 24, 180–190. Reproduced in Fisher (1950, Paper 15).

Frey HC, Rhodes DS (1996). Characterizing, simulating, and analyzing variability and uncertainty: An illustration of methods using an air toxics emissions example. *Human and Ecol Risk Assess*, 2, 762–797.

Fu AQ (2002). Case study: Inference for extreme spatial random rainfall fields. MSc Thesis. Department of Statistics, U. of British Columbia.

Fu AQ, Le ND, Zidek, JV (2003). A statistical characterization of a simulated Canadian annual maximum rainfall field. TR 209-2003, Department of Statistics, U. British Columbia.

Fuentes M (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, 12, 469–484.

Fuentes M, Guttorp P, Challenor P (2003). Statistical assessment of numerical models. TR 76, National Research Center for Statistics and the Environment, U. of Washington.

Fuentes M, Raftery AE (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with output from numerical models. *Biometrics*, 66, 36–45.

Fuller WA (1987). *Measurement Error Models.* New York: Wiley.

Fung KY, Krewski D (1999). On measurement error adjustment methods in Poisson regression. *Environmetrics*, 10, 213–224.

Funtowicz S, Ravetz J (undated). Post-normal science—Environmental policy under conditions of complexity. http://www.nusap.net.

Fu AQ (2002). Case study: inference for extreme spatial random rainfall fields. MSc Thesis, Department of Statistics, U. of British Columbia.

Fu AQ, Le ND, Zidek JV (2003). A statistical characterization of a simulated Canadian annual maximum fainfall field. TR 209-2003, Department of Statistics, U. of British Columbia.

Gamerman D (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. London: Chapman and Hall.

Gaudard M, Karson M, Linder E, Sinha D (1999). Bayesian spatial prediction (with discussion). *J Environ Ecol Statist*, 6, 147–171.

Gelfand A, Zhu L, Carlin BP (2000). On the change of support problem for spatio-temporal data. Research Report 2000-011. Division of Biostatistics, U of Minnesota.

Gelfand AE, Schmidt AM, Banerjee, S Sirmans CF (2004). Nonstationary multivariate process modeling through spatially varying coregionalizations. *Test*, 13, 263–312.

Geweke J (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 24, 1317–1339.

Gilks WR, Richardson S, Spiegelhalter DJ (1995). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Green RH (1979). *Sampling Design and Statistical Methods for Environmental Biologists*. New York: Wiley.

Greenland S (1982). The effect of misclassification in matched-pair case-control studies. *Am J Epidemiol*, 116, 402–406.

Gribik P, Kortanek K, Sweigart I (1976). Designing a regional air pollution monitoring network: An appraisal of a regression experimental design approach. In *Proc Conference on Environmental Modelling and Simulation*, 86–91.

Gumbel EJ (1958). *Statistics of Extremes.* New York: Columbia University Press.

Gustafson P (2004). *Measurement Error, and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* London: Chapman and Hall/CRC.

Guttorp P, Le ND, Sampson PD, Zidek JV (1993). Using entropy in the redesign of an environmental monitoring network. In *Multivariate Environmental Statistics*, (Eds) GP Patil, CR Rao and NP Ross. New York: North Holland/Elsevier Science, 175–202.

Guttorp P, Meiring W, Sampson PD (1994). A space–time analysis of ground-level ozone data. *Environmetrics,* 5, 241–254.

Guttorp P, Sampson PD (1994). Methods for estimate heterogeneous spatial covariance functions with environmental applications. In *Environmental Statistics, Handbook of Statistics, 12*, (Eds) GP Patil and CR Rao. Amsterdam: North Holland.

Guttorp P, Sampson PD, Newman K (1992). Nonparametric estimation of spatial covariance with application to monitoring network evaluation. In *Statistics in Environmental and Earth Sciences*, (Eds) A Walden and P Guttorp. London: Edward Arnold, 39–51.

Haas TC (1990). Lognormal and moving window methods of estimating acid deposition. *J Amer stat Assoc*, 85, 950–963

Haas TC (1992). Redesigning continental-scale monitoring networks. *Atmospheric Environ*, 26A, 3323–3333.

Haas TC (1995). Local prediction of spatio–temporal process with an application to set sulfate deposition. *J Amer Statist Assoc*, 90, 1189–1199.

Haas TC (1996). Multivariate spatial prediction in the presence of nonlinear trends and covariance non-stationarity. *Environmetrics*, 7, 145–165.

Haitovsky Y, Zidek JV (1986). Approximating hierarchical normal priors using a vague component. *J Mult Anal*, 19, 48–66.

Hammersley JM, Clifford P (1971). Markov fields on finite graphs and lattices. Unpublished.

Handcock MS, Stein ML (1993). A Bayesian analysis of kriging. *Technometrics*, 35, 403–410.

Harris B (1982). Entropy. In *Encyclopedia of Statistical Science, Vol 2*, (Eds) S Kotz and NL Johnson. New York: Wiley, 512–516.

Hefferman JE, Tawn JA (2004). A conditional approach for multivariate extreme values (with discussion). *J Roy stat Soc, Ser B*, 66, 497–546.

Helton JC (1997). Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. *J stat Comput Simulation*, 57, 3–76.

Higdon D (1998). A process convolution approach to modelling temperatures in the north Atlantic ocean. *J Environ Ecol Statist*, 5, 173–190.

Higdon D, Swall J, Kern J (1999). Non-stationary spatial modelling. *Bayesian Statistics 6*, (Eds) JM Bernado et al.. Oxford: Oxford University Press, 761–768.

Hjort NL, Omre H (1994). *Topics in Spatial Statistics*. Cambridge: Blackwell.

Holland DM, Caragea PC, Smith RL (2003). Trends in rural sulfur concentrations. Unpublished.

Houghton JP, Segar SA, Zeh JE (1984). *Beaufort sea monitoring program: proceedings of a workshop (September 1983) and Sampling design recommendations*. Unpublished. Prepared for the U.S.National Oceanic and Atmospheric Administration, Department of Commerce, and Department of the Interior.

Howarth RJ, Earle SAM (1979). Application of a generalized power transformation to geochemical data. *J Int Assoc Math Geol*, 11, 45–62.

Huerta G, Sans B, Stroud JR (2004). A spatio–temporal model for Mexico City ozone levels. *J Roy stat Soc, Ser C (Appl Statist*, 53, 231–248.

Hughes JP, Lettenmeier DP (1981). Data requirements for kriging: Estimation and network design. *Water Resources Res*, 17, 1641–1650.

Jaynes ET (1963). Information theory and statistical mechanics, *Statistical Physics*, 3, (Ed) KW Ford. New York: Benjamin, 102–218.

Joe H (1994). Multivariate extreme-value distributions with applications to environmental data. *Can J Statis*, 22, 47–64.

Journel AG (1983). Simple tools applied to difficult problems. In *Statistics and Appraisal*, (Ed) HA David. Iowa State University Press, 237–255.

Journel AG (1988). New distance measures: The route towards truly non-Gaussian geostatistics. *Math Geol*, 20, 459–475.

Journel AG, Huijbregts CJ (1978). *Mining geostatistics.* London: Academic.

Jowett GH (1952). The accuracy of systematic sampling from conveyer belts. *Appl Statist*, 1, 50–59.

Kahneman D, Slovick P, Tversky A (eds.) (1982). *Judgement Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kaiser MS, Cressie N (2000). The construction of multivariate distributions from the Markov Random fields. *J Mult Anal*, 73, 199–200.

Kaiser MS, Cressie N, Lee J (2002). Spatial mixture models based on exponential family conditional distributions. *Statistica Sinica*, 12, 449–474.

Katz RW, Parlange MP, Naveau P (2002). Statistics of extremes in hydrology. *Adv Water Resources*, 25, 1287–1304.

Keifer J (1959). Optimum experimental design. *J Roy Stat Soc, Ser B*, 21, 272–319.

Kennedy WJ, Gentle JE (1980). *Statistical Computing*. New York: Marcel Dekker.

Kharin VV, Zwiers FW (2000). Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere-ocean GCM. *J Climate*, 13, 3760–3788.

Kibria GBM, Sun L, Zidek V, Le ND. (2002). Bayesian spatial prediction of random space–time fields with application to mapping PM2.5 exposure. *J Amer Statist Assoc*, 457, 101–112.

Kitanidis PK (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Res*, 22, 449–507.

Kitanidis PK (1997). *Introduction to Geostatistics: Applications to Hydrology*. Cambridge: Cambridge University Press.

Ko CW, Lee J, Queyranne M (1995). An exact algorithm for maximum entropy sampling. *Oper Res*, 43, 684–691.

Kolmogorov AN (1941). Interpolation and extrapolation of stationary random sequences. *Izvestiia Akedemii Nauk SSSR, Seriia Matematicheskiia*, 5, 3–14. [Translation 1962, Rand Corp, Santa Monica, CA].

Krige DG (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *J Chemical, Metallurgical Mining Soc S Africa*, 52, 119–139.

Laird NM, Ware JH (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Le ND, Sun L, Zidek JV (1998). A note on the existence of maximum likelihood estimates for Gaussian-inverted Wishart models. *Statist Prob Lett*, 40, 133–137.

Le ND, Sun L, Zidek, JV (2001). Spatial prediction and temporal backcasting for environmental fields having monotone data patterns. *Can J Statist*, 29, 516–529.

Le ND, Sun L, Zidek JV (2005). Designing networks for monitoring multivariate environmental fields using data with a monotone pattern. Unpublished.

Le ND, Sun W, Zidek JV (1997). Bayesian multivariate spatial interpolation with data missing by design. *J Roy stat Soc, Ser B*, 59, 501–510.

Le ND, White R, Zidek JV (1999). Using spatial data in assessing the association between air pollution episodes and respiratory health. *Statistics for the Environment 4: Statistical Aspects of Health and the Environment*, (Eds) V Barnett, A Stein, and KF Turkman, Chichester:Wiley, 117–136.

Le ND, Zidek JV (1992). Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *J Mult Anal*, 43, 351–374.

Le ND, Zidek JV (1994). Network designs for monitoring multivariate random spatial fields. In *Recent advances in statistics and probabilities: Proceedings of the 4th International Meeting of Statistics in the Basque Country*, (Eds) JP Vilaplena and ML Duri, Leiden: VSP International Science Publishers, 191–206.

Leadbetter MR, Lindgren G, Rootzén H(1983). *Extremes and Related Properties of Random Sequences and Series*. New York: Springer.

Lee J (1998). Constrained maximum-entropy sampling. *Oper Res*, 46, 655–664.

Lee J, Kaiser MS, Cressie N (2001). Multiway dependence in exponential family conditional distributions. *J Mult Anal*, 73, 199–200.

Ledford AW, Tawn JA (1996). Statistics for near independence of multivariate extremes. *Biometrika*, 83, 169–187.

Ledford AW, Tawn JA (1997). Modelling dependence with joint tail regions. *J Roy stat Soc, Ser B*, 59, 475–499.

Ledford AW, Tawn JA (2003). Diagnostics for dependence within time-series extremes. *J Roy stat Soc, Ser A*, 65, 521–543.

Li KH, Le ND, Sun L, Zidek JV (1999). Spatial–temporal models for ambient hourly PM10 in Vancouver. *Environmetrics*, 10, 321–338.

Liang KY, Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

Lindley DV (1956). On the measure of the information provided by an experiment. *Ann Math Statist*, 27, 968–1005.

Lindley DV (2002). Letter to the editor. *Teaching Statistics*, 24, 22.

Lindstrom MJ, Bates DM (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46, 673–687.

Linthurst RA, Landers DH, Eilers JM, Brakke DF, Overton WS, Meier EP, Crowe RE (1986). *Characteristics of Lakes in the Eastern United States. Vol-*

*ume 1. Population Descriptions and pHysico-Chemical Relationships.* EPA/600/4-86/007a, U.S. Environmental Protection Agency, Washington, DC.

Lippman M (1993). Health effects of tropospheric ozone: Review of recent research findings and their implications to ambient air quality standards. *J Expos Anal Environ Epidemiol*, 3, 103–129.

Little RJA, Rubin, DB (1987). *Statistical Analysis with Missing Data.* New York: John Wiley and Sons.

MacClure M (1991). The case-crossover design: A method for studying transient effects on risk of acute events. *Am J Epidemiology*, 133, 144–153.

Mardia KV, Goodall CR (1993). Spatial–temporal analysis of multivariate environmental monitoring data. In *Mult Environ Statist*, (Eds) GP Patil and CR Rao, Amsterdam: Elsevier Science Publisher BV, 347–386.

Mardia KV, Goodall CR, Redfern EJ, Alonso FJ (1998). The kriged Kalman filter (with discussion). *Test*, 7, 217–285.

Mardia KV, Kent JT, Bibby J (1979). *Multivariate Analysis.* London: Academic.

Marshall AW, Olkin I (1988). Families of multivariate distributions. *J Amer Statist Assoc*, 83, 834–841.

Matern B (1960). *Spatial Variation.* Berlin: Springer-Verlag. Reprint.

Matheron G (1955). Application des methodes statistiques l'evaluation des gisements. *Annales des Mines*, 144, 50–75.

Matheron G (1962). *Traite de geostatistique applique, Vol 1.* Paris: Technip.

Matheron G (1963). Principles of geostatistics. *Econ Geol*, 58, 1246–1266.

Matheron G (1965) *Les variables regionalises et leur estimation.* Paris: Masson.

Matheron G (1969). *Le krigeage universal: recherche d'operateurs optimaux en presence d'une derive.* Fontainbleau: Fascicule 1, Les Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris.

Matheron G (1970). *The theory of regionalized variables and its applications.* Fontainbleau: Fascicule 5, Les Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris.

Matheron G (1971). *La theorie des fonctions aleatoires intrinsiques generalises.* Fontainbleau: Publication N-252, Centre de Geostatistique, Ecole des Mines de Paris.

Matheron G (1972). *Leons sur les Fonctions Alatoires d'Ordre 2.* Ecole des Mines de Paris.

Matheron G (1976). A simple substitute for conditional expectation: the disjunctive kriging. In *Advanced Geostatistics in the Mining Industry, NATO ASI Series C24*, (Eds) M Guarascio, M David and C Huijbeqts, Dordrecht: Reidel, 221–236.

Matheron G (1984). The selectivity of distributions and "the second principle of geostatistics". In *Geostatistics for Natural Resources Characterization, NATO ASI Series C-122*, (Ed) G Verly, Dordrecht: Reidel, 421–433.

Meiring W, Guttorp P, Sampson PD (1998). Space–time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environ Ecol Statist*, 5, 197–222.

Mitchell TJ (1974a). An algorithm for the construction of "D-Optimal" experimental designs. *Technometrics*, 16, 203–210.

Mitchell TJ (1974b). Computer construction of "D-Optimal" first-order designs. *Technometrics*, 16, 211–221.

Monestiez P, Sampson PD, Guttorp P (1993). Modelling of heterogeneous spatial correlation structure by spatial deformation. *Cahiers de Geostatistique*, Fascicule 3, Compte Rendu des Journees de Geostatistique, Fontainebleau. Published by the Ecole Nationale Superieure des Mines de Paris

Müller WG (2001). *Collecting Spatial Data: Optimum Design of Experiments for Random Fields, Second edition.* Hiedelberg: Physica-Verlag.

Myers DE (2002). Space–time correlation models and contaminant plumes. *Environmetrics*, 13, 535–553.

Navidi W (1998). Bidirectional case-crossover designs for exposures with time trends. *Biometrics,* 54, 596–605.

Nott, D and Dunsmuir, WTM (1998). Analysis of spatial covariance structure for Sydney wind patterns. Report S98-6, School of Mathematics, The University of New South Wales.

O'Hagan A (1988) *Probability: Methods and Measurement.* London: Chapman and Hall.

O'Hagan A (1998) Eliciting expert beliefs in substantial practical applications (with discussion). *The Statistician*, 47, 21–35 and 55–68)

Omre H (1987). Bayesian Kriging—merging observations and qualified guess in Kriging. *Math Geol*, 19, 25–39.

Omre H, Halvorsen. K (1989). The Bayesian bridge between simple and universal Kriging. *Math Geol*, 21, 767–786.

Oreskes N, Shrader-Frechete K, Belitz K (1994). Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, 263, 641–646.

Ott WR (1995). *Environmental Statistics and Data Analysis.* Boca Raton, FL: CRC.

Perrichi LR (1981). A Bayesian approach to transformations to normality. *Biometrika*, 68, 35–43.

Pickands J (1975). Statistical inference using extreme order statistics. *Ann Statist*, 3, 119–131.

Pilz J (1991). *Bayesian Estimation and Experimental Design in Linear Regression Models.* New York: Wiley.

Pope CA, Dockery DW (1992). Acute health effects of $PM_{10}$ pollution on symptomatic and asymptomatic children. *Am Rev Respir Dis*, 145, 1123–1128.

Pope CA, Dockery DW, Spengler JD, Raizenne ME (1991). Respiratory health and $PM_{10}$ pollution: a daily time-series analysis. *Am Rev Respir Disease,* 144, 668–674.

Reiss RD, Thomas M (1997). *Statistical Analysis of Extreme Values With Applications to Insurance, Finance, Hydrology and Other Fields.* Cambridge, MA: Birkhäuser, 167–173.

Rényi A (1961). On measures of entropy and information. In *Proceedings Fourth Berkeley Symposium*, (Ed) J Neyman, 547–561.

Richardson S, Best N (2003). Bayesian hierarchical models in ecological studies of health–environment effects. *Environmetrics*, 14, 129–147.

Ro CU, Tang AJS, Chan WH (1988). Wet and dry deposition of sulphur and nitrogen compounds in Ontario. *Atmos Environ*, 22, 2763–2771.

Sampson P, Guttorp P (1992). Nonparametric estimation of nonstationary spatial structure. *J Amer stat Assoc*, 87, 108–119.

Savage LJ (1971 some refers to the 1954 version in text). *Foundations of Statistics.* New York: Dover.

Schmidt AM, Gelfand AE (2003). A Bayesian coregionalization approach for multivariate pollutant data, *J Geophys Res-Atmospheres*, 108, no. D24.

Science Daily (2000). Mother nature cleans up human-made mess: algae proved to be effective bio-monitors in previously contaminated polar lake. http://www.sciencedaily.com/releases/2000/11/001122183430.htm.

Schoenberg IJ (1938). Metric spaces and completely monotone functions. *Ann Math*, 39, 811–841.

Schumacher P, Zidek JV (1993). Using prior information in designing intervention detection experiments. *Ann Statist*, 21, 447–463.

Schwartz J, Dockery DW (1992). Increased mortality in Philadelphia associated with daily air pollution concentrations. *Am Rev Respir Dis*, 145, 600–604.

Sebastiani P, Wynn HP (2000). Maximum entropy sampling and optimal Bayesian experimental design. *J Roy stat Soc, Ser B*, 62, 145–157.

Sen PK (2003). Structure - activity relationship information incorporation in health related environmental risk assessment. *Environmetrics*, 14, 223–234.

Shannon CE (1948). A mathematical theory of communication. *Bell Syst Tech J*, 27, 379–423, 623–656.

Shewry M, Wynn H (1987). Maximum entropy sampling. *Appl Stat*, 14, 165–207.

Shorack GR, Wellner J (1986). *Empirical Processes with Applications in Statistics.* New York: Wiley.

Silvey SD (1980). *Optimal Design.* London: Chapman and Hall.

Singh VP (1998). *Entropy-Based Parameter Estimation in Hydrology.* Dordrecht: Kluwer.

Sirois A, Fricke W (1992). Regionally representative daily air concentrations of acid-related substances in Canada, 1983-1987. *Atmos Environ*, 26A, 593–604.

Smith K (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12, 1–85.

Smith RL (1996). Estimating nonstationary spatial correlations. Unpublished preprint.

Smith RL (2001). *Environmental Statistics.* Notes at the Conference Board of the Mathematical Sciences (CBMS) course at the University of Washington, June 25–29, 2001. http://www.stat.unc.edu/postscript/rs/envnotes.ps

Smith RL (2004). Contribution to the discussion of Hefferman and Tawn (2004).

Stein ML (1998). The kriged Kalman filter. Discussant. *Test*, 7, 217–285.

Stein ML (1999). *Interpolation of Spatial Data—Some Theory for Kriging.* New York: Springer.

Stewart I (1989). *Does God Play Dice: the Mathematics of Chaos.* Oxford: Oxford Basil.

Sun L (2001). Matric-*t* distribution. *Encyclopedia of Environmetrics.* New York: Wiley.

Sun W (1994) Bayesian multivariate interpolation with missing data and its applications. Ph.D. Dissertation. Dept of Statist, U British Columbia.

Sun W (1998). Comparison of a co-kriging method with a Bayesian alternative. *Environmetrics*, 9, 445–457.

Sun W, Le ND, Zidek JV, Burnett, R (1998). Assessment of Bayesian multivariate interpolation approach for health impact studies. *Environmetrics*, 9, 565–586.

Sun L, Zidek JV, Le ND, Ozkaynak H (2000). Interpolating Vancouver's daily ambient PM10 field. *Environmetrics*, 11, 651–663.

Tanner MA (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd Edition.* New York: Springer-Verlag.

Theil H, Fiebig DG (1984). Exploiting continuity: Maximum entropy estimation of continuous distributions. In *Series on Economics and Management Science, Vol 1*, (Eds) WW Cooper and H Theil. Cambridge: Ballinger.

Thomas D, Stram D, Dwyer J (1993). Exposure measurement error: Influence of exposure-disease relationships and methods of correction. *Ann Rev Public Health*, 14, 69–93.

van Eeden C, Zidek JV (2003). Uncertainty, entropy and partial information. *Mathematical Statistics and Applications: Festschrift for Constance van Eeden*, (Eds) M Moore, S Froda, and C Leger. Hayward: IMS Lecture Notes and Monograph Series, Volume 42, 155–167.

Verdinelli I (1991). Advances in Bayesian experimental design. In *Bayesian Statistics 4*, (Eds) J Bernardo, J Berger, AP Dawid, A Smith. Oxford: Clarendon, 1–17.

Verly G (1983). The multigaussian approach and its applications to the estimation of local reserves. *J Int Assoc Math Geol*, 15, 259–286.

Wabba G, Wendelberger J (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Rev*, 108, 36–57.

Wackernagel H (2001) Multivariate geostatistics. In *Encyclopedia of Environmetrics, Vol 3*, (Eds) A El-Shaarawi and W Piegorsch. Chichester: Wiley, 1344–1347.

Wackernagel H (2003). *Multivariate Geostatistics*. Berlin: Springer.

Waller LA, Louis TA, Carlin BP (1997). Bayes methods for combining disease and exposure data in assessing environmental justice. *J Environ Ecol Statist*, 4, 267–281.

West M, Harrison J (1999). *Bayesian Forecasting and Dynamic Models*. New York: Springer.

Whittle P (1954). On stationary processes in the plane. *Biometrika*, 41, 434–449.

Whittle P (1962). Topographic correlation, power-law covariance functions, and diffusion. *Biometrika*, 49, 305–314.

Wikle CW, Cressie N (1999). A dimension-reduced approach to space–time Kalman filtering. *Biometrika*, 86, 815–829.

Wikle LM, Milliff RF, Nychka D, Berliner LM (2001). Spatiotemporal hierarchical Bayesian modeling: tropical ocean surface winds. J Am Stat Assoc, 96, 382–397.

Wu CFJ (1982). On the convergence properties of the EM algorithm. *Ann Stat*, 11, 95–103

Wu S, Zidek JV (1992). An entropy based review of selected NADP/NTN network sites for 1983–86. *Atmos Environ*, 26A, 2089–2103.

Yaglom AM (1957). Some classes of random fields in $n$-dimensional space, related to stationary random processes. *Theor Prob Appl*, 2, 273–320.

Yaglom AM (1986). *Correlation Theory of Stationary and Related Random Functions.* Berlin: Springer-Verlag.

Yakowitz SJ, Szidarovszky F (1985). A comparison of kriging with nonparametric regression methods. *J Mult Anal*, 16, 21–53.

Zeger SL, Liang KY, Albert PS (1988). Models for longitudinal data; A generalised estimating equation approach. *Biometrics*, 44, 1049–1060.

Zhu, Z (2002). Optimal Sampling Design and Parameter Estimation of Gaussian Random Fields. PhD Thesis, Department of Statistics, University of Chicago.

Zhu Z, Stein ML (2005). Spatial Sampling Design for Parameter Estimation of the Covariance Function. *J stat Planning and Inference*, 134, 583–603.

Zhu Z, Stein ML (2006). Two-step Spatial Sampling Design for Prediction with Estimated Parameters. *J Agric Biol Environ Stat*. In press.

Zidek JV (1997). Interpolating air pollution for health impact assessment. *Statistics for the Environment 3: Pollution Assessment and Control*, (Eds) V Barnett and KF Turkman). New York: Wiley, 251–268.

Zidek JV, Le ND, Wong H, Burnett RT (1998a). Including structural measurement errors in the nonlinear regression analysis of clustered data. *Can J Stat*, 26, 537–548.

Zidek JV, Meloche J, Shaddick G, Chatfield C, White R (2003). A computational model for estimating personal exposure to air pollutants with appli-

cation to London's $PM_{10}$ in 1997. TR 2003-3, Statistical and Applied Mathematical Studies Institute, Research Triangle Park.

Zidek JV, Navin FDP, Lockhart R (1979). Statistics of extremes: An alternate method with application to bridge design codes. *Technometrics*, 8, 185–191.

Zidek JV, Sun W, Le ND (2000). Designing and integrating composite networks for monitored multivariate Gaussian pollution fields. *J Roy Stat Soc, Ser C*, 49, 63–79.

Zidek JV, Sun L, Le ND, Özkaynak H (2002). Contending with space–time interaction in the spatial prediction of pollution: Vancouver's hourly ambient $PM_{10}$ field. *Environmetrics*, 13, 1–19.

Zidek JV, White R, Le ND (1998c). Using spatial data in assessing the association between air pollution episodes and respiratory morbidity. In *Statistics for the Environment 4: Statistical Aspects of Health and the Environment*, (Eds) V Barnett and KF Turkman. New York: Wiley, 117–136.

Zidek JV, White R, Le ND, Sun W, Burnett RT (1998b). Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. *J Environ Ecol Statist*, 5, 99–115.

Zidek JV, Wong H, Le ND, Burnett R (1996). Causality, measurement error and collinearity in epidemiology. *Environmetrics*, 7, 441–451.

Zimmerman DL (2004). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. TR 339. Dept of stat and Actuarial Sc, U. Iowa.

# Author Index

# Subject Index