ENVIRONMENTAL STATISTICS

© Richard L. Smith

Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260,
USA

Email address: rls@email.unc.edu

Web reference:
http://www.stat.unc.edu/postscript/rs/envnotes.ps

VERSION 5.0

9 JULY 2001

PREFACE

These notes have been prepared in connection with the Conference Board of the Mathematical Sciences (CBMS) course at the University of Washington, June 25–29, 2001, where the author is principal lecturer. It is intended that the completed version of the notes will be published within a year of the conference.

The aim of the course is to present an overview of the different kinds of statistical methodology being used in contemporary environmental applications of statistics. Since much environmental work involves spatial sampling, there is a heavy emphasis on spatial statistics. I also cover some modern developments in time series analysis which have proved valuable for studying environmental trend, e.g. long-range dependence. Also since many environmental problems are associated with extreme values, I give an overview of some of the specific techniques useful for dealing with these problems. The applications covered have a bias towards "physical" topics such climate change and atmospheric pollution. An overview of the first of these topics is given in Chapter 1, and used as motivation for many of the methodological developments described in detail in later chapters.

Prerequisites for the course are a sound background knowledge of statistics at the graduate level, including linear models, maximum likelihood, Bayesian statistics, etc. Some previous knowledge of time series analysis would be helpful but not essential. I do not assume any previous knowledge of spatial statistics or extreme value theory. Some sections of the notes assume more advanced mathematical knowledge but the reader who is willing to take these on trust can omit these sections without losing information that is vital to subsequent sections. I have identifed such sections by a double asterisk (**).

I would like to thank Peter Guttorp for organizing the CBMS course and for inviting me to be the principal lecturer. It is impossible to name all the other individuals with whom I have collaborated in environmental statistics in some way over the years, but I would like to mention in particular Peter Bloomfield, Stuart Coles, Larry Cox, Jerry Davis, Anthony Davison, Montserrat Fuentes, Amy Grady, Dave Holland, Doug Nychka, Peter Robinson, Jerry Sacks, Jonathan Tawn and Tom Wigley. Previous versions of the notes have been used in connection with courses at the University of North Carolina, Duke University, Cambridge University and the University of Mendoza, the latter being at the invitation of Angela Diblasi. Chapter 8 has expanded from some lecture notes on extreme values in meteorology which were first presented at a course organized by the American Meteorological Society in Reno, Nevada, in 1997. The group at the National Climatic Data Center (Tom Karl, Dave Easterling, and others) has been very helpful in providing access to, and information about, climatic data sets. My personal research has been sponsored by the National Science Foundation and the Environmental Protection Agency; the CBMS series is also sponsored by the NSF. I have also benefitted from extensive interactions with the National Institute of Statistical Sciences and the Geophysical Statistics Project at the National Center for Atmospheric Research. To all of these individuals and organizations, I offer my thanks.

TABLE OF CONTENTS

CHAPTER 1

Introduction: Statistical Problems Associated With Climate Change

The purpose of this introductory chapter is to introduce some of the statistical questions that arise in environmental science through the medium of one rather complex example — the statistical evaluation of hypotheses about climate change. This field of research involves many areas of statistical methodology, such as spatial statistics, estimating trends in correlated time series, and extreme value analysis, which are developed in more detail in later chapters. It also serves to provide some real data sets which we shall use later as practical examples of the application of some advanced statistical techniques.

After some brief background discussion in section 1.1, we discuss a concrete problem, involving trends in hemispheric temperature averages, in section 1.2. Sections 1.3–1.5 illustrate some of the issues involved in analyzing a large data base, in this case the Historical Climatological Network (HCN) which includes daily temperature and precipitation information for 186 stations across the continental USA. We discuss, for example, the need to assess spatial dependence, and in section 1.5 provide some examples of variogram estimation, a topic discussed in much more detail in chapter 2. Section 1.6 aims to put this discussion in context by providing a very brief overview of some of the other environmental statistics problems of considerable interest at the present time. Finally, section 1.7 serves to provide some more technical detail about one of the methodological issues, namely the estimation of trends in series with autoregressive errors.

1.1 Introduction

The problem of climate change is one of the most complex and controversial scientific problems being studied today. In broad terms, the temperature at the earth's surface has been rising steadily since the middle of the nineteenth century, and rising rapidly since the mid-1970s. These statements are supported by extensive observational evidence and are not seriously disputed. However, the causes of this warming trend are by no means universally agreed. By far the best known and most widely publicized explanation is that the rise in temperature is caused primarily by the "greenhouse effect" created by rising levels of certain gases in the atmosphere, in particular carbon dioxide (CO_2), which is attributed largely to anthropogenic causes such as the burning of fossil fuels. This is supported by the fact that CO_2 levels have themselves been steadily rising since the mid-1950s, and by numerical models of the atmosphere and oceans — the so-called general circulation models or GCMs — which have demonstrated a direct association between CO_2 and temperature. However, not everyone accepts that the greenhouse gas effect is

the primary cause of global warming. For example, it is pointed out that the earth went through a "mini ice-age" during the seventeenth and eighteenth centuries, and that some of the warming since that period is an entirely natural effect associated with the ending of that period. On a more specific level, comparisons of GCM output with observed temperatures have shown that greenhouse gases alone cannot explain all fluctuations in temperature; other effects such as the cooling effect due to aerosols (small particles, typically sulfates, such as those emitted during volcanic eruptions) and variations in solar flux must be taken into account. Nevertheless, when all these effects are considered, the component due to greenhouse gases remains strong. A comprehensive review of all these issues up to the end of 1995 was given by Houghton *et al.* (1996); some more recent references addressing the agreement between GCM output and observational data include Santer *et al.* (1996), Hegerl *et al.* (1996), Allen and Tett (1999) and Tett *et al.* (1999).

Statistics enters this discussion in numerous ways. The problem of how to assess the fit between observed data and GCM output is a very complex one. The observed data are spatially distributed across the earth's surface, and in some studies, vertically through the atmosphere as well. Even in the absence of forcing factors such as changes in greenhouse gases, both real data and GCM output show complex temporal dependences which are difficult to characterize physically. Thus even the simplest comparisons require that one take into account both temporal and spatial correlations, and ultimately require spatial-temporal models.

There are many other aspects of climate change that require statistics. GCMs are believed, at least by their supporters, to provide good predictions of mean temperatures and precipitation over large spatial and temporal regions, but they do not perform so well over smaller scales. *Downscaling* is a general term used by climatologists to describe the process of predicting small-scale effects as a function of those operating at large scales. This is again a highly statistical task and many of the techniques which have been developed in recent years require careful analysis of spatial and/or temporal correlations in observational data. Then there are other phenomena which have been observed in the data which require careful statistical evaluation. For example, Karl *et al.* (1996) have developed an number of *indices* of climate change based on such things as length or severity of droughts, numbers of days with very high levels of precipitation, days with either the maximum or the minimum temprature much above or below normal, and so on. One of their most interesting discoveries was that although the mean level of precipitation has not changed greatly, the proportion of days with a very high one-day precipitation level in the USA has increased sharply during the twentieth century. Another claim (Easterling *et al.* 1997) is that the increase in mean global temperatures is due primarily to a decrease in the mean diurnal temperature range, i.e. that daily minumum temperatures are getting warmer while daily maximum temperatures are remaining about the same. Statistical assessment of claims of this nature also requires attention to spatial and temporal correlations, but may introduce other features as well. For example, the claims about the increase in frequency of high-rainfall events are a natural domain for the application of *extreme value analysis*, which is specifically concerned with questions of this nature.

1.2 An Example

Fig. 1.1 [1] shows mean annual temperature anomalies [2] for 1900–1996, computed for the northern hemisphere (NH) and southern hemisphere (SH), together with fitted linear trend and the "best" model fit reflecting anthropogenic influences as well as solar fluctations. The latter is calculated using a method developed by Wigley and Raper (1990, 1991). It shows that the warming trend has not been monotonic; in particular, in the NH, temperatures were decreasing for a period between the 1940s and 1960s and only started rising again at the end of the 1960s; however, since then they have been rising faster than ever before. A similar effect is observed in the SH though it is concentrated into a shorter time-span and it is perhaps less clear that it represents a genuine change to the overall pattern.

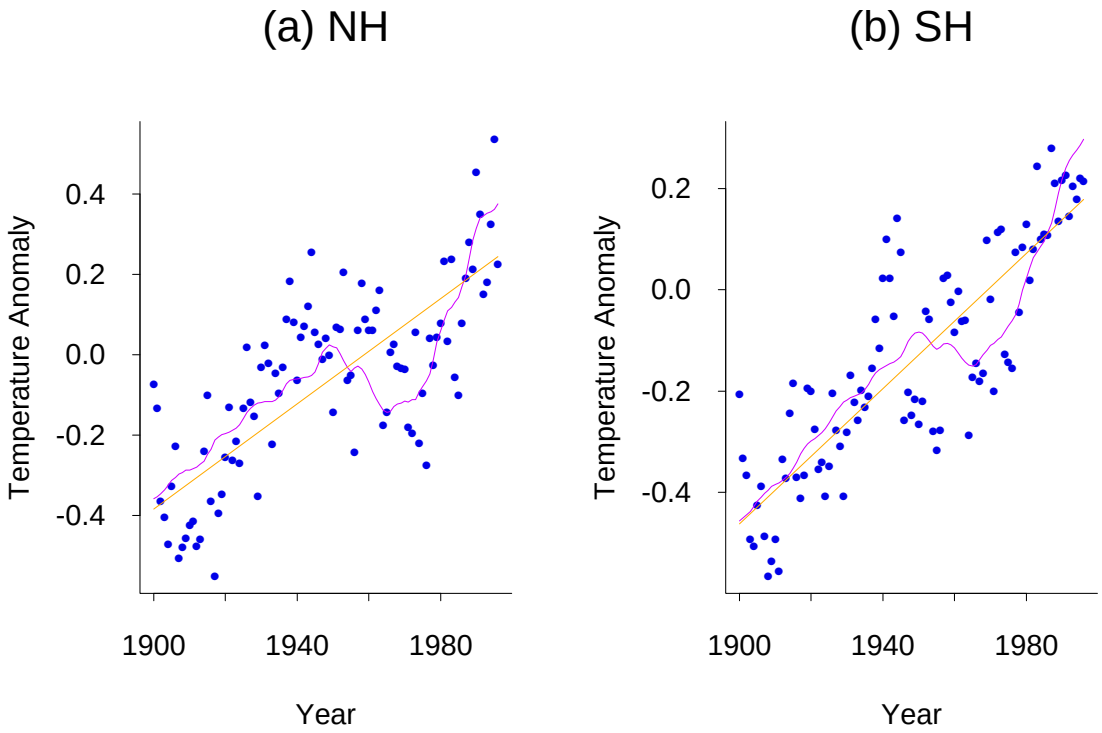


Fig. 1.1. Plot of temperature anomalies in Northern Hemisphere and Southern Hemisphere for 1900–1996, together with fitted straight lines and best fits to trends based on climate models.

The fitted trends that are plotted are based on a linear regression with autoregressive errors. Suppose y_t denotes the observed temperature anomaly in year t. Suppose this is

[1] Based on ongoing work with climate scientists Tom Wigley (National Center for Atmospheric Research) and Ben Santer (Lawrence Livermore National Laboratories). I am grateful to Tom Wigley for the data.

[2] Temperature differences compared with the average over a fixed time interval, approximately 1961–1990.

related to a regressor x_t through the equation

$$y_t = \beta_0 + \beta_1 x_t + \eta_t, \tag{1.1}$$

where the residuals $\{\eta_t\}$ form an autoregressive process of order m, AR(m) for short:

$$\eta_t = \sum_{j=1}^{m} \phi_j \eta_{t-j} + \epsilon_t, \tag{1.2}$$

and the $\{\epsilon_t\}$ are assumed independent $N(0, \sigma^2)$ random variables. Of course if $m = 0$ this is just a standard linear regression with independent errors, but the autoregressive term is introduced to reflect the reality that temperatures are correlated. There are a number of alternative approaches which might be used to assess the significance of trends in climatic time series, some references being Bloomfield (1992), Bloomfield and Nychka (1992) and Smith (1993). Models of the form of (1.1)–(1.2), or extensions in which the $\{\eta_t\}$ process is assumed to be of ARMA form, have also been widely used in the climatological literature, including the references Karl *et al.* (1996) and Easterling *et al.* (1997).

The model defined by (1.1) and (1.2) was fitted by maximum likelihood, for several values of m, and the different models compared using the AIC [3]. For x_t, two models were tried, both the simple linear trend $x_t = t$ and the trend predicted by the climate model, the latter being different for the NH and SH. The results are in the following table:

AR Order m	NH Linear Trend	NH Model Trend	SH Linear Trend	SH Model Trend
0	–259.0	–280.5	–313.7	–320.6
1	–290.4	–297.7	–343.9	–346.2
2	–288.6	–295.7	–343.4	–346.3
3	–286.7	–293.7	–346.2	–348.1
4	–295.9	–301.1	–344.2	–346.1
5	–294.2	–299.3	–342.8	–345.3
6	–292.2	–297.3	–340.9	–343.6

Table 1.1. AIC values for linear and climate model trends fitted to northern hemisphere and southern hemisphere temperature anomalies for 1900–1996, for several values of the autoregressive order m.

The sharp drop in AIC between $m = 0$ and $m = 1$, for all four columns of the table, shows clearly that a model based on independent errors is not adequate. The best model

[3] Akaike Information Criterion. Defined as $-2 \log L + 2 \log p$ where L is the maximized likelihood and p the number of parameters. Widely used as a model selection criterion, especially when comparing large numbers of models.

as selected by AIC has $m = 4$ for the NH and $m = 3$ for the SH, for both the linear and climate model trends. In both cases the trend is highly significant compared with a null hypothesis of no trend — for example, with the linear trend, the slope is .66 ^{o}C per century for the NH, with a standard error of .17, and .67 for the SH, standard error .09. However the climate model trend is clearly superior, especially for the NH. Another comparison is in terms of R^2, i.e. the ratio of residual sums of squares for the fitted statistical model compared with a null model of no trend. In the case of a linear trend this is .52 for the NH, .70 for the SH. Using the climate model trend these figures rise to .62 for the NH and .72 for the SH. That the increase in R^2 is less dramatic for the SH than the NH presumably reflects the fact that both the observed trend and the model-based prediction are closer to linear in the case of the SH than the NH — it does not imply that the climate model fits less well in the SH. On the other hand, it is also apparent from the figure that the model-based trend does not capture all the observed fluctuations in the data — for example, both the NH and SH data show a sharp rise in the 1940s, followed, in the case of the SH data, by a sharp fall in the 1950s, and this is not fully reflected by the model predictions.

A brief description of the procedure used for fitting the model (1.1)–(1.2) is contained in section 1.7.

1.3 Temperature and Rainfall Trends Across the USA

Analyses such as those in section 1.2, based on global or hemispherical data sets, are valuable as broad-brush assessments of global climate change, but they provide very little information about specific meteorological effects. For that, one must study effects on much more localized spatial scales. In the USA, researchers at NCDC [4] have compiled daily records of maximum and minimum temperatures and precipitations for a network of stations known as the Historical Climatological Network (HCN). The analysis that follows is based on 186 stations from the HCN, depicted in Fig. 1.2. For the purpose of some of the plots, the stations have been arranged in a very rough grid, as follows. The ten northernmost stations were arranged in west-to-east order; this forms the top row of the grid. The next ten northernmost stations were arranged in west-to-east order and form the second row of the grid. This continues down to the bottom row, which consists of just six stations. The solid lines of Fig. 1.2 join up the stations in each row. Finally, the whole country is divided into four large regions (dotted lines), which will be used for regional analysis.

[4] National Climatic Data Center, based in Asheville, NC; www.ncdc.noaa.gov.

Fig. 1.2. Map of 186 HCN stations in the continental USA, joined up to indicate grid rows. The dotted vertical and horizontal lines divide the country into four regions for subsequent regional analysis.

For the purpose of this discussion, four time series representing specific meteorological effects were created for each station. The first consists of winter means of daily minimum temperature. Here the winter is defined to mean the months of December, January and February, with December counted as part of the following year's winter. The focus on daily minima in winter was suggested by papers such as Easterling *et al.* (1997), who have suggested that most of the observed global warming is due to an increase in the lowest temperatures. For direct contrast with that, a set of time series consisting of winter means of daily *maximum* temperature was also created. In both of these series, it was arbitrarily decided to count a particular winter's data as being "available" if at least 60 individual daily values were available at that station for the winter in question. Otherwise the winter mean value was recorded as a missing value. The third time series consisted of the annual maximum daily rainfall value for each station and year. This was counted as "available" if at least 240 days' data from the year in question were available. The focus on annual maxima was suggested by the finding of Karl *et al.* (1996) that the frequency of extreme rainfalls has increased over the century. However, for comparison with that, a fourth time series consisting of annual *mean* precipitation at each station was also computed. One feature of the annual maxima is that they clearly do not follow a normal distribution (Fig. 1.7 below) and this poses some interesting problems concerning the handling of time series with highly skewed distributions. The original data in fact contained a very small number of obviously wrong values (daily rainfall of up to 90 inches); unfortunately, there was no clear-cut criterion for deciding exactly which values were in error. Arbitrarily, it was decided to truncate all values in excess of 10 inches so that they are treated as being

exactly 10 inches. Even with such truncation, however, the distribution remains highly skewed.

In Figs. 1.3–1.6, the 186 time series for mean winter daily minima, mean winter daily maxima, maximum annual daily precipitation and mean annual precipitation are plotted, using the grid system described in connection with Fig. 1.2. A few general patterns are observable from these plots. For example, many of the temperature minima series show a steady or decreasing trend over most of the century followed by a clear rise since around 1970, consistent with the pattern anticipated from the NH series in Fig. 1.1. However, the pattern is hard to see clearly because there is a great deal more noise in the individual plots in Fig. 1.3 than there is in Fig. 1.1, as is inevitable given that Fig. 1.3 is based on single station values whereas Fig. 1.1 represents hemispheric averages. On the other hand, it is harder to see any general trend either increasing or decreasing among the mean winter daily maxima, Fig. 1.4. As far as the rainfall series is concerned, Fig. 1.5, it is again hard to see any evidence of overall trend but this may reflect the effect of outlying observations masking whatever trend may exist.

As an example of the appearance of individual series, Fig. 1.7 shows the four plots on a larger scale for one particular station — Portland, Maine, the extreme right-hand plot of the fifth row in Figs. 1.3–1.6. Also shown on this plot are Q-Q plots based on the normal distribution (Chambers *et al* 1983). In computing these plots, attention had to be paid to certain suspect values in the rainfall series — in particular, the recorded daily rainfall total for January 9, 1945 was 50 inches! This seems likely to be spurious and is therefore treated as a censored value for the computation of the Q-Q plot. However the second largest value in the series (9.62 inches on October 21 1996) is treated as genuine. [5]

The Q-Q plots for the minimum and maximum temperatures and for the mean precipitation values follow a straight line quite closely, indicating a good fit to the normal distribution, the only slight doubt being at the lower tail of the distribution for temperature minima. Of course, since these values are averages, one would expect a good fit by the Central Limit Theorem — the plots should not be taken as implying that individual daily values are normally distributed. On the other hand, it is clear from Fig. 1.6(f) that the rainfall maxima data have nothing like a normal distribution.

[5] The most recent version of the data (1999) has corrected the January 9 1945 value to 0.08 inches, but the value for October 21 1996 is still listed as 9.62 inches, so this is presumably correct. The discussion at this point has not been edited, because it shows the need to be vigilant for possible errors even in high-quality data sources.

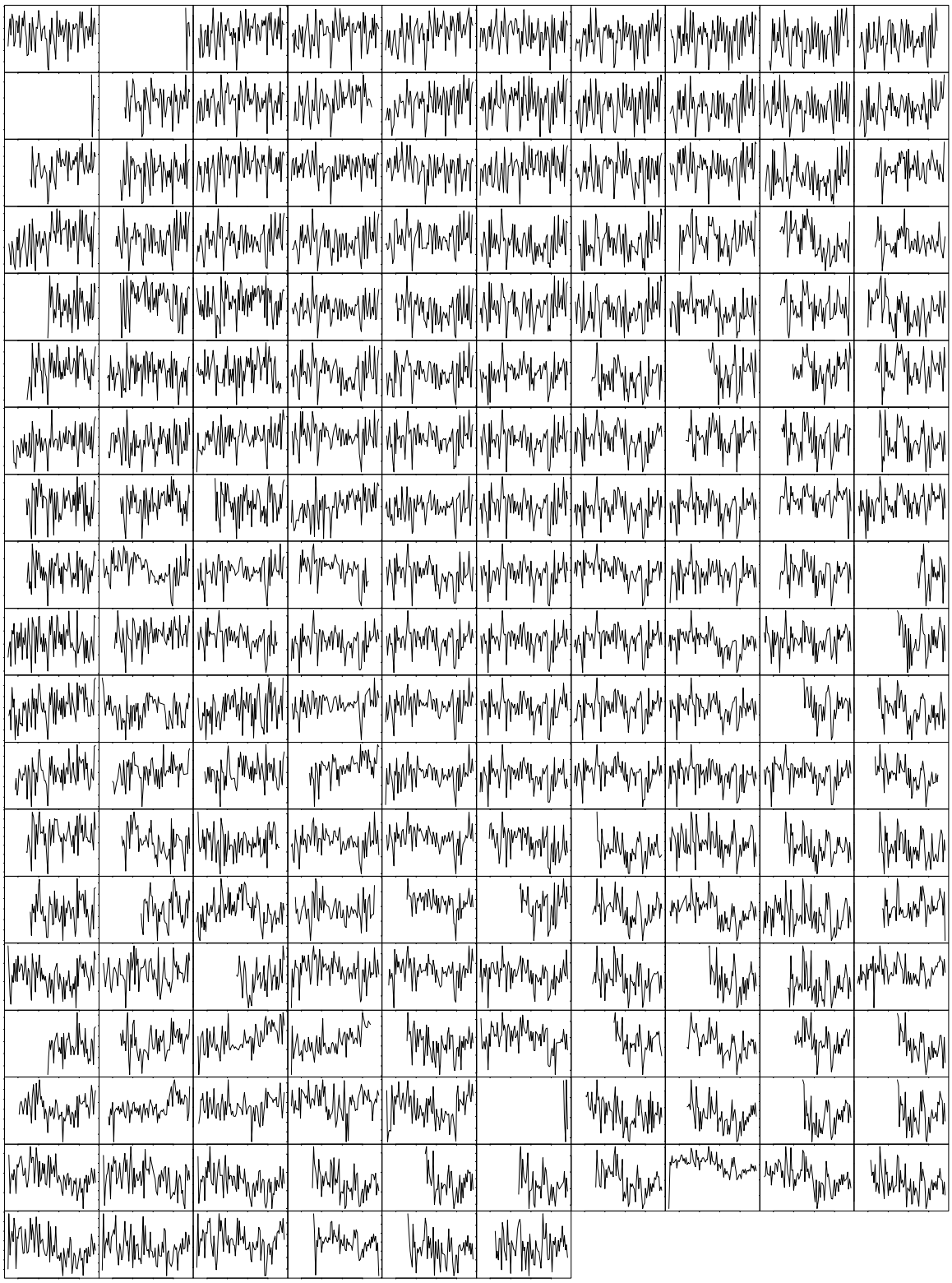


Fig. 1.3. Time series plots of 186 time series, winter means of daily minimum temperature.

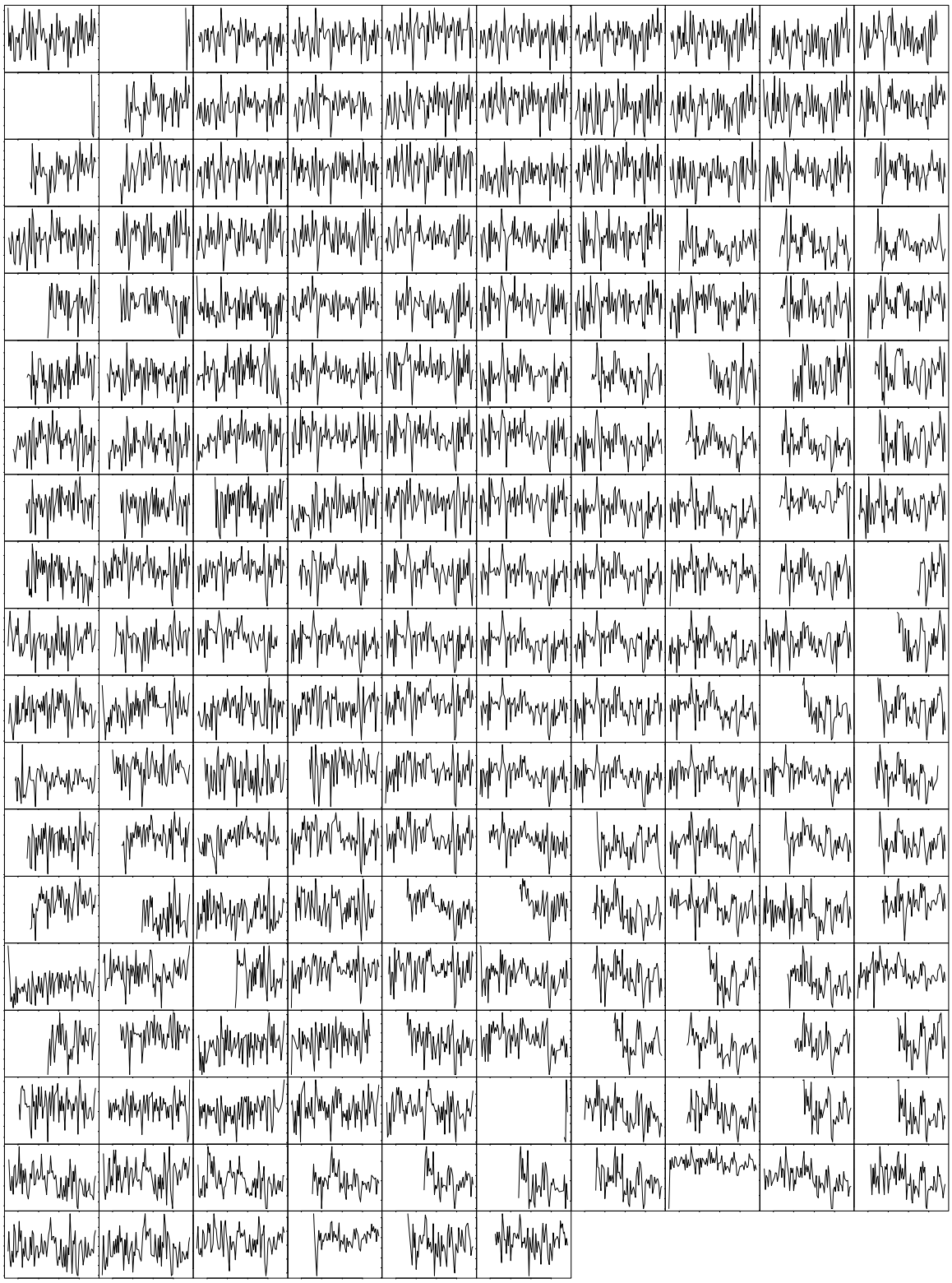


Fig. 1.4. Time series plots of 186 time series, winter means of daily maximum temperature.

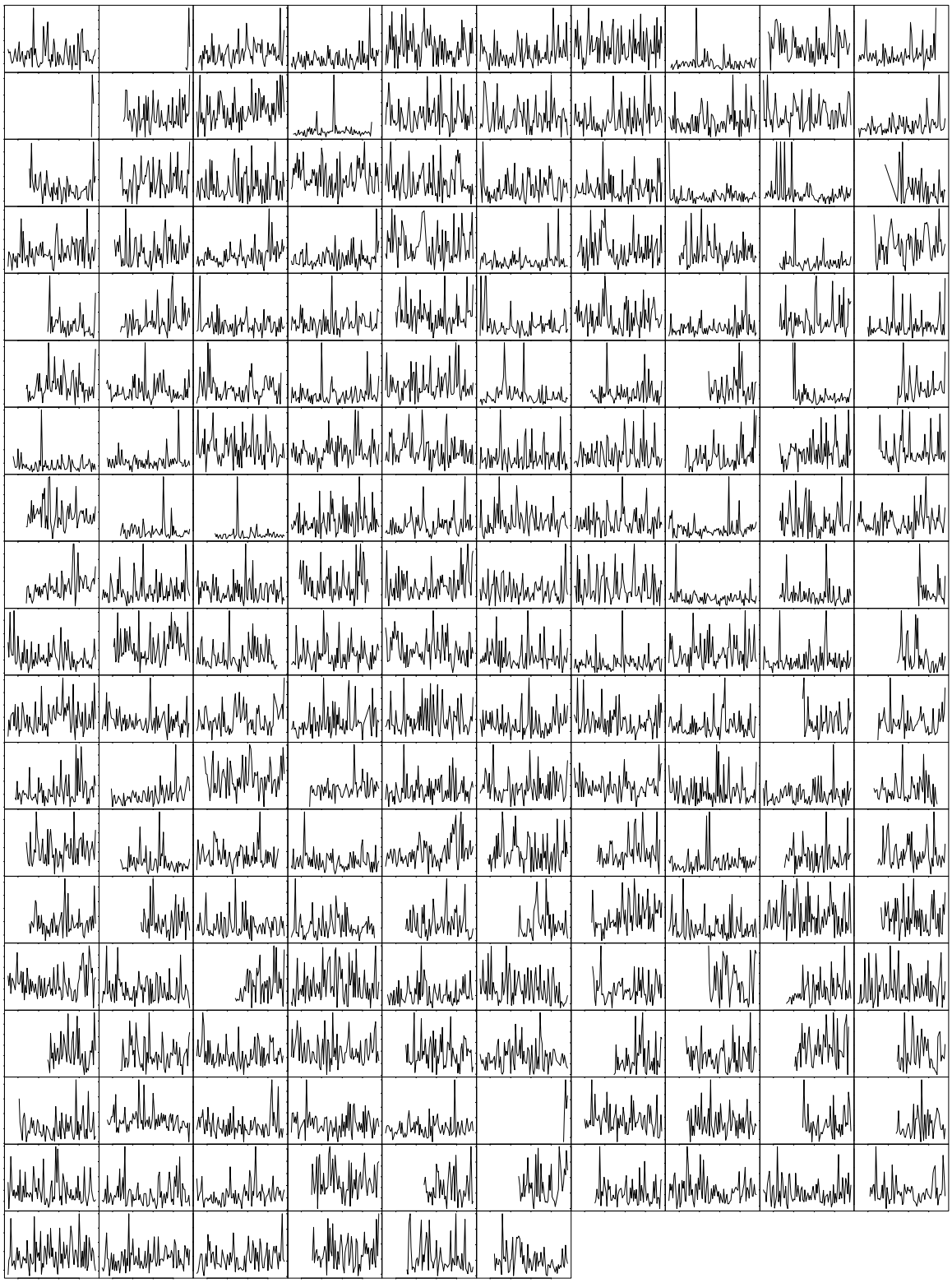


Fig. 1.5. Time series plots of 186 time series, annual maxima of daily precipitation.

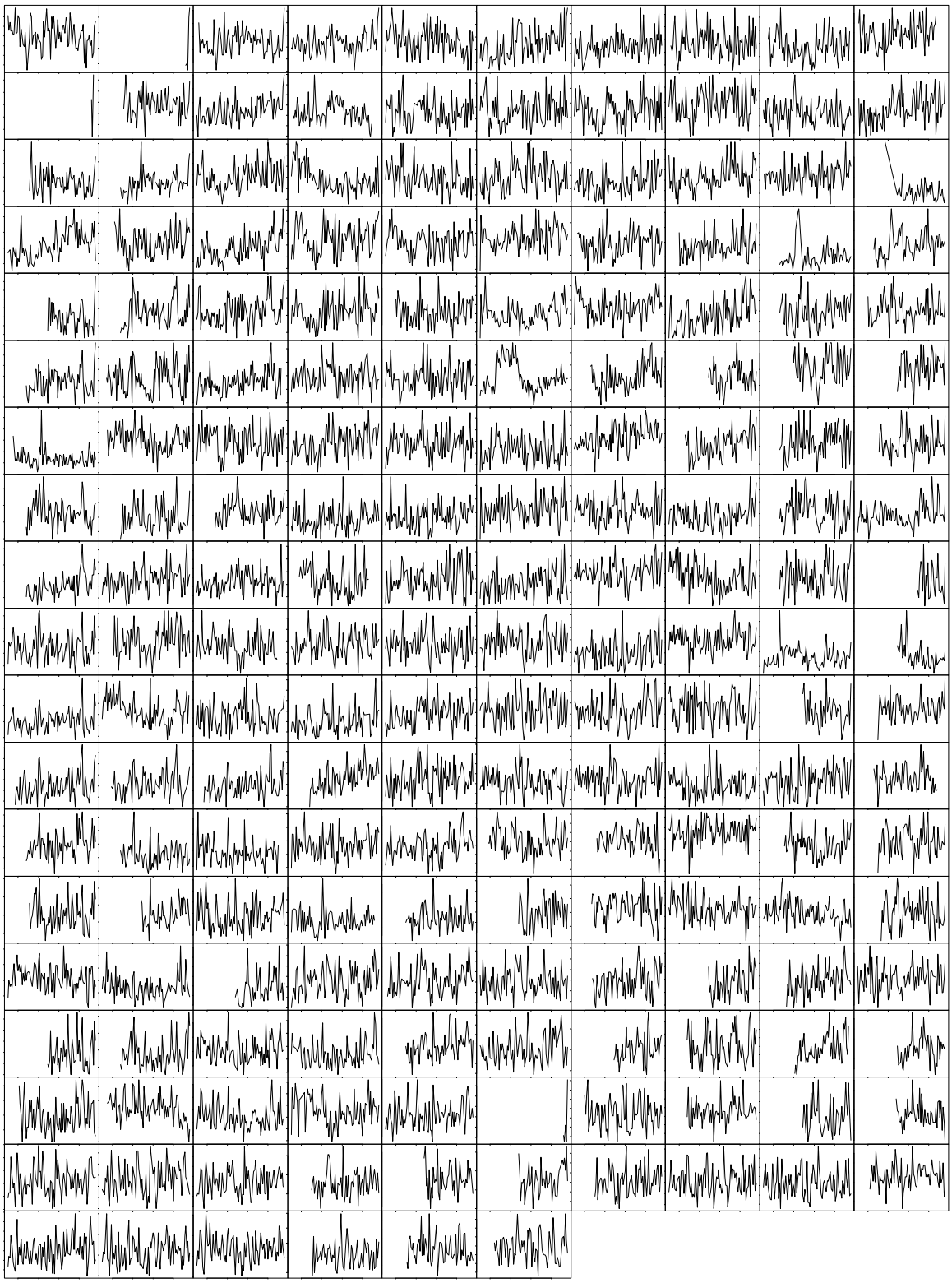


Fig. 1.6. Time series plots of 186 time series, annual mean daily precipitation.

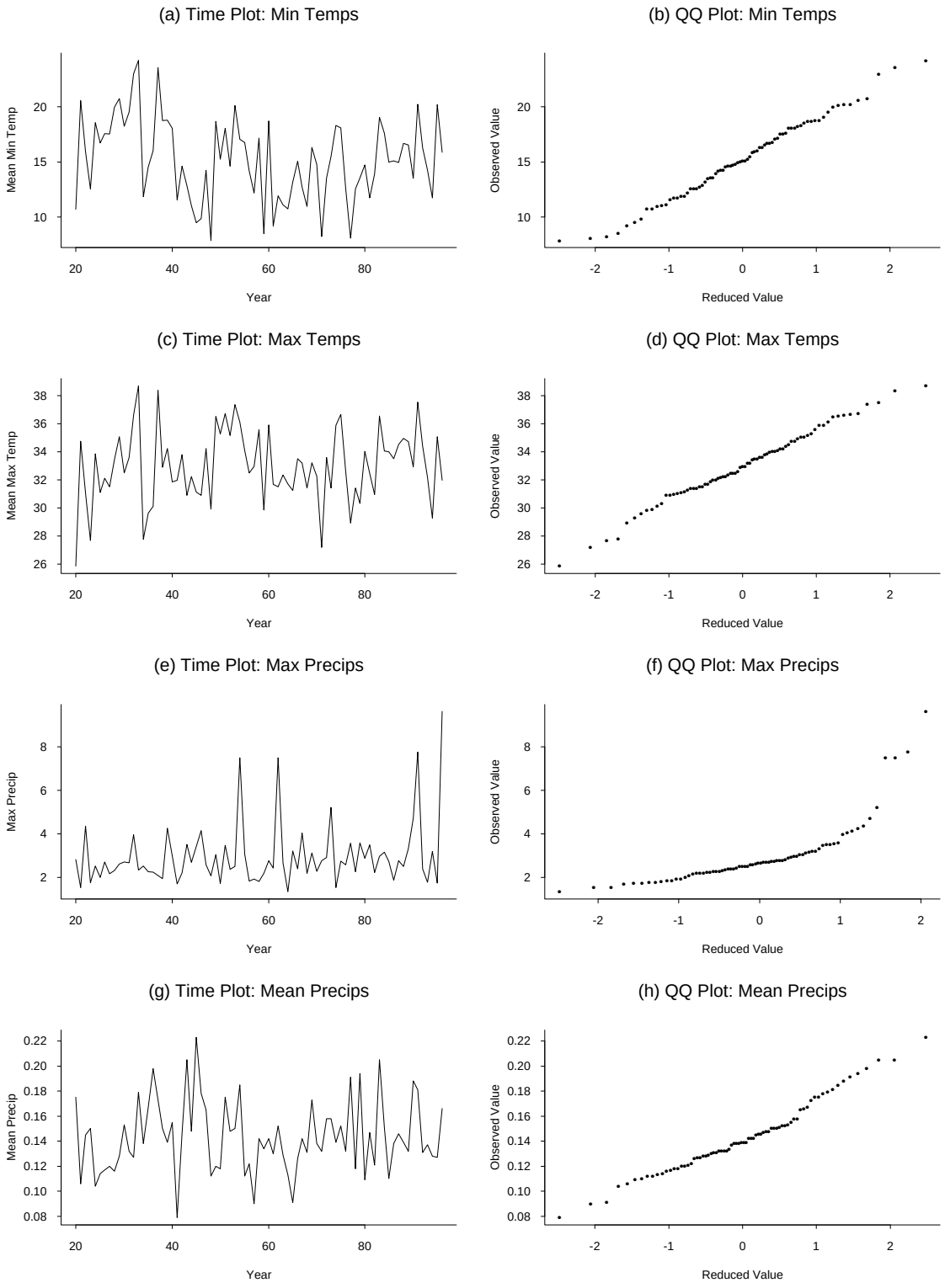


Fig. 1.7. The four series plotted for Portland, Maine, together with normal Q-Q plots. The largest rainfall value (50 inches) has been treated as a censored value.

1.4 Trends in Individual Series

Although our ultimate interest will be in how we can best combine information spatially from series collected in different locations, for the moment we see how far it is possible to go by looking at individual time series. This will provide useful information about the nature of the time series, and will give some guidelines about "what we are looking for" when we come to consider spatial aspects.

For each of the four time series available at each location, and for AR order $m = 0, 1, ..., 4$, the model defined by (1.1) and (1.2) was fitted, with $x_t = t$ to represent a linear trend over the period of the analysis. For the temperature minima and maxima series, since the strongest evidence of a warming trend is available in the data since the mid-sixties, the analysis was confined to the time period 1965–1996. In the case of rainfall series, the evidence available previously has not suggested that the increasing trend be confined to a particular period, so the analysis was carried out for the whole period 1900–1996. In both cases, any series with less than 20 "available" values was omitted from the analysis (see Section 1.3 for the definition of available).

For these series, AIC-based comparisons of different AR models with different orders m tend to select $m = 0$ as the preferred order of the model. For example, with temperature minima, 114 out of the 182 series had minimum AIC for $m = 0$. The corresponding figures for the temperature maxima, rainfall maxima and rainfall means were 103, 125 and 153 respectively. It is interesting to contrast this result with those for the hemispheric averages in section 1.2, where we found very strong evidence of serial dependence and AIC values indicated $m = 4$ and $m = 3$ respectively for the NH and SH series. The contrast between the two sets of results may be the results of much higher noise levels in the individual series, masking whatever correlation is present. This may in itself be an argument for calculating regional averages based on spatial groupings of the individual stations, but for now we stick with the individual series and we assume that all the series have independent errors.

For each series, the statistical significance of the trend was assessed by computing the estimate and standard error of the slope β_1, and then computing the ratio of the two — the t ratio. Grouping the t ratios into certain bands, and classifying according to the number of stations in each band, produced the results given in Table 1.2.

In each case there is a preponderance of values with $t > 0$, and in particular the number of stations for which $t > 2$ is much larger than would be expected if there were no trend (approximately $2.5\% \times 182 = 4.55$). However it is clear that this effect is stronger in the case of temperature minima than temperature maxima, consistent with what other studies have shown. Comparing maximum and mean rainfall levels, there are more stations with significant positive ($t > 2$) trends for mean rainfall than for maximum rainfall, but this is partly balanced by the number of stations with significant negative trends, and in any case the results for rainfall maxima are less reliable in view of the heavy skewness which we have observed in the data.

Series		Range of t values			
	> 2	1 to 2	0 to 1	-1 to 0	-2 to -1	< -2
Minima	37	52	61	22	8	2
Maxima	17	34	68	46	14	3
Max Rain	14	42	55	51	16	4
Mean Rain	35	49	42	35	10	11

Table 1.2. Grouped t values for the slope of a linear trend in individual time series. For example, with the temperature minima series, 37 out of the 182 stations had a t value for the slope bigger than 2, 52 has a t value between 1 and 2, and so on. Regressions are based on time period 1965–1996 for the temperature maxima and minima series, and on 1914–1996 for the rainfall series, fitted by ordinary linear regression with independent errors.

A natural question to ask is what is the spatial distribution of trends over the country. These are depicted in Figs. 1.8–1.11, where in each case the top map shows the distribution classified by the t value of the slope (A for $t > 2$, B for $1 < t \leq 2$, etc.) and the bottom map shows the distribution classified by the slope itself (A for top sextile, B for next sextile, etc.). It is possible to pick out some broad patterns from these. For example, with both the temperature maxima and temperature minima the increasing trends seem to be concentrated in the northern midwestern states whereas in the south east, for example, there are many more stations with a cooling trend. For the rainfall maps it is harder to pick out any consistent pattern. With both temperature and rainfall, there is clearly a lot of variability even between neighboring stations, and this points towards the desirability of some kind of spatial smoothing to produce more satisfactory estimates of local patterns in the trend.

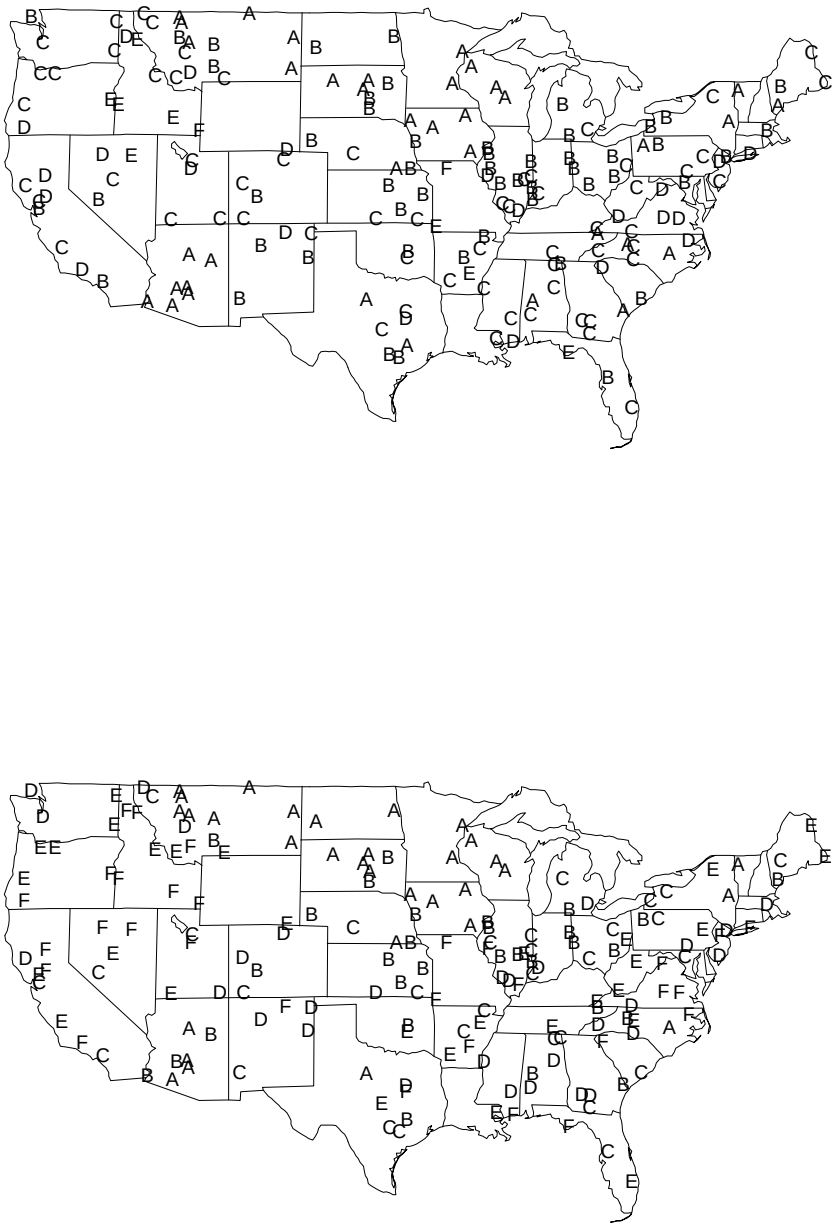


Fig. 1.8. Spatial distribution of trend estimates for means of daily temperature minima. Top plot: Classified by t value, A signifying $t > 2$, B signifying $1 < t \leq 2$ and so on down to F for $t \leq -2$. Bottom plot: Classified by the parameter estimate itself, A for top sextile, B for next sextile, etc.

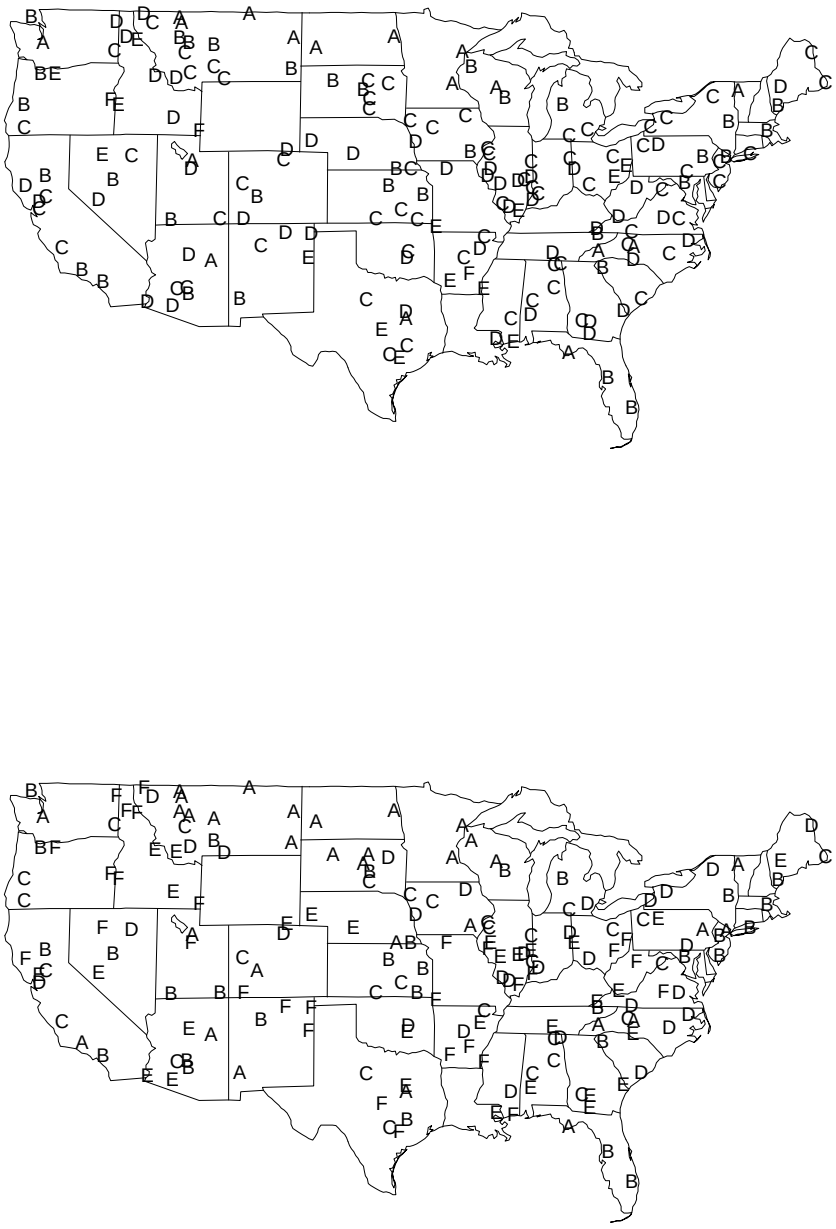


Fig. 1.9. Spatial distribution of trend estimates for means of daily temperature maxima. Same notations as Fig. 1.8.

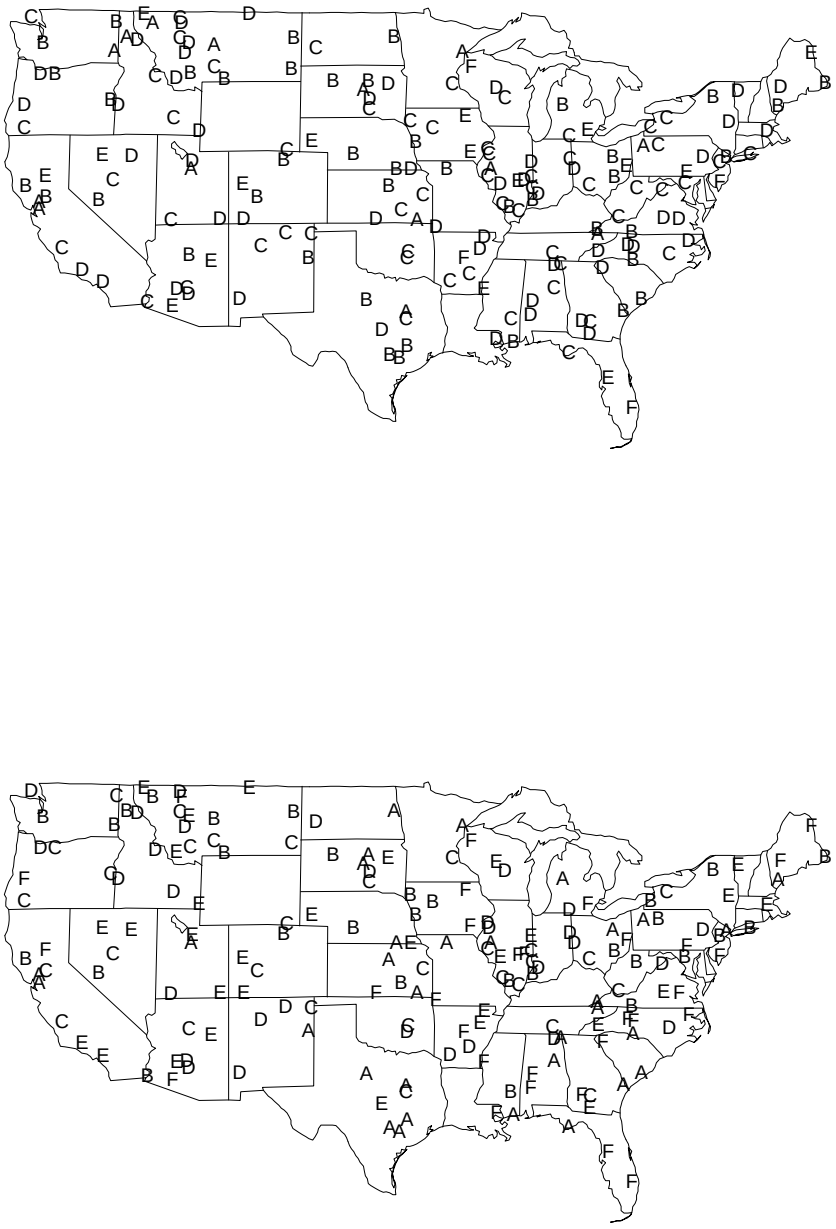


Fig. 1.10. Spatial distribution of trend estimates for maximum daily precipitation. Same notations as Fig. 1.8.

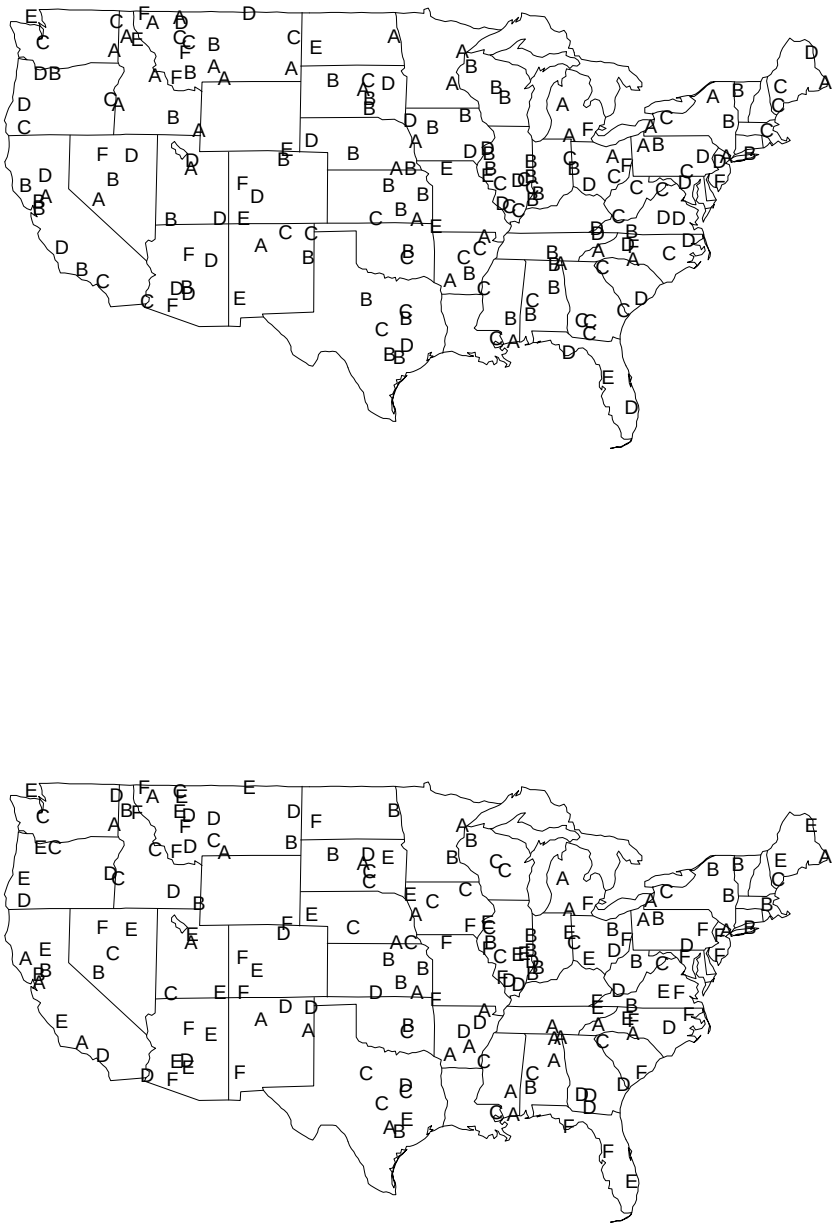


Fig. 1.11 Spatial distribution of trend estimates for mean daily precipitation. Same notations as Fig. 1.8.

1.5 The Need for Spatial Analysis

The discussion in Section 1.4 has shown a number of the difficulties in analyzing the individual series on their own. The high degree of variability in individual series makes the trend estimates unreliable, and although it is possible to pick out some very broad features from maps such as those in Figs. 1.8–1.11, there is still too much random variation from one station to the next to be able to pick out coherent spatial features.

There are (at least) three potential benefits from trying to characterize and exploit the spatial dependence between nearby stations:

1. Formation of regional averages. In practice, most published climatological studies do not rely on single series of measurements but form averages over regions. For example, in one recent analysis Karl and Knight (1998) divided the continental USA into nine climatic regions for a comparison of precipitation trends. Fig. 1.1 represents a cruder classification into four broad regions. The question remains, however, of how to form the averages. Simply averaging over all available stations within a region may not be anything like the best way of estimating the overall average of the region. Kriging methods, which exploit spatial correlations to form optimal linear combinations of the available data, can in principle produce estimates with minimum mean squared error.

2. Improving estimates. If we exploit spatial correlations, we may be able to form better estimates of trend, with smaller standard errors.

3. Spatial smoothing, i.e. trying to improve maps such as those in Figs. 1.8–1.11 by avaeraging over spatially correlated stations. As with spatial averaging, we can expect to do this much more efficiently if we have some knowledge of spatial correlations in the data.

One common method of characterizing spatial dependence is the *variogram*, to be defined more precisely in Chapter 2. The value of the variogram at distance d is the variance of the difference between two stations a distance d apart, it being part of the assumption that this depends on d alone (the homogeneity assumption). There are two standard methods of calculating it, one the obvious average of squared differences between the two stations (the Method of Moments estimator), and the other a robust alternative. In practice, distances are usually grouped into bins before the calculation is carried out. All of this is explained in detail in Chapter 2.

In Figs. 1.12–1.15, the variograms for each of our four meteorological series are plotted, both Method of Moments and Robust estimators, for each of the four regions shown in Fig. 1.1. These plots are based on the standardized residuals (to have mean 0 and variance 1) from the regressions in Section 1.4.

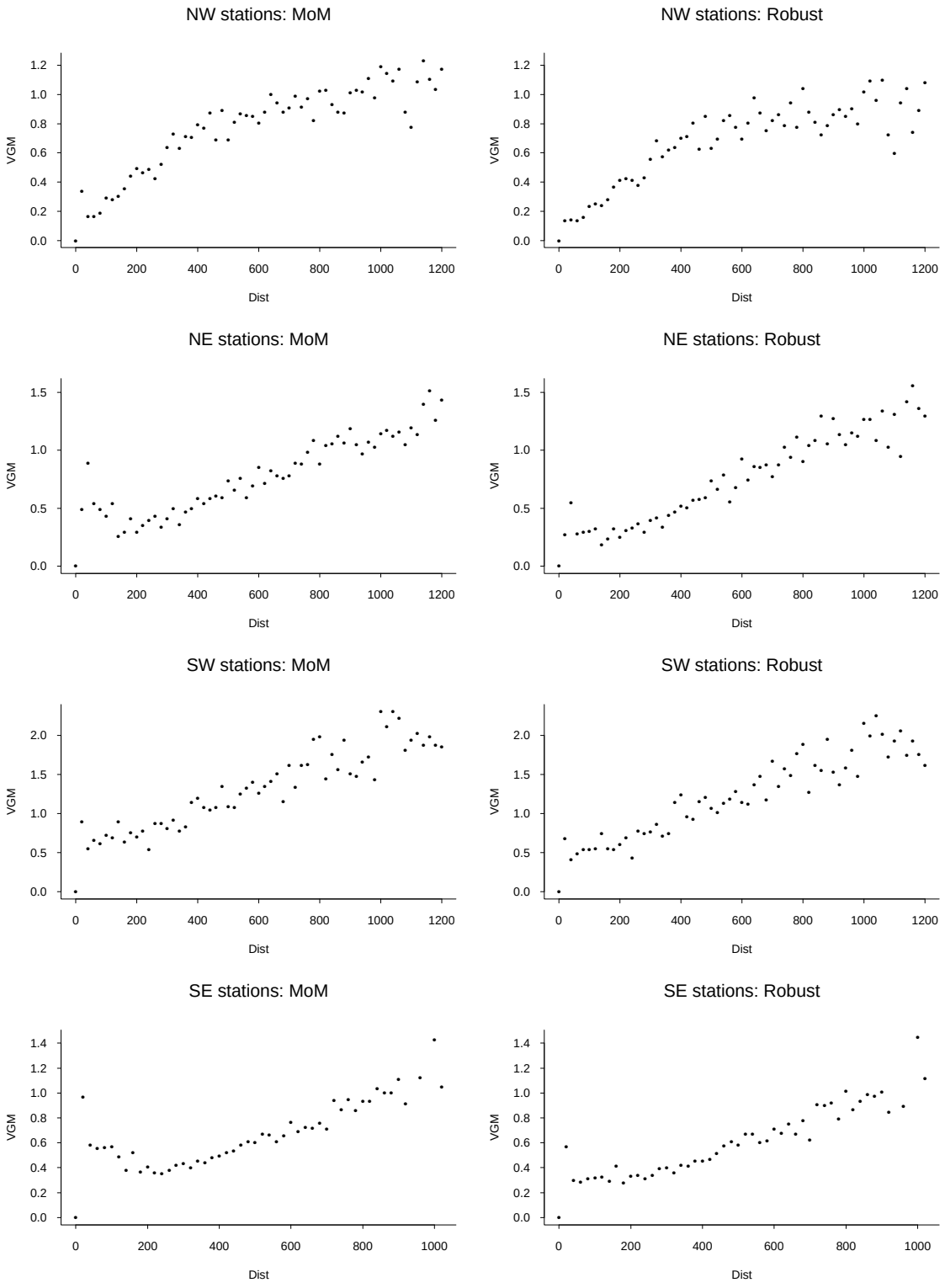


Fig. 1.12 Variogram for winter mean daily minimum temperatures, classified by region, Method of Moments and Robust estimators.

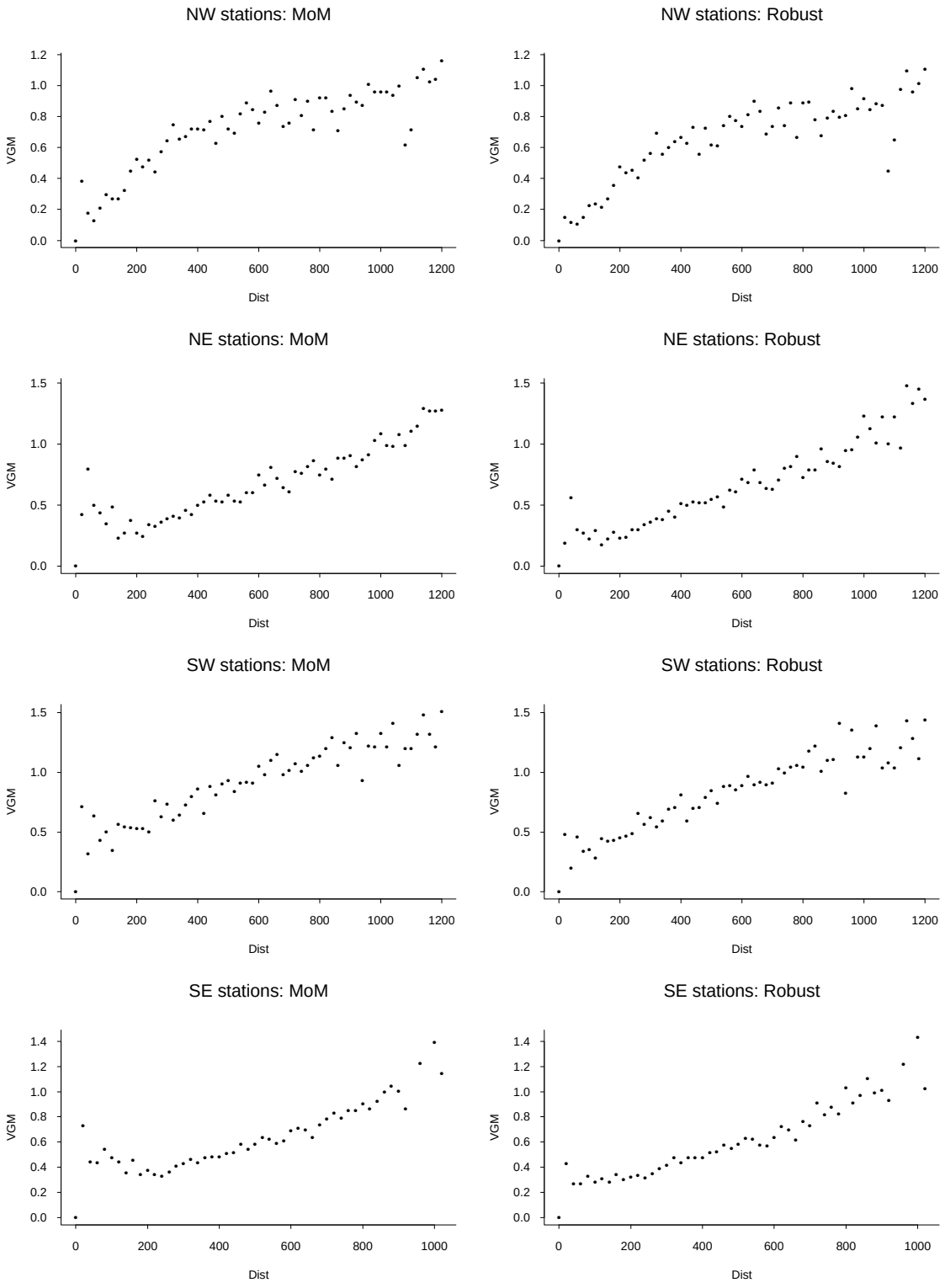


Fig. 1.13 Variogram for winter mean daily maximum temperatures, classified by region, Method of Moments and Robust estimators.

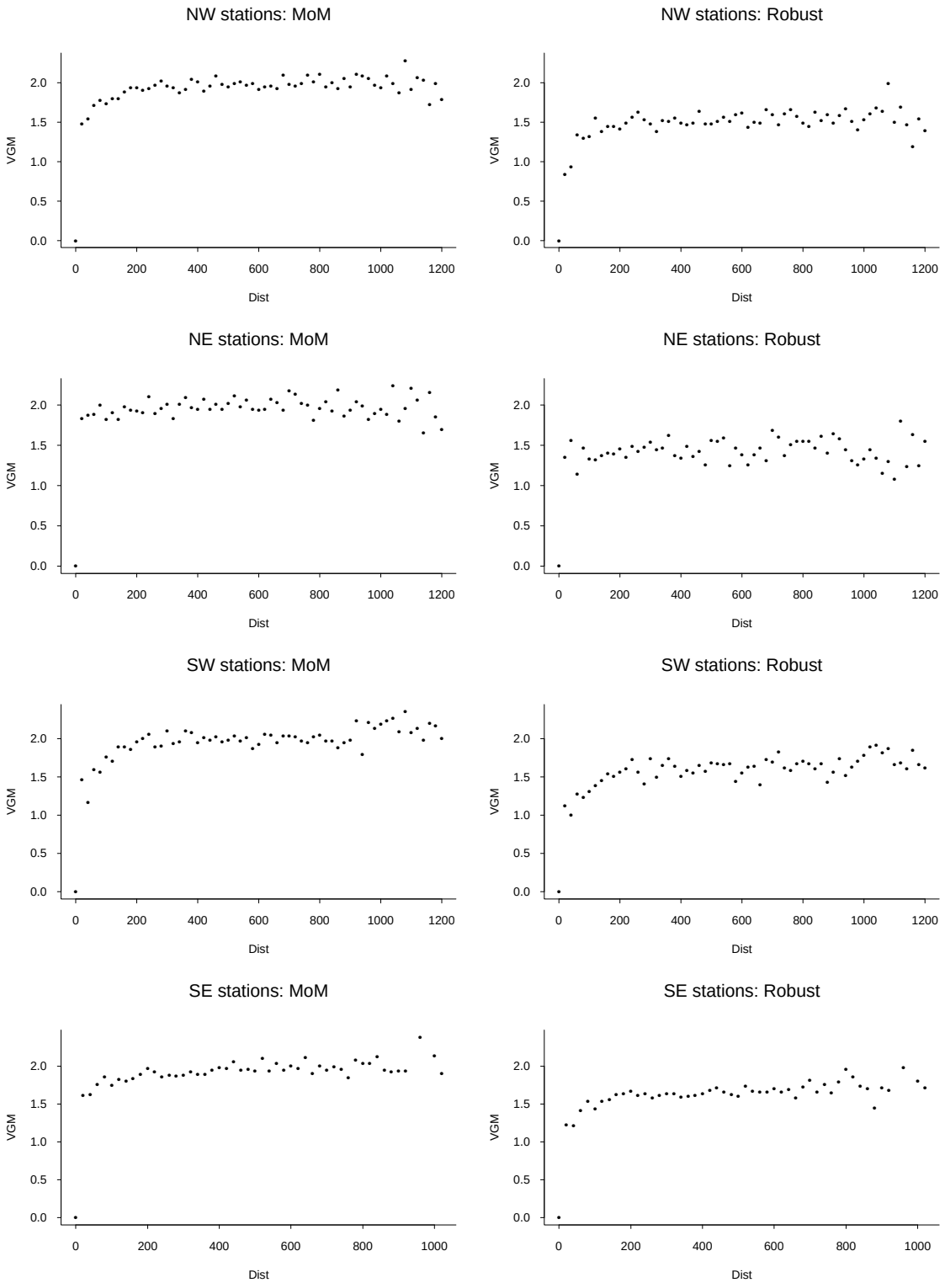


Fig. 1.14 Variogram for annual maximum daily precipitations, classified by region, Method of Moments and Robust estimators.

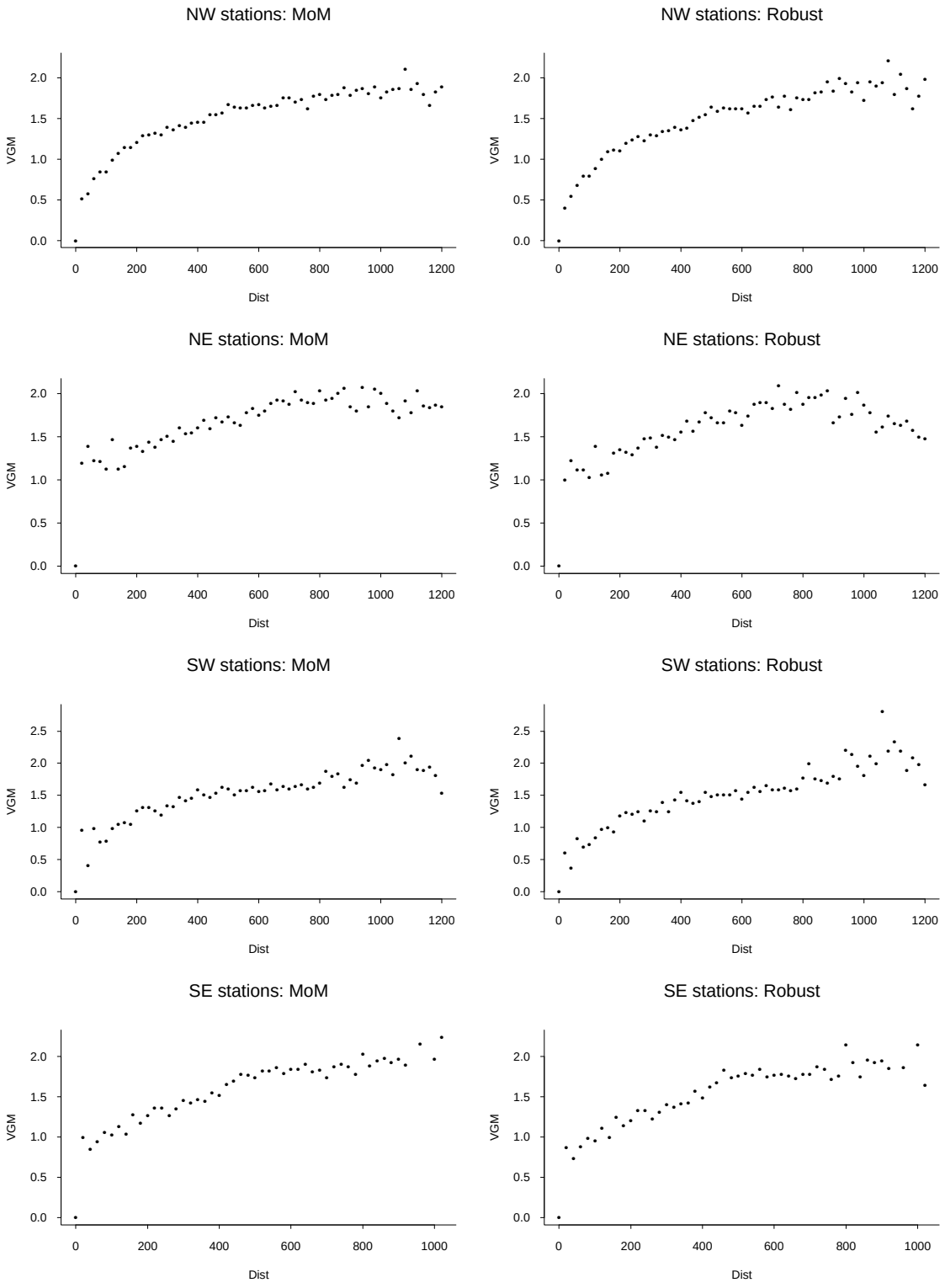


Fig. 1.15 Variogram for annual mean daily precipitations, classified by region, Method of Moments and Robust estimators.

The subdivision into four regions is motivated by the desire to improve the homogeneity of spatial estimates compared with an analysis based on treating the whole country as a single entity. The differences between variogram estimates within different regions shows that some subdivision of this form is necessary, though we cannot tell at this stage whether simply dividing the country into four regions is adequate for this. For example, it appears from Fig. 1.12 that the variogram estimates for the SW region are larger than those in other regions. Since all the residuals are standardized to have common variance 1, this implies that the spatial correlations decay more rapidly in the SW than in the other regions. On the other hand, we do not see a similar pattern in Figs. 1.13–1.15. We can also see that the shape of the variogram varies considerably for the four meteorological series. The faster the variogram approaches its "sill", where it levels off, the shorter the range of spatial correlations. From Fig. 1.14 we can see that the rainfall maxima have very short-range spatial correlations, while from Fig. 1.15, the rainfall means have longer spatial correlations with a "range" of the order 400–800 nautical miles. Both the temperature series show a variogram increasing steadily across the whole range of the plot, implying very long-range spatial correlations.

1.6 An Overview of Environmental Statistics

The problems we have discussed in this chapter are typical of many problems arising in other fields besides those connected with climate. For example, the US Environmental Protection Agency (EPA) has extensive networks of monitoring stations to collect data on atmospheric pollutants such as ozone (O_3), sulfur dioxide (SO_2) and particulate matter of aerodynamic diameter less than or equal to 10 microns (PM_{10}). Among the statistical problems associated with such data sets and the source of detailed examples in Nychka *et al.* (1998) are

(a) Characterizing trend in O_3 or SO_2 series as a function of time and meteorology. These variables are highly dependent on meteorology; therefore, the variability of the meteorological series may mask effects due to changes in the emissions of atmospheric pollutants. By constructing a suitable regression model, we can hope to separate the effects of meteorology and changing emissions and thus characterize the effect of trends in emissions. Many regression strategies have been applied, from parametric linear and nonlinear regression models to nonparametric or semiparametric techniques such as generalized additive modeling (Hastie and Tibshirani 1990) or the method of Sacks *et al.* (1989), the latter of which itself borrows many ideas from the theory of spatial statistics. In many cases, the analysis shows that meteorologically adjusted O_3 or SO_2 levels have been decreasing during the 1980s and 1990s. Although the scientific context is quite different from the ones that govern long-term climatic trends, this shows that "trend estimation" is a very general topic which draws on many areas of statistical methodology.

(b) Spatial variability of ozone. The EPA needs to monitor ozone over rural as well as urban regions, but the density of ozone monitors is typically much lower in rural regions than in cities. Adequate monitoring of rural ozone patterns requires the combination

of data into spatially coherent regions — this includes the identification of appropriate regions. Spatial analysis based on rotated principal components has been used to identify regions and then to calculate regional trends.

(c) Threshold crossings. Prior to 1997, the US ozone standard was based on the number of crossings of daily maximum ozone of a threshold level set at 120 ppb. The new standard introduced during 1997 reduced the threshold to 80 ppb, but based on eight-hour ozone averages rather than daily maxima. Either way, we need to study the frequency of crossings of a high level, a problem closely related to the extreme values problem mentioned earlier in connection with rainfall. Methods from extreme value theory have been adapted to analyze the frequency of high-threshold crossings as a function of time and meteorology. One conclusion from such analyses is that the evidence for a decreasing trend in ozone, after adjusting for meteorology, is often stronger when looking at extreme levels than when looking at the series as a whole.

(d) Design of networks. Often the EPA is required to set up a new network, or to reduce the size of an existing network. A major criterion used in making such decisions is to be able to interpolate between points of the network in the most efficient way. This requires characterizing the spatial correlations in order to calculate optimal interpolators and their mean squared prediction errors, and then determining the network so that the overall mean squared prediction error, in some suitably defined sense, is minimized.

(e) Health effects. A major controversy surrounding particulate matter (PM_{10}) is the effect of small quantities of PM_{10} on human health — some authors have claimed very strong effects on mortality and morbidity, particularly among the elderly population. These claims have strongly influenced the much tighter particulate matter standards introduced by the EPA during 1997. Other authors have pointed out the dependence of the conclusions on seemingly arbitrary aspects of model selection, or have suggested that the strongest effects only occur at high PM_{10} levels already regulated by the EPA, or have disputed whether the effect is really due to PM_{10} as opposed to other atmospheric pollutants. The analyses in Nychka *et al.* (1998) did not employ any spatial analyses since they were based on single-city data sets, but there is a close connection with disease mapping problems requiring spatial modeling both of pollutants and disease incidents, see e.g Zidek (1997).

(f) Combining information. A very general problem of environmental statistics is that of combining different sources of information — for example, from two spatial networks set up for different purposes but intended to measure the same (or similar) quantities. One of the most critical and interesting problems is how to combine data from probability and non-probability samples. The former refers to sampling designs constructed according to proper principles of randomized designs, whereas the latter may often involve data collected in a haphazard or deliberately biased way, e.g. taking samples from a lake with a intention of identifying "hotspots" of high pollution intensity. How can the information from the two types of sample be best combined without introducing a bias from the non-probability sample? This again involves questions of spatial modeling, sometimes supplemented by hierarchical models to represent the variability between different networks.

From this discussion we can see that the general themes which have been identified in our discussion of the climatic data sets are recurrent themes throughout environmental statistics. In praticular, the need for both temporal and spatial modeling of monitoring networks is a central part of the discipline, and there are numerous recurrent themes including the identification of trends, characterizing the distribution of extremes, and sample design problems. All of these problems will be described in more detail in later chapters of the present notes.

1.7** Derivation of the Estimates in Sections 1.2 and 1.4

The purpose of this section is to outline briefly the method used for fitting the regression model with autoregressive errors.

The basic method is numerical maximum likelihood. Equations (1.1) and (1.2) relate the observed time series $\{y_t\}$ to a series of independent errors $\{\epsilon_t\}$, assumed to be normally distributed with mean 0 and variance σ^2. The likelihood L based on $\{\epsilon_t,\ 1 \le t \le T\}$ is given (modulo irrelevant constants) via

$$-\log L = \frac{T}{2}\log\sigma^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{\epsilon_t}{\sigma^2}\right)^2. \tag{1.3}$$

For given parameter values $\beta_0,\ \beta_1,\ \phi_1, ..., \phi_m$, one can calculate the values of ϵ_t and hence evaluate (1.3) as a function of these parameters as well as σ^2. The maximum likelihood estimates are those which minimize (1.3); these may be obtained via any general-purpose routine for unconstrained minimization. The calculations resported here used the DFPMIN routine available in Press *et al.* (1986), using simple differencing approximations to the first-order derivatives of –log L.

There is one complication: if the series is only available for the time span $t = 1, ..., T$, then it will not be possible to calculate the values of $\eta_0,\ \eta_{-1},\ \eta_{1-m}$ used for the first few evaluations of (1.2). This problem has been avoided by conditioning on past values of the series. In other words, we assume that the values of $\{y_t\}$ and $\{x_t\}$ are available for $t \ge 1-m$, but the values for $1-m \le t \le 0$ are used solely for generating the $\{\eta_t\}$, and do not enter the sum (1.3). In the present example, this is not a problem, because the series actually go back well before the nominal starting date of 1900. In traditional time series analysis, there are two ways to deal with this aspect of the problem, (a) the conditional approach, in which all $\eta_t,\ t \le 0$, are assumed equal to 0; (b) the unconditional approach, in which the values of $\eta_1, ..., \eta_m$ are assumed to follow the stationary distribution, and the rest generated conditionally on those. Approach (a) is easier to apply but (b) is theoretically more accurate.

Once the maximum likelihood estimates are determined, the results are interpreted in the usual ways. In particular, the Hessian matrix (matrix of second-order derivatives)

of –log L, evaluated at the maximum likelihood estimates, is inverted; the square roots of the diagonal entries of the inverse matrix provide approximate standard errors of the parameter estimates. The Hessian matrix itself is usually not evaluated exactly but obtained approximately as a by-product of the numerical minimization routine.

The AIC criterion, mentioned in the text and used as a model selection criterion, is given by $-2 \log L + 2p$, where $p = m + 3$ is the number of parameters being estimated. For the general theory of maximum likelihood we refer to any standard textbook treatment, for example Chapter 9 of Cox and Hinkley (1974).

CHAPTER 2

Models for Spatial Correlations

Much of the present theory of spatial statistics has developed from work originally done in the mining industry, regarding spatial sampling of rock formations, and the problem of estimating the total quantity of an ore or mineral in a field, given concentrations of the ore or mineral at a finite set of sampling points. This work, which began in South Africa and reached maturity at the Ecoles des Mines at Fontainebleau near Paris, gave rise to the term *Geostatistics*, by which this whole area of statistics is still sometimes known. The modern environmental applications of Geostatistics go well beyond the original mining applications, but much of the terminology developed by the early researchers is still widely used. One of the earliest papers in this field, by the South African mining engineer Krige (1951), developed the basic equations for optimal linear interpolation in a spatially correlated field; his name was immortalized by Matheron (1962, 1963a, 1963b, 1971) in a series of books and papers which used the term *krigeage* (in French), or its English equivalent *kriging*. A different approach to spatial statistics began with empirical observations that variability in large agricultural field trials was inconsistent with an assumption of independence (Fairfield Smith 1938) and the development of statistical methods to deal with that problem (Papadakis 1937, Bartlett 1938). This led to subsequent developments by Whittle (1954) and Bartlett (1976, 1978); however, the modern development of statistical methods appropriate in this setting began with Besag (1974), and led in a rather different direction from the work in geostatistics. These models are developed in chapter 4. The current chapter, therefore, concentrates largely on the geostatistical approach and its generalizations as they are used in modern environmental statistics. In particular, the chapter focusses on the variogram and on its applications to kriging. However we also discuss estimation techniques from a modern point of view, focussing on maximum likelihood and Bayesian solutions. Among books which have taken a variety of approaches from the very theoretical to the very applied, we mention Journel and Huijbregts (1978), Cliff and Ord (1973, 1981), Ripley (1981, 1988), Upton and Fingleton (1985, 1989), Cressie (1993) and Stein (1999). Cressie's book is particularly comprehensive and is a very valuable reference source.

Section 2.1 outlines a number of topics from the theory of spatial stochastic processes, focussing on concepts of stationarity and isotropy and also discussing various parametric models for spatial correlations. Sections 2.2 and 2.3 give the statistical theory surrounding estimation of the variogram and the fitting of parametric covariance models to spatial data, together with some examples. Section 2.4 discusses kriging and its extensions. Section 2.5 develops the Bayesian approach in more detail, following, in particular, ideas of Le and Zidek. Finally Section 2.6 presents some more detailed examples.

2.1 Spatial Processes

The basic object we consider is a stochastic process $\{Z(s),\ s \in D\}$ where D is a subset of $\mathcal{R}^d$ (d-dimensional Euclidean space), usually though not necessarily $d = 2$. For example, $Z(s)$ may represent the mean winter daily maximum temperature at a specific location s. Let

$$\mu(s) = \mathrm{E}\{Z(s)\}, \quad s \in D,$$

denote the mean value at location s. We also assume that the variance of $Z(s)$ exists for all $s \in D$.

The process Z is said to be *Gaussian* if, for any $k \geq 1$ and locations $s_1, ..., s_k$, the vector $(Z(s_1), Z(s_2), ..., Z(s_k))$ has a multivariate normal distribution.

The process Z is said to be *strictly stationarity* if the joint distribution of $(Z(s_1), Z(s_2), ..., Z(s_k))$ is the same as that of $(Z(s_1+h), Z(s_2+h), ..., Z(s_k+h))$ for any k spatial points $s_1, s_2, ..., s_k$ and any $h \in \mathcal{R}^d$, provided only that all of $s_1, s_2, ..., s_k, s_1 + h, s_2 + h, ..., s_k + h$ lie within the domain D.

The process Z is said to be *second-order stationarity* (also called *weakly stationary*) if $\mu(s) \equiv \mu$ (i.e. the mean is the same for all s) and

$$\mathrm{cov}\{Z(s_1), Z(s_2)\} = C(s_1 - s_2), \quad \text{for all } s_1 \in D,\ s_2 \in D,$$

where $C(s)$ is the covariance function of an observations at location s with one at location 0.

It can immediately be seen that with all variances assumed finite, a strictly stationary process is also second-order stationary. The converse is in general false, but a *Gaussian* process which is second-order stationary is also strictly stationary.

The next concept which we need to introduce is:

The Variogram. Assume $\mu(s)$ is a constant, which we may without loss of generality take to be 0, and then define

$$\mathrm{var}\{Z(s_1) - Z(s_2)\} = 2\gamma(s_1 - s_2). \tag{2.1}$$

The statement (2.1) makes sense only if the left hand side depends on s_1 and s_2 only through their difference $s_1 - s_2$. A process which satisfies this property is called *intrinsically stationary*. The function $2\gamma(\cdot)$ is called the *variogram* and $\gamma(\cdot)$ the *semivariogram*.

Intrinsic stationarity is a weaker property than second-order stationarity. Suppose, first, that the process is second-order stationary. Then it is easy to verify that

$$\begin{aligned}\mathrm{var}\{Z(s_1) - Z(s_2)\} &= \mathrm{var}\{Z(s_1)\} + \mathrm{var}\{Z(s_2)\} - 2\mathrm{cov}\{Z(s_1), Z(s_2)\} \\ &= 2C(0) - 2C(s_1 - s_2)\end{aligned}$$

and so

$$\gamma(h) = C(0) - C(h). \tag{2.2}$$

Conversely, suppose we wanted to find the function $C(\cdot)$ given the function $\gamma(\cdot)$. This could be found from (2.2) once we knew $C(0)$. In an ergodic stationary process, we will have $C(h) \to 0$ as $h \to \infty$, so $C(0)$ may be found as the limit of $\gamma(h)$ as $h \to \infty$. In general, however, there is no guarantee that such limit exists. For example, if $Z(s)$ is standard Brownian motion in one dimension we have $\text{var}\{Z(s_1) - Z(s_2)\} = |s_1 - s_2|$, which tends to ∞ as $|s_1 - s_2| \to \infty$. This is an example of a process which is intrinsically stationary but not second-order stationary. Similar examples exist in higher dimensions, such as the process known as the Brownian sheet.

To summarize: if $\lim_{h' \to \infty} \gamma(h')$ exists then the process is second-order stationary with $C(h) = \lim_{h' \to \infty} \gamma(h') - \gamma(h)$, but if this limit does not exist then the process is not second-order stationary.

For much of the theory of spatial processes, the principal assumption required is intinsic stationarity. From this point of view, the stronger forms of stationarity are not needed. On the other hand, either strict or second-order stationarity are more natural assumptions — for example, as with time series analysis, it is often very useful to think of an observed process as consisting of a deterministic trend superimposed on an underlying stationary field. For this reason, it is a good idea to be cautious when a preliminary analysis of the data indicates that the process is intrinsically stationary but not stationary. It may well be that the process is best approached by first looking for a trend with stationary residuals.

A separate concept is *isotropy*. Suppose the process is intrinsically stationary with semivariogram $\gamma(h)$, $h \in \mathcal{R}^d$. If $\gamma(h) = \gamma_0(||h||)$ for some function γ_0, i.e. if the semivariogram depends on its vector argument h only through its length $||h||$, then the process is *isotropic*.

A process which is both intrinsically stationary and isotropic is also called *homogeneous*.

Isotropic processes are convenient to deal with because there are a number of widely used parametric forms for $\gamma_0(\cdot)$. Here are several examples:

1. *Linear:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1 t & \text{if } t > 0. \end{cases}$$

Here c_0 and c_1 are positive constants. The function tends to ∞ as $t \to \infty$ and so does not correspond to a stationary process.

2. *Spherical:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1 \left\{ \frac{3}{2} \frac{t}{R} - \frac{1}{2} (\frac{t}{R})^3 \right\} & \text{if } 0 < t \leq R, \\ c_0 + c_1 & \text{if } t \geq R. \end{cases}$$

This is valid if $d = 1,\ 2$ or 3, but for higher dimensions it fails the non-positive-definiteness condition (see below). It is a convenient form because it increases from a positive value c_0 when t is small, levelling off at the constant $c_0 + c_1$ at $t = R$. This is of the "nugget/range/ sill" form which is often considered a realistic and interpretable form for a semivariogram (further discussed below).

3. *Exponential:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1(1 - e^{-t/R}) & \text{if } t > 0. \end{cases}$$

Simpler in functional form than the spherical case (and valid for all d) but without the finite range of the spherical form. The parameter R has a similar interpretation to the spherical model, however, of fixing the scale of variability.

4. *Gaussian:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1(1 - e^{-t^2/R^2}) & \text{if } t > 0. \end{cases}$$

5. *Exponential-power form:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1(1 - e^{-|t/R|^p}) & \text{if } t > 0. \end{cases}$$

Here $0 < p \leq 2$. This form generalizes both the exponential and Gaussian forms, and forms the basis for the families of spatial covariance functions introduced by Sacks *et al.* (1989), though in generalizing the results from one dimension to higher dimensions, these authors used a product form of covariance function in preference to constructions based on isotropic processes.

6. *Rational quadratic:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1 t^2/(1 + t^2/R) & \text{if } t > 0. \end{cases}$$

7. *Wave:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1 \left\{1 - \frac{R}{t} \sin\left(\frac{t}{R}\right)\right\} & \text{if } t > 0. \end{cases}$$

The only non-monotonic example in this sequence.

8. *Power law:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1 t^\lambda & \text{if } t > 0. \end{cases}$$

Non-positive-definiteness requires $0 \leq \lambda < 2$. This generalizes the linear case, and is only our second example of a semivariogram that does not correspond to a stationary process.

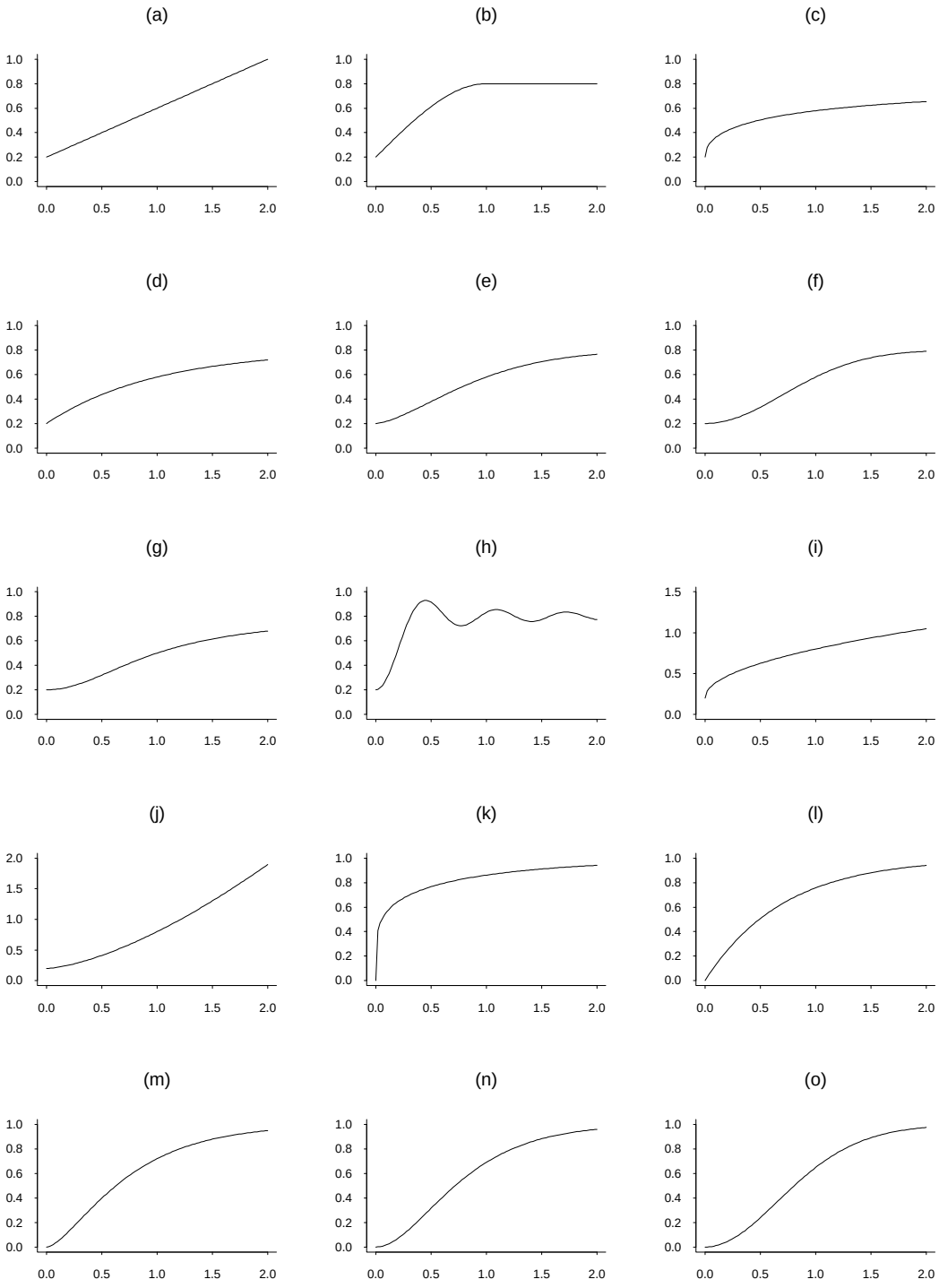


Fig. 2.1. Examples of isotropic variogram functions. (a) Linear. (b) Spherical. (c) Exponential-power, $p = 0.5$. (d) Exponential. (e) Exponential-power, $p = 1.5$. (f) Gaussian. (g) Rational quadratic. (h) Wave. (i) Power law, $\lambda = 0.5$. (j) Power law, $\lambda = 1.5$. (k)–(o) Different forms of Matérn function with θ_2 respectively 0.1, 0.5, 1, 2, 10. The different shapes of the Matérn functions near $t = 0$ can be clearly seen.

9. *The Matérn class:* This was originally given by Matérn (1960), but largely neglected in favor of simpler analytic forms. However, more recently Handcock and Stein (1993) and Handcock and Wallis (1994) demonstrated its flexibility in handling a variety of spatial data sets, including ones related to global warming. The class is best defined in terms of its (isotropic) covariance: we have $C(h) = C_0(||h||)$ where $C_0(0) = 1$ and

$$C_0(t) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{2\sqrt{\theta_2}t}{\theta_1} \right)^{\theta_2} \mathcal{K}_{\theta_2} \left(\frac{2\sqrt{\theta_2}t}{\theta_1} \right).$$

Here $\theta_1 > 0$ is the spatial scale parameter and $\theta_2 > 0$ is a shape parameter. The function $\Gamma(\cdot)$ is the usual gamma function while $\mathcal{K}_{\theta_2}$ is the modified Bessel function of the third kind of order θ_2 (Abramowitz and Stegun 1964, Chapter 9). Fortran and Splus implementations of this function are available. As special cases, $\theta_2 = \frac{1}{2}$ corresponds to the exponential form of semivariogram, and the limit $\theta_2 \to \infty$ results in the Gaussian form.

Fig. 2.1 shows some illustrative examples of isotropic semivariograms, which gives a good idea of the range of different shapes available.

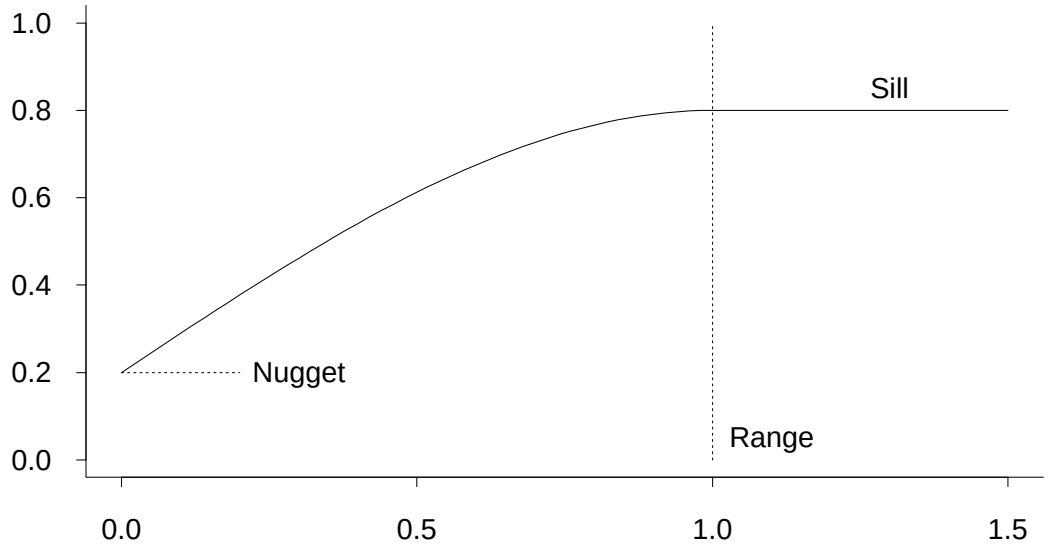


Fig. 2.2. Idealized form of variogram function, illustrating the nugget, sill and range.

Several of these semivariograms have the general shape of Fig. 2.2. We always have $\gamma_0(0) = 0$, but γ_0 increases from a non-negative value near $t = 0$ (the nugget) to a limiting value (the sill) which is either attained at a finite value $t = R$ (the range), or else approached asymptotically as $t \to \infty$. In the latter there is still a scale parameter which we may denote by R, and which may be defined precisely as the value of t at which $\gamma_0(t)$ comes within a specified distance of its limiting value. The cases where the nugget

is strictly positive may appear paradoxical because they imply there is a discontinuity in the covariance function, but in fact this is a well-known feature of spatial data. There are various possible explanations, the simplest being that there is some residual white noise over and above any smooth spatial variation.

Anisotropic cases

There are a number of ways of dealing with anisotropic processes as more or less direct generalizations of isotropic processes. The simplest of these is *geometric anisotropy.* This refers to a semivariogram of the form

$$\gamma(h) = \gamma_0(||Ah||) \tag{2.3}$$

where γ_0 is an isotropic semivariogram and A a $d \times d$ matrix, representing a linear transformation of $\mathcal{R}^d$. Of course, if A is the identity this reduces to the isotropic case. The idea is that the process is not isotropic in the original space, but is in some linearly transformed space, which may for example correspond to stretching one of the coordinates. In the most logical case that A is a positive definite matrix, the contours of equal covariance are ellipses instead of circles.

A generalization of anisotropy arises from the simple observation that if $Z_1, ..., Z_p$ are independent intrinsically stationary processes, then

$$Z = Z_1 + ... + Z_p,$$

is also intrinsically stationary, with semivariogram given by

$$\gamma(h) = \gamma_1(h) + ... + \gamma_p(h),$$

$\gamma_1, ..., \gamma_p$ denoting the semivariograms of $Z_1, ..., Z_p$ respectively. Thus

$$\gamma(h) = \sum_{i=1}^{p} \gamma_0(A_i h), \tag{2.4}$$

where γ_0 is an isotropic semivariogram and $A_1, ..., A_p$ are matrices, is a valid semivariogram generalizing geometric anisotropy. This is called *zonal anisotropy.*

A more complicated idea is to assume that, for some nonlinear function $g(s)$, the process $Z(g(s))$, rather than $Z(s)$ itself, is a stationary isotropic process. This idea can, indeed, handle nonstationary as well as nonisotropic cases and is at the core of a recent proposal by Sampson and Guttorp (1992). However, as this topic develops in quite different ways from the usual geostatistical analysis, we defer discussion of it until a later chapter.

Positive definiteness

One cannot define a spatial covariance or semivariogram function in a totally arbitrary way. The key property which is has to satisfy is *positive definiteness*. In the most general form in which $\text{cov}\{Z(s_1), Z(s_2)\} = C(s_1, s_2)$, which does not suppose any form of stationarity condition, positive definiteness means that the relation

$$\sum_i \sum_j a_i a_j C(s_i, s_j) \geq 0 \tag{2.5}$$

holds for any finite set of points $s_1, ..., s_n$ and arbitrary real coefficients $a_1, ..., a_n$. It is clear that (2.5) is necessary: the left hand side is the variance of $\sum_i a_i Z(s_i)$. That (2.5) is also sufficient is a consequence of *Bochner's theorem*.

In the case of a stationary process in d dimensions, Bochner's theorem implies the spectral representation

$$C(h) = \int ... \int \cos(\omega^T h) G(d\omega) \tag{2.6}$$

where the integral is over $\mathcal{R}^d$ and G is a positive bounded spectral measure.

If

$$\int ... \int |C(h)| dh < \infty$$

then G is automatically differentiable, $G(d\omega) = g(\omega)d\omega$ say, and (2.6) simplifies to

$$C(h) = \int ... \int \cos(\omega^T h) g(\omega) d\omega. \tag{2.7}$$

The necessary and sufficient condition for positive definiteness is then that $g(\omega) \geq 0$ for all ω.

If the process is isotropic ($C(h) = C_0(||h||)$ for some function C_0 of a univariate argument) then the spectral representation simplifies to

$$C_0(t) = \int_{(0,\infty)} Y_d(\omega t) \Phi(d\omega), \tag{2.8}$$

in terms of a function Φ which is nondecreasing on $[0, \infty)$ with $\int \Phi(d\omega) < \infty$, where

$$Y_d(t) = \left(\frac{2}{t}\right)^{(d-2)/2} \Gamma\left(\frac{d}{2}\right) J_{(d-2)/2}(t)$$

and $J_\nu(\cdot)$ denotes the Bessel function of the first kind of order ν.

These results are described in more detail in a number of standard monographs on spatial statistics, including Ripley (1981) and Cressie (1993).

This illustrates the most general strategy for constructing an isotropic stationary covariance function: use (2.8) with arbitrary nondecreasing Φ. Conversely, any conjectured covariance which cannot be written in this form cannot be positive definite and hence is not the covariance of a valid stationary process.

There is a corresponding theory for the variogram. Suppose $\gamma(\cdot)$ is the semivariogram of a second-order stationary process; then, by a combination of (2.2) and (2.5), if $a_1, ..., a_n$ are constants with $\sum a_i = 0$, we have

$$\sum_i \sum_j a_i a_j \gamma(s_i - s_j) \leq 0. \tag{2.9}$$

This is a *conditional non-positive-definiteness* condition. It can easily be seen that (2.9) is a necessary condition for $\gamma(\cdot)$ to be a valid semivariogram in the general (intrinsically stationary) case: the converse result is described in detail by Cressie (1993).

2.2 Estimation

Having now developed the main concepts of spatial covariances and variograms, we begin to consider their estimation. The general scenario is that we have a process $\{Z(s),\ s \in D\}$ observed at a finite number of points $s_1, ..., s_N$.

2.2.1 Estimating the variogram

The simplest estimator is the *method of moments* (MoM) estimator. In the case that the sampling points $s_1, ..., s_N$ lie on a regular lattice, this is defined by

$$2\widehat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(s_i, s_j) \in N(h)} \{Z(s_i) - Z(s_j)\}^2. \tag{2.10}$$

Here $N(h)$ denotes all pairs (s_i, s_j) for which $s_i - s_j = h$ and $|N(h)|$ denotes the cardinality of $N(h)$. In view of the lattice structure, $N(h)$ will either be empty or some reasonably sized subset of the set of all pairs of sampling points, the latter case applying if h is a vector spanning two points of the lattice. Of course, it only makes sense to estimate $\gamma(h)$ for such h vectors.

In the (far more common in practice) case where the points do not lie on a lattice, the same formula (2.10) is applied, but we change the definition of $N(h)$ to

$$N(h) = \{(s_i, s_j) :\ s_i - s_j \in T(h)\},$$

$T(h)$ being some small neighborhood or tolerance region around h.

The size of the tolerance region defining $T(h)$ raises issues similar to smoothing or optimal bandwidth choice in a variety of statistical applications. Journel and Huijbregts

(1978) recommended choosing $T(h)$ large enough to contain at least 30 pairs of points, and this can still be recommended as a rule of thumb.

One objection to the MoM method is that, like many methods based on sample averages, it is not robust against outlying values of Z. A more subtle objection arises from the skewness of the distribution: if we assume the process to be Gaussian, then for a specific s and h, the distribution of $\{Z(s+h) - Z(s)\}^2$ is of the form $2\gamma(h)\chi_1^2$, and the χ_1^2 distribution is highly skewed. However, if $X \sim \chi_1^2$, then $X^{1/4}$ has a nearly symmetric distribution (Fig. 2.3) so that we would expect sample averages of $|Z(s_1) - Z(s_2)|^{1/2}$ to be much better behaved than those of $\{Z(s_1) - Z(s_2)\}^2$. This idea lies at the heart of the proposal made by Cressie and Hawkins (1980). They suggested

$$2\bar{\gamma}(h) = \frac{1}{0.457 + 0.494/|N(h)|} \left\{ \frac{1}{|N(h)|} \sum_{(s_i,s_j)\in N(h)} |Z(s_i) - Z(s_j)|^{1/2} \right\}^4 \tag{2.11}$$

as an approximately unbiased estimator of $2\gamma(h)$.

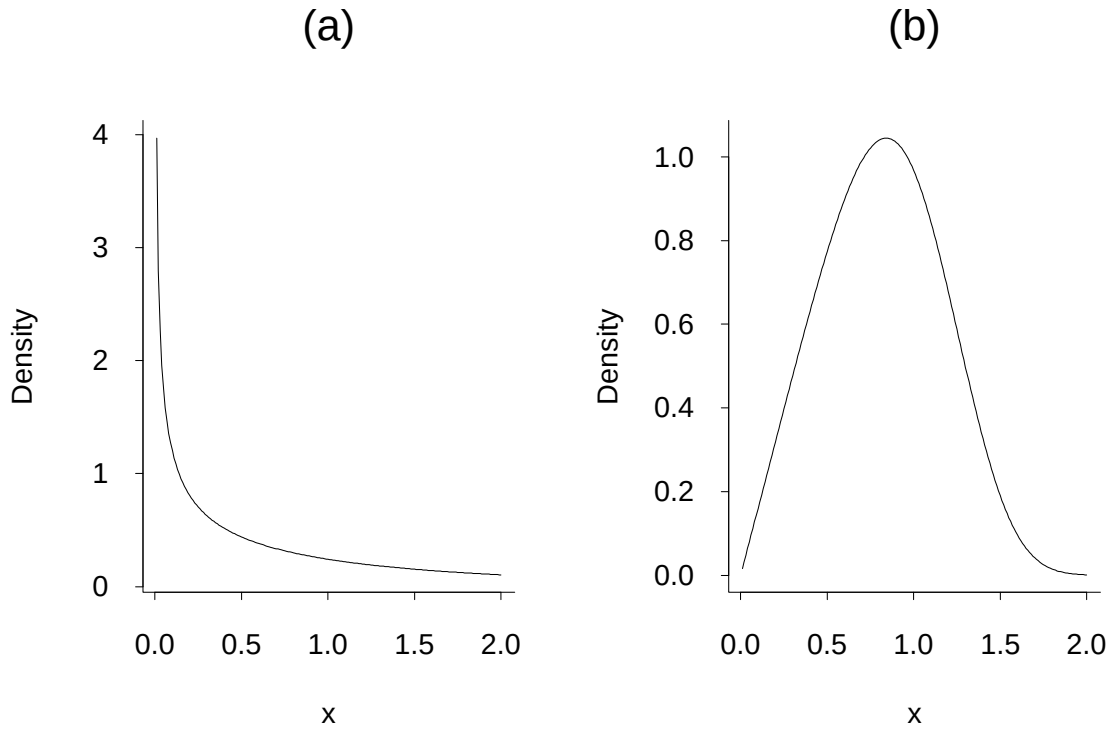


Fig. 2.3. (a) Density of X, (b) density of $X^{1/4}$, when $X \sim \chi_1^2$.

The first factor in (2.11) is a bias correction term, derived as follows. For $X \sim \chi_1^2$, it can be shown that $\mathrm{E}(X^{1/4}) = 2^{1/4}\Gamma(3/4)\pi^{-1/2} = 0.82216$, $\mathrm{var}(X^{1/4}) = 2^{1/2}\{\pi^{-1/2} - \Gamma^2(3/4)\pi^{-1}\} = 0.12192$. Hence the random variable

$$W_n = \frac{1}{n} \sum_{(s_i,s_j)\in N(h)} \frac{|Z(s_i) - Z(s_j)|^{1/2}}{\{2\gamma(h)\}^{1/4}}, \tag{2.12}$$

where $n = |N(h)|$, has mean ν =0.82216 and variance approximately $0.12192/n$. A delta-function argument shows that for twice continuously differentiable f,

$$\begin{aligned} \mathrm{E}\{f(W_n)\} - f(\nu) &= \mathrm{E}\{W_n - \nu\} f'(\nu) + \frac{1}{2}\mathrm{E}\{(W_n - \nu)^2\} f''(\nu) + \ldots \\ &\approx \frac{0.12192}{n} f''(\nu). \end{aligned}$$

Evaluating this for $f(x) = x^4$, we deduce

$$\begin{aligned} \mathrm{E}\{W_n^4\} &= \nu^4 + 6\nu^2 . \frac{0.12192}{n} + \ldots \\ &= 0.457 + \frac{0.494}{n} + \ldots \end{aligned}$$

From this and (2.12) it follows that (2.11) is an approximately unbiased estimator of $2\gamma(h)$.

A numerical study reported by Cressie (1993, pages 80–82) showed that, under a Gaussian model with added Gaussian noise containing 5% contamination, $\bar{\gamma}$ always has smaller bias than $\widehat{\gamma}$ and may also have smaller variance, the latter depending on the signal to noise variance ratio g (interpreting "signal" as the spatial variogram of the Gaussian model and "noise" as the added contaminated component). For large g, $\widehat{\gamma}$ has smaller variance than $\bar{\gamma}$, as might be expected from the fact that $\widehat{\gamma}$ is also the maximum likelihood estimator if we neglect contamination, but for small g this comparison is reversed. Based on these comparisons, Cressie recommended that both estimates be computed and compared. A possible counter-argument is that if $Z(s)$ has a marginal distribution that is far from normal (though not necessarily with any outlier contamination), the whole argument leading to (2.11) as an approximately unbiased estimator would appear to be invalid.

Another possible "robust" estimator is given by

$$2\tilde{\gamma}(h) = \frac{\mathrm{Median}[\{Z(s_i) - Z(s_j)\}^2 : \ (s_i, s_j) \in N(h)]}{0.457}. \tag{2.13}$$

One motivation for (2.13) may be seen by rewriting it in the form

$$2\tilde{\gamma}(h) = \frac{[\mathrm{Median}\{|Z(s_i) - Z(s_j)|^{1/2} : \ (s_i, s_j) \in N(h)\}]^4}{0.457}$$

which corresponds to taking a median instead of a mean in (2.11). Because of the approximate symmetry of the $(\chi_1^2)^{1/4}$ random variable, its mean and median are nearly the same, and the constant 0.457 is derived in the same way as in (2.11), as a first-order bias-correction factor. Another motivation of (2.13) is that it is, modulo a multiplicative factor, the squared interquartile range of $\{Z(s_i) - Z(s_j), \ (s_i, s_j) \in N(h)\}$, which might also be thought of as a natural robust measure of the scale of $Z(s+h) - Z(s)$ over $s \in D$

for fixed h. However, according to Cressie (1993, page 77), it appears that $\bar{\gamma}$ is a more efficient estimator than $\tilde{\gamma}$.

Examples. In chapter 1, we have already seen variograms computed for four meteorological variables and four regions of the USA, by both the MoM and robust methods In each of Figs. 1.11–1.14, the MoM estimate $\widehat{\gamma}$ is on the left hand side and the robust estimate $\bar{\gamma}$ is on the right hand side. The two estimates seem to be rather similar except in the case of annual maximum precipitations (Fig. 1.13), for which the MoM estimate is generally larger than the robust estimate. This is to be expected, because in this case we saw that the distribution is indeed affected by outliers, so one would expect the two estimates to behave in different ways.

Another way to compute these plots is as a "variogram cloud". This method of computing the variogram is available when there are multiple replications of the spatial field. This assumption is satisfied for the data in chapter 1 as we had many years of data which, at least for the present analysis, we are treating as independent from year to year.

In the variogram cloud, one point is plotted for each pair of stations s_i and s_j. The distance between stations s_i and s_j, d_{ij} say, is plotted along the x axis, and an estimate of $\mathrm{Var}\{Z(s_i) - Z(s_j)\}$ is plotted along the y axis. For the latter, we may use either the MoM or the robust method. Recall that the Z values we are using for this comparison are not the raw data but are standardized residuals from a linear regression in time, so the sample means and standard deviations at each station have already been adjusted to be 0 and 1 respectively.

Figs. 2.4 and 2.5 show the variogram clouds for the data of chapter 1, computed for the winter mean daily minimum temperatures (analogous to Fig. 1.11) and the annual maximum daily precipitations (Fig. 1.13). As can be seen, the scatter in the plots is very great, calling into question whether these are homogeneous spatial processes. Our present focus, however, is on the comparison of the MoM and robust estimates, and one way to look at this is directly, by plotting one against the other. Fig. 2.6 shows a plot of robust vs. MoM estimates, for each of the four subdivisions of the USA, corresponding to the variogram clouds in Fig. 2.4. Fig. 2.7 shows the same things computed for the variogram clouds in Fig. 2.5. The 45^o line through the origin is shown to provide a comparison between the two estimates.

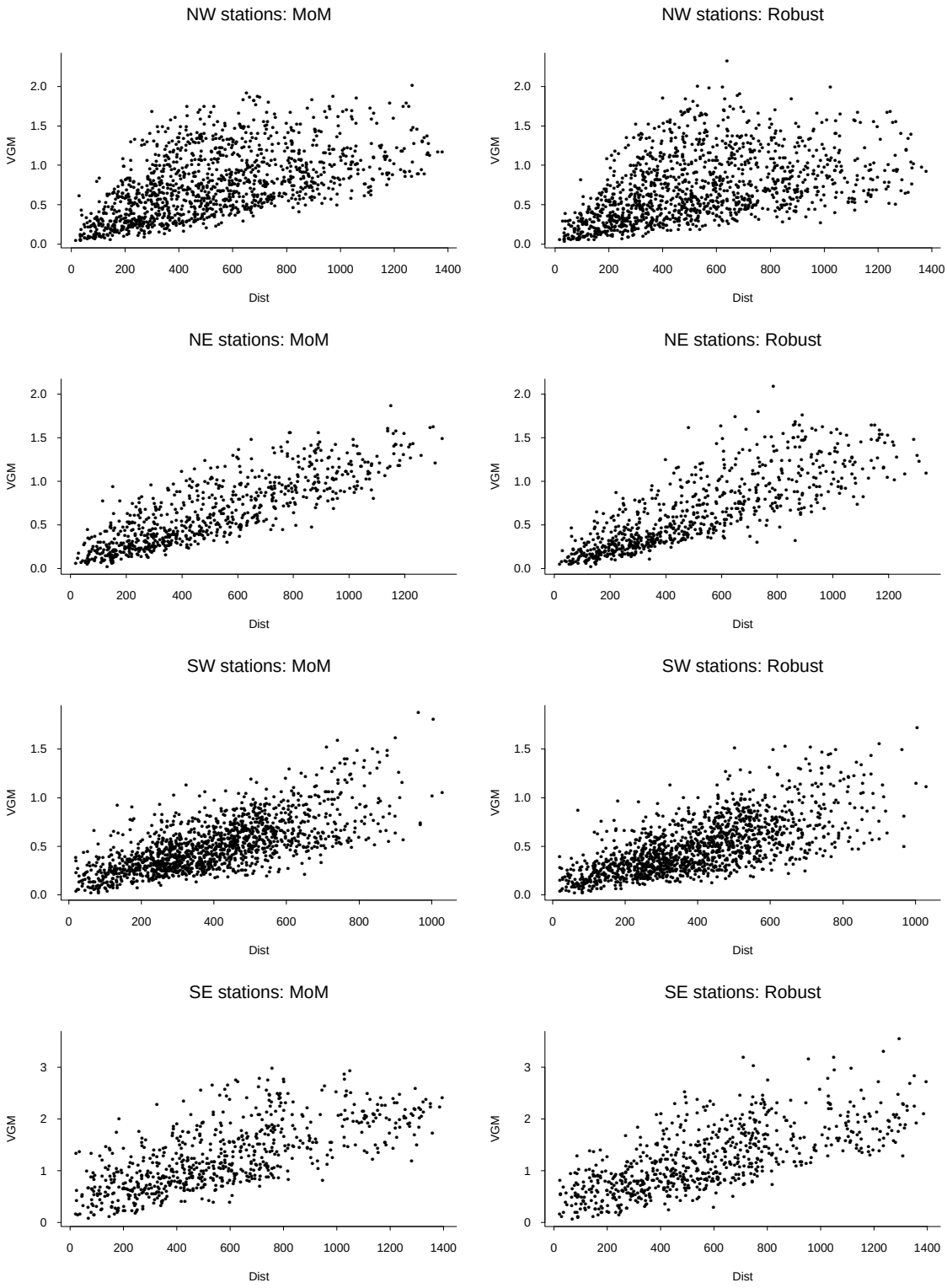


Fig. 2.4. Variogram cloud plots for the data based on mean winter minimum daily temperature.

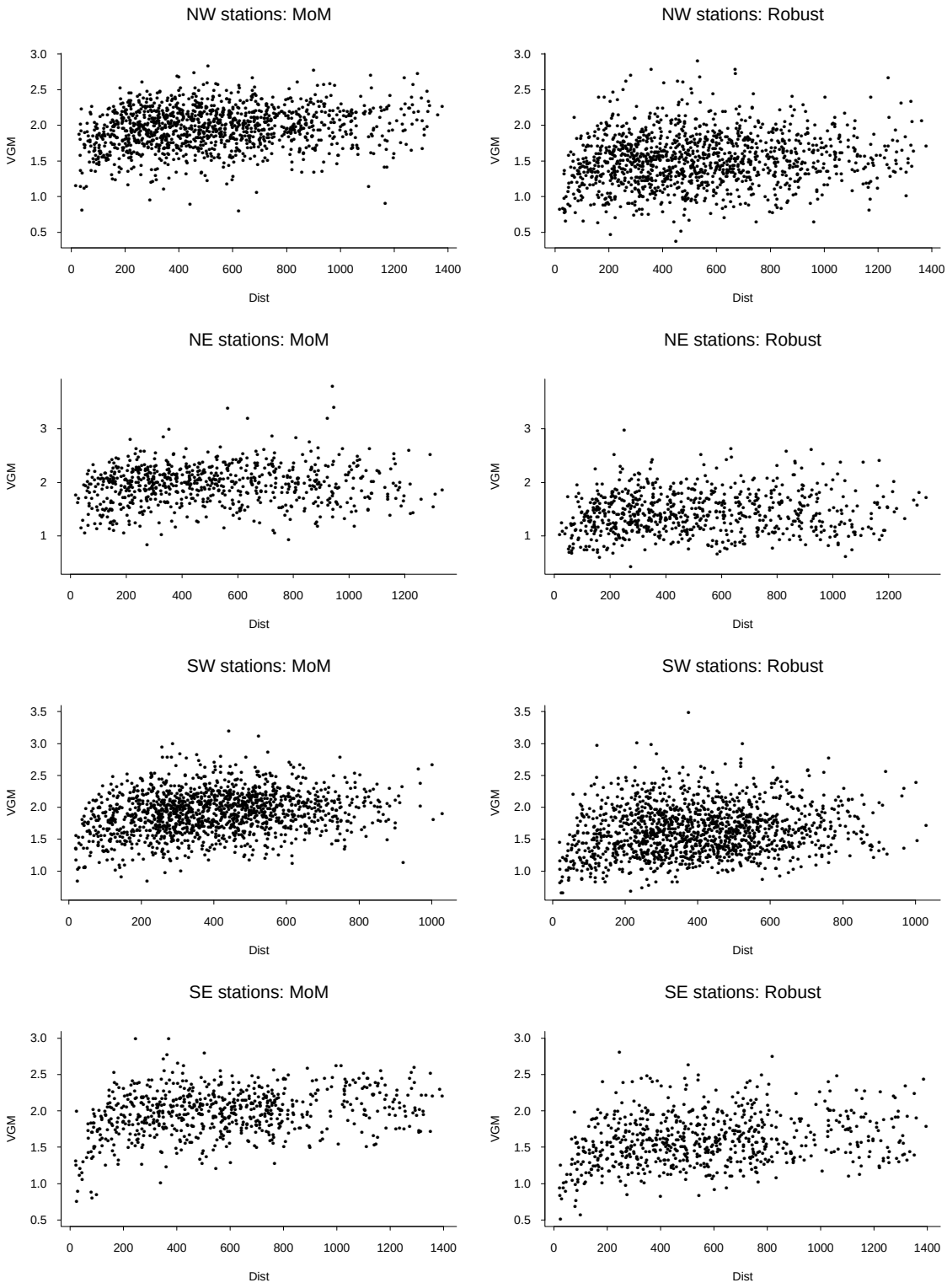


Fig. 2.5. Variogram cloud plots for the data based on annual maximum daily precipitation.

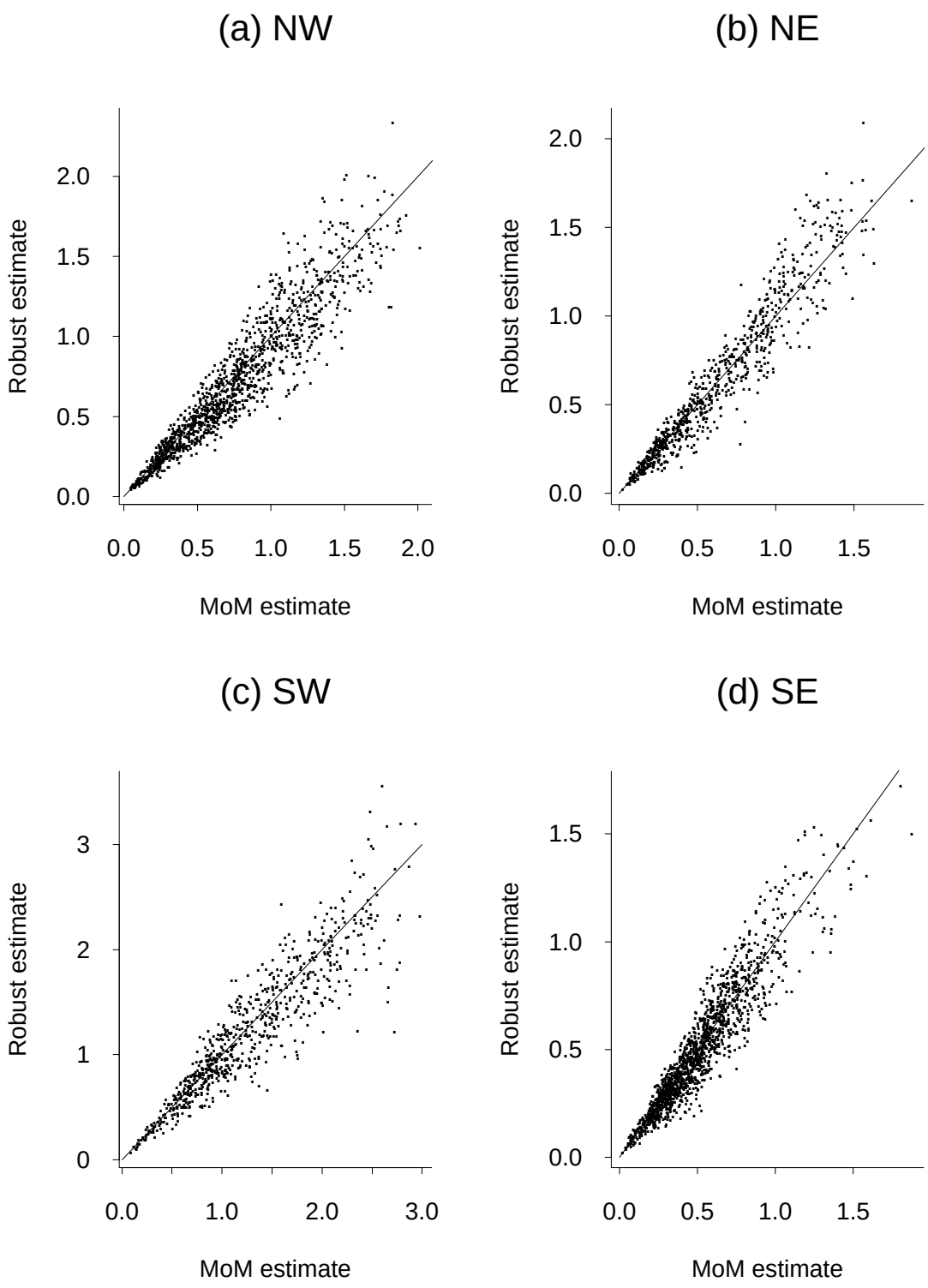


Fig. 2.6. Plot of robust vs. MoM estimators for the variogram cloud in Fig. 2.4.

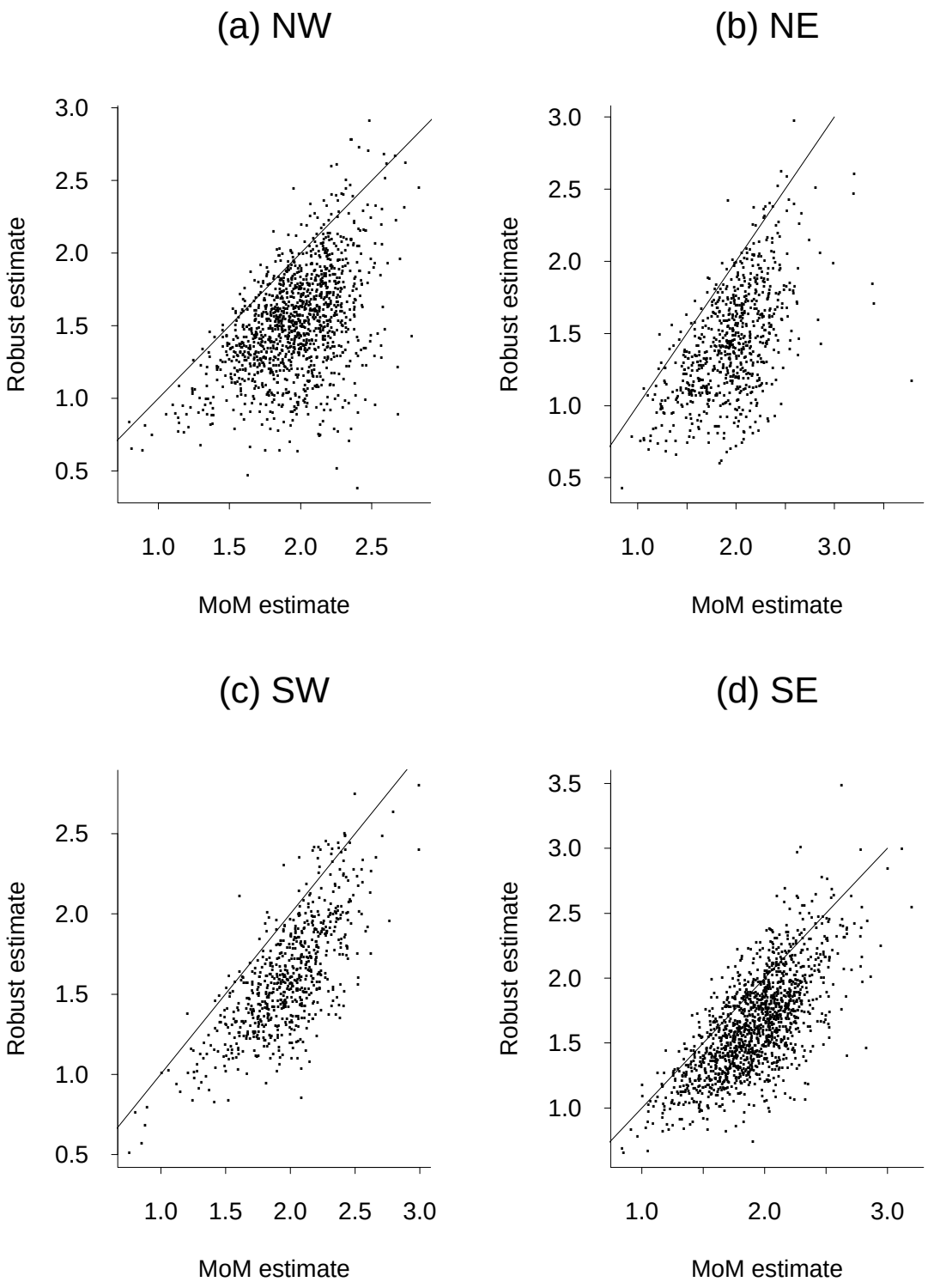


Fig. 2.7. Plot of robust vs. MoM estimators for the variogram cloud in Fig. 2.5.

From Fig. 2.6 it can be seen that the estimates are in good agreement. The scatterplot is tightly clustered around the 45^o line and and the correlation between the MoM and robust estimates is good.

Fig. 2.7 tells a completely different story. In this case, the bulk of the estimates lie below the 45^o line, indicating that the robust estimate is smaller than the MoM estimate. We also observe that there is much greater variability in the scatterplot — in this case, the two estimates seem almost uncorrelated.

Our conclusion is that for a series for which the marginal values are close to normally distributed, as appears to be the case for the data in Fig. 2.4, it makes little difference which estimator is computed. However, for a highly skewed marginal distribution, as in Fig. 2.5, it does make a big difference. It remains open to discussion which is the "correct" estimator for an example like this one, given that ultimately our real interest is in the behavior of the most extreme rainfalls. However, given the tendency of the MoM estimator to be greatly affected by even a small number of outlying values, it is probably more reasonable to use the robust estimator as an indicator of what is going on in the bulk of the distribution, while acknowledging the need for alternative measure to characterize spatial dependence in the extreme values of the process.

2.2.2 Inspecting the variogram cloud for homogeneity

An alternative issue, briefly touched on above, concerns the homogeneity of the process, i.e. the assumption that the spatial process is both stationary and isotropic. It is possible to superimpose the two types of variogram plot in a single figure, as is done in Fig. 2.8 for the NW stations in the winter daily minimum temperature plots — in other words, Fig. 2.8 superimposes the top-left-hand plots of Figs. 1.11 and 2.4. The dots represent the points of the variogram cloud, while the circles represent averaged values of the variogram cloud over subintervals on the distance scale, which is exactly how Fig. 1.11 was computed. The boundaries of the distance subintervals are also indicated, represented as vertical lines on the plot.

Each point of the variogram cloud (i.e. each dot in Fig. 2.7) is the variance between two stations computed from 32 years' data, while the corresponding binned values (the circles in the figure) are averages of all the variogram cloud points within each distance subinterval. One can ask whether the data would support a hypothesis that all the variances within a single bin are equal. It might be possible to test this using, for example, Bartlett's test for the equality of variances in several independent samples. The assumptions of Bartlett's test are not strictly satisfied — for example, the different pairs of stations leading to different dots of the variogram cloud are not actually independent. Nevertheless one might expect that within a single narrow bin, Bartlett's test would give reasonably reliable answers. In this case, it quickly becomes apparent that even crude tests of homogeneity within each distance bin lead to decisive rejection of the null hypothesis, at all but the very largest distances. The process we are sampling from is not spatially homogeneous.

The same conclusion applies to the other three subdivisions of the continental USA and to the other meteorological variables.

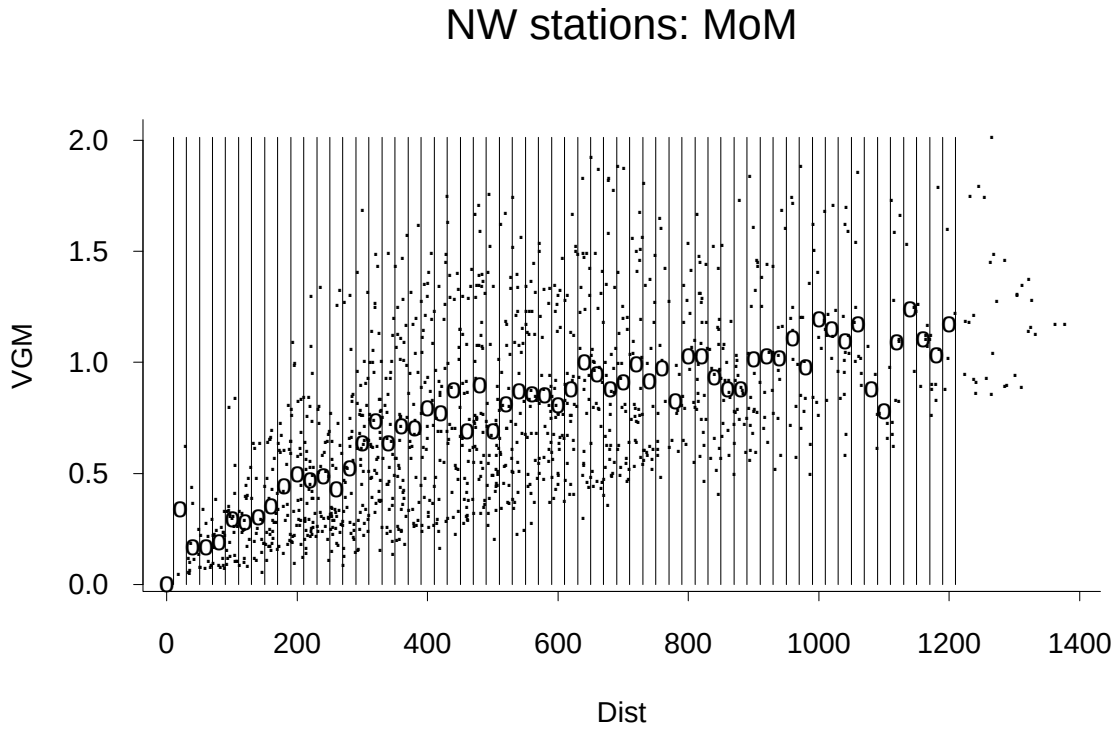


Fig. 2.8. Two forms of variogram plot superimposed.

Given the conclusion that these processes are not spatially homogeneous, we must be cautious in our interpretation of Figs. 1.11–1.14 and 2.4–2.5. The estimated variogram is not necessarily a valid measure of the variance between any two individual stations, but instead, represents an averaged value over the region. Given this alternative interpretation, however, comparisons between the variograms still seem to be justified — for example, temperature averages are correlated across very wide spatial scales, but the range of spatial dependence for precipitation maxima is much smaller. Meanwhile, we defer detailed consideration of inhomogeneous spatial processes to chapter 3.

More details of the calculations

Consider the variogram cloud points corresponding to distances between 590 and 610 nautical miles. We can assume that the ith pair corresponds to locations s_{i1}, s_{i2} and that we have observations $Y_{ij} = Z(s_{i1}, t_{ij}) - Z(s_{i2}, t_{ij})$ for time points t_{ij} at which both observations $Z(s_{i1}, t_{ij}), Z(s_{i2}, t_{ij})$ exist. Because of missing values in the original data set, not all $Z(s,t)$ points are well-defined. Table 2.1 shows values n_i, $\widehat{\sigma}_i^2$ where n_i is the number of observations Y_{ij} and $\widehat{\sigma}_i^2 = S_i/(n_i - 1)$, $S_i = \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$, where $\bar{Y}_{i\cdot}$ is the mean of the values of Y_{ij} as j varies.

i	n_i	$\widehat{\sigma}_i^2$	i	n_i	$\widehat{\sigma}_i^2$	i	n_i	$\widehat{\sigma}_i^2$	i	n_i	$\widehat{\sigma}_i^2$
1	29	1.14082	2	30	.88010	3	29	.91791	4	29	.49816
5	31	.76427	6	31	.63453	7	31	1.46821	8	32	1.22803
9	31	.37191	10	31	.65614	11	31	1.32307	12	29	.56116
13	25	.66362	14	32	1.26053	15	31	.42792	16	24	.67272
17	32	.38503	18	30	.48921	19	32	.84835	20	31	.76461
21	31	.45127	22	30	.48219	23	32	.67729	24	32	1.30333
25	32	.80796	26	32	.60602	27	29	.75231	28	29	.59545
29	32	.77198	30	28	1.34613	31	29	1.15666			

Table 2.1. 31 variogram cloud estimates for pairs of stations corresponding to distances between 590 and 600 N.M. The ith variogram cloud estimate $\widehat{\sigma}_i^2$ is based on n_i pairs of observations.

Standard tests for equality of variances are the likelihood ratio test and Bartlett's modification (see, e.g. Kendall and Stuart (1979), section 24.9). For the likelihood ratio test, we work not with $\widehat{\sigma}_i^2$ but with $\tilde{\sigma}_i^2 = S_i/n_i$; we then define an overall variance $\tilde{\sigma}^2 = \sum S_i/N$ where $N = \sum n_i$, and

$$T = \sum_{i=1}^{r} n_i \log\left(\frac{\tilde{\sigma}^2}{\tilde{\sigma}_i^2}\right),$$

where r is the number of groups (here, 31). According to asymptotic theory, under the null hypothesis that the true variances σ_i^2 are all equal, the distribution of T is approximately χ^2_{r-1}.

Bartlett's modification uses $\widehat{\sigma}_i^2$ in place of $\tilde{\sigma}_i^2$, $\widehat{\sigma}^2 = \sum S_i/(N-r)$ (recomputing T with these values), and then defines

$$T' = \left\{1 + \frac{1}{3(r-1)} \sum_{i=1}^{r} \left(\frac{1}{n_i - 1} - \frac{1}{N-r}\right)\right\}^{-1} T.$$

The distributional approximation is again χ^2_{r-1}, but the distribution of T' is believed to be closer to this than that of T.

In the present example, we find $\tilde{\sigma}^2 = .778$, $T = 71.8$, $\widehat{\sigma}^2 = .805$, $T = 65.1$. The value of $\widehat{\sigma}^2$ is the one used for the superimposed plot with the circles in Fig. 2.7. Based on $T \sim \chi^2_{30}$, we reject the null hypothesis of homogeneity with a p-value of .00003. Based on $T' \sim \chi^2_{30}$, we reject the null hypothesis of homogeneity with a p-value of .00021. Either way, the result points to overwhelming rejection of the null hypothesis.

Similar results are obtained for the vast majority of the vertical bins in Fig. 2.8.

Fig. 2.9. Subdivision of stations (inside dashed rectangle) used for calculations of Table 2.2.

As a second example of these calculations, consider the subset of stations enclosed by the dashed lines in Fig. 2.9. These are 17 stations with latitude between 40 and 45 °N and longitudes between 90 and 100 °W. This is a region for which one might anticipate the process would be reasonably homogeneous. The calculations of T', and the associated p-values, are shown for a variety of distances in Table 2.2. The results suggest that the homogeneity assumption is not good at short distances (less than 140 N.M.), but at longer distances is reasonable.

Distance (N.M.)	T'	$r-1$	p-value
35	27.6	5	.00002
70	30.2	10	.0004
105	15.4	8	.03
140	17.0	14	.20
175	10.8	11	.37
210	13.9	14	.38
245	13.3	15	.51
280	26.2	18	.07
315	23.1	14	.04
350	5.6	6	.35

Table 2.2. Table of T' values for homogeneity test; 17 stations in latitudes 40–45 °N and longitudes 90–100 °W.

2.2.3 Fitting parametric models to the sample variogram

In this section we again assume we are sampling from a homogeneous spatial process, in which the variogram has been estimated for a sequence of distances h by one of the methods of section 2.2.1.

Although the properties of the semivariogram estimators $\widehat{\gamma}(h)$, $\bar{\gamma}(h)$ and $\tilde{\gamma}(h)$ have been extensively investigated for a single value of h, as a function over all h they all lack a very important property: they fail the conditional non-positive-definiteness condition mentioned at the end of section 2.1. Thus it is possible that spatial predictions derived from such estimators will appear to have negative variances. The most common way of avoiding this difficulty is to replace the empirical $\gamma(h)$ by some parametric form which is known to be conditionally non-positive-definite, such as one of the families listed in section 2.1. It may well be considered desirable on general statistical modeling grounds to seek a parametric family which adequately models the observed data, but this provides an additional and specific motivation to do that. Note that in general there is no need to restrict ourselves to isotropic models, though it is usually convenient to consider isotropic models first.

Three methods will be considered:

- Least squares estimation,

- maximum likelihood (ML) and restricted maximum likelihood (REML),

- Bayesian estimators.

In the present subsection we concentrate on the first of these, the ML, REML and Bayesian methods being deferred to subsections 2.2.4–2.2.6.

Suppose we have estimated the semivariogram $\gamma(h)$ at a finite set of values of h, and wish to fit a model specified by the parametric function $\gamma(h;\theta)$ in terms of a finite parameter vector θ. This could, for instance, be any of the isotropic forms considered in Section 2.1 where θ contains the nugget, sill and range and any incidental parameters. For definiteness we shall assume the MoM estimator $\widehat{\gamma}$ has been used and let $\widehat{\boldsymbol{\gamma}}$ denote the vector of estimates, $\boldsymbol{\gamma}(\theta)$ the vector of model values at the same vector of h values.

There are three well-used versions of non-linear least-squares estimators:

- *Ordinary least squares* or OLS, in which we choose θ to minimize

$$\{\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\theta)\}^T \{\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\theta)\}.$$

- *Generalized least squares* or GLS, in which we choose θ to minimize

$$\{\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\theta)\}^T V(\theta)^{-1} \{\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\theta)\}.$$

Here $V(\theta)$ denotes the covariance matrix of $\widehat{\boldsymbol{\gamma}}$ which, since the problem is non-linear, depends on the unknown θ.

- *Weighted least squares* or WLS, in which we choose θ to minimize

$$\{\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\theta)\}^T W(\theta)^{-1} \{\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\theta)\}.$$

In this case $W(\theta)$ is a diagonal matrix whose diagonal entries are the variances of the entries of $\widehat{\boldsymbol{\gamma}}$. Thus WLS allows for the variances of $\widehat{\boldsymbol{\gamma}}$ but not the covariances, while GLS allows for both.

In general, we expect the three estimators OLS, WLS, GLS to be in increasing order of efficiency but in decreasing order of convenience to use. Note, in particular, that OLS is immediately implementable by a nonlinear least squares procedure, whereas WLS and GLS require specification of the matrices $W(\theta)$ and $V(\theta)$.

For a Gaussian process, however, we have the expressions

$$\mathrm{var}[\{Z(s+h) - Z(s)\}^2] = 2\{2\gamma(h)\}^2, \tag{2.14}$$

$$\begin{aligned}&\mathrm{corr}[\{Z(s_1+h_1) - Z(s_1)\}^2, \{Z(s_2+h_2) - Z(s_2)\}^2] \\ &= \frac{\{\gamma(s_1-s_2+h_1) + \gamma(s_1-s_2-h_2) - \gamma(s_1-s_2+h_1-h_2) - \gamma(s_1-s_2)\}^2}{4\gamma(h_1)\gamma(h_2)}\end{aligned} \tag{2.15}$$

which may be used to evaluate the matrices $W(\theta)$ and $V(\theta)$. The derivations of (2.14) and (2.15) will be given below. Thus GLS is possible in principle, but complicated to implement. For example, there is no guarantee that the resulting minimization problem has a unique solution.

As a compromise, Cressie (1985) proposed the approximate WLS criterion: if $\widehat{\gamma}$ is evaluated on a finite set $\{h_j\}$, choose θ to minimize

$$\sum_j |N(h_j)| \left\{ \frac{\widehat{\gamma}(h_j)}{\gamma(h_j;\theta)} - 1 \right\}^2 . \tag{2.16}$$

Note that (2.16) may be derived as the WLS solution under the approximation

$$\mathrm{var}\{\widehat{\gamma}(h)\} \approx \frac{8\gamma^2(h)}{|N(h)|}$$

which follows from (2.14) if we assume that the individual $Z(s_i) - Z(s_j)$ terms, which form the numerator of (2.10), are independent. This assumption of course is not exactly satisfied but may nevertheless be a reasonable approximation if the pairs (s_i, s_j) lying in $N(h)$ are widely spread over the sampling space. The criterion (2.16) is no more difficult to implement than OLS, and may be expected to be substantially more efficient, while avoiding the complications of GLS.

The discussion so far presupposes the MoM estimator $\widehat{\gamma}$, but the criterion based on (2.16) also makes sense with the robust estimator $\bar{\gamma}$.

** *Derivation of (2.14) and (2.15)*

Under the assumption that everything is normally distributed, we have

$$\frac{\{Z(s+h) - Z(s)\}^2}{2\gamma(h)} \sim \chi_1^2$$

which has mean 1 and variance 2. (2.14) follows at once from this.

To derive (2.15), we first derive

$$\begin{aligned} &\mathrm{cov}\{(Z(s_1+h_1) - Z(s_1))^2, (Z(s_2+h_2) - Z(s_2))^2\} \\ &= 2\{\gamma(s_1+h_1-s_2-h_2) - \gamma(s_1-s_2-h_2) - \gamma(s_1+h_1-s_2) + \gamma(s_1-s_2)\}^2. \end{aligned} \tag{2.17}$$

To see this, write $Y_1 = Z(s_1+h_1) - Z(s_1)$, $Y_2 = Z(s_2+h_2) - Z(s_2)$. Note that Y_1 and Y_2 each have mean 0, that their variances are $2\gamma(h_1) = 2\gamma_1$ say and $2\gamma(h_2) = 2\gamma_2$, and that their covariance is

$$\begin{aligned} &C(s_1+h_1-s_2-h_2) - C(s_1-s_2-h_2) - C(s_1+h_1-s_2) + C(s_1-s_2) \\ &= -\gamma(s_1+h_1-s_2-h_2) + \gamma(s_1-s_2-h_2) + \gamma(s_1+h_1-s_2) - \gamma(s_1-s_2) \\ &= c \text{ say} \end{aligned}$$

using (2.2).

Let us also write $Y_1 = \sqrt{2\gamma_1}W_1$, $Y_2 = a_1W_1 + a_2W_2$ where W_1 and W_2 are independent standard normal random variables. We derive the constants a_1 and a_2 by matching up to the variance of Y_2 and the covariance of Y_1 and Y_2: this leads to

$$a_1 = \frac{c}{\sqrt{2\gamma_1}}, \qquad a_2 = \sqrt{2\gamma_2 - \frac{c^2}{2\gamma_1}}.$$

We want to calculate the covariance of Y_1^2 and Y_2^2. However,

$$\mathrm{E}\{Y_1^2\} = 2\gamma_1, \qquad \mathrm{E}\{Y_2^2\} = 2\gamma_2, \tag{2.18}$$

and

$$\begin{aligned} \mathrm{E}\{Y_1^2Y_2^2\} &= \mathrm{E}\left\{2\gamma_1 W_1^2(a_1^2W_1^2 + 2a_1a_2W_1W_2 + a_2^2W_2^2)\right\} \\ &= 6\gamma_1a_1^2 + 2\gamma_1a_2^2 \\ &= 3c^2 + 2\gamma_1\left(2\gamma_2 - \frac{c^2}{2\gamma_1}\right) \\ &= 2c^2 + 4\gamma_1\gamma_2 \end{aligned} \tag{2.19}$$

where in moving from the first to the second lines of this equation we used that $\mathrm{E}\{W_1^4\} = 3$. The result (2.17) then follows by combining (2.18) and (2.19).

Finally, combining (2.17) and (2.14) allows one to calculate the correlation of Y_1^2 and Y_2^2; this is (2.15).

** *Standard errors; Asymptotics*

There appears to have been no discussion in the literature of how to compute standard errors of the parameters estimated by the WLS criterion, or any of the other least squares approaches. Nevertheless it is an important question to consider, for example, in determining whether two variograms fitted to different regions or different time periods are significantly different from one another.

We therefore present a tentative approach here. This, inevitably, gets us into some discussion of asymptotics, which we shall approach in a somewhat heuristic fashion. However, it should be emphasised that the kind of asymptotics considered are *increasing domain asymptotics*, which apply when the region of study is increased with the underlying density of sampling points being constant. The alternative, *infill asymptotics*, which applies as more and more stations are added to a bounded region, leads to rather different results.

In presenting asymptotic results, we use the notation $\to_p$ to denote convergence in probability, $\to_d$ to denote convergence in distribution, and $O_p(\cdot)$ to denote order in probability: if $\{X_n\}$ is a sequence of random variables and $\{a_n\}$ a sequence of positive constants, then $X_n = O_p(a_n)$ means

$$\lim_{M\to\infty} \limsup_{n\to\infty} \Pr\left\{ \left| \frac{X_n}{a_n} \right| > M \right\} = 0.$$

We also use $\mathcal{N}(\mu, \Sigma)$ to denote the multivariate normal distribution with mean μ and covariance matrix Σ.

** *An aside: the "information sandwich" formula*

First we present a general discussion of an asymptotic technique that has come to be known as the "information sandwich" approach.

Suppose we have a statistical model indexed by a finite-dimensional parameter θ, and suppose an estimate $\tilde{\theta}_n$ is constructed by minimizing a criterion function $S_n(\theta)$. The parameter n here is just an index which we are going to let tend to ∞; in most cases, however, n will represent the sample size, and $S_n(\theta)$ will denote some "measure of fit" such as a sum of squares, a likelihood or a quasi-likelihood. We assume the true parameter value is θ_0 and that $\tilde{\theta}_n$ is a consistent estimator. We shall also assume that $S_n(\theta)$ is at least twice continuously differentiable in θ, and that its underlying distribution is sufficiently smooth that the function $H(\theta)$ (defined below) is also continuous in a neighborhood of θ_0. Let $\nabla f(\theta)$ for any f denote the vector of first-order partial derivatives of f with respect to the components of θ, and $\nabla^2 f$ the matrix of second-order partial derivatives.

By a Taylor expansion, we have

$$0 = \nabla S_n(\tilde{\theta}_n) = \nabla S_n(\theta_0) + \nabla^2 S_n(\theta_n^*)(\tilde{\theta}_n - \theta_0)$$

where θ_n^* lies on the straight line joining $\tilde{\theta}_n$ to θ_0. Hence

$$\tilde{\theta}_n = \theta_0 - \{\nabla^2 S_n(\theta_n^*)\}^{-1} \nabla S_n(\theta_0). \tag{2.20}$$

We assume

(A1) $\frac{1}{n}\nabla^2 S_n(\theta) \to_p \ H(\theta)$ as $n \to \infty$ uniformly on some neighborhood of θ_0, where $H(\cdot)$ is a matrix-valued function, continuous near θ_0, with $H(\theta_0)$ invertible,

(A2) $\frac{1}{\sqrt{n}}\nabla S_n(\theta_0) \to_d \ \mathcal{N}(0, V(\theta_0))$ for some covariance matrix $V(\theta_0)$.

Assumptions (A1) and (A2) are satisfied for regular maximum likelihood problems in which $S_n(\theta)$ is the negative log likelihood for the parameter θ, since this then consists of

a sum over n i.i.d. terms, but they are also valid much more generally for a wide variety of estimation criteria.

Since $H(\theta)$ is continuous in θ and invertible at θ_0, it follows that $H(\theta)^{-1}$ is continuous near θ_0, and hence that

$$\left\{ \frac{1}{n} \nabla^2 S_n(\theta_n^*) \right\}^{-1} \to_p \; H(\theta_0)^{-1}. \tag{2.21}$$

Equations (2.20) and (2.21) may then be combined (using the Slutsky lemma) to conclude that

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d \; \mathcal{N}(0, H(\theta_0)^{-1} V(\theta_0) H(\theta_0)^{-1}). \tag{2.22}$$

In maximum likelihood theory, the V and H both define the Fisher information matrix so (2.22) is just a restatement of the well-known asymptotic normality of the maximum lieklihood estimator under regular conditions. In general, however, V and H are not the same, and the phrase *information sandwich* has been coined to describe the matrix $H^{-1}VH^{-1}$.

** *Asymptotic theory for the WLS estimator: Preliminaries*

Suppose the semivariogram $\gamma(h)$ is evaluated on a finite set of h values, say $h_1, ..., h_J$; we assume this set is fixed throughout the discussion. Suppose the region increases to infinity but in such a way that the overall density of points remains approximately constant (and bounded away from 0 and ∞). Then each of the subsets $N(h_j)$ (of pairs of points whose distance apart is within a specified tolerance of h_j) will grow approximately proportionally to n. Denoting the n'th such set by $N_n(h_j)$, we may assume that

$$|N_n(h_j)| = n\phi_{n,j}, \quad \lim_{n\to\infty} \phi_{n,j} = \phi_j,$$

where each ϕ_j is positive and finite, $j = 1, ..., J$. Thus we may define

$$S_n(\theta) = \sum_j n\phi_{n,j} \left\{ \frac{\widehat{\gamma}_n(h_j)}{\gamma(h_j;\theta)} - 1 \right\}^2 \tag{2.23}$$

with $\widehat{\gamma}_n(h_j)$ the estimated semivariogram in the n'th sample. Recalling that

$$2\widehat{\gamma}_n(h_j) = \frac{1}{n\phi_{n,j}} \sum_{s_j:\ (s_j, s_j+h_j)\in N_n(h_j)} \{Z(s_j+h_j) - Z(s_j)\}^2,$$

under Gaussianity assumptions we have by (2.17)

$$\mathrm{cov}\{\widehat{\gamma}_n(h_j), \widehat{\gamma}_n(h_k)\} = \frac{2}{n^2\phi_{n,j}\phi_{n,k}} \sum_{s_j:\ (s_j,s_j+h_j)\in N_n(h_j)} \sum_{s_k:\ (s_k,s_k+h_k)\in N_n(h_k)} \cdot$$
$$\cdot \{\gamma(s_j - s_k + h_j) + \gamma(s_j - s_k - h_k) - \gamma(s_j - s_k + h_j - h_k) - \gamma(s_j - s_k)\}^2.$$

As $n \to \infty$, under some regularity assumptions which we will not attempt to make precise, the magnitude of this expression is of $O(1/n)$. This is because the number of s_j terms in the summand will be of $O(n)$ reflecting the rate of growth of $N_n(h_j)$, but for each fixed s_j, the *total* contribution over all s_k will be bounded, if we assume the covariances decay sufficiently fast. All the covariance models we have considered have covariances decaying at least exponentially fast as the distance between the two points tends to ∞, so this assumption will be satisfied.

Consequently, for asymptotic arguments, we may assume that there exists a $J \times J$ matrix $W(\theta)$ with entries $(w_{jk}(\theta))$, depending on the unknown θ, such that

$$\mathrm{cov}\{\widehat{\gamma}_n(h_j), \widehat{\gamma}_n(h_k)\} \sim \frac{w_{jk}(\theta)}{n} \qquad \text{as } n \to \infty$$

and consequently, if we write $\widehat{\boldsymbol{\gamma}}_n$ for the vector with components $\widehat{\gamma}_n(h_1), ..., \widehat{\gamma}_n(h_J)$, and $\boldsymbol{\gamma}(\theta)$ for the corresponding vector of theoretical values, that under the model with parameter vector θ,

$$\sqrt{n}\{\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}(\theta)\} \to \mathcal{N}(0, W(\theta)). \tag{2.24}$$

** *Application of the information sandwich formula to the distribution of the WLS estimator*

Defining $S_n(\theta)$ by (2.23), we have

$$\nabla S_n(\theta) = 2n \sum_j \phi_{n,j} \left\{ -\frac{\widehat{\gamma}_n^2(h_j)}{\gamma^3(h_j;\theta)} + \frac{\widehat{\gamma}_n(h_j)}{\gamma^2(h_j;\theta)} \right\} \cdot \nabla\gamma(h_j;\theta), \tag{2.25}$$

$$\begin{aligned} \nabla^2 S_n(\theta) = {} & 2n \sum_j \phi_{n,j} \left\{ \frac{3\widehat{\gamma}_n^2(h_j)}{\gamma^4(h_j;\theta)} - \frac{2\widehat{\gamma}_n(h_j)}{\gamma^3(h_j;\theta)} \right\} \cdot \nabla\gamma(h_j;\theta)\nabla\gamma(h_j;\theta)^T \\ & + 2n \sum_j \phi_{n,j} \left\{ -\frac{\widehat{\gamma}_n^2(h_j)}{\gamma^3(h_j;\theta)} + \frac{\widehat{\gamma}_n(h_j)}{\gamma^2(h_j;\theta)} \right\} \cdot \nabla^2\gamma(h_j;\theta). \end{aligned} \tag{2.26}$$

Consider (2.26). In the first term, by consistency of $\widehat{\gamma}_n$, the expression inside parentheses tends to $1/\gamma^2(h_j;\theta)$, while the $\phi_{n,j}$ and $\nabla\gamma$ terms are each of $O(1)$; consequently the whole expression is of $O_p(n)$. In the second term, however, the size of the expression inside parentheses depends on the distance from $\widehat{\gamma}_n$ to γ; since this is of $O_p(1/\sqrt{n})$, the whole second term is of $O_p(\sqrt{n})$, and hence negligible compared with the first term. This leads us to conclude

$$\begin{aligned} \frac{1}{n}\nabla^2 S_n(\theta) \to_p \; & 2\sum_j \frac{\phi_j}{\gamma^2(h_j;\theta)} \nabla\gamma(h_j;\theta)\nabla\gamma(h_j;\theta)^T \\ & = H(\theta) \text{ say.} \end{aligned} \tag{2.27}$$

Similarly, (2.25) may be rearranged to give

$$\nabla S_n(\theta) = 2n \sum_j \frac{\phi_{n,j}}{\gamma^2(h_j;\theta)} \{\gamma(h_j;\theta) - \widehat{\gamma}_n(h_j)\} \nabla\gamma(h_j;\theta)$$
$$- 2n \sum_j \frac{\phi_{n,j}}{\gamma^3(h_j;\theta)} \{\gamma(h_j;\theta) - \widehat{\gamma}_n(h_j)\}^2 \nabla\gamma(h_j;\theta).$$

The second row is negligible compared with the first and so is neglected. The first row consists of a (vector) linear combination of the random variables $\{\gamma(h_j;\theta) - \widehat{\gamma}_n(h_j)\}$ and so, assuming (2.24), its limit at $\theta = \theta_0$ may be expressed in the form

$$\frac{1}{\sqrt{n}} \nabla S_n(\theta_0) \rightarrow_d \ \mathcal{N}(0, R(\theta_0)^T W(\theta_0) R(\theta_0))$$

where, if θ is a Q-dimensional vector with components $\{\theta_q,\ q = 1, ..., Q\}$, $R(\theta)$ is a $J \times Q$ matrix whose (j, q) component is given by

$$r_{jq} = \frac{2\phi_j}{\gamma^2(h_j;\theta)} \frac{\partial\gamma}{\partial\theta_q}(h_j;\theta). \tag{2.28}$$

Consequently, if we define

$$V(\theta_0) = R(\theta_0)^T W(\theta_0) R(\theta_0), \tag{2.29}$$

we have verified conditions (A1) and (A2) of the information sandwich formula, so the final result is given by combining the formulae (2.22), (2.24), (2.27), (2.28) and (2.29).

It should be noted that each of the matrices $W(\theta)$, $H(\theta)$ and $R(\theta)$ is explicitly defined once θ is specified, so the resulting expression for the limiting covariance matrix of $\tilde{\theta}_n$ is computable. In practice, of course, in making these calculations we substitute the estimate $\tilde{\theta}_n$ for the unknown value θ_0.

2.2.4 Maximum likelihood estimation

If we assume that we are sampling from a Gaussian process, then it is straightforward in principle to write down the exact likelihood function and hence to maximize it numerically with respect to the unknown parameters. Kitanidis (1983) and Mardia and Marshall (1984) were the first to advocate estimating spatial processes in this way. The evaluation of the likelihood function requires computing the inverse and determinant of the model covariance matrix — if there are n sampling points, then this is an $n \times n$ matrix, and the process can be slow if n is large. Nevertheless, the present author has successfully implemented this procedure for n up to 500, so computational difficulties do not seem to be adequate reason to avoid this method. Less clear are the sampling properties of maximum likelihood estimates as compared with simpler alternatives such as Cressie's WLS

procedure. In this section we shall first outline the computational procedure, and then discuss some of the pros and cons of maximum likelihood estimation.

We can incorporate deterministic linear regression terms with no essential change in the methodology, so the model we shall consider is

$$Z \sim \mathcal{N}(X\beta, \Sigma), \tag{2.30}$$

with Z an n-dimensional vector of observations, X an $n \times q$ matrix of known regressors ($q < n$; X of full rank), β a q-vector of unknown regression parameters and Σ the covariance matrix of the observations. In many applications we may assume

$$\Sigma = \alpha V(\theta) \tag{2.31}$$

where α is an unknown scale parameter and $V(\theta)$ is a vector of standardized covariances determined by the unknown parameter vector θ. For example, the exponential variogram structure is equivalent to a covariance function

$$\operatorname{cov}\{Z(s_1), Z(s_2)\} = \begin{cases} c_0 + c_1 & \text{if } s_1 = s_2 \\ c_1 \exp(-|s_1 - s_2|/R) & \text{if } s_1 \neq s_2 \end{cases} \tag{2.32}$$

so we may define $\alpha = c_1,\ \phi = c_0/(c_0 + c_1)$ (the nugget:sill ratio), $\theta = (\phi, R)$ and let $V(\theta)$ denote the matrix whose diagonal entries are all $1/(1-\phi)$ and off-diagonal entries are of the form $v_{ij} = \exp(-d_{ij}/R)$ where d_{ij} is the distance between the i'th and j'th sampling points. Of course, we assume $V(\theta)$ is nonsingular.

With Z defined by (2.30), its density is

$$(2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(Z - X\beta)^T \Sigma^{-1}(Z - X\beta)\right\}.$$

Consequently, the negative log likelihood is given by

$$\ell(\beta, \alpha, \theta) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\alpha + \frac{1}{2}\log|V(\theta)| + \frac{1}{2\alpha}(Z - X\beta)^T V(\theta)^{-1}(Z - X\beta). \tag{2.33}$$

As a side calculation, note that if for given V we define $\widehat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Z$ (the GLS estimator of β based on covariance matrix V), we have

$$(Z - X\widehat{\beta})^T V^{-1} X = 0,$$

and so

$$\begin{aligned}(Z - X\beta)^T V^{-1}(Z - X\beta) &= (Z - X\widehat{\beta} + X\widehat{\beta} - X\beta)^T V^{-1}(Z - X\widehat{\beta} + X\widehat{\beta} - X\beta) \\ &= (Z - X\widehat{\beta})^T V^{-1}(Z - X\widehat{\beta}) + (\widehat{\beta} - \beta)^T X^T V^{-1} X(\widehat{\beta} - \beta)\end{aligned} \tag{2.34}$$

which confirms that this choice of β indeed minimizes the generalized sum of squares criterion (2.34) and leads to a sum of squares of generalized residuals which we shall denote by

$$G^2 = (Z - X\widehat{\beta})^T V^{-1}(Z - X\widehat{\beta}). \tag{2.35}$$

Returning to (2.33), we see that if we define $\widehat{\beta}(\theta) = (X^T V(\theta)^{-1} X)^{-1} X^T V(\theta)^{-1} Z$ and the corresponding G^2 by $G^2(\theta)$ from (2.35), we have

$$\ell(\widehat{\beta}(\theta), \alpha, \theta) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\alpha + \frac{1}{2}\log|V(\theta)| + \frac{1}{2\alpha}G^2(\theta). \tag{2.36}$$

It is possible to minimize (2.36) numerically with respect to α and θ, or alternatively to minimize it analytically with respect to α by defining

$$\widehat{\alpha}(\theta) = \frac{G^2(\theta)}{n}.$$

In this case we have to minimize, with respect to θ, the function

$$\begin{aligned} \ell^*(\theta) &= \ell(\widehat{\beta}(\theta), \widehat{\alpha}(\theta), \theta) \\ &= \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\frac{G^2(\theta)}{n} + \frac{1}{2}\log|V(\theta)| + \frac{n}{2}. \end{aligned} \tag{2.37}$$

The quantity (2.36) or (2.37) is often called a *profile negative log likelihood* to reflect the fact that it is computed from the negative log likelihood (2.33) by minimizing analytically over some of the parameters. The method given here is essentially that first proposed by Kitanidis (1983) and by Mardia and Marshall (1984). To calculate (2.37), the key element is the Cholesky decomposition which enables us to write $V = LL^T$ where L is a lower triangular matrix. Hence if we write (2.30) and (2.31) in the form

$$Z = X\beta + \eta, \quad \eta \sim \mathcal{N}(0, \alpha V),$$

and define $Z^* = L^{-1}Z$, $X^* = L^{-1}X$, $\eta^* = L^{-1}\eta$, we have

$$Z^* = X^*\beta + \eta^*, \quad \eta^* \sim \mathcal{N}(0, \alpha I),$$

so that the calculation of $\widehat{\beta}$ reduces to an ordinary least squares problem for (Z^*, X^*). Also, the calculation of $|V|$ is easy because this is just the square of $|L|$, and $|L|$ is just the product of diagonal entries. The author's implementation of this is based on the algorithm of Healy (1968) to calculate the Cholesky decomposition, followed by the SVDFIT algorithm of Press *et al.* (1986, Section 14.3) to solve the ordinary least squares problem, all within the DFPMIN algorithm of Press *et al.* (1986, Section 10.7) to solve the function minimization problem with respect to θ. DFPMIN is a variable metric algorithm requiring the specification of first-order derivatives of the objective function as well as the function

itself, but these can be approximated numerically. In this form, the algorithm is somewhat simpler than the original proposals made by Kitanidis (1983) and Mardia and Marshall (1984).

To summarize the main steps of the algorithm:

1. For the current value of θ, compute $V = V(\theta)$ and hence the Cholesky decomposition $V = LL^T$.

2. Calculate L^{-1}. This is easy, given that L is lower triangular.

3. Calculate $|L|$, which is simply the product of the diagonal entries of L. Hence $|V| = |L|^2$.

4. Compute $Z^* = L^{-1}Z$ and $X^* = L^{-1}X$.

5. Solve the ordinary least squares problem $Z^* = X^*\beta + \eta^*$ — the residual sum of squares is $G^2(\theta)$.

6. Define $g(\alpha, \theta)$ by (2.36), or $g(\theta)$ by 2.27, so that g is the function which we have to minimize.

7. Repeat each of steps 1–6 for each θ (or each (α, θ) pair) for which g has to be evaluated. The minimum will eventually be achieved at a point $\widehat{\theta}$ (or $(\widehat{\alpha}, \widehat{\theta})$) and this defines the maximum likelihood estimator.

8. Define H to be the Hessian matrix, i.e. the matrix of second-order derivatives of g with respect to the unknown parameters, evaluated at the maximum likelihood estimators. This is also known as the observed information matrix, and in the case of a quasi-Newton algorithm such as DFPMIN, may be obtained approximately from the algorithm itself. (The algorithm does not attempt to evaluate H directly, but maintains an approximation of it which is improved as the algorithm continues.) In this case, in accordance with standard maximum likelihood theory, the inverse matrix H^{-1} is an approximation to the sampling covariance matrix of the parameter estimates. In particular, the square roots of the diagonal entries of H^{-1} are approximate *standard errors* of the parameter estimates. Finally, the minimized value of g may be used for *likelihood ratio tests* in comparing one model with another — we shall see numerous examples of this subsequently.

Remarks

1. Effective operation of the algorithm requires reasonable starting values. One solution is to calculate the approximate WLS estimators first, using these as starting values for the MLE procedure. In the author's experience, that level of care is not usually required, but it is important to use starting values that at least represent reasonable guesses of the MLEs. One general piece of advice is to build up gradually from simpler models towards

more complicated ones, using the estimates from simpler models to help gauge starting values for the more complicated models.

2. It is also advisable to remember that efficient operation of quasi-Newton algorithms such as DFPMIN requires that the parameters be at least reasonably well-scaled, e.g. the algorithm will not usually work correctly if one parameter is varying on a scale 10^6 times another. This could require attention, in particular, to choosing a suitable unit for distance.

3. The reader may be wondering why we considered two forms of g, one based on (2.36) and the other based on (2.37), instead of just using (2.37) which has fewer parameters. The reason is that for some examples later on (section 2.6) the covariance matrix does not have the form of (2.31), so in this case the simplification afforded by analytic solution for α is not available.

4. The interpretation of H in step 8 is not precisely in accordance with standard asymptotic theory of maximum likelihood estimates, because it is calculated from a profile log likelihood rather than the original log likelihood function. However, it can be shown that the H matrix in this case has the same interpretation as when it is defined directly from the log likelihood. See Patefield (1977).

Advantages and disadvantages of maximum likelihood estimation

Although maximum likelihood estimation appears to be computationally feasible, opinion is still divided concerning its desirability when compared with simpler methods such as the approximate WLS method due to Cressie (section 2.2.3). Asymptotic properties of maximum likelihood estimators were considered by Mardia and Marshall (1984), who showed that the usual asymptotic properties of consistency and asymptotic normality are satisfied under a form of increasing domain asymptotics (see section 2.2.3 for a parallel discussion of Cressie's WLS method under this form of asymptotics). However, the conditions given by Mardia and Marshall are not particularly easy to check, especially in the case of an irregular lattice of sampling points, and more seriously, there is no indication of how large the samples need to be for asymptotic results to be reliable indicators of sampling properties. Another issue concerns possible multimodality of the likelihood surface. An example given by Warnes and Ripley (1987), and repeated by Ripley (1988), suggests that this can be a problem even with the simplest spatial models. In fact it would appear that the original example given by Warnes and Ripley was in error — Mardia and Watkins (1989) presented an alternative analysis of the same data set, which is discussed in section 2.3 below. Nevertheless, the possibility of multimodality is real, arising from discontinuities in the first derivative of the log likelihood, as shown theoretically by Mardia and Watkins in the case of the spherical variogram model and a variant of the exponential model. They advocated plotting the profile likelihood surface (2.37), as well as or instead of finding the MLE by optimization. In the present author's experience, multimodality is not usually a difficulty in low-dimensional estimation problems, but even with parameter dimensions of the order of 4 or 5, it can happen that parallel runs of the maximum likelihood estimation routine, starting from different initial values, lead to different parameter

estimates. It can also happen that with poor initial values, the algorithm will not converge at all. Given the various difficulties that can arise, it is advisable to be check the results of the algorithm by rerunning from different starting values, and to be cautious about the results if these difficulties arise.

The theoretical benefit of maximum likelihood is that we can expect the estimates to be more efficient than the alternative methods in large samples. However, it is not clear how big a benefit this is. A simulation study by Zimmerman and Zimmerman (1991) compared the MLE, approximate WLS and a number of alternative estimators, concluding that the MLE is only slightly superior to the approximate WLS procedure from this point of view. It has also been pointed out that the MLE procedure depends on the assumption of a Gaussian process and therefore may perform poorly when the true distribution is non-Gaussian, but of course this does not mean that the WLS procedures would necessarily perform better in this case! The simulations of Zimmerman and Zimmerman (1991) do not address this issue since they are restricted to Gaussian processes.

The present author takes the view that the computational complexity of maximum likelihood (or its variant REML — see next subsection) is outweighed by its convenience as a very widely applicable method of estimation, by which a variety of models can be estimated, and compared using either likelihood ratio tests or automatic model selection criteria such as the Akaike Information Criterion (AIC). There is also the advantage that maximum likelihood methods link up naturally with Bayesian procedures, as will be further explored in subsection 2.2.6. The various disadvantages that have been pointed out, such as non-robustness when there are outliers in the data, or the possible multimodality of the likelihood surface, are caveats to keep in mind when using the method, but they are not reasons to abandon maximum likelihood estimation.

Multiple replications

The treatment so far has been based on the assumption that inference must be based on a single realization of the random field Z. Of course, we can expect to get better estimates if there are multiple replications of Z. In the climatological examples of chapter 1, we have treated the data from year to year as independent, which is equivalent to assuming that there are multiple independent replications.

The maximum likelihood procedure in this case is of course only slightly different from that in the single-replication case, but to make the computational procedure explicit, we explain here what the differences are. Suppose there are m replications denoted $Z_1, ..., Z_m$. Then (2.33) is replaced by

$$\ell(\beta, \alpha, \theta) = \frac{mn}{2}\log(2\pi) + \frac{mn}{2}\log\alpha + \frac{m}{2}\log|V(\theta)| + \frac{1}{2\alpha}\sum(Z_i - X\beta)^T V(\theta)^{-1}(Z_i - X\beta).$$

Defining $\bar{Z} = \frac{1}{m}\sum Z_i$, we may write

$$\begin{aligned}&\sum(Z_i - X\beta)^T V(\theta)^{-1}(Z_i - X\beta)\\ &= m(\bar{Z} - X\beta)^T V(\theta)^{-1}(\bar{Z} - X\beta) + \sum(Z_i - \bar{Z})V(\theta)^{-1}(Z_i - \bar{Z}).\end{aligned}$$

This suggests the following modification of the algorithm for $m = 1$, to compute the profile log likelihood function (2.36) or (2.37) in this case:

1. For given θ, solve the GLS problem for $\bar{Z}$, letting $G_0^2(\theta)$ be the resulting generalized residual sum of squares.

2. Calculate

$$G^2(\theta) = G_0^2(\theta) + \frac{1}{m} \sum (Z_i - \bar{Z}) V(\theta)^{-1} (Z_i - \bar{Z}).$$

3. Substitute into (2.36) or (2.37), multiplying by m to obtain the correctly normalized profile log likelihood.

As an example, we consider a data set based on 32 years (1965–1996) of mean winter daily minimum temperatures, confined to the region of latitude 40–45 ^{o} N and longitude 90–100 ^{o}W (recall Fig. 2.9). The individual data points consist, as in several earlier examples, of standardized residuals from a linear regression in time, and we shall treat them as 32 independent observation vectors. It will be recalled from Table 2.2 that there was some doubt about whether the process was homogeneous even within this smaller region, but for the purpose of the present discussion, we shall assume the process is homogeneous. The $X\beta$ component of (2.30) was in most cases omitted from the model, but some models corresponding to a linear spatial trend in the latitude and longitude coordinates were also tried; in this case β is a vector of dimension 2. Maximum likelihood estimates were computed for several models, with the following results:

Model	Spatial Trend	Number of Parameters	NLLH	AIC
Exponential	None	3	–548.3	–1090.6
Exponential	Linear	5	–548.3	–1086.6
2-par Matérn	None	3	–547.1	–1088.2
3-par Matérn	None	4	–548.4	–1088.8
3-par Matérn	Linear	6	–548.4	–1084.8
Gaussian	None	3	–535.3	–1064.6
Wave	None	3	–532.8	–1059.6
Spherical	None	3	–548.2	–1090.4

Table 2.3 Evaluation of negative log likelihood (NLLH) and Akaike Information Criterion (AIC) for several models fitted to 32 years of data at 17 stations.

Each of the exponential, Gaussian, wave and spherical models included a nugget parameter, but the Matérn model was fitted both without a nugget (2-parameter version)

and with (3-parameter). From the NLLH and AIC values tabulated, it can be seen that the Gaussian and wave models are substantially inferior, but the other six models are indistinguishable by the quality of fit. The two models with linear spatial trend would be rejected on the grounds that they do not improve on the models with no trend, while the three-parameter Matérn is similarly rejected on comparision with the two-parameter model. Comparison of the exponential and two-parameter Matérn models leads to the following comparisons:

Model	Parameter	Estimate	Standard error
Exponential	R	18.14	4.94
Exponential	ϕ	.038	.010
Matérn	θ_1	65.7	40.7
Matérn	θ_2	.28	.03

Table 2.4 Parameter estimates and standard errors for the exponential and two-parameter Matérn models. The parameter ϕ is the nugget:sill ratio (recall discussion following (2.32)).

The unit of distance here, in which both R and θ_1 are expressed, is 100 nautical miles, using the approximate conversion factors 1^o latitude = 60 N.M., 1^o longitude = $0.8\times$ 60 = 48 N.M. (0.8 is approximately cos 40 ^{o}). Recall our earlier remarks about scaling; the unit of distance is taken to be 100 N.M. rather than 1 N.M. because the optimization problem is numerically more stable in this case.

2.2.5 Restricted maximum likelihood

The idea of *restricted maximum likelihood* or REML estimation was originally proposed by Patterson and Thompson (1971) in connection with variance components in linear models. However, a number of authors have pointed out that the situation considered by Patterson and Thompson is essentially the same as arises with Gaussian models for spatial data: in both cases there is a linear model with correlated errors, whose covariance matrix depends on some additional parameters. Thus it is natural to try to separate the two parts of the estimation problem, the "linear model" part and the "covariance structure" part. Cressie (1993) is one author who has enthusiastically advocated this approach to spatial analysis.

The motivation behind REML estimation is perhaps best expressed in a very simple case. Suppose $Y_1, ..., Y_n$ are independent univariate random variables, each $\mathcal{N}(\mu, \sigma^2)$ with unknown μ and σ^2. As is well known, the maximum likelihood estimators of μ and σ^2 are $\widehat{\mu} = \bar{Y} = \frac{1}{n}\sum_i Y_i$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_i (Y_i - \bar{Y})^2$. However, this definition of $\widehat{\sigma}^2$ is a biased estimator, whereas the more usual unbiased estimator of σ^2 is $\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2$. This it appears that the maximum likelihood estimator is not the best one to use in

this case. Suppose, however, instead of basing the maximum likelihood estimator on the full joint density of $Y_1, ..., Y_n$, we base it on the joint density of the vector of contrasts, $(Y_1 - \bar{Y}, Y_2 - \bar{Y}, ..., Y_{n-1} - \bar{Y})$, whose distribution does not depend on μ. The "maximum likelihood estimator" of σ^2, under this formulation, turns out to be the unbiased estimator $\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2$. Thus, by constructing an estimate of σ^2 based on an $(n-1)$-dimensional vector of contrasts, we appear to have done better than the usual maximum likelihood estimator based on the full n-dimensional data vector.

This idea can be extended to the general model defined by (2.30) and (2.31). If we let $W = A^T Z$ be a vector of $n - q$ linearly independent contrasts, i.e. the $n - q$ columns of A are linearly independent and $A^T X = 0$, then we find that

$$W \sim \mathcal{N}(0, A^T \Sigma A),$$

and the joint negative log likelihood function based on W is of the form

$$\ell_W(\alpha, \theta) = \frac{n-q}{2}\log(2\pi) + \frac{n-q}{2}\log\alpha + \frac{1}{2}\log|A^T V(\theta) A| + \frac{1}{2\alpha} W^T (A^T V(\theta) A)^{-1} W. \quad (2.38)$$

As pointed out by Patterson and Thompson (1971), it is possible to choose A to satisfy $AA^T = I - X(X^T X)^{-1} X^T$, $A^T A = I$. In this case a further calculation, first given by Harville (1974), shows that (2.38) may be simplified to

$$\begin{aligned} \ell_W(\alpha, \theta) = &\frac{n-q}{2}\log(2\pi) + \frac{n-q}{2}\log\alpha - \frac{1}{2}\log|X^T X| + \frac{1}{2}\log|X^T V(\theta)^{-1} X| \\ &+ \frac{1}{2}\log|V(\theta)| + \frac{1}{2\alpha} G^2(\theta), \end{aligned} \quad (2.39)$$

where $G^2(\theta)$ is the same as in (2.37). This is minimized with respect to α by setting $\tilde{\alpha} = G^2(\theta)/(n-q)$ in which case (2.39) reduces to

$$\begin{aligned} \ell^*_W(\theta) &= \ell_W(\tilde{\alpha}, \theta) \\ &= \frac{n-q}{2}\log(2\pi) + \frac{n-q}{2}\log\frac{G^2(\theta)}{n-q} - \frac{1}{2}\log|X^T X| + \frac{1}{2}\log|X^T V(\theta)^{-1} X| \\ &\quad + \frac{1}{2}\log|V(\theta)| + \frac{n-q}{2}. \end{aligned} \quad (2.40)$$

Comparing (2.40) with (2.37), it can be seen that there are two substantive changes: the coefficient of $\log G^2(\theta)$ has been changed from $n/2$ to $(n-q)/2$, and there is an additional term of $\frac{1}{2}\log|X^T V(\theta)^{-1} X|$.

** *Derivation of (2.39)*

We follow Harville (1974). Recall that A is an $n \times (n-q)$ matrix and let G denote the $n \times q$ matrix $V^{-1}X(X^T V^{-1} X)^{-1}$, so that $\widehat{\beta} = G^T Z$. Let $B = [A|G]$; in other words,

the $n \times n$ matrix formed by placing the matrices A and G alongside one another. Then

$$\begin{aligned} |B| &= |B^T B|^{1/2} \\ &= \left| \begin{pmatrix} A^T A & A^T G \\ G^T A & G^T G \end{pmatrix} \right|^{1/2} \\ &= |A^T A|^{1/2} |G^T G - G^T A (A^T A)^{-1} A^T G|^{1/2} \end{aligned}$$

the last line depending on a well-known result for the determinant of a block matrix (see for instance, Mardia, Kent and Bibby (1979), formula A.2.3j, page 457). However after noting that $A^T A = I$, $AA^T = I - X^T (X^T X)^{-1} X$, it may quickly be verified that

$$G^T G - G^T A (A^T A)^{-1} A^T G = (X^T X)^{-1}.$$

Thus $|B| = |X^T X|^{-1/2}$.

Recall that the density of Z under (2.30)–(2.31) is

$$f_Z(z) = (2\pi)^{-n/2} \alpha^{-n/2} |V|^{-1/2} \exp\left\{ -\frac{1}{2\alpha} (Z - X\beta)^T V^{-1} (Z - X\beta) \right\}. \tag{2.41}$$

Define $Z^\dagger = B^T Z = (Z^T A, Z^T G)^T = (W^T, \widehat{\beta}^T)^T$. The Jacobian of the transformation from Z to $Z^\dagger$ is $|B|^{-1} = |X^T X|^{1/2}$. Moreover, using (2.34)–(2.35) we have that

$$(Z - X\beta)^T V^{-1} (Z - X\beta) = G^2(\theta) + (\widehat{\beta} - \beta)^T X^T V^{-1} X (\widehat{\beta} - \beta).$$

Now $G^2(\theta)$ is a function of elements orthogonal to $\widehat{\beta}$ and hence is itself a function of W. Thus a change of variables in (2.41) leads to

$$\begin{aligned} f_{W,\widehat{\beta}}(w, \widehat{\beta}) = |X^T X|^{1/2} (2\pi)^{-n/2} \alpha^{-n/2} |V|^{-1/2} \cdot \\ \cdot \exp\left\{ -\frac{1}{2\alpha} G^2(\theta) - \frac{1}{2\alpha} (\widehat{\beta} - \beta)^T X^T V^{-1} X (\widehat{\beta} - \beta) \right\}. \end{aligned} \tag{2.42}$$

Now integrate (2.42) with respect to $\widehat{\beta}$, leading to

$$f_W(w) = |X^T X|^{1/2} (2\pi)^{-(n-q)/2} \alpha^{-(n-q)/2} |V|^{-1/2} |X^T V^{-1} X|^{-1/2} \exp\left\{ -\frac{1}{2\alpha} G^2(\theta) \right\}$$

from which (2.39) follows at once.

2.2.6 Bayesian procedures

Bayesian procedures to spatial statistics have been considered by a number of authors, in particular Le and Zidek (1992) and Handcock and Stein (1993). The latter authors considered the model defined by (2.30) and (2.31) with the improper prior density

$$\pi(\beta, \alpha, \theta) \propto \frac{\pi(\theta)}{\alpha} \tag{2.43}$$

for some prior $\pi(\theta)$. The posterior density takes the form

$$\pi(\beta, \alpha, \theta|Z) \propto \frac{\pi(\theta)}{\alpha}(2\pi)^{-n/2}\alpha^{-n/2}|V(\theta)|^{-1/2}\exp\left\{-\frac{1}{2\alpha}(Z-X\beta)^T V(\theta)^{-1}(Z-X\beta)\right\}.$$

Again defining $\widehat{\beta}(\theta) = (X^T V(\theta)^{-1}X)^{-1}X^T V(\theta)^{-1}Z$ and ignoring constants, equation (2.34) leads to

$$\begin{aligned}\pi(\beta, \alpha, \theta|Z) \propto & \frac{\pi(\theta)}{\alpha}\alpha^{-n/2}|V(\theta)|^{-1/2}\exp\left\{-\frac{G^2(\theta)}{2\alpha}\right\}\cdot \\ & \cdot\exp\left\{-\frac{1}{2\alpha}(\beta-\widehat{\beta})^T X^T V(\theta)^{-1}X(\beta-\widehat{\beta})\right\}.\end{aligned} \tag{2.44}$$

Integrating out with respect to β, we obtain

$$\pi(\alpha, \theta|Z) \propto \frac{\pi(\theta)}{\alpha}\alpha^{-n/2}|V(\theta)|^{-1/2}\exp\left\{-\frac{G^2(\theta)}{2\alpha}\right\}\cdot\alpha^{q/2}|X^T V(\theta)^{-1}X|^{-1/2} \tag{2.45}$$

and a further integration with respect to α leads to

$$\pi(\theta|Z) \propto \pi(\theta)|V(\theta)|^{-1/2}G^2(\theta)^{-(n-q)/2}|X^T V(\theta)^{-1}X|^{-1/2}, \tag{2.46}$$

which is the same as equation (3.2) of Handcock and Stein (1993).

Comparing (2.44) with (2.40), it can be seen that if we ignore $\pi(\theta)$ in (2.44), the mode of the posterior density of θ is precisely the REML estimator. This was first pointed out by Harville (1974) and indeed follows at once from (2.42), on writing out the joint density of Z in this form and then integrating out with respect to β — the result is exactly the same as if we integrate (2.42) with respect to $\widehat{\beta}$. However, a fully Bayesian approach involves not maximizing (2.44), but integrating with respect to the components of θ, and in this respect the two methods are quite different. The integration with respect to θ must be performed numerically.

2.2.7 MINQE estimation

Another method of estimation is the method of minimum norm quadratic estimation, or MINQE for short, which was originally developed by C.R. Rao, see for example Rao (1979). In comparison with the other methods we have considered, MINQE is restricted in scope, being confined to a particular class of spatial estimation problems, but within those classes of problems, it seems competitive with the other methods. The following description is based on the accounts of Kitanidis (1983) and Stein (1987).

Suppose we write the universal kriging model in the form

$$Z = X\beta + \eta,$$

where the semivariogram of η is $\gamma(\cdot;\theta)$, and suppose $\gamma(\cdot;\theta)$ is of the form

$$\gamma(h;\theta) = \sum_{k=1}^{K} \theta_k \gamma_k(h), \tag{2.47}$$

in other words, γ is a linear combination of k *known* semivariograms $\gamma_1, ..., \gamma_K$, with unknown weights $\theta_1, ..., \theta_K$. Just as with REML estimation, suppose we define $W = A^T Z$ to be a vector of orthogonal contrasts to X, where we assume that the columns of X include a constant term, so that the covariance of W is of the form $-A^T \Gamma(\theta) A = \Psi(\theta)$ say, where $\Gamma(\theta)$ is the matrix with entries $\gamma(s_i - s_j;\theta)$, $s_1, ..., s_n$ being the sampling points. We also let $\Psi_1, ..., \Psi_K$ denote the corresponding Ψ matrices when $\gamma = \gamma_k$, for each of $k = 1, ..., K$. The problem is therefore to estimate the coefficients $\{\theta_k\}$ when observed data have the covariance matrix

$$\Psi(\theta) = \sum_{k=1}^{K} \theta_k \Psi_k. \tag{2.48}$$

Suppose, for a given vector $p^T = (p_1, ..., p_K)$, we choose to estimate $p^T\theta$ by the quadratic form $Y^T HY$. For this to be unbiased, we require E$\{Y^T HY\}$=E$\{tr(HYY^T)\}$
$= \sum \theta_k tr(H\Psi_k)$ and hence $tr(H\Psi_k) = p_k$. The idea behind MINQE is to choose the minimum variance unbiased estimator of this form. Typically, the variance of $Y^T HY$ is of the form $tr(HVHV)$ for some matrix V. For example, in the case of a Gaussian process it is easily verified directly that Var$\{Y^T HY\} = 2\ tr(H\Psi(\theta)H\Psi(\theta))$.

In practice, the usual scheme is to fix $V = \Psi(\alpha)$ for some prior guess α of θ. A Lagrange multiplier solution to the resulting contrained optimization problem leads us to define $A_i = \Psi(\alpha)^{-1}\Psi_i\Psi(\alpha)^{-1}$. The estimator $\widehat{\theta}$ will be an unbiased estimator of θ if

$$Y^T A_i Y = tr(A_i \Psi(\widehat{\theta})) = \sum_j \widehat{\theta}_j tr(A_i \Psi_j) \tag{2.49}$$

for all i. If we let B denote the matrix with entries $b_{ij} = tr(A_i\Psi_j)$, and let C with entries c_{ij} denote the inverse of B (assumed to exsit), then the solution of (2.49) is

$$\widehat{\theta}_i = \sum c_{ij} Y^T A_j Y. \tag{2.50}$$

For the case where $\Psi(\theta)$ cannot be written as a linear function of θ, Stein suggested replacing the definition of A_i with

$$A_i = \Psi(\alpha)^{-1} \left\{ \frac{\partial}{\partial \alpha_i} \Psi(\alpha) \right\} \Psi(\alpha)^{-1}$$

but in this case the method seems to be even less well motivated compared with general procedures such as maximum likelihood or REML.

The advantage of this method, compared with maximum likelihood or REML or even the approximate WLS procedure, is that for fixed α it is a linear estimation procedure and therefore does not require any iterated procedures. Moreover, in many cases it appears that the estimate is not too sensitive to the specification of α. On the other hand, the optimality properties of the procedure hold only when $\alpha = \theta$, and this suggests iterating the procedure, using the current estimate $\widehat{\theta}$ to define α for the next iteration. Kitanidis (1983) showed that if this procedure is iterated to convergence then the MINQE estimator satisfies the same equations as those satisfied by the maximum likelihood estimator. Zimmerman and Zimmerman (1991) proposed a compromise in which a crude estimator of θ was taken for the first estimate and then iterated once only. With these refinements, the method is competitive in performance with the ML and REML procedures, but more computationally demanding than the approximate WLS procedure.

2.3 Examples

Table 2.5 gives a data set analyzed by Cressie (1989,1993). The data consist of water levels in the Wolfcamp aquifer in south-west Texas. Eighty-five measurements were taken by drilling into the ground and locating the height; the x and y coordinates represent miles from an arbitrary origin and the z coordinate is the water level in feet above sea level. The original interest in this example was the proposal to build a nuclear waste repository in Deaf Smith County, which is near the center of the mapped region. It is believed that any leakage of nuclear waste will flow with the water in the aquifer so there is interest in reconstructing the shape of the whole water surface. In fact the analysis showed fairly directly that there is a steady slope from the south-west to the north-east and that leakage from the proposed site would flow directly into the city of Amarillo. This is described in some detail in Chapter 4 of Cressie (1993). We concentrate here on the spatial model-fitting aspect of the problem.

Cressie himself gave two analyses. The first assumed that the process was intrinsically stationary, but because of the obvious anisotropy, used a geometrically anisotropic model. From this, he used a kriging algorithm (section 2.4) to reconstruct the surface. This appeared satisfactory but gave a very irregular reconstructed surface. A second technique was to use median polish kriging (also to be discussed in section 2.4) as a crude method of removing an underlying trend. The residuals from this trend did appear to follow a stationary isotropic model and were satisfactorily fitted by the spherical variogram model. Ordinary kriging was then applied to these residuals and added to the trend surface obtained by median polish kriging to obtain a second reconstructed surface. This was similar in general characteristics to the first method but did produce a noticeably smoother surface.

In Fig. 2.10 we show the variograms ($V(t) = 2\gamma(t)$) for the raw data computed by both the method of moments (MoM) and the robust method. To account for the anisotropy, separate variograms were computed for pairs of points whose relative orientation lay in the SW-NE quadrant and those in the SE-NW quadrant. The figure is similar to Fig. 4.3

of Cressie (1993) except that Cressie only used the MoM method. The unit of distance (t) here is five miles. Also shown are fitted variograms by the approximate WLS method based on (2.16) using the power law model. In contrast to Cressie's analysis, which forced a common value of the power λ, this analysis fitted the power law model separately to all four plots. In fact the two MoM fits produced $\lambda = 2.0$ and $\lambda = 1.9$ which is consistent with a common λ; Cressie claimed $\lambda = 1.99$ but did not remark on how close this is to the upper boundary for this to be an intrinsically stationary model (the upper boundary is $\lambda = 2$). In contrast, the two fits using the robust method of variogram calculation produced *substantially* different variograms (note that the figures are not all plotted on the same scale) and estimates $\lambda = 2.8$ for the SW-NE direction and 2.6 for the SE-NW direction, both well beyond the permitted range. Of course, we could force a valid fit by constraining $\lambda < 2$ in the WLS algorithm, but this would not address the question of whether an intrinsically stationary model is reasonable. I believe that the discrepancies between the MoM and robust variograms, and the results of the power law fit, provide ample evidence that it is not.

Fig. 2.11 indicates an analysis somewhat different from Cressie's, based on a model of the structure of (2.30) in which the X matrix represented regressors given by the x and y coordinates, i.e. we are assuming a linear trend surface with correlated errors. The form of correlation function was chosen so as to be consistent with an exponential variogram model. For an initial analysis, this was estimated by an ordinary least squares regression analysis and the residuals from that regression analyzed as a spatial model. In this case there was no evidence of anisotropy and Fig. 2.11 shows the MoM and robust variogram estimators together with the WLS fit (separately for each variogram) and the ML and REML fits. For the ML and REML fits, the original model (2.30) was taken, i.e. we did not rely on residuals from an OLS fit of β. The ML and REML fits are both reasonably close to the WLS fit – in fact, based on a visual inspection the ML fit appears closer both to the WLS fit and to the individual variogram points.

Fig. 2.12 continues this analysis by showing both the Matérn and wave models fitted by ML and REML. The Matérn fit is based on $\widehat{\theta}_2 = 0.29$ (ML fit) or 0.27 (REML fit) – in this model there appears to be no need for a nugget parameter, which was in fact estimated as 0. The wave model was suggested by the apparent oscillatory shape of the variogram points at large t. In fact, the fitted variogram does not seem to follow this shape too well but still gives the best model as judged by maximum likelihood over all models. However, there is no significant difference among the leading models, as shown in Table 2.6.

x	y	z	x	y	z
42.7827	127.6228	1.464	103.2663	20.3424	1.591
−27.3969	90.7873	2.553	−14.3107	31.2654	2.540
−1.1629	84.8960	2.158	−18.1345	30.1812	2.352
−18.6182	76.4520	2.455	−18.1215	29.5324	2.528
96.4655	64.5806	1.756	−9.8880	38.1448	2.575
108.5624	82.9232	1.702	−12.1634	39.1108	2.468
88.3636	56.4535	1.805	11.6575	18.7335	2.646
90.0421	39.2582	1.797	61.6912	32.9491	1.739
93.1727	33.0585	1.714	69.5790	33.8084	1.674
97.6110	56.2789	1.466	66.7221	33.9326	1.868
90.6295	35.0817	1.729	−36.6545	150.9146	1.865
92.5526	41.7524	1.638	−19.5510	137.7840	1.777
99.4900	59.1578	1.736	−21.2979	131.8254	1.579
−24.0674	184.7664	1.476	−22.3617	137.1368	1.771
−26.0629	114.0748	2.200	21.1472	139.2620	1.408
56.2784	26.8483	1.999	7.6846	126.8375	1.527
73.0388	18.8814	1.680	−8.3323	107.7769	2.003
80.2668	12.6159	1.806	56.7072	171.2644	1.386
80.2301	14.6180	1.682	59.0005	164.5486	1.089
68.8384	107.7742	1.306	68.9689	177.2482	1.384
76.3992	95.9938	1.722	70.9023	161.3814	1.030
64.4615	110.3964	1.437	73.0024	162.9896	1.092
43.3966	53.6150	1.828	59.6624	170.1054	1.161
39.0777	61.9981	2.118	61.8725	174.3018	1.415
112.8045	45.5477	1.725	63.7081	173.9145	1.231
54.2590	147.8199	1.606	5.6271	79.0873	2.300
6.1320	48.3277	2.648	18.2474	77.3919	2.238
−3.8047	40.4045	2.560	85.6882	139.8170	1.038
−2.2305	29.9111	2.544	105.0765	132.0318	1.332
−2.3618	33.8200	2.386	−101.6428	10.6511	3.510
−2.1889	33.6821	2.400	−145.2365	28.0233	3.490
63.2243	79.4992	1.757	−73.9931	87.9727	2.594
−10.7786	175.1135	1.402	−94.4818	86.6261	2.650
−18.9889	171.9169	1.364	−88.8498	76.7099	2.533
−38.5788	158.5274	1.735	−120.2590	80.7648	3.571
83.1450	159.1156	1.376	−86.0245	54.3633	2.811
−21.8025	15.0255	2.729	−72.7910	43.0922	2.728
−23.5646	9.4144	2.766	−100.1737	42.8988	3.136
−20.1130	22.0927	2.736	−78.8354	40.8214	2.553

Table 2.5 Wolfcamp aquifer data.

x	y	z	x	y	z
−16.6265	17.2562	2.432	−83.6906	46.5048	2.798
29.9075	175.1288	1.024	−95.6166	35.8218	2.691
100.9157	22.9781	1.611	−87.5548	29.3927	2.946
101.2954	22.9639	1.548			

Table 2.5 (continued)

Model	Order	ML	REML
Exponential	1	148.0	137.1
Gaussian	1	147.5	136.6
Matérn	1	148.3	137.6
Wave	1	149.4	137.7
Spherical	1	148.4	137.5
Matérn	2	152.3	
Matérn	3	156.1	

Table 2.6 ML and REML fits to various models. The tabulated value is log maximum ML or REML.

The last two rows in Table 2.6 show (for ML estimation in the Matérn model only) the results of extending the analysis to include a quadratic or cubic trend; in both cases there is some improvement in the fit but not significant as judged by the usual χ^2 test.

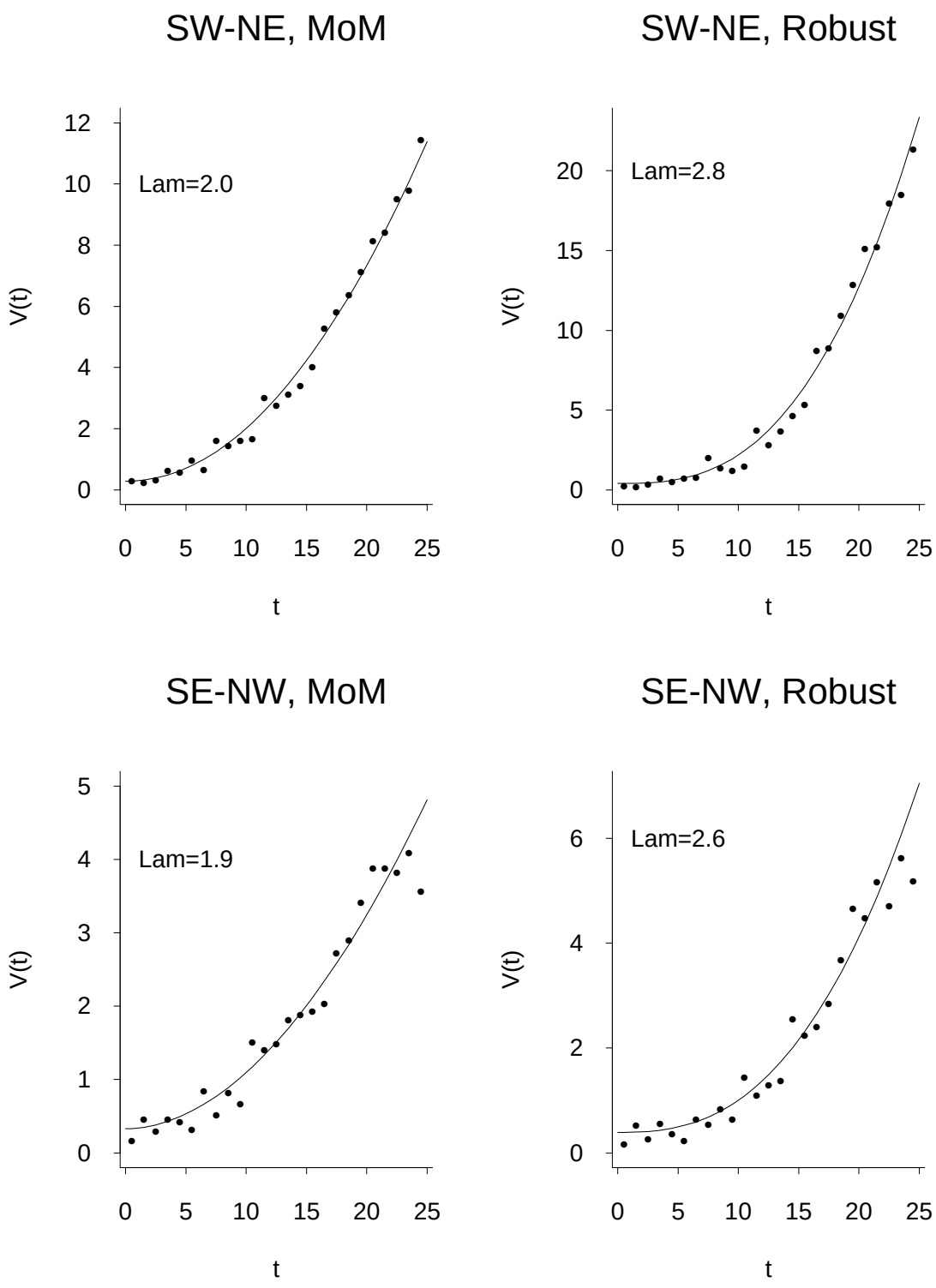


Fig. 2.10. Fitted variograms from Texas data: Raw data with power law models.

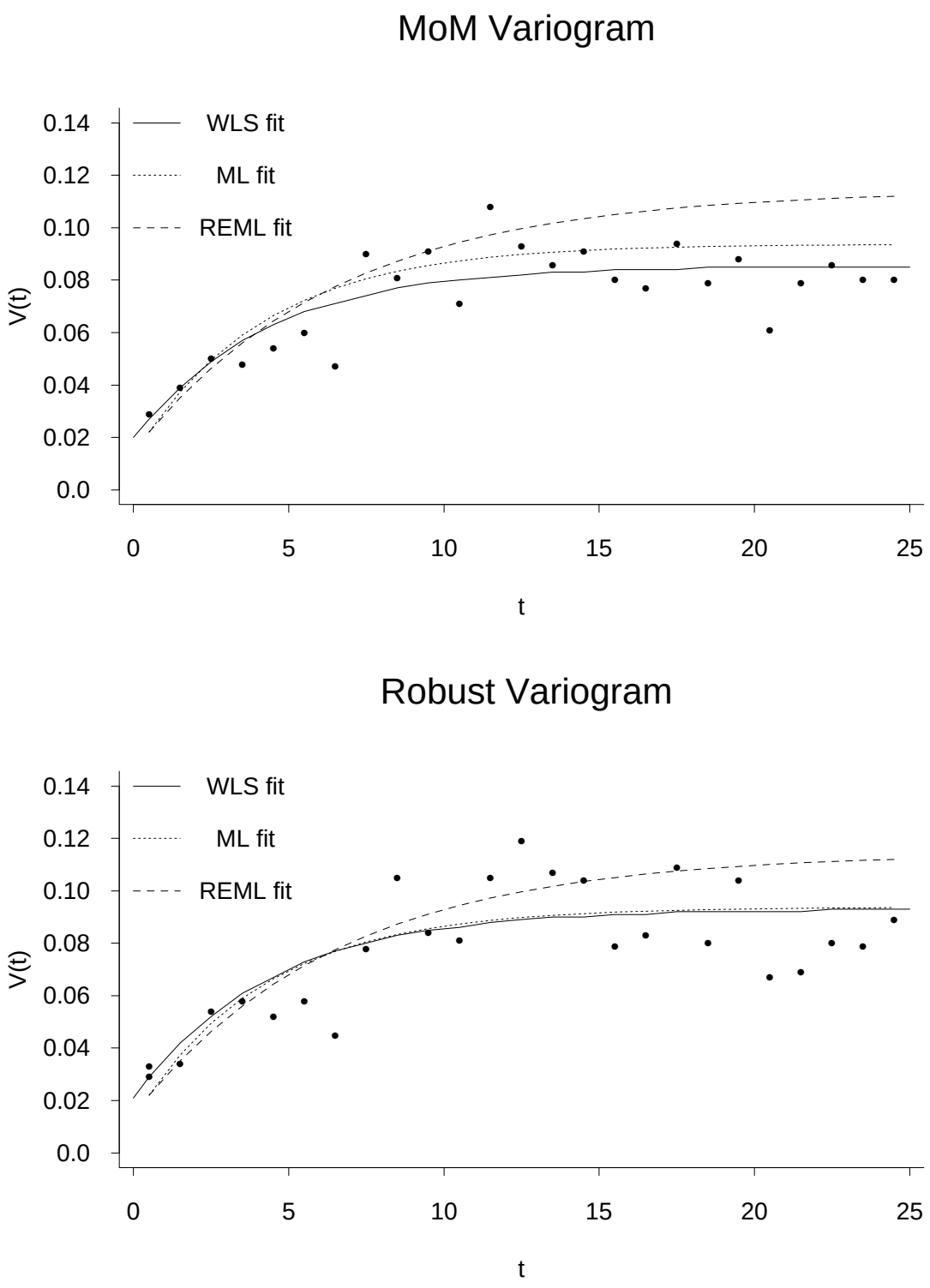


Fig. 2.11 Fitted variograms from detrended Texas data: Exponential variograms fitted by WLS, ML and REML methods, with MoM and robust variograms.

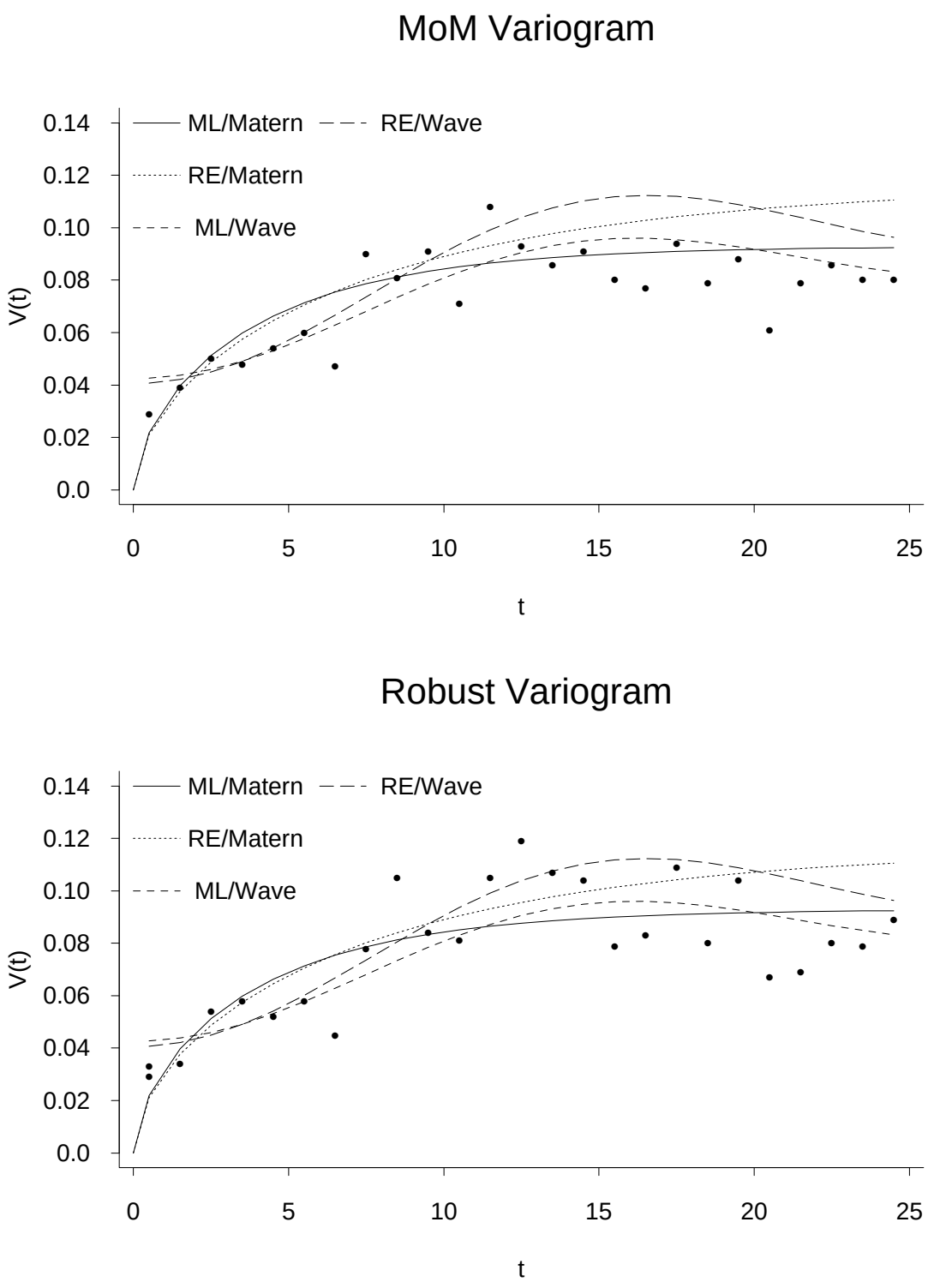


Fig. 2.12. Matérn and wave models fitted to detrended Texas data.

Our second example is based on the data in Table 2.7, which were originally given by Davis (1973) and have been re-analyzed by, amongst others, Ripley (1981, 1988) and Handcock and Stein (1993). The data are 51 measurements of the height of the earth's surface within a 310-foot square. The x and y coordinates have been expressed in units of 50 feet. For the following analysis, the variable z will be replaced throughout by $z/100$ for reasons of numerical stability.

x	y	z	x	y	z
.3	6.1	870	5.2	3.2	805
1.4	6.2	793	6.3	3.4	840
2.4	6.1	755	.3	2.4	890
3.6	6.2	690	2.0	2.7	820
5.7	6.2	800	3.8	2.3	873
1.6	5.2	800	6.3	2.2	875
2.9	5.1	730	.6	1.7	873
3.4	5.3	728	1.5	1.8	865
3.4	5.7	710	2.1	1.8	841
4.8	5.6	780	2.1	1.1	862
5.3	5.0	804	3.1	1.1	908
6.2	5.2	855	4.5	1.8	855
.2	4.3	830	5.5	1.7	850
.9	4.2	813	5.7	1.0	882
2.3	4.8	762	6.2	1.0	910
2.5	4.5	765	.4	.5	940
3.0	4.5	740	1.4	.6	915
3.5	4.5	765	1.4	.1	890
4.1	4.6	760	2.1	.7	880
4.9	4.2	790	2.3	.3	870
6.3	4.3	820	3.1	.0	880
.9	3.2	855	4.1	.8	960
1.7	3.8	812	5.4	.4	890
2.4	3.8	773	6.0	.1	860
3.7	3.5	812	5.7	3.0	830
4.5	3.2	827	3.6	6.0	705

Table 2.7 Davis' data

Ripley (1981) showed contour plots of fitted surfaces from linear up to quintic, which demonstrate that there is no simple dominant trend as there appears to be in the previous example. Fig. 2.13 shows MoM and robust variograms with fitted power law curves, computed separately for the SW-NE and SE-NW quadrants as in Fig. 2.10. In this case

there does not seem to be an argument about the validity of an intinsically stationary assumption (all four fitted values of λ are well below 2) but the strong anisotropy is disturbing. The same plots fitted to the residuals from a linear trend are better (Fig. 2.14) but still not satisfactory – it appears that there is much more persistence in the SE-NW direction than the SW-NE. In this case an exponential model has been fitted.

Warnes and Ripley (1987) made the claim, repeated by Ripley (1988), that this was an example of a multimodal likelihood. They fitted an isotropic exponential variogram with no nugget to the raw data (i.e. no trend) and produced an apparently irregular profile likelihood for the range paranmeter R. However the same model was fitted to the same data by Mardia and Watkins (1989) who found no trace of multimodality. The present author's calculation, shown in Fig. 2.15(a), supports the conclusion of Mardia and Watkins. For this, the profile likelihood was evaluated for values of R in multiples of 0.001 from 5.5 to 6.5. It increased monotonically to a maximum at $R = 6.12$, and then decreased monotonically, exactly as claimed by Mardia and Watkins. However, they did show that multimodality can be a problem when the log likelihood is not everywhere twice differentiable, as happens for the spherical model for example. In any case, as is clear from Fig. 2.13, fitting an exponential variogram with no trend is not a sensible analysis for this data set.

Handcock and Stein (1993) analyzed the same data set by a Bayesian analysis based on the Matérn covariance function. They used a linear trend in x and y together with one additional covariate, the horizontal distance from the survey point to the closest stream. For the present analysis, a linear trend in x and y has been used, though we know from Fig. 2.14 that this is not fully satisfactory either. Fig. 2.15(b) shows a profile likelihood plot in θ_2 which shows that the maximum is attained at about $\theta_2 = 1.19$. In contrast, the posterior density shown in Fig. 3.2 of Handcock and Stein (1993) has a mode attained slightly below $\theta_2 = 1$ and falls off much more sharply on either side of the mode. (For example, from their plot is appears that the posterior density at $\theta_2 = 1.5$ is only about 10% of its maximum value, whereas in Fig. 2.15(b), the value of L at $\theta_2 = 2$ is still half its maximum value.) This shows that the two methods are not in practice equivalent.

Further fits based on the Matérn covariance function with higher-order trend produced log L=77.84 in the quadratic case ($\widehat{\theta}_2 = 1.37$) and log L=85.84 in the cubic case ($\widehat{\theta}_2 = 1.61$), compared with log L=72.74 in the linear trend model. In each case the improvement is highly significant as judged by a χ^2 test with respectively 3 and 4 degrees of freedom, which reinforces the unsatisfactory nature of a simple trend model for this example.

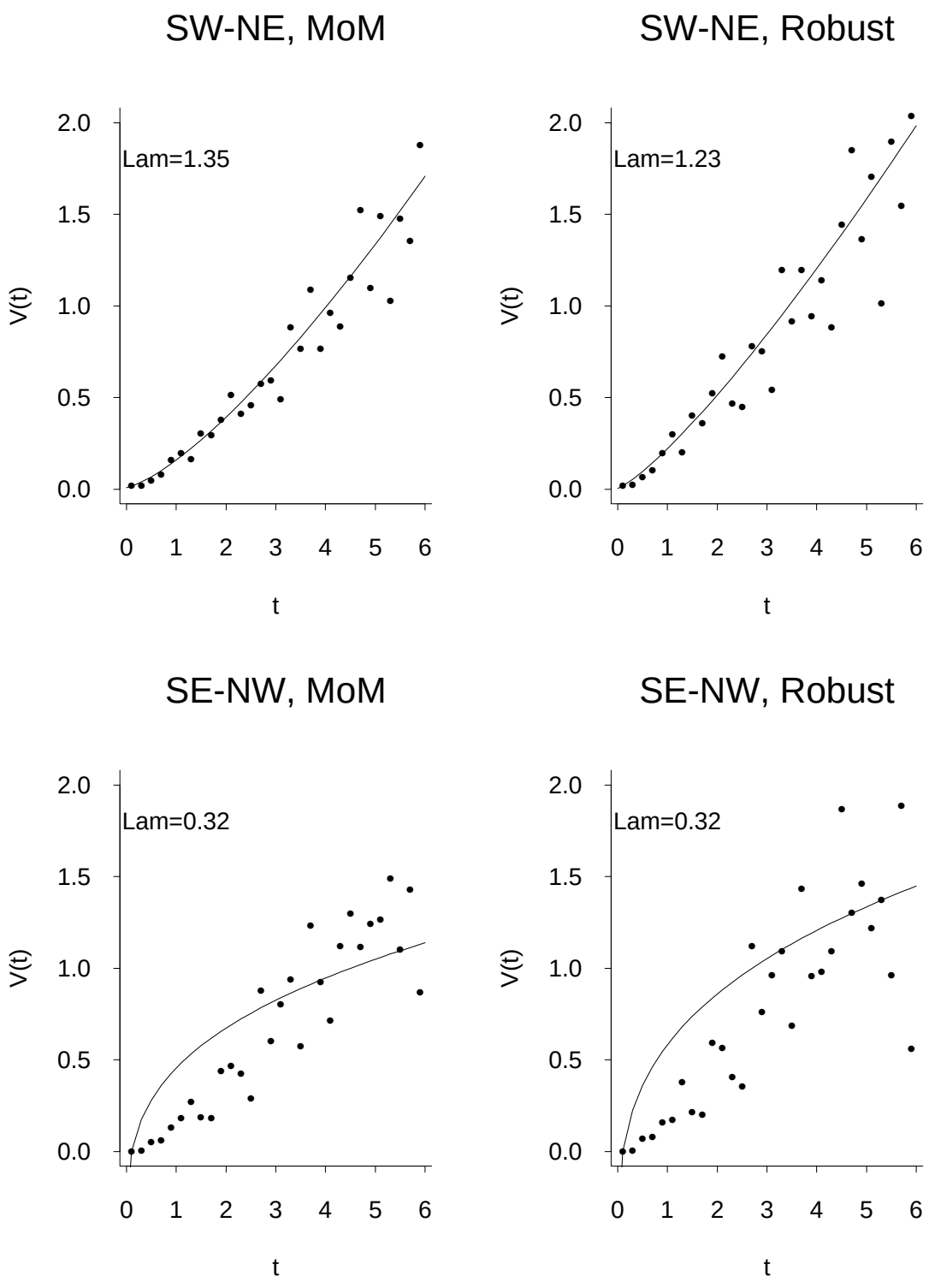


Fig. 2.13. Fitted variograms to Davis data: Raw data with power law models.

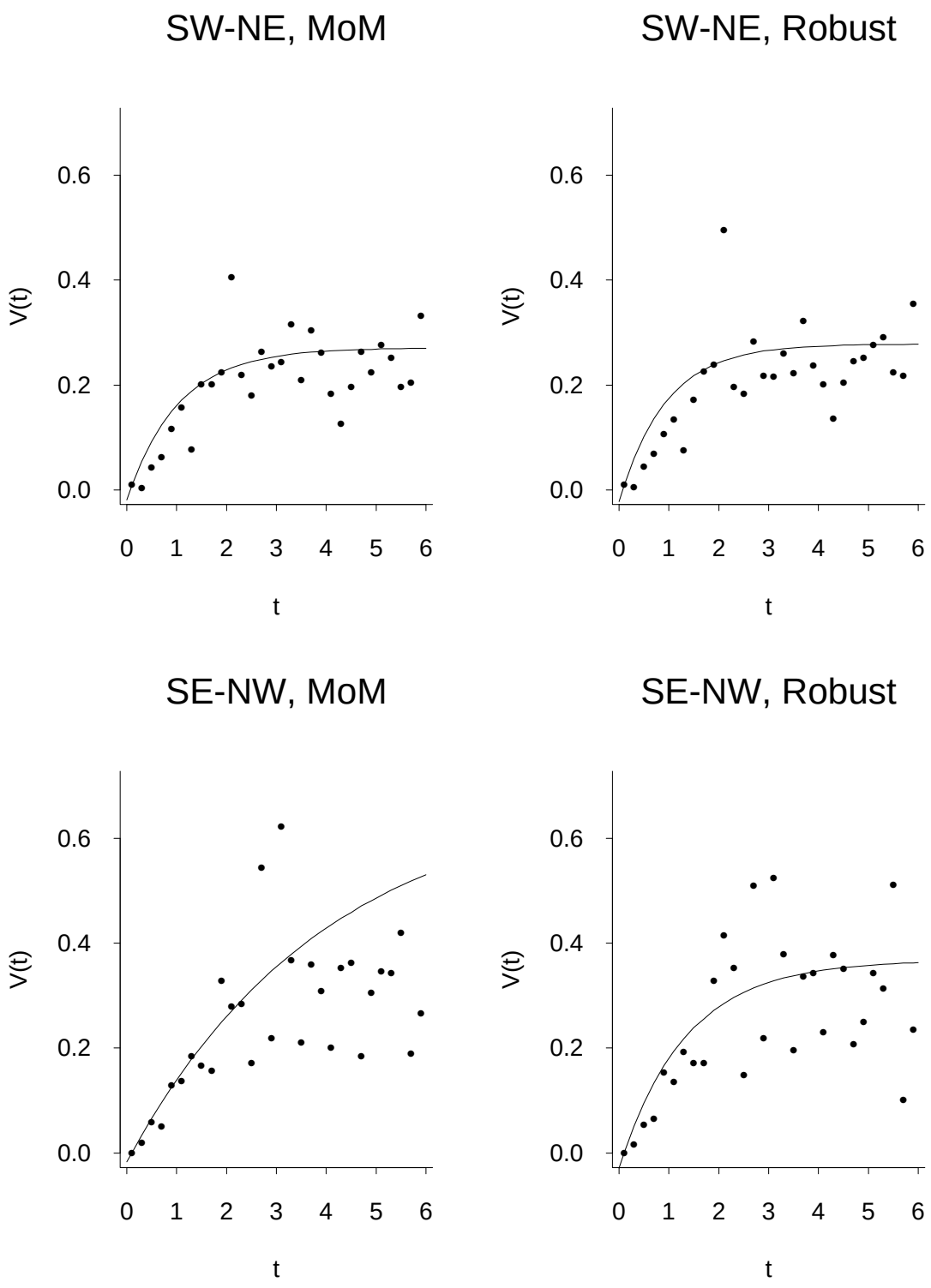


Fig. 2.14. Fitted variograms to detrended Davis data with exponential variogram models.

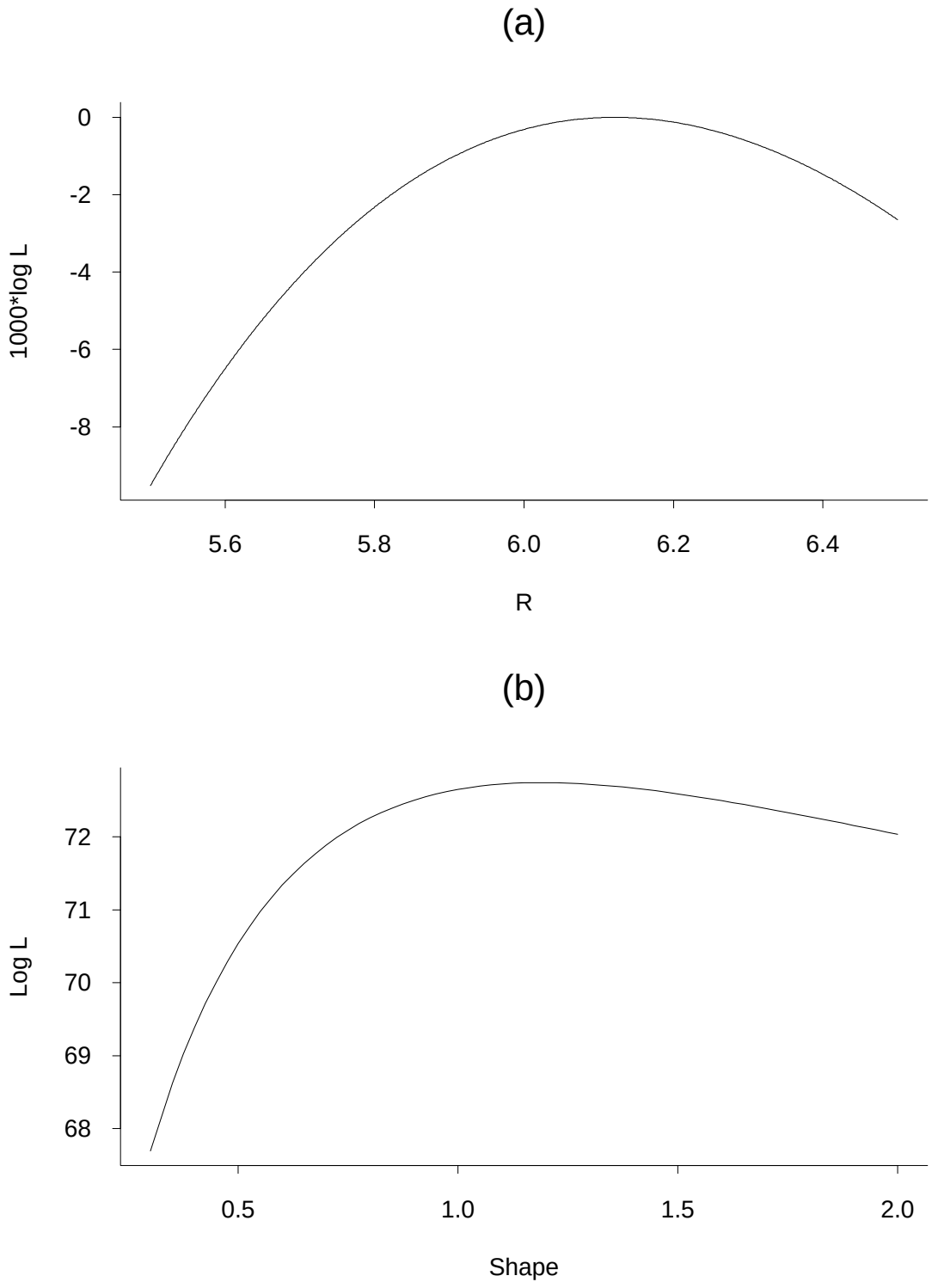


Fig. 2.15. Profile log likelihoods for Davis data. (a) Exponential model fitted to raw data, no trend. (b) Matérn model fitted to data with linear trend.

2.4 Kriging: Prediction and Interpolation

We now turn to the central topic of this subject: the use of spatial covariance models for prediction and interpolation. The name frequently used for this process is *kriging*, though as commonly used, that term refers only to the construction of a spatial predictor in terms of known model parameters, whereas our approach will (ultimately) take the model parameters into account as well, and in that sense, is more general than traditional kriging.

The problem may be stated in the following form: given observations of a vector field $z(s_1), z(s_2), ..., z(s_n)$, predict the value $z(s_0)$, for some $s_0 \notin \{s_1, ..., s_n\}$. A generalization is to predict the joint value at several points, or an integral such as $Z(A) = \int_A z(s)ds$ for some set A — if $z(\cdot)$ measures the density of an ore, then $Z(A)$ measures the total quantity of ore over a region A. However, as we shall see, these problems are generally dealt with as a direct generalization of the methodology for a single point, so we concentrate on that in our initial discussion.

We shall take three approaches to this: an approach based on Lagrange multipliers, an approach based on conditional inference, and a Bayesian approach. The Lagrange multipler approach is the most direct derivation of the kriging estimator and is the one most commonly given in textbooks, but it does not give so much insight into what is going on. The conditional inference approach extends the ideas involved in REML estimation (section 2.2.5) and shows how the kriging predictor may be derived as a conditional mean in an appropriate space of predictors. Finally the Bayesian approach is given, which leads to the same answers as the standard kriging predictor when the model parameters θ are known (in the notation of sections 2.2.4–2.2.6), but it also extends to the case where these parameters are unknown. The reader new to the subject is recommended to pick one of the three approaches and work through the formulae in detail: once the method is fully understood from one approach, it is relatively straightforward to check that the other approaches lead to the same answers.

2.4.1. Lagrange multiplier approach

Let us write the vector $Z = (z(s_1), ..., z(s_n))^T$ and $z_0 = z(s_0)$. We need to know the joint covariance matrix of Z and z_0; let us suppose

$$\mathrm{Cov}\left\{\begin{pmatrix} Z \\ z_0 \end{pmatrix}\right\} = \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma_0^2 \end{pmatrix}, \tag{2.51}$$

where Σ is the covariance matrix of Z, σ_0^2 is the variance of z_0 and τ is the vector of cross-covariances between Z and z_0. For some of our calculations, following (2.31), we shall write

$$\Sigma = \alpha V(\theta), \quad \tau = \alpha w(\theta), \quad \sigma_0^2 = \alpha v_0(\theta) \tag{2.52}$$

in terms of the scale parameter α and functions V, w and v_0 of a finite-dimensional parameter θ.

The basic model will be assumed to be as in (2.30), with $Z = X\beta + \eta$ for some matrix of covariates X, and we also assume $z_0 = x_0^T\beta + \eta_0$ for some given vector x_0, the vector of covariates at z_0 (or s_0, if we are thinking in terms of the original stations). Both η and η_0 represent random errors with mean 0. This is the *universal kriging* problem; the special case

$$\mathrm{E}\{Z\} = \mu\mathbf{1}, \quad \mathrm{E}\{z_0\} = \mu, \tag{2.53}$$

in which $\mathbf{1}$ denotes the n-vector of ones and μ is some overall constant, is the *ordinary kriging* problem in which there is an unknown common mean but no other regression coefficient. We consider predictors of form

$$\widehat{z}_0 = \lambda^T Z, \tag{2.54}$$

subject to the constraint

$$\lambda^T X = x_0^T. \tag{2.55}$$

The reason for the constraint (2.55) will appear momentarily.

Let us consider the prediction error in (2.54). We may write

$$\begin{aligned} z_0 - \widehat{z}_0 &= x_0^T\beta + \eta_0 - \lambda^T(X\beta + \eta) \\ &= \eta_0 - \lambda^T\eta \end{aligned} \tag{2.56}$$

where we have used the constraint (2.55) — in other words, the reason for this constraint is to make the procedure work without assuming β is known. The reader might at this point be wondering why we make such a big fuss about β being unknown when we are implicitly assuming that θ (or the covariances Σ and τ) are known — this is a valid point to raise, but we return to it later.

If we assume (2.55) and hence (2.56), the mean squared prediction error becomes

$$\mathrm{E}\{(z_0 - \widehat{z}_0)^2\} = \sigma_0^2 - 2\lambda^T\tau + \lambda^T\Sigma\lambda. \tag{2.57}$$

We are therefore led to the following constrained optimization problem: *minimize (2.57) subject to (2.55).*

Solution to the constrained optimization problem

Consider the Lagrangian

$$L = \sigma_0^2 - 2\lambda^T\tau + \lambda^T\Sigma\lambda - 2(\lambda^T X - x_0^T)\nu, \tag{2.58}$$

where 2ν is a vector of Lagrange multipliers.

According to the theory of Lagrange multipliers, subject to suitable differentiability conditions (which are trivially valid for this problem, because the problem is quadratic

in λ), the optimal λ will be attained at some stationary point of L. Differentiating with respect to the components of λ in (2.58), this is achieved when

$$0 = -\tau + \Sigma\lambda - X\nu,$$

or in other words $\lambda = \Sigma^{-1}(\tau + X\nu)$. To find ν, substitute back in (2.55) to get

$$\nu = (X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau).$$

The final result is

$$\lambda = \Sigma^{-1}\tau + \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau), \tag{2.59}$$

or the predictor

$$\widehat{z}_0 = \lambda^T Z = (x_0 - X^T\Sigma^{-1}\tau)^T\widehat{\beta} + \tau^T\Sigma^{-1}Z. \tag{2.60}$$

The resulting prediction error variance (2.57) becomes

$$\begin{aligned}
&\sigma_0^2 - 2\lambda^T\tau + \lambda^T\Sigma\lambda \\
&= \sigma_0^2 - 2\tau^T\Sigma^{-1}\tau - 2\tau^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) \\
&\quad + \tau^T\Sigma^{-1}\tau + 2\tau^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) \\
&\quad + (x_0 - X^T\Sigma^{-1}\tau)^T(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) \\
&= \sigma_0^2 - \tau^T\Sigma^{-1}\tau + (x_0 - X^T\Sigma^{-1}\tau)^T(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau).
\end{aligned} \tag{2.61}$$

Since it will come in useful later, we also give the extension of this calculation to handle the prediction covariance between two stations. Suppose, instead of a single unobserved z_0, we have to predict two variables $z_a = x_a^T\beta + \eta_a$, $z_b = x_b^T\beta + \eta_b$ corresponding to two unobserved stations s_a and s_b. Suppose the joint covariance matrix is given by

$$\mathrm{Cov}\left\{\begin{pmatrix} Z \\ z_a \\ z_b \end{pmatrix}\right\} = \begin{pmatrix} \Sigma & \tau_a & \tau_b \\ \tau_a^T & \sigma_{aa} & \sigma_{ab} \\ \tau_b^T & \sigma_{ab} & \sigma_{bb} \end{pmatrix}, \tag{2.62}$$

and the optimal predictors are $\widehat{z}_a = \lambda_a^T Z$, $\widehat{z}_b = \lambda_b^T Z$ where

$$\begin{aligned}
\lambda_a &= \Sigma^{-1}\tau_a + \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_a - X^T\Sigma^{-1}\tau_a), \\
\lambda_b &= \Sigma^{-1}\tau_b + \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_b - X^T\Sigma^{-1}\tau_b).
\end{aligned} \tag{2.63}$$

The mean squared prediction error is then

$$\begin{aligned}
&\mathrm{E}\{(z_a - \lambda_a^T Z)(z_b - \lambda_b^T Z)\} \\
&\quad = \sigma_{ab} - \lambda_a^T\tau_b - \lambda_b^T\tau_a + \lambda_a^T\Sigma\lambda_b \\
&\quad = \sigma_{ab} - \tau_b^T\left\{\Sigma^{-1}\tau_a + \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_a - X^T\Sigma^{-1}\tau_a)\right\} \\
&\qquad - \tau_a^T\left\{\Sigma^{-1}\tau_b + \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_b - X^T\Sigma^{-1}\tau_b)\right\} \\
&\qquad + \tau_b^T\Sigma^{-1}\tau_a + \tau_b^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_a - X^T\Sigma^{-1}\tau_a) \\
&\qquad + \tau_a^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_b - X^T\Sigma^{-1}\tau_b) \\
&\qquad + (x_a - X^T\Sigma^{-1}\tau_a)(X^T\Sigma^{-1}X)^{-1}(x_b - X^T\Sigma^{-1}\tau_b) \\
&\quad = \sigma_{ab} - \tau_b^T\Sigma^{-1}\tau_a + (x_a - X^T\Sigma^{-1}\tau_a)(X^T\Sigma^{-1}X)^{-1}(x_b - X^T\Sigma^{-1}\tau_b).
\end{aligned} \tag{2.64}$$

Essentially the same calculations are given in a number of other books on spatial statistics, e.g. pp. 48–49 of Ripley (1981) or pp. 154–155 of Cressie (1993). Note that the formulae are sometimes expressed in terms of the variogram rather than the covariance matrices but, at least for stationary processes, it is straightforward to pass from one to the other. Stationarity itself plays no role in the prediction formulae we have derived, though it is usual, for the reasons explained in earlier sections, to work with either stationary or intrinsically stationary processes.

2.4.2. Conditional inference approach

This is "conditional" in the sense that it exploits the decomposition implicit in the derivation of REML estimation: by decomposing the vector Z into a component which is essentially $\widehat{\beta} - \beta$ and a component which is orthogonal to that, and conditioning on the latter, we can derive the kriging predictor from this point of view.

As a first step, we assume β is known. We use the following classical result of multivariate analysis (see, e.g., Mardia, Kent and Bibby (1979), p. 63): if we consider a partitioned multivariate normal vector

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right\},$$

then the conditional distribution of X_1 given X_2 is normal with mean

$$\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \tag{2.65}$$

and variance

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{2.66}$$

Applying this result with $\Sigma_{11} = \sigma_0^2$, $\Sigma_{12} = \tau$, $\Sigma_{22} = \Sigma$ in our previous notation, we deduce that the conditional distribution of z_0 given Z is normal with mean

$$x_0^T\beta + \tau^T\Sigma^{-1}(Z - X\beta) \tag{2.67}$$

and variance

$$\sigma_0^2 - \tau^T\Sigma^{-1}\tau. \tag{2.68}$$

When β is unknown, the obvious solution is to substitute $\widehat{\beta}$ for β in (2.67). This leads to the proposed predictor

$$\widehat{z}_0 = x_0^T\widehat{\beta} + \tau^T\Sigma^{-1}(Z - X\widehat{\beta}) = \lambda^T Z \tag{2.69}$$

say, where (as the reader may easily check) λ is given by (2.59). Thus, this argument leads very quickly to the correct formula for $\widehat{z}_0$, though it does not so far prove that it has any properties which might make it desirable as a predictor. Note that the equation (2.55), which formed a key part of our earlier derivation, follows directly from (2.69). Because

of this, the prediction error $z_0 - \lambda^T Z$ has mean $x_0^T\beta - \lambda^T X\beta = 0$, and variance given by (2.61), as before.

The key step of this proof is to note that $z_0 - \lambda^T Z$ is independent of $Z - X\widehat{\beta}$. Since $Z - X\widehat{\beta}$ is in one to one correspondence with the vector $W = A^T Z$ which is used to define the REML estimator (section 2.2.5), this establishes that the conditional distribution of z_0 given W is normal with mean $\lambda^T Z$ and variance given by (2.61).

To establish the independence just referred to, write $Z - X\widehat{\beta} = RZ$ where

$$R = I - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1},$$

and note that the covariance of RZ and $z_0 - \lambda^T Z$ is

$$\mathrm{E}\{R(Z - X\beta)(z_0 - x_0^T\beta - (Z - X\beta)^T\lambda)\} = R(\tau - \Sigma\lambda)$$

and because everything is jointly normal, it will suffice to prove that the latter quantity is 0. However

$$\tau - \Sigma\lambda = -X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau),$$

and hence

$$\begin{aligned} R(\tau - \Sigma\lambda) &= -\{I - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\}X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) \\ &= -X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) + X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) \\ &= 0, \end{aligned}$$

which establishes the desired result.

To complete the proof that this solves the kriging problem, we need to show that if $\tilde{z}_0 = \tilde{\lambda}^T Z$ is some other predictor, satisfying (2.55) with $\tilde{\lambda}$ in place of λ, then $\mathrm{E}(\tilde{z}_0 - z)^2 \geq \mathrm{E}(\hat{z}_0 - z)^2$, with equality if and only if $\tilde{\lambda} = \lambda$.

Since the transformation from Z to $(\hat{\beta}, W^T)^T$ is one-one, we may write $\tilde{z}_0 = \hat{z}_0 + c_1^T\hat{\beta} + c_2^T W$ for arbitrary vectors c_1 and c_2. However, in that case $\mathrm{E}(\tilde{z}_0 - z_0) = \mathrm{E}(\hat{z}_0 - z_0) + c_1^T\beta + 0 = c_1^T\beta$, which is identically 0 if and only if $c_1 = 0$. By uncorrelatedness of $\hat{z}_0 - z_0$ and W,

$$\mathrm{E}\{\tilde{z}_0 - z_0)^2\} = \mathrm{E}\{\hat{z}_0 - z_0)^2\} + \mathrm{E}\{(c_2^T W)^2\},$$

where the second term is ≥ 0. Because the matrix A is of rank $n - q$, the second term is 0 only if $c_2 = 0$. Therefore, the two properties of unbiasedness and minimising the mean squared prediction error can be satisfied only if $c_1 = c_2 = 0$. This completes the proof.

2.4.3. Bayesian approach

The fact that the preceding argument is equivalent to a Bayesian approach has been noted in other contexts, e.g. it lies at the heart of Meinhold and Singpurwalla's (1983)

derivation of the Kalman filtering equations from purely Bayesian considerations. The Bayesian approach generalizes automatically to the case in which the variogram parameters are unknown, whereas the classical approach essentially makes the assumption that these are known and only deals with the question of uncertainty of model parameters in a very peripheral way. This is one major reason for viewing the problem in Bayesian terms, and the close parallels between this and the more traditional approaches of sections 2.4.1 and 2.4.2 adds to its justification.

The model throughout this discussion is the same as in section 2.4.1, writing the covariances in the form of (2.52) and taking (2.43) as the prior. The choice of $\pi(\theta)$ is largely arbitrary, but the equivalence of Bayesian and least-squares approach works only for the "classical" noninformative prior on (β, α).

Simplest case: β, α, θ all known

This follows as in section 2.4.2:

$$\{z_0|Z,\beta,\alpha,\theta\} \sim \mathcal{N}[(x_0 - X^T\Sigma^{-1}\tau)^T\beta + \tau^T\Sigma^{-1}Z, \sigma_0^2 - \tau^T\Sigma^{-1}\tau]. \tag{2.70}$$

Note that Σ, τ and σ_0^2 may all be written in terms of α and θ, using (2.52).

We shall now improve upon (2.70) by, successively, removing the conditioning on β, α and θ. We write $\pi(x|y)$ for the generic density of one variable x conditioned on another variable y where the variables x and y will be different from one usage to the next.

To remove the conditioning on β, we write

$$\pi(z_0|Z,\alpha,\theta) = \int \pi(z_0|Z,\beta,\alpha,\theta)\pi(\beta|Z,\alpha,\theta)d\beta \tag{2.71}$$

where the first factor inside the integral is given by (2.70) and the second derived from (2.44). Note that it follows from (2.44) that the posterior distribution of β, given Z, α and θ, is multivariate normal with mean $\widehat{\beta} = \widehat{\beta}(\theta)$ (i.e. the GLS estimator of β given the covariance matrix $V(\theta)$), and covariance matrix $\alpha(X^TV(\theta)^{-1}X)^{-1}$. Combining this with (2.70), we find that the conditional distribution of z_0 given α and θ is multivariate normal with mean

$$\begin{aligned}\widehat{z}_0(\theta) &= (x_0 - X^T\Sigma^{-1}\tau)^T\widehat{\beta} + \tau^T\Sigma^{-1}Z \\ &= (x_0 - X^TV(\theta)^{-1}w(\theta))^T\widehat{\beta} + w(\theta)^TV(\theta)^{-1}Z\end{aligned} \tag{2.72}$$

and covariance matrix

$$\begin{aligned}&(x_0 - X^T\Sigma^{-1}\tau)^T(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) + \sigma_0 - \tau^T\Sigma^{-1}\tau \\ &\quad = \alpha\{(x_0 - X^TV(\theta)^{-1}w(\theta))^T(X^TV(\theta)^{-1}X)^{-1}(x_0 - X^TV(\theta)^{-1}w(\theta)) \\ &\qquad + v_0(\theta) - \tau^TV(\theta)^{-1}\tau\} \\ &\quad = \alpha V_0(\theta) \quad \text{say.}\end{aligned} \tag{2.73}$$

The next step is to remove the conditioning on α. Similarly to (2.71), we have

$$\pi(z_0|Z,\theta) = \int \pi(z_0|Z,\alpha,\theta)\pi(\alpha|Z,\theta)d\alpha. \tag{2.74}$$

The posterior distribution of α, given Z and θ, may be obtained from (2.45): the result is that $G^2(\theta)/\alpha$ has a χ^2_{n-q} distribution. Define

$$\widehat{\alpha}(\theta) = \frac{G^2(\theta)}{n-q}.$$

Then with slight abuse of notation, we have

$$(\alpha|Z,\theta) \sim \widehat{\alpha}(\theta)\frac{n-q}{\chi^2_{n-q}}.$$

Conditionally on Z and θ, we then have

$$\begin{aligned}\frac{z_0 - \widehat{z}_0(\theta)}{\sqrt{\widehat{\alpha}(\theta)V_0(\theta)}} &= \frac{z_0 - \widehat{z}_0(\theta)}{\sqrt{\alpha V_0(\theta)}} \cdot \sqrt{\frac{\alpha}{\widehat{\alpha}(\theta)}} \\ &\sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-q}/(n-q)}} \\ &\sim t_{n-q}\end{aligned} \tag{2.75}$$

since the numerator and denominator in (2.75) are conditionally independent given Z and θ.

The final result agrees with equation (3.1) of Handcock and Stein (1993), except that they have a factor $n/(n-q)$ multiplying $\widehat{\alpha}(\theta)$, which results from a slightly different definition of the latter quantity.

Finally, following Handcock and Stein, we integrate over θ to obtain

$$\pi(z_0|Z) = \int \pi(z_0|Z,\theta)\pi(\theta|Z)d\theta \tag{2.76}$$

where the first factor in the integrand is determined by (2.48) and the second by (2.46). This part has to be carried out numerically, but should be straightforward since for most models of interest the dimension of θ is 2 or at most 3. Handcock and Stein give several examples.

An aside: Besag's candidate's formula

An interesting alternative version of (2.76) is the formula

$$\pi(z_0|Z) = \frac{\pi(z_0|z,\theta)\pi(\theta|Z)}{\pi(\theta|z_0,Z)}, \tag{2.77}$$

given by Besag (1989), allegedly based on a student's answer to an examination question. Note that in (2.77), θ is fixed (and arbitrary) — there is no integration in this formula, at least not explicitly.

The derivation of (2.77) is an immediate consequence of the fact that

$$\pi(z_0|Z)\pi(\theta|z_0,Z) \quad \text{and} \quad \pi(z_0|z,\theta)\pi(\theta|Z)$$

are each equal to the joint density of z_0 and θ, conditional on Z. The formula is likely to be especially useful when $\pi(\theta|Z)$ and $\pi(\theta|z_0,Z)$ can each be calculated analytically, e.g. as the result of a conjugate prior calculation. This is rarely the case in spatial statistics, and in practice (2.77) does require numerical integration with respect to θ, to determine the normalizing constants. It is therefore not clear whether Besag's candidate's formula brings any practical benefits for kriging, but the formula seems worth knowing about.

2.4.4. Prediction at multiple sites

The formulae (2.63) and (2.64) are very easily extended to multiple sites. For example, if $s_a, s_b, ..., s_q$ are several sites for prediction with associated covariate vectors $x_a, x_b, ..., x_q$ and covariances $\tau_a, ..., \tau_q, \sigma_{aa}, ..., \sigma_{qq}$ defined by analogy with (2.62), then the optimal predictor of

$$c_a z_a + c_b z_b + ... + c_q z_q,$$

where $z_a, ..., z_q$ are the unobserved values of the process at $s_a, ..., s_q$ and $c_a, ..., c_q$ are arbitrary scalars, is

$$(c_a\lambda_a + c_b\lambda_b + ... + c_q\lambda_q)^T Z$$

with $\lambda_a, ...$ given by an obvious extension of (2.63). The prediction variance is obtained as the sum of terms of the form

$$c_a c_b \mathrm{cov}(z_a - \lambda_a^T Z, z_b - \lambda_b^T Z)$$

with (a,b) ranging over all possible pairs of indices and the covariances given by (2.64).

The extension to predicting quantities of form

$$Z(A) = \int_A z(s)ds,$$

where A is some subset of the observation space, should now be clear. The point predictor is

$$\widehat{Z}(A) = \int_A \widehat{z}(s)ds, \tag{2.78}$$

where $\widehat{z}(s)$ is the predictor at the site s, and the prediction error variance is

$$\mathrm{E}\{(Z(A) - \widehat{Z}(A))^2\} = \int_A \int_A \mathrm{E}[\{z(s_1) - \widehat{z}(s_1)\}\{z(s_2) - \widehat{z}(s_2)\}]ds_1 ds_2, \tag{2.79}$$

where the individual covariance terms are derived from (2.64).

These calculations have been presented for the most commonly analyzed scenario in which β is unknown and α and θ are known. The case where all three parameters are unknown is, at the present time, probably best handled by Bayesian techniques, where the required generalizations from the case where β alone is unknown to where all the parameters are unknown follows as in section 2.4.3.

2.4.5. Frequentist corrections for unknown covariance structure

Let us now revert to the frequentist viewpoint and consider various strategies that may be adopted for the case when θ is unknown.

As already pointed out, when θ is known, the predictor $\widehat{z}_0(\theta)$ defined by (2.72) is the best linear unbiased predictor of z_0, with mean squared prediction error $m(\theta)$ given by (2.73). For the purpose of the present discussion, in the case that α is also unknown we absorb it into θ.

When θ is unknown and estimated by $\widehat{\theta}$, using any of the estimation methods we have discussed, the obvious strategy is to use $\widehat{z}_0(\widehat{\theta})$ as a predictor of z_0 and to cite $m(\widehat{\theta})$ as its MSPE. This suffers from the objection that $m(\widehat{\theta})$ makes no allowance for the discrepancy between $\widehat{\theta}$ and θ and may therefore be expected to underestimate the true MSPE. There are essentially three strategies one can adopt in response to this objection,

(i) Ignore it, i.e. use $\widehat{z}_0(\widehat{\theta})$ as the predictor of z_0 and $m(\widehat{\theta})$ as the MSPE even when θ is unknown,

(ii) Estimate the discrepancy between $m(\widehat{\theta})$ and the true MSPE, and use that to derive a corrected MSPE,

(iii) Adopt the Bayesian procedure (2.76) or (2.77) in the hope that it will have good properties from a frequentist as well as Bayesian point of view.

Strategies of the form (ii) have been considered by a number of authors, in particular Prasad and Rao (1990) and Harville and Jeske (1992) in the case of variance components models, and by Zimmerman and Cressie (1992) in the (more general) cases arising from spatial covariance matrices. The main approximations can be viewed as variants of the delta method, but there are many such variants and no clear-cut guidelines as to which performs best. The following discussion is intended to do no more than present the main outlines of the arguments; for details we must refer to the original papers.

Suppose $e_1 = z_0(\theta) - z_0$ is the prediction error using the optimal predictor when θ is known, and let $e_2 = z_0(\widehat{\theta}) - z_0(\theta)$ denote the additional prediction error arising from θ being unknown. We shall assume all distributions are multivariate normal. A very general

argument for these kinds of models shows that e_1 is independent of the current observation vector Z, and hence of e_2 (which is a function of Z). It follows at once that

$$\mathrm{E}\{(z_0(\widehat{\theta}) - z_0)^2\} = \mathrm{E}\{(e_1 + e_2)^2\} \geq \mathrm{E}\{e_1^2\} = m(\theta),$$

so that $m(\theta)$ is indeed an underestimate of the true MSPE when θ is unknown. Moreover, if we approximate

$$e_2 \approx (\widehat{\theta} - \theta)^T \nabla \widehat{z}_0(\theta)$$

(with ∇ denoting gradient) and apply the same argument a second time, we deduce

$$\begin{aligned}
\mathrm{Var}\{e_2\} &\approx \mathrm{E}[\nabla \widehat{z}_0(\theta)^T (\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T \nabla \widehat{z}_0(\theta)] \\
&= \mathrm{E}[tr\{\nabla \widehat{z}_0(\theta)^T (\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T \nabla \widehat{z}_0(\theta)\}] \\
&= \mathrm{E}[tr\{(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T \nabla \widehat{z}_0(\theta) \nabla \widehat{z}_0(\theta)^T\}] \\
&= tr[\mathrm{Cov}\{\widehat{\theta}\} \cdot \mathrm{Cov}\{\nabla \widehat{z}_0(\theta)\}]
\end{aligned}$$

where in the middle of this argument we used the matrix identity $tr(AB) = tr(BA)$. Therefore, it appears that we ought to estimate the mean squared prediction error by

$$\mathrm{Var}\{\widehat{z}_0(\widehat{\theta})\} \approx m(\widehat{\theta}) + \mathrm{tr}[\mathrm{Cov}\{\widehat{\theta}\} \cdot \mathrm{Cov}\{\nabla \widehat{z}_0(\theta)\}] \tag{2.80}$$

and this will improve on the crude approximation $m(\widehat{\theta})$.

Harville and Jeske (1992), extending the earlier argument given by Prasad and Rao (1990), argued that even (2.80) is not the final answer, because although (2.80) adjusts for the difference between $m(\theta)$ and the true MSPE, there is an additional bias corresponding to the fact that $m(\widehat{\theta})$ typically underestimates $m(\theta)$. This additional bias turns out to be asymptotically equivalent to the bias in (2.80) itself, so an improved approximation is to double the correction term:

$$\mathrm{Var}\{\widehat{z}_0(\widehat{\theta})\} \approx m(\widehat{\theta}) + 2\ \mathrm{tr}[\mathrm{Cov}\{\widehat{\theta}\} \cdot \mathrm{Cov}\{\nabla \widehat{z}_0(\theta)\}]. \tag{2.81}$$

Zimmerman and Cressie (1992) cited this result and a series of conditions required for its asymptotic validity, but they were cautious about recommending its use in practice. A general conclusion which they reached was that when spatial correlation is weak the improved approximations to the MSPE can indeed improve substantially on $m(\widehat{\theta})$, but when spatial correlation is strong it is often preferable to use $m(\widehat{\theta})$ as the estimator of MSPE.

The third strategy mentioned at the beginning of this subsection is to continue to use the Bayesian solution in the belief that Bayesian predictive distributions also have good frequentist properties. At the present time, there is a growing literature on the use of second-order asymptotic theory to improve upon naïve predictive distributions, but none

of this literature so far has derived specific solutions to problems involving unknown covariance matrices. General references include Komaki (1996) and Smith (1997), who have explicitly considered Bayesian predictive procedures as a means of obtaining predictive distributions with good frequentist properties, and Barndorff-Nielsen and Cox (1996), whose approach is not at all Bayesian. An early example of the kind of result showing that a Bayesian predictive procedure may improve on a simple estimative procedure was the paper of Aitchison (1975). Based on the results of Komaki and Smith, one can say that there is good reason to believe that some form of Bayesian solution will provide a good solution to the problem, but the results may well be sensitive to the specification of the prior distribution and also to the loss function. Detailed examination of these issues in the context of spatial prediction would appear to be an important area for future research.

2.4.6. Model misspecification in kriging

Some of the other issues involved in applying kriging, and some other approaches to issues that we have discussed, will be reviewed briefly here, without detailed discussion.

We have given detailed attention to the estimation of covariance structure and how either Bayesian or frequentist prediction intervals should be modified to take account of this structure being estimated. A somewhat simpler approach is to act as if the covariance structure was known, but to develop a sensitivity analysis to the misspecification of covariance structure. This approach was taken by Warnes (1986). Using Taylor expansions of covariance matrices and their inverses, he developed first-order approximations to the error in universal kriging when a parameter of the covariance model is misspecified. As examples he worked out the consequences of this for the exponential and Gaussian models, using the data of Table 2.7 (without any trend in the model) as illustration. His results imply that the Gaussian model is much more sensitive to misspecification than the exponential model though it was not clear from his discussion to what extent this was a feature of that particular data set.

An earlier paper to take a "perturbation approach" to this problem was Diamond and Armstrong (1984). They examined the properties of kriging procedures under assumptions of the form $1 - \delta < g(h)/\gamma(h) < 1 + \delta$ for all t, where $\gamma(h)$ is the true semivariogram at distance h and $g(h)$ is the semivariogram assumed by the analyst. Using numerical analysis methods, they showed that if Γ is the $n \times n$ matrix with entries $\gamma(s_i - s_j)$, where $s_1, ..., s_n$ are the sampling points, then the relative error in the kriging coefficients may be bounded by an expression which depends only on δ and the condition number of Γ. They gave corresponding expressions for the change in the prediction variance resulting from misspecification of γ, and also considered the effects of misspecifying the sampling points $\{s_i\}$, and misspecifying the regression component in universal kriging.

A quite different approach to the whole question of kriging with misspecified covariances has been taken in several papers of M. Stein, see e.g. Stein (1987, 1988), Stein and Handcock (1989). Stein has considered the whole problem from the point of view

of "infill asymptotics", i.e. assume the domain of observation is fixed but the number of observations increases so as to fill up the domain in a dense way.

As an example of this approach, consider the paper Stein (1987). Stein considered a one-dimensional model with the variogram

$$\gamma(t;\theta) = \theta_1 t + \theta_2(t - \frac{1}{2}t^2). \tag{2.82}$$

Stein considered this model using the MINQE method of estimation which was described in subsection 2.2.7.

Considering the case in which n observations are equally spaced over the interval $[0,1]$, Stein showed that the estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are not consistent as $n \to \infty$, but the combination $\widehat{\theta}_1 + \widehat{\theta}_2$ is consistent as an estimator of $\theta_1 + \theta_2$. He interpreted this by noting that $\theta_1 + \theta_2$ is the gradient of $\gamma(t)$ at $t = 0$, i.e. the *local* behavior of $\gamma(t)$ near $t = 0$ is successfully estimated by this procedure, but the global behavior with $t \neq 0$ is not. However, he also argued that for kriging purposes, i.e. for estimating $Z(s)$ at some s other than one of the sampling points $0, 1/(n-1), ..., 1$, the local behavior of the variogram near $t = 0$ is what matters, and we do get asymptotically efficient predictors based on the estimated θ_1 and θ_2.

Stein (1988) showed that this was part of a general phenomenon in which the required property was *compatibility* of the assumed and true variograms. That is to say, if the two variograms are compatible then kriging based on the assumed variogram will be asymptotically efficient under infill asymptotics.

In more detail, suppose we observe a spatial process on a bounded region $R \subset \mathcal{R}^d$ and let $C(s,t),\ s,t \in R$ denote the covariance function. Stein defined two covariance functions $C_0(\cdot,\cdot)$ and $C_1(\cdot,\cdot)$ to be *compatible* if the Gaussian processes with zero means and covariance functions C_0 and C_1 are mutually absolutely continuous on R. Note that it is not required that the actual observed process be Gaussian, but the definition of compatibility is framed in terms of a Gaussian process with the same covariance structure. Suppose s is a fixed point in R and $\{s_i,\ i = 1, 2, ..., \}$ a sequence of sampling points in R. It is assumed that s is not a member of $\{s_i\}$ but that it is a limit point of $\{s_i\}$. Let $e_i(N)$ $(i = 0, 1)$ denote the prediction error at s of the optimal kriging predictor derived under the covariance function C_i, and let $V_0(\cdot)$ denote the variance of a random quantity under the true covariance function C_0. Note that, since the optimal kriging predictors are linear functions of the data points, $V_0(e_i(N))$ will be the same for a non-Gaussian process as it is for a Gaussian process with the same covariance C_0.

Under all these conditions Stein's (1988) main result (Theorem 1) is the following: if

$$\lim_{N\to\infty} V_0(e_0(N)) = 0,$$

then

$$\lim_{N\to\infty} \frac{V_0(e_0(N))}{V_0(e_1(N))} = 1.$$

In words, assuming compatibility, the kriging predictor derived under the incorrect covariance function C_1 will be asymptotically efficient relative to that derived under the correct function C_0.

Precise conditions required to ensure compatibility are not easy to specify, but loosely it requires that $C_0(s,t)$ and $C_1(s,t)$ behave similarly as $t \to s$. For example, suppose we have a stationary process in one dimension, and the covariance function C_0 is $2m$ times continuously differentiable on a region $[0,T]$ for some $T > 0$, and $C_0^{(2m+1)}(0+) \neq 0$. A necessary though not always sufficient condition for another stationary covariance $C_1(\cdot)$ to be compatible with $C_0(\cdot)$ is that C_1 is also $2m$ times continuously differentiable on $[0,T]$ and $C_0^{(2m+1)}(0+) = C_1^{(2m+1)}(0+)$. For example, $C_0(t) = \lambda_0 e^{-|t|/\lambda_0}$ and $C_1(t) = \lambda_1 e^{-|t|/\lambda_1}$ are compatible: in this case $m = 0$ and $C_0'(0+) = C_1'(0+) = -1$. This example was given by Stein and Handcock (1989), who remarked that similar conditions also apply in higher dimensions.

Using the notion of compatibility, Stein and Handcock (1989) argued against the spherical variogram function on the grounds that it is incompatible with the exponential variogram function – even though the parameters may be chosen so that the first-order derivatives of the two variograms agree at the origin, this is insufficient for compatibility. They also argued that compatibility explains the results of Warnes (1986). Two exponential variograms with different ranges may be made compatible with appropriate choices of the scale parameters. However, this is not true for the Gaussian variogram. Thus a Gaussian variogram with misspecified range parameter cannot give efficient kriging merely by adjusting the scale parameter. However, it is also clear (because the orders of differentiability at the origin are different) that the Gaussian and exponential variograms are never compatible with each other.

These results of Stein and his co-authors should be contrasted with the more traditional results based on *increasing domain asymptotics*, in which the size of the sampling region is assumed to increase as the sample size tends to ∞. This was the form of asymptotics used, for example, in our discussion of the approximate WLS procedure in section subsection 2.2.3, and for Mardia and Marshall's (1984) proof of the asymptotic properties of MLE, which we mentioned in subsection 2.2.4. Which kind of asymptotics are really appropriate is still a matter for debate. The increasing domain asymptotics certainly give results which are more compatible with traditional asymptotic theory; on the other hand, Stein and others have argued that infill asymptotics are a more realistic representation of actual sampling procedures. Probably there is no ultimate resolution of this argument, but the user of either kind of asymptotics needs to be aware of the kinds of assumptions that are being made to justify the procedures.

Zimmerman and Zimmerman (1991) compared several estimators and corresponding kriging predictors in a Monte Carlo study. They assumed one of two basic variogram

models – linear and exponential – and considered seven estimators, including Cressie's WLS estimators based on both the method of moments and robust estimates of variogram, the OLS estimate based on the method of moments variogram, and ML and REML. They concluded, broadly, that the OLS or either version of WLS estimator perform, in practice, just as well as the more computationally demanding ML and REML. However, all their Monte Carlo results are based on Gaussian processes and there are certainly questions to be asked about the extent of applicability of their results.

2.4.7. Median polish kriging

This is a quite different approach to the whole subject. Instead of attempting "optimal" reconstruction of an unobserved stochastic process, this method is based on robustly reconstructing an irregular surface. The account given here is adapted from Cressie (1993), pages 183–194.

The basic idea is to write the process in the form

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

with μ and unknown mean function which is assumed to be of form

$$\mu(x, y) = a + r(x) + c(y) \tag{2.83}$$

writing $s = (x, y)$ in terms of its coordinates. The key point of (2.83) is additivity, which may not always be appropriate since this property is not invariant under rotations.

Consider first the case where data lie on a grid (not necessarily a regular grid, i.e. the distances between the horizontal and vertical lines of the grid need not be constant). Suppose we have data $\{Y_{k\ell},\ 1 \le k \le p,\ 1 \le \ell \le q\}$ with mean function $\mu_{k\ell} = a + r_k + c_\ell$ with the constants $\{r_k\}$ and $\{c_\ell\}$ arbitrary except for a normalizing condition (either the mean or the median of each set of constants must be 0). The main idea is that of median polish analysis of variance, originally proposed by John Tukey as a means of robust analysis of a two-way table.

To get an idea of this, let us first quickly review standard (least squares) ANOVA. The estimates are $\widehat{a} = \bar{Y}_{\cdot\cdot}$, $\widehat{r}_k = \bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot}$, $\widehat{c}_\ell = \bar{Y}_{\cdot\ell} - \bar{Y}_{\cdot\cdot}$, where as usual the dot represents averages over the missing components. This may be easily observed to have the following property: if we define residuals

$$\begin{aligned} R_{k\ell} &= Y_{k\ell} - \widehat{a} - \widehat{r}_k - \widehat{c}_\ell \\ &= Y_{k\ell} - \bar{Y}_{k\cdot} - \bar{Y}_{\cdot\ell} + \bar{Y}_{\cdot\cdot}, \end{aligned}$$

then the mean residuals over rows and columns are zero:

$$\frac{1}{p}\sum_k R_{k\ell} = 0 \text{ for each } \ell, \quad \frac{1}{q}\sum_\ell R_{k\ell} = 0 \text{ for each } k. \tag{2.84}$$

Median polish kriging attempts to replace (2.84) by the property that the row and column medians (rather than means) of the residuals are zero. More precisely, we seek estimators $\tilde{a}$, $\tilde{r}_k$, $\tilde{c}_\ell$ such that

$$\begin{aligned} &\text{med}_\ell(Y_{k\ell} - \tilde{a} - \tilde{r}_k - \tilde{c}_\ell) = 0 \text{ for each } k, \\ &\text{med}_k(Y_{k\ell} - \tilde{a} - \tilde{r}_k - \tilde{c}_\ell) = 0 \text{ for each } \ell. \end{aligned} \tag{2.85}$$

The algorithm that achieves this is the following. We create a $(p+1) \times (q+1)$ matrix $Y^{(i)}$ where the first p rows and q columns contain the current estimates of the residuals at the i'th iteration, the first p entries of the last column contain the current estimates of the r_k's, the first q entries of the last row contain the current estimates of the c_ℓ's, and the $(p+1, q+1)$ entry contains a. The initial values are $Y^{(0)}_{k\ell} = Y_{k\ell}$ if $k \le p$ and $\ell \le q$, otherwise 0.

For i odd, for each $k \in \{1, 2, ..., p+1\}$, we define

$$\begin{aligned} Y^{(i)}_{k\ell} &= Y^{(i-1)}_{k\ell} - \text{ med}\{Y^{(i-1)}_{k\ell'},\ 1 \le \ell' \le q\}, \\ Y^{(i)}_{k,q+1} &= Y^{(i-1)}_{k,q+1} + \text{ med}\{Y^{(i-1)}_{k\ell'},\ 1 \le \ell' \le q\}. \end{aligned}$$

For i even, for each $\ell \in \{1, 2, ..., q+1\}$, we define

$$\begin{aligned} Y^{(i)}_{k\ell} &= Y^{(i-1)}_{k\ell} - \text{ med}\{Y^{(i-1)}_{k'\ell},\ 1 \le k' \le p\}, \\ Y^{(i)}_{p+1,\ell} &= Y^{(i-1)}_{p+1,\ell} + \text{ med}\{Y^{(i-1)}_{k'\ell},\ 1 \le k' \le p\}. \end{aligned}$$

Thus on an odd iteration, we adjust all the rows (including the last one) so that the median residuals are all zero, then on an even iteration, we similarly adjust all the columns. This process continues until convergence, or until the change from one iteration to the next is no greater than some specified very small number ϵ. The latter criterion is the usual one adopted in practice and usually reaches its conclusion after only a few iterations. For example, the Splus function `twoway`, with `trim=0.5` to select medians (the default), works in precisely this way. At the end, $\tilde{a}$ is the final value of $Y^{(i)}_{p+1,q+1}$, $\tilde{r}_k$ is the final value of $Y^{(i)}_{k,q+1}$, and $\tilde{c}_\ell$ is the final value of $Y^{(i)}_{p+1,\ell}$. The whole routine works in exactly the same way in circumstances under which there may be more than one observations corresponding to each (k, ℓ) combination, or if there are missing observations.

Properties of the algorithm

Does it converge in a finite number of steps? Not necessarily, but it does if two modifications are made. First, we must assume that all the data are recorded to finite precision – of course this will always be satisfied in practice. The second condition is

related to the definition of the median of an even number of observations. Usually, the mean of $2m$ ordered data points $x_1 \leq ... \leq x_{2m}$ is defined to be $(x_m + x_{m+1})/2$. Under this definition, the algorithm may not converge. However, if the median is unambiguously defined to be either x_m or x_{m+1}, then it does converge in a finite number of steps. In practice, as already remarked, we take the usual definition of the median and impose an ϵ-stopping criterion.

A second question concerns the properties of the solution. The median of n univariate observations $x_1, ..., x_n$ may be defined as the value of a that minimizes $\sum_i |x_i - a|$. The analogous property that we might expect of the median polish kriging estimators $\tilde{a}$, $\{\tilde{r}_k\}$ and $\{\tilde{c}_\ell\}$ is that they minimize the L_1 criterion

$$\sum_k \sum_\ell |Y_{k\ell} - a - r_k - c_\ell|$$

over all choices of a, $\{r_k\}$ and $\{c_\ell\}$. It appears that in most cases, but not all, this condition is indeed achieved.

References for this section are Fink (1978) for convergence, Kemperman (1984) and Sposito (1987) for the equivalence between the median polish and L_1 minimzation criteria.

Use of median polish for spatial trend estimation

Suppose the data points do indeed lie on a rectangular grid. The median polish algorithm will estimate the trend at each grid point. This is then extended to a complete surface reconstruction by linear interpolation within grid cells.

Now consider the more common case in which the data points do not lie on an exact grid. This is usually handled by means of a *low-resolution map*, i.e. we place a grid over the observations such that there is approximately one observation per grid cell. We then apply the median polish algorithm with each data point identified with the nearest point of the grid. As noted above, the algorithm can handle multiple or missing data points, so it is not necessary that there be exactly one data point for each grid cell.

Usually, construction of the median polish surface is followed by ordinary kriging of the residuals. The residuals may still be expected to be spatially correlated, but with the large-scale variability absorbed by the median polish, one would expect them to fit a stationary, and in many cases isotropic, model much more easily than the original data. The reconstructed residual surface is then added to the median polish trend to obtain the predicted surface corresponding to the original data.

One point should be noted in connection with this, when the data do not fall on a regular grid. In constructing the residual associated with a data point (x, y), it is important to use the estimated trend at the point (x, y), which is constructed by linear interpolation between the grid points, rather than the trend at the grid point to which (x, y) is associated.

By doing this, after kriging the residuals and reconstructing the original surface, the final answer will be an interpolator, i.e. a surface which goes exactly through all the observed data points. This is a property of ordinary or universal kriging and is usually considered desirable.

Bias reduction in the estimated variogram

One further point is made in favor of median polish kriging: that covariances calculated from data that are centered by median polishing may be less biased than those from data centered by mean polishing. Cressie and Glonek (1984) gave some specific calculations justifying this assertion in the case of one-dimensional data.

2.5 Hierachical Models for Trends

In this section we consider a generalization of the kriging problem, which arises in numerous contexts when combining spatial data analysis with some other form of analysis such as a regression of the data at each station against some covariates relevant to that station. Examples include:

(i) The data analyses of chapter 1, and specifically the calculations leading to Figs. 1.7–1.10, in which a time trend was measured at each meteorological station, but no attempt was made to smoothe the results by combining data from different stations. It seems reasonable to assume, however, that any variations in the trend will occur smoothly over space, so the question arises of how best to construct a smoothed surface.

(ii) A different example has arisen in connection with the analysis of SO_2 data from the EPA's CASTNET network (Holland *et al.* 2000). Weekly records of SO_2 have been collected at each of 35 stations in the eastern USA, along with associated meteorological records (principally temperature and precipitation). At each station, it has been possible to fit a regression model to express log SO_2 as a function of meteorology, long-term trend and an irregular annual seasonal effect. Both linear regression and GAM (Hastie and Tibshirani 1990) procedures have been used for this. As a result, it has been possible to compute an estimate of the adjusted SO_2 trend at each station, along with its standard error. The trends are negative across the region, but are strongest in the midwest region of the United States, i.e. the western portion of the region under study. However, in many contexts one would like to be able to report a regional trend — the averaged trend over some specified region of the map. The question therefore arises of how best to smoothe the trend estimates at individual stations, and to interpolate between stations.

In both of these contexts, we may assume there is some smooth underlying (but unobserved) field $Z(s)$, where $Z(s)$ measures the "true" trend at location s. At station s_j we make an observation $\widehat{Z}(s_j)$, corresponding to the estimated trend at this station, after

performing some kind of regression analysis. Thus our underlying point of view is that the regression procedure amounts to measuring the true trend with error. We will write

$$\widehat{Z}(s_j) = Z(s_j) + e_j \tag{2.86}$$

where $e_j \sim N(0, \sigma_j^2)$ represents the error in the regression analysis. For σ_j, we shall in practice use the standard error obtained from the regression procedure, though in the discussion to follow we shall treat this as known, ignoring the fact that in practice σ_j is itself estimated from the residuals of the fitted model.

Concerning the process $Z = \{Z(s_j),\ j = 1, ..., n\}$, we adopt the standard universal kriging assumptions from (2.30), in other words

$$Z = X\beta + \eta, \tag{2.87}$$

where $X\beta$ represents a spatial regression component and η a vector of spatially correlated errors, assumed to have a normal distribution with mean 0 and covariance matrix $\alpha V(\theta)$, with individual entries $\alpha v(s_i, s_j; \theta)$, specified in terms of a scaling constant α and a finite-dimensional parameter θ.

In this model, η represents a vector of perturbations in the unobserved spatial field Z, whereas $e_1, ..., e_n$ have the role of measurement errors. Since they represent completely different sources of variation they may be treated as independent, so henceforth we make that assumption. It is less clear that $e_1, ..., e_n$ may be treated as independent of one another, but for the moment we shall make that assumption as well.

The model defined by (2.86)–(2.87) is of hierarchical structure, with the "state equation" (2.87) representing the hypothesized true state of nature and the "measurement equation" (2.86) generating the observations $\widehat{Z}(s_j)$ as a function of the state of nature. More general forms of hierarchical model would require the use of special filtering techniques (such as Markov chain Monte Carlo algorithms) to reconstruct an estimate of the true surface Z, but in the present case, with both parts of the model being assumed Gaussian, it is possible to take a direct approach via kriging.

Combining (2.86) and (2.87) into one equation, we have

$$\widehat{Z} \sim N(X\beta, \alpha V(\theta) + W) \tag{2.88}$$

where W is the diagonal matrix with entries $\sigma_1^2, ..., \sigma_n^2$. The parameters α, β and θ may therefore be estimated via any of the methods used in section 2.2, for instance a maxiimum likelihood or REML procedure. Note, however, that the form of the log likelihood must be based on (2.36), or (2.39) in the case of REML estimation, with $\alpha V(\theta)$ replaced by $\alpha V(\theta) + W$. The analytic maximization with respect to α, which leads to (2.37) or (2.40) is not available in this case.

Once θ has been estimated, one can estimate the surface Z, at both monitored and unmonitored sites, by standard kriging, after noting that $\mathrm{Cov}\{Z(s), \widehat{Z}(s_j)\}$ is just $v(s, s_j; \theta)$.

In introducing the above we noted that the assumption that the measurement errors are uncorrelated from site to site may not be justified, since this amounts to an assumption of no spatial correlation between individual measurements. This assumption may be removed if we replace the diagonal matrix W by a general covariance matrix, representing the (assumed known) covariances of all of $e_1, ..., e_n$. For the case of the temperature data sets, W is easy to estimate, because the trend estimates have been obtained from simple linear regression over a fixed time period at each station, so the correlations between estimates $\widehat{Z}(s_i)$, $\widehat{Z}(s_j)$ is the same as the correlation of the raw data at stations s_i and s_j. In the case of the SO_2 regressions, the correlations are not so easy to characterize given the more complicated form of regression analysis. However, Holland *et al.* (2000) proposed a jackknife procedure to estimate correlations between estimates from different series, and these correlations have been used in the discussion to follow.

In estimating and comparing different models, a number of different features have been considered,

(a) Five forms of covariance function $V(\theta)$ — exponential, Gaussian, wave, spherical and Matérn,

(b) Option to include or exclude a nugget term (note that the measurement error variances $\{\sigma_j^2\}$ themselves have the interpretation of nugget effects but the general program includes the possibility of *estimating* an additional nugget effect in V),

(c) Polynomial terms up to fourth order modelled using the $X\beta$ regression component,

(d) An additional option was included to permit the simplest form of anisotropic model, geometric anisotropy as given by (2.3). In this case there is no loss of generality in taking the transformation matrix A to be of the form

$$A = \begin{pmatrix} D\cos\phi & D\sin\phi \\ -D^{-1}\sin\phi & D^{-1}\cos\phi \end{pmatrix}$$

corresponding to rotation through an angle ϕ followed by expansion of one axis by D together with compression of the perpendicular axis by D^{-1}. The resulting covariance function has elliptical contours; see for example Fig. 2.16 when the circular contours for $D = 1$ are contrasted with a typical case in which $D > 1$.

As an illustration of these methods, we consider the temperature trends example (i). The raw data consist of the slopes of estimated linear trends, along with their standard errors, fitted to 32 years (1965–1996) of winter mean daily minimum temperatures, at 182 stations in the continental USA. This data set has already been considered in Chapter 1. In particular, Fig. 1.8(a) depicts the estimated trends, and Fig. 1.8 (b) the associated t

statistics, where a coding was used of A to indicate the top sextile of values, B the next sextile, and so on down to F. The range of estimated values runs from –.294 to +.288 (degrees F per year) and an alternative plot, shown in Fig. 2.17(a) to follow, shows the values arranged in 10 equal intervals — plotted as 0 for a value between –.294 and –.236, 1 for a value between –.236 and –.178, and so on up to 9 for a value between .230 and .288. The same scale is used for the smoothed values in Fig. 2.17(b). In the following discussion, we explain the procedure by which the smoothed values are produced.

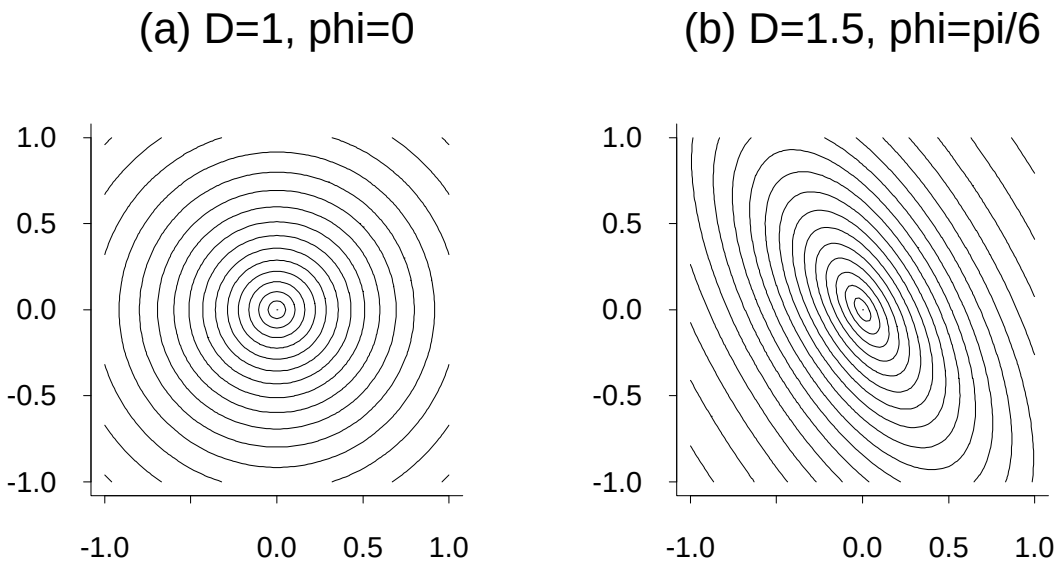


Fig. 2.16. Contour plots for isotropic and geometrically anisotropic covariance functions. (a) Isotropic case, $D = 1$. (b) Anisotropic case, $D = 1.5$, $\phi = \frac{\pi}{6}$.

A variety of models was fitted by maximum likelihood, the negative log likelihood (NLLH) values being given in Table 2.8. All these models are based on the Gaussian covariance function after an earlier fit based on the Matérn covariance function has resulted in $\theta_2 > 50$, at which the algorithm used to evaluate the Matérn covariance function defaults to the Gaussian form (recall that the Gaussian form arises in the limit as $\theta_2 \to \infty$). The model was based on the version of the model which does not use sample correlations between stations, i.e. equation (2.88) with W the diagonal matrix of standard errors in the linear regression equation. The model was tested for a nugget effect, but none was found. Specifically, when a parameter ϕ representing the nugget:sill ratio was included in the model, the value of ϕ quickly converged to 0, indicating lack of significance. The model was also tested for geometric anisotropy, using the two-parameter transformation described under (d) above, and with this the results were inconclusive. The main geometrically

anisotropic model to be included in this comparison is indicated as GA in Table 2.8.

Likelihood ratio tests based on the fitted NLLH values indicate that a cubic spatial trend is the most appropriate for this data set. For example, in testing the quadratic trend (degree 2) against no trend (degree 0), the likelihood ratio test statistic is $T = 2 \times (411.59 - 404.56) = 14.06$ with 8–3=5 degrees of freedom, which is statistically significant when compared against the asymptotic χ^2_5 distribution which applies to the null distribution of this test statistic. The P value is .015. A further test of degree 2 as the null hypothesis and degree 3 as the alternative yields $T = 14.14$, 4 degrees of freedom, P value .007, which is statistically significant. However testing degree 3 as the null against degree 4 as the alternative yields $T = 6.4$, 5 degrees of freedom, P=.27, not statistically significant. Regarding the GA model with degree 3, testing this against the isotropic model of degree 3 results in $T = 5.5$, 2 degrees of freedom, P value .064, which one may regard as borderline — not significant at the .05 level, but perhaps worth computing so as to compare the results with those of the isotropic model.

Degree of Polynomial Trend	Number of Parameters	NLLH
0	3	–404.56
1	5	–405.80
2	8	–411.59
3	12	–418.46
3 GA	14	–421.21
4	17	–421.66

Table 2.8. Comparison of model fits for temperature trends data

When the smoothed estimates of trends at individual stations are computed, they are indeed much smoother than the original estimates, as is shown in Fig. 2.17 (b). As an example, consider the station at Trenton, Missouri, marked 0 in Fig. 2.17(a). At this point, our original estimate of trend was –.294 with standard error .138, based on 22 years' data at this station (1965–1986). After smoothing, the estimate is .079, prediction standard error .020, coded as 6 on Fig. 2.17(b). The changes are not so dramatic at most other stations, but in general there is a much higher spatial coherence between estimates in Fig. 2.17(b) than there is in Fig. 2.17(a). These comparisons are based on the isotropic model with cubic trend.

In Fig. 2.18(a), a contour plot is shown for trends across the United States, with prediction standard errors in Fig. 2.18(b). The corresponding plots are shown in Fig. 2.19 for the geometrically anisotropic version of the model. The two plots are quite similar, and show that the strongest positive trends are in the northern midwest region of the country, with much of the rest at near 0 trend, and a slight negative trend in the south-east.

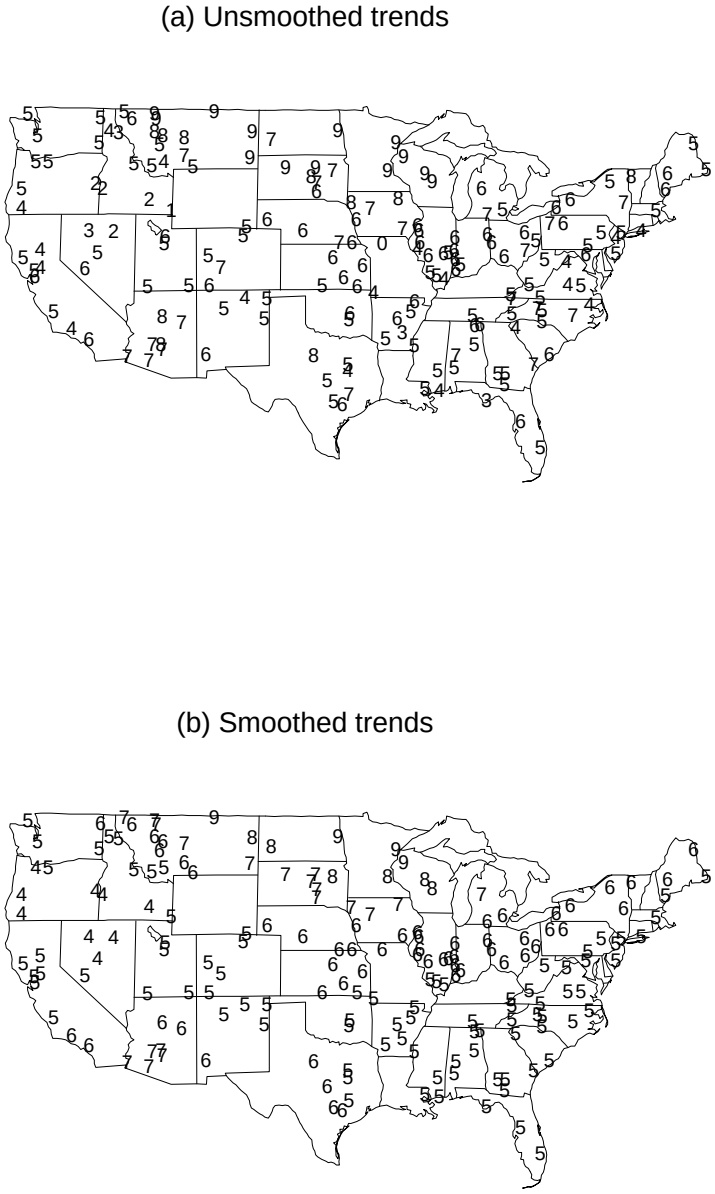
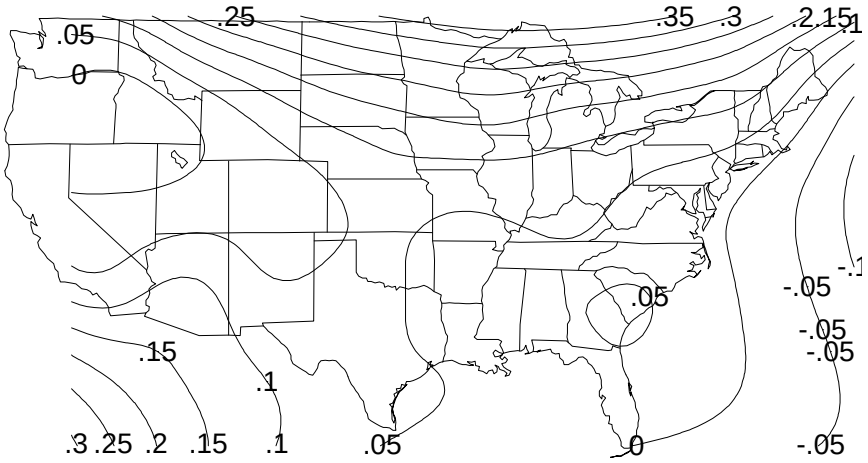


Fig. 2.17. Unsmoothed (top plot) and smoothed (bottom plot) spatial trends for temperature data set. The estimates have been mapped to a common scale and coded as 0—9 to correspond to equal deciles of the range of values.



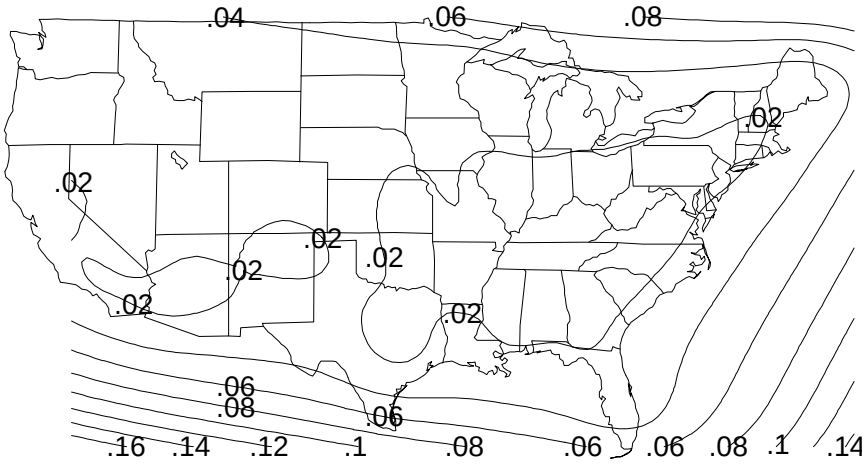


Fig. 2.18. Contour plots for smoothed spatial trends of meteorological data and standard errors based on isotropic model with Gaussian covariance model and cubic trend

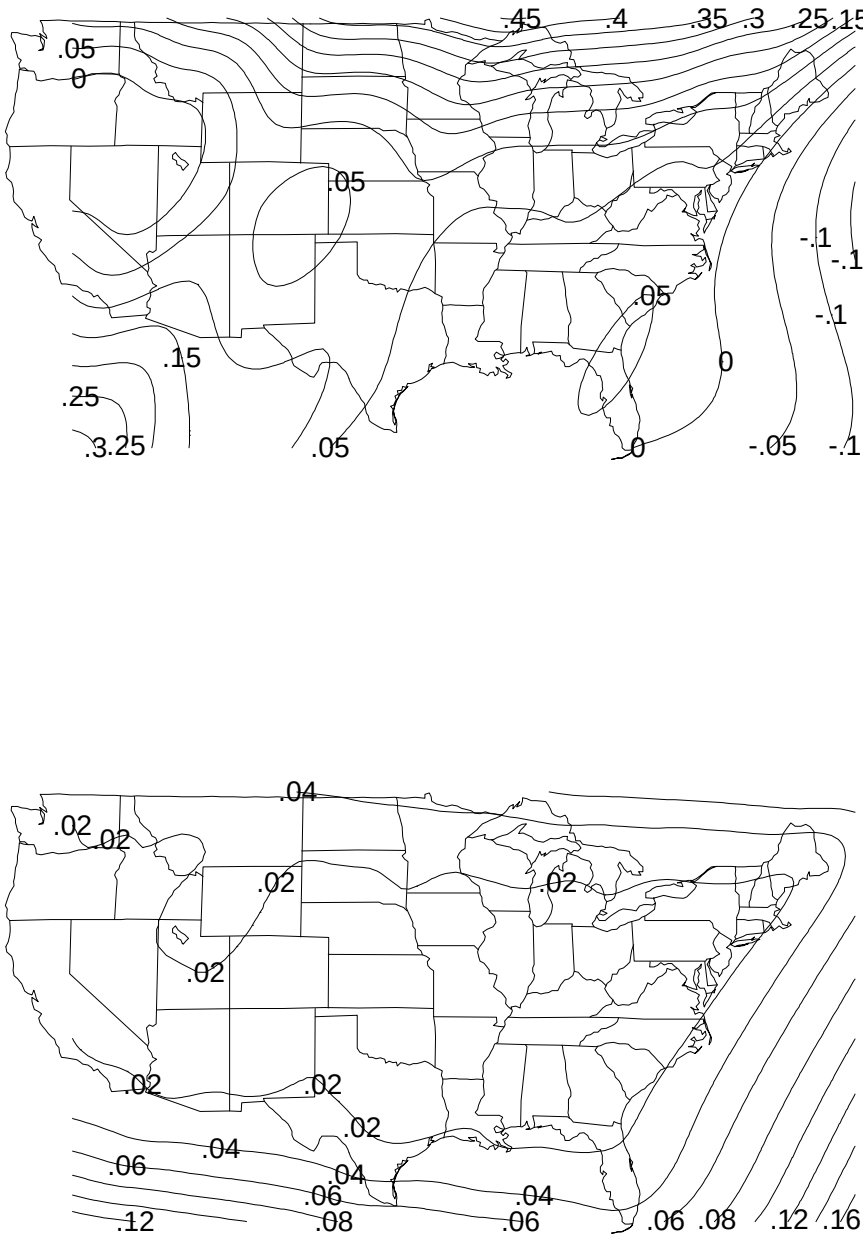


Fig. 2.19. Contour plots for smoothed spatial trends of meteorological data and standard errors based on geometrically anisotropic model with Gaussian covariance model and cubic trend

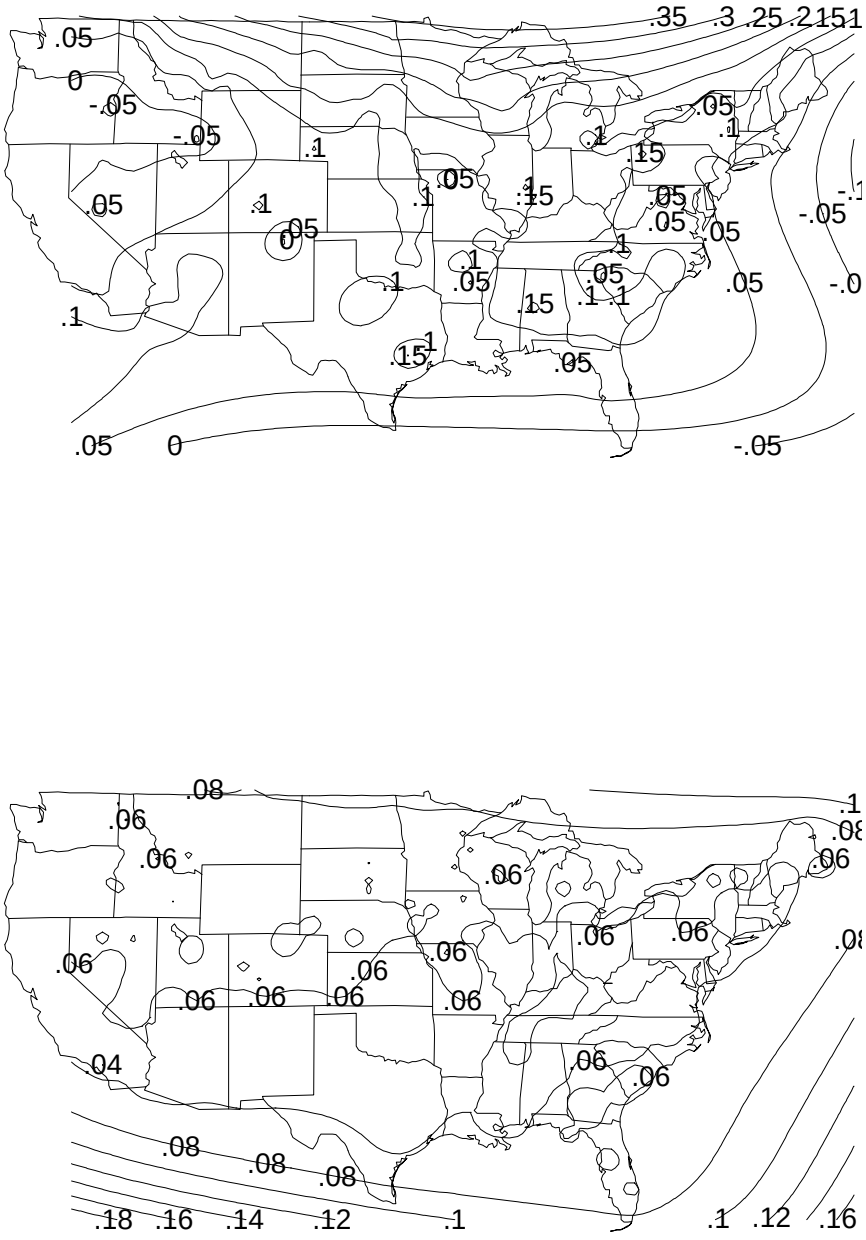


Fig. 2.20. Contour plots for smoothed spatial trends of meteorological data and standard errors based on isotropic model with Matérn covariance model and cubic trend, including correlations of measurement errors

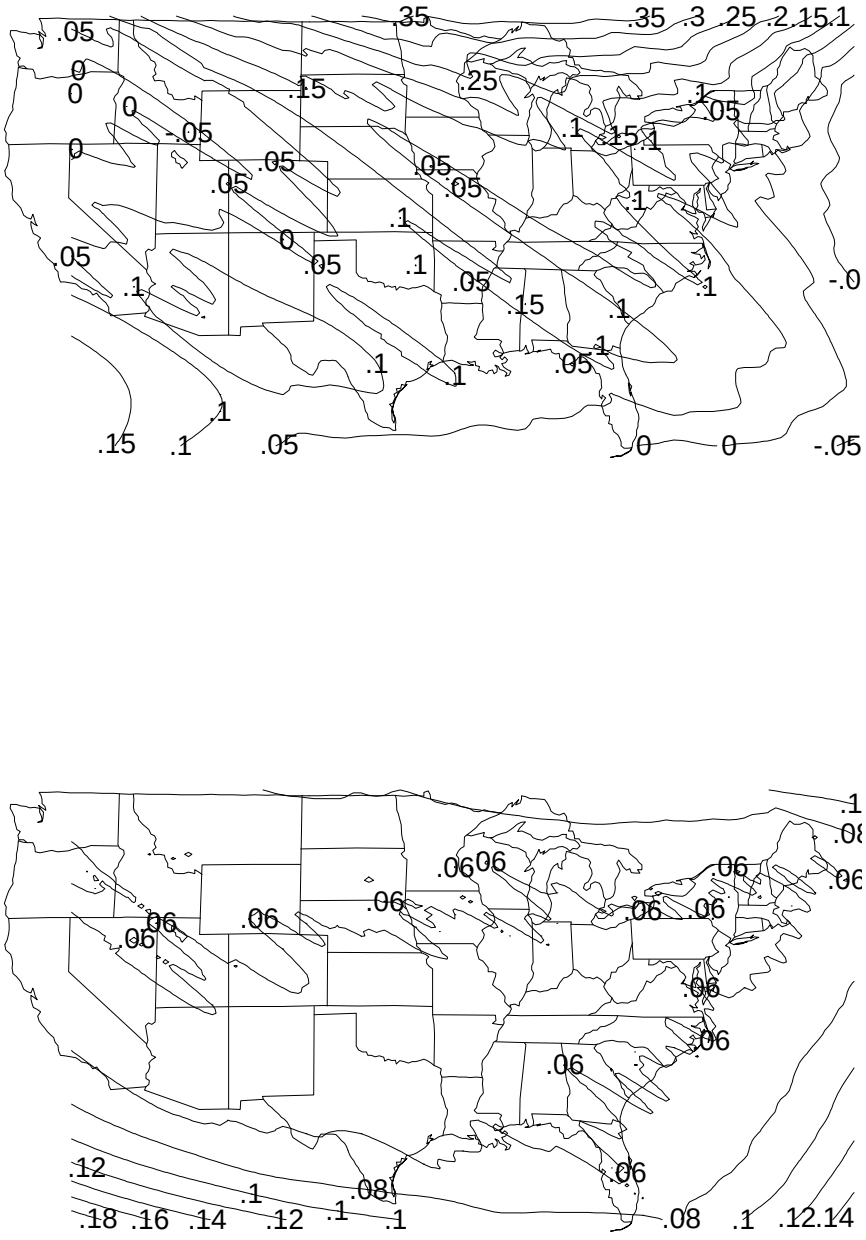


Fig. 2.21. Contour plots for smoothed spatial trends of meteorological data and standard errors based on geometrically anisotropic model with Matérn covariance model and cubic trend, including correlations of measurement errors

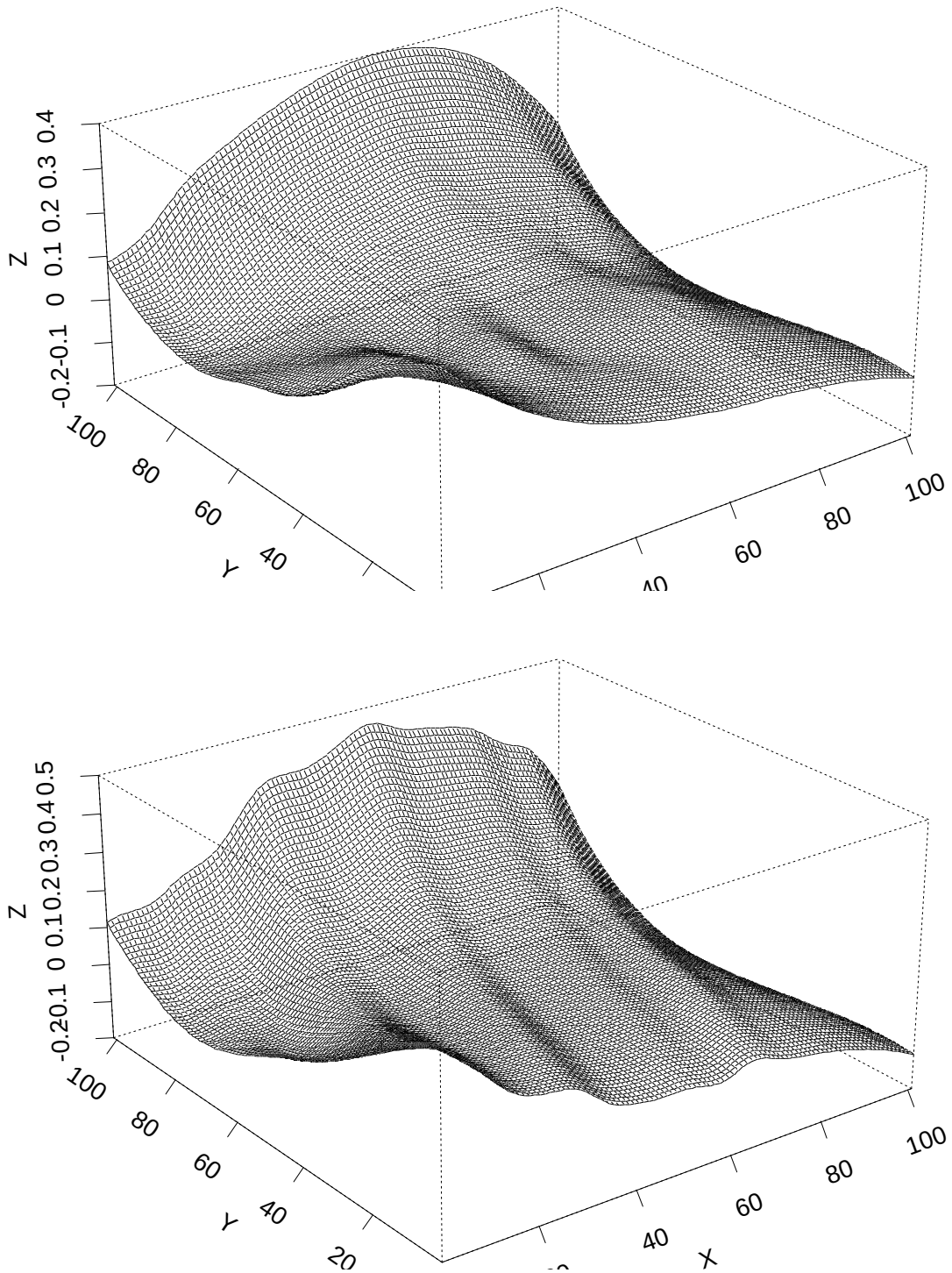


Fig. 2.22. Surface plots for smoothed spatial trends of meteorological data. Top plot: Isotropic model with cubic trends, diagonal W matrix Bottom plot: Geometrically anisotropic model with cubic trends, diagonal W matrix

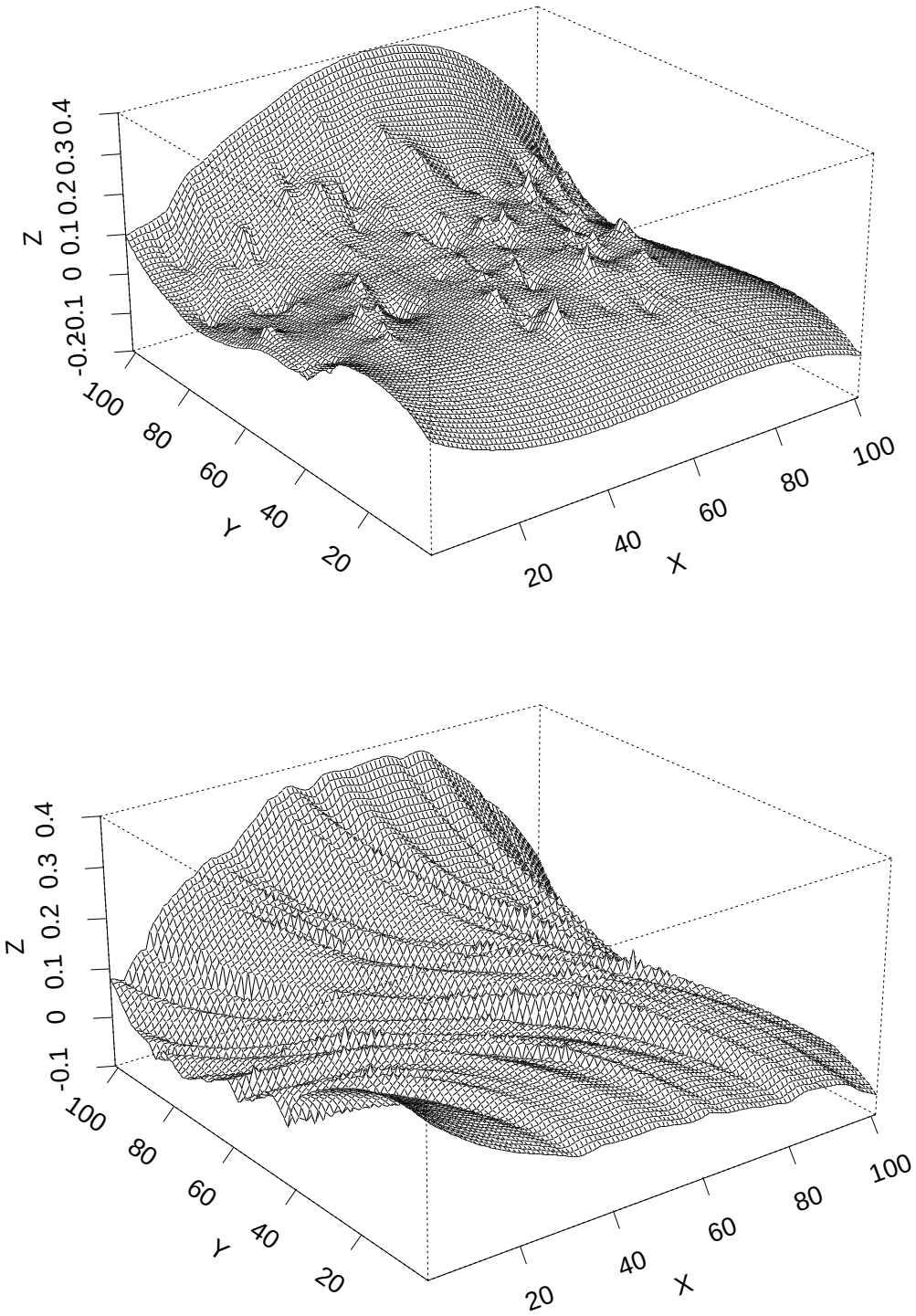


Fig. 2.23. Same as Fig. 2.22, but with non-diagonal W matrix

What happens when the diagonal matrix W in (2.88) is replaced by a non-diagonal matrix, corresponding to the full covariance matrix of errors in the regression analysis? In this case W is easy to estimate, since all of the regression estimates are linear combinations of the original observations with the same weights at each station, so the correlations within the W matrix are the same as the raw spatial correlations among residuals at each station.

When the model is fitted in this way, using a Matérn covariance structure, we again find that a cubic trend is appropriate, but the estimated Matérn shape parameter $\widehat{\theta}_2$ is quite different — about 0.20, as shown in Table 2.9. This is based on an isotropic model; for the geometrically anisotropic version, $\widehat{\theta}_2$ is even smaller, about 0.12. In this case, the difference between the isotropic and anisotropic forms is not statistically significant as judged by the NLLH values, but the NLLH values are significantly different from those in Table 2.8, in other words, including the correlations in the W matrix has apparently improved the fit of the model to the data.

Parameter	Isotropic		Anisotropic	
	Estimate	S.E.	Estimate	S.E.
$\log \alpha$	–5.82	0.14	–5.82	0.14
θ_1	0.74	0.40	1.37	1.82
$\log D$	—	—	1.48	0.64
ϕ	—	—	–.95	0.04
θ_2	0.20	0.18	0.12	0.12
NLLH	–423.24		–424.94	

Table 2.9. Parameter estimates for two model fits to temperature trends data with correlations of measurement errors.

Contour plots in Figs. 2.20 and 2.21 show that the broad shape of the reconstructed surface is similar to the reconstructions in Fig. 2.18 and 2.19, but is noticably more irregular (especially in Fig. 2.19). An alternative way of representing the contour plots in Figs. 2.18–2.21 is as a surface (perspective) plot, and this is shown in Figs. 2.22 and 2.23 (reconstructed surface only; not the standard errors). This confirms that the overall shape of the reconstructed surface is the same for all four models, but it becomes successively more irregular as additional parameters are added to the model.

At this stage, we do not have a clear view of which is the best model among those considered. Given the statistically significant differences between the model fits in Tables 2.8 and 2.9, it would appear that including the non-diagonal W matrix improves the fit, but at the cost of a much more irregular reconstructed surface (which may be right, but which is certainly harder to interpret). Given that the entries of the W matrix were estimated simply by calculating pairwise correlations among the trend estimates from the original

data series, rather than by fitting any smoothed parametric model, it is possible that the lack of smoothness arises from the estimation of W. On this basis, it is not clear that including the non-diagonal entries of W really is an improvement. However, on the basis of the estimates given, we conclude that the non-diagonal W matrix does improve the fit but that in this case, the isotropic model for Z suffices, in other words, the best predictions among those considered are those in Fig. 2.20 and the top plot of Fig. 2.23.

Our final example in this chapter is concerned with trend modeling of atmospheric sulfur dioxide (SO_2) concentrations (Holland *et al.* 2000). The need to monitor trends in SO_2 arose from the Clean Air Act Amendments of 1990, which mandated reductions on SO_2 emissions — as a result of this act, emissions by 2010 are projected to be down to 9 million tons per year, compared to an estimated 19 million tons per year which would have occurred without regulation. To monitor this, EPA has set up CASTNET, a network of 35 stations in the eastern United States, which are deliberately set in rural locations, away from major population centers and known emissions sources. This is to avoid the effects of purely local influences on the measured concentrations.

The analysis of Holland *et al.* (2000) proceeded in two stages: first, the estimation of a trend at each station, and second, the combination of single-station trends into spatial estimates and regional averages. The second stage employed spatial analysis and will therefore be our main concern here, but we briefly describe the first stage to provide relevant background information.

The measured concentration of SO_2 is known to vary seasonally and also to be influenced by meteorological factors, in particular, temperature and the wind speed and direction. To adjust for these factors, a generalized additive model (GAM) was fitted, as follows: for week i in year j at station ℓ, the SO_2 level $S_{ij\ell}$ satisfies

$$\log S_{ij\ell} = \mu_\ell + g_{1,\ell}(w_i) + g_{2,\ell}(y_{ij}) + g_{3,\ell}(t_{ijl}) + g_{4,\ell}(u_{ij\ell}, v_{ij\ell}) + \epsilon_{ij\ell}, \tag{2.89}$$

where $g_{1,\ell}, g_{2,\ell}$ and $g_{3,\ell}$ are smooth functions of week w_i, year y_{ij} (measured in decimal parts of years) and mean temperature $t_{ij\ell}$, and $g_{4,\ell}$ is a smooth function of the vector mean $(u_{ij\ell}, v_{ij\ell})$ where u and v are the mean east-west and north-south components of wind velocity. The data points are assumed to be defined weekly since this is the frequency of measurement of SO_2. We assume the errors $\{\epsilon_{ij\ell}\}$ are independent normal random variables with mean 0 and variance σ_ℓ^2 for each station ℓ. The model (2.89) may be fitted using the GAM software in S-Plus and the estimated function g_2 used to define a station trend (estimated percent reduction from 1989 to 1995) for each station. The resulting percentages, plotted at the approximate locations of the stations, are shown in Fig. 2.24. Also shown on this plot are the three grids (midwest, mid-Atlantic and South) that will be used subsequently for regional trend analysis.

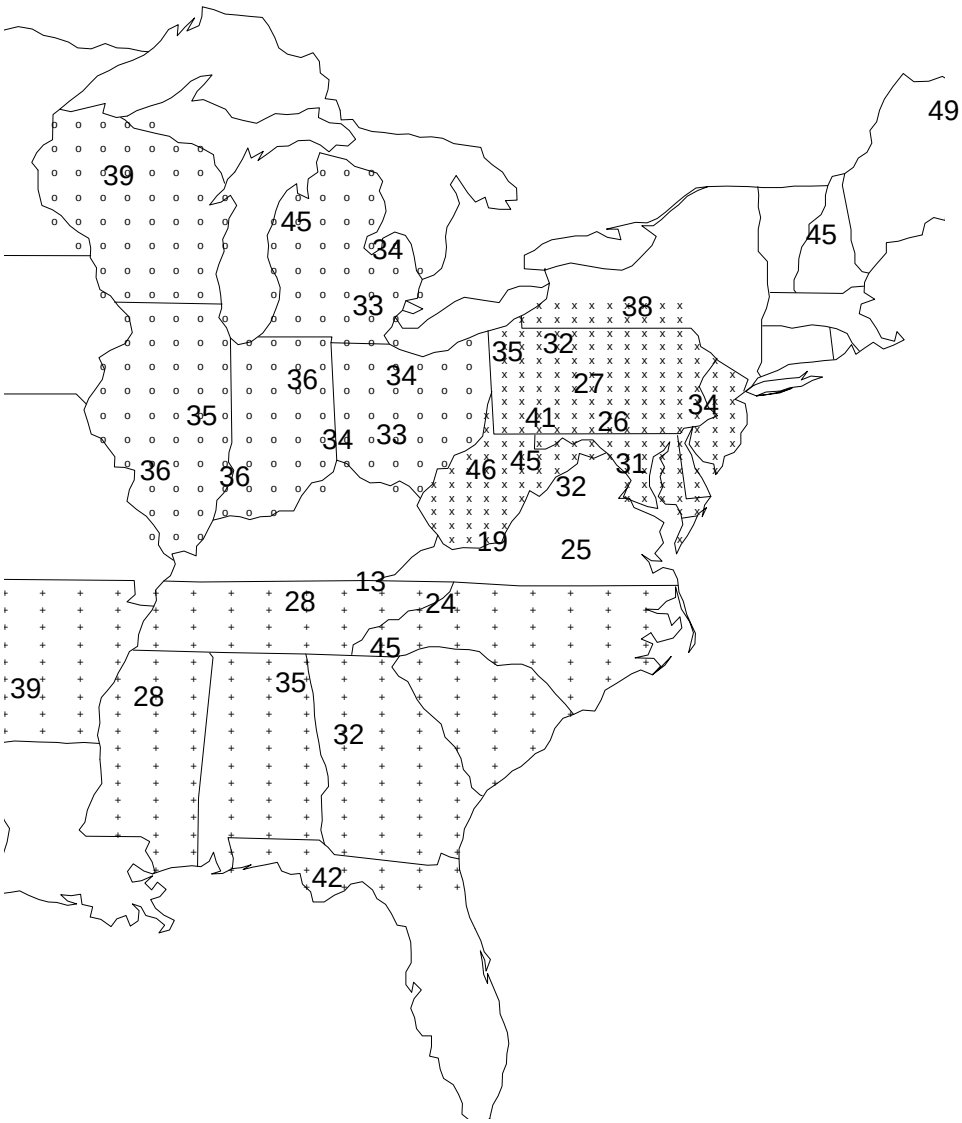


Fig. 2.24. Estimated downward trends (percentage decrease from 1989–1995) based on GAM model fitted to individual stations. The superimposed grids correspond to the three grids used for regional calculations: midwest (o), mid-Atlantic (x) and South (+).

The GAM analysis provided a standard error for each individual trend estimate, but it did not provide covariances between trend estimates at different stations. Because these covariances are important in subsequent analysis, we need some means oc calculating them. The method used was a jackknife calculation (Efron and Tibshirani 1993): for each of the 81 months of data, a *pseudo-estimate* for each site was computed by omitting that month's data and recomputing the estimates based on the remaining months. Sample variances and covariances of the pseudo-estimates may be used to approximate the variances and covariances of the original estimates. The method assumes there is no significant time-series dependence within the series, but autocorrelation analysis suggests this is a reasonable assumption.

We now turn to the second phase of the analysis: estimation of a spatial covariance model for the assumed underlying smoothed trend $Z(s)$, using the GAM-based estimates to give $\widehat{Z}(s)$ at each station s. A variety of models, including different forms of spatial covariance function, different assumptions for the degree of polynomial trend (0 or 1 — no higher-order terms were found significant) and either diagonal or non-diagonal W matrix (the diagonal form being based on ignoring the correlations between station estimates found by the bootstrap procedure) were fitted and are tabulated in Table 2.10. Also tried was a geometrically anisotropic (GA) form of the model. The Matérn model was also fitted but in every case resulted in $\widehat{\theta}_2 > 50$, which defaults to the Gaussian case.

Degree of Polynomial Trend	W matrix	Model	Number of Parameters	NLLH
0	Diag.	Exponential	2	87.51
0	Diag.	Gaussian	2	87.00
0	Diag.	Wave	2	87.25
0	Diag.	Spherical	2	87.14
1	Diag.	Exponential	4	87.14
1	Diag.	Gaussian	4	84.64
1	Diag.	Wave	4	83.88
1	Diag.	Spherical	4	84.76
1	Diag.	Gaussian GA	6	83.89
0	Non-diag.	Gaussian	2	83.23
0	Non-diag.	Gaussian GA	4	81.97
1	Non-diag.	Gaussian	4	82.00
1	Non-diag.	Gaussian GA	6	80.82

Table 2.10. Comparison of model fits for SO_2 trends data

Conclusions from Table 2.10 are (i) there is very little to choose between the four covariance models — exponential, Gaussian, wave and spherical; (ii) there is significant

evidence of a trend in the diagonal-W case but this is much less strong in the non-diagonal-W case; (iii) the GA model does not appear significant compared with the isotropic model, (iv) using NLLH as a method of comparison, it looks as though the model fit is improved by including the off-diagonal entries of W.

As an example of the resulting trend estimates and mean squared prediction errors, Fig. 2.25 shows contour plots of these for the model with Gaussian covariance, isotropic, linear trend and diagonal W matrix; Fig. 2.26 shows the plots for the same model but with non-diagonal W. The measurement stations are shown as dots on the contour plots. In this case the distance scale used for the analysis was degrees of latitude or longitude — unlike the earlier temperature example, the scale was not initially converted to nautical miles. It may appear less reasonable to treat degrees of latitude and degrees of longitude on an equal footing, but since we have tested for geometric anisotropy and found no effect, it would appear that the distinction does not have much practical effect. In any case, it should be noted that the estimated value of R here, measured in degrees, is quite small relative to the total scale of the data, implying that variations in the SO_2 trend are spatially dependent only over quite small distances of the order of 1–2 degrees.

The plots both show the sharpest decline in SO_2 level at roughly the point where the three states of Ohio, Pennsylvania and West Virginia come together, but the peak is sharper in Fig. 2.26 than in Fig. 2.25 and in other respects the two plots are not exactly the same shape. This does not provide any verification of which model is better but it does show that the distinction between the two modeling techniques (either with or without the covariances in the W matrix) cannot be ignored.

Estimating regional trends

Much of the interest in this kind of analysis lies in the possibility of using the smoothed surfaces obtained from the spatial analysis to estimate regional trends. As an example, we consider the "regions" defined by the three sub-grids of hypothetical measurement sites (midwest, mid-Atlantic and South) represented in Fig. 2.24.

The methodology essentially follows that of section 2.4.4: having obtained a spatially smoothed estimate of $Z(s)$ at each site s, the integral $Z(A)$ over a region A is estimated by pointwise integration as in (2.78), with a mean squared prediction error as in (2.79). However, the integrals in (2.78) and (2.79) were for practical calculation replaced by summation over the grids shown in Fig. 2.24. Note that there has been an implicit change of notation here: we have used $\widehat{Z}(s)$ in this section as the estimated trend (from the GAM analysis) at site s, but the $\widehat{Z}(s)$ used in the integrals (2.78) and (2.79) is the predicted (kriged) value after the spatial analysis.

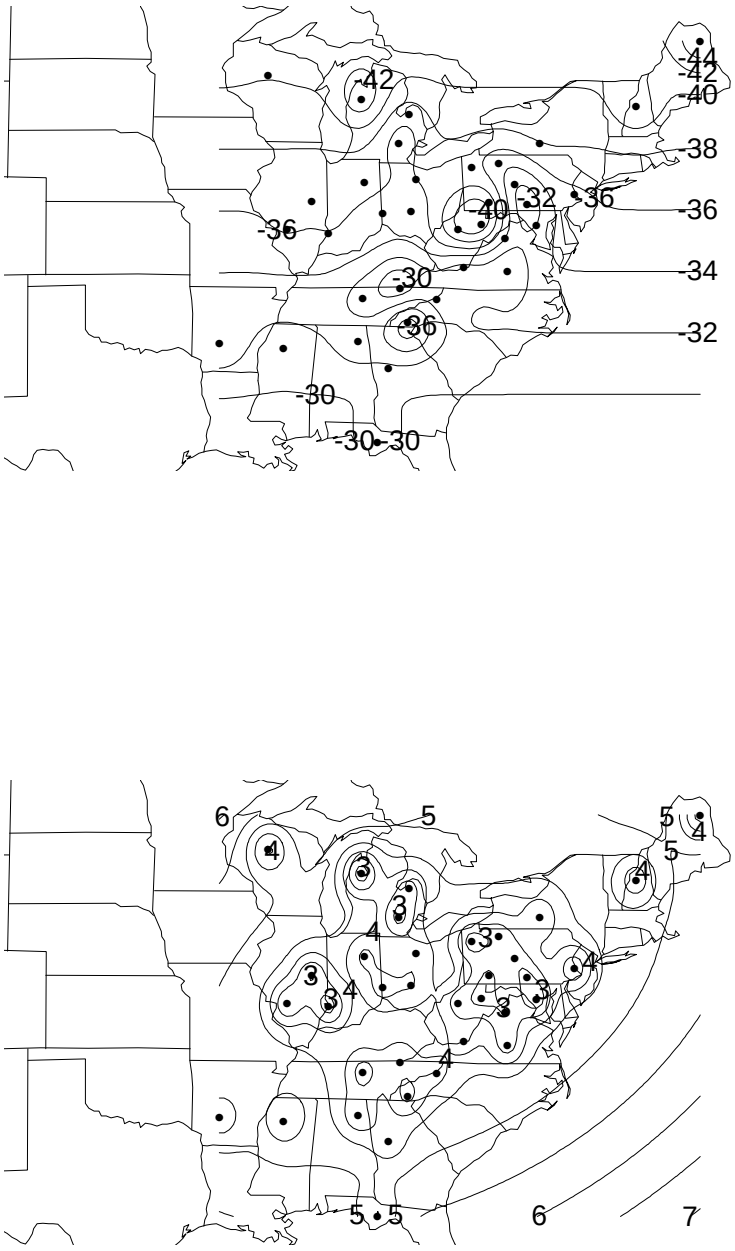


Fig. 2.25. Contour plots for smoothed spatial trends of SO_2 data and standard errors based on isotropic model with Gaussian covariance model and linear trend; diagonal W matrix.

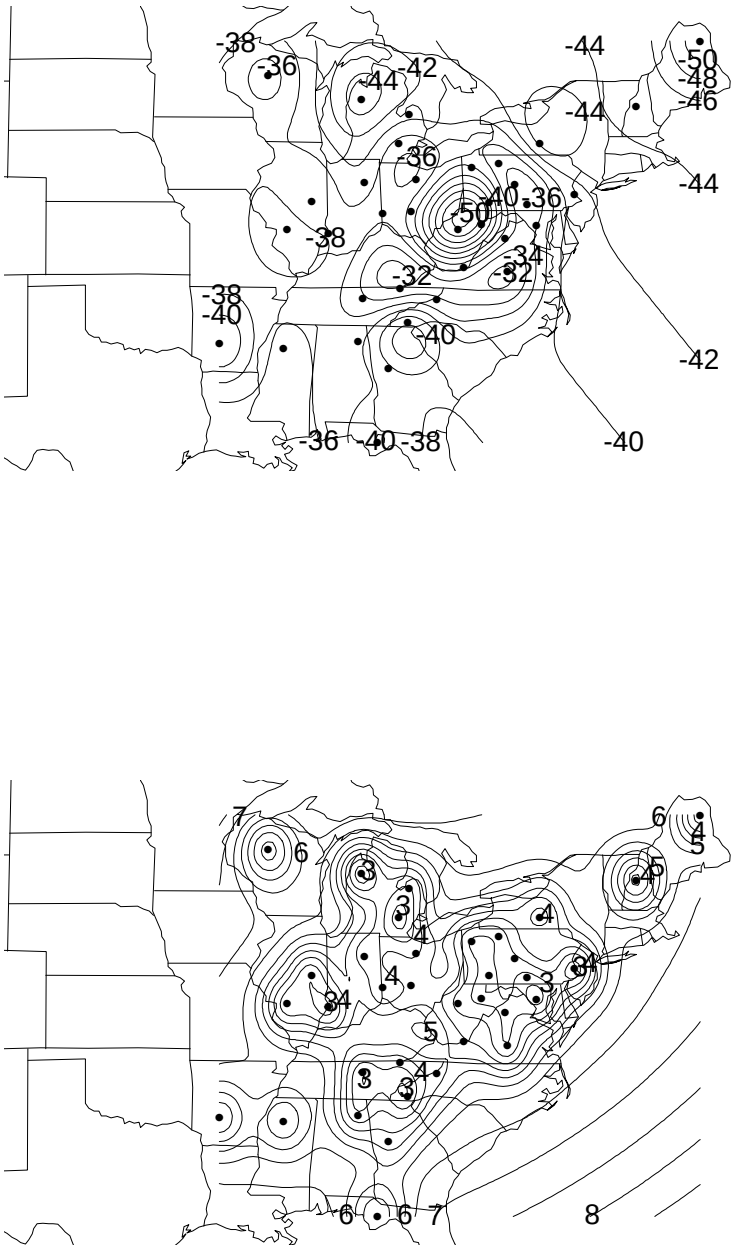


Fig. 2.26. Contour plots for smoothed spatial trends of SO_2 data and standard errors based on isotropic model with Gaussian covariance model and linear trend; non-diagonal W matrix.

For the model being considered here (including linear trend, and so slightly different from the values given in Holland *et al.* (2000)), the estimated regional trends and mean squared prediction errors (MSPEs) are given in Table 2.11. In this table the MSPEs are simply the square roots of the estimated prediction variances from (2.79) — the Bayesian versions of the calculation in columns 4 and 5 will be explained momentarily.

Region	Estimated Trend	Ordinary MSPE	Bayes Trend	Bayesian MSPE
Midwest	–38.76	2.59	–38.60	2.63
Mid-Atlantic	–40.60	2.23	–40.17	2.34
South	–37.62	2.97	–36.95	3.11

Table 2.11. Regional trends and mean squared prediction errors by non-Bayesian and Bayesian calculations

The calculations of MSPE in column 3 of Table 2.11 are based on equation (2.79), which in turn depends on (2.64) for the covariance of predictions at two sites. However, as noted already in section 2.4, such calculations of prediction errors do not take account of the fact that the covariance parameters of the process are themselves unknown parameters.

In Section 2.4.3, we presented an alternative approach to the problem based on Bayesian calculations, which had the advantage that the uncertainty of all the model parameters (in the notation used earlier, β, α and θ) could be incorporated in a single model. The earlier discussion was for prediction at a single site, but the concepts extend easily to the case of multiple sites, and we shall follow them here.

Similar to the earlier Bayesian analysis, we assume a joint prior density of (β, α, θ) of the form $\pi(\theta)/\alpha$ (recall (2.43)) which is of improper form for β and α but allows for an arbitrary prior $\pi(\theta)$. In the present example, the only unknown parameter in the covariance function is the range parameter, so this becomes θ. It might be natural to assume $\pi(\theta)$ improper on the range $0 < \theta < \infty$ but it has been shown that, for this parameter, an improper prior density leads to an improper posterior. Therefore, we must assume a proper prior density for θ, and Holland *et al.* (2000) assumed an inverse gamma prior, $IG(a, b)$ with density $b^a \theta^{-a-1} e^{-b/\theta}/\Gamma(a)$ on $0 < \theta < \infty$, $a > 0$, $b > 0$. Holland *et al.* assumed $a = 2$, $b = 1.73$ where the choice $a = 2$ is intended to represent a reasonable compromise between the improper case $a = 0$ and an over-informative prior, while b was chosen so that the prior mean $b/(a-1)$ matches the estimate range from the maximum likelihood analysis. In the present discussion we have chosen the slightly simpler values $a = b = 2$, but the principle is the same. The joint posterior density of (α, θ) is given by (2.45) but with the change that the expression $G^2(\theta)/\alpha$ must be replaced by

$$(Z - X\widehat{\beta})^T \{\alpha V(\theta) + W\}^{-1} (Z - X\widehat{\beta}). \tag{2.90}$$

Compare (2.35). The distinction arises from the matrix W, which was not present in section 2.4. Because of this, we cannot integrate out α analytically, as we did in going from (2.45) to (2.46), and must work directly with (2.45), incorporating the change just mentioned.

Once we obtain the posterior density of (α, θ) given the original data Y, say $\pi(\alpha, \theta|Y)$ the posterior density for the quantity of interest, $Z(A)$ say, is given by

$$\pi(Z(A) \mid Y) = \int \pi(Z(A)|Y, \alpha, \theta)\pi(\alpha, \theta|Y)d\alpha d\theta. \tag{2.91}$$

(2.91) is similar to (2.76), but adapted to the present set up.

In practice, we proceed as follows:

(i) Since it is not possible to obtain $\pi(\alpha, \theta|Y)$ analytically, we proceed by a Markov chain Monte Carlo procedure (see, for example, Gilks *et al.* (1996)). This yields a Monte Carlo sample of values $(\alpha_m, \theta_m,\ 1 \leq m \leq M)$, whose sampling distribution approximates the posterior distribution desired.

(ii) Instead of calculating the full posterior density from (2.91), we concentrate on the posterior mean and variance of $Z(A)$. The posterior mean is derived from the iterated expectation formula,

$$\mathrm{E}\{Z(A)\} = \mathrm{E}[\mathrm{E}\{Z(A)|\alpha, \theta\}], \tag{2.92}$$

where the inner expectation in (2.92) is just the conditional expectation of $Z(A)$ when α and θ are known, i.e. the standard kriging calculation, while the outer expectation evaluates the mean of this as a function of (α, θ). The inner expectation is performed analytically but the outer expectation must be performed numerically, averaging over the Monte Carlo sequence $(\alpha_m, \theta_m,\ 1 \leq m \leq M)$,

For the conditional variance of $Z(A)$, we use the iterated expectation formula for variances:

$$\mathrm{Var}\{Z(A)\} = \mathrm{E}[\mathrm{Var}\{Z(A)|\alpha, \theta\}] + \mathrm{Var}[\mathrm{E}\{Z(A)|\alpha, \theta\}], \tag{2.93}$$

where again, for each of the two terms in (2.93), the inner calculation of conditional mean or variance is performed analytically using the standard kriging formulae, but the outer calculation is numerical. The second term of (2.93) is the additional contribution to the prediction variance resulting from the fact that α and θ are *a priori* unknown, and for this reason, the estimated variance by this method should be larger than that based on (2.79).

In practice, it does not seem to make much difference — at least, not for this example. The fourth and fifth columns of Table 2.11 show the posterior means and posterior standard deviations of $Z(A)$ using the Bayesian formulae, and they are not very much different from the values obtained without the Bayesian calculation. The main virtue of the calculation

in the present case is that it has shown that the neglect of parameter uncertainty in the kriging formulae has not had a great impact on the results.

One point that should be made about this calculation is that it is still not a fully Bayesian calculation — although we assumed prior densities for α and θ, the GAM model estimates from the first part of the analysis are still treated as fixed. Also, we have not made any attempt to incorporate uncertainty about the W matrix, which could be an important part of the calculation. In view of these points, there is scope for a more comprehensive Bayesian analysis.

To conclude this section, we return to the interest in calculating regional trends. Fig. 2.27 shows estimated regional mean trends and prediction intervals (based on estimated mean $\pm 2 \times MSPE$, non-Bayesian calculation) for each of the three regions as well as the 35 stations, and each of eight possible models. The sensitivity to different model choices is by no means negligible, though the variability between models is not excessive after taking into account the uncertainty of the estimates as accounted for by the prediction intervals.

As far as the implications for implementation of the Clean Air Act Amendments are concerned, Holland *et al.* (2000) note that the estimated reduction in atmospheric SO_2 levels is much greater than the estimated reduction in all emissions, estimated from independent EPA sources as –28%, –18% and –10% respectively for the Midwest, Mid-Atlantic and South. However, they suggest that the discrepancy arises from the rural location of the stations: in rural locations, the main component of SO_2 arises as a result of long-range transport from power stations, and the estimated reductions in atmospheric SO_2 are quite consistent with the estimated emissions reductions from this source.

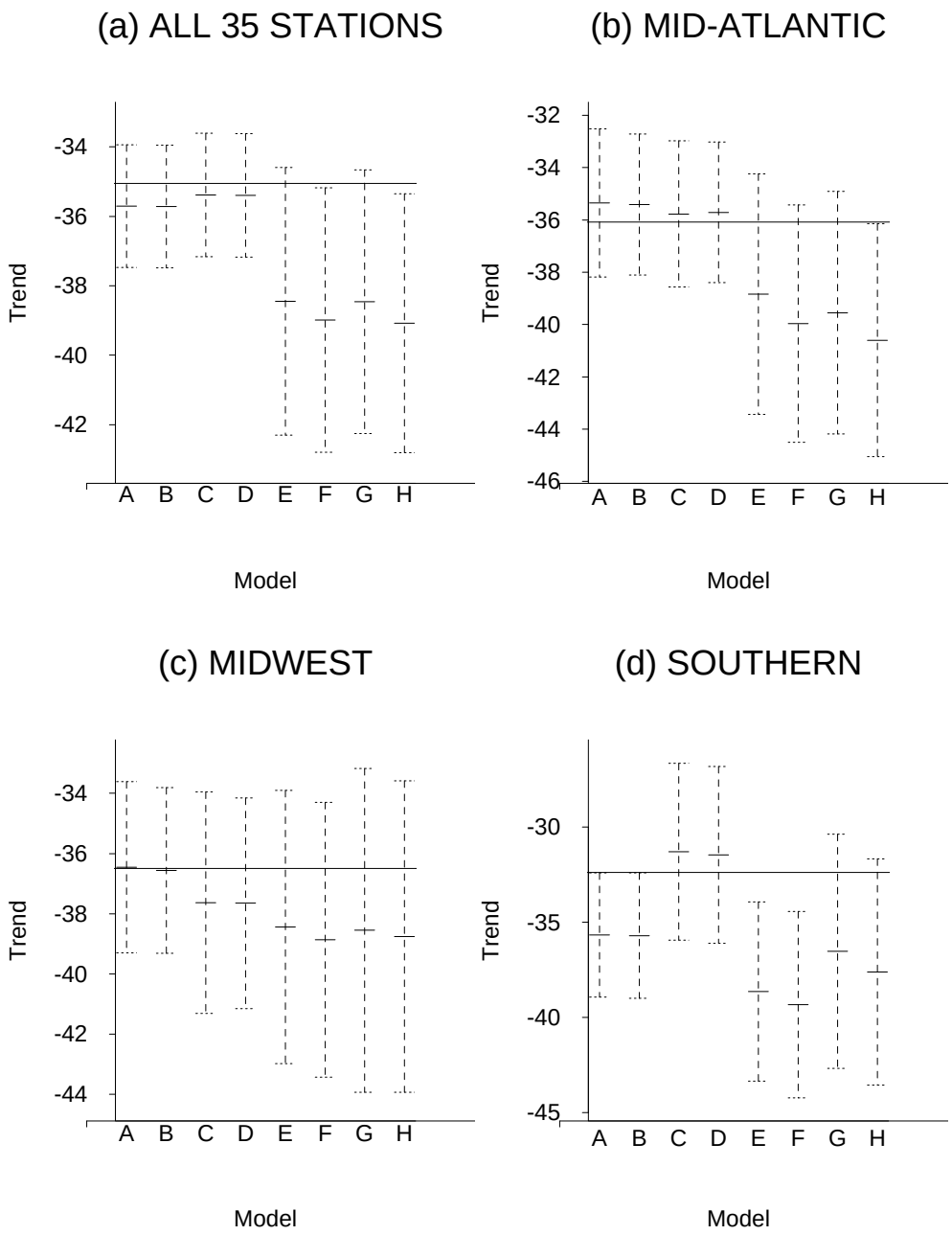


Fig. 2.27. Comparison of regional average predictions for eight models: A, exponential covariance function, no spatial trend, diagonal W matrix; B, Gaussian covariance function, no spatial trend, diagonal W matrix; C, exponential covariance function, linear spatial trend, diagonal W matrix; D, Gaussian covariance function, linear spatial trend, diagonal W matrix; E–H, as A–D but including general W matrix. Predictions are calculated for (a) overall mean of 35 stations, (b) Mid-Atlantic grid, (c) Midwest grid, (d) South grid. In each case the mean prediction and 95% prediction error bounds (non-Bayesian calculation) are shown for each of the eight models. The horizontal line on each plot represents the observed mean of all 35 stations (plot (a)) or of the stations lying within the grid in question (plots (b)–(d)).

CHAPTER 3

Nonstationary Spatial Processes

In this chapter we consider a number of different approaches to processes which are not spatially stationary. Of course, it is not possible to give a complete categorization of all the techniques available, as the form of model appropriate for a particular application depends very much on the details of the application. In this chapter, we shall highlight five techniques:

(a) Moving-window methods, in which the predictor or interpolator at a particular location, is based on a "window" of observations centered at that location,

(b) Methods based on an eigenfunctions expansion of the covariance function,

(c) Deformation methods, in which it is assumed that the process is stationary and isotropic only after some nonlinear deformation of the sampling space,

(d) Bayesian methods, in which inhomogeneity is expressed in the form of the prior distribution,

(e) Models defined in terms of kernel smoothers.

These methods are dealt with, in turn, in sections 3.1–3.5.

3.1 Moving-Window Approaches

The idea of a moving-window approach is that to fit a spatial model and to perform kriging at a sampling location s, we should restrict ourselves to a "window" of sampling stations close to s, within which it is reasonable to assume a homogeneous model. Thus the method retains all the mathematical techniques of homogeneous processes, while not assuming that homogeneity applies across the whole sampling region. As a result, it appears to be a reasonable compromise between the methods of chapter 2, which assumed homogeneity everywhere, and the more sophisticated models for inhomogeneous processes, which are considered in later sections of the present chapter. The principal advocate of this methodology has been Haas (1990, 1995, 1998).

For the present description we shall largely follow Haas (1995), who develops the method in the context of spatio-temporal processes. However the essential ideas of the method apply in a purely spatial context as well, as in the earlier paper Haas (1990). Apart from including the time domain, the paper of Haas (1995) develops a number of

ideas in more detail than in the earlier work, in particular the idea of cross-validation as a means of choosing the window size, and this is the main reason we follow this paper here.

Specifying the model

Suppose we have spatio-temporal data $Z(t,s)$ where t denotes time and s denotes space. Specifically, we have a sample $\{Z(t_i,s_i)\}$ at n time-space points $\{(t_i,s_i),\ 1 \le i \le n\}$. In most environmental applications this will consist of a fixed time series of observations at each measuring station s_i, but this format is not required for the methodology.

The method requires the specification of two parameters, the *time window* m_T and the *sampling fraction* f_c. Once these parameters are specified, the window is defined as follows. Suppose we want to predict or interpolate at a specific time t_0 and location s_0. Restrict the observations to those which lie within the time window $(t_0 - \frac{m_T}{2}, t_0 + \frac{m_T}{2})$. Within that window, pick out observations in order of space, i.e. first select all the observations at the spatial location closest to s_0, then those at the location second closest to s_0, and so on until a fixed number $n_c = nf_c$ of observations has been selected. Prediction at (t_0, s_0) will be based entirely on this group of n_c observations.

The next step is to consider the form of regression model suitable for both the mean and standard deviation of Z. Haas considered a general model of the form

$$Z(t,s) = \mu(t,s,\beta) + \psi(\mu(t,s,\beta))R(t,s) \tag{3.1}$$

in terms of additional functions μ and ψ, where μ is typically a regression function of covariates such as meteorology, in terms of additional parameters β, which are also estimated separately within each window.

For model fitting and kriging, it is necessary to specify a suitable spatio-temporal corelation structure for the residual process $R(s,t)$, restricted to the given window. The basic covariance model assumed by Haas is

$$C\{R(t_1,s_1), R(t_2,s_2)\} = C_T(t_2 - t_1)C_S(s_2 - s_1) \tag{3.2}$$

where C_T denotes the temporal covariance function and C_S the spatial covariance function, each of which has been assumed stationary within the window. For the functions C_T and C_S, he assumed the "spherical" form of isotropic covariance structure considered in section 2.1. In other words, the same form of covariance is assumed for both the spatial and temporal scale, though of course the parameters may be quite different for the two functions. In the case of the spatial component, it is possible to relax the isotropy assumption, while still retaining stationarity, by allowing geometric anisotropy.

The product form of equation (3.2), in which the spatio-temporal covariance function is written as a product of a function of space and a function of time, is known as the *separability* assumption and is widely discussed in the context of time-space processes. It

is an assumption which is very widely used because of its convenience, though it is often criticized as unrealistic when applied to actual time-space data. For the moment, we shall accept this assumption, but when we discuss time-space processes in more detail later on, we shall examine this aspect much more critically.

One specific way in which (3.2) may fail is if the processes are strongly seasonal, so that different correlation functions apply in different seasons. Haas discussed this aspect specifically and proposed doing separate analyses by season when it occurs.

Once the model functions for μ, ψ, C_T and C_S are parametrically specified, under as assumption of joint normality, we could in principle estimate the model by maximum likelihood. Haas avoided this but instead described an algorithm including first OLS and later GLS regression to estimate the parameters of μ and ψ, along with the approximate WLS procedure to estimate the parameters of C_T and C_S. We shall not describe the algorithm but instead refer to Haas' paper for the details. All the estimation is restricted to observations within the space-time window about (t_0, s_0).

Finally, once the model is fitted, kriging formulae as in section 2.4 are used to calculate an optimal predictor at (t_0, s_0), say $Z^*(t_0, s_0)$, and its prediction standard error, $S_e(t_0, s_0)$.

Checking the model and selecting the window size

The key idea here is *cross-validation.* Select a cross-validation subsample of time-space locations $\{x_i = (t_i, s_i),\ i = 1, ..., n_{CV}\}$. For each x_i, let $Z^*(x_i)$ with prediction standard error $S_e(x_i)$ constructed from a data set in which $Z(x_i)$ has been removed (this process includes re-estimation of the entire model based on the reduced data set). One may then construct *cross-validation residuals* $R_i^{(CV)}$ and *standardized cross-validation residuals* $R_i^{(SCV)}$ through the formulae

$$R_i^{(CV)} = Z^*(x_i) - Z(x_i), \qquad R_i^{(SCV)} = \frac{Z^*(x_i) - Z(x_i)}{S_e(x_i)}. \tag{3.3}$$

Both forms of residuals are valuable in checking the model. For instance, if the model is fitted correctly the standardized residuals $R_i^{(SCV)}$ should be approximately independent $N(0, 1)$ random variables, and this could be checked using normal quantile plots. This idea was discussed in some detail by Haas (1990).

A second usage of residuals is to aid selection of the initial model parameters m_T and f_c, which determine the size of the window. For this, is is more convenient to use the unstandardized residuals. Haas recommended calculating a variance ratio

$$\frac{\bar{S}_e^2}{MSE} \tag{3.4}$$

where $\bar{S}_e^2$ is the mean square of the S_e values, in other words the estimated prediction standard errors according to the kriging formulae, averaged over the cross-validation data

points, while MSE is the observed mean squared error of the $R_i^{(CV)}$ values. If everything is working correctly, (3.4) should be close to 1. In practice, $\bar{S}_e^2$ may be biased by two factors, (i) the lack of homogeneity in the process, and (ii) estimation error in determining the parameters of the process. The first tends to be the dominant effect if the window width is too large and the second if the window width is too small. Both act, in general, to make $\bar{S}_e^2$ an underestimate of the true mean squared prediction error. However, this also suggests using the variance ratio as a criterion for selecting m_T and f_c.

Focussing specifically on f_c, Haas gave a number of examples in which (3.4) was evaluated for several values of f_c, the objective being to find a value for which the variance ratio was close to 1.

However, one might question whether using (3.4) as a cross-validation criterion is the best or most reasonable approach. In other contexts in which cross-validation is used in statistics, the focus is on minimizing prediction error variance itself (rather than minimizing the error with which we can estimate the prediction error variance). It is not clear whether this criterion would lead to substantially different estimates from those used by Haas.

Reconstructing the full covariance matrix

One disadvantage of the woving-window approach is that it does not lead to a single model to describe the whole data set. For example, different covariance functions are fitted to different portions of the data set, and if we simply combine these together to form an overall estimated covariance matrix, the result may not be positive definite. This problem was addressed specifically by Haas (1998), who proposed the following solution.

Suppose we are interested in estimating the covariance matrix at $n+q$ spatial-temporal locations $\{x_i = (t_i, s_i),\ i = 1, ..., n+q\}$ where $x_1, ..., x_n$ are the observed sampling locations and $x_{n+1}, ..., x_{n+q}$ are q additional sampling locations for whom we want to calculate covariances with the rest of the data, for instance because we want to do kriging at those points. The objective is to contruct an estimate $\widehat{\Sigma}$ of the covariance matrix at these $n+q$ locations.

It is possible to use the moving-window approach to construct an initial estimated matrix $P = (p_{ij})$, as follows. For locations x_i and x_j, define m_{ij} to be the midpoint between x_i and x_j. Let p_{ij} be the covariance between $Z(x_i)$ and $Z(x_j)$ according to the model fitted to the data centered at m_{ij}. Then we might expect the individual p_{ij} to be good estimates of the true covariances σ_{ij} say, but there is no guarantee that the matrix P is positive definite. We seek a matrix $\widehat{\Sigma}$ with entries $\{\widehat{\sigma}_{ij}\}$ that is "close" to P in some suitably defined sense, but also positive definite or at least positive semi-definite.

In considering how close P is to $\widehat{\Sigma}$, Haas considered two possible metrics:

(a) *Frobenius norm*: $||P - \widehat{\Sigma}||_F = \{\sum_i \sum_j (p_{ij} - \widehat{\sigma}_{ij})^2\}^{1/2}$,

(b) *the 2-norm*: $||P-\widehat{\Sigma}||_2 = \{\rho((P-\widehat{\Sigma})^T(P-\widehat{\Sigma}))\}^{1/2}$, where $\rho(A)$ denotes the largest eigenvalue of a symmetric matrix A.

Higham (1988), following Halmos (1972), showed that if P is a *normal* matrix — which includes any symmetric matrix with real components — then there exists a unique positive semi-definite matrix which is closest to P in both the Frobenius norm and the 2-norm. Moreover, Higham (1988) gave an algorithm to calculate this matrix. Haas' (1998) proposal was, in effect, to use this procedure to determine an estimated covariance matrix $\widehat{\Sigma}$ that is close to P while also satisfying the property of being positive semi-definite.

3.2 The EOF method and extensions

A very general method of representing the covariance function of a stochastic process without any stationarity conditions is through *empirical orthogonal functions*, usually abbreviated to EOFs. In the signal processing literature the method is also known as the *Karhunen-Loève expansion*; when restricted to a finite set of observations, it is equivalent to the well-known statistical method of principal components. The method appears to have been discovered independently by a number of researchers in the 1940s, for example Kosambi (1943), Loève (1945, 1946), Karhunen (1946, 1947), Obukhov (1947, 1954). Holmström (1963) may have been the first to consider explicitly its application to atmospheric sciences; Cohen and Jones (1969) developed the method from a statistician's perspective. Other works surveying its origins include North (1984) and Yaglom (1987). In section 3.2.1 we shall initially follow Cohen and Jones (1969), who developed the method from familiar statistical concepts. Examples from climate analysis are given in section 3.2.2.

In recent work, statisticians have experimented with hybrids between the classical EOF method and other approaches to spatial analysis; for example, Nychka and Saltzman (1998) and Holland *et al.* (1999) used covariance functions which are a mixture of a traditional stationary, isotropic covariance function and an EOF expansion, and Nychka, Wikle and Royle (1999) have tried an alternative form of expansion in terms of wavelet basis functions. We shall review these developments more briefly in sections 3.2.3 and 3.2.4.

3.2.1 The EOF expansion

Cohen and Jones (1969) introduced the method in the context of doing regression about a variable of interest y say, as a linear function of a field $X(s)$, where s ranges over some domain $\mathcal{D}$. In practice, of course, the field X would be sampled at only finitely many points, but this point is discussed later. In the specific example which they used to illustrate the method, y was temperature at Washington National Airport and X was the pressure field over the northern hemisphere, but the ideas would work equally well if y was the value of $X(s_0)$ at some point $s_0 \notin \mathcal{D}$. Thus, it is possible to view this as an alternative approach to the classical kriging problem, but without stationarity assumptions.

Suppose, then, we have replicated observations $(X_i, y_i,\ i = 1, 2, ..., n)$, where $X_i(s)$, $s \in \mathcal{D}$ are independent realizations of a random field on $\mathcal{D}$ and y_i are observations of the quantity of interest. By decomposing the distribution of y_i into components in the linear space spanned by $X_i(s),\ s \in \mathcal{D}$ and an orthogonal component ϵ_i, we write

$$y_i = \int_{\mathcal{D}} X_i(s)B(s)ds + \epsilon_i, \quad 1 \leq i \leq n, \tag{3.5}$$

with ϵ_i uncorrelated with $X_i(s),\ s \in \mathcal{D}$.

By easy generalization of the standard normal equations of linear regression theory, the least squares estimator of $B(s)$ solves

$$\sum y_i X_i(t) = \int_{\mathcal{D}} \sum_i X_i(x)X_i(t)B(s)ds. \tag{3.6}$$

For simplicity, assume the y_i and $X_i(s)$ all have mean 0.

As $n \to \infty$, we have $n^{-1}\sum y_i X_i(t) \to C_y(t)$ and $n^{-1}\sum_i X_i(x)X_i(t) \to C(s,t)$ where $C_y(t)$ is the covariance of y_i and $X_i(t)$ and $C(s,t) = \mathrm{Cov}\{X_i(s), X_i(t)\}$. Hence by (3.6), in the limiting case the optimal $B(s)$ solves

$$C_y(t) = \int_{\mathcal{D}} C(s,t)B(s)ds. \tag{3.7}$$

Thus, we require to solve an integral equation with kernel $C(s,t)$.

In practice, it is likely that we would have only a finite number of observations of $X_i(s)$ for each i, so we pretend that $\mathcal{D}$ is split up into m segments of areas $w_1, ..., w_m$ centered at m observation points $s_1, ..., s_m$. Approximating the integral in (3.5) by a finite sum, we then get

$$y_i = \sum_j w_j X_i(s_j)B(s_j) + \epsilon_i, \quad 1 \leq i \leq n,$$

or, writing x_{ij} in place of $w_j X_i(s_j)$ and β_j in place of $B(s_j)$, we get the equation

$$y_i = \sum_j x_{ij}\beta_j + \epsilon_i, \quad 1 \leq i \leq n,$$

which is of course exactly the form usually assumed in linear regression.

To make further progress with (3.7), we need the *Karhunen-Loève expansion* of the covariance function $C(s,t)$, see e.g. Yaglom (1987). The basic idea is to solve the integral equation

$$\int_{\mathcal{D}} C(s,t)\psi(t)dt = \lambda\psi(s), \tag{3.8}$$

in terms of eigenvalues λ and corresponding eigenfunctions ψ. According to the general theory of integral equations with positive definite symmetric kernels, such an equation has a countable set of eigenvalues, which we may without loss of generality assumed to be ordered as $\lambda_1 \geq \lambda_2 \geq ... \geq 0$, and corresponding $\psi_1, \psi_2, ...$ Under very mild conditions the system $\{\psi_\nu,\ \nu \geq 1\}$ is *complete* and *orthonormal* where the second condition means that

$$\int_{\mathcal{D}} \psi_\mu(s)\psi_\nu(s)ds = \begin{cases} 1 & \text{if } \nu = \mu \\ 0 & \text{if } \nu \neq \mu \end{cases}$$

and the first condition means that any square integrable function g may be expanded as an infinite linear combination of $\psi_\nu,\ \nu \geq 1$, in the form

$$g(s) = \sum_\nu a_\nu \psi_v(s), \tag{3.9}$$

where

$$a_\nu = \int_{\mathcal{D}} g(s)\psi_\nu(s)ds.$$

In particular, if we apply such expansion to $C(s,t)$ itself, we get

$$C(s,t) = \sum_\nu \lambda_\nu \psi_\nu(s)\psi_\nu(t). \tag{3.10}$$

Another consequence of (3.9) is that we can expand the random process $X(s)$ (f which $X_i(s),\ i = 1, 2, ...$, are independent copies) in multiples of the ψ_ν:

$$X(s) = \sum_\nu z_\nu \lambda_\nu^{1/2} \psi_\nu(s), \tag{3.11}$$

where $z_\nu,\ \nu = 1, 2, 3, ...$ are uncorrelated random variables with mean 0 and common variance 1 — in the case of a Gaussian process, they are independent $N(0,1)$ random variables.

The simplest way to see (3.11) (in the Gaussian case) is to start with this as the definition of the process $X(t)$, and then verify that

(a) Each $X(s)$ is a linear combination of the z_μ and therefore has a normal distribution; the same is true for any finite linear combination of the form $\sum_j c_j X(s_j)$; therefore, any finite set $X(s_1), X(s_2), ..., X(s_m)$ has a multivariate normal distribiution, which fulfils the definition for $X(s),\ s \in \mathcal{D}$ to be a Gauusian process.

(b) If $X(s)$ is defined for each s by (3.11), then $\text{Cov}\{X(s), X(t)\} = C(s,t)$ for each s, t. This follows very quickly by combining (3.11) with (3.10).

Therefore, the process defined by (3.11) is a Gaussian process with the required covariance function, so we might as well take (3.11) as the definition of the process X.

In the finite case where the process is only observed at finitely many points $s_1, ..., s_m$ with corresponding weights $w_1, ..., w_m$, the integral equation (3.8) is approximated by

$$\sum_{k=1}^{m} C(s_j, s_k)\psi_\nu(s_k)w_k = \lambda_\nu \psi_\nu(s_j). \tag{3.12}$$

If we write

$$C(s_j, s_k) = \frac{\Gamma_{jk}}{\sqrt{w_j w_k}}, \qquad \psi_\nu(s_j) = \frac{v_j^{(\nu)}}{\sqrt{w_j}},$$

then (3.12) becomes

$$\sum_{k=1}^{m} \Gamma_{jk} v_k^{(\nu)} = \lambda_\nu v_j^{(\nu)}, \tag{3.13}$$

in other words, λ_ν is the ν'th eigenvalue of the matrix Γ with entries Γ_{jk}, and the corresponding eigenvector is $v^{(\nu)} = \begin{pmatrix} v_1^{(\nu)} & \dots & v_m^{(\nu)} \end{pmatrix}^T$. This is a principal components decomposition of the covariance matrix Γ, and in practice would typically be found by applying a principal components analysis to the sample covariance matrix.

We now return to the integral equation (3.7) which gives the optimal coefficients $B(s)$ for predicting a random variable y_i. If we write

$$B(s) = \sum_\nu \beta_\nu \psi_\nu(s) \tag{3.14}$$

and substitute in (3.7), formally interchanging the integral and sum, we get

$$\begin{aligned} C_y(t) &= \sum_\nu \int_{\mathcal{D}} C(s,t)\beta_\nu \psi_\nu(s) ds \\ &= \sum_\nu \beta_\nu \lambda_\nu \psi_\nu(t), \end{aligned}$$

and so

$$\beta_\nu = \frac{1}{\lambda_\nu} \int C_y(t)\psi_\nu(t) dt. \tag{3.15}$$

Equations (3.14) and (3.15) together define the function $B(s)$ which solves the integral equation (3.7).

We can also combine (3.5) with (3.11) (writing $X_i(s)$ for $X(s)$, $z_{i\nu}$ for z_ν) to get

$$\begin{aligned} y_i &= \sum_\nu z_{i\nu} \lambda_\nu^{1/2} \int_{\mathcal{D}} \psi_\nu(s) B(s) ds + \epsilon_i \\ &= \sum_\nu z_{i\nu} \lambda_\nu^{1/2} \beta_\nu + \epsilon_i \end{aligned} \tag{3.16}$$

In practice, we replace (3.11) by

$$X_i(s_j)w_j^{1/2} = \sum_{\nu=1}^{m} z_{i\nu}\lambda_n u^{1/2}\psi_\nu(s_j)w_j^{1/2}$$
$$= \sum_{\nu=1}^{m} z_{i\nu}\lambda_n u^{1/2}v_j^{(\nu)},$$

and hence (3.16) becomes, on truncating the sum at some N,

$$y_i = \sum_{\nu=1}^{N} z_{i\nu}\lambda_\nu^{1/2}\beta_\nu + \epsilon_i \tag{3.17}$$

which is equivalent to a *principal components regression* for y_i. To apply the method in practice we must choose N, but in principal components regression there are many ways to do this corresponding to the familiar tools that are used in deciding how many variables to include in a regression equation, e.g. sequential F tests, AIC, BIC, Mallows' C_p criterion, and so on.

Summary

1. For a very general class of covariance functions $C(s,t)$ without stationarity conditions, one can find a complete orthonormal basis of eigenfunctions ψ_ν with corresponding eigenvalues λ_ν; the process $X(s)$ may then be expanded through (3.11) and the covariance $C(s,t)$ through (3.10).

2. For predicting a random variable y_i in terms of $X_i(s)$, $s \in \mathcal{D}$, the optimal linear predictor is of the form $\int_{\mathcal{D}} X_i(s)B(s)ds$, when $B(s)$ solves the integral equation (3.7) and is given explicitly by (3.14) and (3.15).

3. In practice, if we observed the process $X_i(s)$ at only a finite number of locations $s = s_1, ..., s_m$ with weights $w_1, ..., w_m$, and if we define a matrix Γ with entries $\Gamma_{jk} = C(s_j, s_k)\sqrt{w_j w_k}$, then finding the Karhunen-Loève expansion is equivalent to a principal components decomposition for Γ, and the optimal prediction problem for y_i is equivalent to fitting a principal components regression.

3.2.2 Applications to climate change

The EOF method is not widely used as a conventional statistical methodology. However, it has become very popular as a tool in climate research, particularly, related to detection and attribution questions. A number of recent references on this approach include Hegerl *et al.* (1996), Hegerl and North (1997), Hasselmann (1997), Allen and Tett (1999). Here, we give a brief description of the method of Allen and Tett (1999), refering the reader to the cited references for earlier versions of their approach.

The problems of detection and attribution in climate change essentially have to do with how one can measure the agreement between an artificially generated climate signal, such as would typically be produced by a numerical general circulation model, and real data as measured by surface observation stations or satellites. Typically, both the observational and model-generated data are aggregated into grid cells across the earth's surface, but there are several thousand grid cells to cover the whole surface. For example, one common scheme used 5° latitude and longitude grid cells, implying 36 latitude classifications (from 90°N to 90°S in 5° increments, and similarly 72 longitude increments, giving a total of 2,592 grid cells. In the following discussion we shall use ℓ to denote the number of grid cells. One common kind of study is to generate m possible "response patterns" $x_k,\ k = 1, ..., m$ using the climate model, each one an ℓ-dimensional time series corresponding to the response to one particular kind of forcing factor. For example, x_1 may be the response due to the increase in carbon dioxide, x_2 the response due to changes in sulfate aerosols, x_3 the response due to change in solar forcing, and so on — each of these factors being believed to influence the observed climate change, according to model theories of the causes of climate change. Assuming that the observed climate signal is a linear combination of the components due to different forcing factors, this suggests a model of the form

$$y = X\beta + u, \tag{3.18}$$

where y is $\ell \times 1$, $X = (x_1 \quad ... \quad x_m)$ is a $\ell \times m$ matrix consisting of all the modeled responses to forcing factors, and β is the vector of coefficients which we are trying to estimate. If the residual vector u in (3.18) has covariance matrix C_N, then the optimal generalized least squares (GLS) estimator of β is

$$\hat{\beta} = (X^T C_N^{-1} X)^{-1} X^T C_N^{-1} y, \tag{3.19}$$

and

$$\mathrm{Var}\{\hat{\beta}\} = (X^T C_N^{-1} X)^{-1} \tag{3.20}$$

The crucial difficulty is that C_N is unknown and therefore we need some reliable method of estimating it.

One way of deriving (3.19) and (3.20) from the standard ordinary least squares (OLS) regression equations is to define a matrix P so that $PC_N P^T = I_\ell$, and then to write the regression equation (3.18) in the form $Py = PX\beta + Pu$ where the covariance matrix of Pu is I_ℓ. The OLS estimator for β is then $\hat{\beta} = (X^T P^T P X)^{-1} X^T P^T P y$ with covariance $(X^T P^T P X)^{-1}$. But if P and C_N are invertible we will have $P^T P = C_N^{-1}$, which leads to (3.19) and (3.20).

The approach that has been widely developed in climate studies is to take model observations from runs of the GCM that are conducted in stationary conditions without forcing factors. Typically such runs are 1,000 to 2,000 years long and therefore provide long enough data series to allow one to estimate climate variability with reasonable accuracy, assuming that the climate model gives an adequate representation of how the real climate

would look if there were no external influences due to greenhouse gases or other effects. Thus, with observations Y_N from n years' of unforced model runs centered to 0 mean, one could estimate

$$\hat{C}_N = \frac{1}{n} Y_N Y_N^T. \tag{3.21}$$

The difficulty with (3.21) is that typically $\ell > n$ and so $\hat{C}_N$ is a singular matrix. Even though this in itself is not a fatal objection, e.g. valid versions of (3.19) and (3.20) can be given involving generalized inverses, the real difficulty is that with so many degrees of freedom in the covariance function, the low-amplitude components of covariance are not reliably estimated, and therefore, any direct attempt to apply (3.20) by substituting $\hat{C}_N$ for C_N will give poor estimates for $\hat{\beta}$ and innaccurate estimates of uncertainty.

The solution is to define a $\kappa \times \ell$ transformation matrix $P^{(\kappa)}$ corresponding to the κ largest EOFs of Y_N, so that $P^{(\kappa)} \hat{C}_N {P^{(\kappa)}}^T = I_\kappa$, and then to define the OLS regression equation

$$P^{(\kappa)} y = P^{(\kappa)} X \beta + P^{(\kappa)} u,$$

which leads to an estimator

$$\tilde{\beta} = (X^T {P^{(\kappa)}}^T P X)^{-1} X^T {P^{(\kappa)}}^T P^{(\kappa)} y \tag{3.22}$$

and estimated covariance

$$(X^T {P^{(\kappa)}}^T P^{(\kappa)} X)^{-1}. \tag{3.23}$$

In practice, numerous refinements of this basic methodology have been proposed to protect the resulting estimates from various sources of bias. However, the statistical basis for these refinements is in many cases rather unclear. Among the issues discussed by Allen and Tett (1999) are:

(1) The direct estimate of covariance (3.23) is likely to be an underestimate of the true covariance matrix, and an alternative is

$$\hat{\mathrm{Var}}\{\tilde{\beta}\} = (X^T \hat{C}_{N_1}^{-1} X)^{-1} X^T \hat{C}_{N_1}^{-1} C_{N_2} C_{N_1}^{-1} X)(X^T \hat{C}_{N_1}^{-1} X)^{-1} \tag{3.24}$$

with $\hat{C}_{N_1}$ the estimate of C_N from the EOF construction and $\hat{C}_{N_2}$ an independent estimate of the covariances of the κ largest EOFs from a separate run of the climate model. The use of independent unforced model runs to correct the estimate of variance was apparently first suggested by Hegerl *et al.* (1996).

(2) The estimates are corrected for serial correlation, using a method of Zwiers and von Storch (1995).

(3) The model runs for the forcing conditions (columns of X) are also subject to random noise and therefore it would be desirable to correct for that source of bias.

(4) There remains the question of how in practice to choose κ. Allen and Tett proposed a residual test based on

$$r^2 = (y - X\tilde{\beta})^T \hat{C}_N^{-1} (y - X\tilde{\beta}) \sim \chi^2_{\kappa - m}, \tag{3.25}$$

where rejection of r^2 based on (3.25) is taken as an indication that κ is too large.

3.2.3 Combining stationary models and EOFs

Nychka and Saltzman (1998) suggested an alternative form of covariance model combining a traditional geostatistical model (stationary, isotropic) with a truncated EOF expansion, in the form

$$C(s_1, s_2) = \sigma(s_1)\sigma(s_2) \left\{ \rho e^{-||s_1 - s_2||/\theta} + \sum_{\nu=1}^{M} \lambda_\nu \psi_\nu(s_1) \psi_\nu(s_2) \right\} \tag{3.26}$$

which permits the standard deviation to vary with location s according to a general function $\sigma(s)$, and a leading term which corresponds to a stationary isotropic model of exponential covariance type. The remaining terms depend on eigenvalues λ_ν and eigenfunction ψ_ν of the covariance operator in simlar fashion to (3.10), and essentially allow various degrees of nonstationarity according to the value of the index M.

By analogy with (3.11), if the process is Gaussian then we also have for the observed process $Z(s)$,

$$Z(s) = \sigma(s) \left\{ \rho^{1/2} Z_0(s) + \sum_{\nu=1}^{M} a_\nu \lambda_\nu^{1/2} \psi_\nu(s) \right\} \tag{3.27}$$

in which $Z_0(s)$ is a stationary isotropic process with covariance $e^{-||s_1 - s_2||/\theta}$ and the random coefficients a_ν are standard normal random variables, independent both of each other and of the process Z_0.

As an example, Nychka and Saltzman (1998) fitted the model (3.26) to ozone data in the neighborhood of Chicago, using both real data collected by ozone monitors and synthetic data generated by a computer model (the regional oxidant model or ROM). In both cases, multiple replications of the field were used both to estimated the site-by-site variances $\sigma^2(s_j)$ and to separate the correlation function into stationary and nonstationary components; it would probably not be appropriate to attempt these models in the classical "geostatistics" set-up of one onservation of the entire field. Nychka and Saltzman estimated $\rho = 0.5$ for the ROM data, 0.25 for the observational data, in either case a substantial departure from stationarity; they used $M = 5$ in the expansion.

More details of the method were provided by Holland *et al.* (1999), who applied the same ideas in modeling sulfur dioxide measurements from CASTNet (the EPA's *Clean Air*

Status and Trends Network), which covers the whole eastern U.S. They extended (3.26) and (3.27) to include a nugget effect:

$$C(s_1, s_2) = \sigma(s_1)\sigma(s_2)\left[\rho\left\{(1-\alpha)\delta(s_1 - s_2) + \alpha e^{-||s_1 - s_2||/\theta}\right\} + \sum_{\nu=1}^{M} \lambda_\nu \psi_\nu(s_1)\psi_\nu(s_2)\right], \tag{3.28}$$

where $0 < \alpha \leq 1$ and $\delta(y)$ is 1 if $y = 0$, 0 otherwise,

$$Z(s) = \sigma(s)\left\{(\alpha\rho)^{1/2} Z_0(s) + \sum_{\nu=1}^{M} a_\nu \lambda_\nu^{1/2} \psi_\nu(s)\right\} + \epsilon(s) \tag{3.29}$$

where $\epsilon(\cdot)$ is a white noise process, $\epsilon(s) \sim N[0, \sigma^2(s)(1-\alpha)\rho]$ independently at each site s.

To estimate the model (3.28) they proposed the following procedure. Assuming for the moment that $\alpha, \sigma(s), \rho$ and θ are all known, define

$$R_{j,k} = \hat{C}(s_j, s_k) - \sigma(s_j)\sigma(s_k)\left[\rho\left\{(1-\alpha)\delta(s_j - s_k) + \alpha e^{-||s_j - s_k||/\theta}\right\}\right], \tag{3.30}$$

where $C(s_j, s_k)$ is the sample covariance of sites s_j and s_k. Then form an eigenvalue-eigenvector decomposition of the matrix R, retaining the M largest eigenvalues and associated eigenvectors. These then define the λ_ν and $\psi_n u$ of (3.28). The values of α and $\sigma(s_j)$ were assumed known (in the example of Holland *et al.*, they were estimated from a preliminary thin-plate spline fit to the site-by-site variances), and ρ and θ were estimated by a grid-search algorithm: for each candidate pair of values for ρ and θ, $\lambda_1, ..., \lambda_M$ and $\psi_1, ..., \psi_M$ were estimated from the R matrix given by (3.30); they in turn were substituted into (3.28) to obtain a model-based estimate $\hat{C}(s_j, s_k)$; the final objective function to minimize was

$$\sum_{j \leq k} \{C(s_j, s_k) i \hat{C}(s_j, s_k)\}^2.$$

3.2.4 Wavelet expansions

Nychka, Wikle and Royle (1999) proposed alternative expansions using wavelet basis functions. Since the method is still developing, we present only an outline of the approach here.

Their approach is motivated by a number of practical features of applying spatial prediction in large systems:

- Large networks, of say 5,000 stations, cannot be handled by traditional geostatistics and kriging methods because the numerical difficulties of storing and performing matrix operations such as inversion on matrices of order $5,000 \times 5,000$ are intractable;

- To model such systems realistically it is essential to use a nonstationary covariance matrix;

- Efficient numerical methods are based on iterative procedures such as conjugate gradient algorithms for solving large systems of linear equations;

- The methods assume that the data exist on a finite regular grid though not necessarily that all the grid points are observed;

- They may be easily extended to spatial-temporal systems involving a nontrivial temporal covariance structure.

The basic model assumes an expansion of the process

$$Z(s) = \sum_{\nu=1}^{MN} a_\nu \psi_\nu(s), \tag{3.31}$$

in which we assume an $M \times N$ grid, the basis functions ϕ_ν are given and the a_ν, $\nu = 1, ..., MN$ are random coefficients whose joint distribution is $N_{MN}[0, \Sigma_a]$ for some covariance matrix Σ_a. If we define a matrix Ψ with entries $\Psi_{\mu,\nu} = \psi_\nu(s_\mu)$, then the covariance matrix for the vector with coefficients $Z(s_\mu)$ is given by

$$\Sigma_Z = \Psi \Sigma_a \Psi^T. \tag{3.32}$$

Unlike the case where the ψ_ν are eigenfunctions of the covariance function of Z, we can no longer assume that Σ_a is a diagonal matrix; nevertheless, one of the motivations of the approach is that in practice, one can approximate (3.32) very well with a Σ_a which is either diagonal or of some simple structure such as tri-diagonal. Nychka *et al.* gave numerical examples based on diagonal Σ_a which showed that one could obtain a very good approximation through (3.32) of some standard stationary, isotropic covariance functions such as exponential and Gaussian; the approximation performed well at the center of the grid but not so well at the edges.

The actual method they used was based on a specific approach to multiresolution basis functions using the W transform (Kwong and Tang 1994). We shall not attempt to explain the details of this; the method is motivated more by the need to define a representation which facilitates linear algebra computations than one which is good in a functional approximation sense, though it appears to perform well from the latter point of view as well.

3.3 Deformation methods

In this section we review methods based on the idea that a process might be stationary and isotropic only after some deformation of the space of observations, and we describe

methods which estimate both the deformation itself and the spatial covariance structure in the deformed space. The original ideas on this method were due to Sampson and Guttorp (1992) and extended, with a somewhat different estimation method, by Guttorp and Sampson (1994) and Guttorp, Meiring and Sampson (1994). Their methods are described in section 3.3.1. Their estimation methods were not likelihood-based but an alternative, aimed at producing maximum likelihood estimates based on some suitable parameterization of the deformation, is given in section 3.3.2. Tow applications of this methodology are presented in sections 3.3.3 and 3.3.4. Finally, section 3.3.5 outlines some of the more recent developments, including attempts to reformulate the problem in a fully Bayesian context.

3.3.1 The Sampson-Guttorp approach

Throughout this section, we will refer to either the covariance function

$$C(s_1, s_2) = \text{ Cov } \{Z(s_1), Z(s_2)\}, \quad s_1, s_2 \in D,$$

or the *dispersion*

$$D(s_1, s_2) = \text{ Var } \{Z(s_1) - Z(s_2)\}, \quad s_1, s_2 \in D.$$

If either $C(s_1, s_2)$ or $D(s_1, s_2)$ depends on s_1 and s_2 only through the Euclidean distance between the two stations $||s_1 - s_2||$, then we shall call the process *homogeneous*. This corresponds to the concepts of stationarity (or intrinsic stationarity if defined from the dispersion) and isotropy that were discussed in Chapter 2. Classical geostatistics, as reviewed in Chapter 2, is based primarily on homogeneous models, though there are some fairly simple non-homogeneous models that can be handled by the same methods as classical goestatistics. Two fairly well-known examples (Journel and Huijbregts 1978) are based on either

$$D(s_1, s_2) = 2\gamma_0(||A_0(s_1 - s_2)||),$$

which is known as *geometric anisotropy*, or its extension

$$D(s_1, s_2) = 2\sum_{j=0}^{J-1} \gamma_j(||A_j(s_1 - s_2)||).$$

known as *zonal anisotropy*. Here $A_0, A_1, ...$, are arbitrary matrices and $\gamma_0, \gamma_1, ...$, isotropic semivariogram functions. However, this is still quite a restrictive class of models.

Geometric anistropy is based on the notion that a simple linear transformation of the observation space will make the process homogeneous. Sampson and Guttorp (1992) proposed a much more radical extension based on nonlinear transformations.

$$D(s_1, s_2) = 2\gamma_0(f(s_1), f(s_2)) \tag{3.33}$$

with γ_0 again an isotropic semivariogram and f a smooth nonlinear map from $\mathcal{R}^d$ to $\mathcal{R}^{d'}$. In principle one may permit $d' \neq d$ though in most of the Sampson-Guttorp work it is assumed that $d' = d$ and we shall continue to assume that here. The idea behind (3.33) is that the map f takes the coordinates from the real, geographical or "G" space, into an alternative dispersion or "D" space in which the process is homogeneous. This approach may not be universally applicable to inhomogeneous processes, and in subsequent discussion we shall see some limitations to it, but it has become very widely applied because it seems to capture the nature of the nonstationarity in many environmental data sets. For an entertaining early example of how the interpretation of a spatial analysis may be completely changed by such a transformation, see Lewis (1989).

The original paper of Sampson and Guttorp (1992) adopted the following strategy, briefly summarized here, based on three steps:

(a) A mapping of the n sampling points from the G space into the D space is found to minimize a stress criterion

$$\min_{\delta} \frac{\sum_{i<j}\{\delta(d_{ij}) - h_{ij}\}^2}{\sum_{i<j} h_{ij}^2}$$

where d_{ij} is the observed dispersion between sites i and j, h_{ij} is the distance between sites i and j in D space and the minimization is taken over all monotonically increasing functions δ. This formulation of the problem permits it to be solved by a multidimensional scaling (MDS) algorithm.

(b) The mapping of the N sampling points is then extended to a *smooth* function from the entire G space into the D space, using a representation based on thin plate splines.

(c) The function δ is replaced by a smooth function g (so $d_{ij} \approx g(h_{ij})$), which satisfies the positive definiteness condition required for g to be the variogram of a homogeneous process. For this purpose, Sampson and Guttorp used a very general representation of g as a mixture of Gaussian-type variograms.

This approach has a number of *ad hoc* features and some clear disadvantages, e.g. the restriction that δ be monotone in step (a) (necessary for the MDS algorithm to make sense) means that we can only consider monotonically increasing variogram functions in step (c), and for some purposes, that may be an undesirable restriction. Guttorp and Sampson (1994) mentioned some other undesirable features of the approach, and Guttorp, Meiring and Sampson (1994) proposed an alternative version which seems more appealing: combining steps (a) to (c) into a single step, they proposed choosing the deformation f and the semivariogram γ_0 to minimize

$$\sum_{i,j} \left(\frac{d_{ij} - \hat{d}_{ij}}{\hat{d}_{ij}} \right)^2 + \lambda\{J(f_1) + J(f_2)\}, \tag{3.34}$$

where d_{ij} is the empirical (sample) dispersion between sites i and j, $\hat{d}_{ij}$ is the modeled dispersion based on (3.33), f_1 and f_2 are the x and y coordinates of the map f in (3.33), and $J(f_i)$ is the "bending energy" functional of a scalar function f_i, defined in the case of a two-dimensional map by

$$J(f_i) = \int\int \left\{ \left(\frac{\partial^2 f_i}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f_i}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f_i}{\partial y^2} \right)^2 \right\} dxdy. \tag{3.35}$$

The interpretation of $\lambda > 0$ in (3.34) is that it is a smoothing parameter — the larger λ, the smoother the map f is required to be. The motivation for the last term in (3.34) comes from the theory of thin-plate splines (Green and Silverman 1994), in which penalty functions of the form of $\lambda J(f)$, with λ an adjustable smoothing parameter, are essentially taken as the definition of smoothing splines.

Application of the methodology to ozone model assessment

As an example of the application of this approach to a fairly complicated real-data problem, we outline the main steps in the analysis of ozone fields by Meiring, Guttorp and Sampson (1998).

The purpose of this paper was to propose a method for assessing the agreement between real data obtained from 32 monitoring sites in a region of northern California, and the output of a numerical model known as SARMAP. The data on hourly ozone levels exhibited numerous features that required specialized statistical modeling. For example, there is a strong diurnal effect which influences not only the means and the variances of the ozone levels, but also the spatial correlations. In addition to the spatial correlation structure, the data showed temporal correlations which were different at each site. These features made it necessary not only to consider the problem as one of fitting a spatial-temporal model, but also meant that simple models for the spatial-temporal correlation structure, such as separability (where the spatial-temporal correlations factor into a product of a spatial correlation and a temporal correlation) could not capture the true structure of the data. Because of the importance of getting the diurnal effect right, a separate model was fitted to each of the 24 hours of the day.

With these preliminaries, the actual model fitted by Meiring *et al.* followed these steps:

(1) A square root transformation was applied to transform the data for each site-hour combination to approximate normality. The square root transformation was chosen after examining normal QQ plots of residuals of the raw data and of data with a square root and logarithmic transformation. The square root transformation was not universally the best of the three, but was judged to be the best overall.

(2) For each site-hour combination, the overall mean of the transformed data was estimated, and subtracted from the data to give residuals of the form

$$w_t(x_i) = \sqrt{z_t(x_i)} - h_k(x_i)$$

where $z_t(x_i)$ is the original observation at time t and site x_i, and $h_k(x_i)$ for $k = 1, ..., 24$ is the overall mean of all $\sqrt{z_t(x_i)}$ values correspnding to the k'th hour of the day.

(3) An AR(2) model was fitted to each of the $w_t(x_i)$ series, with coefficients fitted separately at each site, thus

$$w_t(x_i) = \alpha_1(x_i)w_{t-1}(x_i) + \alpha_2(x_i)w_{t-2}(x_i) + y_t(x_i),$$

with $y_t(x_i)$ a residual process. After fitting this model, the residuals are approximately uncorrelated in time but of course they are still spatially correlated.

(4) Letting $\sigma_{xx}(t)$ denote the variance of $y_t(x)$, and empirical dispersion function

$$D_y(u, x) = \text{Var}\left\{\frac{y_t(u)}{\sqrt{\sigma_{uu}(t)}}\frac{y_t(x)}{\sqrt{\sigma_{xx}(t)}}\right\}$$

was computed, separately for each of the 24 hours of the day. (In fact, to improve the precision of the estimates and to avoid some problems with missing values, the dispersions were calculated using all observations within a three-hour moving window, but still a separate estimate was computed for each of the 24 hours.) The variance functions $\sigma_{xx}(t)$ were assumed to be constant in space but again dependent on the hour of the day.

(5) The empirical dispersions $D_y(u, x)$ were then used as the input to estimating the model (3.33), using an exponential semivariogram function with nugget for γ_0, and λ chosen by visual inspection of the fitted variograms and deformation maps. A cross-validation approach might have been used, but given the amount of computation required, a visual approach was preferred. However, the same parameter λ was used for each of the 24 hours. Apart from that aspect, a completely separate model was fitted for each hour.

(6) The model fitted in step 5 was used to construct estimated (krigged) values of $y_t(x)$, together with appropriate prediction variances, for each time point t and each location x in a subgrid lying within a single grid cell of the SARMAP model. In addition, the mean functions $h_k(x)$ and the AR(2) coefficients $\alpha_1(x)$ and $\alpha_2(x)$ were interpolated to subgrid points using a simple numerical interpolator (the function `interp` in S-PLUS — it appears that no kriging was used for this step, nor did the authors attempt to allow for the uncertainty in the interpolated value). A technicality here is that the AR(2) coefficients must be constrained to remain within the stationarity region of the AR(2) model — the achieved this by first rewriting the model in terms of partial autocorrelations ($\alpha_1 = \phi_1(1 - \alpha_2)$, $\alpha_2 = \phi_2$) and then applying the actual interpolation to the functions

$$\beta_i(x) = \log\left\{\frac{1 + \phi_i(x)}{1 - \phi_i(x)}\right\}, \quad i = 1, 2,$$

to ensure that the interpolated $\phi_i(x)$ remained in the interval $(-1, 1)$.

(7) The interpolated values of $y_t(x)$, $h_k(x)$, $\alpha_1(x)$ and $\alpha_2(x)$ were used to reconstruct values of $z_t(x)$ for each time point t ad subgrid location x, by reversing the sequence of steps used to construct $y_t(x_i)$ from $z_t(x_i)$. In addition, prediction variances and covariances between locations were estimated. The variance and covariance calculations are actually the most technically involved part of the computation, since they involve reversing the square root transformation and hence required moments up to fourth order of the interpolated process. The calculations were described in detail by Meiring *et al.* in their paper, but we shall not attempt to reproduce them here.

(8) Finally, the sample average of the estimated $z_t(x)$, over all subgrid locations within a single grid cell of the SARMAP model, were computed as an estimate of the overall ozone level within the grid cell at time t, together with a variance of the overall average value computed using the variances and covariances computed in step 7.

The usefulness of this kind of analysis lies essentially in its ability to tell how well the deterministic numerical model is able to reproduce the monitoring data — a good numerical model can be used to examine the effects of possible changes in emissions control strategies and can be very important in understanding what factors are important in determining the level of ozone. In the numerical examples in their paper, Meiring *et al.* reported generally good agreement between the observation-based predictions and the output of the numerical model; the main specific discrepancy they noted was that the model tends to overestimated ozone levels during the night and early morning when the levels are generally low.

3.3.2 Maximum likelihood fitting

As already noted, the original paper of Sampson and Guttorp (1992) used an *ad hoc* approach to fitting the model, while the alternative criterion (3.34), though not *ad hoc*, is not a likelihood-based approach and therefore could be expected to be inferior in terms of statistical efficiency compared with maximum likelihood or Bayesian approaches. The possibility of fitting these models with a maximum likelihood approach was first pointed out by Mardia and Goodall (1993) and later extended by Smith (1996). The present discussion is largely based on the latter reference.

Consider the model represented as follows:

1. We have N replications $Z_1, ..., Z_N$ of a spatial field observed at each of n sites (thus $Z_k = (Z_k(s_1), ..., Z_k(s_n))$, where $s_1, ..., s_n$ are the sampling sites, for $k = 1, 2, ..., N$). These are assumed independent with

$$Z_k \sim N_n(\mu, \Sigma), \tag{3.36}$$

N_n denoting the n-dimensional normal distribution, μ an arbitrary n-vector of means and Σ an $n \times n$ covariance matrix.

2. We assume $\sigma_{ij} =$ Cov $\{Z_k(s_i), Z_k(s_j)\}$ of the form

$$\sigma_{ij} = C_0(f(s_i), f(s_j))$$

where C_0 is a homogeneous covariance function and f is represented by a linear combination of radial basis functions ((3.39) below).

3. For the initial analysis we assume C_0 has the Matérn structure

$$C_0(t) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{2\sqrt{\theta_2}t}{\theta_1} \right)^{\theta_2} \mathcal{K}_{\theta_2} \left(\frac{2\sqrt{\theta_2}t}{\theta_1} \right). \tag{3.37}$$

Here $\theta_1 > 0$ is the spatial scale parameter and $\theta_2 > 0$ is a shape parameter. The function $\Gamma(\cdot)$ is the usual gamma function while $\mathcal{K}_{\theta_2}$ is the modified Bessel function of the third kind of order θ_2 (Abramowitz and Stegun 1964, Chapter 9). This form was used by Handcock and Wallis (1994) for their analysis of climate data and it seems to be very widely applicable as a simple parametric form for spatial correlations. Nevertheless it is not universally appropriate and we shall propose a more general representation for two-dimensional isotropic covariances in (3.46) below.

The analysis being discussed here does not give any separate attention to analysis of means and variances at each spatial location, assuming that these are estimated by simple sample estimates, but concentrates on the spatial correlations. Thus, from now on, it is assumed that the process has been standardized to have mean 0 and variance 1 at each site.

With this simplification, the negative log likelihood based on $Z_1, ..., Z_N$ reduces to

$$L = \frac{N}{2} \log |\Sigma| + \frac{N-1}{2} \text{ tr } \left(\Sigma^{-1} \hat{\Sigma} \right) \tag{3.38}$$

where $\hat{\Sigma}$ is the usual $n \times n$ sample correlation matrix.

The traditional formulation of thin-plate splines (Green and Silverman 1994, Chapter 7) requires a function f to pass through a finite number of data points $z_i = f(x_i, y_i)$ $(i = 1, ..., n)$, to minimize the bending energy $J(f)$ given by (3.35). The solution to this problem may be represented in the form (Green and Silverman 1994, page 142)

$$f(x, y) = a + bx + cy + \sum_{i=1}^{n} \delta_i \eta_i(x, y) \tag{3.39}$$

where

$$\sum \delta_i = \sum \delta_i x_i = \sum \delta_i y_i = 0 \tag{3.40}$$

and

$$\eta_i(x, y) = r^2 \log r, \qquad r = \{(x - x_i)^2 + (y - y_i)^2\}^{\frac{1}{2}}. \tag{3.41}$$

Thus (3.39) represents f as a sum of linear terms and n *radial basis functions* η_i with *centers* at the observed data points (x_i, y_i). The constraints (3.40) ensure that the problem does not become overdetermined.

Smoothing splines differ from interpolating splines in that the fitted function f is no longer required to pass exactly through the given data points z_i: this is usually more appropriate in a statistical context when there is noise in the data. The usual formulation of smoothing splines is to minimize a function of the form

$$S(f) = \sum \{z_i - f(x_i, y_i)\}^2 + \lambda J(f) \tag{3.42}$$

where z_i is a possibly noisy observation of the function $f(x_i, y_i)$ and $\lambda > 0$ is a smoothing parameter.

Although the exact solution to (3.42) is computable (Green and Silverman 1994, page 147–148), in practice we do not usually have an *a priori* fixed value of λ and an alternative approach is simply to restrict the representation (3.39) to a subset of radial basis functions. Thus we assume

$$\delta_i = 0 \quad \text{for } i \notin \{i_1, ..., i_m\} \tag{3.43}$$

where $i_1, ..., i_m$ are some subset of indices to be determined. This approach is similar to the way radial basis functions have been used in non-linear time series analysis (Casdagli 1989, Smith 1993, Judd and Mees 1995), where they are an alternative to neural net representations.

In the present context f is a bivariate function, so we apply the RBF approach to each of its two components, $f^{(1)}$ and $f^{(2)}$ say. There is a potential difficulty with this in that a function constructed in this way may not be bijective. Non-bijective functions are a problem in the deformation approach because they correspond to a mapping which folds over itself, which in most contexts seems counterintuitive. The difficulty was noted by Sampson and Guttorp (1992) who suggested that, in most cases, the problem of folding can be avoided by choosing a sufficiently smooth map, which is equivalent to keeping m, the number of active RBFs, fairly small.

Some further simplification is possible. First, the constant a in (3.39) is unnecessary — this is so because the resulting covariance functions depend only on *differences* between coordinates in the D space and are therefore unaffacted by locations shifts in D space. So we set $a = 0$. Second, in the case $m = 0$, the model is invariant under orthogonal rotations. This suggests that, in the case $m > 0$ as well, we simplify the parametrization to

$$\begin{aligned} f^{(1)}(x, y) &= b_1^2 x + \rho b_1 b_2 y + \sum_1^n \delta_i^{(1)} \eta_i(x, y), \\ f^{(2)}(x, y) &= \rho b_1 b_2 x + b_2^2 y + \sum_1^n \delta_i^{(2)} \eta_i(x, y), \end{aligned} \tag{3.44}$$

where $b_1 > 0,\ b_2 > 0,\ \rho \in \mathcal{R}$, and each of the sequences $\{\delta_i^{(1)},\ i = 1, ..., n\}$ and $\{\delta_i^{(2)},\ i = 1, ..., n\}$ satisfy the constraints (3.40), (3.43). Finally we note that with f still permitting arbitrary scale changes, we may without loss of generality set $\theta_1 = 1$ in (3.37). Thus, whenever $m \geq 3$, the final model has $2m - 2$ free parameters $b_1,\ b_2,\ \rho,\ \theta_2,\ \delta_{i_1}^{(1)},\ \delta_{i_1}^{(2)}, ...,\ \delta_{i_{m-3}}^{(1)}$, $\delta_{i_{m-3}}^{(2)}$.

The model (3.37) does not allow for a *nugget effect*. This could be permitted, for example, by allowing the value for $C_0(0)$ to be greater than the limiting $t \to 0$ value obtained from (3.37). In many applications, no nugget effect is observed with the Matérn covariance structure, but with the climate data of section 3.3.3, it turns out that such an effect is needed. A much broader extension is to abandon the parametric form entirely and to represent C_0 nonparametrically. A general representation for a d-dimensional isotropic covariance function (recall Chapter 2) is given by

$$C_0(h) = \int_0^\infty Y_d(wh)\Phi(dw)$$

where $\Phi(\cdot)$ is a general positive measure on $[0, \infty)$ and

$$Y_d(t) = \left(\frac{2}{t}\right)^{\frac{d-2}{2}} \Gamma\left(\frac{d}{2}\right) J_{\frac{d-2}{2}}(t)$$

where J_v is the modified Bessel function of order v. In particular, when $d = 2$ this reduces to

$$C_0(h) = \int_0^\infty J_0(wh)\Phi(dw). \tag{3.45}$$

In practice the measure Φ may be assumed concentrated on a finite number of atoms, so (3.45) reduces to

$$C_0(h) = \sum_{c=1}^{C} \phi_c J_0(w_c h) \tag{3.46}$$

in terms of $2C$ parameters $\phi_1, w_1, ..., \phi_C, w_C$. Representations of the form of (3.46) were appraently first introduced by Shapiro and Botha (1991) and have been the basis of a number of subsequent proposals, e.g. Hall *et al.* (1994), Lele (1995), Cherry *et al.* (1996). Sampson and Guttorp (1992) used a similar representation but with $J_0(t)$ replaced by the Gaussian-type kernel e^{-t^2}. This was derived from the slightly less logical requirement that the function C_0 should be a positive definite isotropic covariance function in all dimensions simultaneously, rather than just in the specific dimension d in which we happen to be working. An advantage of the present approach is that it allows non-monotone functions C_0.

An important feature of the present approach is that only a subset of centers, represented by the indices $i_1, ..., i_m$ in (3.43), is included in the model. This is contrast to Mardia and Goodall (1993), who implicitly assumed that all the centers are included. Using all the

centers leads to intractable computational problems when the number of centers is large, and might also be expected to result in badly overfitted models. Therefore, we must limit the number of centers included, but the computational complexity of the problem rules out any attempt at an exhaustive search over subsets.

To simplify this problem, the n centers are first arranged in order, with indices $i_1, ..., i_n$. The problem then reduces to the selection of m, the number of centers to be included in the model. The way we do the ordering is somewhat arbitrary. One reasonable approach, given say r indices $i_1, ..., i_r$, would be to select i_{r+1} to maximize the increase in log likelihood when indices $i_1, ..., i_{r+1}$ are included compared with the log likelihood when indices $i_1, ..., i_r$ are included. Something like this approach was actually used for the ozone example given in section 3.3.3, but not for the much larger climate data example of section 3.3.4, where the approach taken was just to select the order of indices to achieve roughly even geographical dispersion.

A more detailed analysis is possible for the choice of m, the number of centers to include. One approach is via the maximized log likelihoods, with either a sequence of likelihood ratio tests or some automatic model selection criteerion such as AIC. An alternative, possibly superior if the ultimate objective of the analysis is prediction at sites off the network, is to use cross-validation. However, this is also very computationally intensive if applied in the usual way of leaving out one station at a time and refitting the model based on the remaining stations. Instead, an approach will be described based on leaving out one quarter of the observations at each stage.

Suppose we write a typical data vector in the form

$$Z = \begin{pmatrix} Z^{(1)} \\ Z^{(2)} \end{pmatrix}$$

where $Z^{(1)}$ represents the approximately one quarter of the observations omitted, and $Z^{(2)}$ the remainder. We partition both the fitted correlation matrix Σ and the sample correlation matrix $\hat{\Sigma}$ in the obvious way,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix}.$$

The model is re-fitted using just the $Z^{(2)}$ components, and used to predict $Z^{(1)}$. Since we are not attempting to model the station means, we assume without loss of generality that they are all 0. The optimal predictor of $Z^{(1)}$ is then $\hat{Z}^{(1)} = AZ^{(2)}$, where $A = \Sigma_{12}\Sigma_{22}^{-1}$. The mean squared prediction error is given by

$$\begin{aligned} &\mathrm{E}\left\{(Z^{(1)} - \hat{Z}^{(1)})^T(Z^{(1)} - \hat{Z}^{(1)})\right\} \\ &= \mathrm{E}\left[\ \mathrm{tr}\ \left\{Z^{(1)}Z^{(1)T} - 2AZ^{(2)}Z^{(1)T} + AZ^{(2)}Z^{(2)T}A^T\right\}\right]. \end{aligned} \tag{3.47}$$

If we average (3.47) over the N data points, a sample-based estimate becomes

$$\text{tr } \left\{ \hat{\Sigma}_{11} - 2\Sigma_{12}\Sigma_{22}^{-1}\hat{\Sigma}_{21} + \Sigma_{12}\Sigma_{22}^{-1}\hat{\Sigma}_{22}\Sigma_{22}^{-1}\Sigma_{21} \right\}. \tag{3.48}$$

This calculation is repeated four times, with a different quarter of the stations omitted on each occasion. Finally, the four cross-validation scores obtained from (3.48) are added, to obtain an overall CV score for the model.

3.3.3 An example based on ozone data

This example uses the same data source as in the paper of Nychka and Saltzman (1998); the precise data set used here was originally compiled by Andrew Royle. Twenty-one monitoring stations from the greater Chicago area are shown in Fig. 3.1, together with the locations of the cities of Chicago and Gary, Indiana. The data consisted of a sample covariance matrix based on 89 vector observations, which we assume to be independent. As previously discussed, the analysis will be simplified by ignoring any variation in the station variances and focussing exclusively on the sample correlation matrix.

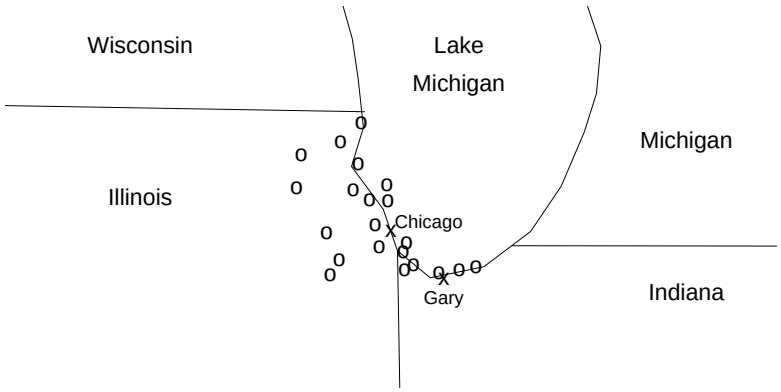


Fig. 3.1. Map of ozone stations.

The model described in section 3.3.2 treats the station locations as centers in a radial basis function representation, and relies heavily on the selection of a suitable subset of stations. The first step involved placing the posisble centers in order, which involved a certain amount of trial and error, but with the intention that centers which make a large contribution to the model are introduced early in the analysis. The rest of the discussion will take this step for granted and concentrate on the determination of m, the number of centers to be included in the model.

Number of Centers	NLLH	CV	Number of Centers	NLLH	CV
0	598.7	3.34	13	772.3	4.01
4	649.3	3.35	14	777.9	4.33
5	672.3	3.47	15	782.2	5.22
6	689.2	3.49	16	793.0	7.68
7	701.3	3.53	17	802.0	6.44
8	745.3	3.33	18	805.6	3.92
9	753.7	3.44	19	813.6	7.10
10	754.3	3.44	20	813.9	5.32
11	765.7	3.66	21	815.8	4.33
12	772.1	3.74			

Table 3.1: Minimum L values and CV scores for a sequence of models, ozone data.

Table 3.1 lists the 19 models, starting with $m = 3$ (for which all the $\delta_i^{(1)}$ and $\delta_i^{(2)}$ coefficients in (3.44) are 0) and proceeding up to the full model $m = 21$. The table shows both the minimum negative log likelihood (NLLH) values and the CV scores. All the models are based on the Matérn form of covariance (3.37), with $2m - 2$ independent parameters. It can be seen that a likelihood ratio test or AIC approach would lead to a large value of m being chosen, such as $m = 19$. However, the approach based on minimizing the CV score leads to $m = 8$. This only just beats the models with $m = 3$, a linear transformation from G to D space!

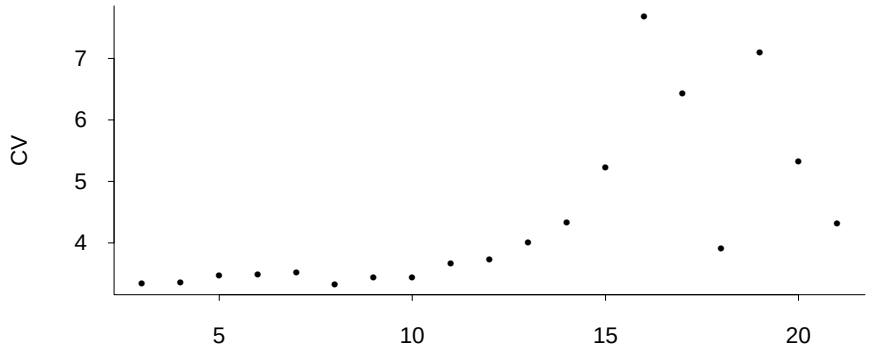


Fig. 3.2. Cross-validations scores for the ozone data.

A plot of the CV scores is shown in Fig. 3.2. We might expect these to decrease steadily for the first few values of m, to reach a minimum, and then to increase again. This is far from the observed form of the plot, a situation which may be due to the irregular spatial distribution of the stations combined with the strong nonlinearity of the optimization problem. However, it is clear that the CV scores in the right-hand part of the plot (say, for $m > 12$) are substantially larger than those in the left-hand part, which indicates that we should not allow m to be too large. The subsequent discussion is based on $m = 8$.

The D space under this transformation is shown in Fig. 3.3. The striking feature of this plot is that the three stations in the lower right-hand corner of Fig. 3.1 have been pulled a considerable distance from the other 18, which reflects the fact that the spatial correlations between the two groups, although still positive, are much smaller than those within the larger group. The explanation may lie in the distribution of sources. Ozone in Chicago itself, and in the suburbs to the north and west, is caused primarily by traffic, whereas there are a number of industrial sources in the neighborhood of Gary, IN, where the three discrepant monitors are located.

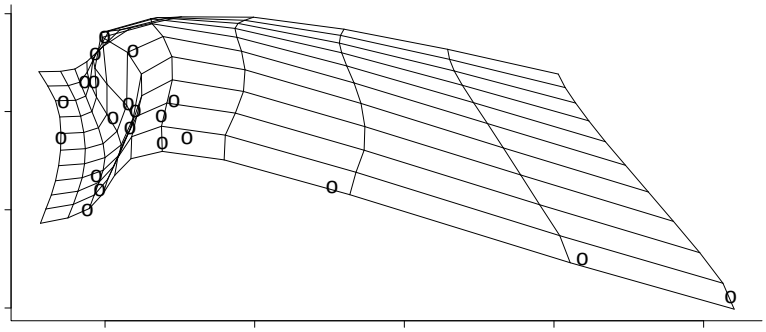


Fig. 3.3. D-space for the ozone data.

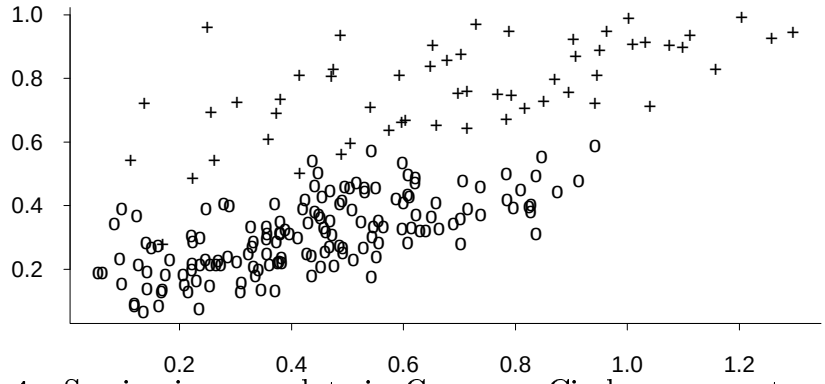


Fig. 3.4. Semivariogram plots in G-space. Circles represent pairs of points which lie entirely in the main cluster of 18 points; crosses represent pairs where at least one of the pair is one of the three outlying points.

Fig. 3.4 shows the semivariogram plot in G space, and Fig. 3.5 the same plot in D space, together with the fitted Matérn curve. The variability of the plot is much reduced by transforming to D space and fits the Matérn curve reasonable well. Each pair of stations is represented by one point on the plot, those that involve one of the three discrepant stations being marked by a cross, the remainder by a circle. The effect of the transformation is to move the crosses to the right of the picture and the circles to the left, showing that the plot consists of two distinct clusters.

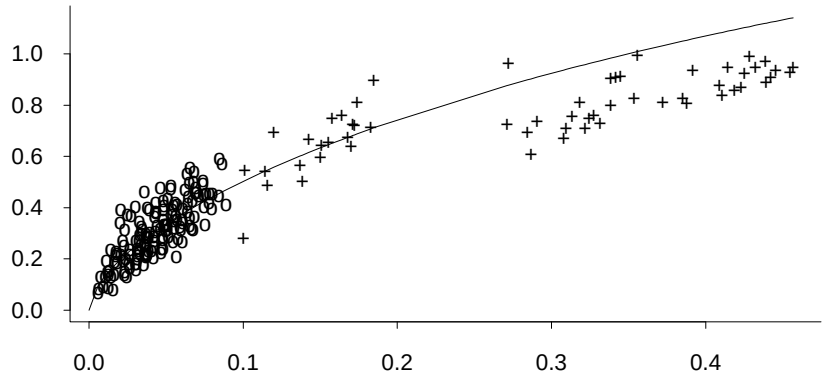


Fig. 3.5. Semivariogram plot in D space, corresponding to Fig. 3.4 after applying the transformation.

On the strength of these results, it appears that the methodology leads to estimators of spatial covariance which take into account the discrepant behavior of ozone at the three southeasternmost stations.

3.3.4. An example using climatic data

For our second example, we consider temperature measurements from 138 stations scattered over the continental United States. The stations form part of the Historical Climatological Network (HCN) discussed in Chapter 1. As ultimate objectives of the study, we might consider how to measure climate trends taking into account that these are not the same everywhere — it may be necessary to consider a model in which the climate trend varies smoothly with spatial location, but the accurate estimation of such a model would then require that one should take into account spatial correlations as well as spatial variability in means and trends. However with that objective in mind, we now focus exclusively on the spatial correlations and do not consider further the consequences of this for climate change. A previous example in which spatial analysis has been used to inform decisions about climate change is the paper by Handcock and Wallis (1994).

From this data set, a correlation matrix was constructed based on 40 years of annual average temperatures at each of the 138 stations. The 40 years were selected as those for which a reasonably complete record was available at all 138 stations. This correlation matrix will then be treated as a sample correlation matrix on the assumption that observations from different years form independent, identically distributed random vectors.

Obviously this appraoch will fail to take into account the effect of both temporal correlations between years, and long-term temperature trends such as might arise from global climate change, but our aim in the present study is to uncover spatial structure in the data rather than to produce a definitive analysis taking into account all aspects of variability.

For this data set, the same kinds of models were fitted by the same methods as for the ozone data. A key issue is again the order $i_1, i_2, ...$, in which the possible centers of the radial basis functions are introduced into the model (cf. (3.42)) and in this case there is even less scope to determine an optimal ordering. Not only are the combinatorial problems of subset selection much greater with 138 stations than with 21, but the time taken to compute each value of the log likelihood (which includes factorizing a 138×138 matrix) is much greater, making the whole procedure extremely computationally intensive. For this reason, a single ordering of the centers was determined prior to any model fitting, mainly chosen so that at each stage of the model fitting process, the centers in the model provide reasonable geographical coverage over the whole region being studied.

Based on this, and employing a log likelihood criterion for selecting the number m of centers included in the model, an initial model selection was made with $m = 21$. For this data set the Matérn covariance function again provided a reasonable fit, but it was found essential to include a parameter representing the "nugget effect".

Fig. 3.6 shows plots of the sample semivariogram in both G and D space, with the fitted Matérn curve for the latter. Although the transformation from G to D space unquestionably improves the fit as measured by the log likelihood, it must be admitted that there is not much evidence from this in Fig. 3.6, especially when we contrast with this with the very noticable improvement seen with the ozone data between Figs. 3.4 and 3.5! Maps of the G and D space are shown in Fig. 3.7 and it is evident that the main effect of the transformation is to pull a group of stations in the southwestern states (California, Nevada, Arizona) away from the rest of the country.

However there is also evidence in Fig. 3.6 that the semivariogram is *decreasing* at very large distances. The Bessel function representation (3.46) allows for this possibility, and we therefore use it in subsequent analysis. After some further experimentation a Bessel model with four components was fitted, with results shown in Fig. 3.8. The map of the D space is similar to that in Fig. 3.7, but the semivariogram plot now shows evidence of two distinct clusters of points, with the semivariogram flat or decreasing in the right-hand cluster.

In discussing these results with a climatologist colleague (Professor Peter J. Robinson of the Department of Geography, University of North Carolina at Chapel Hill) it was suggested that there might be a climatological explanation based on the patterns of air circulation over the continent. There is a tendency for weather patterns to move northwards up the west coast of the USA, then eastwards over the northern Rockies, and then to fan out over the rest of the country. This might well induce a negative correlation between the region southwest of the Rockies and the rest of the country. However, it was also pointed

out that this pattern of air circulation is much more prevalent during the summer months than the winter.

This suggested a re-analysis based on sample correlations computed separately for the summer (June, July, August) and winter (December, January, February) data, with results shown in Figs. 3.9 and 3.10 respectively. The distortion of the map created by the southwestern states is indeed much greater in the summer than the winter, and the evidence for the semivariogram to be decreasing at large lags is also much greater for the summer than for the winter. These results therefore reinforce the climatological explanation.

In fact, examination of raw sample correlations shows a much higher proportion of negative correlations (usually involving the three southwestern states) than could be explained by chance variation on the assumption that the true correlations are always non-negative. This however points to a limitation of the whole approach taken in this section. If it were indeed the case that in one part of the country, spatial correlations are negative over certain distance ranges, while in another part, correlations are always non-negative regardless of the distance between two stations, then we could not expect the model (3.33) to capture this adequately. To do that, we would need to model which explicitly allowed for the homogeneous semivariogram function γ_0 to be different in different parts of the space.

The other feature of this example that should be made clear is that, for models and data sets of the size being considered here, the question of uniqueness of local maxima of the log likelihood function is something that definitely needs to be considered. Indeed, all the evidence is that the local maxima are not unique, since re-runs of the same model based on different starting values typically result in different estimates of the coefficients of the radial basis functions. Thus in this case the concerns originally raised by Warnes and Ripley (1987) are seen to be valid. In most cases, when two fits of the same model produce different answers, the log likelihood values are very close, and the resulting maps and semivariogram plots also very similar in appearance. Nevertheless, the algorithm sometime stops at parameter values which are clearly a long way from optimality. Therefore as a practical measure, it is recommended that the same model be re-run from several different starting values before accepting any of the model fits as definitive.

Summarizing, the results for the climatological data are considerably more complicated than those for the ozone data. Nevertheless the fitted models provide considerable insight into the true spatial structure of the data, a statement which is reinforced by the interpretation in terms of streamflow patterns.

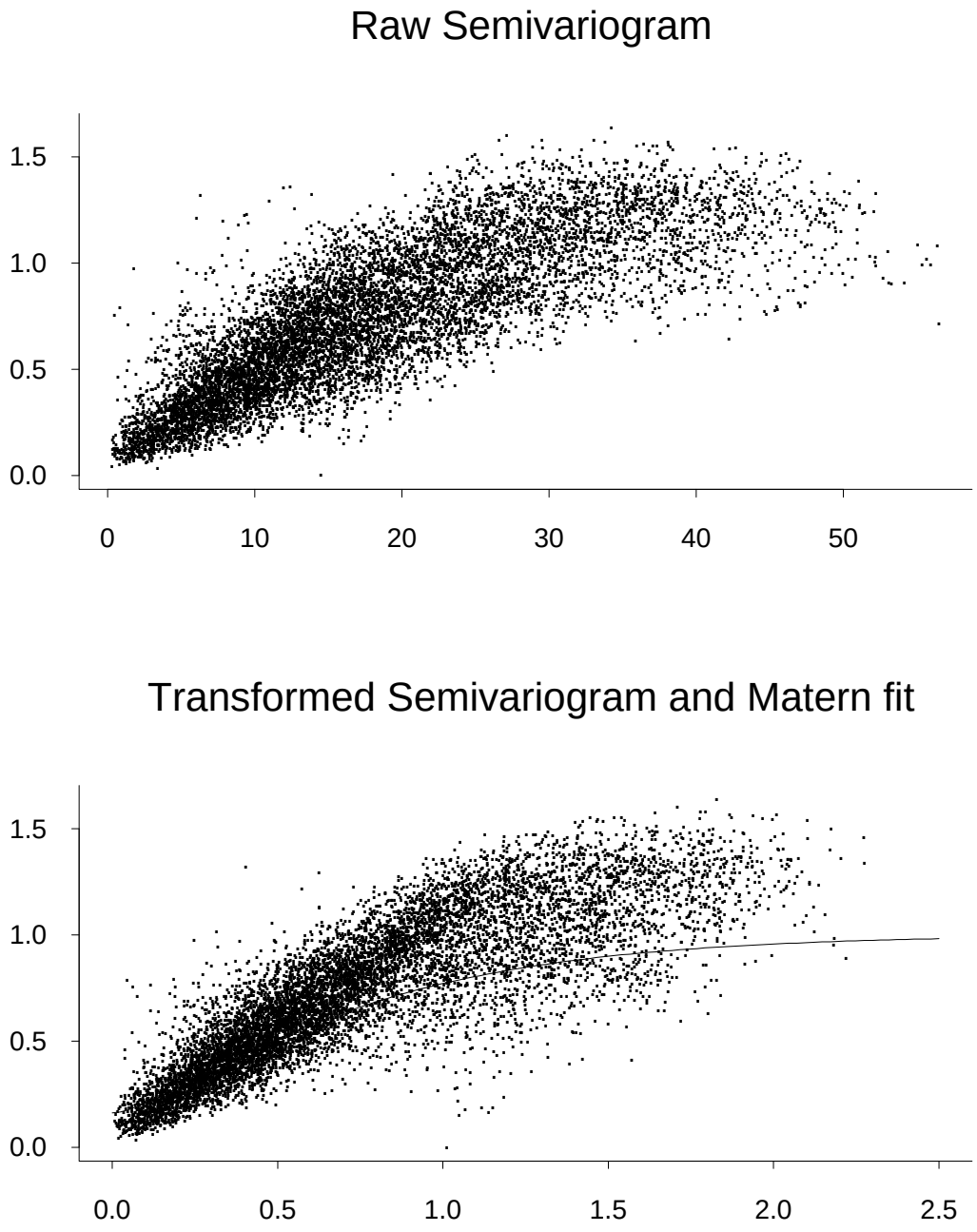


Fig. 3.6. Semivariogram plots for the climatic data, before and after transforming from G space to D space, with fitted Matérn curve in the latter case.

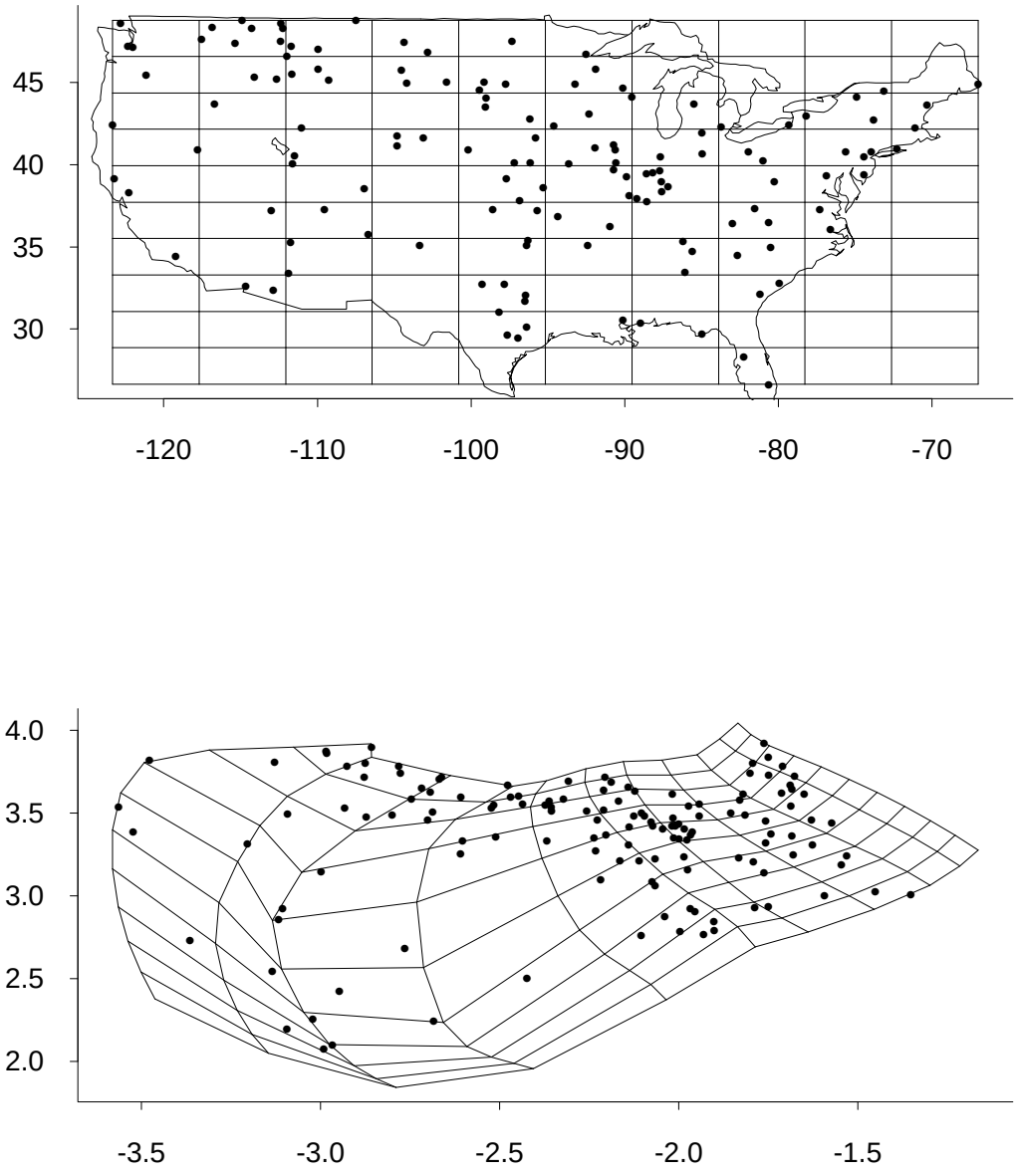


Fig. 3.7. G-space and D-space for the climatic example: all-year data and Matérn covariance.

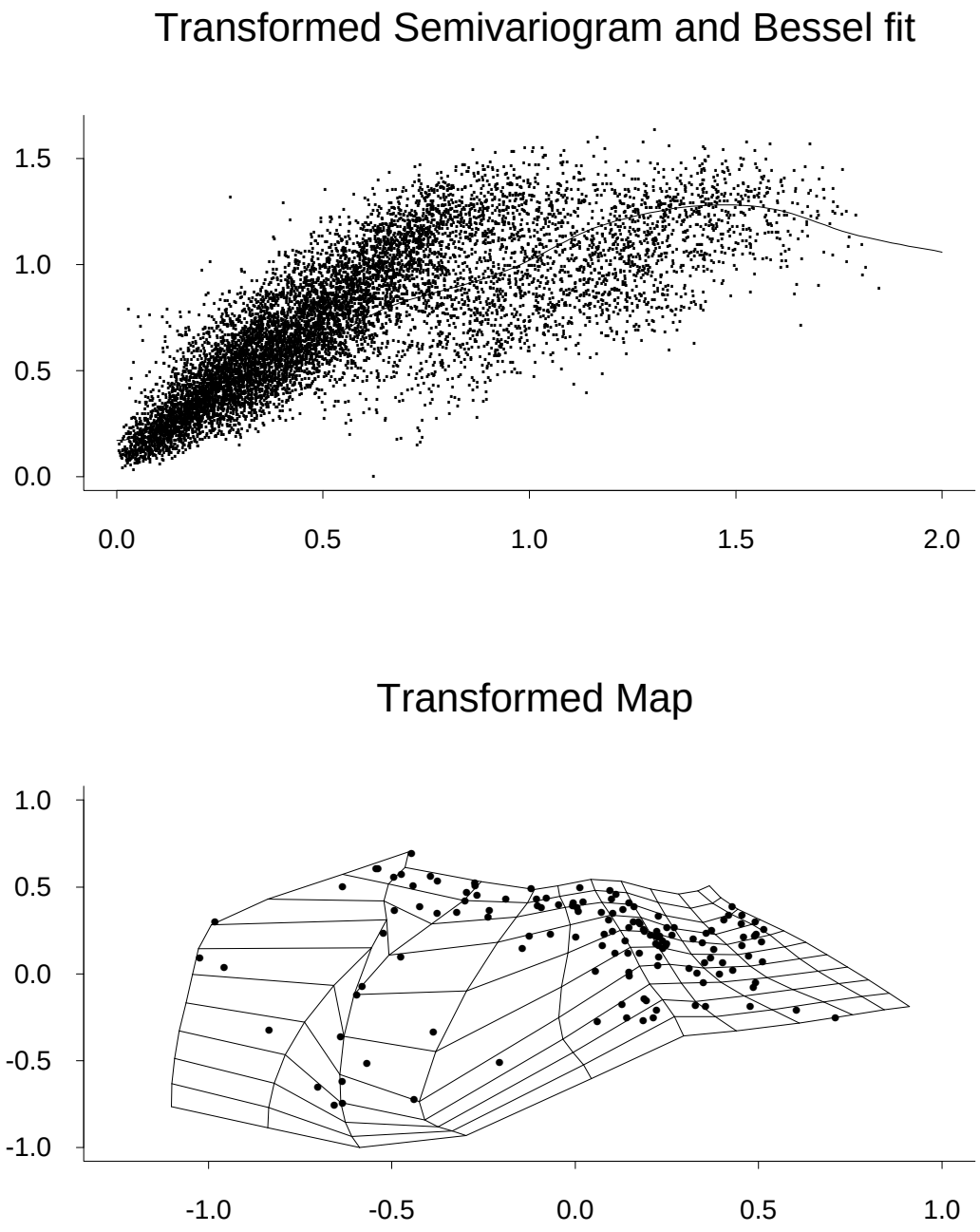


Fig. 3.8. Semivariogram in D space and representation of D space using Bessel function fit to the covariance function: all-year data.

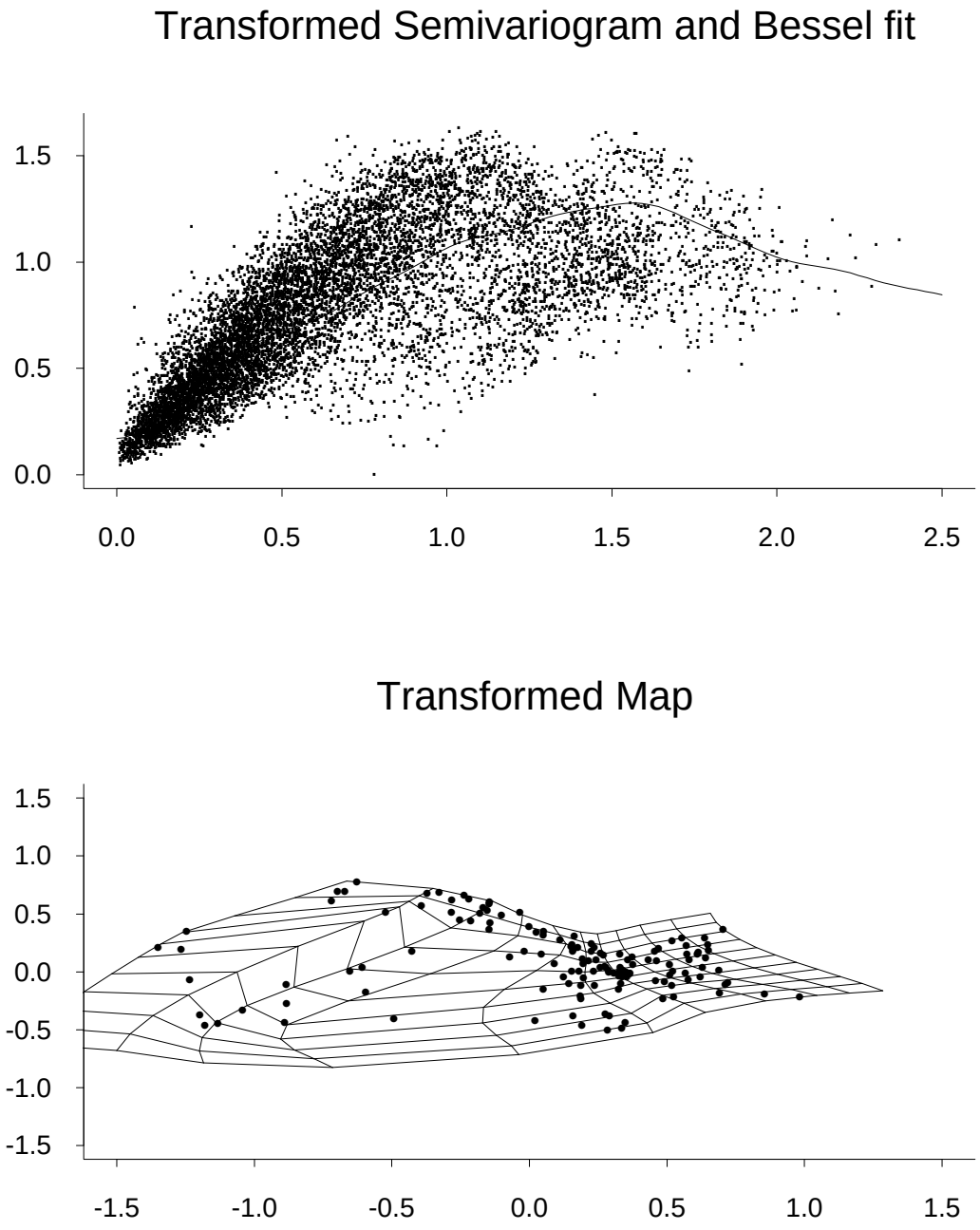


Fig. 3.9. Semivariogram in D space and representation of D space using Bessel function fit to the covariance function: summer data.

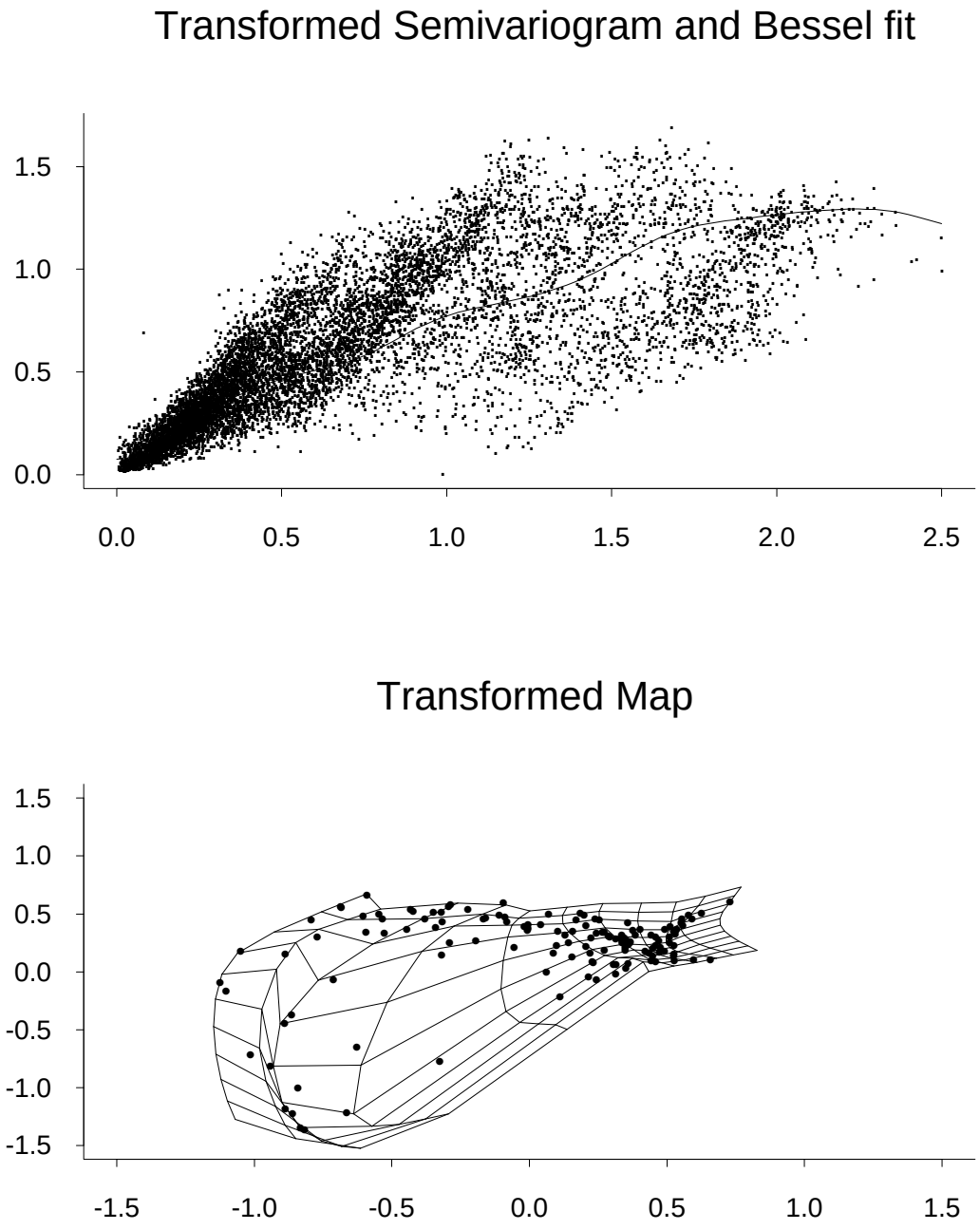


Fig. 3.10. Semivariogram in D space and representation of D space using Bessel function fit to the covariance function: winter data.

3.3.5 Bayesian approaches

A natural development of maximum likelihood methods is to take a Bayesian approach, since this also requires computing a likelihood function and in the context of a very high-dimensional data problem, sampling from the posterior using a Monte Carlo approach may not be very much more computationally demanding than trying to find the maximum of a complicated and possibly multi-model likelihood function. Bayesian methods have the advantage that when used in prediction, they correctly allow for the uncertainty in the parameters being estimated, and also, as a result of the well-known "shrinkage" effect that occurs when estimating high-dimensional parameters, they may also perform better simply as estimators of the spatial structure. Therefore, there are natural reasons to consider a Bayesian approach to the kinds of models defined in this section. Some of the disadvantages are the obvious computational complexity of the problem, and the difficulty of specifying a suitable parametric model, with associated prior distributions, in a rational way without appearing to make some totally arbitrary decision. Two recent papers have attempted to do this, however, taking quite different approaches to the problem, and we shall review both of these briefly here.

Damien, Sampson and Guttorp (2001) considered a model of the form

$$Z(x,t) = \mu(x,t) + \nu(x)^{1/2} E_\tau(x) + E_\epsilon(x,t), \tag{3.49}$$

in which $\mu(x,t)$ is a deterministic mean function (actually taken to be constant in time, so it is really $\mu(x)$), $\nu(x)$ is a location-dependent variance function (actually constant in space in their examples, so $\nu(x) \equiv \nu$, though this is not essential for their formulation), E_τ is some smooth spatial process and $E_\epsilon(x,t)$ represents local measurement error (independent for each (x,t) combination). Thus the model as specified incorporates no temporal correlation at all, though it could easily be extended to do so.

With this specification, the covariances at the same time point are of the form

$$\mathrm{Cov}\{Z(x,t), Z(y,t)\} = \begin{cases} \{\nu(x)\nu(y)\}^{1/2} \mathrm{Corr}\{E_\tau(x), E_\tau(y)\}, & x \neq y, \\ \nu(y) + \sigma_\epsilon^2, \text{x=y}, \end{cases}$$

As elsewhere in this section, we assume a deformation stucture for the spatial correlations, which leads to

$$\sigma_{ij} = (v_i v_j)^{1/2} \rho_\theta(||\xi_i - \xi_j||), \tag{3.50}$$

where ξ_i is the transformation of x_i in the D space and ρ_θ is some homogeneous spatial correlation function indexed by parameter θ. After estimating a sample mean the likelihood function is of the form

$$L(S|\Sigma) = |2\pi\Sigma|^{-(T-1)/2} \exp\left\{-\frac{T}{2}\mathrm{tr}(\Sigma^{-1}S)\right\}, \tag{3.51}$$

where Σ is the modeled covariance matrix and S is the sample covariance matrix.

One possible approach would be to parametrize the $x_i \to \xi_i$ transformation using radial basis functions or some other "basis function" approach, as in section 3.3.2, in order the reduce the problem to a parametric model for which standard MCMC methods are easily applicable. Damian *et al.* did not do that but instead treated the $\{\xi_i\}$ directly as unknown random quantities, with a prior distribution designed to penalize configurations with high bending energy. The precise formula for bending energy was of the form

$$\xi^{(1)^T} K_1 \xi^{(1)} + \xi^{(2)^T} K_2 \xi^{(2)},$$

where $\xi^{(1)}$ is the vector of first coordinates of the $\{\xi_i\}$, $\xi^{(2)}$ similarly for the second coordinates, and K_1 and K_2 are matrices whose entries are determined by the x_i values. Damian *et al.* cited earlier references including Bookstein (1991) and Sampson *et al.* (1991) for the precise details of this construction. The "prior" on $\Xi = \{\xi_i,\ i = 1, ..., n\}$ is then taken to be of form

$$\pi(\Xi) \propto \exp\left(-\frac{\xi^{(1)^T} K_1 \xi^{(1)} + \xi^{(2)^T} K_2 \xi^{(2)}}{2\tau^2}\right). \tag{3.52}$$

For the remaining parameters of the model, Damain *et al.* took a conventional approach of assigning vague priors. In the specific formulation that they adopted in their examples, the vector of standard deviations ν_i was replaced by a single parameter ν, and the ρ_θ model in (3.50) was taken to be the standard exponential decay function, $\rho_\theta(d) = \exp(-\theta d)$, so there is only a one-dimensional parameter θ to worry about. It seems clear that these specifications could easily be extended to include alternative forms of covariance structure or more complicated representations for the mean and variance functions of the process.

With this specification of the model, the full joint distribution of all the random quantities involved is obtained by combining (3.51), (3.52) with the prior $\pi(\nu, \theta)$ say for ν and θ, to give a combined likelihood $\times$ prior function of the form

$$L\{S|\Sigma(\Xi, \nu, \theta)\}\pi(\Xi)\pi(\nu, \theta), \tag{3.53}$$

where the notation is chosen to indicate that the modeled covariance matrix Σ is an explicit function of the D-space locations Ξ and the parameters ν and θ.

Generating a posterior sample from the density proportional to (3.53) essentially involves alternately sampling Ξ given (ν, θ) and sampling (ν, θ) given Ξ. The actual sampling strategy follows the by now standard Hastings-Metropolis method of generating a trial value for the parameters being updated followed by an accept-reject step. The authors remarked that in the trial sampling step for ξ_i, it is beneficial to impose a spatial structure on the increments, so that neighboring points have a tendency to move together rather than being resampled independently.

One rather unclear feature of this method is the treatment of the parameter τ in (3.52). This parameter controls the amount of smoothness in the deformation and is therefore analogous to the smoothing parameter λ in (3.34) or the number of basis functions m in

(3.43). In their examples, Damian *et al.* simply set $\tau = 1$ without specifically justifying that value. Presumably cross-validation is also a possibility in this case, if it can be implemented computationally, but it would also be worth exploring a hierarchical structure whereby τ itself is treated as a random parameter, with a presumably widely dispersed prior distribution, and estimated through a third level of Monte Carlo sampling. This would be consistent with the approach often taken towards similar problems in Bayesian hierarchical models.

An alternative viewpoint of essentially the same set of problems has been taken by Schmidt and O'Hagan (2000). Like Damian *et al.* (2001), they restricted themselves to a fairly simple version of the problem without temporal correlations and assuming spatial covariances of the form (3.50). The likelihood function is again given by (3.51) though with the small change in their case that the determinant term $|\Sigma|^{-(T-1)/2}$ is replaced by $|\Sigma|^{-T/2}$. For the covariance function ρ_θ, they assumed a finite mixture of Gaussian terms

$$\rho_\theta(h) = \sum_{k=1}^{K} a_k \exp(-b_k h^2), \tag{3.54}$$

where $a_k > 0$, $\sum a_k = 1$ and $b_1 > b_2 > ... > b_K > 0$, similar to the original paper of Sampson and Guttorp (1992). As already remarked in section 3.3.2, one could in principle obtain a more general representation through expansions of the form of (3.46).

For the prior distribution on the set of configurations Ξ — where we adopt the same notation as Damian *et al.* (2001) for ease of comparison of the two papers — if we write the D-space coordinates $\Xi = (\xi_1 \quad \xi_2 \quad \dots \quad \xi_n)$ as a $2 \times n$ matrix and G-space coordinates similarly as X, then Schmidt and O'Hagan take the prior distribution of Ξ to be multivariate normal with mean X and a covariance matrix of form $V \otimes R_d$ with V a 2×2 matrix representing the dependence between the two coordinates, and R_d and $n \times n$ matrix representing correlations between points. For R_d they took an exponential correlation function of form

$$R_d(x, x') = \exp(-b_d ||x - x'||^2),$$

with b_d again a smoothing parameter (analogous to τ in (3.52)) which controls the amount of bending. Like Damian *et al.*, Schmidt and O'Hagan did not attempt any formal inference about b_d but suggested it could be set equal to $\frac{1}{2a}$, where a is a typical squared distance in G space. For the 2×2 matrix V, they remarked that identifiability considerations allow one to restrict attention to diagonal matrices, and therefore proposed a prior distribution in which the diagonal elements independently have inverse gamma priors. The resulting prior distribution on Ξ is proper and this avoids certain problems of identifiability which remain in the specified model (e.g. if all the values of Ξ are multiplied by a constant and the parameter b_k in (3.54) adjusted accordingly, we get exactly the same model covariance function Σ). Under a Bayesian analysis, the posterior weights given to statistically equivalent configurations will be the in the same ratio as the prior weights, but provided the overall prior distribution is proper, this will not cause any fundamental problems.

For the priors on the parameters a_k and b_k in (3.54), Schmidt and O'Hagan assumed uniform distribution of $a_1, ..., a_K$ over the simplex $a_k/ge0$, $\sum a_k = 1$, and independent log-normal priors for b_k.

One difficulty that they note with the model is that it does not incorporate a nugget effect. Within the framework of (3.54), a nugget could be defined by letting $b_1 \to \infty$. However, as it stands, this would be inconsistent with the assumed log-normal prior for all the b_k. To get around this difficulty, Schmidt and O'Hagan proposed setting $b_1 = \infty$, while retaining log-normal priors for $b_2, ..., b_K$.

Another possible extension of the model (noted at the end of their paper) is to allow K also to be a variable parameter with some prior distribution. In this case, the number of parameters of the model is not fixed *a priori*, and to fit the model by MCMC sampling, one needs an extension of MCMC known as the reversible jump sampler (Green 1995). Conceptually, this should not be difficulty, though Schmidt and O'Hagan do not consider it in their paper.

In the remainder of their paper, Schmidt and O'Hagan describe in considerable detail an MCMC algorithm for efficient updating of the posterior distribution, a procedure to obtain predictive distributions for observations at unobserved locations, and some simulated and real-data examples. We shall not attempt to describe any of these features in detail, but it is worth remarking that, as with Damian *et al.* (2001), the sampling of points in D space is not taken independently but constrained so that there is a tendency for nearby points to move together. In the method of Schmidt and O'Hagan, this is achieved through a principal components decomposition of the sample covariance matrix S, which they interpret as ensuring that groups of sites which are highly correlated in G space tend to move together in D space.

To conclude this section, we note some comparisons and contrasts between the methods of Damian *et al.* (2001) and Schmidt and O'Hagan (2000), as well as the maximum likelihood approach of Smith (1996). Maximum likelihood estimators do not have good statistical properties in very high dimensions and this is essentially the reason why Smith (1996) did not take the full configuration matrix Ξ as a vector of unknown parameters, preferring the parameterize the deformation f in terms of a finite number of radial basis functions. In the present writer's view, this is still the simplest approach conceptually, though there is a lot of arbitrariness in the selection of a radial basis function, the location of the centers, and the number of centers included. The practical justification for the method relies on the assumption that the first two decisions — choice of basis function and locations of centers — do not have a huge sensitivity on the results provided "reasonable" choices are made in each case, while the third decision, how many centers to include, has the same intepretation of a smoothness parameter as arises somewhere in every model of this type. The two Bayesian approaches we have reviewed do not rely on low-dimensional parametrizations of Ξ, but impose some structure in the prior distribution. In both cases, the prior must impose some smoothness constraint on the deformation and the way this is done still appears to be somewhat arbitrary. Apart from the smoothness parameter, the

overall form of the prior is quite different for the two approaches, with that of Damian *et al.* much more explicity tied to the traditional "bending energy" formulations of spline theory. Finally, although the paper by Schmidt and O'Hagan appears to go further in developing a detailed and efficient computational algorithm, the ability of either method to handle large data sets would still appear to be untested at the present time.

3.4 The Le-Zidek approach

An alternative Bayesian approach was introduced by Le and Zidek (1992) and Caselton, Kan and Zidek (1992), and has been further developed in a series of papers, e.g. Brown, Le and Zidek (1994), Le, Sun and Zidek (1997), Zidek, Sun and Le (2000). It differs from the Bayesian approach of section 2.4 by allowing a more comprehensive prior structure on the covariances, in particular, exploiting the well-known Wishart conjugate prior for the inverse of an unknown covariance matrix. Other developments involving Wishart priors were due to Loader and Switzer (1992), Monestiez and Switzer (1991) and Mardia and Goodall (1993), but the Le-Zidek development is the most general and comprehensive, so we concentrate on that here.

Le and Zidek (1992) pointed out two disadvantages of classical kriging: the dependence on parametric, isotropic assumptions about a covariance matrix Σ, and the fact that the kriging step proceeds under the assumption that Σ is known; this nearly always means that kriging variances are underestimated because they do not account for uncertainty in Σ.

Previous Bayesian approaches had been tried, for example Omre (1987) computed a Bayesian interpolator, and Omre and Halvorsen (1989) extended this to include linear trend with unknown coefficients. The resulting model was identical to an approach due to Fedorov and Mueller (1988, 1989) (see Chapter 6), though the Fedorov-Mueller approach was non-Bayesian. Omre, Halvorsen and Bertig (1989) elaborated on the earlier Omre-Halvorsen approach to illustrate two competing linear structures. However, in this approach the final estimation of the spatial covariance matrix is based on an empirical Bayesian step, with resulting estimates treated as if they were known a priori. It is therefore not clear whether this answers the traditional criticism of kriging, i.e. that uncertain components treated as if known.

Another approach due to Loader and Switzer (1992) used Bayesian arguments to motivate an interpolation procedure, though their approach is not fundamentally Bayesian. They did consider anisotropic models. However they did not explicitly determine the posterior distribution of unobservables, and their approach was limited to producing predictions at a single site, not considering the covariance of predictions at different sites.

The new approach introduced by Le and Zidek (1992) aimed to avoid these objections by putting everything in a fully Bayesian context. However, as we shall see, it still requires

some parametric modeling of the Wishart matrix Ψ, and therefore does not entirely answer the objections that they raised to earlier approaches.

In this section, we first give some relevant mathematical background on multivariate and matrix-valued random variables and a review of Bayesian multiple regression theory, and then we develop the Le-Zidek formulae in detail.

3.4.1 Review of multivariate distribution theory

Although the basic distribution theory is standard (e.g. Johnson and Kotz (1972), Mardia, Kent and Bibby (1979), Press (1982), Anderson (1984)), the results are not often all collected in one place, so we do this for reference here.

We begin with the multivariate normal distribution: the p-dimensional random vector X has a multivariate normal distribution with mean μ and nonsingular covariance matrix Σ (notation: $X \sim N_p(\mu, \Sigma)$) if its density at $X = x$ is

$$(2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}. \tag{3.55}$$

For the purpose of the discussion here, we do not need to consider the case of singular Σ, when the multivariate normal distribution is still well-defined but does not have a density.

A convenient generalizion of this which we shall call the *matrix normal distribution* (this does not seem to be standard terminology, though it is a natural concept) applies when $X = (x_{ij})$ and $M = (\mu_{ij})$ are both $(p \times q)$-dimensional matrices and the covariance matrix Σ is of the form

$$\mathrm{Cov}\{x_{ij}, x_{k\ell}\} = w_{ik}v_{j\ell}, \tag{3.56}$$

with $W = (w_{ik})$ a $p \times p$ matrix and $V = (v_{j\ell})$ a $q \times q$ matrix. In that case we write

$$X \sim N_{pq}(M, W \otimes V),$$

the $\otimes$ notation meaning the Kronecker product of W and V, i.e. the matrix whose entries are given by (3.56). Some elementary facts about Kronecker products include

$$\begin{aligned} |W \otimes V| &= |W|^q|V|^p, \\ (W \otimes V)^{-1} &= W^{-1} \otimes V^{-1}, \end{aligned} \tag{3.57}$$

and we shall also write $W^{-1} = G = (g_{ik})$, $V^{-1} = F = (f_{j\ell})$. To extend (3.55) to apply to the matrix normal distribution, we note that the quadratic form in the exponential of (3.55) may be written as

$$\sum_i \sum_j \sum_k \sum_\ell (x_{ij} - \mu_{ij})(x_{k\ell} - \mu_{k\ell})g_{ik}f_{j\ell} = \mathrm{tr}\{G(X-\mu)F(X-\mu)^T\}, \tag{3.58}$$

where $\mathrm{tr}(A)$ denotes the trace of the matrix A, so the density is

$$(2\pi)^{-pq/2}|G|^{q/2}|F|^{p/2}\exp\left[-\frac{1}{2}\mathrm{tr}\{G(X-\mu)F(X-\mu)^T\}\right]. \tag{3.59}$$

A side benefit of the representation (3.59) is that likelihood calculations involving a $(p\times q)$-dimensional matrix normal distribution require the calculation of determinant and inverse for a $p\times p$ and a $q\times q$ matrix, but not directly for a $(pq)\times(pq)$ matrix, a considerable computational saving if p and q are large. This is relevant, for example, for the calculation of the likelihood function in separable spatial-temporal processes (Chapter 5).

We now turn to the *Wishart distribution*: for integer m and p, the random matrix D has a Wishart distribution repesented by

$$D\sim W_p(A,m). \tag{3.60}$$

if D may be represented as $D=\sum_{j=1}^m Z_jZ_j^T$ when $Z_1,...,Z_m$ are independent $N_p(0,A)$. The density is proper when $m>p-1$, and is given by

$$\frac{c_{p,m}|D|^{(m-p-1)/2}}{|A|^{m/2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(DA^{-1})\right\} \tag{3.61}$$

where $c_{p,m}$ is the constant

$$c_{p,m}=\left\{2^{mp/2}\pi^{p(p-1)/4}\prod_{j=1}^p\Gamma\left(\frac{m+1-j}{2}\right)\right\}^{-1}. \tag{3.62}$$

Johnson and Kotz (1972) also write this as

$$c_{p,m}=\left\{2^{mp/2}\Gamma_p\left(\frac{m}{2}\right)\right\}^{-1}$$

where

$$\Gamma_p\left(\frac{m}{2}\right)=\pi^{p(p-1)/4}\prod_{j=1}^p\Gamma\left(\frac{m+1-j}{2}\right). \tag{3.63}$$

In Bayesian statistics, the Wishart distribution is often taken as a conjugate prior for the inverse of a multivariate normal covariance matrix, and this is sometimes distinguished by the terminology *inverse Wishart distribution*: if (3.60) holds and $C=D^{-1}$, $B=A^{-1}$, then we write

$$C\sim W_p^{-1}(B,m), \tag{3.64}$$

with density proportional to $|C|^{-(m+p+1)/2}\exp\{-\frac{1}{2}\mathrm{tr}(C^{-1}B)\}$.

Another useful concept is the *matrix t* distribution (Dickey (1967), Johnson and Kotz (1972)), which is the distribution of $T=J^{-1}X$ when $X\sim N_{pq}(0,I_p\otimes Q)$, $JJ^T\sim$

$W_p(P, m+p-1)$ independent of X (P and Q are respectively positive definite $p \times p$ and $q \times q$ matrices; I_p is the $p \times p$ identity matrix), and which has density given equivalently by

$$\pi^{-pq/2}\Gamma_q\left(\frac{m+p+q-1}{2}\right)\left\{\Gamma_q\left(\frac{m+q-1}{2}\right)\right\}^{-1} \cdot \\ \cdot |Q|^{(p+q-1)/2}|P|^{q/2}|Q+T^TPT|^{-(m+p+q-1)/2} \tag{3.65}$$

or

$$\pi^{-pq/2}\Gamma_q\left(\frac{m+p+q-1}{2}\right)\left\{\Gamma_p\left(\frac{m+p-1}{2}\right)\right\}^{-1} \cdot \\ \cdot |P|^{-(m+p-1)/2}|Q|^{-p/2}|P^{-1}+TQ^{-1}T^T|^{-(m+p+q-1)/2}. \tag{3.66}$$

We write $T \sim t(p, q; P, Q, m)$.

A key result to the derivations of Caselton, Kan and Zidek (1992) and Le and Zidek (1992) is the following decomposition: suppose $C \sim W_p^{-1}(B, m)$, where C and B are decomposed as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}, \tag{3.67}$$

where C_{11} and B_{11} are $a \times a$, C_{12} and B_{12} are $a \times b$, etc. ($p = a + b$). Define

$$\begin{aligned} C_{1|2} &= C_{11} - C_{12}C_{22}^{-1}C_{21}, \\ \tau &= C_{12}C_{22}^{-1}, \\ B_{1|2} &= B_{11} - B_{12}B_{22}^{-1}B_{21}, \\ \eta &= B_{12}B_{22}^{-1}. \end{aligned} \tag{3.68}$$

Then

$$C_{22} \sim W_b^{-1}(B_{22}, m-a), \tag{3.69}$$

$$C_{1|2} \sim W_a^{-1}(B_{1|2}, m), \tag{3.70}$$

$$\tau|C_{1|2} \sim N_{ab}(\eta, C_{1|2} \otimes B_{22}^{-1}), \tag{3.71}$$

where, also, C_{22} is independent of $(C_{1|2}, \tau)$.

Since the proofs of (3.69)–(3.71) are not readily accessible, in the remainder of this subsection we outline them. This material may be omitted by the reader not interested in the details.

We use the following standard results:

Result 1 (inverse of a partitioned matrix; Mardia, Kent and Bibby (1979), section A.2.4):

If A and $B = A^{-1}$ are each partitioned as in (3.67), then

$$\begin{aligned} B_{11} &= (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} = A_{1|2}^{-1}, \\ B_{12} &= -B_{11}A_{12}A_{22}^{-1}, \\ B_{21} &= -A_{22}^{-1}A_{21}B_{11}, \\ B_{22} &= (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} = A_{2|1}^{-1}, \end{aligned}$$

Equivalently,

$$B_{12} = -A_{11}^{-1}A_{12}B_{22}, \quad B_{21} = -B_{22}A_{21}A_{11}^{-1}.$$

Result 2 (Mardia, Kent and Bibby (1979, Theorem 3.4.6):

Suppose $D \sim W_p(A, m)$, $m > a$, where D and A are partitioned as in (3.67), then

$$D_{2|1} \sim W_b(A_{2|1}, m - a), \tag{3.72}$$

and $D_{2|1}$ is independent of (D_{11}, D_{12}).

Proof of (3.69):

Writing $A = B^{-1}$, $D = C^{-1}$, we have $D \sim W_p(A, m)$. By Result 2, (3.72) holds. However, Result 1 shows that $A_{2|1} = B_{22}^{-1}$, $D_{2|1} = C_{22}^{-1}$. With these substitutions, (3.72) is equivalent to (3.69).

Proof of (3.70):

We have $D_{11} \equiv \sum_{j=1}^{m} Z_j^{(1)} Z_j^{(1)^T}$ where $Z_j^{(1)}$ consists of the first a entries of the vector $Z_j \sim N(0, A)$. Therefore $Z_j^{(1)} \sim N(0, A_{11})$ and $D_{11} \sim W_a(A_{11}, m)$. But $D_{11}^{-1} = C_{1|2}$, $A_{11}^{-1} = B_{1|2}$, hence the result.

Proof of (3.71):

Writing

$$\begin{aligned} \tau = C_{12}C_{22}^{-1} &= -D_{11}^{-1}D_{12}, \\ B_{12}B_{22}^{-1} &= -A_{11}^{-1}A_{12}, \\ B_{22}^{-1} &= A_{2|1}, \\ C_{1|2} &= D_{11}^{-1}, \end{aligned}$$

the result reduces to showing that

$$D_{11}^{-1}D_{12}|D_{11} \sim N_{ab}(A_{11}^{-1}A_{12}, D_{11}^{-1} \otimes A_{2|1}). \tag{3.73}$$

Writing $Z_j^{(2)}$ for the last b entries of Z_j, the conditional distribution of $Z_j^{(2)}$ given $Z_j^{(1)}$ has mean $A_{21}A_{11}^{-1}Z_j^{(1)}$ and covariance matrix $A_{2|1} = A_{22} - A_{21}A_{11}^{-1}A_{12}$ (recall section 2.4), so we can write

$$Z_j^{(2)} = A_{21}A_{11}^{-1}Z_j^{(1)} + Z_j^{(3)},$$

where $Z_j^{(3)}$ has distribution $N(0, A_{2|1})$ independently of $Z_j^{(1)}$. Then

$$\begin{aligned} D_{12} &= \sum_{j=1}^{m} Z_j^{(1)} Z_j^{(2)^T} \\ &= \sum_{j=1}^{m} Z_j^{(1)} Z_j^{(1)^T} A_{11}^{-1}A_{12} + \sum_{j=1}^{m} Z_j^{(1)} Z_j^{(3)^T} \\ &= D_{11}A_{11}^{-1}A_{12} + \sum_{j=1}^{m} Z_j^{(1)} Z_j^{(3)^T}. \end{aligned} \tag{3.74}$$

Consider the distribution of the second term in (3.74), where we first condition on the entire sequence $Z_j^{(1)}$, $1 \le j \le m$. For a single j, $Z_j^{(1)} Z_j^{(3)^T}$ is a $p \times p$ matrix which is the Kronecker product of two p-dimensional vectors; therefore, its distribution, conditional on $Z_j^{(1)}$, is normal with mean 0 and covariance matrix $\left(Z_j^{(1)} Z_j^{(1)^T} \right) \otimes A_{2|1}$. Summing over all j, the conditional distribution of $\sum_{j=1}^m Z_j^{(1)} Z_j^{(3)^T}$ given $Z_j^{(1)}$, $1 \le j \le m$, is normal with mean 0 and covariance matrix $D_{11} \otimes A_{2|1}$. But this depends on $Z_j^{(1)}$, $1 \le j \le m$, only through D_{11}, therefore by an iterated expectation step, the conditional distribution given D_{11} is the same. Substituting in (3.74), the conditional distribution of D_{12} given D_{11} is

$$N_{ab}(D_{11}A_{11}^{-1}A_{12}, D_{11} \otimes A_{2|1}).$$

The result (3.73) follows immediately from this.

As a final step in the argument, the independence statement that follows (3.71) follows from the independence statement at the end of Result 2: we have that $D_{2|1}$ is independent of (D_{11}, D_{12}). But $D_{2|1} = C_{22}^{-1}$, $D_{11} = C_{1|2}^{-1}$, $D_{12} = -D_{11}C_{12}C_{22}^{-1} = -C_{1|2}^{-1}\tau$, so C_{22} is independent of $(C_{1|2}^{-1}, C_{1|2}^{-1}\tau)$, which is equivalent to what was asserted.

3.4.2 Bayesian inference for multivariate regression

This material is also standard, though the complete results are not often all collected togther, e.g. Press (1989) gives the derivations but in the slightly simpler case that the prior distibutions are vague. Therefore, we re-derive the results here in the general case. Other references are Lindley and Smith (1972), Box and Tiao (1973).

We assume $y_1, ..., y_n$ are independent with

$$y_j \sim N_p(Bx_j, \Sigma).$$

Here x_j is a $q \times 1$ vector of covariates and B is a $p \times q$ matrix of regression coefficients. The prior distribution on (B, Σ) is assumed to be of the form

$$\begin{aligned} \Sigma &\sim W_p^{-1}(\Psi, m), \\ B|\Sigma &\sim N_{pq}(B^0, \Sigma \otimes F^{-1}). \end{aligned} \tag{3.75}$$

Define $V = \Sigma^{-1}$. Combining (3.59) and (3.61), we see that the joint density of (B, V) is proportional to

$$|V|^{(m+q-p-1)/2} \exp\left[-\frac{1}{2}\text{tr}\left\{V((B-B^0)F(B-B^0)^T + \Psi)\right\}\right]. \tag{3.76}$$

The hyperparameters which must be specified to define this joint density are (m, Ψ, B^0, F). The vague prior formulation of Press (1989) corresponds to $m = -q, \Psi = 0, B^0 = 0, F = 0$.

The likelihood, i.e. the joint density of $y_1, ..., y_n$ given B and V, is proportional to

$$\begin{aligned} &|V|^{n/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(y_j - Bx_j)^T V(y_j - Bx_j)\right\} \\ =&|V|^{n/2} \exp\left[-\frac{1}{2}\text{tr}\{V(S_{yy} - BS_{xy} - S_{yx}B^T + BS_{xx}B^T)\}\right] \end{aligned} \tag{3.77}$$

where we define

$$\begin{aligned} S_{yy} &= \sum y_j y_j^T, \qquad S_{xy} = \sum x_j y_j^T, \\ S_{yx} &= \sum y_j x_j^T, \qquad S_{xx} = \sum x_j x_j^T. \end{aligned} \tag{3.78}$$

Combining (3.78) and (3.77), the posterior distribution of (B, V) given $Y = (y_1, ..., y_n)$ is proportional to

$$\begin{aligned} |V|^{(m+n+q-p-1)/2} \exp\Bigg[-\frac{1}{2}\text{tr}\Bigg\{V\Bigg(&S_{yy} - BS_{xy} - S_{yx}B^T + BS_{xx}B^T \\ &+ (B-B^0)F(B-B^0)^T + \Psi\Bigg)\Bigg\}\Bigg]. \end{aligned} \tag{3.79}$$

Let us rewrite the expression

$$S_{yy} - BS_{xy} - S_{yx}B^T + BS_{xx}B^T + (B-B^0)F(B-B^0)^T \tag{3.80}$$

in the form

$$(B-B^*)G(B-B^*)^T + H. \tag{3.81}$$

It is readily checked that (3.80) and (3.81) are the same if we define

$$\begin{aligned} G &= S_{xx} + F, \\ B^* &= (S_{yx} + B^0F)G^{-1}, \\ H &= S_{yy} - B^*S_{xy} - S_{yx}B^{*T} + B^*S_{xx}B^{*T} + (B^* - B^0)F(B^* - B^0)^T. \end{aligned} \tag{3.82}$$

Therefore, the joint density (3.79) is equivalent to

$$|V|^{(m+n+q-p-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left\{V((B-B^*)G(B-B^*)^T + H + \Psi)\right\}\right]. \tag{3.83}$$

Comparing (3.83) with (3.76), therefore, the posterior density amounts to updating the prior hyperparameters as follows:

$$\begin{aligned} m &\to m+n, \\ \Psi &\to \Psi + H, \\ B^0 &\to B^*, \\ F &\to G. \end{aligned} \tag{3.84}$$

In particular, the posterior density of Σ given Y is $W_p^{-1}(\Psi + H, m+n)$ and the posterior density of B given Y and Σ is $N_{pq}(B^*, \Sigma \otimes G^{-1})$.

We can also represent the marginal posterior of B in matrix-t form: if $J = \Sigma^{-1/2}$ then $J(B-B^*) \sim N_{pq}(0, I_p \otimes G^{-1})$ conditionally on J, so the distribution of $B - B^*$ is $t(p, q; (\Psi+H)^{-1}, G^{-1}, m+n+1)$. Using (3.66), the posterior density of B is proportional to

$$|\Psi + H + (B-B^*)G(B-B^*)^T|^{-(m+n+q)/2}. \tag{3.85}$$

The final piece of "standard theory" we review is the predictive distribution of a new observation, cf. Press (1982). Suppose a new p-dimensional observation y^* is to be taken at a covariate vector x^* such that

$$y^*|x^*, B, \Sigma \sim N_p(Bx^*, \Sigma). \tag{3.86}$$

Under the Bayesian framework, the conditional distribution of y^* given $y_1, ..., y_n$ is obtained by integrating out the distribution (3.86) with respect to the posterior distributions of B and Σ; this is also known as the predictive distribution. It may be derived as follows. Conditionally on Σ, we have $B \sim N_{pq}(B^*, \Sigma \otimes G^{-1})$, and then (3.86) implies

$$y^*|x^*, \Sigma, y_1, ..., y_n \sim N_p(B^*x^*, (1 + x^{*T}G^{-1}x^*)\Sigma).$$

Hence the Bayesian predictive mean of y^* is B^*x^*, and the marginal predictive distribution of $y^* - B^*x^*$ given $x^*, y_1, ..., y_n$ is $t(p, 1; (\Psi+H)^{-1}, 1 + x^{*T}G^{-1}x^*, m+n-p+1)$. By (3.65), the predictive density of y^* is proportional to

$$\left\{1 + x^{*T}G^{-1}x^* + (y^* - B^*x^*)^T(\Psi+H)^{-1}(y^* - B^*x^*)\right\}^{-(m+n+1)/2}. \tag{3.87}$$

3.4.3 Details of the Le-Zidek approach

The fundamental idea of the approach is to assume there are a set of $p = u + g$ locations at which we would like to know the value of the field being measured, but for practical measurement purposes, only the last g locations are "gauged" while the first u are "ungauged". The complete field is assumed to be represented by n independent p-dimensional vectors $y_j,\ j = 1, ..., n$, but only the last g components of each y_j are actually observed. The model for y_j is

$$y_j \sim N_p(Bx_j, \Sigma) \tag{3.88}$$

where x_j is a q-dimensional covariate vector, B is a $p \times q$ matrix of regression coefficients and Σ is a $p \times p$ residual covariance matrix.

The vector y_j and the matrices B and Σ may be partitioned into gauged and ungauged sites:

$$y_j = \begin{pmatrix} y_j^{(1)} \\ y_j^{(2)} \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

As in section 3.4.1 we write

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \qquad \tau = \Sigma_{12}\Sigma_{22}^{-1}.$$

Recall from section 2.4 that the conditional mean of $y_j^{(1)}$ given $y_j^{(2)}$ is $B_1 x_j + \tau(y_j^{(2)} - B_2 x_j)$, i.e. τ is the slope of the regression equation, and the conditional variance is $\Sigma_{1|2}$.

Note that we may also write Σ in the form

$$\Sigma = \begin{pmatrix} \Sigma_{1|2} + \tau\Sigma_{22}\tau^T & \tau\Sigma_{22} \\ \Sigma_{22}\tau^T & \Sigma_{22} \end{pmatrix}.$$

By analogy with the framework of section 3.4.2, we assume the following joint prior for B and Σ:

$$\begin{aligned} \Sigma &\sim W_p^{-1}(\Psi, m), \\ B|\Sigma &\sim N_{pq}(B^0, \Sigma \otimes F^{-1}). \end{aligned} \tag{3.89}$$

Using (3.69)–(3.71), the prior distribution for Σ may be equivalently rewritten:

$$\Sigma_{22} \sim W_g^{-1}(\Psi_{22}, m - u), \tag{3.90}$$

$$\Sigma_{1|2} \sim W_u^{-1}(\Psi_{1|2}, m), \tag{3.91}$$

$$\tau|\Sigma_{1|2} \sim N_{ug}(\eta, \Sigma_{1|2} \otimes \Psi_{22}^{-1}), \tag{3.92}$$

where $(\Psi_{22}, \Psi_{1|2}, \eta)$ represents the same decomposition of the prior covariance matrix Ψ as $(\Sigma_{22}, \Sigma_{1|2}, \tau)$ does of Σ; in particular $\eta = \Psi_{12}\Psi_{22}^{-1}$. Also Σ_{22} is *a priori* independent of $(\Sigma_{1|2}, \tau)$.

The observed data are given by

$$D = (y_j^{(2)}, z_j, \quad 1 \le j \le n)$$

being i.i.d. realizations of

$$y_j^{(2)}|B, \Sigma, x_j \sim N_g(B_2 x_j, \Sigma_{22}).$$

To generate the full posterior distribution of (B, Σ), we first note the following elementary fact about Bayesian statistics: if the model parameter θ factorizes as (θ_1, θ_2) with θ_1 and θ_2 *a priori* independent, and if the distribution of the observed data D depends only on θ_2, then the posterior distributions of θ_1 and θ_2 are also independent, the posterior for θ_1 being the same as the prior and the posterior for θ_2 the same as if we did not consider θ_1 at all. In obvious notation,

$$\frac{\pi_1(\theta_1)\pi_2(\theta_2)f(D|\theta_2)}{\int\int \pi_1(\theta_1)\pi_2(\theta_2)f(D|\theta_2)d\theta_1 d\theta_2} = \pi_1(\theta_1) \cdot \frac{\pi_2(\theta_2)f(D|\theta_2)}{\int \pi_2(\theta_2)f(D|\theta_2)d\theta_2}.$$

We apply this reasoning with $\theta_1 = (\Sigma_{1|2}, \tau, B_1 - B_1^0 - \tau(B_2 - B_2^0))$, $\theta_2 = (B_2, \Sigma_{22})$, noting that the distribution of the observed data is defined entirely in terms of the mean B_2 and the covariance matrix Σ_{22}. Also, $B_1 - B_1^0 - \tau(B_2 - B_2^0)$ is *a priori* independent of B_2 and has the prior distribution

$$\{B_1 - B_1^0 - \tau(B_2 - B_2^0)\}|\Sigma \sim N_{uq}(0, \Sigma_{1|2} \otimes F^{-1}). \tag{3.93}$$

Using the results of section 3.4.2, we have

$$\Sigma_{22}|D \sim W_g^{-1}(\Psi_{22} + H_{22}, m - u + n), \tag{3.94}$$

$$B_2|D, \Sigma_{22} \sim N_{gq}(B_2^*, \Sigma_{22} \otimes G^{-1}), \tag{3.95}$$

where

$$\begin{aligned} G &= S_{xx} + F, \\ B_2^* &= (S_{y_2 x} + B_2^0 F)G^{-1}, \\ H_{22} &= S_{y_2 y_2} - B_2^* S_{xy_2} - S_{y_2 x} B_2^{*T} + B_2^* S_{xx} B_2^{*T} + (B_2^* - B_2^0)F(B_2^* - B_2^0)^T, \end{aligned} \tag{3.96}$$

using obvious notation, e.g. B^0 and B^* are partitioned in the same way as B, $S_{y_2 x}$ is the sum of products matrix between $y_j^{(2)}$ and x_j, $1 \le j \le n$, and so on.

The equations (3.91)–(3.96) together define the complete posterior distribution of (B, Σ) given D. One consequence worth noting is the marginal posterior distribution of B_1 given D and Σ: combining (3.93) and (3.95), this is

$$B_1|D, \Sigma \sim N_{uq}(B_1^*, (\tau \Sigma_{22} \tau^T) \otimes G^{-1} + \Sigma_{1|2} \otimes F^{-1}), \tag{3.97}$$

where $B_1^* = B_1^0 + \tau(B_2^* - B_2^0)$.

The calculations may be extended to predictive distributions of future observations, as follows. Suppose a new observation y^* is to be taken corresponding to covariates x^*. We partition y^* into $y^{(1)^*}$ (first u coordinates) and $y^{(2)^*}$ (last g coordinates) and consider separately the predictive distributions of $y^{(2)^*}$ and $y^{(1)^*}$ given $y^{(2)^*}$. The predictive density of $y^{(2)^*}$ is given by (3.87) with the obvious changes of notation, i.e.

$$\{y^{(2)^*} - B_2^* x^*\}|x^*, D \sim t(g, 1; (\Psi_{22} + H_{22})^{-1}, 1 + x^{*T} G^{-1} x^*, m - u + n - g + 1). \quad (3.98)$$

The conditional distribution of $y^{(1)^*}$ given $y^{(2)^*}$ is normal with mean $B_1 x^* + \tau(y^{(2)^*} - B_2 x^*)$ and covariance matrix $\Sigma_{1|2}$. We also have from (3.93) that

$$\{B_1 x^* - \tau B_2 x^*\}|D, \tau, \Sigma_{1|2} \sim N_u[B_1^0 x^* - \tau B_2^0 x^*, (x^{*T} F^{-1} x^*)\Sigma_{1|2}].$$

Hence

$$y^{(1)^*}|y^{(2)^*}, D, \tau, \Sigma_{1|2} \sim N_u[B_1^0 x^* + \tau(y^{(2)^*} - B_2^0 x^*), (1 + x^{*T} F^{-1} x^*)\Sigma_{1|2}].$$

We also have that $\tau|\Sigma_{1|2}$ is given by (3.92), so after integrating out this conditional distribution,

$$\begin{aligned} y^{(1)^*}|y^{(2)^*}, D, \Sigma_{1|2} \sim N_u[&B_1^0 x^* + \eta(y^{(2)^*} - B_2^0 x^*), \\ &\{1 + x^{*T} F^{-1} x^* + (y^{(2)^*} - B_2^0 x^*)^T \Psi_{22}^{-1} (y^{(2)^*} - B_2^0 x^*)\}\Sigma_{1|2}]. \end{aligned}$$

Finally using (3.91) as well, we deduce

$$\begin{aligned} &\{y^{(1)^*} - B_1^0 x^* + \eta(y^{(2)^*} - B_2^0 x^*)\}|y^{(2)^*}, D \\ &\quad \sim t(u, 1; \Psi_{1|2}^{-1}, 1 + x^{*T} F^{-1} x^* + (y^{(2)^*} - B_2^0 x^*)^T \Psi_{22}^{-1} (y^{(2)^*} - B_2^0 x^*), m - u + 1). \end{aligned} \quad (3.99)$$

Equations (3.98) and (3.99) are equivalent (in different notations) to the main results (16) and (17) of Le and Zidek (1992).

It is important to understand the structure of this result. The predictive distribution of $y^{(2)^*}$ is just the standard result for a predictive distribution in a multivariate normal family, as in (3.87). On the other hand, the conditional distribution of $y^{(1)^*}$ given $y^{(2)^*}$ is derived entirely from the prior — although (3.99) has been derived as the conditional distribution given D, it is in fact independent of D. However, the marginal predictive distribution of $y^{(1)^*}$, which may be derived by combining (3.98) and (3.99) and then integrating out $y^{(2)^*}$, does depend on D through (3.98).

3.4.4 Hierarchical models

As recognized by Le and Zidek (1992) and developed further by Brown, Le and Zidek (1994a), the theory that has been presented in section 3.4.3 is only a part of what is required

to produce realistic spatial interpolations. The conditional distribution of the ungauged sites given the gauged sites is unaffected by the data, i.e. the posterior distribution is the same as the prior distribution, so we clearly need more structure on the prior parameters B^0 and especially Ψ to develop a workable method. However, this could be done through a hierarchical approach. For example, one possible model is to assume that there are no covariates, the overall mean is given by a common μ, and the Ψ matrix is based on a common correlation coefficient ρ between each pair of stations, i.e.

$$\Psi = \begin{cases} \sigma^2 & \text{if } i = j, \\ \rho\sigma^2 & \text{if } i \neq j, \end{cases} . \tag{3.100}$$

In this case, we can modify the structure of section 3.4.3 very slightly to assume $B^0 = (\mu, \mu, ..., \mu)^T$ (a $p \times 1$ vector dependent on a single parameter μ) where $\mu|\Sigma$ has a simple normal distribution $N(\mu_0, 1/f)$ with f an arbitrary scalar; the limit $f \to 0$ corresponds to an uninformative prior. They also refer to the possibility of a regional model in which the stations are classified into geographical regions and the parameters μ, σ^2 and ρ are different for each region, with spatial correlation 0 between different regions. Another possibility, not explicitly mentioned by Le and Zidek but seemingly natural for their approach, is to model Ψ by one of the standard geostatistical covariances, for example, $\Psi = (\psi_{ij})$ where

$$\Psi = \begin{cases} \sigma^2 & \text{if } i = j, \\ \alpha\sigma^2 e^{-d_{ij}/\beta} & \text{if } i \neq j, \end{cases} \tag{3.101}$$

where d_{ij} is the distance between stations i and j, β is the range parameter of the exponential variogram model, and $\alpha \in [0, 1]$ is a parameter whose value reflects the nugget effect ($\alpha = 1$ is no nugget effect).

The models may be characterized by defining a vector of hyperparameters θ where, for example, in (3.100), $\theta = (\sigma^2, \rho)$, and in (3.101), $\theta = (\sigma^2, \alpha, \beta)$ (or (σ^2, β) if the possibility of a nugget effect is ignored). Our model is then of the form

$$\begin{aligned} y_j^{(2)}|B, \Sigma, x_j &\sim N_g(B_2 x_j, \Sigma_{22}), \\ \Sigma|\Psi, m \sim W_p^{-1}(\Psi, m), \quad B|\Sigma, B^0, F &\sim N_{pq}(B^0, \Sigma \otimes F^{-1}), \end{aligned} \tag{3.102}$$

where the first equation in (3.102) represents the first stage of the hierarchchy, i.e. the distribution of the observed data in terms of B and Σ, while the second equation represented the second stage of the hierarchy, given the joint distribution of B and Σ. Implicit in the hyperparametric approach is that we represent all the unknowns of the second stage (B^0, Ψ, F and m) as functions of a set of hyperparameters θ.

There are then two ways we may proceed. One way is to add a third stage to the hierarchy in (3.102), specifying either a proper or improper hyperprior for θ and calculating a fully Bayesian posterior distribution using Markov chain Monte Carlo methods, as in many modern approaches to hierarchical models. A second approach, slightly simpler, is to estimate the parameters θ by an exact or approximate maximum likelihood scheme,

and then to treat the estimates as fixed values, in effect ignoring the uncertainty in θ. This is sometimes known as the "empirical Bayes" approach, or "type II maximum likelihood". Le and Zidek favored this approach because it is computationally simpler to implement and (they argued) the posterior distributions for B and Σ are generally not too sensitive to slight misspecifications of θ. In spite of their preference for this approach, it should perhaps be pointed out that the actual interpolation step (the conditional distributions at the ungauged sites given the gauged sites) is rather heavily dependent on the prior distribution and therefore may be affected by uncertainty in θ rather more than the predictive distributions at the gauged sites. For this reason, it may still be desirable to consider the fully Bayesian approach.

It should be pointed out that it is possible to integrate out the values of B_{22} and Σ_{22} in (3.102). Since the second row of (3.102) implies (using (3.90))

$$\Sigma_{22}|\Psi, m \sim W_g^{-1}(\Psi_{22}, m-u), \quad B_2|\Sigma, B^0, F \sim N_{gq}(B_2^0, \Sigma_{22} \otimes F^{-1}),$$

letting $Y^{(2)} = (y_1^{(2)}, ..., y_n^{(2)})$ denote the $g \times n$ observed data matrix and $X = (x_1, ..., x_n)$ the corresponding matrix of covariates, we have

$$Y^{(2)}|B_2, \Sigma_{22}, B^0, \Psi, F, m \sim N_{gp}[B_2 X, \Sigma_{22} \otimes I_n],$$

and hence

$$Y^{(2)}|\Sigma_{22}, B^0, \Psi, F, m \sim N_{gp}[B_2^0 X, \Sigma_{22} \otimes (I_n + X^T F^{-1} X)],$$

This is of the form

$$\Sigma_{22}^{-1/2}(Y^{(2)} - B_2^0 X)|\Sigma_{22}, B^0, \Psi, F, m \sim N_{gp}[0, I_p \otimes (I_n + X^T F^{-1} X)],$$

so in the definition of the matrix t distribution in section 3.4.1, we identify T with $Y^{(2)} - B_2^0 X$ and J with $\Sigma^{-1/2}$, and thereby deduce

$$Y^{(2)} - B_2^0 X|B^0, \Psi, F, m \sim t(g, n; \Psi_{22}^{-1}, I_n + X^T F^{-1} X, m-u-g+1),$$

so for example, using (3.65), the marginal density of $Y^{(2)}$ is proportional to

$$\begin{aligned} &\Gamma_n\left(\frac{m-u+n}{2}\right)\left\{\Gamma_n\left(\frac{m-u-g}{2}\right)\right\}^{-1} |\Psi_{22}|^{n/2} \cdot \\ &\quad \cdot |I_n + X^T F^{-1} X + (Y^{(2)} - B_2^0 X)^T \Psi_{22}^{-1}(Y_2 - B_2^0 X)|^{-(m-u+n)/2}. \end{aligned} \tag{3.103}$$

(We have retained the Γ_n terms in (3.103) because of the possibility that m is itself regarded as a hyperparameter, as it sometimes is.)

At this point, there would appear to be two ways of proceeding. One is to use (3.103) directly to define the likelihood of the observed data given the hyperparameters, using it either as the input to a Bayesian approach or else maximizing with respect to the

hyperparameters to obtain type II maximum likelihood estimators. This possibility is mentioned by Brown, Le and Zidek (1994a) but not pursued by them, apparently because of computational difficulties. Instead, following a paper by Chen (1979), they advocated an application of the EM algorithm (Dempster, Laird and Rubin 1977), in which the model is retained in the form of (3.102), and the algorithm alternates between maximizing the likelihood of θ given Σ_{22} (the M step), and computing the posterior mean of the sufficient statistic $(\Sigma_{22}^{-1}, \log|\Sigma_{22}|)$ given the current values of θ and the observed data (the E step). We omit the details of this, referring to section 4 of Brown, Le and Zidek (1994a).

Brown, Le and Zidek (1994a) also considered the possibility of multivariate observations, for example, several pollutants measured at each station. This approach requires no mathematical generalization of what has already been given, because for instance, if there are K measured variables at each of the p stations of the full set of (gauged and ungauged) stations, we simply repeat the same theory with Σ and Ψ now $(pK) \times (pK)$ matrices. However, specifying parametric models for such large matrices is likely to prove a difficult challenge in itself, and to simplify matters, they suggested a Kronecker product form for this:

$$\Psi = \Lambda \otimes \Omega, \tag{3.104}$$

with Λ a $p \times p$ matrix of intersite covariances and Ω a $K \times K$ matrix of covariances between the variables. This approach involves some oversimplification — for example, it is a consequence of (3.104) that the correlations between measuring stations are the same for all the pollutants. Brown, Le and Zidek argued that this is reasonable because the same air transport processes apply to all the pollutants. (However, this may not be true. For example, ozone is a product of photochemical reactions taking place over large spatial scales and time scales of several hours, whereas fine particulate matter tends to vary over much shorter time and space scales. It might be possible to extend (3.104) to a sum of Kronecker product matrices, each representing a different physical process.)

Le, Sun and Zidek (1997) made a further extension of the theory to *data missing by design*. This model is designed to cope with the situation, quite common in practice, that not all of the variables are measured at all of the measuring stations. They assumed that of the gK possible combinations of gauged sites and measured variables, some number L of them are deliberately missing from all the observations. One could, in principle, still model the prior covariance matrix through the Kronecker product model (3.104), but because of the missing variables, only a $(gK - L) \times (gK - L)$ submatrix of (3.104) would be modeled. At this point, there would appear to be two ways that one could proceed:

(i) Maximize the likelihood (3.103) or apply the EM algorithm but using only the $(gK - L) \times (gK - L)$ submatrix of Σ corresponding to the observed data,

(ii) Use the EM algorithm based on the full $(gK) \times (gK)$ matrix Σ, in which the E step is modified to take account of the missing data.

Le, Sun and Zidek (1997) considered only the second method and did not say why they rejected the first, but presumably, it was because if g and K are large, the M step would

be too computationally burdensome, requiring the determinant of a $(gK - L) \times (gK - L)$ matrix without any simplification along the lines of (3.57). Instead, they showed how one could apply the E step to $(\Sigma_{22}^{-1}, \log|\Sigma_{22}|)$ after taking account of the missing data — in effect, this is another application of the gauged and ungauged sites analysis of section 3.4.3, but with the role of the ungauged sites now taken by the L site-pollutant combinations for which data are unavailable. We refer to their paper for the details of this.

A limitation of this approach is that is appears to require that the same vector of observations is available at all time points — it would be nice if the theory also worked for *data missing at random*, but it would appear that it does not.

3.4.5 Discussion and applications

Brown, Le and Zidek (1994a) mention several reasons why their approach is preferable to traditional kriging:

• Traditional kriging ignores the uncertainty in the estimation of the covariance structure, but this approach incorporates that into the prior distributions. The use of multivariate and matrix-valued t distributions is analogous to the use of the t distribution in elementary statistics to allow for the uncertainty in σ^2.

• Typical geostatistical models are based on oversimplified parametric models assuming stationary and isotropic covariances. Bayesian methods allow parametric models to be updated based on the incoming data.

• The use of covariates allow for additional effects which may be either time or space dependent. For example, they would allow the use of a spatial trend $f(s)$ (where s is location) represented by some parametric function of s as in the papers of Fedorov and Mueller (1988, 1989).

• Traditional kriging considers only the interpolation of the field one location at a time, whereas this approach allows full multivariate posterior distributions to be given for all the ungauged locations. This is valuable, for instance, in the construction of simultaneous prediction intervals.

As an example, Brown, Le and Zidek (1994a) considered the interpolation of an air pollution field from seven stations in Ontario, where three variables were measured at each station: O_3 (ozone), SO_4 (sulfate) and NO_3 (nitrate). (In fact the three variables were not always measured at the same location, but they ignored this feature in that ppaer, unlike the later paper of Le, Sun and Zidek (1997)). They used monthly averaged data for 72 months beginning in January 1983, and took logarithms of the monthly averages as the basic variables used in the analysis. The time-dependent covariates they used were 1, t, $\cos(2\pi t/12)$ and $\sin(2\pi t/12)$, where t is time in months. Thus the model allows for a linear time trend and sinusoidal seasonal effect. They found that after adjusting for these effects,

the resulting data for SO_4 and NO_3 exhibited no seasonal variation or serial correlation — there was some residual correlation for O_3 but this was ignored.

In their analysis they had data from 7 stations but considered the possibly of predicting the air pollution field at a further 400 stations. If the $m > p$ condition is to be satisfied for a proper Wishart prior, we therefore require $m > 407$. The EM algorithm produced a point estimate of $m = 467$, which does not imply a great many degrees of freedom to spare. In practice, they performed further spatial smoothing by inserting the estimated Σ_{22} matrix into the interpolation algorithm of Sampson and Guttorp (1992), which allows for nonstationarity by incorporating a nonlinear deformation function into the analysis. This approach to nonstationary models is discussed in detail in Section 3.3. This allows the predictions of the model to be extended to the entire region of interest. They did point out, however, that incorporating the Sampson-Guttorp approach in this way deviates from a strictly Bayesian view of the method, because, for example, it takes no account of estimation errors in the Sampson-Guttorp method. Yet another possibility might be to use maximum likelihood or Bayesian versions of the Sampson-Guttorp method, which would allow for a strictly hierarchical approach to be retained, but at the cost of several further layers of computation.

Another application of the methodology was given by Sun, Zidek, Le and Özkaynak (2000). This paper was about the interpolation of the PM_{10} field from 10 monitoring stations in the vicinity of Vancouver, British Columbia. The measurements were based on the Tapered Element Oscillating Microbalanace (TEOM) instrument, which effectively measures the accumulation of particles on a filter, with readings typically being taken hourly. Sun *et al.* considered daily aggregate data for 1996, filling in missing values by interpolation. They considered the interpolation of the field to 299 additional locations. They found that an AR(1) model provided a good fit to the temporal dependence, with a common AR parameter of 0.34 being estimated for each of the 10 locations. Residuals from the AR(1) model were used for the spatial analysis. The time trend was represented as a sum of day-of-week and week-of-year parameters, and the Sampson-Guttorp (1992) method was again used to obtain a spatial interpolator allowing for nonstationarity in the underlying field. A final feature of their method was the use of cross-validation to assess the fit of the model. Each of the 10 stations was omitted in turn, and a sequence of predictions obtained at that station after refitting the model (including the Sampson-Guttorp step) to the other 9 stations. In this way, they were able to assess the true coverage probabilities of the estimated prediction intervals. Combining all the stations, they concluded that nominal 99%, 95% and 80% achieved respectively 96%, 91% and 78% true coverage probabilities. They suggested that the discrepancies between nominal and actual coverage proportions may be due more to extreme values among the PM_{10} measurements than to lack of fit of the spatial model.

They did highlight some difficulties in this approach. Taking residuals from a time series analysis and then doing a spatial analysis creates the possibility of a phenomenon which they called "spatial leakage", in which the spatial correlations of the residuals from the time series analysis may be different from those of the original data. In practice, for

the daily data, they computed both sets of spatial correlations and found them to be in good agreement, but they remarked that the spatial leakage problem was much worse with hourly data. An alternative approach might be to incorporate the temporal correlations directly into the analysis, for example by assuming the $g \times n$ observed data matrix Y has covariance matrix $\Sigma \otimes \Gamma$, with Σ the $g \times g$ spatial covariance matrix and Γ the $n \times n$ temporal covariance matrix. In spatial-temporal analysis this condition on a covariance matrix is known as separability (see Chapter 5), and is appropriate when the temporal correlation structure is the same at all sites, as appears to be the case here. This feature is easily incorporated into the analysis given earlier — for example, in (3.103), the matrix $I_n + X^T F^{-1} X$ should be replaced by $\Gamma + X^T F^{-1} X$, and in (3.78), the definitions of S_{yy}, etc. should be replaced by

$$\begin{aligned} S_{yy} &= Y\Gamma^{-1}Y^T, \qquad S_{xy} = X\Gamma^{-1}Y^T, \\ S_{yx} &= Y\Gamma^{-1}X^T, \qquad S_{xx} = X\Gamma^{-1}X^T. \end{aligned} \tag{3.105}$$

Of course, the predictive distributions of sections 3.4.2 and 3.4.3 would also have to be modified to account for the temporal correlations, and we do not give those details here. Sun *et al.* hinted that they considered such an approach, but remarked that it would require further *ad hoc* assumptions and their method appeared to work quite well for the application in question.

3.5 Kernel-based models

Recently an alternative approach has become popular, initially as an alternative representation for stationary processes, but easily extendable to nonstationary processes. Higdon (2001) has given a nice review of the concepts; other recent references include Higdon (1998), Higdon *et al.* (1999), Barry *et al.* (1996) Ver Hoef and Barry (1998), Ver Hoef *et al.* (2000), Fuentes (2001) and Fuentes and Smith (2001).

The initial idea for a stationary process was to write a given process as a convolution of white noise with a smoothing kernel:

$$z(s) = \int_{\mathcal{S}} K(u-s)w(u)du, \ s \in \mathcal{S}, \tag{3.106}$$

where $\mathcal{S}$ is some domain of observation (for example, the whole of two-dimensional space), $K(\cdot)$ is a smoothing kernel and $w(\cdot)$ is a white noise process, which will be defined in a moment.

The motivation for defining a spatial process as an integral of white noise can be said to go back to Whittle (1954), who gave a similar representation for discrete spatial processes. Specifically, Whittle considered processes $\{\xi_{s,t}\}$ defined (for integer s, t) by a relationship of form

$$\sum_{j,k} a_{j,k}\xi_{s-j,t-k} = \epsilon_{s,t}, \tag{3.107}$$

with $\{\epsilon_{s,t}\}$ uncorrelated random variables of mean 0 and variance 1. Equation (3.107) may be thought of as the spatial equivalent of an autoregressive process in time series analysis. Defining

$$L(z_1, z_2) = \sum_{j,k} a_{j,k} z_1^j z_2^k,$$

for complex variables $z_1,\ z_2$, Whittle demonstrated that if $L(z_1, z_2) \neq 0$ whenever $|z_1| = |z_2| = 1$, then we can rewrite (3.107) in the form

$$\xi_{j,k} = \sum_{j,k} b_{j,k} \epsilon_{s-j,t-k} \tag{3.108}$$

where $b_{j,k}$ is the coefficient of $z_1^j z_2^k$ in the expansion of $1/L(z_1, z_2)$. Equation (3.108) can be thought of as the discrete analog of (3.106).

In the continuous case (3.106), it is necessary to be a little careful over what we mean by the white noise process $w(u)$, since the process does not exist as a real function in the usual sense. The following definition will suffice for all our purposes. Assuming a two-dimensional process (an equivalent definition may easily be given for other dimensions), for any $\epsilon > 0$, let $u_{\epsilon,i,j} = (i\epsilon, j\epsilon)$ for $i = 0, \pm 1, \pm 2, ..., j = 0, \pm 1, \pm 2, ...$, and let $w_{\epsilon,i,j}$ be independent $N(0, \epsilon^2)$ random variables. For any measureable function $a(u)$ for which $\int_{\mathcal{S}} a^2(u)du < \infty$, define

$$\int_{\mathcal{S}} a(u)w(u)du = \lim_{\epsilon \to 0} \sum_{i,j:u_{\epsilon,i,j} \in \mathcal{S}} a(u_{\epsilon,i,j}) w_{\epsilon,i,j}, \tag{3.109}$$

where the limit in (3.109) may most easily be interpreted as convergence in distribution. Note that an immediate consequence of this definition is that $\int_{\mathcal{S}} a(u)w(u)du$ has mean 0 and variance

$$\lim_{\epsilon \to 0} \epsilon^2 \sum_{i,j:u_{\epsilon,i,j} \in \mathcal{S}} a^2(u_{\epsilon,i,j}) = \int_{\mathcal{S}} a^2(u)du,$$

and that $z(s)$ defined by (3.106) has mean 0 and covariance function

$$\begin{aligned}
C(h) &= \mathrm{E}\{z(s)z(s-h)\} \\
&= \mathrm{E}\left\{ \int_{u \in \mathcal{S}} \int_{u' \in \mathcal{S}} K(u-s)K(u'-s+h)w(u)w(u')dudu' \right\} \\
&= \lim_{\epsilon \to 0} \mathrm{E}\left\{ \sum_{(i,j:u_{\epsilon,i,j} \in \mathcal{S})} \sum_{(i',j':u_{\epsilon,i',j'} \in \mathcal{S})} K(u_{\epsilon,i,j} - s) w_{\epsilon,i,j} K(u_{\epsilon,i',j'} - s + h) w_{\epsilon,i',j'} \right\} \\
&= \lim_{\epsilon \to 0} \epsilon^2 \sum_{i,j:u_{\epsilon,i,j} \in \mathcal{S}} K(u_{\epsilon,i,j} - s) K(u_{\epsilon,i,j} - s + h) \\
&= \int_{\mathcal{S}} K(u-s)K(u-s+h)du \\
&= \int_{\mathcal{S}} K(u-h)K(u)du.
\end{aligned} \tag{3.110}$$

Thus the covariance function is stationary, and furthermore, is isotropic if $K(u)$ depends only on $||u||$. If either

(i) $\int K(s)ds < \infty$ and $\int K^2(s)ds < \infty$,

or

(ii) $C(h)$ is integrable and non-negative definite,

then there is a one-to-one relationship between $K(\cdot)$ and $C(\cdot)$, so the process may be equally well specified in terms of either. (In the non-isotropic case, the relationship is no longer one-to-one because $C(\cdot)$ does not uniquely determine $K(\cdot)$, but it is still legitimate to seek a kernel $K(\cdot)$ for which (3.110) is consistent with empirical covariances.)

Thiébaux and Pedder (1987, Chapter 5) gave examples of how (3.106) could be used to define a spatial process, and Higdon (2001) discussed the whole concept in detail. If $\mathcal{S} = \mathcal{R}^2$ and we define

$$\tilde{K}(\omega) = \int_{\mathcal{R}^2} e^{i\omega^T x} K(x)dx, \quad \tilde{C}(\omega) = \int_{\mathcal{R}^2} e^{i\omega^T x} C(x)dx,$$

then (3.110) is equivalent to

$$\tilde{C}(\omega) = \tilde{K}^2(\omega), \tag{3.111}$$

and this relationship may be used to calculate either of K or C from the other (first performing a Fourier transform, then applying (3.111), and then inverting the transform). Here are three examples of this procedure, the first of which is taken directly from Higdon (2001), the other two apparently new.

1. Suppose

$$K(u) = \exp\left(-\frac{||u||^2}{2\tau}\right),$$

for some $\tau > 0$. Then

$$\begin{aligned}
\tilde{K}(\omega) &= \int \exp\left(i\omega^T u - \frac{1}{2\tau}u^T u\right) du \\
&= \int \exp\left\{-\frac{\tau}{2}\omega^T\omega - \frac{1}{2\tau}(u - i\tau\omega)^T(u - i\tau\omega)\right\} du \\
&= 2\pi\tau \exp\left(-\frac{\tau}{2}\omega^T\omega\right).
\end{aligned}$$

Therefore if we define $\tau' = 2\tau$,

$$\begin{aligned}
\tilde{C}(\omega) &= 4\pi^2\tau^2 \exp(-\tau\omega^T\omega) \\
&= \pi\tau \cdot 2\pi\tau' \exp\left(-\frac{\tau'}{2}\omega^T\omega\right)
\end{aligned}$$

and hence

$$C(u) = \pi\tau \exp\left(-\frac{1}{2\tau'}||u||^2\right).$$

The specific example Higdon gave of this was for $\tau = 1$, when $C(u) \propto \exp(-||u||^2/4)$.

2. In a d-dimensional process let

$$C_{\phi,\alpha,\nu}(u) = \frac{\pi^{d/2}\phi}{2^{\nu-1}\Gamma(\nu + d/2)\alpha^{2\nu}}(\alpha||u||)^{\nu}K_{\nu}(\alpha||u||)$$

denote the Matérn covariance function, where here K_ν is a Bessel function. The Fourier transform of the covariance function, also known as the spectral density (Fuentes, 2001), is given by

$$\tilde{C}_{\phi,\alpha,\nu}(\omega) = \phi(\alpha^2 + ||\omega||^2)^{-\nu-d/2}. \tag{3.112}$$

Hence the corresponding kernel function is

$$\tilde{K}_{\phi,\alpha,\nu}(\omega) = \phi^{1/2}(\alpha^2 + ||\omega||^2)^{-\nu/2-d/4}$$

which is the same as (3.112) with (ϕ, α, ν) replaced by $(\phi^{1/2}, \alpha, \nu/2 - d/4)$. Hence

$$K_{\phi,\alpha,\nu}(u) = C_{\phi^{1/2},\alpha,\nu/2-d/4}(u).$$

Thus the kernel is itself a Matérn covariance function, but with different parameters. If we denote the parameters of the covariance function by $(\phi_C, \alpha_C, \nu_C)$ and the corresponding parameters of the kernel function by $(\phi_K, \alpha_K, \nu_K)$, the relationship is

$$\phi_K, \phi_C^{1/2}, \quad \alpha_K = \alpha_C, \quad \nu_K = \frac{\nu_C}{2} - \frac{d}{4}. \tag{3.113}$$

We require $\nu_C > 0$ for C to be a legitimate covariance function. However, the kernel K can be a well-defined positive function even if $\nu_K < 0$; therefore, (3.113) is valid for all $\nu_C > 0$.

3. In $d = 2$ let $K(u) = \frac{2}{\pi}(1 - ||u||^2)I(||u|| < 1)$. This is the Epanechnikov kernel, considered further below, where the constant $\frac{2}{\pi}$ ensures $\int K(u)du = 1$. Then $C(h) = 0$ unless $||h|| < 2$, so we write $||h|| = 2t$ where $0 \le t \le 1$. Define

$$\begin{aligned}
c_0 &= \frac{8}{15} - \frac{8t^2}{3}, \\
c_1 &= \frac{8t}{3}, \\
c_2 &= -\frac{16}{15} + \frac{8t^2}{3}, \\
c_3 &= -\frac{8t}{3}, \\
c_4 &= \frac{8}{15}.
\end{aligned} \tag{3.114}$$

Also let

$$B_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt,$$

cf. Abramowitz and Stegun (1964), equation (6.6.1). Then

$$C(h) = \frac{16}{\pi^2}\sum_{k=0}^{4} c_k B_{1-t^2}\left(\frac{3}{2}, \frac{k}{2}+\frac{1}{2}\right). \tag{3.115}$$

A derivation of (3.115) is given at the end of this section.

Higdon pointed out a number of features of this modeling approach which might make it desirable:

- It is possible to specify the process directly in terms of K and thus not have to worry about whether c is positive definite,
- The definition is extendable to non-normal $w(s)$, e.g. Wolpert and Ickstadt (1998), Ickstadt and Wolpert (1999),
- The domain $\mathcal{S}$ can be restricted for certain purposes, e.g. to model edge effects,
- Dimension reduction — the process $w(s)$ could be restricted to a small number of locations $s = s_1, ..., s_m$ to create a convenient parametric representation for the entire process z,
- Nonstationary covariances — the process (3.106) can be extended as

$$z(s) = \int_{\mathcal{S}} K_s(u)w(u)du, \ s \in \mathcal{S}, \tag{3.116}$$

  with a general space-dependent kernel K_s,

- Space-time models, e.g. (3.106) could be extended to

$$z(s,t) = \int_{\mathcal{S}} K(u-s)w(u,t)du, \ s \in \mathcal{S}, \tag{3.117}$$

  with t a time variable,

- Dependent spatial processes, e.g. a pair of form

$$\begin{aligned} z_1(s) &= \int_{\mathcal{S}} K_1(u-s)w(u)du, \\ z_2(s) &= \int_{\mathcal{S}} K_2(u-s)w(u)du, \end{aligned} \tag{3.118}$$

with different kernels $K_1,\ K_2$ but a common white noise process u.

Given that the emphasis of this chapter is on nonstationary models, for the rest of the present discussion we shall give particular attention to equation (3.116) and its various ramifications.

One very nice feature of the model (3.116) discovered by Higdon *et al.* (1999) is that when the kernel $K_s(u)$ is a Gaussian kernel for each s, the covariance function of the process is explicitly computable. The details are as follows. First, by a Gaussian kernel we mean that for each s, there is some 2×2 covariance matrix $\Sigma(s)$ such that

$$K_s(u) = \frac{1}{2\pi}|\Sigma(s)|^{-1/2} \exp\left(-\frac{1}{2}u^T \Sigma(s)^{-1} u\right).$$

If we parameterize

$$\Sigma(s) = \begin{pmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{pmatrix}, \quad \Sigma(s') = \begin{pmatrix} a'^2 & \rho' a' b' \\ cr\rho' a' b' & b'^2 \end{pmatrix}, \tag{3.119}$$

then Higdon *et al.* show

$$\rho(s, s') \propto \frac{1}{q_1} \exp\left\{-\frac{1}{q_2}(s - s')^T W (s - s')\right\}, \tag{3.120}$$

where

$$\begin{aligned} W &= \begin{pmatrix} b^2 + b'^2 & -(\rho ab + \rho' a' b') \\ -(\rho ab + \rho' a' b') & a^2 + a'^2 \end{pmatrix}, \\ q_1 &= 2\pi a a' b b' \sqrt{(1-\rho^2)(1-\rho'^2)} \sqrt{-\frac{(\rho^2-1)b^2 + (\rho'^2-1)b'^2}{(\rho^2-1)(\rho'^2-1)b^2 b'^2}} \cdot \\ &\quad \cdot \sqrt{\frac{2\rho\rho' a a' b b' + a^2((\rho^2-1)b^2 - b'^2) + a'^2((\rho'^2-1)b'^2 - b^2)}{a^2 a'^2((\rho^2-1)b^2 + (\rho'^2-1)b'^2)}}, \\ q_2 &= -2(2\rho\rho' a a' b b' + a^2((\rho^2-1)b^2 - b'^2) + a'^2((\rho'^2-1)b'^2 - b^2). \end{aligned} \tag{3.121}$$

It is perhaps worth pausing to consider a number of possible applications of (3.120) and (3.121). The attractive feature about this is that since there is an explicit formula for the covariance function of the process, it is possible to write down the full covariance matrix and hence the likelihood function for the process at any configuration of sampling points. To complete the specification of the model, one must determine how the parameters $a,\ b,\ \rho,\ a',\ b',\ \rho'$ in (3.119) depend on sampling points $s,\ s'$. One possibility would be to define the functions

$$\phi_1(s) = \log a(s), \quad \phi_2(s) = \log b(s), \quad \phi_3(s) = \log\left\{\frac{1+\rho(s)}{1-\rho(s)}\right\}, \tag{3.122}$$

and let $\phi_k(s),\ k = 1, 2, 3$, vary smoothly over space according to a thin-plate spline or an expansion in radial basis functions, similar to (3.39). The specific motivation for defining ϕ_k in this way is that (3.122) allows $-\infty < \phi_k < \infty$ and, conversely, is easily inverted so that the functions $a,\ b$ and ρ satisfy the needed constraints $a > 0,\ b > 0,\ -1 < \rho < 1$. One could use the radial basis function representation to represent each $\phi_k(s)$ as a parametric function and then fit by maximum likelihood, or alternatively, by analogy with (3.34), choose $\phi_k,\ k = 1, 2, 3$ to minimize the functional

$$\sum_{i,j} \left(\frac{r_{ij} - \rho_{ij}}{\rho_{ij}} \right)^2 + \sum_{k=1}^{3} \lambda_k J(\phi_k),$$

in which r_{ij} and ρ_{ij} are respectively the sample and population covariance (or correlation) between sites i and j, J is the bending energy functional and $\lambda_k,\ k = 1, 2, 3$ are prescribed smoothing constants. A disadvantage of this approach is that it is very closely tied in with the Gaussian form of covariance function, to which it reduces when the ϕ_k functions are constants, and this does not allow the flexibility of either the Matérn covariance function or of a general nonparametric representation such as (3.46); moreover, Stein (1999) has argued that the Gaussian covariance function has undesirable theoretical properties, and while it is not clear that Stein's arguments preclude the application of the Gaussian covariance function to practical problems involving small numbers of monitoring stations, it would perhaps be unwise to build too elaborate a theory tied specifically to this form of covariance. Another aspect of the theory is that one could easily extend it to deal with the case when the observed process is a finite sum of independent processes of the form (3.116), and this would be a nonstationary extension of the expansion of a stationary covariance function in terms of Gaussian covariances, cf. Sampson and Guttorp (1992), Schmidt and O'Hagan (2000).

The actual development of Higdon *et al.* (1999) proceeded along rather different lines:

1. Instead of a representation of the form of (3.122), they considered a reparameterization of (a, b, ρ) in terms of the focus points and an overall scaling parameter of the ellipse defined by $x^T \Sigma(s) x = 1$. The focus points and the scaling parameter were assumed to be realizations of another spatial stochastic process.

2. The whole structure was embedded in a Bayesian hierarchical model including additional random components for the mean effect and an additional measurement error variance. The model could then be fitted by MCMC techniques.

In more recent work, Higdon (2001) has considered models in which the integral in (3.106) is replaced by a sum over a finite number of white noise components. The method has obvious extensions to the nonstationary case (3.116). The direct analog of (3.106) in this case is the model

$$z(s) = \sum_{j=1}^{m} w_j K(s - u_j), \tag{3.123}$$

with $w_1, ..., w_m$ i.i.d. $N(0,1)$ and $s_1, ..., s_m$ supposed sampling locations for the w_j. The process (3.123) is assumed to be sampled at n locations $s = s_1, ..., s_n$, and embedded in a linear model

$$y = \mu 1_n + Kw + \epsilon, \tag{3.124}$$

where μ is a constant overall mean, 1_n is the n-vector of ones, K is an $n \times m$ kernel matrix with entries $K(s_i - w_j)$, $w \sim N_m(0, I_m)$, $\epsilon \sim N_n(0, \sigma_\epsilon^2 I_n)$. The overall covariance function of this process is of form $KK^T + \sigma_\epsilon^2 I_n$ which is of "mixed models" form, so the process can be fitted by well-established maximum likelihood or REML methods for this form of covariance matrix. Alternatively, as pointed out by Higdon, one could take a Bayesian approach.

Extensions considered by Higdon (2001) include:

(i) The model (3.124) may be extended to a multi-resolution process, of the form

$$y = \mu 1_n + \sum_{\ell=1}^{p} K_\ell w_\ell + \epsilon, \tag{3.125}$$

with p separate kernels and white noise processes superimposed,

(ii) spatial-temporal processes

$$y_t = \mu 1_n + K w_t + \epsilon_t, \tag{3.126}$$

in which t is time and the individual components of w_t, say $w_{j,t}$, $1 \le j \le m$, could be taken as independent in time or as the realization of some time series (independent for each j). Higdon considered the case in which each $w_{j,t}$ process was a random walk in time t.

Representing a nonstationary process as a kernel integral of stationary processes

An alternative representation discussed by Fuentes (2001) and Fuentes and Smith (2001) is to replace the white noise process in (3.106) with a stationary process, whose parameters may, however, themselves be allowed to vary over space, thus creating a non-stationary process. The basic representation formula is

$$z(s) = \int_{\mathcal{S}} K(s-u) z_{\theta(u)}(s) du, \tag{3.127}$$

where $z_{\theta(u)}(\cdot)$ is a stationary process (for example, a Matérn process) with parameters $\theta(u)$ possible varying according to location u.

Although it is possible, as will shortly be shown, to rewrite the model (3.127) as an integral of white noise and hence to apply similar techniques to those developed by Higdon, the actual modeling process is different, since Higdon treats the kernel K as a function

of unknown parameters to be estimated, whereas the intention behind (3.127) is to use a simple predetermined form of kernel and concentrate on the estimation of the process $z_{\theta(u)}(\cdot)$. If $\theta(u)$ varies slowly as a function of u, then the process is nearly stationary over small regions (exactly stationary if $\theta(u)$ is constant), but by allowing $\theta(u)$ to vary, one allows for the possibility that the locally stationary process will look quite different in different portions of the sampling region. Thus, the method is consistent with the moving windows approach of Haas, but by representing the whole process as a well-defined stochastic model, avoids the problem of positive definiteness over the whole sampling space, which is a difficulty in the Haas approach.

There are, however, two possible interpretations of (3.127), and it is worthwhile to clarify these before proceeding:

(1) we could write $z_{\theta(u)}(s)$ in terms of its spectral representation (cf. Yaglom (1987)),

$$z_{\theta(u)}(s) = \int e^{is^T x} \sqrt{f_{\theta(u)}(x)} w(x) dx, \tag{3.128}$$

in which f_θ denotes the (spatial) spectral density of the process z_θ and w is a white noise process on $\mathcal{R}^2$, or

(2) We could assume the process $z_{\theta(u)}(s)$ is a given spatial process in s for each u, but is *independent* for different u. In that case, one has the covariance function

$$\mathrm{Cov}\{z(s), z(s')\} = \int K(s-u) K(s'-u) C_{\theta(u)}(s-s') du \tag{3.129}$$

as given by Fuentes and Smith (2001), where $C_\theta(\cdot)$ is the covariance function of the process $z_\theta(\cdot)$. This function is well-defined for any (s, s'), and is guaranteed to be a positive definite covariance. Moreover, when $\theta(u)$ is constant it is indeed a stationary process, though in that case, the covariance function is not C_θ but rather $C_\theta C_K$, where C_K is defined by (3.110). In general, we are interested in cases where $\theta(u)$ is not constant and in this case the process is genuinely nonstationary.

The alternative model defined by (3.127) and (3.128) can be rewritten in the form

$$z(s) = \int K_s(x) w(x) dx, \tag{3.130}$$

where

$$K_s(x) = e^{is^T x} \int K(s-u) \sqrt{f_{\theta(u)}(x)} du, \tag{3.131}$$

and is therefore consistent with (3.116).

In this case, the covariance function of the process is

$$\mathrm{Cov}\{z(s), z(s')\} = \int K_s(x) \overline{K_{s'}(x)} dx \tag{3.132}$$

in which the overbar denotes complex conjugate (necessary here, because $K_{s'}(x)$ is a complex function). When $\theta(u)$ is a constant θ, (3.131) reduces to $K_s(x) = e^{is^T x}\sqrt{f_\theta(x)}$, and (3.132) to

$$\int e^{i(s-s')^T x} f_\theta(x) dx = C_\theta(s - s'), \tag{3.133}$$

since the left hand side of (3.133) is exactly the representation of the covariance function C_θ in terms of its spectral density (Yaglom 1987).

By analogy with the reduction of (3.127) and (3.128) to (3.130), we can attempt a similar reduction of the model which leads to (3.129), but the result in this case is a little different. Allowing for the process $z_{\theta(u)}(s)$ to be independent for each u, one can write the spectral representation (3.128) in the form

$$z_{\theta(u)}(s) = \int e^{is^T x}\sqrt{f_{\theta(u)}(x)} w_u(x) dx, \tag{3.134}$$

where $w_u(x)$ is an independent white noise process for each u. (3.127) and (3.134) combine to produce the representation

$$z(s) = \int\int K(s - u) e^{is^T x}\sqrt{f_{\theta(u)}(x)} w_u(x) dx du. \tag{3.135}$$

One can rewrite (3.135) in the form

$$z(s) = \int e^{is^T x} dW_s(dx),$$

where, proceeding formally,

$$dW_s(x) = \left\{\int K(s - u)\sqrt{f_{\theta(u)}(x)} w_u(x) du\right\} dx,$$

and the process $W_s(x)$ has orthogonal increments in the sense that for any s, s', whenever $x \neq x'$,

$$\mathrm{E}\left\{dW_s(x) dW_{s'}(x')\right\} = 0,$$

but the process is not orthogonal with respect to s because

$$\mathrm{E}\left\{dW_s(x) dW_{s'}(x)\right\} = \int K(s - u) K(s' - u) f_{\theta(u)}(x) dx.$$

In this case the process is not reducible to one of the form (3.116).

Just as (3.117) extends Higdon's model to a spatial-temporal process, so it is possible to make a similar extension with (3.127), so that

$$z(s, t) = \int_{\mathcal{S}} K(s - u) z_{\theta(u)}(s, t) du, \tag{3.136}$$

with each $z_{\theta(u)}(s,t)$ a spatial-temporal process. For example, it is possible that $z_{\theta(u)}(s,t)$ for each u will have a separable covariance function (i.e. one which factors into a product of spatial and temporal covariances), but the overall process $z(s,t)$ will not have this structure.

Example

Fuentes and Smith (2001) give an application to SO_2 modeling over the eastern U.S. We give only an outline of the analysis here and refer to the paper for the full details.

The background of this example lies in the E.P.A.'s need to understand the effects of possible changes in emissions control policies on observed levels of atmospheric pollutants. A numerical model that incorporates meteorology and atmospheric chemistry, known as Models-3, is run to simulate the SO_2 levels in the week of July 11, 1995, which is one particular week when very high SO_2 levels were observed. If the model succeeds in reproducing the observed monitoring data, then it can be re-run under alternative assumptions about emissions, to simulate the effect of new pollution regulations. However, a critical part of the assessment of model accuracy is its ability to reproduce observed monitoring data under the actual emissions scenario of the week in question. Fig. 3.11 shows Models-3 output for this week, averaged over 36 km^2 grid cells, while Fig. 3.12 shows the observed monitoring data, averaged over the week from hourly measurements, at 38 monitors which are part of the Clean Air Status and Trends Network (CASTNet). We shall use a spatial model fitted to the Models-3 data to "predict" the values at each of 6 sites, shown in Fig. 3.13. Comparison of the predictive distributions with the observed monitor values will then serve as a check on the accuracy of Models-3.

The modeling strategy is based on the covariance (3.129) but with the integral replaced by a finite sum, so we actually assume

$$\mathrm{Cov}\{z(s), z(s')\} = \frac{1}{M} \sum_{m=1}^{M} K(s - s_m) K(s' - s_m) C_{\theta(s_m)}(s - s'), \tag{3.137}$$

where the locations s_m were taken on a 9×9 grid (so $M = 81$ in this example). The kernel K was taken to be of scaled Epanechnikov form, so

$$K(s) = \frac{2}{\pi h^2} \left(1 - \frac{||s||^2}{h^2}\right) I(||s|| < h),$$

where in this example $h = 229$ km. The value of h was chosen to give reasonable overlap between the 81 sampling locations of s_m and does not reflect any "optimal bandwidth" considerations.

The stationary covariance $C_{\theta(\cdot)}(\cdot)$ in (3.137) was taken to be of Matérn form, and initial analysis of the results indicated that the Matérn sill parameter ϕ varied substantially over

the spatial region of the study. This was modeled hierarchically: if we write s_m as s_{ij} $(1 \leq i \leq 9, 1 \leq j \leq 9)$ where i and j are scaled longitude and latitude coordinates, then

$$\phi(s_{ij}) = a + r_i + c_j + \epsilon_\phi(s_{ij}), \tag{3.138}$$

where r_i and c_j are random longitude and latitude effects, and $\epsilon_\phi(\cdot)$ is a residual spatial process which is itself taken to be a Gaussian process with a Matérn covariance function. The whole analysis employs Bayesian hierarchical modeling concepts which have two advantages in this kind of analysis: first that they allow us to fit a nested model of the form (3.138) within the overall modeling equation (3.137), and second, that the Bayesian structure allows us to calculate predictive distributions for individual observations in a manner that correctly allows for the estimation of unknown model parameters.

A final complication is the "change of support" problem. We have specified a covariance function between points, not grid boxes. A point covariance function is needed if we are to obtain predictive distributions at individual locations, as is the objective. However, the Models-3 data are on grid boxes, not at individual locations. The model fitting must reflect this discrepancy between the scale of model-based grid-cell averages and monitoring data which are taken at individual locations. However, if one calculates averages over grid boxes B and B', one has

$$\mathrm{Cov}\left\{ \frac{1}{|B|}\int_B z(s)ds, \frac{1}{|B'|}\int_{B'} z(s')ds' \right\} = \frac{1}{|B| \cdot |B'|}\int_B \int_{B'} C(s,s')ds'ds. \tag{3.139}$$

The actual fitting took (3.137) as the definition of the pointwise covariance $C(s,s')$, but converted it to the form (3.139) before computing the likelihood function for the Models-3 data. The integrations in (3.139) were replaced by discrete sums for computational convenience.

Fig. 3.14 shows a map of the fitted posterior model of the sill parameter, and Fig. 3.15 shows the fitted longitude and latitude effects, and the estimated semivariogram of ϵ_ϕ, corresponding to (3.138). The fitted model is then used to compute predictive distributions for six CASTNet monitoring locations in the states of Florida, Michigan, North Carolina, Indiana, Maine and Illinois (Fig. 3.13). Fig. 3.16 shows posterior densities for each of these locations. It is particularly noticable that the Indiana site has a very dispersed predictive distribution. This site is near the Indiana coal fields and the predictive distribution may reflects a large local heterogeneity of the SO_2 field in that area. The North Carolina SO_2 level also seems somewhat elevated, and with a relatively large predictive variance (though not nearly as large as for the Indiana site), which may reflect the influence of the Tennessee power plants.

Finally, the right-hand plot in Fig. 3.17 shows the predictive means and 90% credible intervals for the predictions based on Models-3, plotted against the observed monitoring data. As a comparison, the left-hand plot shows a much more simple-minded comparison, which simply involves plotting the Models-3 output for the nearest grid cell against the

observed monitoring data. This crude comparison takes no account of the change of support problem, nor does it contain any measure of uncertainty. The right-hand plot shows, much more clearly than the left-hand plot, the true difficulty with this example: whereas the fit between model output and monitoring data is very good at four of the sites, at the remaining two (Indiana and North Carolina) the model clearly overpredicts the actual SO_2 data, implying that further understanding of these locations is needed in order to improve the model.

Appendix: Derivation of (3.115)

With $K(u) = \frac{2}{\pi}(1 - ||u||^2)I(||u|| < 1)$, rewrite (3.110) as

$$C(h) = \int K\left(u - \frac{h}{2}\right) K\left(u + \frac{h}{2}\right) du,$$

where we assume $||h|| = 2t,\ 0 \le t \le 1$ (if $t > 1$ the integral is 0). Writing $u = (x, y)$, there is no loss of generality in assuming $\frac{h}{2} = (t, 0)$, and we consider only the integral over $x > 0, y > 0$, since the other three quadrants are the same by symmetry. Moreover, within this quadrant the integrand is positive only when $(x + t)^2 + y^2 < 1$, so we may write

$$C(h) = \frac{16}{\pi^2} \int_0^{1-t} \int_0^{\sqrt{1-(x+t)^2}} \left\{1 - (x - t)^2 - y^2\right\} \left\{1 - (x + t)^2 - y^2\right\} dy dx.$$

First perform the integral with respect to y: we get

$$\begin{aligned} C(h) = \frac{16}{\pi^2} \int_0^{1-t} \sqrt{1 - (x + t)^2} \Big[&\left\{1 - (x - t)^2\right\} \left\{1 - (x + t)^2\right\} \\ &- \frac{1}{3} \left\{2 - (x - t)^2 - (x + t)^2\right\} \left\{1 - (x + t)^2\right\} + \frac{1}{5} \left\{1 - (x + t)^2\right\}^2 \Big] dx. \end{aligned} \tag{3.140}$$

After some manipulation, the term inside square brackets in (3.140) is seen to be

$$\sum_{k=0}^{4} c_k (x + t)^k, \tag{3.141}$$

with $c_0, ..., c_4$ given by (3.114). Moreover, the substitution $x + t = \sqrt{1 - u}$ shows that

$$\begin{aligned} \int_0^{1-t} \sqrt{1 - (x + t)^2} (x + t)^k &= \int_0^{1-t^2} u^{1/2} (1 - u)^{k/2 - 1/2} \\ &= B_{1-t^2}\left(\frac{3}{2}, \frac{k}{2} + \frac{1}{2}\right). \end{aligned} \tag{3.142}$$

The result (3.115) follows after combining (3.140), (3.141) and (3.142).

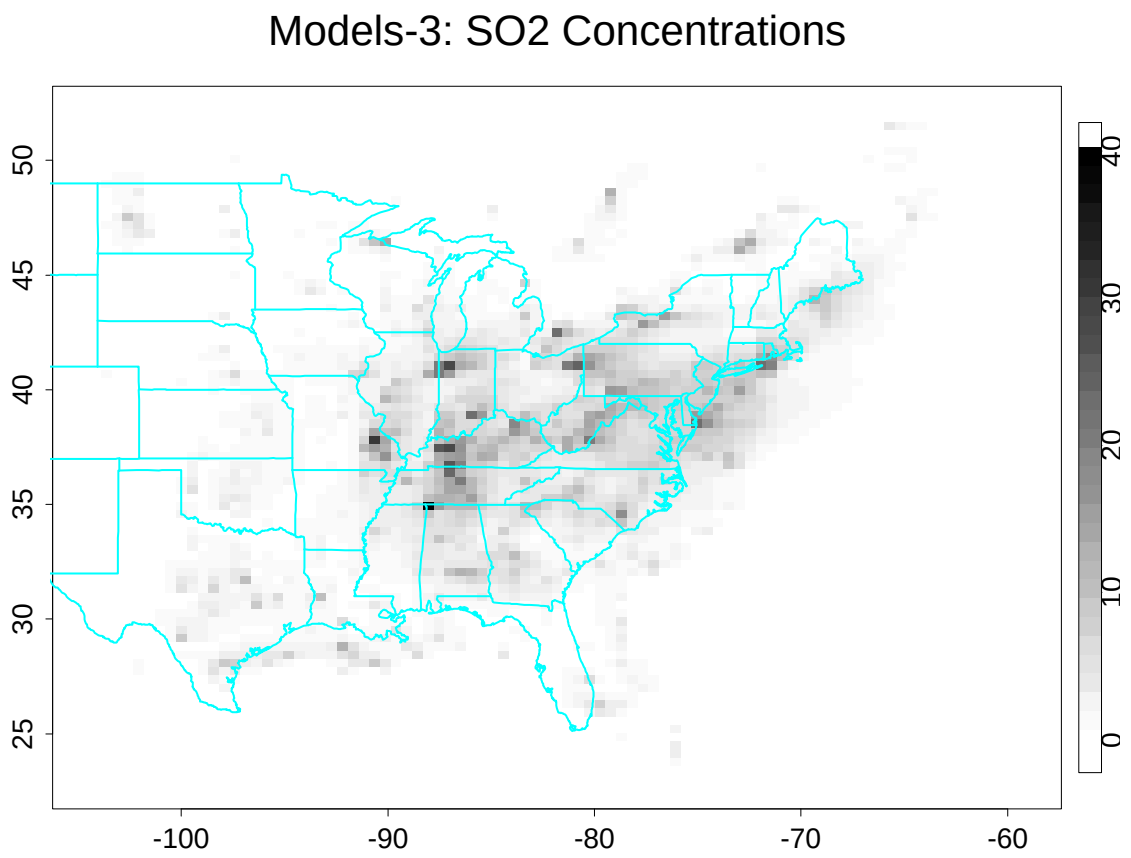


Fig. 3.11 Models-3 output, mean SO_2 concentrations, week of July 11 1995

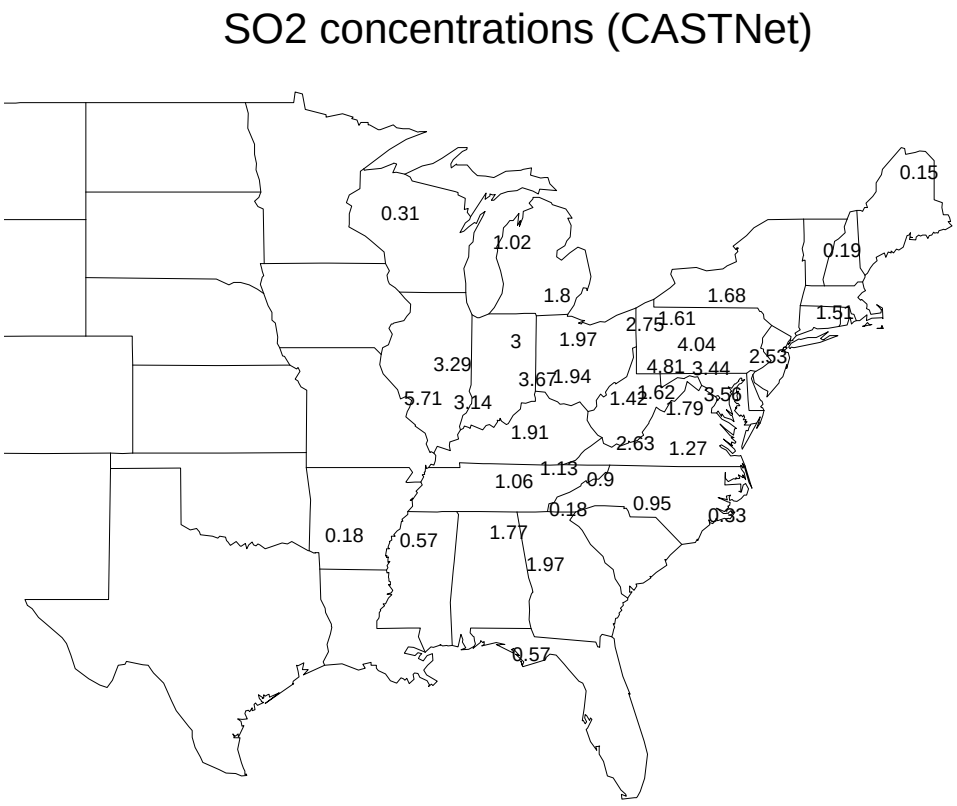


Fig. 3.12 CASTNet data, mean SO_2 concentrations, week of July 11 1995

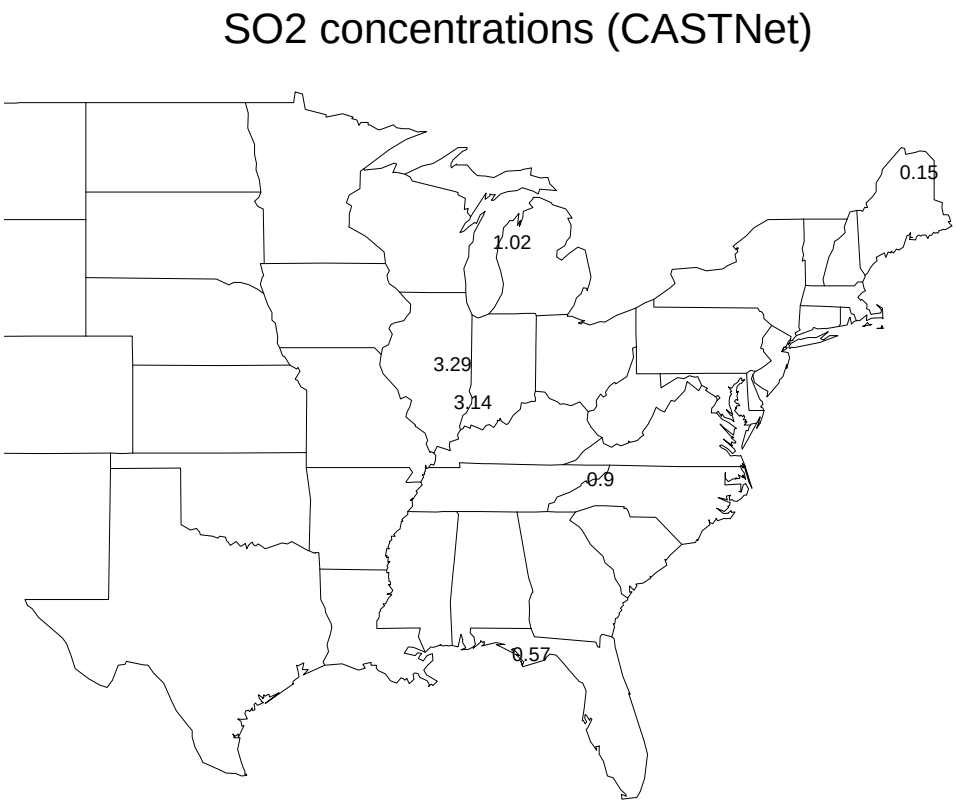


Fig. 3.13 Mean SO_2 concentrations at selected sites, week of July 11 1995

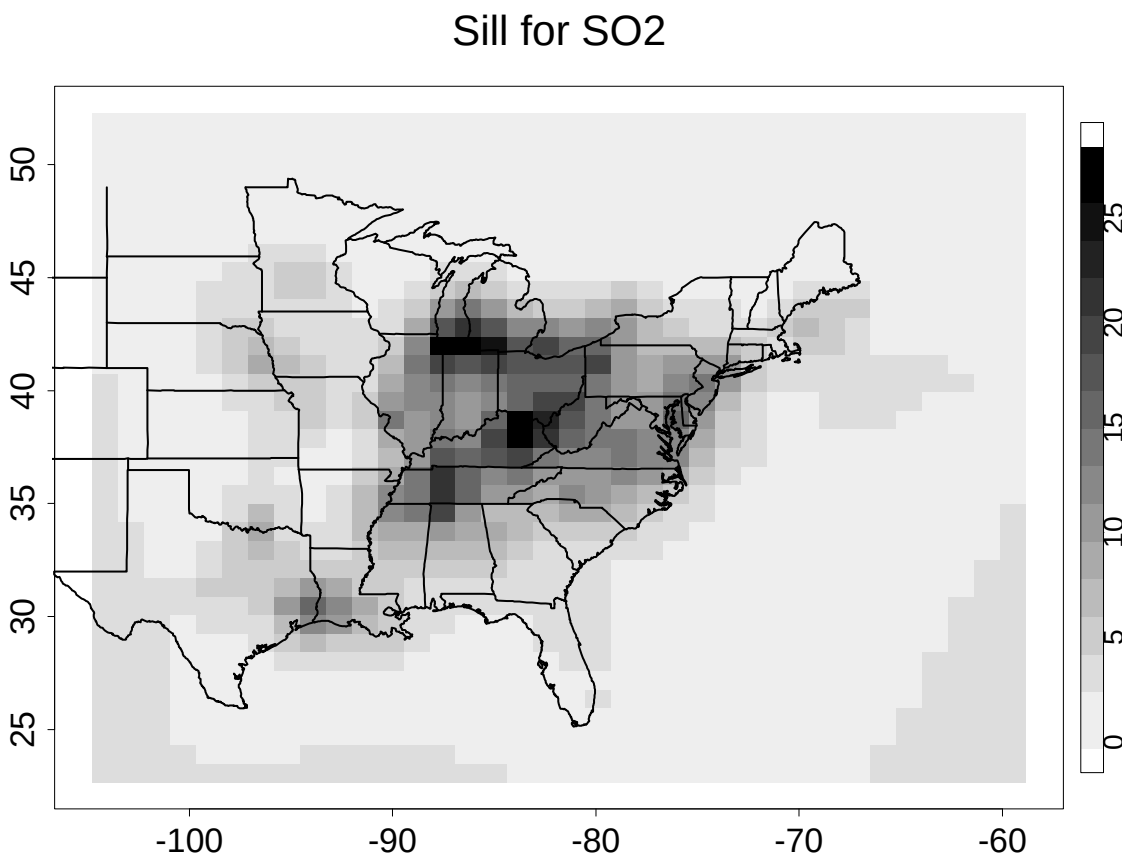


Fig. 3.14 Modes of posterior distribution of Matérn sill parameter from Models-3 data

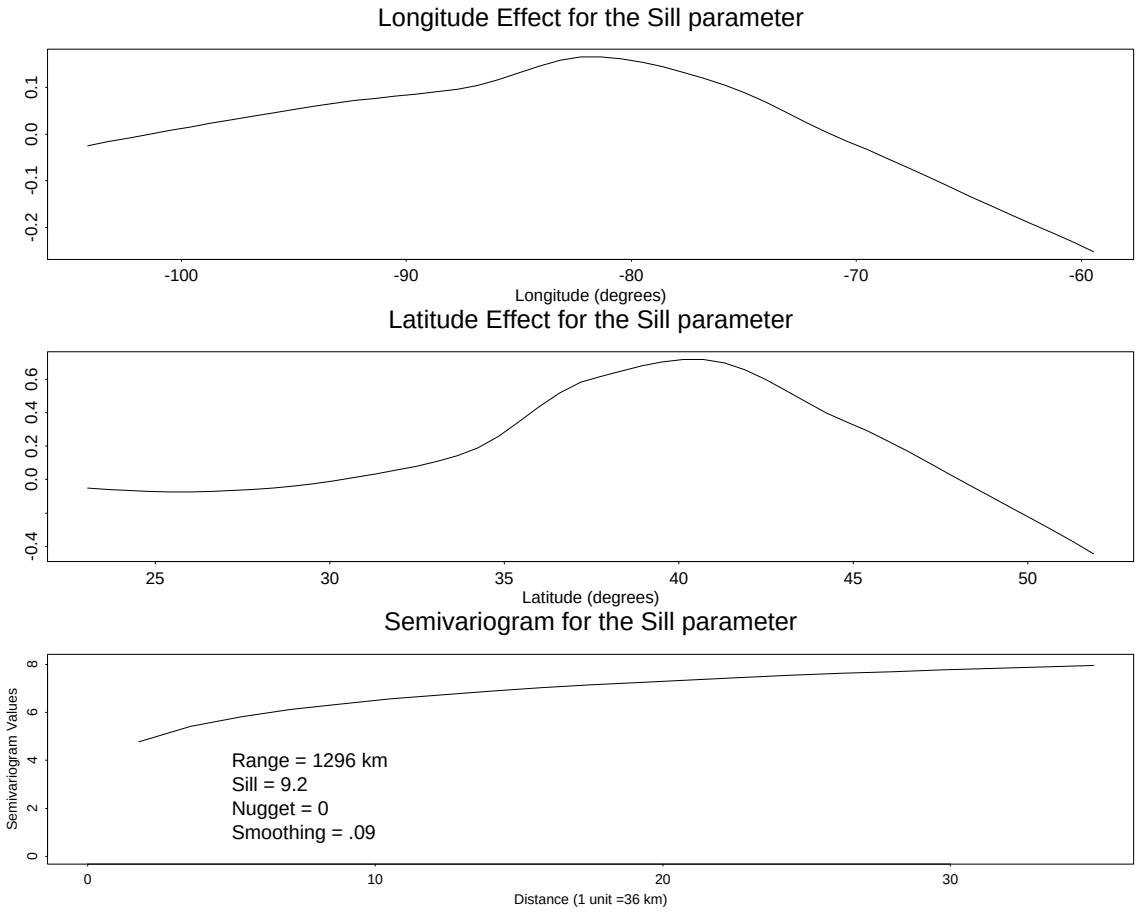


Fig. 3.15 Spatial characteristics of Matérn still parameter

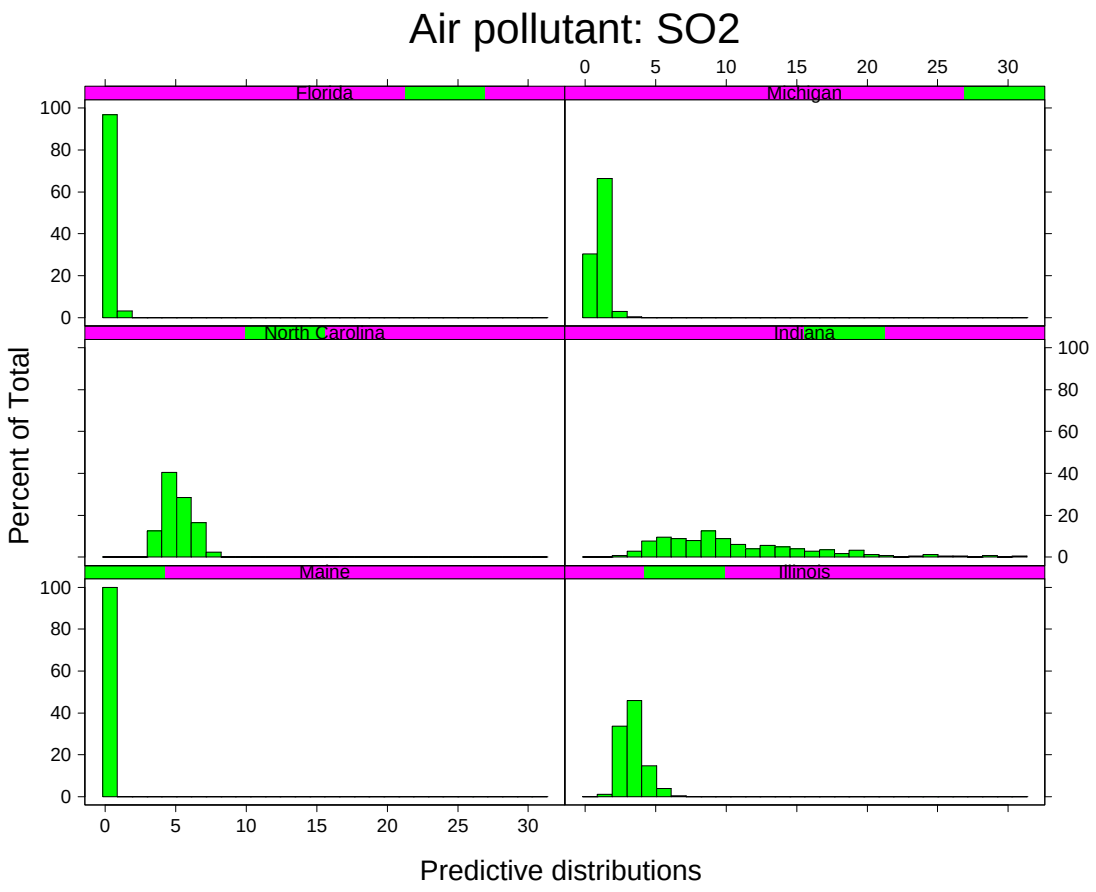


Fig. 3.16 Predictive distributions of Models-3 concentrations at 6 selected sites

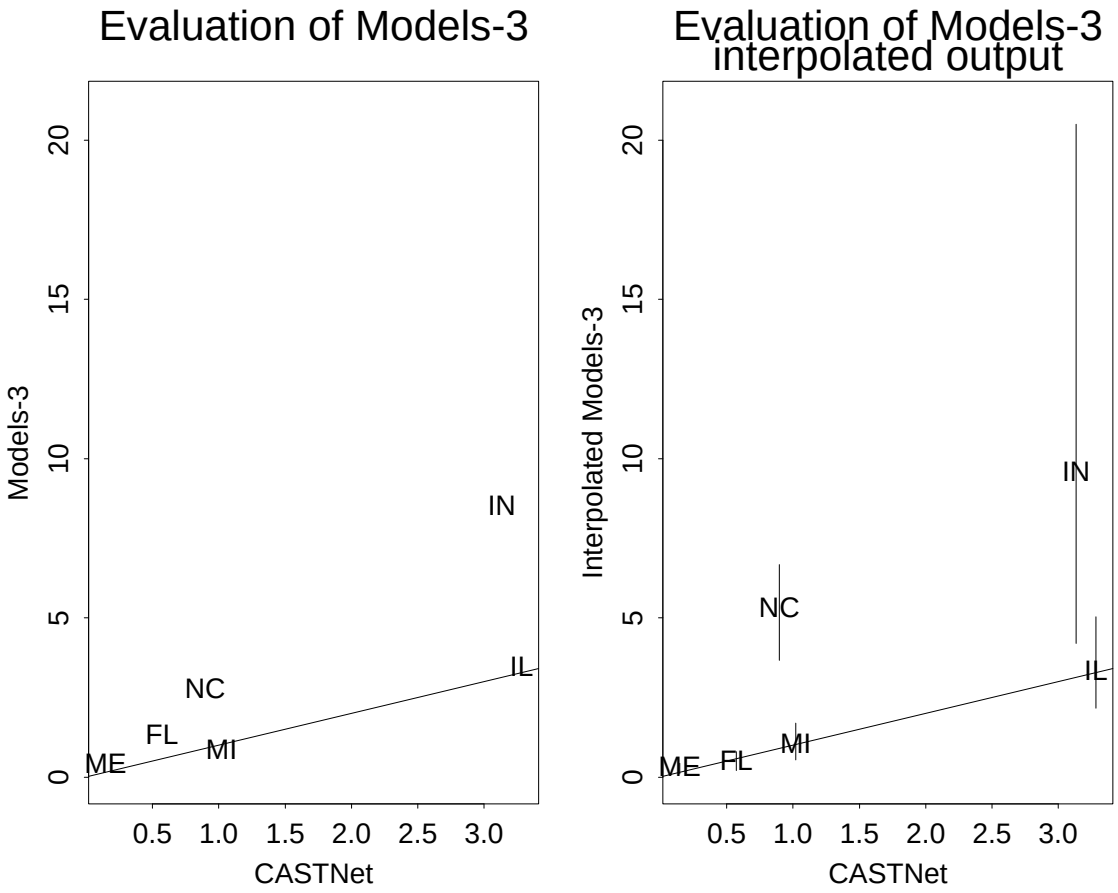


Fig. 3.17 Comparisons of CASTNet predictions with Models-3 output: (L) crude comparison, (R) after allowing for change of support.

CHAPTER 4

Models Defined by Conditional Probabilities

Chapters 2 and 3 were concerned primarily with Gaussian models, i.e. ones in which all the joint distributions could be described by multivariate normal distributions, in which the covariance function of the process was defined in a wide variety of ways. The alternative broad strategy for defining spatial models is the *conditional probabilities* approach, in which the model is defined in terms of the conditional distribution of the observation at one location given its values at other locations. The most convenient models of this kind are those when the spatial locations have some form of lattice structure with "neighbors" defined by the links of the lattice, and the conditional probabilities at a single site are a function solely of the values at a neghboring site. Such models are called *Markov random fields* and have been much studied by both probabilists and statisticians. They were originally developed in statistical physics — for example, the famous Ising model for phase transitions was of this form and the first example of what we now call Markov chain Monte Carlo methods was developed in the context of simulating from a statistical physics system (Metropolis *et al.* 1953). The idea of using models of this type as statistical models has its origins in work by Whittle (1954) and Bartlett (see, e.g. Bartlett (1978) for discussion and earlier references) but the most significant breakthrough was the paper by Besag (1974) which laid out both a probabilistic structure for Markov random fields and methods of inference.

From a modern-day perspective, Markov random fields are often used as prior distributions as part of a hierarchical model structure. Although such models are often based on lattices, the lattice need not be regular (e.g. some spatial analyses have been based either on states within the USA, or counties within a state, where one certainly does not have any regular lattice structure, though it is possible to define "neighbors", e.g. by the condition that two states or two counties have a common border) and in some cases there is no lattice structure at all. For this reason, the title of this chapter reflects that the models may be more general than just models defined on lattices, though by far the most commonly applied models do involve some kind of lattice structure whether is is regular or not.

4.1 Markov random fields as spatial models

In this section we outline the basic ideas of Markov random fields on finite lattices, and their inference, following the seminal paper of Besag (1974).

4.1.1 Introduction to lattice models

As in earlier chapters, we are concerned with a spatial process in which values are measured at a set of points. In contrast with earlier cases where the sampling points are distributed arbitrarily, however, we now consider models where they are on a fixed lattice, such as a square or triangular lattice. Typically we shall assume there are n lattice points and write $i, j, k, ...$ for the points of the lattice, $X_i, X_j, X_k, ...$, for the measurements made at those points. Associated with each point i there is a set of neighbors of i, which we denote by N_i. We assume the neighborhood relationship is symmetric in the sense that $j \in N_i$ if and only if $i \in N_j$. Fig. 4.1 shows three typical examples of lattice arrangements.

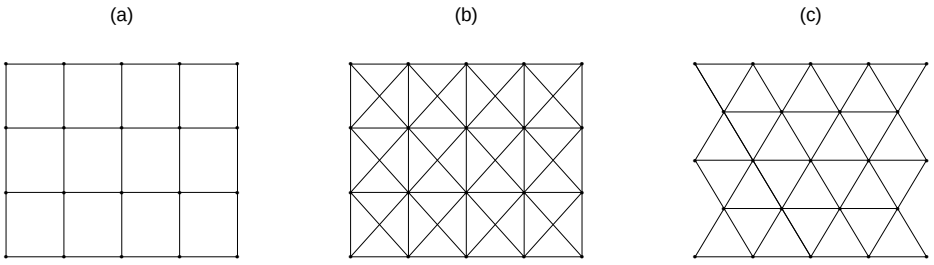


Fig. 4.1. Three examples of lattices. (a) Square lattice, first-order neighbor scheme. (b) Square lattice, second-order neighbor scheme. (c) Triangular lattice.

Within such a structure, we could in principle determine a totally arbitrary spatial model by specifying all possible joint probabilities $p(x_1, ..., x_n)$, interpreted as a probability mass function in the case of discrete random variables and as a probability density function in the continuous case. Alternatively, one can try to define a model by specifying conditional probabilities of the form $p(x_i|x_j,\ j \neq i)$. Throughout this chapter, the letter p will be used interchangably to denote a probability mass function or a density function, and to denote either marginal, joint or conditional probabilities — the meaning will be clear from the arguments of the function. Prior to the early 1970s, researchers had studies spatial distributions both from the conditional-probability viewpoint and the joint-pronbability viewpoint, but there was no clear understanding of the link between them. In many contexts, the meaning of a probability model is easier to understand if it is expressed in terms of conditional probabilities, but if one just writes down conditional probabilities in an arbitrary way, there is no clear-cut route to translate the conditional probabilities into joint probabilities, and indeed no guarantee that any such joint probability distribution exists. However, by the early 1970s, it had become clear that in an important class of special cases, known as *Markov random fields* (MRFs), there was a precise link between the conditional and joint probabilities. This result became known as the Hammersley-Clifford theorem after its originators, though the original proof of Hammersley and Clifford remained unpublished for many years. Besag provided a short proof of a special case of the Hammersley-Clifford theorem, one which is however adequate for almost all statistical applications, and was the first to develop general methods of statistical estimation for such processes. These in turn have had an important influence on modern ideas such as the Gibbs sampler, first introduced explicitly by Geman and Geman (1984), but whose basic concepts relied heavily on earlier work by Besag and others.

An example of a lattice model is the *auto-logistic* model: $X_1, ..., X_n$ are 0–1 random variables and

$$\begin{aligned} \Pr\{X_i = 1 | X_j = x_j,\ j \neq i\} &= \Pr\{X_i = 1 | X_j = x_j,\ j \in N_i\} \\ &= \frac{\exp(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j)}{1 + \exp(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j)}. \end{aligned} \tag{4.1}$$

Another model is the *auto-normal*:

$$X_i | (X_j = x_j,\ j \neq i) \sim N\left(\mu_i + \sum_{j \in N_i} \beta_{ij}(x_j - \mu_j), \sigma^2\right). \tag{4.2}$$

Another model that at first sight may seem the same as (4.2) is

$$X_i = \mu_i + \sum_{j \in N_i} \beta_{ij}(X_j - \mu_j) + \epsilon_i, \qquad \epsilon_i \text{ independent } N(0, \sigma^2), \tag{4.3}$$

which is known as the *simultaneous equation model*. Simultaneous equation models are widely used in econometrics but are in fact somewhat different from Markov random field models — the precise connection between (4.2) and (4.3) will be given later.

The obvious question about any of (4.1)–(4.3) is to decide whether the equation specifies a legitimate probability model, in the sense that the equations uniquely determine the joint probability distribution of all the random variables. In the case of (4.1) and (4.2) the answer is: yes, provided $\beta_{ij} = \beta_{ji}$ for all i and j. This is most easily seen by writing down an explicit formula for the joint density. In the case of (4.1) this is

$$p(x_1, ..., x_n) = K \exp\left(\sum_k \alpha_k x_k + \frac{1}{2} \sum_j \sum_{k \in N_j} \beta_{jk} x_j x_k\right) \tag{4.4}$$

where K is a normalizing constant chosen to make the probabilities sum to 1. For, if (4.4) holds, we have

$$\begin{aligned} \frac{\Pr\{X_i = 1 | X_j = x_j,\ j \neq i\}}{\Pr\{X_i = 0 | X_j = x_j,\ j \neq i\}} &= \frac{p(x_1, ..., x_{i-1}, 1, x_{i+1}, ..., x_n)}{p(x_1, ..., x_{i-1}, 0, x_{i+1}, ..., x_n)} \\ &= \exp\left(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j\right) \end{aligned} \tag{4.5}$$

where the factor $\frac{1}{2}$ in (4.4) has disappeared because the index pair (i, j) is counted twice in (4.4) but only once in (4.5). This is where the symmetry assumption comes in: otherwise β_{ij} in (4.5) would have to be replaced by $(\beta_{ij} + \beta_{ji})/2$.

Similarly, the joint density consistent with (4.2) is

$$p(x_1, ..., x_n) = (2\pi\sigma^2)^{-(n/2)}|B|^{1/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j,k}(x_j - \mu_j)b_{jk}(x_k - \mu_k)\right\} \tag{4.6}$$

where the matrix B has entries $\{b_{jk}\}$ given by

$$b_{jk} = \begin{cases} 1 & \text{if } j = k, \\ -\beta_{jk} & \text{if } j \in N_k, \\ 0 & \text{otherwise.} \end{cases} \tag{4.7}$$

To see that this is consistent with (4.2), fix index i and complete the square in the exponent as

$$\begin{aligned} &-\frac{1}{2\sigma^2}(x_i - \mu_i)^2 + \frac{1}{\sigma^2}\sum_{j\in N_i}(x_i - \mu_i)\beta_{ij}(x_j - \mu_j) + ... \\ &= -\frac{1}{2\sigma^2}\left\{x_i - \mu_i - \sum_{j\in N_i}\beta_{ij}(x_j - \mu_j)\right\}^2 + ... \end{aligned}$$

where in each case ... denotes terms that do not depend on x_i. This conditional density implies (4.2). Note that we have again used the assumption $\beta_{ij} = \beta_{ji}$.

To write (4.3) in a similar way, suppose again we define a matrix B with entries $\{b_{jk}\}$ given by (4.7) and we assume that B is invertible, but this time we do not need to assume symmetry. Then (4.3) can be written in vector form as

$$B(\mathbf{X} - \boldsymbol{\mu}) = \boldsymbol{\epsilon}$$

implying that $\mathbf{X}$ is a multivariate normal random variable with mean $\boldsymbol{\mu}$ and covariance matrix $B^{-1}B^{-T}\sigma^2$, or in terms of the joint density

$$p(x_1, ..., x_n) = (2\pi\sigma^2)^{-(n/2)}|B| \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x} - \boldsymbol{\mu})^T B^T B(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{4.8}$$

which can also be written in the form (4.6), but with a different definition of the matrix B. In other words, any model specified by (4.3) can also be rewritten in the form (4.2), but with different coefficients. One advantage of the form (4.2) is identifiability, since the matrix B in (4.8) is only identifiable through $B^T B$.

These examples show that the general problem of verifying whether a family of conditional distributions is consistent with some joint distribution is not trivial. In general, suppose we are given a family of one-dimensional conditional distributions of the form

$p(x_i|x_j,\ j \neq i)$. To calculate the joint distribution $p(x_1, ..., x_n)$, we first fix some reference state $(x_1^*, ..., x_n^*)$ and then characterize the ratio in the form

$$\begin{aligned} \frac{p(x_1, ..., x_n)}{p(x_1^*, ..., x_n^*)} &= \prod_{i=0}^{n-1} \frac{p(x_1^*, ..., x_i^*, x_{i+1}, x_{i+2}, ..., x_n)}{p(x_1^*, ..., x_i^*, x_{i+1}^*, x_{i+2}, ..., x_n)} \\ &= \prod_{i=0}^{n-1} \frac{p(x_{i+1}|x_1^*, ..., x_i^*, x_{i+2}, ..., x_n)}{p(x_{i+1}^*|x_1^*, ..., x_i^*, x_{i+2}, ..., x_n)}. \end{aligned} \tag{4.9}$$

Since we can carry out the operation (4.9) for any $(x_1, ..., x_n)$, we can calculate all the joint probabilities relative to the reference state, and then renormalize to obtain the complete joint probability distribution. However, this implies some restrictions on the class of allowable functions. One condition is that the operation (4.9) must be invariant to permutations of the coordinates. This amounts to a consistency condition on the conditional probabilities. Another consistency condition is that the result must be invariant to the choice of the reference state, e.g. if we chose an alternative $\mathbf{x}^{**}$, also having positive probability, then as calculated by (4.9), we have

$$\frac{p(\mathbf{x})}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x})}{p(\mathbf{x}^{**})} \cdot \frac{p(\mathbf{x}^{**})}{p(\mathbf{x}^*)}. \tag{4.10}$$

Conditions of this form were discussed by Brook (1964). The final complication in using (4.9) to specify joint probabilities is in the calculation of the normalizing constant needed to complete the calculation. If there are n sites each of which may be in one of m possible states, then there are m^n possible states of the system, a prohibitive number unless n is extremely small. This fact in itself has motivated much of the modern theory of MCMC methods, which provide a means of estimating such a constant by simulation.

4.1.2 Markov random fields and the Hammersley-Clifford Theorem

We saw in section 4.1.1 that lattice models can be easily defined and provide a good intuitive understanding of how a variable at a given location depends on its neighbors, but it is not so easy to specify exactly when models defined by conditional probabilities lead to consistent systems of joint probabilities. The most important class of models when this can be done is the class of Markov random field models, and the mechanism for defining consistent systems of probability distributions is provided by a result known as the Hammersley-Clifford theorem. This result was first given by J.M. Hammersley and P. Clifford in an unpublished paper from 1971, and subsequently rederived in different forms by a number of other authors. Besag (1974) gave an elementary proof in the case of a finite lattice subject to a further condition (the positivity condition), and we concentrate on that approach here. For further information about the Hammersley-Clifford theorem and its historical development, we refer to the published discussion from Besag (1974), Clifford (1990) which included a copy of the original proof, and Smith (1997a).

A Markov random field is characterized by the property that all conditional probabilities take the form

$$p(x_i|x_j,\ j \neq i) = p(x_i|x_j,\ j \in N_i) \tag{4.11}$$

where N_i denotes the set of neighbors of i under some lattice structure. Of course (4.11) includes cases in which N_i consists of all other sites in the network (the complete graph), but in most cases of practical interest we are working on a network where each site has a fixed small number of neighbors, and in such cases the Hammersley-Clifford theorem provides a real simplification.

We define a *clique* to be any subset of sites with the property that each member of the clique is a neighbor of each other member. In most cases the cliques are extremely simple sets: for example with a square lattice, the cliques are single sites, and pairs of sites which are mutual neighbors; there are no other cliques.

We shall assume (initially) that there are only finitely many values x_i available at each site, and that one of these is 0. The latter condition is not a restriction because we can simply relabel the values at each site without changing the structure of the model.

The final condition and controversial is *positivity*: given any state $\mathbf{x} = (x_1, ..., x_n)$ such that each component $x_1, ..., x_n$ has positive marginal probability, we assume that the joint probability $p(\mathbf{x})$ is also positive. Since there are only finitely many possible states per site, we can simply eliminate all states of zero marginal probability and write the condition as $p(\mathbf{x}) > 0$ for all $\mathbf{x}$.

Following Besag, we define

$$q(\mathbf{x}) = \log\left\{\frac{p(\mathbf{x})}{p(\mathbf{0})}\right\}. \tag{4.12}$$

Then Besag claimed that there exist functions $g_i(x_i)$, $g_{ij}(x_i, x_j)$, etc., such that

$$\begin{aligned} q(\mathbf{x}) = &\sum_i x_i g_i(x_i) + \sum_{i<j} x_i x_j g_{ij}(x_i, x_j) + \sum_{i<j<k} x_i x_j x_k g_{ijk}(x_i, x_j, x_k) \\ &+ ... + x_1 x_2 ... x_n g_{12...n}(x_1, x_2, ..., x_n). \end{aligned} \tag{4.13}$$

To see (4.13), for example, we define $g_i(x_i)$ by

$$x_i g_i(x_i) = q(0, ..., 0, x_i, 0, ..., 0) - q(\mathbf{0}),$$

and similar if more complicated differencing operations define higher-order g's as well.

The Hammersley-Clifford theorem, in the form stated by Besag, is now as follows:

For a Markov random field, $g_{ij...s}(x_i, x_j, ..., x_s)$ is non-zero if and only if $\{i, j, ..., s\}$ form a clique. Subject to this restriction, the g's are arbitrary.

The importance of this result is that in conjunction with (4.13) and (4.12), it enables us immediately to write down the general form of joint probability distribution for any MRF.

Therefore, given a conjectured family of conditional distributions, we can immediately check whether they are consistent with some member of this family.

Proof. For any state $\mathbf{x}$, let $\mathbf{x}_i$ denote the same state with the i'th coordinate set equal to 0. The following argument is the same for any i, so without loss of generality set $i = 1$.

From (4.13) we have

$$q(\mathbf{x}) - q(\mathbf{x}_1) = x_1 \left\{ g_1(x_1) + \sum_{j>1} x_j g_{ij}(x_1, x_j) + ... \right\}.$$

Suppose $k \notin N_1$. Then $q(\mathbf{x}) - q(\mathbf{x}_1)$ is independent of the k'th component x_k for all $\mathbf{x}$. Putting $x_i = 0$ for $i \neq 1$ or k, we see $g_{1k}(x_1, x_k) = 0$. With different choices of the vector $\mathbf{x}$, the same argument also shows that all higher-order g's are 0 as well.

The converse argument (that any family of g's satisfying the conditions of the theorem suffices for a MRF) follows simply by observing that the q's given by (4.13) define a MRF. This completes the proof.

The controversy over the positivity condition arose because Hammersley apparently believed the theorem was true without such a condition, though subsequent developments showed that this was not true. Some of this was included in Hammersley's own published discusson of Besag (1974). The result has also been extended to two other cases, under a similar positivity condition:

• the case where the number of possible states per site is countable, provided $\sum e^{q(\mathbf{x})} < \infty$,

• the case of continuous random variables, $p(\mathbf{x})$ being interpreted as a probability density but otherwise the same equations (4.12) and (4.13) holding, provided $e^{q(\mathbf{x})}$ is integrable.

4.1.3 Specific Spatial Models

We have already seen some specific examples of lattice models, namely, the auto-logistic and auto-normal schemes given by (4.1) and (4.2) respectively. These are particular instances of what Besag called *auto-models*, in which the q function is of form

$$q(\mathbf{x}) = \sum_i x_i g_i(x) + \sum_{i<j} \beta_{ij} x_i x_j \tag{4.14}$$

in which $\beta_{ij} = 0$ unless i and j are neighbors. For such models, the associated conditional probabilities are of the form

$$\frac{\Pr\{X_i = x_i | X_j = x_j,\ j \neq i)}{\Pr\{X_i = 0 | X_j = x_j,\ j \neq i)} = \exp\left[x_i \left\{ g_i(x_i) + \sum_j \beta_{ij} x_j \right\} \right], \tag{4.15}$$

in which $\beta_{ij} = 0$ unless i and j are neighbors, and also $\beta_{ij} = \beta_{ji}$ for all i, j.

The auto-logistic and auto-normal are the two best-known examples, but there are other examples in which the conditional distributions take other exponential family forms, such as the auto-Poisson and auto-exponential models. The auto-Poisson model takes the form that each X_i, conditionally on its neighbors $X_j = x_j$, $j \in N_i$, has a Poisson distribution with mean μ_i, of the form

$$\mu_i = \exp\left(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j\right) \tag{4.16}$$

but there is an important restriction: in order for (4.16) to satisfy $\sum_{\mathbf{x}} \exp\{q(\mathbf{x})\} < \infty$, we must have $\beta_{ij} \leq 0$ for each (i, j) pair. Intuitively, the auto-Poisson model would seem ideally suited to applications such as disease counts, but the non-positivity condition on the $\{\beta_{ij}\}$ implies that it is only usable in cases of negative interactions between neighboring sites. This is a severe restriction and has led to alternative conditionally Poisson schemes being developed in recent years, see section ?????

Models for square lattices

When the sites take the form of a square lattice, it is natural to label each site with a two-dimensional coordinate (i, j), where i and j are integers, and to denote the corresponding random variables $X_{i,j}$, etc. Also, we can simplify the classification of models by restricting ourselves to those which obey some stationarity or homogeneity assumptions. A *first-order model* is one in which the site (i, j) has just four neighbors: $(i-1, j)$, $(i+1, j)$, $(i, j-1)$ and $(i, j+1)$ (Fig. 4.1(a)). One simple model allows for differential row and column effects in the form

$$q(\mathbf{x}) = \alpha \sum x_{i,j} + \beta_1 \sum x_{i,j} x_{i+1,j} + \beta_2 \sum x_{i,j} x_{i,j+1}, \tag{4.17}$$

where the isotropic model $\beta_1 = \beta_2$ is a special case of (4.17). A *second-order model* includes diagonal cross-links between site (i, j) and the neighbors $(i-1, j-1)$, $(i-1, j+1)$, $(i+1, j-1)$ and $(i+1, j+1)$ (Fig. 4.1(b)). One can, of course, continue the discussion to define third- and higher-order models.

In the case of a second-order model, one can naturally extend (4.17) to

$$\begin{aligned} q(\mathbf{x}) = \alpha \sum x_{i,j} + \beta_1 \sum x_{i,j} x_{i+1,j} + \beta_2 \sum x_{i,j} x_{i,j+1} \\ + \gamma_1 \sum x_{i,j} x_{i+1,j+1} + \gamma_2 \sum x_{i,j} x_{i+1,j-1} \end{aligned} \tag{4.18}$$

but this is not the most general form of model, since in this case there are cliques of three neighbors so one could include terms of the form $x_{i,j} x_{i-1,j} x_{i,j-1}$, and so on, thus going beyond the framework of auto-models.

For Gaussian models, one can re-express these models in terms of the conditional mean of a random variable given its neighbors, e.g. analogously to (4.17) one has

$$\mathrm{E}\{X_{i,j}|X_{i',j'},\ (i',j') \neq (i,j)\} = \alpha + \beta_1(X_{i-1,j} + X_{i+1,j}) + \beta_2(X_{i,j-1} + X_{i,j+1}) \quad (4.19)$$

or, analogously to (4.18),

$$\begin{aligned}\mathrm{E}\{X_{i,j}|X_{i',j'},\ (i',j') \neq (i,j)\} = \alpha + \beta_1(X_{i-1,j} + X_{i+1,j}) + \beta_2(X_{i,j-1} + X_{i,j+1}) \\ + \gamma_1(X_{i-1,j-1} + X_{i+1,j+1}) + \gamma_2(X_{i-1,j+1} + X_{i+1,j-1}).\end{aligned} \quad (4.20)$$

4.2 Inference in lattice models

Section 4.1 showed how to define a large class of lattice models. We now consider statistical inference about the parameters of such models. Again, the initial ideas on this subject were given by Besag (1974), and extended by Besag (1975) who introduced the idea of *maximum pseudolikelihood estimation* (MPLE). However, with the development of modern MCMC techniques, it is now possible to talk about estimation in more general frameworks, including Monte Carlo approximations to the exact maximum likelihood estimates (MCMLE).

In this section, we review a number of different methods.

4.2.1 Coding methods

This idea was discussed by Besag (1974). Suppose we have a square lattice with a first-order neighborhood scheme (Fig. 4.1(a)). Suppose we condition on the odd points, i.e. all points (i,j) such that $i-j$ is odd. Conditionally on those points, the even points are independent, so one can write down an exact joint density for the even points

$$\prod_{(i,j):\ i-j\ \text{even}} p(x_{i,j}|x_{i',j'},\ (i',j') \neq (i,j)) \quad (4.21)$$

and interpret this as an exact likelihood for the even data points, conditionally on all the odd data points. Thus one can maximize (4.21) with respect to any unknown parameters, interpreting the results as conditional maximum likelihood estimators. These estimators will not be efficient, however, because they are based on only half the data points.

One can, of course, define a complementary scheme based on the joint density of all the odd points conditioned on the even points. For more complicated lattices, it is more difficult to find subsets of the lattice points which are conditionally independent given the rest, but in principle, the idea is applicable in any regular lattice, and a number of specific examples were given by Besag and some discussants of his paper.

4.4.2 Pseudolikelihood

Besag (1975) extended the coding idea of Besag (1974), to suggest a general scheme for estimating lattice processes.

In the coding scheme for a regular lattice, one can define a conditional likelihood as in (4.21), or the equivalent formula based on odd lattice points. One obvious way to combine the two is simply to multiply the two conditional likelihoods together, resulting in

$$\prod_{(i,j)} p(x_{i,j}|x_{i',j'},\ (i',j') \neq (i,j)) \tag{4.22}$$

where the product is taken over all lattice points (i,j). However (4.22) is in principle a much more general idea, not dependent on any particular form of lattice structure, but simply defining a "likelihood function" as the product of all one-dimensional conditional distributions. To distinguish this from a true likelihood function, it is widely known as the *pseudolikelihood*. Maximum pseudolikelihood estimates are chosen so as to maximize the pseudolikelihood with respect to unknown parameters of the model.

The advantage of maximum pseudolikelihood estimates is that they are extremely easy to compute. The disadvantage is that we do not know very much about their sampling properties, except in a few particular instances. The use of standard approximations for standard errors via the observed information matrix, and for comparing models via likelihood ratio tests, are not valid when applied to pseudolikelihoods. In the case of standard errors, one way round the problem would be to use the information sandwich approach (subsection 2.2.3), but we are not aware of any instances of this method actually being applied in the present context. More detailed statistical properties are reviewed in a later section.

4.2.3 Exact and approximate MLEs for Gaussian processes

The calculation of exact MLEs for Gaussian processes is simpler than for other kinds of conditionally defined processes, because in the case of a joint normal density such as (4.6), the normalizing constant for the probability measure is explicitly defined. In other kinds of models, such as (4.4), there is no exact formula for the normalizing constant K, short of summation over all possible states, which is prohibitively slow in a large lattice. Thus we can in principle compute the MLEs for the auto-normal process directly, by maximizing (4.6) with respect to the unknown parameters, whereas we cannot do the same thing with a non-Gaussian likelihood such as (4.4). In practice, even (4.6) may be hard to evaluate in a large lattice, given the need to evaluate $|B|$.

An alternative approach is to evaluate $|B|$ approximately, using methods first given by Whittle (1954). Whittle's ideas were in fact developed for simultaneous autoregression models of the structure of (4.3), but Besag pointed out that, because of the close similarity

of the likelihood expressions in (4.6) and (4.8), we could effectively use the same approximations with auto-normal models. As discussed by Besag, $\frac{1}{n}\log|B|$ is approximated by the coefficient of $z_1^0 z_2^0$ in the power series expansion

$$\log\left(1 - \sum_{j,k} \beta_{jk} z_1^j z_2^k\right). \tag{4.23}$$

For example, in the case of model (4.19), (4.23) is the absolute term in the power series expansion of

$$\log\left\{1 - \beta_1(z_1 + z_1^{-1}) - \beta_2(z_2 + z_2^{-1})\right\}$$

while in the case of model (4.20), the corresponding expression is

$$\log\left\{1 - \beta_1(z_1 + z_1^{-1}) - \beta_2(z_2 + z_2^{-1}) - \gamma_1(z_1 z_2 + z_1^{-1} z_2^{-1}) - \gamma_2(z_1 z_2^{-1} + z_1^{-1} z_2)\right\}.$$

The method assumes that the covariance matrix B is derived from a stationary process. This assumption was criticized by Guyon (1982), who argued that neglecting edge effects was important in dimension $d \geq 2$. Guyon proposed an alternative scheme, also based on Whittle (1954), using estimates of the spectral density, but correcting those estimates for edge effects.

4.2.4 Simulated maximum likelihood

Suppose now we are faced with computing maximum likelihood estimates in a model with a joint density of form

$$p(\mathbf{x};\theta) = C(\theta)F(\mathbf{x};\theta) \tag{4.24}$$

where F is a known function of data values $\mathbf{x}$ in terms of unknown parameters θ, and $C(\theta)$ is a normalizing constant defined by the property that the sum or integral of p is 1, but not directly computable. The auto-logistic model, when written as (4.5), is precisely of this form, but so are a wide class of other models of similar structure.

To use (4.24) for maximum likelihood estimation, we need to be able to evaluate $C(\theta)$, or at any rate the ratio $C(\theta)/C(\theta_0)$ for any θ relative to some fixed reference value θ_0.

We now approximate this ratio through a simulation scheme, as follows. Fix θ_0 and let $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)}$ denote M simulated realizations from the stochastic process when θ_0 is the true parameter. For the moment, we do not consider how such simulations might be generated. Also let $\mathbf{X}$ denote the actual data which are observed.

We then have that

$$\frac{1}{M}\sum_{m=1}^{M} \frac{F(\mathbf{X}^{(m)};\theta)}{F(\mathbf{X}^{(m)};\theta_0)} \cdot \frac{F(\mathbf{X};\theta_0)}{F(\mathbf{X};\theta)} \tag{4.25}$$

is an unbiased estimate of

$$\frac{C(\theta_0)}{C(\theta)} \cdot \frac{F(\mathbf{X};\theta_0)}{F(\mathbf{X};\theta)},$$

in other words, the likelihood ratio of θ_0 to θ.

To see this, the key step is the calculation

$$\begin{aligned} \mathrm{E}_{\theta_0}\left\{\frac{F(\mathbf{X}^{(m)};\theta)}{F(\mathbf{X}^{(m)};\theta_0)}\right\} &= \sum_{\mathbf{x}} \frac{F(\mathbf{x};\theta)}{F(\mathbf{x};\theta_0)} \cdot C(\theta_0)F(\mathbf{x};\theta_0) \\ &= C(\theta_0)\sum_{\mathbf{x}} F(\mathbf{x};\theta) \\ &= \frac{C(\theta_0)}{C(\theta)}. \end{aligned} \tag{4.26}$$

In Monte Carlo maximum likelihood estimation (MCMLE), the usual procedure is to generate a single sequence $\mathbf{X}^{(1)},...,\mathbf{X}^{(M)}$ from some given θ_0, then to minimize the sum (4.25) analytically with respect to θ. The simulation procedure is more efficient the closer θ_0 is to the true MLE $\widehat{\theta}$, so sometimes the procedure is repeated several times, using the estimate from one minimization at the initial θ_0 for the next. One good use of the pseudolikelihood method (subsection 4.4.2) is to generate the initial θ_0.

We still have to describe how to generate the simulations $\{\mathbf{X}^{(m)}\}$ for given θ_0. For lattice models, the most convenient methods are of MCMC type, of which the two most widely used variants are the following:

Gibbs sampling. Start with arbitrary $\mathbf{x}^{(0)} = (x_1^{(0)},...,x_n^{(0)})$. Generate a new value of x_1, denoted $x_1^{(1)}$, from the conditional distribution of X_1 given $X_2 = x_2^{(0)},...,X_n = x_n^{(0)}$. Then generate a new value of x_2, denoted $x_2^{(1)}$, from the conditional distribution of X_2 given $X_1 = x_1^{(1)}, X_3 = x_3^{(0)},...,X_n = x_n^{(0)}$. Continue up to the generation of $x_n^{(1)}$ from the conditional distribution of X_n given $X_1 = x_1^{(1)},...,X_{n-1} = x_{n-1}^{(1)}$. This completes one ieration of the sampler. Then, starting from the new vector $\mathbf{x}^{(1)}$, return to x_1 and repeat the whole process to generate $\mathbf{x}^{(2)}$. Repeat many times.

The Hastings-Metropolis algorithm. Again we start with an arbitrary $\mathbf{x}^{(0)}$ and generate a new "trial value" $\mathbf{x}'$ from some distribution $q(\mathbf{x}';\mathbf{x}^{(0)})$ which depends on $\mathbf{x}^{(0)}$. Typically, but not necessarily, $\mathbf{x}'$ is formed from $\mathbf{x}^{(0)}$ by just changing one component. Then form the ratio

$$\alpha = \frac{q(\mathbf{x}^{(0)};\mathbf{x}')F(\mathbf{x}';\theta_0)}{q(\mathbf{x}';\mathbf{x}^{(0)})F(\mathbf{x}^{(0)};\theta_0)}.$$

If $\alpha \geq 1$ then we accept $\mathbf{x}'$; in other words, set $\mathbf{x}^{(1)} = \mathbf{x}'$. If $\alpha < 1$, we perform an independent random drawing: with probability α, accept $\mathbf{x}'$ and set $\mathbf{x}^{(1)} = \mathbf{x}'$; otherwise, reject $\mathbf{x}'$ and set $\mathbf{x}^{(1)} = \mathbf{x}^{(0)}$.

We shall not dwell on the details of either the Gibbs sampler or the Hastings- Metropolis procedure since they have by now been covered in many books and papers. Moreover, there are many variants on the basic procedure, such as those based on the Swendsen-Wang algorithm (Swendsen and Wang 1987). In practice it would be unusual to generate the samples $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(M)}$, required for the MCMLE procedure, in exactly the way that has just been described. For example, it is usual to discard some of the initial iterations as *warm-up iterations*, and once past the warm-up stage, it is usual to group the iterations together in *blocks* and only take the first or last member from each block.

The basic idea of MCMLE was apparently first stated by Penttinen (1984) and elaborated by Geyer and Thompson (1992). References on Gibbs sampling and Hastings-Metropolis methods include Hastings (1970), Geman and Geman (1984), Besag *et al.* (1991), Gelman and Rubin (1992), Geyer (1992), Besag and Green (1993), Smith and Roberts (1993), Tierney (1994), Carlin and Louis (1996) and Gilks *et al.* (1996).

4.2.5 Bayesian methods

Under the Bayesian approach, a prior density $\pi(\theta)$ is specified; Bayesian analysis then proceeds by computing a posterior density

$$\pi(\theta|\mathbf{X}) \propto p(\mathbf{X};\theta)\pi(\theta) \tag{4.27}$$

where the constant of proportionality is chosen so that when (4.27) is integrated respect to θ, the answer is 1.

In the case where both $p(\mathbf{X};\theta)$ and $\pi(\theta)$ are available in closed form, the problem may be solved by Monte Carlo sampling of θ. Both the Gibbs sampler and the Hastings-Metropolis algorithm, described in the previous subsection, may be adapted to this purpose, and are widely used.

Under a model of form (4.24), there is an additional complication: we cannot evaluate (4.27) directly because we do not know how to calculate $C(\theta)$. In that case, we must proceed indirectly, replacing (4.27) with

$$\frac{\pi(\theta|\mathbf{X})}{\pi(\theta_0|\mathbf{X})} \approx \frac{\pi(\theta)}{\pi(\theta_0)} \cdot \left\{ \frac{1}{M} \sum_{m=1}^{M} \frac{F(\mathbf{X}^{(m)};\theta)}{F(\mathbf{X}^{(m)};\theta_0)} \right\}^{-1} \cdot \frac{F(\mathbf{X};\theta)}{F(\mathbf{X};\theta_0)} \tag{4.28}$$

where $\{\mathbf{X}^{(m)},\ 1 \le m \le M\}$ form a sample of values from the distribution of $\mathbf{X}$ given $\theta = \theta_0$. Thus MCMC comes in twice, once to generate the $\{\mathbf{X}^{(m)}\}$ values in (4.28), and a second time to sample from the resulting approximate posterior density of θ.

4.3 Examples

Table 4.1 shows a famous data set due originally to Mercer and Hall (1911), but since re-analyzed by many other authors, including Whittle (1954), Besag (1974) and Cressie (1993). The data are derived from an agricultural field trial and show wheat yields on 500 plots arranged in a 20×25 array. The left-hand 12 columns are given in the top half of the table, and the right-hand 13 columns in the bottom half.

3.63	4.15	4.06	5.13	3.04	4.48	4.75	4.04	4.14	4.00	4.37	4.02	
4.07	4.21	4.15	4.64	4.03	3.74	4.56	4.27	4.03	4.50	3.97	4.19	
4.51	4.29	4.40	4.69	3.77	4.46	4.76	3.76	3.30	3.67	3.94	4.07	
3.90	4.64	4.05	4.04	3.49	3.91	4.52	4.52	3.05	4.59	4.01	3.34	
3.63	4.27	4.92	4.64	3.76	4.10	4.40	4.17	3.67	5.07	3.83	3.63	
3.16	3.55	4.08	4.73	3.61	3.66	4.39	3.84	4.26	4.36	3.79	4.09	
3.18	3.50	4.23	4.39	3.28	3.56	4.94	4.06	4.32	4.86	3.96	3.74	
3.42	3.35	4.07	4.66	3.72	3.84	4.44	3.40	4.07	4.93	3.93	3.04	
3.97	3.61	4.67	4.49	3.75	4.11	4.64	2.99	4.37	5.02	3.56	3.59	
3.40	3.71	4.27	4.42	4.13	4.20	4.66	3.61	3.99	4.44	3.86	3.99	
3.39	3.64	3.84	4.51	4.01	4.21	4.77	3.95	4.17	4.39	4.17	4.17	
4.43	3.70	3.82	4.45	3.59	4.37	4.45	4.08	3.72	4.56	4.10	3.07	
4.52	3.79	4.41	4.57	3.94	4.47	4.42	3.92	3.86	4.77	4.99	3.91	
4.46	4.09	4.39	4.31	4.29	4.47	4.37	3.44	3.82	4.63	4.36	3.79	
3.46	4.42	4.29	4.08	3.96	3.96	3.89	4.11	3.73	4.03	4.09	3.82	
5.13	3.89	4.26	4.32	3.78	3.54	4.27	4.12	4.13	4.47	3.41	3.55	
4.23	3.87	4.23	4.58	3.19	3.49	3.91	4.41	4.21	4.61	4.27	4.06	
4.38	4.12	4.39	3.92	4.84	3.94	4.38	4.24	3.96	4.29	4.52	4.19	
3.85	4.28	4.69	5.16	4.46	4.41	4.68	4.37	4.15	4.91	4.68	5.13	
3.61	4.22	4.42	5.09	3.66	4.22	4.06	3.97	3.89	4.46	4.44	4.52	
4.58	3.92	3.64	3.66	3.57	3.51	4.27	3.72	3.36	3.17	2.97	4.23	4.53
4.05	3.97	3.61	3.82	3.44	3.92	4.26	4.36	3.69	3.53	3.14	4.09	3.94
3.73	4.58	3.64	4.07	3.44	3.53	4.20	4.31	4.33	3.66	3.59	3.97	4.38
4.06	3.19	3.75	4.54	3.97	3.77	4.30	4.10	3.81	3.89	3.32	3.46	3.64
3.74	4.14	3.70	3.92	3.79	4.29	4.22	3.74	3.55	3.67	3.57	3.96	4.31
3.72	3.76	3.37	4.01	3.87	4.35	4.24	3.58	4.20	3.94	4.24	3.75	4.29
4.33	3.77	3.71	4.59	3.97	4.38	3.81	4.06	3.42	3.05	3.44	2.78	3.44
3.72	3.93	3.71	4.76	3.83	3.71	3.54	3.66	3.95	3.84	3.76	3.47	4.24
4.05	3.96	3.75	4.73	4.24	4.21	3.85	4.41	4.21	3.63	4.17	3.44	4.55
3.37	3.47	3.09	4.20	4.09	4.07	4.09	3.95	4.08	4.03	3.97	2.84	3.91
4.09	3.29	3.37	3.74	3.41	3.86	4.36	4.54	4.24	4.08	3.89	3.47	3.29
3.99	3.14	4.86	4.36	3.51	3.47	3.94	4.47	4.11	3.97	4.07	3.56	3.83
4.09	3.05	3.39	3.60	4.13	3.89	3.67	4.54	4.11	4.58	4.02	3.93	4.33
3.56	3.29	3.64	3.60	3.19	3.80	3.72	3.91	3.35	4.11	4.39	3.47	3.93
3.57	3.43	3.73	3.39	3.08	3.48	3.05	3.65	3.71	3.25	3.69	3.43	3.38
3.16	3.47	3.30	3.39	2.92	3.23	3.25	3.86	3.22	3.69	3.80	3.79	3.63
3.75	3.91	3.51	3.45	3.05	3.68	3.52	3.91	3.87	3.87	4.21	3.68	4.06
4.49	3.82	3.60	3.14	2.73	3.09	3.66	3.77	3.48	3.76	3.69	3.84	3.67
4.19	4.41	3.54	3.01	2.85	3.36	3.85	4.15	3.93	3.91	4.33	4.21	4.19
3.70	4.28	3.24	3.29	3.48	3.49	3.68	3.36	3.71	3.54	3.59	3.76	3.36

Table 4.1. Mercer-Hall data, first 12 columns (top) and last 13 (bottom).

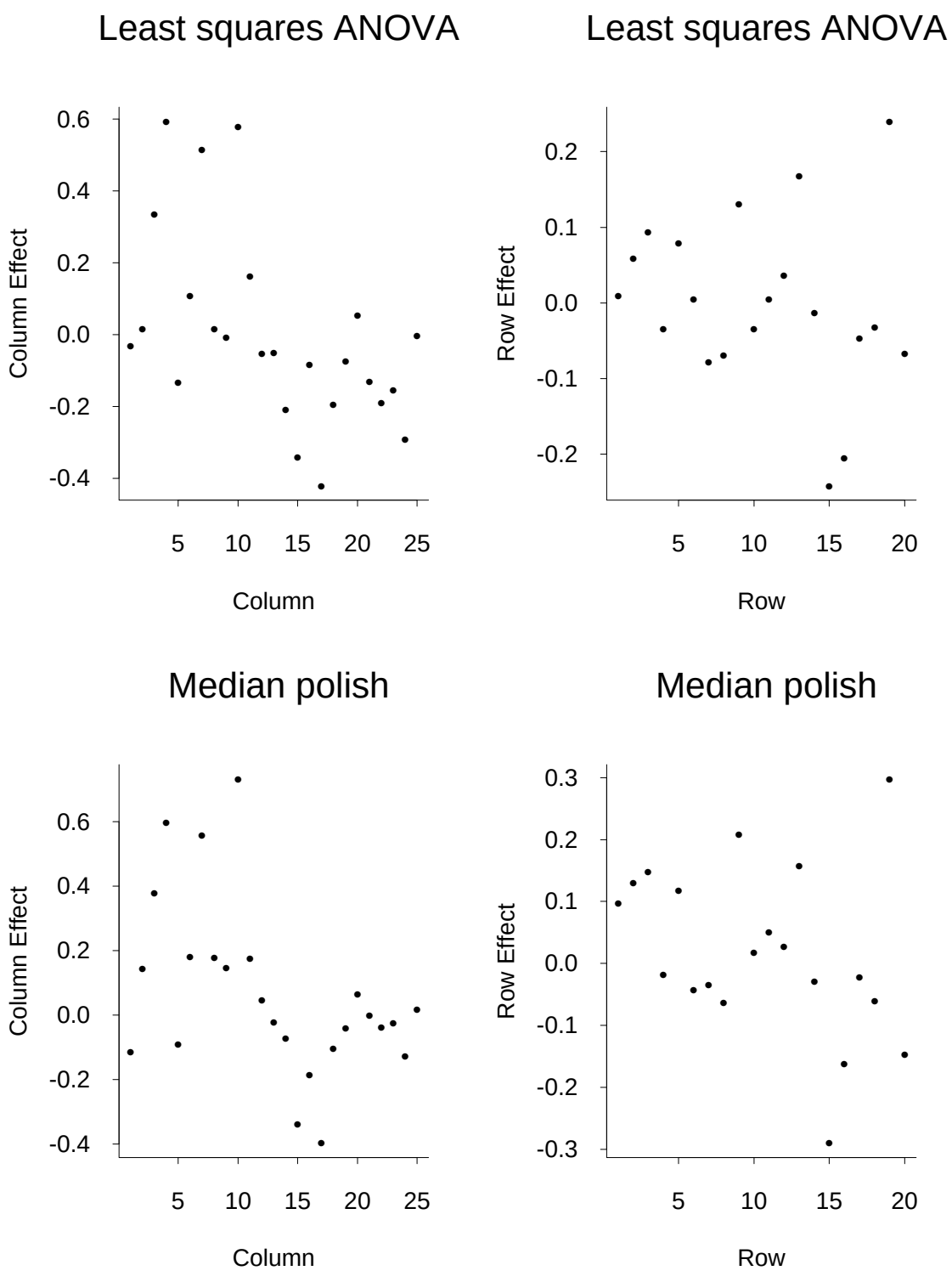


Fig. 4.2. Row and columns effects for Mercer-Hall data, computed by least-squares ANOVA (top plots) and by median polish (bottom plots)

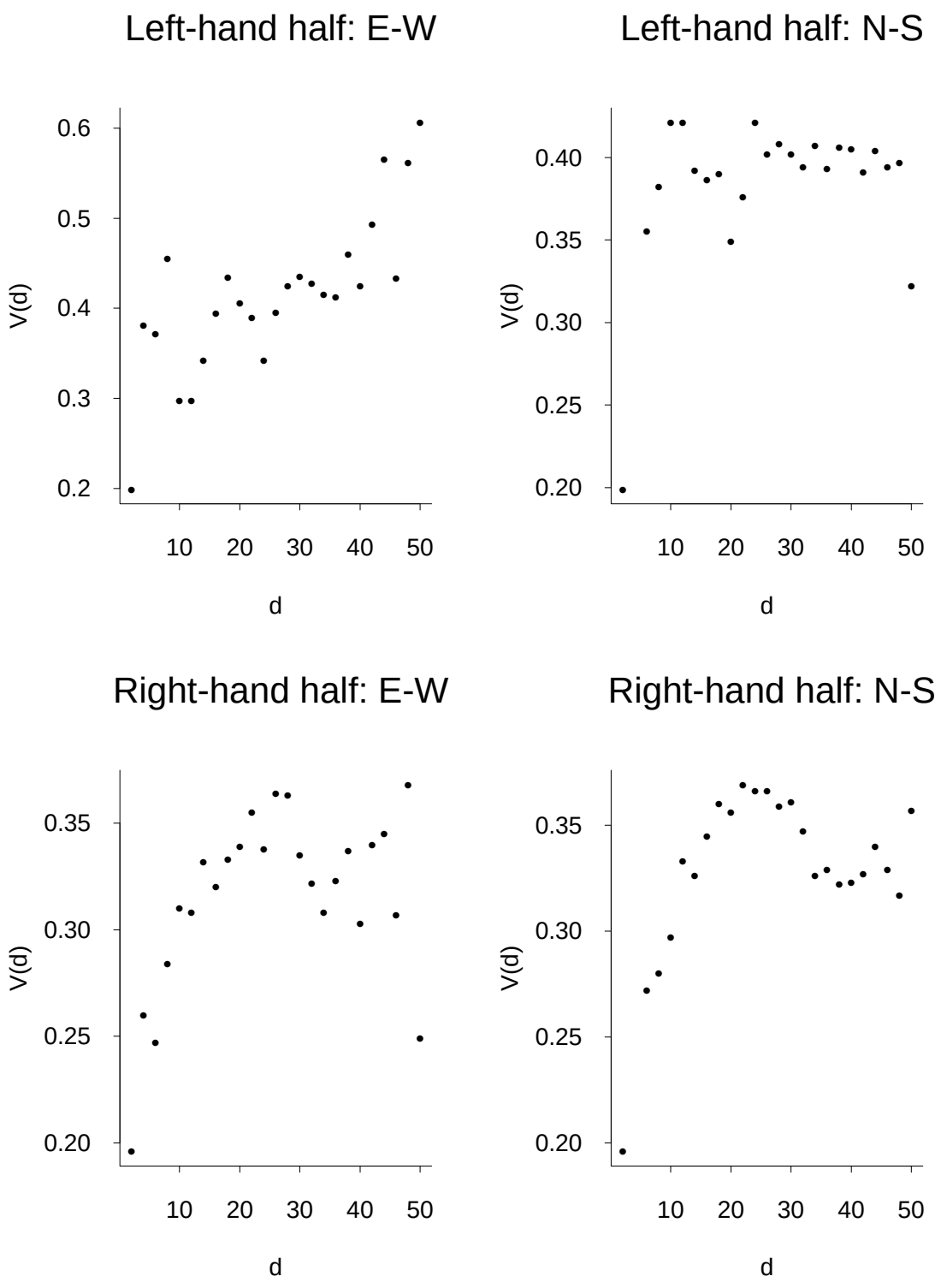


Fig. 4.3. Variogram estimates for Mercer-Hall data, computed separately for E-W and N-S directions, and for the left-hand and right-hand halves of the data.

Some preliminary analysis suggests that the data do not come from a stationary process. For example, Fig. 4.2 shows estimated row and column effects from a two-way analysis of variance, computed both by standard least squares ANOVA (top half of plot) and by the robust median polish method which was described in section 2.4 (bottom half). Both sets of estimates show that the column effects are very variable in the left-hand half of the plot, but much less so in the right-hand half. There is also some suggestion of variability in the row effects, with more inter-row variation in the later rows, but this is not nearly so marked. This suggests that, as an alternative to analyzing the whole data set as a single homogeneous model, we should consider the left-hand and right-hand halves separately, and this has been done, with the split exactly as in Table 4.1. In Fig. 4.3, variogram estimates are plotted, computed separately in the E-W and N-S directions, and for the left-hand and right-hand halves of the data. The plots confirm the greater variability of the left-hand half of the plot in the E-W direction.

	$\widehat{\beta}_1$	$\widehat{\beta}_2$
Full data set:		
Coding, first analysis	0.332	0.128
Coding, second analysis	0.354	0.166
S.E.	0.03	0.03
Whittle method	0.368	0.107
MLE	0.364	0.114
S.E.	0.024	0.025
Left-hand half:		
MLE	0.400	0.000
S.E.	0.029	0.033
Right-hand half:		
MLE	0.275	0.191
S.E.	0.041	0.043

Table 4.2. Estimates for model (4.19) in Mercer-Hall data, using both coding and Whittle methods (quoted from Besag, 1974) and by exact MLE, the latter being computed both for the full data set and spearately for the left and right halves.

In Table 4.2, estimates of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ in model (4.19) are shown, taken from Besag's (1974) results for the coding and Whittle methods, and the present author's calculations of exact MLE. Two sets of coding estimates are given, one for "odd" sites and the other for "even" sites. In Table 4.3, similar calculations are given for model (4.20). In this case there are four different configurations for the coding estimates, as described in detail in Besag's original paper.

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$
Full data set:				
Coding, first analysis	0.344	0.043	0.079	–0.062
Coding, second analysis	0.318	0.085	0.016	0.011
Coding, third analysis	0.407	0.243	–0.067	–0.034
Coding, fourth analysis	0.361	0.236	–0.092	–0.041
S.E.	0.05	0.06	0.07	0.06
Whittle method	0.381	0.160	–0.015	–0.056
MLE	0.380	0.171	–0.020	–0.060
S.E.	0.024	0.046	0.037	0.035
Left-hand half:				
MLE	0.400	0.019	0.000	–0.025
S.E.	0.029	0.075	0.055	0.055
Right-hand half:				
MLE	0.329	0.274	–0.042	–0.116
S.E.	0.042	0.053	0.051	0.049

Table 4.3. Estimates for model (4.20) in Mercer-Hall data, similar to Table 4.2.

It can be seen that for analyses based on the full data set, Whittle's estimates are in all cases very close to the exact MLEs — the coding estimates appear to be more variable. However there are also significant differences between the left and right halves of the data, especially in the parameter β_2 (as expected).

One advantage of exact maximum likelihood is that it enables us to test directly between nested models, using likelihood ratio tests. Under the full data set, the likelihood ratio statistic for testing model (4.19) against model (4.20) is 2.86, with 2 degrees of

freedom, a P-value of 0.24. For the left and right halves of the data, the corresponding statistics are 0.22 (P=0.90), 5.19 (P=0.07). Only the last of these comes close to being statistically significant, but overall there is no reason to reject model (4.19).

In summary, for a data set of this size, given present-day computing speeds, I would recommend using an exact maximum likelihood procedure, which has the advantage of permitting direct calculations of standard errors and likelihood ratio test statistics. In the present data set, the main feature is the apparent nonstationarity which has been dealt with by the rather *ad hoc* procedure of splitting the data into two halves. This procedure indeed confirms that the two halves are statistically different, but of course it does not show that the stationary model is necessarily applicable to either half of the data; the best one can say is that there is no obvious evidence to contradict such an assumption.

CHAPTER 6
Design of a Monitoring Network

In previous chapters, we have seen how a spatially distributed quantity may be estimated from a finite set of monitoring stations, and we have characterized the error in such estimation procedures. In this chapter, we consider the design of the network: assuming that the position of the monitoring stations is, at least to some extent, under the control of the agency responsible for the measurements, where should they be placed in order to maximize the usefulness of the network? This problem is faced every time an organization such as the U.S.E.P.A. is charged with collecting data on a new pollutant, and it also arises in many other contexts, e.g. the positioning of meteorological stations.

In practice, many considerations based on cost, political influence, geographical convenience, and so on, may influence the design of a network. So it is inevitable that mathematical formulations of optimal design are only one part, possibly not very significant, of the decision-making process. Nevertheless, even when economic and political constraints are taken into account, the monitoring agency usually has a considerable amount of discretion, and it makes sense to exercise that discretion in a way that maximizes the information that is gained from the network — however that is defined. Therefore, in this chapter we shall concentrate on mathematical formulations of the problem, and will discuss computational approaches to the solutions of such problems.

Mathematical formulations of this problem have generally followed one of two broad strategies, though there are several more *ad hoc* approaches. The two broad strategies may be characterized as the maximum entropy approach and the optimal design approach. The maximum entropy criterion may be derived from Bayesian considerations (section 6.1) but is also sometime intuitively justified as a measure of the information in an experiment, or equivalently, the reduction in uncertainty of a variable of interest as the result of performing an experiment. This approach has been particularly developed by Zidek and co-workers, e.g. Caselton and Zidek (1984), Caselton, Kan and Zidek (1992), Zidek, Sun and Le (2000). After some mathematical preliminaries in section 6.2, we develop this approach in detail in section 6.3. The second approach to network design derives from the theory of optimal design of experiments, including such concepts as D-optimality and A-optimality, which we review in sectoin 6.4. This theory was first applied in a network design context by Fedorov and Müller (1989). However, the direct "message" of that paper, which was an attempt to show how optimal designs developed for regression applications could be directly applied in spatial analysis contexts, did not hold up as researchers started to look at more realistic models. Nevertheless, there has been extensive recent research on how classical design criteria could be applied in the context of spatial models which are typical in environmental applications. An excellent recent book on this topic is Müller (2000), and we review this whole area in section 6.5.

Besides the two mathematical formulations, numerous more *ad hoc* approaches have been proposed, motivated by specific applications such as the United States National Atmospheric Deposition Program/National Trends Network (NADP/NTN) which was studied extensively in the late 1980s and early 1990s as part of the effort to reduce acid rain over North America, and more recently in connection with EPA networks for ozone and particulate matter. Some of these *ad hoc* methods have been developed considerably further down the road of practical application than the more theoretical approaches. We review these methods in section 6.6.

A different but related problem is that of "data assimilation". This refers to a large class of methods, mostly developed by atmospheric scientists, that involve integrating real data into a numerical model. The canonical example is that of short-term weather forecasting, where numerical models of the weather are constantly evolving but must incorporate real data from observing networks when it becomes available. Sometimes, the weather forecaster has discretion over where to take observations — for instance, an observational aircraft may be due to fly through part of the weather system, but the route it takes may be chosen to maximimize the usefulness of the observations it gets. In this context, there is a question of determining the optimal route. Some recent work in this problem is reviewed in section 6.7.

Finally, section 6.8 presents a summary of the chapter and some overall conclusions.

6.1. A Bayesian formulation of optimal design

An early discussion of the design of experiments from a Bayesian point of view was given by Lindley (1956). Lindley proposed a criterion which amounts to maximization of the expected Shannon information to be gained from an experiment, when the objective "is not to reach decisions but rather to gain knowledge about the world". Later Bernardo (1979) showed that this criterion can also be derived from a Bayesian decision-theoretic point of view, provided one makes certain "reasonable" assumptions about the form of the utility function. In this initial section, we shall outline Bernardo's argument.

Suppose we are considering conducting an experiment E which will yield data X distributed according to a model

$$X \sim p(\cdot \mid \theta).$$

Here $p(\cdot \mid \theta)$ is a known distributional family depending on the true value θ of a random unknown parameter Θ. Note that, as in all Bayesian formulations, we are treating any unknown parameter as the realization of a random variable rather than as a fixed constant. Also suppose that the object of interest is a function of Θ, denoted Ψ or $\Psi(\Theta)$, with a true value denoted by ψ. Suppose the prior density of Θ is $\pi_\Theta(\theta)$ and that of Ψ is $\pi_\Psi(\psi)$. The marginal density of X will be denoted $p_X(x) = \int p(x \mid \theta)\pi_\Theta(\theta)d\theta$. Since Θ will not enter the following discussion except through Ψ, henceforth we write $\pi(\cdot)$ instead of $\pi_\Psi(\cdot)$. The posterior density of Ψ given data $X = x$, evaluated at $\Psi = \psi$, will be $\pi(\psi \mid x)$.

Bernardo (1979), following earlier work by Lindley (1956), proposed the following measure of the information contained in E when the prior density is $\pi(\cdot)$:

$$I\{E,\pi\} = \int p_X(x) \int \pi(\psi \mid x) \log \frac{\pi(\psi \mid x)}{\pi(\psi)} d\psi dx. \tag{6.1}$$

Given that the inner integral is essentially an information divergence between the prior and posterior densities, the expression (6.1) may be interpreted as an "expected gain of information" which results from performing the experiment E.

In a monitoring network context, if we let E denote a decision to select a specific set of monitoring sites, X as the data generated by such a network, and ψ as a specific parameter or random variable of interest, then (6.1) suggests a criterion which we may use to discriminate among different proposals for E, with the "best" network being the one which maximizes $I\{E,\pi\}$ for some given prior density π. In section 6.2, we shall see how some authors have used this in the actual design of a network.

In the remainder of the present section, we shall outline Bernardo's elegant derivation of (6.1) from some basic principles of Bayesian decision theory.

Decision-theoretic formulation

According to the Bayesian view of the world, Ψ is a random variable, and the outcome of the experiment E may be represented by the statistician's reporting a probability distribution, $\pi^\dagger(\psi)$, to represent her "belief" about ψ after conducting the experiment. The dagger $\dagger$ here may be thought of as a symbol representing "reported".

In a decision-theoretic formulation of the problem, there will be a utility function, $u(\pi^\dagger(\cdot), \psi)$, which represents the gain in reporting a density $\pi^\dagger$ when the true value of Ψ is ψ.

Suppose we perform an experiment producing data x, and let $\pi(\cdot \mid x)$ denote the posterior distribution of Ψ given x. The expected utility is then

$$\int u(\pi^\dagger(\cdot), \psi)\pi(\psi \mid x) d\psi. \tag{6.2}$$

We suppose that the utility function possesses the following properties:

(a) u is *proper* if (6.2) is maximized over all probability distributions $\pi^\dagger$ by setting $\pi^\dagger(\psi) = \pi(\psi \mid x)$.

This captures the natural (for a Bayesian) property that the optimal solution to the decision problem is the posterior density. Another way to think of it is that this condition guarantees the coherence of the procedure.

(b) u is *local* if $u(\pi^\dagger(\cdot), \psi)$ depends on the function $\pi^\dagger(\cdot)$ only through its value at ψ, i.e. $\pi^\dagger(\psi)$.

This says that the utility function at a particular value of ψ should not depend on other values of Ψ which have not occurred. As pointed out by Bernardo in his paper, this is reminiscent of the likelihood principle in the theory of statistical inference.

Bernardo's theorem says the following: *If u is proper and local, then it must be of the form*

$$u(\pi^\dagger(\cdot), \psi) = A \log \pi^\dagger(\psi) + B(\psi), \tag{6.3}$$

where A is constant and B is a function of ψ alone.

Sketch of proof of (6.3):

Since u is local, $u(\pi^\dagger(\cdot), \psi) = u(\pi^\dagger(\psi), \psi)$. Therefore, the property that u is proper becomes: *Among all functions $\pi^\dagger(\cdot)$, and for all probability densities $\pi(\cdot)$, the value of $\int u(\pi^\dagger(\psi), \psi)\pi(\psi)d\psi$, subject to $\int \pi^\dagger(\psi)d\psi = 1$, is maximized when $\pi^\dagger(\cdot) = \pi(\cdot)$.*

By the principle of Lagrange multipliers, the optimal $\pi^\dagger(\cdot)$ is an extreme point of the functional

$$F\{\pi^\dagger(\cdot)\} = \int u(\pi^\dagger(\psi), \psi)\pi(\psi)d\psi - A\left[\int \pi^\dagger(\psi)d\psi - 1\right] \tag{6.4}$$

for some constant A.

However, for this to be achieved, we must have

$$F_1\{\pi^\dagger(\cdot)\} \equiv \frac{\partial}{\partial \epsilon} F\{\pi^\dagger(\cdot) + \epsilon\tau(\cdot)\}|_{\epsilon=0} = 0 \tag{6.5}$$

for all sufficient smooth and small functions τ. To see that (6.5) must be true, expand (6.4) in a Taylor series for small ϵ and fixed τ: $F\{\pi^\dagger(\cdot)+\epsilon\tau(\cdot)\} = F\{\pi^\dagger(\cdot)\}+\epsilon F_1\{\pi^\dagger(\cdot)\}\tau(\cdot)+o(\epsilon)$. If $F_1 \neq 0$, the second term can be made either positive or negative for sufficiently small $|\epsilon|$, so $\pi^\dagger$ cannot be an extreme value of the functional F. Therefore $F_1 = 0$, which is (6.5).

Evaluating the derivative in (6.5), this condition reduces to

$$\int u_1(\pi^\dagger(\psi), \psi)\pi(\psi)\tau(\psi)d\psi - A\int \tau(\psi)d\psi = 0 \tag{6.6}$$

where u_1 denotes the first-order partial derivative of u with respect to its first argument.

However, for (6.6) to be true for all τ in a sufficiently broad class, we must have

$$u_1(\pi^\dagger(\psi), \psi)\pi(\psi) = A \tag{6.7}$$

for each ψ.

For u to be a proper utility function, (6.7) must be true when $\pi^\dagger(\psi) = \pi(\psi)$, and so

$$u_1(\pi(\psi), \psi)\pi(\psi) = A \tag{6.8}$$

for each ψ.

But (6.8) is just a differential equation in the (scalar) quantity $\pi(\psi)$, and it may easily be checked that this has solution

$$u(\pi(\psi), \psi) = A \log \pi(\psi) + B(\psi)$$

where $B(\psi)$ is a constant depending only on ψ. This is (6.3).

Interpretation

The expected utility before doing the experiment is

$$\int \{A \log \pi(\psi) + B(\psi)\}\pi(\psi)d\psi, \tag{6.9}$$

where $\pi(\psi)$ is the prior distribution.

The expected value of the utility after doing the experiment is

$$\int \left\{ \int \{A \log \pi(\psi \mid x) + B(\psi)\}\pi(\psi \mid x)d\psi \right\} p_X(x)dx, \tag{6.10}$$

where $\pi(\psi \mid x)$ is the posterior distribution given $X = x$ and $p_X(x)$ is the marginal distribution of X. The difference between (6.10) and (6.9) is therefore the expected gain in utility as a result of doing the experiment.

However, the term involving $B(\psi)$ is the same in (6.9) and (6.10) — this follows at once from the elementary identity $\pi(\psi) = \int \pi(\psi \mid x)p_X(x)dx$.

Ignoring $B(\psi)$, then, the difference between (6.10) and (6.9) is

$$A \int \int [\log\{\pi(\psi \mid x)\}\pi(\psi \mid x) - \log\{\pi(\psi)\}\pi(\psi)]\, d\psi p_X(x)dx. \tag{6.11}$$

However, let us also note the identity

$$A \int \int \log \pi(\psi)\, \{\pi(\psi) - \pi(\psi \mid x)\}\, d\psi p_X(x)dx = 0, \tag{6.12}$$

which follows, after exchanging the order of integrals, because $\int \pi(\psi \mid x) p_X(x) dx = \pi(\psi)$ for each ψ.

Substituting from (6.12), (6.11) becomes

$$A \int \int [\log\{\pi(\psi \mid x)\} - \log\{\pi(\psi)\}] \, d\psi \pi(\psi \mid x) p_X(x) dx,$$

which is (6.1).

The alternative representation (6.11) is of interest in itself, since it may be interpreted as the expected information after the experiment minus the information before the experiment, and this is another way of thinking about the purpose of the experimental design as "maximizing the information contained in the experiment".

6.2 Information in the multivariate normal and t distributions

As a technical preliminary to the results of section 6.3, we show how to calculate the information in the multivariate normal and t distributions. We refer back to section 3.4 for the basic distribution theory.

Suppose $X \sim N_p(\mu, \Sigma)$. The density of X, evaluated at $X = x$, is given by (2.X1) and may be denoted $f(x)$. Then

$$\log f(X) = -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu).$$

But $\mathrm{E}\{(X - \mu)^T \Sigma^{-1} (X - \mu)\} = p$, so

$$I(X) = \mathrm{E}\{\log X\} = -\frac{p}{2} \{1 + \log(2\pi)\} - \frac{1}{2} \log |\Sigma|. \tag{6.13}$$

In the case of the t distribution, we shall actually work with the matrix t distribution discussed in section 3.4.1, though for the present purposes, we shall only need the multivariate form of it.

Suppose $T \sim t(p, q; P, Q, m)$ and let $f(t)$ denote the density at $T = t$. By (2.X11),

$$\begin{aligned} \log f(t) = &-\frac{pq}{2} \log \pi + \log \Gamma_q \left(\frac{m + p + q - 1}{2} \right) - \log \Gamma_q \left(\frac{m + q - 1}{2} \right) \\ &+ \frac{m + p + q}{2} \log |Q| + \frac{q}{2} \log |P| - \frac{m + p + q - 1}{2} \log |Q + t^T P t|. \end{aligned}$$

We evaluate $\mathrm{E}\{\log |Q + t^T P t|\}$ as the limit of $\frac{d}{dr} \mathrm{E}\{|Q + t^T P t|^r\}$ as $r \to 0$. From the fact that (2.X11) is a density, we deduce

$$\int |Q + t^T P t|^{-(m+p+q-1)/2} dt = \pi^{pq/2} \frac{\Gamma_q((m + q - 1)/2)}{\Gamma_q((m + p + q - 1)/2)} |Q|^{-((m+q-1)/2)} |P|^{-q/2},$$

and so, replacing m by $m-2r$,

$$\int |Q+t^TPt|^{-(m-2r+p+q-1)/2}dt$$
$$= \pi^{pq/2}\frac{\Gamma_q((m-2r+q-1)/2)}{\Gamma_q((m-2r+p+q-1)/2)}|Q|^{-((m-2r+q-1)/2)}|P|^{-q/2}.$$

Taking the ratio of the last two expressions,

$$\mathrm{E}|Q+T^TPT|^r = \frac{\Gamma_q((m-2r+q-1)/2)}{\Gamma_q((m-2r+p+q-1)/2)}\cdot\frac{\Gamma_q((m+q-1)/2)}{\Gamma_q((m+p+q-1)/2)}\cdot|Q|^r.$$

Therefore,

$$\mathrm{E}\log|Q+T^TPT| = \frac{d}{dr}\Big\{\log\Gamma_q((m-2r+q-1)/2)$$
$$-\log\Gamma_q((m-2r+p+q-1)/2)+r\log|Q|\Big\}_{r=0}.$$

However,

$$\log\Gamma_q\left(\frac{m}{2}\right) = \frac{q(q-1)}{4}\log\pi+\sum_{j=1}^{q}\log\Gamma\left(\frac{m+1-j}{2}\right)$$

by (2.X9), and since $\frac{d}{dx}\log\Gamma(x)=\psi(x)$, where $\psi(\cdot)$ is the digamma function, we have

$$\frac{d}{dr}\log\Gamma_q\left(\frac{m-2r}{2}\right) = -\sum_{j=1}^{q}\psi\left(\frac{m-2r+1-j}{2}\right).$$

Hence

$$\mathrm{E}\log|Q+T^TPT| = -\sum_{j=1}^{q}\psi\left(\frac{m+q-j}{2}\right)+\sum_{j=1}^{q}\psi\left(\frac{m+p+q-j}{2}\right)+\log|Q|.$$

Finally, we deduce that when $T\sim t(p,q;P,Q,m)$, the information in T is

$$I(T) = -\frac{pq}{2}\log\pi+\log\left\{\frac{\Gamma_q((m+p+q-1)/2)}{\Gamma_q((m+q-1)/2)}\right\}-\frac{p}{2}\log|Q|-\frac{q}{2}\log|P|$$
$$-\frac{m+p+q-1}{2}\sum_{j=1}^{q}\left\{\psi\left(\frac{m+p+q-j}{2}\right)-\psi\left(\frac{m+q-j}{2}\right)\right\}. \tag{6.14}$$

We leave it as an exercise to show that one would get exactly the same answer starting from (2.X12).

6.3 Information- and entropy-based criteria of optimal design

In this section we outline the development of a theory of optimal design based on information or entropy criteria, concentrating on the work of Zidek and co-authors.

6.3.1 First formulation: Caselton and Zidek (1984)

Caselton and Zidek (1984) considered the problem of dividing $p = u + g$ sites into two subsets, where g sites will be gauged and the rest ungauged. Suppose the p-dimensional random vector Y represents the values of the random field at all p sites, but this is subdivided into $Y^{(1)}$ and $Y^{(2)}$ corresponding to the ungauged and gauged sites respectively. We assume that Y has a multivariate normal distribution with mean μ and covariance matrix Σ where μ and Σ are known. Thus, the method takes account of the prediction uncertainty of $Y^{(1)}$ given $Y^{(2)}$ but not of any additional uncertainty arising from the estimation of μ and Σ.

In the context of section 6.1, we identify ψ with $Y^{(1)}$, the variable we are trying to predict (since this is a Bayesian theory, we do not in any case make any distinction between parameters and random variables), and X with $Y^{(2)}$, the measured data. Then according to (6.1) or its equivalent form (6.11), the information gained about $Y^{(1)}$ as a result of measuring $Y^{(2)}$ may be written in the form

$$I(Y^{(1)}|Y^{(2)}) - I(Y^{(1)}) \tag{6.15}$$

where $I(X)$ denotes the information in a random variable X, i.e. $\int f(x) \log f(x) dx$ where $f(x)$ is the density of X evaluated at x.

The conditional distribution of $Y^{(1)}$ given $Y^{(2)}$ is normal with variance $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, while the unconditional variance of $Y^{(1)}$ is Σ_{11}. Therefore, according to (6.13),

$$I(Y^{(1)}|Y^{(2)}) - I(Y^{(1)}) = -\frac{1}{2}\log|\Sigma_{1|2}| + \frac{1}{2}\log|\Sigma_{11}|.$$

The optimal design, according to this criterion, is the one which minimizes

$$\begin{aligned} \log|\Sigma_{1|2}| - \log|\Sigma_{11}| &= \log|\Sigma_{11}^{-1/2}\Sigma_{1|2}\Sigma_{11}^{-1/2}| \\ &= \log|\Sigma_{11}^{-1/2}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\Sigma_{11}^{-1/2}| \\ &= \log|I - R| \end{aligned} \tag{6.16}$$

where $R = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$.

Another way to write (6.16) is

$$\sum \log(1 - \rho_i^2), \tag{6.17}$$

where $\rho_1^2, \rho_2^2, \ldots$ are the eigenvalues of R or, equivalently, the squared *canonical correlations* between U and G. Therefore, another way to express the criterion is to choose the subdivision that makes the canonical correlations between U and G large, in the precise sense of minimizing (6.17).

6.3.2 Second formulation: Caselton, Kan and Zidek (1992)

Caselton, Kan and Zidek (1992) — henceforth CKZ — extended the theory of Caselton and Zidek (1984) (CZ) to consider the case where μ and Σ are *a priori* unknown. However, at the same time they also considered a number of different formulations of the problem, in particular one which leads to a different selection criterion from CZ.

CKZ formulated the design problem in terms of entropy, where, following Jaynes (1963), the entropy in a random variable Y with density $f(y)$ is defined as

$$H(Y) = \mathrm{E}\left\{ -\log \frac{f(Y)}{m(Y)} \right\}, \tag{6.18}$$

where m is an appropriate reference measure taken to be an invariant measure by CKZ. In the subsequent development the role of the measure m is ignored, though CKZ admitted that its arbitrariness was a slightly awkward feature of the theory. If we do ignore m, then the entropy of a random variable is simply the negative of the information as defined in section 6.1; "entropy" is also commonly equated to "uncertainty". Note that the theory of section 6.1 also involves an unspecified additive constant, deriving from $B(\psi)$ in (6.9), but since Bernardo's criterion is based solely on *differences* of information between two distributions for the same random variable, the unspecified additive constant does not affect the conclusions in that case.

Using obvious notation, we write $H(U)$ for the entropy in the ungauged stations U, $H(U|G)$ for the conditional entropy given G (i.e. the entropy in the conditional distribution of the measurements in U given the measurements in G), and so on. In this notation, the CZ criterion is the same as choosing U to minimizie

$$H(U|G) - H(U). \tag{6.19}$$

However, CKZ criticized (6.19) as ignoring the information in G: it might be appropriate if the number of stations in G was negligible in comparison to U but in the specific application of CKZ — which involved reducing an existing network of 81 stations to some smaller number — it was important to consider the information in the gauged stations themselves as well as their role in predicting values at the ungauged stations. Noting that one can decompose the total entropy as

$$H(U, G) = H(U|G) + H(G), \tag{6.20}$$

where the left-hand side of (6.20) represent the information in the entire system and is therefore constant, they proposed the criterion of minimizing $H(U|G)$ or equivalently:

$$\text{Choose } G \text{ to maximize } H(G). \tag{6.21}$$

Note that in the case of a multivariate normal distribution, (6.20) is equivalent to the decomposition

$$\frac{1}{2}\log|\Sigma| = \frac{1}{2}\log|\Sigma_{1|2}| + \frac{1}{2}\log|\Sigma_{22}|,$$

which is a well-known identity for the determinant of a partitioned matrix (Mardia, Kent and Bibby 1979, section A.2.3, page 457). Thus, criterion (6.21), applied to the model of CZ, would lead one to choose the gauged set G which maximized Σ_{22}.

In fact, CKZ considered a much more complicated problem, in which μ and Σ are initially unknown and estimated through a Bayesian scheme as in section 3.4. For the development given in that paper, they assumed that previous observations are available from the entire system, i.e. U and G. In many situations, that would not be a realistic assumption, though for the specific applied problem considered by CKZ — that of minimizing an existing network — it is realistic. They did have complete measurements for all 81 stations in the network, though as we shall see, they still had difficulties associated with not having enough data to estimate the model satisfactorily. In any case, we present CKZ's mathematical development here because it leads to a very elegant conclusion, which should surely be used as the basis for more general developments under alternative model and data assumptions.

Following the Bayesian scheme outlined in section 3.4.2, CKZ assumed the prior distribution

$$\begin{aligned}\Sigma &\sim W_p^{-1}(\Psi, m),\\ \mu|\Sigma &\sim N_p(\mu^0, f^{-1}\Sigma),\end{aligned}$$

where the matrix Ψ, the vector μ^0 and the scalars m and f are hyperparameters. The situation is slightly simpler than section 3.4.2 because we are not considering regression coefficients, but the same theory obviously applies to the present setting.

Suppose we have n complete data vectors $y_1, ..., y_n$, each independently sampled from $N_p(\mu, \Sigma)$, and write Y for the complete data vector. As shown in section 3.4.2, the posterior distribution given Y is of the form

$$\begin{aligned}\Sigma &\sim W_p^{-1}(\hat{\Psi}, \hat{m}),\\ \mu|\Sigma &\sim N_p(\hat{\mu}^0, \hat{f}^{-1}\Sigma),\end{aligned} \tag{6.22}$$

where, for instance, $\hat{m} = m + n$, $\hat{f} = f + n$ and $\hat{\mu}^0$ and $\hat{\Psi}$ are the same as B^* and $\Psi + H$ in the notation of (2.X30).

Now consider a hypothetical future observation $y^{(2)^*}$, where by the notation we mean that the future data are observed only on the gauged portion of the network. According to the $H(G)$ criterion, the objective is to choose G so that the entropy in the predictive distribution of $y^{(2)^*}$ is maximized.

To calculate the entropy in the conditional distribution, we note first that

$$\begin{aligned} \Sigma_{22}|Y &\sim W_g^{-1}(\hat{\Psi}_{22}, \hat{m}-u), \\ \mu_2|\Sigma, Y &\sim N_g(\hat{\mu}_2^0, \hat{f}^{-1}\Sigma_{22}), \\ y^{(2)^*}|\mu, \Sigma &\sim N_g(\mu_2, \Sigma_{22}), \end{aligned} \tag{6.23}$$

where the first equation in (6.23) follows from the first equation in (6.22) in analogous fashion to equation (2.X36) from section 3.4.2. Then, integrating out first with respect to μ_2 and then Σ_{22}, we deduce,

$$\begin{aligned} y^{(2)^*}|\Sigma, Y &\sim N_g(\hat{\mu}_2^0, (1+\hat{f}^{-1})\Sigma_{22}), \\ y^{(2)^*}|Y &\sim t(g, 1; \hat{\Psi}_{22}^{-1}, 1+\hat{f}^{-1}, m-u-g+1). \end{aligned} \tag{6.24}$$

The only parameter in (6.24) to be affected by the partitioning is $\hat{\Psi}_{22}$, and by (6.14), the entropy is of the form

$$H(G) = \frac{1}{2}\log|\hat{\Psi}_{22}| + \text{ constants.}$$

Therefore, the network design problem reduces to choosing G to maximize $\log\hat{\Psi}_{22}$.

CKZ applied this criterion to the optimal reduction of a set of $p = 81$ stations in a wet deposition monitoring network. The data consisted of logged monthly mean sulfate levels which were treated as independent normal observations, for a total of $n = 48$ months. With m degrees of freedom for the prior Wishart distribution, the requirement of a proper posterior implies $m + n > p - 1$, or $m > 32$. This underlines the data adequacy problem mentioned earlier — even though they did have prior data on the full network, it was not really sufficient to estimate a full 81×81 covariance matrix. In practice, they chose four values of m arbitrarily (33, 36, 42 and 48), and examined the sensitivity to m in their subsequent discussion. For the prior value of Ψ, they used a simple intraclass correlation structure as in (2.X46), with σ^2 and ρ estimated from the combined data on all stations. An alternative formulation might be to estimate Ψ by fitting one of the standard spatial models for the variogram, which would be analogous to the approach of Loader and Switzer (1992). CKZ argued against this approach as making too strong a prior assumption about the form of the spatial covariances, but nevertheless, it seems that their own approach also puts rather strong prior weight on the assumption that Σ is of intra-class correlation structure, which is equally if not more unrealistic.

The actual algorithm adopted by CKZ consisted of dropping one station at a time, where the station to be dropped was selected by the criterion of maximizing $\hat{\Psi}_{22}$ for the remaining stations. This algorithm does not necessarily produce the optimal G over the set of all possible subsets of a fixed size g, but a complete solution to that problem would involve searching over a prohibitively large set. In comparing the different values

of m, they argued that the influence of m in the selected design was minimal, but they also showed that the total uncertainty could be decomposed into components representing prediction error, uncertainty about the model, and so on, and that the influence of m on these individual components was considerable (for example, $m = 33$, the smallest value admitted by their study, implies the greatest uncertainty about the model, and this could influence the optimal design if one used that as a design criterion).

In a subsequent extension of the methodology, Wu and Zidek (1992) proposed avoiding the problem created by the high dimensionality of the covariance matrix by dividing the data points into clusters, where each cluster had fewer than 48 data points, and treating the design problem as a separate task within each cluster. This way, the degrees of freedom problem associated with the Wishart prior was avoided, and it was possible to use improper priors without getting an improper posterior distribution. For the subdivision into clusters, Wu and Zidek applied the K-means clustering algorithm of Hartigan and Wong (1979), as adapted by Krzanowski and Lai (1988). Somewhat controversially, the variable to which they applied the K-means algorithm was not the spatial location of the data point, but the response variable of interest (they considered a total of 9 measurements for different ions, and therefore repeated the procedure 9 times, each one producing a different clustering of the stations). This procedure was in no way guaranteed to produce spatially contiguous clusters, though in most of their examples, the clusters did correspond to rough spatial groupings of the stations. They argued in favor of their approach, rather than a more obvious clustering based on spatial coordinates, on the grounds that by starting with clusters that are statistically homogeneous, the potential for reducing the network with minimal loss of information would be much greater.

Returning to the CKZ paper, at the end of their paper they gave further discussion of the two competing criteria for design. They argued that the total uncertainty in the system may be decomposed into components due to prediction of the unobserved portion of the network, learning about the model, and the measurement uncertainty in G. Since the total uncertainty in the system is a constant, choosing G to maximize the uncertainty in G is equivalent to minimizing the uncertainty in the other two components — in this sense, the criterion takes into account the benefits of learning about the model as well as predicting the unobserved portion of the network.

However, it is not entirely clear how this reasoning would apply in the limiting case when n, the number of prior observations, tends to infinity. In that case, $\hat{\Psi}$ converges to the true covariance matrix Σ and there is no model uncertainty at all. However, as already noted, in that case the CKZ criterion would choose G to maximize Σ_{22}, which is still not the same as the CZ criterion.

Perhaps an even simpler argument, also mentioned in passing by CKZ, is to ignore the role of U altogether, and simply view the $H(G)$ criterion as placing the monitors where they will give most information. In an environmental regulatory context, the p candidate locations for monitors could represent suspected violators of air pollution standards, and

the network design problem could be interpreted as gaining the maximum amount of evidence as a prelude to possible legal action.

6.3.3 Incorporating costs: Zidek, Sun and Le (2000)

A more recent paper by Zidek, Sun and Le (2000) — henceforth ZSL — has further elaborated these criteria and also considered how to incorporate costs of measurement into the analysis.

Extending an earlier analysis by Le, Sun and Zidek (1997) (recall section 3.4.4 for discussion of that), they considered a network of 31 stations in Ontario, actually obtained by combining three earlier networks, and also considered the possibility of extending them by adding up to 15 additional sites. Four pollutants were measured, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3) and sulfates (SO_4), but only the last two were considered in this analysis as they are the two pollutants currently considered most damaging on human health. Because the decision process includes both the possibility of adding a new monitor to an existing site and opening up a whole new site, they introduced the terminology of *pseudosites* to encompass both possibilities — each "pseudosite" consists of a specific site-monitor combination. Thus the existing network consists of 62 pseudosites (31 for O_3 and 31 for SO_4), not all of which are monitored, and there is also the possibility for adding up to 30 additional pseudosites. The costs of opening a new pseudosite were estimated at \$3,000 per year (in Canadian dollars at 1997 prices) for adding a new monitor to an existing site and \$8,000 for setting up a monitor at a completely new site (they also discussed other costs such as operating costs but we shall not refer to that part of the discussion here). The structure of the model is exactly of the "data missing by design" form, analyzed in detail by Le, Sun and Zidek (1997).

Following the CKZ theory described earlier, ZSL formulated the network design problem as choosing the new sites to maximize $\log|\Phi|$ where Φ is the prior (i.e. based on existing data) covariance matrix for all the gauged pseudosites after modifying the network. Since the problem in this instance was where to add new monitors, rather than where to delete existing ones, they decomposed $log|\Phi|$ into elements representing the existing monitors and the supposed new monitors. With this decomposition, the design criterion becomes to maximize

$$\log|\Phi_{\mathrm{add|g}}| = \log|\Phi_{\mathrm{add}} - \Phi_{\mathrm{add,\ g}}\Phi_{\mathrm{g}}^{-1}\Phi_{\mathrm{g,\ add}}|$$

where the notation reflects the subdivision of stations into existing stations "g" and added stations "add".

The actual details of the analysis involved the multivariate extension of the model defined by Brown, Le and Zidek (1994), as modified by Le, Sun and Zidek (1997) to take into account data missing by design, and also using the method of Sampson and Guttorp (1992) to extend the estimated covariance matrix to cover the whole region of interest. We omit discussion of these steps since they have already been outlined in section 3.4. Instead, we concentrate on the problem of incorporating costs into the optimal design criterion.

As noted already, there is a cost associated with a potential monitoring site s, denoted $C(s)$, which is not the same for all possible sites s. The ZSL theory also leads to an entropy measure $E(s)$, which has the intepretation of the uncertainty associated with the site s. Thus, it seems natural to consider a combined objective of maximizing $O(s) = E(s) - DE\ C(s)$ where DE is a cost to entropy conversion factor. When $DE = 0$ the problem reduces to that of finding the subset of possible added sites to maximize Φ, and this is one for which an efficient algorithm is now avaiable (Ko *et al.* 1995). However when $DE > 0$ this idea does not work and one is again reduced to an approach based on adding one pseudosite at a time. ZSL remark that if DE was too large, one would never add any stations at all, but in practice, stations do get added to networks, so the practical value of DE cannot be too large. By repeating the analysis for various small values of DE, including 0, they were able to assess the sensitivity of the selected network to the exact value of this parameter. In discussing specific numerical results they noted that the order of selecting pseudosites is not too sensitive to DE, but the cutoff point (i.e. when it is no longer beneficial to add new stations) is sensitive to the exact value of DE, as might be expected.

At the end of their paper, ZSL refer to some additional possibilities —

- ranking potential pseudosites based on the ratio $E(s)/C(s)$ (suggested by Dr. Lawrence Phillips),

- choosing the additional pseudosites to maximize the entropy of the expanded network subject to a constraint on total cost.

However they also mention the possibility of applying a new combinatorial optimization algorithm of Anstreicher *et al.* (1996) to refine the current ZSL approach, in particular, avoiding the one-at-a-time selection algorithm and directly calculating the optimum G. The possibility of using improved algorithms for maximum entropy sampling was also mentioned by Bueso *et al.* (1998), and further illuminated in a discussion by Lee (1998). Evidently, this is still a developing area of research.

6.3.4 Possibilities for extension to a fully hierarchical model

This subsection is purely speculative since the idea has not been tried in practice, but nevertheless, it seems worthwhile to outline how the ideas of CKZ and ZSL could, at least in principle, be extended to a fully hierarchical approach.

The model of section 6.3.2 assumed hyperparameters Ψ, μ^0, f and m. As discussed in section 3.4.4, a fully Bayesian hierarchical approach would allow all of these to be functions of a hyperparameter θ, with a prior density $\pi(\theta)$. The model may then be represented symbolically as

$$\begin{aligned} Y|\mu,\Sigma &\sim f(Y|\mu,\Sigma), \\ (\mu,\Sigma)|\theta &\sim g(\mu,\Sigma|\theta), \\ \theta &\sim \pi(\theta). \end{aligned} \tag{6.25}$$

A Bayesian hierarchical analysis proceeds by alternately sampling (μ, Σ) from the conditional density of (μ, Σ) given (Y, θ), and θ from the conditional density of θ given (Y, μ, Σ). In this way, we generate a Monte Carlo sample $\{\theta_a,\ 1 \leq a \leq A\}$ from the posterior density $\pi(\theta|Y)$. Let us suppose that this has been done and consider how to estimate the entropy measure for a subset G generating a future random variable $y^{(2)^*}$.

The full predictive density is

$$\int\int\int f(y^{(2)^*}|\mu, \Sigma) g(\mu, \Sigma|\theta, Y) \pi(\theta|Y) d\mu d\Sigma d\theta. \tag{6.26}$$

However, the integral with respect to μ and Σ is possible analytically, since it derives from the $t(g, 1; \hat{\Psi}_{22}^{-1}, 1 + \hat{f}^{-1}, \hat{m} - u - g + 1)$ predictive distribution for $y^{(2)^*} - \hat{\mu}_2^0$ which we have already seen — here $\hat{\Psi}_{22}$, $\hat{\mu}_2^0$, $\hat{f}$ and $\hat{m}$ are the parameters of the normal-Wishart posterior distribution given θ. Therefore, we can write down the density directly from (2.X11) and simplify (6.26) to

$$f_{\text{pred}}(y^{(2)^*}|Y) = \int f_{\text{pred}}(y^{(2)^*}|\theta, Y) \pi(\theta|Y) d\theta \tag{6.27}$$

which we would in practice estimate by

$$\hat{f}_{\text{pred}}(y^{(2)^*}|Y) = \frac{1}{A} \sum_{a=1}^{A} f_{\text{pred}}(y^{(2)^*}|\theta_a, Y). \tag{6.28}$$

Note the use of the subscript "pred" here to denote a predictive density. The evaluation of entropy, however, requires integration with respect to $y^{(2)^*}$, and this seems to require further Monte Carlo sampling. The following procedure is therefore proposed.

1. Generate a random sample $\theta_1, ..., \theta_A$ from the posterior distribution of θ given Y.

2. Fix some reference value $\tilde{\theta}$, for example (though not necessarily), the sample mean of $\theta_1, ..., \theta_A$.

3. Generate independent samples $z_1, ..., z_B$ from $N_g(0, I_g)$ and $S_1, ..., S_B$ from $W_g^{-1}(I_g, \hat{m} - u)$. Here I_g is the $g \times g$ identity matrix.

4. For each $b \in \{1, ..., B\}$, define $\Sigma_{22}(b) = \hat{\Psi}_{22}^{1/2} S_b \hat{\Psi}_{22}^{1/2}$ ($\hat{\Psi}_{22}^{1/2}$ is the matrix square root of $\hat{\Psi}_{22}$), and $y^{(2)^*}(b) = \hat{\mu}_2^0 + (1 + \hat{f}^{-1})^{1/2} \Sigma_{22}^{1/2}(b) z_b$. Thus, $\Sigma_{22}(b)$ is a random matrix from $W_g^{-1}(\hat{\Psi}_{22}, \hat{m} - u)$ and $y^{(2)^*}(b)$ is a random vector from the predictive distribution of $y^{(2)^*}$ given Y and $\theta = \tilde{\theta}$.

5. Defining $\hat{f}_{\text{pred}}(y^{(2)^*}|Y)$ as in (6.28), the Monte Carlo estimate of the entropy in G is given by

$$\hat{H}(G) = -\frac{1}{B} \sum_{b=1}^{B} \log \hat{f}_{\text{pred}}(y^{(2)^*}(b)|Y) \cdot \frac{\hat{f}_{\text{pred}}(y^{(2)^*}(b)|Y)}{\hat{f}_{\text{pred}}(y^{(2)^*}(b)|\tilde{\theta}, Y)}, \tag{6.29}$$

where $\hat{f}_{\text{pred}}(y^{(2)*}|\tilde{\theta}, Y)$ is the analytic predictive density derived from the multivariate t distribution.

Note that a feature of this procedure is the use of the same sequence of random numbers in step 3, whatever the design G being evaluated. This seems desirable, because it should minimize the influence of the random sequence in comparing different G's. Nevertheless, it must be admitted that this procedure lacks the simple pristine elegance of the $|\hat{\Psi}_{22}|$ criterion!

6.4. Optimal design theory and the General Equivalence Theorem

The major alternative theoretical approach, to the one based on entropy or information theory outlined in sections 6.1–6.3, is based on the theory of optimal experimental design. This theory originated in papers of Kiefer and Wolfowitz in the 1950s and has been well treated in a number of excellent books, e.g. Fedorov (1972), Silvey (1980), Atkinson and Donev (1992). However, the connection with spatial statistics and prediction was not made until two fundamental papers of Fedorov and Müller (1988, 1989). Subsequent work has see much further development of this approach, including a recent book by Müller (2000) which is recommended as the most complete treatment of this whole approach to network design theory.

This section gives a very brief outline of the main principles of optimal design theory, focussing on the famous Equivalence Theorem of Kiefer and Wolfowitz. In section 6.5 we outline the Fedorov-Müller approach, followed by more recent developments.

The traditional formulation of optimal design theory is for a linear regression problem in which certain variables x_i are chosen by the experimenter and p covariates of interest are known functions of x_i, denoted $f_1(x_i), ..., f_p(x_i)$. The i'th data point is then

$$y_i = \sum_{j=1}^{p} f_j(x_i)\beta_j + \epsilon_i, \tag{6.30}$$

where $\beta_1, ..., \beta_p$ are unknown coefficients and, as usual in linear regression theory, ϵ_i are uncorrelated errors with mean 0 and common variance σ^2.

Writing

$$Y_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, F_n = \begin{pmatrix} f_1(x_1) & \dots & f_p(x_1) \\ \vdots & \vdots & \vdots \\ f_1(x_n) & \dots & f_p(x_n) \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \tag{6.31}$$

the optimal estimator of β is of course $\hat{\beta}_n = (F_n F_n)^{-1} F_n^T Y_n$, with covariance matrix $(F_n^T F_n)^{-1}\sigma^2$.

For the purpose of developing optimal design theory, it is helpful to rewrite this well-known result in a more abstract form. Let $\mathcal{X}$ denote the space of all possible values for the design points x_i, and let ξ_n denote the measure on $\mathcal{X}$ which puts mass $\frac{1}{n}$ at each of the design points $x_1, ..., x_n$. Then with $f(x) = (\, f_1(x) \quad \dots \quad f_p(x)\,)^T$,

$$\frac{1}{n} F_n^T F_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i) f(x_i)^T = \int_{\mathcal{X}} f(x) f(x)^T d\xi_n(x), \tag{6.32}$$

where the integral in (6.32) is interpreted as a Stieltjes integral with respect to the discrete measure ξ_n.

Define

$$M(\xi) = \int_{\mathcal{X}} f(x) f(x)^T d\xi(x),$$

which is defined for any positive measure ξ, discrete or continuous, subject to the usual measurability and finiteness conditions. The abstraction referred to above is to generalize the definition of $M(\xi)$ from equally weighted discrete measures on n data points to any finite measure ξ.

For the linear regression (6.30), with ξ_n defined by (6.32), the covariance matrix of the best linear unbiased estimator $\hat{\beta}_n$ is given by

$$M(\xi_n)^{-1} \frac{\sigma^2}{n}.$$

Thus, choosing a good design means making $M(\xi_n)$ "large" in some suitably defined sense.

All the standard optimal design criteria are of the form: choose ξ to minimize $\Psi\{M(\xi)\}$ where $\Psi\{\cdot\}$ is some functional on the space of non-negative definite symmetric $p \times p$ matrices. In a real design problem with n observations, ξ is restricted to discrete probability measures whose weights are multiples of $\frac{1}{n}$, but for the purpose of developing a general theory, we allow ξ to be an arbitrary probability measure on $\mathcal{X}$.

Typical optimality criteria include the following:

• *D-optimality:* $\Psi\{M(\xi)\} = -\log |M(\xi)|$. The original motivation for this was that the D-optimal design minimizes the volume of a confidence ellipsoid of fixed significance level for β.

• *A-optimality:* $\Psi\{M(\xi)\} = \mathrm{tr}\{M(\xi)^{-1}\}$. This minimizes the average variance of the parameter estimates.

• *E-optimality:* $\Psi\{M(\xi)\}$ is the maximum of $a^T M(\xi)^{-1} a$ over all vectors a such that $a^T a = 1$. This is interpretable as minimizing the variance of the least well estimated contrast subject to a normalizing condition on the contrast.

• *G-optimality:* $\Psi\{M(\xi)\} = \max_{x\in\mathcal{X}} d(x,\xi)$ where $d(x,\xi) = f(x)^T M(\xi)^{-1} f(x)$. Here, $\sigma^2 d(x,\xi)$ is the variance of the estimated response function $f(x)^T\beta$ at x, so the G-optimal design is the one which minimizes this variance over all x.

A side comment about G-optimality is that the mean of $d(x,\xi)$ over the whole design is p:

$$\begin{aligned}\int_{\mathcal{X}} d(x,\xi)d\xi(x) &= \int_{\mathcal{X}} \mathrm{tr}\{M(\xi)^{-1}f(x)f(x)^T\}d\xi(x)\\ &= \mathrm{tr}\{M(\xi)^{-1}M(\xi)\}\\ &= \mathrm{tr}(I_p) = p.\end{aligned}$$

Therefore, $\max_{x\in\mathcal{X}} d(x,\xi) \geq p$ for any design ξ. *In particular*, if we can find a design ξ^* for which $\max_{x\in\mathcal{X}} d(x,\xi^*) = p$, ξ^* must be G-optimal.

D-, A- and E-optimality can all be rewritten in terms of the eigenvalues of $M(\xi)$, say $\lambda_1, ..., \lambda_p$, as follows:

• *D-optimality:* minimize $\prod \frac{1}{\lambda_i}$,

• *A-optimality:* minimize $\sum \frac{1}{\lambda_i}$,

• *E-optimality:* minimize $\max \frac{1}{\lambda_i}$.

These may all be regarded as special cases of the criterion

$$\Psi_k(\xi) = \left(\frac{1}{p}\sum_{i=1}^{p} \lambda_i^k\right)^{1/k}, \quad 0 < k < \infty, \tag{6.33}$$

in which $k = 1$ corresponds to A-optimality and the limits $k \to 0$, $k \to \infty$ give D-optimality and E-optimality respectively. (6.33) shows that there is, at least in theory, a continuum of optimal design criteria with many intermediate cases apart from the three given so far.

We now turn to the *General Equivalence Theorem*, originally given by Kiefer and Wolfowitz (1960) and subsequently much extended. Our treatement follows most closely the book by Atkinson and Donev (1992); Silvey (1980) presented what is probably the slickest proof of the result.

To present the General Equivalence Theorem, we must first define what it means by the derivative of a functional $\Psi\{M(\xi)\}$ with respect to ξ. Suppose δ_x is a unit point mass at x, and consider modifying ξ into

$$\xi' = (1-\alpha)\xi + \alpha\delta_x,$$

where $0 < \alpha < 1$. Then

$$M(\xi') = (1-\alpha)M(\xi) + \alpha M(\delta_x).$$

The derivative of Ψ, in the direction δ_x, is

$$\phi(x,\xi) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \left[\Psi\{(1-\alpha)M(\xi) + \alpha M(\delta_x)\} - \Psi\{M(\xi)\} \right].$$

The General Equivalence Theorem asserts that the following three conditions are equivalent:

(1) ξ^* minimizes $\Psi\{M(\xi)\}$,

(2) $\phi(x,\xi^*) \geq 0$ for all x,

(3) $\phi(x,\xi^*)$ achieves its minimum at points of the design, i.e. at points x which have positive point measure under ξ^*.

The rough intuition behind this is that (2) asserts that Ψ is not decreased by adding an infinitesimal point mass at x, for any x, so in that sense, the design ξ^* is locally optimal. However, most of the Ψ functions we consider, including D-optimality, are convex functions of ξ, so any locally optimal design is also globally optimal. For a full proof, the reader is referred to any of the books we have cited.

Example. Consider $\Psi\{M(\xi)\} = -\log|M(\xi)|$. Consider the case when $\xi = \xi_n$, the point measure with mass $\frac{1}{n}$ at each of $x_1, ..., x_n$. Let $F_n = (\, f(x_1) \quad \dots \quad f(x_n)\,)^T$ as in (6.31). Assume a new design ξ_{n+1} is created by adding a point mass at x, so that $F_{n+1} = (\, F_n^T \quad f(x)\,)^T$ and

$$\xi_{n+1} = \frac{n}{n+1}\xi_n + \frac{1}{n+1}\delta_x. \tag{6.34}$$

We also have

$$F_{n+1}^T F_{n+1} = F_n^T F_n + f(x) f(x)^T.$$

Therefore,

$$\begin{aligned} |F_{n+1}^T F_{n+1}| &= |F_n^T F_n| \cdot |I_p + (F_n^T F_n)^{-1} f(x) f(x)^T| \\ &= |F_n^T F_n| \{1 + f(x)^T (F_n^T F_n)^{-1} f(x)\} \end{aligned}$$

where we have used the matrix identity $|I_n + B^T C| = |I_m + C B^T|$ applicable whenever B and C are both $m \times n$ matrices (see, e.g., Mardia, Kent and Bibby (1979), section A.2.3, page 458).

Since $|F_n^T F_n| = n^p |M(\xi_n)|$, we have

$$\Psi\{M(\xi_{n+1})\} - p\log(n+1) = \Psi\{M(\xi_n)\} - p\log n - \log\left\{1 + \frac{f(x)^T M(\xi)^{-1} f(x)}{n}\right\},$$

and hence

$$\begin{aligned}
&\lim_{n\to\infty} n[\Psi\{M(\xi_{n+1})\} - \Psi\{M(\xi_n)\}] \\
&= \lim_{n\to\infty} np \log \frac{n+1}{n} - \lim_{n\to\infty} n \log \left\{1 + \frac{d(x,\xi)}{n}\right\} \\
&= p - d(x,\xi_n).
\end{aligned}$$

Since, for large n, any design ξ may be approximated by one concentrated on n equally weighted points, we conclude

$$\phi(x,\xi) = p - d(x,\xi).$$

With this interpretation, condition (2) for the D-optimality of a design ξ^* reduces to

$$d(x,\xi) \leq p \text{ for all } x, \tag{6.35}$$

and we have already seen that (6.35) implies G-optimality of the design ξ. Therefore, provided we extend the notion of design to allow arbitrary measures ξ, the General Equivalence Theorem implies that D-optimality and G-optimality are equivalent.

Atkinson and Donev (1992) give a table of $\phi(x,\xi)$ functions for a variety of design criteria Ψ, including:

- For A-optimality, $\phi(x,\xi) = \text{tr}\{M(\xi)\}^{-1} - f(x)^T M(\xi)^{-2} f(x)$,

- For E-optimality, $\phi(x,\xi) = \lambda_{\min} - f(x)^T r r^T f(x)$, where $\lambda_{\min}$ is the smallest eigenvalue and r is the corresponding unit eigenvector.

Another consequence of the General Equivalence Theorem is that it implies an algorithm for constructing the optimal design. Suppose we add one design point at a time and, given the current design measure ξ_n, define ξ_{n+1} by (6.34) where x is chosen to minimize $\phi(x,\xi_n)$. Then as $n \to \infty$, the design measure ξ_n converges to the optimal design ξ^*. However, this strategy (which amounts to a simple steepest descent algorithm) does not generally lead to a very fast rate of convergence.

It should be pointed out that neither the General Equivalence Theorem nor the algorithm just described necessarily leads to the optimal design over n data points for any fixed n — they should be interpreted as limiting results for $n \to \infty$. Nevertheless, for large n it seems reasonable to assume that the design ξ_n created by the algorithm is a reasonable approximation to the optimal design on n points, and this is the philosophy we shall adopt subsequently.

6.5 Applications of optimal design theory to the design of spatial networks

6.5.1 The Fedorov-Müller approach

Fedorov and Müller (1988, 1989) made the connection between optimal design theory and choosing a network to optimize spatial prediction by exhibiting a particular class of spatial models which could be viewed from either of two points of view, as a regression model or as a spatial random field model, and showing how two different points of view led essentially to the same conclusions. In doing this, they opened up the possibility of applying optimal design theory to more general problems of spatial prediction in random fields.

The Fedorov-Müller (FM) model assumes observations of the form

$$y_{it} = f(x_i)^T \theta_t + \epsilon_{it}, \quad 1 \le i \le n, \quad 1 \le t \le T, \tag{6.36}$$

where y_{it} denotes the observation at time t at location x_i, $f(x)$ is a known vector of regressors at location x, and ϵ_{it} are uncorrelated random errors with mean 0 and variance 1. Each θ_t is a vector of coefficients which, in their model, is different at each time point t. In fact the model is of simple "random effects" type: for each t, θ_t is sampled from a distribution which has mean θ_0 and covariance matrix D_0, and θ_t at different time points t are uncorrelated. Thus, combining all the y_{it} at a single time point t into a vector y_t, and similarly combining the $f(x_i)$ vectors into a matrix, F, and the errors ϵ_{it} into a vector ϵ_t, (6.36) may also be written

$$y_t = F^T \theta_t + \epsilon_t, \quad \mathrm{E}\{\theta_t\} = \theta_0, \quad \mathrm{Cov}\{\theta_t\} = D_o, \tag{6.37}$$

or equivalently

$$\mathrm{E}\{y_t\} = \theta_0, \quad \mathrm{Cov}\{y_t\} = I + F^T D_0 F. \tag{6.38}$$

Note that (6.38) is of the form of a very particular kind of spatial covariance model, which one could analyze from a "geostatistical" point of view without explicit consideration of the regression.

Within this model framework, FM considered three cases:

(a) θ_0, D_0 known, θ_t to be estimated for each t,

(b) D_0 known, θ_0 and each θ_t to be estimated,

(c) θ_0, D_0 both unknown.

Under assumption (a), the optimal estimator of θ_t is

$$\hat{\theta}_t = (D_0^{-1} + nM)^{-1}(D_0^{-1}\theta_0 + F y_t),$$

where $M = n^{-1}FF^T = n^{-1}\sum_{i=1}^{n} f(x_i)f(x_i)^T$. In this case,

$$\mathrm{E}\{(\hat{\theta}_t - \theta_t)(\hat{\theta}_t - \theta_t)^T\} = (D_0^{-1} + nM)^{-1}. \tag{6.39}$$

In case (b), the optimal estimator of θ_t is just the usual least squares estimator,

$$\bar{\theta}_t = (nM)^{-1} F y_t,$$

and the optimal estimator of θ_0 is

$$\hat{\theta}_0 = \frac{1}{T} \sum_{t=1}^{T} \bar{\theta}_t.$$

The corresponding variances are

$$\mathrm{E}\{(\bar{\theta}_t - \theta_t)(\bar{\theta}_t - \theta_t)^T\} = n^{-1} M^{-1}, \tag{6.40}$$

$$\mathrm{E}\{(\hat{\theta}_0 - \theta_0)(\hat{\theta}_0 - \theta_0)^T\} = T^{-1}(D_0 + n^{-1} M^{-1}). \tag{6.41}$$

In case (c), no exact theory for the estimation of θ_0 and D_0 exists, but FM advocated sample-based estimators, arguing that as $T \to \infty$ these would be consistent estimators and therefore, for large T, the theory would be similar to case (a).

The optimal design problem is based on choosing $x_1, ..., x_n$ so as to minimize some suitable functional on one of (6.39), (6.40) or (6.41), and FM considered only the D-optimality criterion, but with three variants corresponding to each of the three covariance matrices. They therefore considered

$$\Psi_1 = -\log|nM|, \quad \Psi_2 = -\log|D_0^{-1} + nM|, \quad \Psi_3 = \log|D_0 + n^{-1} M^{-1}|.$$

Here, Ψ_3 would be appropriate if our main objective were to estimate θ_0, e.g. for the purpose of estimating long-term characteristics of the system, whereas Ψ_1 or Ψ_2 would be more appropriate for interpolation at a specific time instance t.

In accordance with the General Equivalence Theorem, observations should be placed at locations which maximize $\psi(x, \xi)$, where

$$\begin{aligned}
&\text{for } \Psi_1 : \psi(x,\xi) = f(x)^T M(\xi)^{-1} f(x), \\
&\text{for } \Psi_2 : \psi(x,\xi) = f(x)^T \{D_0^{-1} + nM(\xi)\}^{-1} f(x), \\
&\text{for } \Psi_3 : \psi(x,\xi) = f(x)^T M(\xi)^{-1} \{D_0 + n^{-1} M(\xi)^{-1}\}^{-1} M(\xi)^{-1} f(x).
\end{aligned} \tag{6.40}$$

Now we turn to the alternative viewpoint of the problem, in which we view y_t as a random field with covariance given by (6.38), and consider optimal prediction at a new location x. The case θ_0 known is just the calculation of a conditional distribution in the multivariate normal distribution, while the case θ_0 unknown corresponds to the traditional formulation of kriging in which the process has unknown constant mean. For the present discussion, we consider only the case θ_0 known. FM in fact solve this problem by direct calculations on the covariance matrix (6.38), but a simpler solution is just to use the

equivalent random effects representation from which it follows at once that the optimal predictor at a point x is $f(x)^T\hat{\theta}_t$, with prediction variance $1 + f(x)^T(D_0^{-1} + nM)^{-1}f(x)$.

However, a rational criterion when adding a point to a network is to add a new monitor as the location where the prediction variance is greatest — or alternatively, if the objective is to reduce the size of the network, to remove the point for which the prediction variance given the other points is smallest. Both of these correspond to designing the network to include points for which $f(x)^T(D_0^{-1} + nM)^{-1}f(x)$ is large, and by the general equivalence theory, such a strategy is consistent (for large n) with choosing the network to minimize Ψ_2.

Thus, a formal equivalence between D-optimality theory and a more intuitive criterion, based on choosing monitor sites at places where prediction variance is maximized, is established.

Despite the elegance of this conclusion, both the model and the result are rather artificial for a real design problem. The covariance function (6.36) allows for monitors to be placed quite close to each other, or even at exactly the same location, still without achieving perfect correlation among the observations. Moreover, most solutions of D-optimality problems in fact lead to measures concentrated on a fairly small number of data points, i.e. they will involve multiple replications. But in a real system, there are no independent replications of the field at a single time instance, and taking repeated measurements at the same location is a pointless waste of data. FM tried to address this problem by placing an upper bound on the measure ξ, thereby forcing points to be separated, but this is clearly an artificial solution in the context of their model, and leaves open the question of how to solve the design problem for more realistic spatial models.

6.5.2 Designs for estimating a regression function in a spatially correlated field

This section, which is based on Chapter 5 of Müller (2000), discusses the problem of optimal design for estimating a regression function in a spatially correlated field, when the correlation function is known. The alternative case when the correlation function is unknown, and whose estimation is one of the objectives of the experiment, is considered in Chapter 6 of Müller (2000) and section 6.5.3 here.

There is an extensive literature on design of experiments in cases with correlated errors, for example Sacks and Ylvisaker (1966, 1968, 1970), Sacks *et al.* (1989), Su and Cambanis (1994). However most of the problems for which theoretical solutions are available are restricted to particular types of covariance functions and one-dimensional problems. The emphasis here is much more on computational solutions for the kinds of covariance functions typically found in spatial statistics.

There are two reasons why the theory of section 6.5.1 is unrealistic for a real monitoring problem: the covariance function (6.38) is implausible for a smooth random field, and related to that, the problem mentioned at the end of section 6.5.1, i.e. that D-optimality

typically leads to designs concentrated on a small number of points whereas the present context clearly requires taking observations at n different (and, typically, widely separated) data points. To achieve this, we need some alternative formulations of the problem.

The basic model is considered to be of the form

$$y_i = \eta(x_i, \beta) + \epsilon_i, \tag{6.41}$$

where $\eta(\cdot, \cdot)$ is a known nonlinear function dependent on an unknown parameter β, and $\mathrm{Cov}\{\epsilon_i, \epsilon_j\} = c(x_i, x_j)$ is known for all $x_i,\ x_j$. Of course, the case when $\eta(x, \beta)$ reduces to a linear function of the form $f(x)^T\beta$ is a special case of this. Write y and $\eta(\beta)$ for the vectors formed by stacking the invidual y_i and $\eta(x_i)$ in a column, and $C(A)$ the matrix formed from all $c(x_i, x_j)$, when A is the set $\{x_1, ..., x_n\}$. We also write n as n_A to signify the number of data points in A. The nonlinear generalized least squares criterion chooses β to minimize

$$S\{\beta, C(A)\} = \{y - \eta(\beta)\}^T C(A)^{-1}\{y - \eta(\beta)\}.$$

The covariance matrix of the estimator $\hat{\beta}$ is asymptotically equivalent to $M(A)^{-1}$, where

$$M(A) = \sum_{x \in A} \sum_{x' \in A} \dot{\eta}(x)[C(A)^{-1}]_{x,x'}\dot{\eta}^T(x'), \tag{6.42}$$

where $\dot{\eta}(x)$ is the column vector of partial derivatives of $\eta(x, \beta)$ with respect to β. By analogy with section 6.4 (but using slightly different notation, Φ in place of $-\Psi$), we assume the objective is to choose the finite set A to achieve

$$\max_{A \subset \mathcal{X},\ n_A \leq n} \Phi\{M(A)\},$$

where n is the given permitted number of monitors and $\mathcal{X}$ is the sampling space from which the values x_i must be selected.

In the case of continuous design measures, for the examples in in section 6.5.1, Ψ was a convex fuction of ξ. With the switch of sign, this means we would like Φ to be a concave or at least differentiable function of ξ, so that derivative-based methods can again be used to find the optimum. However, the difficulty is that this is no longer true when M is given by (6.42) and the design is restricted to those giving equal weight to n different design points.

Here, we give an outline to one solution of this problem, referring to Müller (2000) or Pázman and Müller (2000) for the full details. Other algorithms for the same or related problems are discussed by Pázman and Müller (1998), Müller and Pázman (1998, 1999) and also reviewed by Müller (2000).

Since the use of the exact information matrix (6.42) leads to a nondifferentiable function, the first step is to introduce a class of *approximate information matrices*, defined on

all measures ξ with finite support S_ξ, which are close to the true information matrix but nevertheless differentiable. One such choice is

$$M^{(\epsilon)}(\xi) = \sum_{x \in S_\xi} \sum_{x' \in S_\xi} \dot{\eta}(x)[C(S_\xi) + W^{(\epsilon)}(\xi)]^{-1}_{x,x'} \dot{\eta}^T(x'), \tag{6.42}$$

where ϵ is a small positive constant and $W^{(\epsilon)}(\xi)$ is a diagonal matrix with entries

$$[W^{(\epsilon)}(\xi)]_{x,x} = \log\left\{\left(\frac{\xi^{(\epsilon)}}{\xi(x)}\right)^{\epsilon}\right\},$$

and $\xi^{(\epsilon)} = (\sum_{x \in S_\xi} \xi(x)^{1/\epsilon})^\epsilon$ is a continuous approximation to $\xi_{\max} = \max\{\xi(x), x \in S_\xi\}$.

The idea behind this definition is that $M^{(\epsilon)}(\xi)$ is very close to $M(A)$ when ξ is uniformly distributed over a finite set A, but for $\epsilon > 0$, gives relatively little weight to data points x for which $\xi(x) < \xi_{\max}$, and therefore, should lead to a optimal design with roughly equal weights.

However, even with this modification there is still no way to force $n_{S_\xi} \leq n$, so we introduce another modification design to force all $\xi(x)$ to be at least κ when $x \in S_\xi$. The idea is that if we take $\kappa \approx 1/n$, we will get a design that indeed gives approximately equal weight to n design points.

The proposed modification replaces (6.42) with

$$M^{(\epsilon)}_\kappa(\xi) = \sum_{x \in S_\xi} \sum_{x' \in S_\xi} \dot{\eta}(x)[C(S_\xi) + W^{(\epsilon)}_\kappa(\xi)]^{-1}_{x,x'} \dot{\eta}^T(x'), \tag{6.43}$$

where $W^{(\epsilon)}_\kappa(\xi)$ is a diagonal matrix with entries

$$[W^{(\epsilon)}_\kappa(\xi)]_{x,x} = \log\left\{\left(\frac{\xi_\epsilon}{\xi_\epsilon(x)}\right)^{\epsilon}\right\},$$

where

$$\xi_\epsilon(x) = \{\kappa^{1/\epsilon} + \xi(x)^{1/\epsilon}\}^\epsilon - \kappa,$$
$$\xi_\epsilon = \{\kappa^{1/\epsilon} + \sum_{x \in S_\xi} \xi(x)^{1/\epsilon}\}^\epsilon - \kappa.$$

Note that $\xi_\epsilon(x)$ is a differentiable approximation to the function which takes value $\xi(x) - \kappa$ when $\xi(x) > \kappa$ and 0 when $\xi(x) < \kappa$. When $\kappa = 0$, $M^{(\epsilon)}_\kappa(\xi)$ is the same as $M^{(\epsilon)}(\xi)$. Moreover, when $\kappa > 0$ and $\xi(x) \geq \kappa$ for all $x \in S_\xi$, $\lim_{\epsilon \to 0} M^{(\epsilon)}_\kappa(\xi) = M(\xi)$ (independent of κ), so this really does approximate the original criterion.

For some purposes we need an additional refinement: in (6.43), replace $W^{(\epsilon)}_\kappa(\xi)$ by $\rho W^{(\epsilon)}_\kappa(\xi)$ for $\rho \geq 1$. This is explained further below.

The actual algorithm proposed by Müller (2000) is as follows:

1. Fix some discrete approximation $\bar{\mathcal{X}}$ to $\mathcal{X}$, and start with a desigh giving positive weights to all points in $\bar{\mathcal{X}}$. This could be equally weighted over the whole of $\bar{\mathcal{X}}$, or alternatively, a small perturbation of a previous design ξ where the points in $\bar{\mathcal{X}} - S_\xi$ get very small but positive weights.

2. At iteration s and with current design $\xi_{(s)}$, try to increase $M_\kappa^{(\epsilon)}(\xi_{(s)})$ by adding weight to $x_{(s)}$, where $x_{(s)}$ is determined by a method desribed below.

3. Continue until no further increase is possible, resulting in a candidate optimal design ξ^+. However, as a check against whether ξ^+ is a local rather than a global optimum, repeat the iteration using a larger value of ρ. If the new iteration converges to the same ξ^+, stop, because this means a global optimum has been achieved. If the new iteration produces a different ξ^+, increase ρ again and repeat the process. However, ρ must not be made too large, because large values of ρ tend to produce nonsensical designs.

The steps for determining the next design point $x_{(s)}$, when the current design is $\xi_{(s)}$, are as follows. For ease of notation we write ξ in place of $\xi_{(s)}$ and x_r or x_q in place of $x_{(s)}$.

(a) Define $U_\kappa(\xi) = [C(\bar{\mathcal{X}}) + \rho V_\kappa(\xi)]^{-1}$, where $V_\kappa(\xi)$ is the diagonal matrix with diagonal entries $\log\{\min(\xi_{\max}, \kappa)/\min(\xi(x), \kappa)\},\ x \in \bar{\mathcal{X}}$.

(b) Define

$$a(x) = \sum_{z \in \bar{\mathcal{X}}} [U_\kappa(\xi)]_{x,z} \dot{\eta}(z),$$

$$g(\xi, x) = a(x)^T \nabla\Phi \left[\sum_{x' \in \bar{\mathcal{X}}} \sum_{z \in \bar{\mathcal{X}}} \dot{\eta}(x') [U_\kappa(\xi)]_{x',z} \dot{\eta}(z)^T \right] a(x).$$

Here $\nabla\Phi(A)$ denotes the gradient of the matrix of the function Φ, i.e. the matrix with entries $\partial\Phi(A)/\partial a_{ij}$ where $\{a_{ij}\}$ are the entries of the matrix A.

(c) Compute

$$q(\xi) = \max_{x:\ \xi(x) \le \kappa} \frac{g(\xi, x)}{\xi(x)}, \tag{6.44}$$

and let x_q denote the x for which (6.44) is achieved.

(d) If $\xi_{\max} > \kappa$, compute

$$r(\xi) = \max_{x:\ \xi(x) > \kappa} \frac{\left\{ g(\xi, x) - [\sum_{x' \in \mathcal{X}} g(\xi, x')] \mathcal{I}_{B_\xi}(x)/n_{B_\xi} \right\}}{\xi(x) - \kappa} \tag{6.45}$$

and let x_r denote the x for which (6.45) is maximized. Here B_ξ denotes the set of x for which $\xi(x) = x_{\max}$ and $\mathcal{I}_B(\cdot)$ is the indicator of the set B.

(e) Place the next observation at $x = x_r$ if $q(\xi) < -\epsilon'$ or if $|q(\xi)| \leq \epsilon'$ and $r(\xi) \geq 0$; otherwise, take x_q. Here ϵ' is a fixed very small constant.

6.5.3 Other design objectives: Estimating the variogram

The algorithm of section 6.5.2 is designed specifically for the problem of estimating regression coefficients when the spatial covariance function is known. There are at least two other "design" problems of importance: design for efficient estimation of the variogram (or spatial covariance function), and design for optimal prediction and interpolation. In this section, we briefly discuss the second problem before giving more detailed treatment of the first.

Designs for optimal prediction are considered in section 5.4 of Müller (2000), who considers criteria of the form

$$\sum_{j=1}^{q} w_j \mathrm{Var}\{\hat{y}(x_j^*)\}, \tag{6.46}$$

where $\hat{y}(x_j^*)$ denotes the predictor at a point x_j^* and w_j is a weight function. Here, $x_1^*, ..., x_q^*$ are q given locations at which a good predictor is required, not to be confused with the sampling locations $x_1, ..., x_n$. If $q \leq n$ we can trivially minimize (6.46) by fixing a sampling location at each x_j^*, but if $q > n$ the problem is not trivial.

In the case of a regression model with uncorrelated homoscedastic errors, the formula for $\mathrm{Var}\{\hat{y}(x_j^*)\}$ is

$$\sigma^2\{1 + x_j^{*T}(X^T X)^{-1} x_j^*\}, \tag{6.47}$$

and the problem of minimizing the maximum of (6.47) over all x_j^* is just the problem of G-optimality, which, as discussed in section 6.4, is asymptotically equivalent to D-optimality. However, for a general correlated case, the formula (6.47) must be replaced by equation (2.61) from chapter 2, which gave the variance of the universal kriging procedure. This is not readily amenable to the standard optimal design techniques for finding the optimal sampling points, though of course, if viewed simply as a criterion function by which to evaluate competing designs, the combination of (6.46) and (2.61) is perfectly feasible.

Other approaches to design for optimal interpolation have been given by Pesti *et al.* (1994) and Benedetti and Palma (1995). A much earlier reference is Bras and Rodriguez-Iturbe (1976).

Now let us turn to design to optimize estimation of the variogram or covariance function.

Warrick and Myers (1987) considered the standard variogram estimator in which there are n sampling locations and hence $n(n-1)/2$ interpoint distances, which are classified

into N bins $B_1, ..., B_N$, each centered on a particular distance h_i. If f_i denotes the number of (x_j, x_k) pairs in bin B_i, the traditional estimator of the semivariogram is

$$\hat{\gamma}(h_i) = \frac{1}{2f_i} \sum_{(x_j,x_k)\in B_i} \{y(x_j) - y(x_k)\}^2.$$

Alternative estimators, such as the Cressie-Hawkins robust estimator, lead to similar design considerations and are not treated separately. The bins B_i may be defined purely by Euclidean distances or, in anisotropic cases, may include angles as well.

Warrick and Myers defined five characteristics of a good design:

1. For each distance-angle class, the number of pairs should be as large as possible, particularly for short distances,

2. The average of the distances in each class should be close to the plotted distance,

3. The variance of the distances in each class should be small,

4. The average of the angles in each class should be close to the plotted angle,

5. The variance of the angles in each class should be small.

The actual criterion they proposed was of the form

$$SS = a \sum_{i=1}^{N} w_i (f_i - f_i^*)^2 + b \sum_{i=1}^{N} m_{1i} + c \sum_{i=1}^{N} m_{2i}, \tag{6.48}$$

where f_i^* are target values for f_i, m_{1i} and m_{2i} are respectively the variances of the distances and angles in bin B_i, and the w_i and a, b, c are chosen weights. Obviously there is considerable arbitrariness over the choice of these quantities and for their computational examples, they concentrated on the case $b = c = 0$, $w_i = 1$ and all f_i^* the same (though the latter specification, in particular, is questionable, given extensive more recent theory which has suggested that the behavior of the variogram at short distances is critical to efficient spatial prediction, e.g. Stein (1988, 1999)).

The algorithm proposed by Warrick and Myers essentially starts with M arbitrary locations and then adds a further $N - M$ points at random, adding and deleting points at random until no practical further reduction is possible. As pointed out by Müller and Zimmerman (1999), the criterion (6.48) is essentially a multidimensional scaling criterion, for which specialized procedures are available (Cox and Cox 1995).

Several subsequent authors extended these ideas to consider not only nonparametric estimation of the semivariogram but also parametric estimation, in cases where the semivariogram function is specified as a function $\gamma(\cdot, \theta)$ with θ a finite-dimensional parameter. Here we concentrate on the paper of Müller and Zimmerman (1999).

Müller and Zimmerman considered the variant of the standard variogram estimator in which there are $N = n(n-1)/2$ bins and every $f_i = 1$, the so-called variogram cloud esitmator. If the i'th interdistance pair corresponds to data locations x_r and x_s, we define $\hat{\gamma}_i = \frac{1}{2}\{y(x_r) - y(x_s)\}^2$ as an estimator of the theoretical semivariogram function $\gamma(x_r - x_s, \theta)$. Since in the Gaussian case an analytical expression for the covariance of γ_i and γ_j is known (recall equations (2.14) and (2.15) in chapter 2), this creates the possibility of a Generalized Least Squares algorithm of the form

$$\hat{\theta}_{\text{GLS}} = \arg\min_{\theta}\{\hat{\gamma} - \gamma(\theta)\}^T \Sigma^{-1}(\theta)\{\hat{\gamma} - \gamma(\theta)\}, \tag{6.49}$$

where $\hat{\gamma}$ and $\gamma(\theta)$ are the vectors formed from the individual $\hat{\gamma}_i$ and $\gamma(x_r - x_s, \theta)$ and $\Sigma(\theta)$ is the covariance function derived from (2.14) and (2.15).

Unfortunately, as pointed out by Müller and Zimmerman, (6.49) as an estimator of the variogram is inconsistent, and they proposed an alternative iterative scheme which involves a sequence of estimators θ_m, where θ_{m+1} is chosen to minimize

$$\{\hat{\gamma} - \gamma(\theta)\}^T \Sigma^{-1}(\theta_m)\{\hat{\gamma} - \gamma(\theta)\}, \tag{6.50}$$

and the algorithm is iterated to convergence. This estimator is denoted $\bar{\theta}$ and has asymptotic covariance given by $M(\xi, \theta)^{-1}$, where

$$M(\xi, \theta) = G^T(\xi, \theta)\Sigma^{-1}(\xi, \theta)G(\xi, \theta), \tag{6.51}$$

where ξ denotes the design, Σ is the covariance matrix as in (6.49) and (6.50), and the $N \times p$ matrix G consists of all partial derivatives of the form $\partial\gamma_r/\partial\theta_j,\ 1 \le r \le N,\ 1 \le j \le p$.

The optimal design may then be defined as the design measure ξ which maximizes $\Phi\{M(\xi, \theta)\}$ for some functional Φ, e.g. $\Phi(M) = \log|M|$. However, in this case there is an extra complication because M depends on θ which is unknown, and indeed the purpose of the experiment is to estimate θ. However, this problem also arises in connection with nonlinear regression (e.g. the books on optimal design by Silvey (1980) and Atkinson and Donev (1992) both have a chapter on this), and is usually solved by some form of iterative criterion in which the design is constructed sequentially using the best available estimate of θ at each step. The solution discussed by Müller and Zimmerman (1999) assumed that some trial value $\hat{\theta}_0$ is available which is taken as the basis for determining the optimal design.

The key step in the Müller-Zimmerman development is to show how M changes as the result of adding a new sampling location x to the existing design ξ. They are then able to optimize the selection of x to maximize the increase in $\Phi\{M(\xi')\}$ when ξ is modified into ξ' by the additional of a sampling point at x.

If we add a new data point, we add n new interpoint distances and the change in M is to M^*, where M^* is of the form

$$M^* = M + \Gamma V \Gamma^T,$$

where Γ is $p \times n$ and V is $n \times n$. The new data point is therefore chosen at the location x which maximizes $\Phi(M^*)$.

If M^* has to be evaluated from scratch for each candidate x, this algorithm could be computationally very tedious, e.g. if $\Phi(M^*) = \log |M^*|$ then most standard algorithms for the computation of $|M^*|$ require $O(N^3) = O(n^6)$ calculations. Fortunately, however, we do not need to repeat this entire calculation for each x, because of the formula

$$|M^*| = |M| \; |V| \; |V^{-1} + \Gamma^T M^{-1} \Gamma|, \tag{6.52}$$

where only the second and third components have to be re-evaluated for each x, and these are both $O(n^3)$ rather than $O(n^6)$ operations.

Müller and Zimmerman (1999) considered several variants on the basic algorithm, e.g. adding one point at a time, adding several points together, or alternative addition and deletion of data points (a procedure suggested by Fedorov), and also included extensive comparisons with alternative algorithms.

One problem that Müller and Zimmerman (1999) do not address is how to combine design criteria for the estimation of the variogram and the regression coefficients in a spatially correlated regression. See also Müller (2000), Chapter 6, who also reviews these developments and proposes weighted averages of different design criteria.

For the remainder of this section, I outline a possible alternative approach which combines estimation of the regression and spatial variance components into a single estimation procedure which therefore allows the application of standard design criteria such as D-optimality (though whether an all-purpose criterion such as D-optimality is appropriate in a context where the different parameters may have very different interpretations and levels of importance is a question I do not consider). This approach is essentially similar to the Müller-Zimmerman method just discussed, but uses maximum likelihood estimation and the Fisher information matrix.

Suppose the current data are represented by an n-dimensional vector Y with mean $\mu(\theta)$ and covariance matrix $V(\theta)$, both parametric functions of some p-dimensional θ. Let $M(\theta)$ denote the Fisher information matrix, i.e. the matrix with entries $m_{rs}(\theta)$, $1 \leq r \leq p$, $1 \leq s \leq p$,, where

$$m_{rs}(\theta) = \mathrm{E} \left\{ \frac{\partial^2 \log f(Y;\theta)}{\partial \theta_r \partial \theta_s} \right\}.$$

Then it follows that

$$m_{rs}(\theta) = \frac{1}{2} \mathrm{tr} \left(\frac{\partial V}{\partial \theta_r} V^{-1} \frac{\partial V}{\partial \theta_s} V^{-1} \right) + \frac{\partial \mu^T}{\partial \theta_r} V^{-1} \frac{\partial \mu}{\partial \theta_s}. \tag{6.53}$$

As the result (6.53) may not be entirely well known, a proof is given separately below.

Suppose now we are considering adding a new observation y^* at a location x. The joint distribution of $(\,Y^T \quad y^*\,)^T$ is normal with mean $(\,\mu^T \quad \nu\,)^T$ and covariance matrix $\begin{pmatrix} V & \tau \\ \tau^T & \phi \end{pmatrix}$ where ν and ϕ are the mean and variance of y^* and τ is the vector of cross-covariances between Y and y^*. Note that ν and ϕ depend on the new location x while τ is a joint function of the new location x and the existing design ξ, but we shall not indicate this explicitly in the notation.

The conditional distribution of y^* given $Y = y$ is of the form $N(\beta, \alpha)$, where $\beta = \nu + \tau^T V^{-1}(y - \mu)$ and $\alpha = \phi - \tau^T V^{-1}\tau$. Applying formula (6.53) just to this, the (r, s) contribution to the information matrix from y^*, conditional on $Y = y$, is

$$\frac{1}{2\alpha^2}\frac{\partial\alpha}{\partial\theta_r}\frac{\partial\alpha}{\partial\theta_s} + \frac{1}{\alpha}\frac{\partial\beta}{\partial\theta_r}\frac{\partial\beta}{\partial\theta_s}. \tag{6.54}$$

However, in (6.54), β is a function of y, so we need to take a further expectation to evaluate the unconditional expectation of (6.54).

Defining $\omega = V^{-1}\tau$, we have $\beta = \nu + \omega^T(y - \mu)$, so

$$\frac{\partial\beta}{\partial\theta_r} = \left(\frac{\partial\nu}{\partial\theta_r} - \omega^T\frac{\partial\mu}{\partial\theta_r}\right) + \frac{\partial\omega^T}{\partial\theta_r}(y - \mu),$$

so multiplying $\frac{\partial\beta}{\partial\theta_r}\frac{\partial\beta}{\partial\theta_s}$ and taking expectations gives the result

$$\left(\frac{\partial\nu}{\partial\theta_r} - \omega^T\frac{\partial\mu}{\partial\theta_r}\right)\left(\frac{\partial\nu}{\partial\theta_s} - \omega^T\frac{\partial\mu}{\partial\theta_s}\right) + \frac{\partial\omega^T}{\partial\theta_r}V\frac{\partial\omega}{\partial\theta_s}.$$

Substituting back in (6.54), the change in $m_{rs}(\theta)$ as a result of adding the new observation y^* is

$$\frac{1}{2\alpha^2}\frac{\partial\alpha}{\partial\theta_r}\frac{\partial\alpha}{\partial\theta_s} + \frac{1}{\alpha}\left\{\left(\frac{\partial\nu}{\partial\theta_r} - \omega^T\frac{\partial\mu}{\partial\theta_r}\right)\left(\frac{\partial\nu}{\partial\theta_s} - \omega^T\frac{\partial\mu}{\partial\theta_s}\right) + \frac{\partial\omega}{\partial\theta_r}V\frac{\partial\omega}{\partial\theta_s}\right\} \tag{6.55}$$

Collecting the entries of (6.55) into a matrix $U(\theta)$, we see that the new information matrix after adding the observation y^* is of the form $M(\theta) + U(\theta)$. If the design objective is to maximize a functional $\Phi\{M(\theta)\}$, therefore, a reasonable algorithm would be to choose the new design point x to maximize

$$\Phi\{M(\theta) + U(\theta)\} - \Phi\{M(\theta)\}. \tag{6.56}$$

As in the Müller-Zimmerman procedure discussed previously, it is desirable to simplify the calculation of (6.56) so that $\Phi\{M(\theta) + U(\theta)\}$ does not have to be re-evaluated from scratch for each candidate design point x. In the case $\Phi(M) = \log|M|$ (D-optimality), this is possible, as follows. Looking carefully at the form of (6.55), we see that it is a sum of three terms each of which is of the form $u_r u_s$ for scalar components u_r and u_s.

Therefore, it is possible to write $U(\theta) = W(\theta)W(\theta)^T$ where W is a $n \times 3$ matrix. But then, application of the $|I + B^T C| = |I + C^T B|$ formula (Mardia, Kent and Bibby (1979), equation (A.2.3n) — of course this also lies behind (6.52) above) leads to

$$|M + U| = |M + WW^T| = |M| \cdot |I + W^T M^{-1} W|, \tag{6.57}$$

so the only determinant that needs to be re-evaluated for each x is that of a 3×3 matrix.

The algorithm suggested here, like the one of Müller and Zimmerman (1999), is cruder than the procedures considered in section 6.5.2, because they only allow for adding and deleting points in the class of designs on n or fewer data points, whereas the algorithm in section 6.5.2 essentially operates on a far larger class of designs and is therefore much more likely to find a global optimal design. It remains an open question whether something like the approximate information matrix idea is applicable in the present context, though remarks by Müller and Zimmerman (1999) and Müller (2000) suggest it will not be easy.

Appendix: Derivation of (6.53)

Ignoring an irrelevant constant, the negative log likelihood function is

$$\ell = \frac{1}{2} \log |V| + \frac{1}{2}(y - \mu)^T V^{-1} (y - \mu). \tag{6.58}$$

We assume y and μ have entries y_i and μ_i respectively, $V = (v_{ij})$ and $V^{-1} = (v^{ij})$. So (A1) can also be written

$$\ell = \frac{1}{2} \log |V| + \frac{1}{2} \sum_i \sum_j (y_i - \mu_i)(y_j - \mu_j) v^{ij}. \tag{6.59}$$

If V has cofactors V_{ij} then $\partial |V| / \partial v_{ij} = 2V_{ij} = v^{ij}|V|$ if $i \neq j$, and $\partial |V| / \partial v_{ii} = V_{ii} = v^{ii}|V|$. See, e.g., Mardia, Kent and Bibby (1979), equations (A.2.3d), (A.2.3e), (A.9.3). Also, $\sum_i \sum_j v_{ij} v^{ij} = n$, and it follows that

$$\sum_i \sum_j v_{ij} \frac{\partial v^{ij}}{\partial \theta_r} + \sum_i \sum_j \frac{\partial v_{ij}}{\partial \theta_r} v^{ij} = 0.$$

Hence,

$$\begin{aligned}
\frac{\partial}{\partial \theta_r}(\log |V|) &= \frac{1}{|V|} \frac{\partial |V|}{\partial \theta_r} \\
&= \frac{1}{|V|} \left\{ \sum_i \frac{\partial |V|}{\partial v_{ii}} \frac{\partial v_{ii}}{\partial \theta_r} + \sum_{i<j} \frac{\partial |V|}{\partial v_{ij}} \frac{\partial v_{ij}}{\partial \theta_r} \right\} \\
&= \sum_i \sum_j v^{ij} \frac{\partial v_{ij}}{\partial \theta_r} \\
&= -\sum_i \sum_j v_{ij} \frac{\partial v^{ij}}{\partial \theta_r}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta_r} &= -\frac{1}{2}\sum_i\sum_j v_{ij}\frac{\partial v^{ij}}{\partial \theta_r} - \sum_i\sum_j (y_i-\mu_i)\frac{\partial \mu_j}{\partial \theta_r}v^{ij} \\
&\quad + \frac{1}{2}\sum_i\sum_j (y_i-\mu_i)(y_j-\mu_j)\frac{\partial v^{ij}}{\partial \theta_r}, \\
\frac{\partial^2 \ell}{\partial \theta_r \partial \theta_s} &= -\frac{1}{2}\sum_i\sum_j \frac{\partial v_{ij}}{\partial \theta_s}\frac{\partial v^{ij}}{\partial \theta_r} - \frac{1}{2}\sum_i\sum_j v_{ij}\frac{\partial^2 v^{ij}}{\partial \theta_r \partial \theta_s} \\
&\quad + \sum_i\sum_j \frac{\partial \mu_i}{\partial \theta_s}\frac{\partial \mu_j}{\partial \theta_r}v^{ij} - \sum_i\sum_j (y_i-\mu_i)\frac{\partial^2 \mu_j}{\partial \theta_r \partial \theta_s}v^{ij} \\
&\quad - \sum_i\sum_j (y_i-\mu_i)\frac{\partial \mu_j}{\partial \theta_r}\frac{\partial v^{ij}}{\partial \theta_s} - \sum_i\sum_j (y_i-\mu_i)\frac{\partial \mu_j}{\partial \theta_s}\frac{\partial v^{ij}}{\partial \theta_r} \\
&\quad + \frac{1}{2}\sum_i\sum_j (y_i-\mu_i)(y_j-\mu_j)\frac{\partial^2 v_{ij}}{\partial \theta_r \partial \theta_s}
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{E}\left\{\frac{\partial^2 \ell}{\partial \theta_r \partial \theta_s}\right\} &= -\frac{1}{2}\sum_i\sum_j \frac{\partial v_{ij}}{\partial \theta_s}\frac{\partial v^{ij}}{\partial \theta_r} - \frac{1}{2}\sum_i\sum_j v_{ij}\frac{\partial^2 v^{ij}}{\partial \theta_r \partial \theta_s} \\
&\quad + \sum_i\sum_j \frac{\partial \mu_i}{\partial \theta_s}\frac{\partial \mu_j}{\partial \theta_r}v^{ij} + \frac{1}{2}\sum_i\sum_j v_{ij}\frac{\partial^2 v^{ij}}{\partial \theta_r \partial \theta_s} \qquad (6.60) \\
&= -\frac{1}{2}\sum_i\sum_j \frac{\partial v_{ij}}{\partial \theta_s}\frac{\partial v^{ij}}{\partial \theta_r} + \sum_i\sum_j \frac{\partial \mu_i}{\partial \theta_s}\frac{\partial \mu_j}{\partial \theta_r}v^{ij}.
\end{aligned}$$

The final result in (6.60) is of the form

$$-\frac{1}{2}\mathrm{tr}\left(\frac{\partial V}{\partial \theta_s}\frac{\partial V^{-1}}{\partial \theta_r}\right) + \frac{\partial \mu^T}{\partial \theta_r}V^{-1}\frac{\partial \mu}{\partial \theta_s}.$$

However, from $VV^{-1} = I_n$ it follows that

$$\frac{\partial V^{-1}}{\partial \theta_r} = -V^{-1}\frac{\partial V}{\partial \theta_r}V^{-1},$$

so the final formula for (6.60) — the (r,s) entry of the Fisher information matrix — is

$$m_{rs}(\theta) = \frac{1}{2}\mathrm{tr}\left(\frac{\partial V}{\partial \theta_r}V^{-1}\frac{\partial V}{\partial \theta_s}V^{-1}\right) + \frac{\partial \mu^T}{\partial \theta_r}V^{-1}\frac{\partial \mu}{\partial \theta_s}.$$

and this is (6.53).

6.6. Other approaches to network design

Apart from the systematic approaches to network design based on entropy (sections 6.1–6.3) or optimal design theory (sections 6.4 and 6.5), there have been numerous more *ad hoc* approaches to network design, some of which have been quite successful at producing algorithms for practical implementation. We focus on four developments here. Section 6.6.1 presents a relatively straightforward approach due to Haas (1992). In section 6.6.2, we discuss a series of papers by Oehlert, which are perhaps of as much interest for the approach they take to spatial-temporal modeling as for the actual application to network design. In section 6.6.3, we describe two approaches due to Nychka and Saltzman, which are focussed on efficient computation rather than the achievement of any theoretical optimality criteria. Section 6.6.4 presents a recent Bayesian approach due to P. Müller and co-authors, which makes the link with modern algorithms for Bayesian Monte Carlo sampling.

6.6.1 Haas's approach

Haas (1992) proposed a procedure for optimally selecting new locations in a monitoring network, using the NADP/NTN wet sulfate deposition network as an example. He considered a subregion approach, i.e. the nation is divided into subregions and a separate network optimization performed within each, arguing that this is appropriate both because subregions are more likely to be homogeneous in terms of physical processes and ecology, and also recognizing different policy objectives in different parts of the country. As optimization criteria, he proposed (a) the mean relative error of estimation (estimate standard error divided by estimate), and (b) the standard deviation of the relative error estimate at the subregion's center. The idea was that (a) is a measure of the quality of the network for spatial prediction, while (b) also takes account the error in estimating the spatial model, which may lead to quite different designs as already discussed in section 6.5.3.

For the actual spatial methodology, Haas (1992) used the earlier methods of Haas (1990a, 1990b) which are based on a combination of spatial regression and variogram estimation on the residuals within moving windows, though the basic concepts of Haas's design methodology could be applied in conjunction with any of the standard spatial prediction techniques. In the absence of any analytic method for assessing criterion (b) above, Haas used simulation. The problem of adding 10 new sites to an existing network was essentially formulated as a 20-dimensional nonlinear optimization problem (each new site is associated with two dimensions, its latitude and longitude) and standard nonlinear optimization methods applied. This methodology, although computationally intensive, was apparently successful at finding suitable design locations, but in the practical examples considered by Haas, the overall reduction of prediction standard error as a result of adding new stations was slight.

6.6.2 Oehlert's spatial-temporal model and characterization of designs by predictive variance

In a series of papers, Oehlert (1993, 1995, 1996) proposed a novel spatial-temporal model for sulfate deposition data, and applied it to various aspects of optimal network design.

Oehlert (1993) developed a model based on an assumption that the observed process is homogeneous within rectangles which are taken to be of 1 degree latitude and 1.5 degrees longitude. If y_i denotes the complete vector of observations at monitoring station i, then the assumed model for y_i is of the form

$$y_i = \alpha_{j(i)} + \beta_{j(i)}t + L + N_i + \delta_i, \tag{6.61}$$

where α_j and β_j are overall mean and time-trend parameters for rectangle j, t is a vector of regressors indicating time of observation (normalized to mean 0), L represents a random long-term trend that is assumed common to all stations, N_i is a random short-term temporal trend that is specific to station i, and δ_i denotes a constant term representing the bias in station i, where by bias we are referring to such things as elevation effects or proximity to sulfate sources, which could make station i different from the underlying field.

For the long-term random trend L, Oehlert assumed an ARMA(1,1) term with correlation structure $\rho_k = \rho_1\phi^{k-1}$ for $k \geq 1$. For the collection of short-term random series N_i, he assumed a covariance matrix of the structure $S \otimes C$ where S represents the spatial correlations and C the temporal correlations. For the temporal component, he assumed an MA(1) process which leads to a tridiagonal structure for C. For S, he assumed that spatial correlations were estimated either by a kernel smoothing process (though this method cannot be guaranteed to produce a positive definite covariance) or else the standard exponential spatial covariance structure $s_{ij} = \sigma^2$ when $i = j$, $\sigma^2 re^{-cd_{ij}}$ when $i \neq j$, where d_{ij} is the distance between stations i and j. The δ_i were treated as random effects with mean 0 and variance σ_δ^2.

For the estimation of the model (6.61), Oehlert proposed a two-stage process. First, a simple ordinary least squares regression was fitted at each site to obtain estimates a_i and b_i of the underlying mean and trend parameters $\alpha_{j(i)}$ and $\beta_{j(i)}$. However, the vectors of these estimates over all stations have complicated covariance structures, denoted Σ_a and Σ_b, which are functions of all the parameters mentioned above.

The second stage of the process involves taking the a_i and b_i coefficients as the raw data, and feeding them into an overall network model. It is assumed that the expected value of b_i is of the form $\beta_{j(i)} + M_k$, where M_k is an additional "network bias" indexed by the network k (the data do come from different networks), and similarly for a_i. For the covariance structure of the α_j and β_j parameters, Oehlert took a smoothness prior approach where these parameters are assumed to have specific forms of prior distributions, e.g.

$$\beta \sim N[0, \lambda_\beta A^T A], \tag{6.62}$$

where the matrix A has one column for each rectangle and one row for each pair of adjacent rectangles, and entries a_{ij} are 0 unless rectangle j is one of the adjacent pair i, and in that

case is valued either 1 or –1. The parameter λ_β (and a corresponding λ_α) is interpretable as a smoothness parameter. The monitor biases M_k are assumed independent with mean 0 and variance λ_M.

Based on this model, it is then possible to estimate the individual α_j and β_j, and the standard errors of the estimates, essentially by solving the generalized least squares equations in which the vaues a_i and b_i are treated as the input data and the covariance structure combines Σ_a, Σ_b and the parameters λ_α, λ_β and λ_M. Since this involves quite a few adjustable parameters, we need some method of determining those parameters. Oehlert proposed an indirect generalized cross-validation procedure (IGCV) of the form

$$IGCV = \frac{\frac{1}{n}(y - H_g y)^T \Sigma^{-1} (y - H_g y)}{\frac{1}{n}\mathrm{tr}(I - H_g)} \tag{6.63}$$

(Altman 1990, Engle *et al.* 1986) where H_g is the g'th predictor matrix (i.e. the predictors are of the form $\hat{y}_g = H_g y$) and Σ is the assumed covariance matrix for y. The idea is that g indexes a class of predictors and we choose the g which minimizes (6.63).

Despite the use of the IGCV criterion, it is clear that the selection of suitable hyperparameters is a critical aspect of the analysis, and in subsequent work, Oehlert often compared the results of different analyses assuming different hyperparameters.

Oehlert (1995) used the model of Oehlert (1993) to estimate the detectabilities of regional trend in current and in modified networks, based on hypothesized future trends and other model parameters. For example, one could ask what is the probability of detecting a certain trend by a given date (the detectability question), or what is the probability that an estimated change is within say 20% of its true value (quantifiability). Both of these questions hinge on determining the standard errors of trend estimates.

The model was fitted both to raw sulfate deposition data and to precipitation-adjusted data, with much better results in the second case. The procedure for precipitation adjustment was to fit a regression, at each site, of log sulfate deposition on a monthly indicator variable and the monthly mean of precipitation volume. Each month's value was then adjusted to the overall mean precipitation level, and the adjusted annual values were formed by taking volume-weighted means of the adjusted monthly values.

As criteria for assessing the performance of a network, Oehlert considered:

(a) the sum of regional variances for five large regions — this is a measure of the region-wide variability

(b) the sum of the largest 40 individual rectangle variances — this is a measure of local variability (the number 40 is arbitrary, but represents about one fifth of the available rectangles)

Within this framework, he considered the effects of deleting 10 stations sequentially, i.e. at each step, delete the station which produces the smallest increase in either of the criteria (a) or (b). In fact, he showed it was possible to delete up to 10 stations with very little change in the criteria, because there are dense stations in the Midwest/Great Lakes region and most of the deleted stations come from there.

Conversely, he also considered the effects of adding up to 10 stations, again taking a sequential approach. As candidate locations for adding stations, he took the center of each rectangle. He argued that a reduction of 10% to 15% in the variance sums was possible, mostly by adding stations along the east coast, though in this case, the results are sensitive to which of the two criteria, (a) or (b), is adopted.

Oehlert (1996) extended this analysis to the case of 100 deleted stations, out of a current network of 249. Again a sequential approach was adopted, and he showed that with optimal selection of the stations, the increases in criteria variances ranged between 7% and 34%, whereas in comparison, if stations were simply deleted at random the increases in variance were often of the order of a factor of 2. In this case, there are differences in the design depending on which of criteria (a) or (b) is adopted. Criterion (a) tends to keep stations in and near small regions, whereas (b) leads to a much more uniform distribution of stations. Neither method would delete stations from the boundary, however. Sequential selection using any of the possible criteria is better than random deletion, for all criteria. Oehlert emphasized that the purpose of the design selection is to optimize the variance of long-term trend esimation, and the results would not necessarily apply to spatial prediction on short temporal scales. A possibly controversial feature of his conclusions was that since they were based on the US network but also used Canadian data, many of the proposed deletions were near the Canadian border. This seems to assume that while the US is trying to save costs by reducing the size of its network, the Canadian government is happy to continue to maintain the same network in its borders!

6.6.3 The computational approach of Nychka and Saltzman

Nychka and Saltzman (1998) studied network design in the context of ozone monitors in the Midwest United States. Most of the existing ozone monitors were put into place to monitor the Environmental Protection Agency (EPA) ozone standard, which at the time of the study was based on exceedances by daily ozone maxima over the standard of 120 ppb (since reduced to a standard of 80 ppb for eight-hour daily averages). The enforcement of the standard is based on readings at the monitors and does not raise any issues for spatial interpolation. However, for broader purposes there are two reasons why spatial interpolation is desirable. The first is to compute some measure of the total ozone impact, for instance as an indicator of the total human health impact. This could be based on the average ozone over a region, or it could be based on some form of spatially weighted average (for instance, one weighted according to population density). In order to estimate such averages, we need a means of interpolating between monitors. The second reason why spatial modeling is desirable is from the point of view of assessing the agreement between real ozone data and the output of numerical models such as the EPA's Regional Oxidant

Model (ROM — since replaced by a more sophisticated model known as Models-3). Numerical models are important because they are used to assess the possible consequences of changes in pollution control strategies, something which cannot be assessed from pollution data. However, it is also important to validate the models on real data. The spatial interpolation problem arises because models typically produce simulations of average ozone over large grid boxes, and do not necessarily produce reliable predictions at specific sites. Therefore, it is desirable to interpolate the observed data over large grid boxes to make a meaningful comparison. Davis *et al.* (2000) have made a detailed study of ozone data and ROM output.

One feature observed by Nychka and Saltzman (1998) is that spatial isotropic models are not appropriate for these data. The solution they adopted was to fit a model of the form (3.5), which combines a standard (stationary, isotropic) geostatistical model with additional component from an empirical orthogonal functions decomposition. This model was fitted to the ROM data, before being used to aid in selecting stations for the real data observational network — thereby neatly sidestepping the issue of design for efficient estimation of spatial parameters. Instead, they focussed on two broad objectives, (a) prediction of the overall average ozone level with minimum variance, (b) bounding the error in spatial prediction at individual locations. They pointed out that these criteria are, respectively, similar to the A-optimality and G-optimality criteria of optimal design theory, but beyond that, they did not use any concepts of design theory.

Within this framework they considered three specific problems that had arisen in practical work with the EPA: to reduce the size of a 20-station urban network in and around Chicago, to expand a network of rural stations, and to asses the impact of possible changes in the network in the Great Lakes region.

First approach: regression and variable selection

The first approach they considered was to use a simple regression model to estimate the overall ozone level based on 20 monitors or some subset thereof. A model with the structure

$$Y = \alpha + X_J \beta_J + \epsilon, \tag{6.64}$$

could in practice perform almost as well as a full spatial prediction model, if the objective is to predict only a single variable Y. Here, the response Y is taken to be the average of the 20 stations in the full network, X_J is the subset of stations that are retained after reducing the network, and ϵ is treated as an independent random error at each time point. Nychka and Saltzman argued that the sums of squares and cross products that are involved in fitting the model (6.64) are sample estimates of the population variances and covariances that are involved in kriging, so regression based on (6.64) forms an easily implemented approximation to the spatial prediction problem.

The problem therefore reduces to a problem in classical regression theory, namely, to find the best set of predictors of a given size, in the sense of minimizing the residual

sum of squares. There are a number of algorithms for doing this without evaluating all possible subsets. One is the leaps and bounds algorithm of Furnival and Wilson (1974), implemented in the S language as `leaps` (Becker *et al.* 1988). Another procedure is Tibshirani's (1995) *lasso* procedure, whose objective is to minimize the residual sum of squares subject to a constraint of the form $\sum |\beta_j| \leq t$. Although this problem is not obviously related to the one of regression subset selection, in practice it tends to produce solutions for which several of the β_j are 0, and the size of the set for which $\beta_j \neq 0$ can be controlled by reducing the value of t. Nychka and Saltzman suggested scaling the X variables so that the least squares solution has $\sum |\beta_j| = 1$. The size of the subset can then be reduced by fixing some $t \in (0, 1)$.

These algorithms can identify designs that allow for drastic reductions in the number of observing stations. In their example, Nychka and Saltzman found a subset of 5 stations for which the residual standard deviation in predicting the overall overage was 2.5 ppb. This is to be compared with the unconditional variance of the overall average which was 16.1 ppb. In other words, the regression achieves an R^2 of nearly 98% based on a 75% reduction in the size of the network.

Second approach: use of space-filling designs

The second approach used by Nychka and Saltzman is more appropriate in the context of adding stations to an existing network or when the number of candidate sites is too large to apply something like the leaps or lasso procedures. Their basic philosophy was this: rather than try to optimize the network with respect to a very specific covariance function or design objective, try to find a design which fills up the available space (in some suitably defined sense), thus producing something that could be useful for a variety of purposes. They argued that such networks would typically produce near-optimal performance for prediction even when compared with networks designed for specific purposes. It should be noted, however, that such a philosophy would not be appropriate if estimation of the spatial model itself was an objective, since in that case, it is desirable to have some small interpoint distances to produce reliable variogram estimates at short distances (see section 6.5.3 for further discussion on this point).

To define what is meant by a space-filling design, it is necessary to specify some suitable metrics. Nychka and Saltzman defined

$$d_p(x, D) = \left(\sum_{u \in D} ||x - u||^p \right)^{1/p} \tag{6.65}$$

as the "distance" between a point x and a set of observing stations D. Here, p can be any negative number. For $p < 0$, $d_p(x, D)$ tends to 0 as x approaches any point of D, and as $p \to -\infty$, it is simply the distance from x to the closest point of D.

The second metric is of the form

$$C_{p,q}(D) = \left(\sum_{x \in C} d_p(x, D)^q \right)^{1/q}, \tag{6.66}$$

where $q > 0$ and the sum is taken over all x in the (assumed discrete) space of possible locations C. Then the objective is to choose D to minimize $C_{p,q}(D)$. When $p \to -\infty$ and $q \to \infty$, this becomes the "minimax" distance criterion, i.e. minimize the maximum distance from any point in C to the nearest point in $D \subset C$.

It should perhaps be pointed out that other authors have take some other approaches to the definition of space-filling designs. For example, Trujillo-Ventura and Ellis (1991) discuss space-filling as one of several possible objectives of a network, where their definition of space-filling is to choose the locations $x_1, ..., x_N$ to maximize

$$\sum_{i=1}^{N} \min_{j \neq i} |x_i - x_j|^{1/2}. \tag{6.67}$$

It is not clear whether this leads to anything substantially different from (6.65) and (6.66).

To implement this approach, Nychka and Saltzman proposed a simple Monte Carlo algorithm based on random additions and deletions to D (Royle and Nychka 1998), and argued that in many cases, such designs perform well from the point of view of spatial predictions. They even cited theoretical justification for this in Johnson *et al.* (1990). They also argued that $d_p(x, D)$ can have the rough interpretation of a spatial prediction error though they also said that some alternative metrics may produce even closer agreement, for example in this connection they mentioned

$$d(x, D) = \sum_{u \in C} \left[\sum_{x \in D} \left\{ 1 - \exp\left(-\frac{||x - u||}{\theta} \right) \right\}^{-1} \right]^{-1},$$

as a *covariance filling* criterion based on the exponential covariance function.

In their example, Nychka and Saltzman considered the effect of adding to an existing set of 163 stations within an overall candidate set of 420 locations. Such sets are too big to apply the leaps or lasso procedures. By adding only 5 stations to the network, the residual prediction standard deviation was reduced from 3.9 ppb to 3.0 ppb. Adding 10 stations to the original set of 163 only results in a marginally bigger decrease, to 2.8 ppb.

In the final part of their paper, they developed these methods further to consider optimal additions and deletions to a network of 168 stations in the Great Lakes region. Reducing the size of the network by half increased the median prediction error by only 10%.

Apart from the Monte Carlo algorithms used by Nychka and co-authors, there is a substantial literature on both random and deterministic arrangements designed to have space-filling properties. Examples of deterministic designs include Fibonacci sequences and latin hypercube designs. Bates *et al.* (1996) and Chapter 4 of Müller (2000) are among the authors to have surveyed these methods.

6.6.4 The Bayesian approach of P. Müller

In a series of papers, e.g. Bielza *et al.* (1999), Müller (1999), Sansó and Müller (1997), P. Müller and collaborators have developed an approach to optimal design based on Monte Carlo simulation over the design space, using modern ideas such as Hastings-Metropolis sampling to do this in an efficient way. We follow Sansó and Müller (1997) here, which used this idea to study the problem of reducing a set of 80 rainfall stations in Venezuela to one of 40 stations.

The approach is decision-theoretic: it requires a specific utility function to measure the quality of predictions. Suppose the current set of stations is divided into subsets D of the stations to be retained and the complement D^c which is to be discarded. We assume a variable y_i is defined at all locations and let y_D denote the observed data, i.e. the values of y_i for $i \in D$. Suppose $\hat{y}_i(y_D)$ is an optimal kriging predictor of y_i given y_D. The utility function proposed by Sansó and Müller is of the form

$$u(D, y) = C \sum_{i \in D^c} \mathcal{I}\left(|y_i - \hat{y}_i(y_D)| \leq \delta\right) - \sum_{i \in D} c_i + C_0,$$

where δ is a specified target for the prediction accuracy, C is interpreted as the payoff for getting a prediction that is within δ of the true value, c_i is the cost of operating station i and C_0 is an adjustable constant — for the algorithm to be described, it is necessary that $u(D, y) > 0$ for all D and y, and this can be forced by choosing appropriate C_0.

The actual criterion is to choose D to maximize $U(D)$, where $U(\cdot)$ is the risk function derived from u:

$$U(D) = \mathrm{E}\{u(D, y)\}.$$

In their application, y_i was taken to be the logarithm of the annual rainfall total and they assumed a model (for the total rainfall field) of the form

$$y \sim N[X\beta, \sigma^2 V(\lambda)],$$

where $X\beta$ is a regression component with covariates latitude and longitude, and $V(\lambda)$ was taken to be the exponential correlation function, $v_{ij} = \exp\{-\lambda d_{ij}\}$ where d_{ij} is the distance between stations i and j. A Bayesian analysis proceeds using conjugate priors for β and σ^2 and numerical integration with respect to λ, along the same lines as Handcock and Stein (1993). It turned out that the posterior distribution of λ had very small dispersion so in subsequent analysis, λ is fixed at its posterior mean value, which significantly simplifies the calculations. We let $\theta = (\beta, \sigma^2)$ denote those parameters for which significant posterior variability is assumed.

A naïve simulation approach would generate random samples (θ_m, y^m), $1 \leq m \leq M$ from the posterior distribution of θ and the resultant predictive distribution of $y|\theta$, and approximate

$$U(D) \approx \frac{1}{M} \sum_m u(D, y^m), \tag{6.68}$$

followed by choosing D to maximize (6.68). This is a highly inefficient procedure and seems unlikely to lead to satisfactory results.

Instead, the idea of Sansó and Müller was to embed the design selection problem into an artificial Bayesian inference problem in order to use modern sampling techniques for Bayesian inference. Consider a probability density $h(d, \theta, y)$ defined so that

$$h(D, \theta, y) \propto p(\theta)p(y|\theta)u(d, \theta, y), \tag{6.69}$$

where the $p(\theta)$ denotes the posterior density of θ given the past data, and $p(y|\theta)$ is the conditional distribution of some future observation vector given θ. The idea is to interpret (6.2) as a joint distribution of θ, y and D itself, and then to create a "posterior distribution" of D by Monte Carlo sampling. Some reasonable approximation to the mode of the marginal posterior distribution of D will then be our guess as to the best design.

In more detail, the idea is to perform a Hastings-Metropolis sampler on D (see, e.g. Tierney (1994) or Gilks *et al.* (1996)) combined with a more conventional updating sampling from the posterior distribution of θ. The trial distribution used for the Metropolis updating step was based on simple switching of one station between D and D^c — Sansó and Müller remarked that this is not really a satisfactory approach because it is unlikely that such an algorithm would explore a sufficiently large range of the design space, but they also remarked that this deficiency seems to be shared by most of the other network design algorithms at the present time, when the number of possible sites is moderately large. A Monte Carlo sample of values of D was generated, and a cluster analysis performed using a hierarchical clustering algorithm. The mode was selected from amongst the values in the largest cluster.

In discussion, they identified two major problems with this procedure, (a) the difficulty of defining a suitable updating distribution for D, (b) finding the posterior mode among the resulting sample. As partial solution to the first problem, they also considered a procedure where up to eight stations were randomly added to or deleted from the network as a single trial step, arguing that this would produce faster coverage over the space of possible designs.

6.7. Designs for data assimilation

This section gives a short discussion of a very extensive problem in atmospheric science, known as data assimilation. It arises particularly in the context of numerical weather forecasting, where the problem arises of how best to incorporate real observational data into a numerical weather model. However, it is also relevant to other applications of environmental modeling, such as ozone forecasting using such systems as the EPA's Models-3 system. The treatement given here essentially follows Berliner, Lu and Snyder (1999), henceforth BLS, though with some modifications.

Suppose X_0 is some p-dimensional vector representing the state of the weather at time t_0. At some time $t_1 > t_0$, we will take a q-dimensional observation Y whose distribution will depend on X_1, the state of the weather at time t_1. This observation in turn will be used to forecast the weather X_2 at some time $t_2 > t_1$. The problem is how to decide which Y to take, from a set of possible choices, to optimize the prediction of X_2.

This formulation was motivated by the so-called FASTEX study (Fronts and Tropical Storm-Tracks Experiment — Snyder (1996), Joly *et al.* (1997)), which involved the prediction of storm conditions in western Europe based on observations taken over the North Atlantic. At the center of the experiment were two long-range aircraft, which could fly through the developing storm system and take additional observations, which would supplement the standard observational network. The problem was to design the flight path so that the experiment would yield maximum information in the sense of improving subsequent weather forecasts. In the above formulation, the observations Y should be understood as including both the aircraft observations and the standard ground-based observations — only a portion of Y, that collected by the aircraft, is in any way under the experimenter's control.

The problem is extremely complicated for a variety of reasons. First, the dimensions are very high: $p \approx 10^7$, $q \approx 10^5$. Standard statistical techniques such as the Kalman filter are not easily implemented in such high dimensions. Indeed, it seems that the effective implementation of the ideas to be discussed here will require new developments in numerical analysis, but we shall ignore that aspect in our discussion.

A second complication is that the observed systems are in reality highly nonlinear, and indeed chaotic, in the sense that small perturbations in the initial state X_0 have the potential to create much larger perturbations in future observations of the system. This aspect, however, is also largely ignored in the discussion to follow, though it may imply that the matrices A_0 and A_1 below are not well conditioned.

A typical evolution equation might be written in the form

$$X_1 = F_0(X_0), \tag{6.70}$$

where F_0 is a nonlinear function, found numerically from a weather forecasting model. If we represent the current state of knowledge about X_0 by a multivariate normal distribution,

$$X_0 \sim N_p(\mu_0, U_0), \tag{6.71}$$

then a linear approximation to (6.70) suggests

$$X_1 - F_0(\mu_0) \approx \nabla F_0(\mu_0) \cdot (X_0 - \mu_0),$$

(∇F_0 denotes the gradient matrix of F_0) then combined with (6.71) we have, approximately,

$$X_1 \sim N_p(F_0(\mu_0), \nabla F_0(\mu_0) U_0 \nabla F_0(\mu_0)^T), \tag{6.72}$$

One could extend (6.72) by allowing for additional system noise, leading to an equation

$$X_1 \sim N_p(F_0(\mu_0), V_0 + \nabla F_0(\mu_0) U_0 \nabla F_0(\mu_0)^T), \tag{6.73}$$

The matrix V_0 is not present in the BLS development, but it seems a natural addition, for two reasons. First, even though the dynamics of the model are deterministic, in practice we do not know either F_0 or ∇F_0 analytically, and must estimate them from numerical experiments. The matrix V_0 can be thought of a representing our uncertainty about the true dynamics.

Second, we have already mentioned that the dimension p of the numerical model is very high, possibly too high to permit any realistic statistical analysis. It may be necessary to reduce p is some way, for example, by grouping the weather system variables into coarser grid cells than are used in the actual numerical model. However, using coarse grid cells creates the problem of how to deal with phenomena that are known to exist at a smaller scale than that of the grid cell. In the numerical modeling literature, this leads to so-called closure methods or "parameterizations" which can often be thought of as stochastic adjustments to the model. A famous example in another context is the role of clouds in climatological models. These models typically are contructed with grid cells too large to allow for cloud modeling, but it is certain that clouds exert a significant influence on heat flow through the atmosphere. One way to bring clouds into the model is to develop a stochastic formulation of the impact of clouds conditional on the boundary conditions which the model does produce. In the context of (6.73), the covariance matrix V_0 could represent perturbations due to phenomena that are deterministic but which cannot be adequately represented by the model on the grid scale considered, so for practical purposes they are stochastic.

After these initial discussions, we now propose a specific model, which is based on linearizing F as well as the operations producing Y and X_2, to produce:

$$X_0 \sim N_p(\mu_0, U_0), \tag{6.74}$$
$$X_1 = A_0 X_0 + \epsilon_0, \quad \epsilon_0 \sim N_p(0, V_0), \tag{6.75}$$
$$Y = B X_1 + \eta, \quad \eta \sim N_q(0, W), \tag{6.76}$$
$$X_2 \sim A_1 X_1 + \epsilon_1, \quad \epsilon_1 \sim N_p(0, V_1), \tag{6.77}$$

where ϵ_0, ϵ_1 and η are independent and the matrices A_0, A_1, B, V_0, V_1, W are all assumed known. Since Y is partially under the experimenter's control, we can assume that B and W are functions of the design D, and on occasion we will write explicitly $B = B_D$ and $W = W_D$ to emphasize this dependent. All the other quantities in (6.74)–(6.77), however, are fixed.

To analyze this model, we first note that X_0 only enters the discussion as providing a prior distribution for X_1, so we can combine (6.74) and (6.75) immediately into

$$X_1 \sim N_p(\mu_1, U_1), \tag{6.78}$$

where

$$\mu_1 = A_0\mu_0, \quad U_1 = V_0 + A_0 U_0 A_0^T, \tag{6.79}$$

analogous to (6.73).

Next, combining (6.79) with (6.76) yields

$$Y \sim N_q(B\mu_1, W + BU_1B^T), \tag{6.80}$$

while combining (6.79) with (6.77) yields

$$X_2 \sim N_p(A_1\mu_1, V_1 + A_1U_1A_1^T). \tag{6.81}$$

For the covariance of Y and X_2, since η and ϵ_1 are independent we have

$$\begin{aligned} \mathrm{E}\{(Y - B\mu_1)(X_2 - A_1\mu_1)\} &= \mathrm{E}\{B(X_1 - \mu_1)(X_1 - \mu_1)^T A_1^T\} \\ &= BU_1A_1^T. \end{aligned} \tag{6.82}$$

Therefore, combining (6.80)–(6.82), the joint distribution of Y and X_1^T is

$$\begin{pmatrix} Y \\ X_2 \end{pmatrix} \sim N_{q+p}\left[\begin{pmatrix} B\mu_1 \\ A_1\mu_1 \end{pmatrix}, \begin{pmatrix} W + BU_1B^T & BU_1A_1^T \\ A_1U_1B^T & V_1 + A_1UA_1^T \end{pmatrix}\right].$$

Hence, the conditional distribution of X_2 given Y has mean

$$\mathrm{E}\{X_2|Y\} = A_1\mu_1 + A_1U_1B^T(W + BU_1B^T)^{-1}(Y - B\mu_1), \tag{6.83}$$

and covariance matrix

$$S = V_1 + A_1U_1A_1^T - A_1U_1B^T(W + BU_1B^T)^{-1}BU_1A_1^T. \tag{6.84}$$

If we denote the dependence on the design D by writing S_D in place of S, B_D in place of B amd W_D in place of W, then the design problem becomes to choose D to maximize

$$\Phi(S_D) = \Phi\left\{V_1 + A_1U_1A_1^T - A_1U_1B_D^T(W_D + B_DU_1B_D^T)^{-1}B_DU_1A_1^T\right\} \tag{6.85}$$

for some suitably chosen design criterion Φ. For example, $\Phi(S) = |S|$ would be analogous to D-optimality, $\Phi(S) = \mathrm{tr}(S)$ to A-optimality, and so on.

The derivation of S given here is a little simpler than that of BLS, partly because we adopted a simpler model — in (6.75) and (6.77) we assumed the functions A_0X_0 and A_1X_1 were linear functions of X_0 and X_1 respectively, whereas BLS did not assume that initially. However, to derive their updating equations they were effectively forced to assumed linear updating equations, and their equation (2.22) is the same as (6.84), except for the system noise matrix V_1 which was not present in their formulation.

BLS proceeded at this point using the A-optimality criterion, but there are some additional mathematical properties that might still make it more appropriate to focus on D-optimality. For example, using (6.84) we can write

$$\begin{aligned}|S| =&|V_1 + A_1U_1A_1^T| \cdot |W + BU_1B^T|^{-1} \cdot \\ &\cdot |W + BU_1B^T - BU_1A_1^T(V_1 + A_1U_1A_1^T)^{-1}A_1U_1B^T|.\end{aligned} \tag{6.86}$$

The first factor in (6.86) is the determinant of a $p \times p$ matrix, but this is not affected by the design and can therefore be ignored in design calculations. The second and third factors in (6.86) are both determinants of $q \times q$ matrices, which should be much easier to calculate. We still have to invert the $p \times p$ matrix $V_1 + A_1U_1A_1^T$, but this has to be done only once, not repeated for each candidate design.

In fact, in a typical case we will be able to subdivide Y as $(\, Y_1^T \quad Y_2^T \,)^T$ where Y_1 of dimension $q - d$ represents the fixed weather observations and only Y_2 of dimensions d is under the experimenter's control. If we partition W and B correspondingly as

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

Then

$$W + BU_1B^T = \begin{pmatrix} W_{11} + B_1U_1B_1^T & W_{12} + B_1U_1B_2^T \\ W_{21} + B_2U_1B_1^T & W_{22} + B_2U_1B_2^T \end{pmatrix}$$

and hence

$$\begin{aligned}|W + BU_1B^T| &= |W_{11} + B_1U_1B_1^T| \cdot \\ &\cdot |W_{22} + B_2U_1B_2^T - (W_{21} + B_2U_1B_1^T)(W_{11} + B_1U_1B_1^T)^{-1}(W_{12} + B_1U_1B_2^T)|.\end{aligned} \tag{6.87}$$

The first determinant in (6.87) is not affected by the design D, so for optimization purposes we only have to take account of the second factor.

A similar simplification is possible for the third factor in (6.86), if we replace the matrix U_1 in (6.87) by

$$U_1 - U_1A_1^T(V_1 + A_1U_1A_1^T)^{-1}A_1U_1$$

which is also independent of the design D.

With these simplifications, the calculation of a $q \times q$ determinant is reduced to one of $d \times d$. If $q \approx 10^5$, $d \approx 50$, this could be quite a saving!

Example

BLS gave a numerical example of their procedures by re-analyzing a numerical experiment due to Lorenz and Emanuel (1998), henceforth LE. This involved a set of differential equations

$$\frac{dx_i}{dt} = -x_{i-2}x_{i-1} + x_{i-1}x_{i+1} - x_i + F, \tag{6.88}$$

where F is a forcing constant and x_i are assumed to satisfy a periodicity condition

$$x_{i+n} = x_i. \tag{6.89}$$

An interpretation is that these are n equally spaced weather readings around a hypothetical equator. For numerical experiments, BLS, following LE, set $n = 40$ and $F = 8$, at which the dynamics of the system are fairly chaotic. In their framework, a time period of $t = 0.2$ is one "day", and they assumed measurements every six or twelve hours.

In the formulation of LE, sites 1–20 are "ocean" sites and 21–40 "land". The land sites are routinely measured but the ocean sites are not. However, the possibility exists to measure one additional ocean site each time, and the question arises of where to take it to optimize the system's forecasting capability. LE assumed a current projection or "analysis" is available of the values of x_i, but at the sites where measurements are taken, x_i is replaced by y_i, and the whole system projected forwards using the numerical model. Since the whole exercise was a simulation, they were then able to generate "forecast errors" for a period up to 10 days ahead, by comparing the hypothetical weather forecasts with the values obtained using the true x_i. They then considered the effect of adding one ocean observation using two strategies, (a) select the ocean site at random, (b) select the ocean site to be where the current forecasting error is maximized. Strategy (b) would not be realizable in practice, because it still uses the simulations to determine where the forecasting error is largest, but if it could be justified as an appropriate strategy, maybe other methods such as ensemble forecasting (i.e. rerun the simulation from a selection of starting values in the neighborhood of the analysis values x_i) could be used to identify the appropriate location. In the event, LE produced a number of numerical results under strategies (a) and (b), finding that (a) did not greatly reduce the forecast errors (compared with no ocean sampling), but (b) reduced the forecast errors significantly.

As BLS pointed out, this conclusion could have been anticipated from optimal design theory, since the criterion of LE was based on looking at forecast errors at individual times and places, and this is similar to G-optimality. However, we have seen (section 6.4) that G-optimality is equivalent to D-optimality, and may be achieved in practice by adding observations to sites of high forecast errors. Hoewever, it is not clear how LE would extend their strategy to the case of multiple ocean observations.

The experiments by BLS took a similar starting point to those of LE, but used the A-optimal design found by from their equivalent of (6.84). They also compared their method with another data assimilation strategy due to Palmer *et al.* (1998). A one-sentence summary of their conclusions would be that their strategy was comparable with the others when used to reconstruct the current state of the system, but generally superior when used to forecast ahead in time.

6.8. Summary and conclusions

This chapter has surveyed a variety of approaches to the design optimality problem which have complementary strengths. The maximum entropy approaches represent the most sophisticated *formulation* of the optimal design problem but there are still some shortcomings in their actual implementation. For example, in general they employ simple strategies of adding and dropping stations one at a time, which may not lead to the optimal subset over a large class of candidate monitoring sites. However, in cases when the maximum entropy criterion reduces to the determinant of a posterior covariance matrix, there have been recent developments of optimal subset algorithms which may point towards new approaches. Another weakness of the implementation of maximum entropy ideas is that although they employ in principle the concepts of Bayesian hierarchical models, in practice the top-level (hyperparameter) stage of the model is limited to the so-called type II maximum likelihood or empirical Bayes method, in which the hyperparameters are replaced by point estimates without further consideration of their variability. This restriction is probably not too important in the "network reduction" context, in which prior data are available over all the sampling sites of interest, since in that case, the second or normal-Wishart stage of the hierarchy surely produces adequate posterior distributions for the means and covariances of those sites. However, when the normal-Wishart sampling ideas are extended to include prediction off the network, as originally laid out by Le and Zidek (1992), it really is a key point that the conditional distributions of the ungauged sites, given the gauged sites, are not in any way updated by the existing data on gauged sites, in other words, one is relying on the top level of the hierarchical model to gain any information about those conditional distributions. From this point of view, it is worth noting that they nowhere consider the role of the design in estimating hyperparameters, such as the parameters of a fitted variogram model. As we saw in section 6.5.3, this aspect of the design problem has been considered in other parts of the literature.

Another question related to the maximum entropy approach is whether a sophisticated criterion like maximizing the entropy of the joint predictive distribution at all the ungauged sites is really a better criterion than something much easier to state and to understand, such as choosing the design to minimize the maximum variance of the prediction error. As we saw in section 6.6, a number of the *ad hoc* approaches to network design use that or similar criteria, and it is hard to argue against such approaches.

In contrast to the maximum entropy idea, the more classical design optimality criteria use simpler concepts to define optimal designs, but have developed more sophisticated algorithms for their computation. The original idea which motivated the whole theory of optimal experimental design was that by generalizing the concept of a design measure to arbitrary normed positive measures over the design space, it was possible to use functional-analytic methods of optimization to calculate optimal designs. Optimal design theory was first applied to the spatial monitoring context by Fedorov and Müller (1989), but the main conclusion of that paper — which is that classical D-optimal designs may be directly applicable in a spatial context — does not hold when more realistic spatial models are considered. The difficulty is that classical optimality criteria tend to produce designs which involve replications at a relatively small number of design points, which is fine in a classical experimental design context that really does involve independent replications,

but not in a spatial sampling context where there is nothing to be gained by repeated sampling at the same locations. Instead, recent research has tried to adapt the functional-analytical algorithms of classical design theory to context of optimal sampling in a random field. That there is some hope of doing that is shown by the recent papers of Pázman and W. Müller as reviewed in the book of Müller (2000) and section 6.5.2 here, though the most sophisticated algorithms still only apply to the context of estimating regression parameters in cases when the covariance function of the spatial field is known. When the problem is extended to include estimation of covariance or variogram parameters, recent work reviewed in section 6.5.3 shows that it is possible to apply optimal design criteria also in this context, though the algorithms at the moment are restricted to one-at-a-time adding or deleting of points and cannot therefore be guaranteed to find globally optimal designs. As briefly discussed in this section, it seems possible to apply such methods directly to the Fisher information matrix for all the unknown parameters including the regression coefficients, but this idea does not appear to have been tried in practice so far.

Maximum entropy approaches to design of experiments have been gaining increasing attention in the design literature, e.g. a paper by Shewry and Wynn (1987) showed how to apply maximum entropy sampling in some simple contexts and this was also used in the paper by Sacks *et al.* (1989) on the optimal design of computer experiments. A recent paper by Sebastiani and Wynn (2000) has introduced ideas quite similar to Caselton, Kan and Zidek (1992) in more general design contexts. The relationship between different theoretical concepts of design, in particular between design optimality and decision theoretic criteria, has been explored from a theoretical point of view by Dawid and Sebastiani (1999).

Of the other ideas reviewed in section 6.6, the first method due to Nychka and Saltzman, based on regression subset selection, seems simple and appealing in cases when it is applicable, which are, essentially, only for the network reduction problem (because the method assumes availability of data at all the sampling locations of interest) and only then when the network is not too large. The second Nychka-Saltzman method is based on the idea of space-filling designs, which other authors have also proposed as a simple alternative to more sophisticated design optimality ideas when the objective is optimal spatial prediction — these designs seem less appropropriate for the purpose of estimating the covariance structure. Of the other recent contributions to the problem, the work of P. Müller and his co-authors (e.g. Sansó amd Müller 1997) makes the connection with modern computational methods for Bayesian hierarchical models, using the utility function for a particular design to define a "prior distribution" over designs and sampling the design along with the model parameters. The idea has some connections with simulated annealing which, though not mentioned in the present review, is also a standard algorithmic approach to solving hard combinatorial optimization problems (Brooks and Morgan (1995) have a nice review of simulated annealing from a statistical point of view).

The possibility of applying any of these ideas in the context of data assimilation, as discussed in section 6.7, is fairly new at the present time, but it is clear that that problem

could be generalized to take into account the effect of estimating the model, entropy rather than design-optimality criteria, and a variety of other aspects of the problem.

Perhaps the most obvious gap on the literature at this point is the paucity of approaches that go beyind the traditional contexts of multivariate normal distributions and regression models. The work of P. Müller could in principle be applied in any context where it is possible to define a utility function, and some other specialized approaches have been developed for specific problems, e.g. Schumacher and Zidek (1993) consider optimal designs to maximize the power of an F test in the context of assessing of some environmental "intervention" such as drilling a new oil well. However, most of the other methods use normal theory to formulate a design criterion. In an environmental regulatory context, it may be more appropriate to focus on issues such as standards enforcement. Current approaches to sampling environmental fields, such as the EMAP surface waters procedures (Baker *et al.* (1997)) use probability sampling methods but make essentially no use of any kind of "optimal design" concept. However the development of such concepts, suitable for that context, would surely require much more emphasis on extreme values and the probability of detecting the violation of an environmental standard than is traditional in the design of experiments. These are largely unexplored questions, at least from a formal mathematical point of view, but there is surely much scope for future research.

CHAPTER 7

Trends in Climatological Time Series

Many climatological time series contain apparent trends. Indeed, the detection of such trends is often the first step in making statements about climate change. However, it is obvious that climatological time series are also autocorrelated, and an attempt to identify trends without taking this into account would be seriously flawed. Of course, the problem of detecting trends in time series is an old and classical problem of time series analysis, but a number of recent developments have been stimulated specifically by climatological applications.

In this chapter, we focus on three approaches. Section 7.1 describes what we might call classical approaches, in which the residuals from the trend are assumed to follow a traditional time series model of autoregressive or ARMA (autoregressive–moving average) structure, though we also include some references in which the model was taken to be a fractionally differenced ARMA process, which incorporates long-range dependence. The phenomenon of long-range dependence — in which the autocorrelations of a time series are assumed to decay to 0 more slowly than is consistent with any ARMA process — is of importance because many climatological time series appear to show this behavior. Therefore, Section 7.2 develops the properties of long-range dependent processes in more detail, using spectral methods. Finally in this chapter, Section 7.3 discusses an alternative modeling approach applied to bivariate time series, in which the objective is less concerned with detailed time series modeling and more with the discrimination among a variety of trend terms representing different combinations of climate forcing signals. The final conclusion of this analysis is that to provide adequate explanation of observed trends in northern and southern hemispheric average temperatures, it is necessary to include all of the three most important forcing factors of current climate research: greenhouse gases, sulfate aerosols and solar fluctuations.

7.1 Classical time series approaches

7.1.1 The Cochrane-Orcutt model

An early and classical reference to regression with autocorrelated errors was Cochrane and Orcutt (1949). In modern terminology, their contribution was to advocate the use of generalized least squares (GLS) estimates over ordinary least squares (OLS) estimate of the regression parameters in such circumstances. If the model is written in matrix form as

$$y = X\beta + \nu,$$

where ν has mean 0 and covariance matrix U, then the GLS estimate chooses β to minimize

$$(y - X\beta)^T U^{-1} (y - X\beta), \tag{7.1}$$

in contrast with OLS which omits the U^{-1}. In the case of an AR(1) model,

$$\nu_t = \phi \nu_{t-1} + \epsilon_t,$$

with $|\phi| < 1$ and ϵ_t independent with variance σ_ϵ^2, Cochrane and Orcutt showed that in stationarity, assuming a data series of length n,

$$U = \frac{\sigma_\epsilon^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \dots & \dots & \phi^n \\ \phi & 1 & \phi & \dots & \dots & \phi^{n-1} \\ \phi^2 & \phi & 1 & \dots & \dots & \phi^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi^n & \phi^{n-1} & \phi^{n-2} & \dots & \dots & 1 \end{pmatrix},$$

$$U^{-1} = \frac{1}{\sigma_\epsilon^2} \begin{pmatrix} 1+\phi^2 & -\phi & 0 & \dots & \dots & 0 \\ -\phi & 1+\phi^2 & -\phi & \dots & \dots & 0 \\ 0 & -\phi & 1+\phi^2 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & \dots & -\phi & 1+\phi^2 \end{pmatrix}.$$

In fact, one can do more than this: Cochrane and Orcutt simply discussed how to estimate β without also considering the estimation of ϕ and σ_ϵ^2, but it is easily verified that $|U| = \sigma_\epsilon^{2n}$, so that a minor modification of (7.1),

$$\ell(\beta, \phi, \sigma_\epsilon^2) = \frac{n}{2} \log \sigma_\epsilon^2 + \frac{1}{2} (y - X\beta)^T U^{-1} (y - X\beta), \tag{7.2}$$

is in fact the joint negative log likelihood of all the unknown parameters. Estimation of the full model proceeds by minimizing (7.2) jointly with respect to $(\beta, \phi, \sigma_\epsilon^2)$.

From a modern point of view, an exact maximum likelihood procedure will implicitly subsume the Cochrane-Orcutt calculations, since exact MLE includes GLS for the regression coefficients. In practice, as already discussed in Chapter 1, it is common to use some approximate form of likelihood, e.g. a conditional likelihood, and such methods are always asymptotically equivalent to maximum likelihood in large samples, though exact MLE may be superior in small samples. Brockwell and Davis (1991) is an excellent reference on exact likelihood for ARMA models. Karl *et al.* (1996), amongst others, have applied these ideas to the analysis of meteorological time series.

7.1.2 The Bloomfield and Bloomfield-Nychka approaches

Turning now to more modern references that have explicitly considered the role of time series regression in climate change studies, one of the first studies was by Bloomfield (1992), who considered models of form

$$y_t = \mu + x_t(\theta) + \nu_t, \tag{7.3}$$

where y_t is the observed mean temperature in year t, $x_t(\theta)$ is some signal defined by a parameter θ, and ν_t is a stationary time series representing the error. The signal could be something simple such as a linear trend, $x_t(\theta) = \theta t$, or it could be more complicated, such as the output of a climate model.

For the noise series ν_t, Bloomfield considered an AR(p) process

$$\nu_t = \sum_{r=1}^{p} \phi_r \nu_{t-r} + \epsilon_t, \tag{7.4}$$

with $\{\epsilon_t\}$ i.i.d., and fitted the model (7.3) for various orders p. In an example using the then-current version of the IPCC (Intergovernmental Panel on Climate Change) temperature series from 1861 to 1989 (Nicholls *et al.* 1996), fitting a linear trend, he found that $p = 1$ was clearly inadequate but the AIC criterion selected $p = 4$, which seemed a good fit overall.

An alternative model considered by Bloomfield was the fractional ARIMA process (Granger and Joyeux 1981, Hosking 1981). The simplest form of this is fractionally differenced noise,

$$(I - B)^d \nu_t = \epsilon_t, \tag{7.5}$$

where B is the backshift operator ($B^k \nu_t = \nu_{t-k}$), and $(I - B)^d$ is interpreted as a binomial expansion,

$$(I - B)^d = I - dB + \frac{d(d-1)}{2} B^2 - \frac{d(d-1)(d-2)}{6} B^3 + ... \tag{7.6}$$

The range of stationarity and invertibility is $-\frac{1}{2} < d < \frac{1}{2}$, with $d > 0$ interpreted as "long-range dependence".

The ARIMA(p, d, q) process refers to the case that $(I - B)^d \nu_t$ is not white noise but an ARMA(p, q) process,

$$\left(I - \sum_{r=1}^{p} \phi_r B^r \right) (I - B)^d \nu_t = \left(I + \sum_{s=1}^{q} \theta_s B^s \right) \epsilon_t. \tag{7.7}$$

In practice, fractional noise and fractional ARIMA processes are usually fitted not directly from the expansion (7.6), but in terms of their correlation functions or equivalently spectral densities; the spectral density of fractional noise is

$$\frac{\sigma_\epsilon^2}{2\pi} |1 - e^{i\lambda}|^{-2d}. \tag{7.8}$$

Bloomfield selected the ARIMA$(0, d, 1)$ model for the IPCC series, in which he found $d \approx 0.25$. Based on this he found an estimated linear trend of 0.37 (95% confidence band: 0.24 to 0.50). The corresponding results based on an AR(4) model were very little different: a point estimate of 0.38 and a confidence interval 0.25 to 0.51.

Another well-known global temperature series is the Hansen-Lebedeff (1987, 1988) series of surface mean temperatures from 1880 to 1987. Repeating the same analyses to this series, Bloomfield found a linear trend of 0.57 (95% confidence interval, 0.37 to 0.76) based on the ARIMA$(0, d, 1)$ model. In this case, AIC for an AR(p) model selected $p = 8$, for which the estimated trend was 0.58 and confidence interval 0.36 to 0.80. Once again, there is very little difference for the two time series approaches.

In another analysis, Bloomfield fitted model (7.3) in which $x_t(\theta)$ was taken as the output of a climate model indexed by "climate sensitivity" θ. The climate sensitivity, usually denoted $\Delta T_{2\times}$, is defined as the rise in the earth's mean temperature, in equilibrium conditions, corresponding to a doubling of atmospheric carbon dioxide since pre-industrial conditions. It plays the role of a variable parameter in most climate models; however, considerable time and energy has been expended in finding the best value of climate sensitivity based on the agreement between climate model output and real data. Bloomfield used the climate model of Wigley and Raper (1990) and, by fitting the model (7.3) as a nonlinear regression model with time series errors, deduced a point estimate of $\Delta T_{2\times}$ of 1.39°C, with a 95% confidence interval of 0.69 to 2.19. This was in good agreement with estimates available at the time; the advantage of a formal time series fitting is that the range of variability can be expressed precisely as a confidence interval, in preference to more informal assessments of uncertainty. In 2001, it is generally accepted that $\Delta T_{2\times}$ lies within 1.5°C and 4.5°C based on more sophisticated models using finer discretizations and allowing for other forcing factors.

Bloomfield and Nychka (1992) developed a more comprehensive approach to assessing the influence of time series dependence on an estimated trend, using spectral densities.

The idea is based on the following calculations. Suppose one has a linear estimator $\tilde{\theta}$, calculated from observations $\{x_t,\ t = 1, ..., T\}$ through a formula of form

$$\tilde{\theta} = \sum y_t u_t \tag{7.9}$$

where the u_t are fixed constants. We deduce

$$\mathrm{Var}\{\tilde{\theta}\} = \sum_{s=1}^{T} \sum_{t=1}^{T} u_s u_t \gamma_{|s-t|}. \tag{7.10}$$

where $\{\gamma_k\}$ denotes the autocovariance function.

One could try to evaluate (7.10) using the sample autocovariances, but this is generally considered to be a bad idea because the sample autocovariances have rather poor sampling

properties. A better idea is to substitute the theoretical γ_k's based on a fitted model, for example, the parameters estimated in an AR(p) model fit. An alternative approach, however, is based on the spectral density. From the formula

$$\gamma_k = \int_{-\pi}^{\pi} e^{i\lambda k} f(\lambda) d\lambda,$$

where $f(\lambda)$ is the spectral density, we see that

$$\begin{aligned} \sum_{s=1}^{T}\sum_{t=1}^{T} u_s u_t \gamma_{s-t} &= \sum_{s=1}^{T}\sum_{t=1}^{T} u_s u_t \int_{-\pi}^{\pi} e^{i\lambda(s-t)} f(\lambda) d\lambda \\ &= \int_{-\pi}^{\pi} U(\lambda) f(\lambda) d\lambda \end{aligned} \tag{7.11}$$

where

$$U(\lambda) = \left| \sum_{t=1}^{T} u_t e^{i\lambda t} \right|^2 . \tag{7.12}$$

For the AR(p) model,

$$f(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \left| 1 - \sum_{r=1}^{p} \phi_r e^{ir\lambda} \right|^{-2},$$

while in the fractional ARIMA case,

$$f(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \left| \frac{1 + \sum_{s=1}^{q} \theta_s e^{is\lambda}}{1 - \sum_{r=1}^{p} \phi_r e^{ir\lambda}} \right|^2 \cdot |1 - e^{i\lambda}|^{-2d}, \tag{7.13}$$

from which the integrand in (7.11) is easily evaluated and the integral itself may then be found by numerical integration. As an example, Fig. 7.1 shows the functions $U(\lambda)$ and $f(\lambda)$ for a temperature series of 149 years in Amherst, MA (Lund *et al.* 1995), for which an AR(2) model has been fitted and the estimator $\tilde{\theta}$ in (7.9) is just the standard least squares linear regression estimator. The main point here is that the function $U(\lambda)$ gives nearly all its weight to a very small region near $\lambda = 0$; therefore, it is important to estimate the spectral density well in this region. In contrast, misspecification of $f(\lambda)$ far away from $\lambda = 0$ is likely to have hardly any influence on the estimated variance of $\tilde{\theta}$.

In an investigation of the possibilities for using (7.11) to assess the errors in a linear estimator of trend, Bloomfield and Nychka (1992) defined a "catalog of spectra": they considered three autoregressive models corresponding to $p = 1$, 2 and 8; the ARIMA(0,d,0) process and two theoretical spectral densities derived by Wigley and Raper (1990) for the dynamics of a simple mechanical model of the earth's climate. The latter had a larger concentration of power at low frequencies than the autoregressive models, but were still bounded spectral densities, whereas (7.13) is unbounded as $\lambda \to 0$.

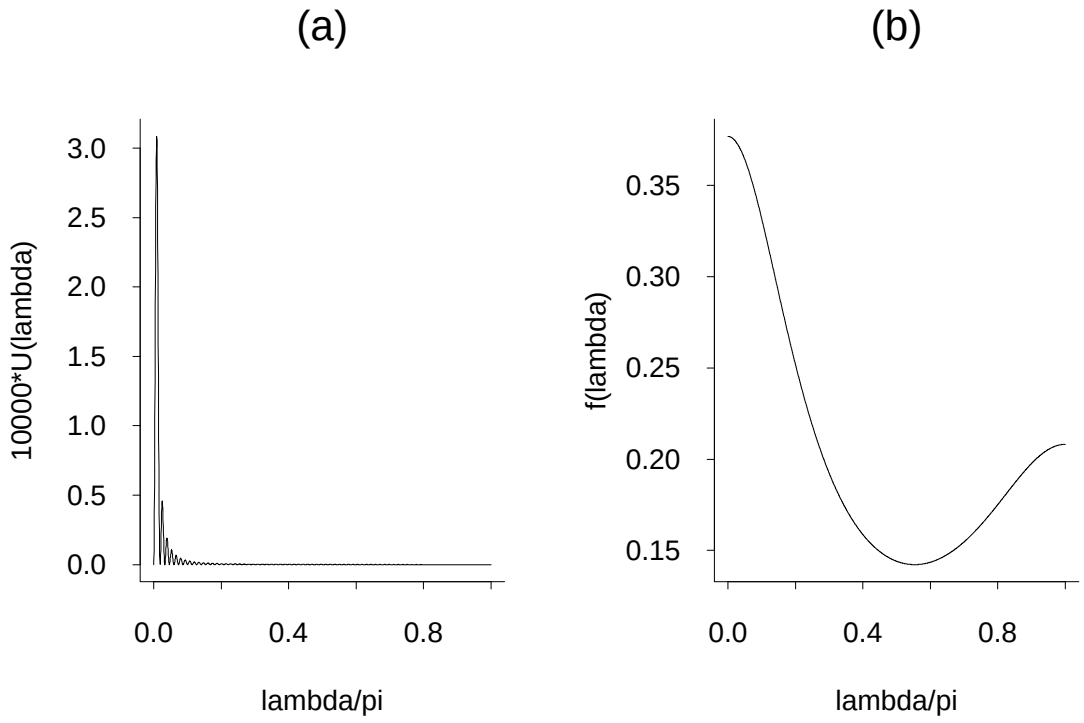


Fig. 7.1. Illustration of (7.11): (a) $U(\lambda)$, (b) $f(\lambda)$, using fitted AR(2) model, for 149 years of annual temperature averages in Amherst, MA.

Plots of the raw spectral densities showed considerable variation from one model to another at very low frequencies, but the estimated standard errors by (7.11) were fairly consistent over the range of realistic models. For example, when $\tilde{\theta}$ is the least squares estimate of linear trend over the IPCC series 1861–1989, Bloomfield and Nychka found a numerical value of $\tilde{\theta} = 0.367$°C, with standard errors according to different types of spectra as in Table 7.1. The results are fairly consistent across the AR(8), fractional noise, and the second Wigley-Raper model — the latter being a "multi-box" representation of the earth's atmosphere whereas the first Wigley-Raper model was "one box" and therefore less realistic. In contrast, assuming a low order of time series (white noise, AR(1) or AR(2)) gave much smaller standard errors, but such estimates are presumably not very realistic.

In conclusion, ignoring time series correlation or using unrealistically low-order models (AR(1) or AR(2)) is likely to lead to serious underestimation of the standard error of the trend, but among more realistic models, similar results are obtained by several different methods.

Another paper that took a spectral approach to the problem was by Kuo, Lindberg and Thomson (1990). They analyzed the Hansen-Lebedeff temperature series alongside a time series of atmopheric CO_2 measurements and in each case fitted a linear trend with time series errors, where the spectral density of the time series component was estimated nonparametrically using a multiple-window procedure. They claimed significant upward trends for both series — in particular, for the Hansen-Lebedeff series, their estimate of

the overall linear trend was .554 with a standard error of .096 (°C per century). This is consistent with Bloomfield's results reported earlier, for which the point estimate was .57 and the standard error about 0.1. They also examined the coherence — effectively, a measure of the correlation between two spectral densities — for the temperature and CO_2 series. For this, they claimed the somewhat surprising result that temperature leads CO_2 by about five months.

Spectrum	Standard Error
White noise	0.028
AR(1)	0.059
AR(2)	0.059
AR(8)	0.107
Fractional	0.097
Wigley-Raper 1	0.072
Wigley-Raper 2	0.101

Table 7.1. Standard errors of linear trend estimate from IPCC series by various spectral assumptions; from Bloomfield and Nychka (1992).

As a matter of general methodology, one might consider applying formula (7.11) with $f(\lambda)$ replaced by a nonparametric estimate of the spectral density. The difficulty is that, for a time series of length T when $\tilde{\theta}$ is the least squares estimate of a linear trend, the greatest power of U is obtained very near frequency $1/T$, which is a region not well estimated by traditional spectral density techniques. Multi-taper methods such as those considered by Kuo *et al.* are a method of improving on simple periodogram-based estimates of the spectral density, but it is still not clear that they resolve this fundamental difficulty with using nonparametric spectral estimates in this context.

7.1.3 Other "classical" approaches

Tol and de Vos (1993) attempted a direct regression of temperature on CO_2, using a regression equation of the form

$$y_t = \alpha + \beta x_{t-L} + \nu_t, \tag{7.14}$$

in which y_t was the IPCC series, x_t was mean CO_2 level and the lag L was taken to be 20 years. In their first analysis they took ν_t to be AR(1), though as Bloomfield had already shown, such an assumption is not adequate for this series. They then repeated the analysis using an ARMA(2,2) model for ν_t, both with and without the βx_{t-20} term, finding that the model including the CO_2 signal fitted much better than the model without, though the standard error was substantially larger under the ARMA(2,2) model than under the

AR(1) model (point estimate $\hat{\beta} = .015$ under both models; standard errors .002 under AR(1), .005 under ARMA(2,2)).

In further analysis, Tol and de Vos extended (7.14) to a model of form

$$y_t = \alpha + \beta y_{t-1} + \text{ forcing terms } + \text{ error}, \tag{7.15}$$

where the forcing terms, apart from CO_2, included sunspot numbers as a measure of solar variability, a "dust veil index" measuring atmospheric aerosols, a linear trend, and an ENSO (El Niño–Southern Oscillation) signal. Based on this they estimated $\Delta T_{2\times}$ to be 3.12°C with a standard error of 1.26. Further analysis of updated series by Tol (1994) reduced both the estimate of $\Delta T_{2\times}$ and its standard error: 2.8°C with a standard error of 0.8. Tol also applied the same model separately to the northern hemisphere and southern hemisphere series, though did not try to model them jointly (as we shall a little later).

The papers discussed so far have all broadly supported the hypothesis of a significant trend in global temperature. A contrary view was expressed by Woodward and Gray (1993, 1995). They considered models of the form (7.3) (with $x_t(\theta) = \theta t$) similarly to Bloomfield (1992), obtaining similar results, but as an alternative model, they also considered the ARIMA(p, d, q) model with integer d, as in standard time series texts such as Box, Jenkins and Reinsel (1994). In particular, they found a good fit to an ARIMA(9,1,0) model, with no overall positive or negative trend, for both the IPCC and Hansen-Lebedeff data sets. They pointed out that if such a model were correct, it would not be reasonable to forecast future temperatures as increasing even though observed temperatures may have been increasing for some time.

Woodward and Gray (1995) extended this analysis to include a formal test of the "linear trend plus autoregressive noise" model against an ARIMA alternative. They proposed a bootstrap test for this and found in simulations that, with parameter values for the two models similar to those fitted to the observed series, there was a reasonable level of discrimination between the two models. They then applied the bootstrap test several times to real data series, finding in every case a preference for the ARIMA model. From this, they concluded that there is no evidence of an anthropogenically induced climate signal.

These ideas are closely related to unit-root testing (Dickey and Fuller 1979, 1981, Dickey *et al.* 1986). An example of a unit root process would be

$$\Phi(B)(\nabla X_t - \mu) = \epsilon_t, \tag{7.16}$$

where $\nabla X_t = X_t - X_{t-1}$ and $\Phi(B)$ is some polynomial function of the lag operator B whose roots lie outside the unit circle (so that the model $\Phi(B)X_t = \epsilon_t$ would be stationary). In (7.16), the presence of the term ∇X_t makes it a nonstationary unit root model. As an extension of their methodology, Woodward and Gray (1995) discussed how to test the null hypothesis $\mu = 0$ in (7.16) against the alternative $\mu \neq 0$ (the previous analyses had assumed $\mu = 0$). For several real data series, they found that this hypothesis was not rejected, i.e. there is no evidence of a deterministic drift even within a unit root model.

The difficulty with the Woodward-Gray approach lies in the credibility of the ARIMA (or unit-root) model for climate series. Such models originated in economics, where there is generally no bar to a permanent long-term shift in a price or some other economic indicator. However, when applied to the climate, they imply that (a) there is no such thing as a "stationary distribution" for a climatic variable such as temperature, (b) over the course of a very long time period, climatic series will shift arbitrarily far from their starting values. Neither assumption fits very well with our physical notions of climate. Moreover, when the ARIMA(p, d, q) model is extended to include fractional d, estimates of d are typically between 0 and $\frac{1}{2}$ — this is necessary if the series is to be stationary, but most actual estimation methods do not restrict themselves *a priori* to this range. Thus, when the range of models is extended to include long-range dependence, we typically get results that are consistent with long-range dependence but still within the class of stationary series. Even this hypothesis is not without its problems — for example, the theoretical spectral density derived from dynamical considerations by Wigley and Raper (1990) is bounded, albeit putting high power on low frequencies — but the presence of natural cycles of very long amplitudes such as the Milankovich cycles suggests that in practice, there will be very low frequency variation which is indistinguishable from long-range dependence in practice, but still consistent with an overall stationary system.

7.2 Approaches based on long-range dependence

There are two reasons or trying to extend the analysis of time series with trends beyond autoregressive or ARMA processes. First, if we are even to consider the possibility that trends in climatological time series are caused by natural variability, then it is clear that this variability must be persistent over very long time scales — otherwise, we could not expect a warming pattern to persist for over a century, as seems to be the case in the current record. This suggests looking for models which, while consistent with stationary time series, have correlations that decay very slowly at large lags. A second reason for looking at models of this form is that the spectral formula for the variance of a trend estimate implies the need to estimate the spectral density well for frequencies very close to 0, and this suggests looking for estimation methods that focus particularly on that range of frequencies.

At the present time, there are three main approaches to long-range dependence. The first, already mentioned, is based on the fractional ARIMA model (7.7), first introduced independently by Granger and Joyeux (1980) and Hosking (1981), developed further by Hosking (1984), Haslett and Raftery (1989) and a number of other authors. These models are parametric and can be estimated by maximum likelihood techniques — Beran (1994) reviewed them in detail and, as already mentioned, Bloomfield (1992) used them in the context of climatological trend analysis. Their main disadvantage in this context is that since, from a spectral point of view, the fractional ARIMA model attempts to model the whole of the spectral density, if the model is misspecified across the whole range of frequencies, it may result in biased estimated in the specific region of interest, which corresponds to very low frequencies. Alternatively, trying to find a model which is correctly

specified across the whole frequency range may result in relatively large values of the ARMA orders p and q, and an unparsimonious model. This objection hardly applies to Bloomfield's model for the IPCC series, for which $p = 0$, $q = 1$, but it is a possible objection to the general use of the fractional ARIMA model.

The second method is based on spectral estimation of long-range dependence, in which estimation is confined to a narrow window of frequencies near 0. This method was applied to climatological time series by Smith (1993) and Smith and Chen (1996), and is the main focus of the discussion to follow.

The third method is more recent and is based on wavelets, cf. McCoy and Walden (1996), Craigmile *et al.* (2000). This method may be particularly advantageous when the residual time series may be nonstationary in addition to having long-range dependence. We do not pursue this theme in the present discussion.

7.2.1 The spectral approach

Now, we discuss some details of the spectral approach. To begin with some definitions: suppose $\{y_t,\ t = 0, \pm 1, \pm 2, ...\}$ is a stationary time series with autocovariances $\gamma_k = \mathrm{Cov}\{y_t, y_{t-k}\}$ for $k \geq 0$. In cases where a spectral density exists, it may be derived from the autocovariances through the formula

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \gamma_n e^{-in\lambda} \\ &= \frac{\gamma_0}{2\pi} + \frac{1}{2\pi} \sum_{n=1}^{\infty} \gamma_n \cos(n\lambda), \end{aligned}$$

which has inverse

$$\begin{aligned} \gamma_n &= \int_{-\pi}^{\pi} e^{in\lambda} f(\lambda) d\lambda \\ &= \frac{1}{2} \int_{0}^{\pi} \cos(n\lambda) f(\lambda) d\lambda. \end{aligned}$$

The *periodogram* may be defined by

$$I_T(\lambda) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} y_t e^{it\lambda} \right|^2$$

and is an approximately unbiased estimator of $f(\lambda)$ for each λ. In practice, the periodogram is usually calculated only at the *Fourier frequencies*, $\lambda_j = 2\pi j/T$ for $j = 0, 1, 2, ..., T/2$. In addition, in many applications two modifications are often made to the calculation of the raw periodogram. *Tapering* means multiplying a finite data sequence $\{y_t,\ 1 \leq t \leq T\}$ by a tapering function $h(t)$, such that $h(t) \to 0$ as $t \to 1$ or T. This has the effect of avoiding problems such as *leakage* caused by a sharp cutoff at the ends of the

data, see e.g. Bloomfield (1976) or any modern book on spectral methods. The second very common modification of the periodogram is *smoothing* which involves averaging a periodogram over neighboring Fourier frequencies to obtain a smoother function. In the present discussion we shall mostly work with the raw periodogram, without applying either tapering or smoothing, largely because the periodogram is being viewed primarily as an intermediate step in the estimation of long-range dependence.

Long-range dependence may be defined by either of

$$\begin{aligned} \gamma_k &\sim ak^{2d-1}, \ k \to \infty, \\ f(\lambda) &\sim b\lambda^{-2d}, \ \lambda \to 0. \end{aligned} \tag{7.17}$$

In principle more general definitions are possible, e.g. by including additional logarithmic or other so-called slowly-varying functions in either part of (7.17), but in practice this definition usually suffices. The relation between a and b is

$$b = \frac{a}{\pi}\Gamma(2d)\cos(\pi d), \tag{7.18}$$

which follows from a Fourier series identity (Zygmund 1959, Chapter V, (2.22)).

Suppose (7.17) holds and consider the sample mean based on T observations, $\bar{Y}_T = T^{-1}\sum_{t=1}^{T} y_t$. An asymptotic argument shows that

$$\begin{aligned} \mathrm{Var}(\bar{Y}_T) &\sim \frac{a}{d(2d+1)}T^{2d-1} \\ &= \frac{\pi b}{d(2d+1)\Gamma(2d)\cos(\pi d)}T^{2d-1}. \end{aligned} \tag{7.19}$$

For the regression coefficient of a normalized linear trend,

$$\tilde{\theta}_T = \frac{\sum y_t(t - \frac{T+1}{2})}{\sum(t - \frac{T+1}{2})^2}, \tag{7.20}$$

a similar argument leads to

$$\begin{aligned} \mathrm{Var}(\tilde{\theta}) &\sim \frac{36a(1-2d)}{d(1+2d)(3+2d)}T^{2d-3}, \\ &= \frac{36\pi b(1-2d)}{d(1+2d)(3+2d)\Gamma(2d)\cos\pi d}T^{2d-3}. \end{aligned} \tag{7.21}$$

The formulae led Smith (1993) to propose the following procedure. Suppose we estimate a linear trend of the form $\alpha + \theta\{t - (T+1)/2\}$ using OLS estimators $\tilde{\alpha}_T = \bar{y}_T$ and $\tilde{\theta}_T$ from (7.20), then form residuals

$$e_t = y_t - \tilde{\alpha}_T - \tilde{\theta}_T\left\{t - \frac{T+1}{2}\right\}.$$

Suppose now we estimate the long-range dependence parameters b and d from the residuals e_t. Then the variance of $\tilde{\alpha}_T$ and $\tilde{\theta}_T$ may be estimated from (7.19) and (7.21). In particular, (7.21) may be used to estimate a standard error for $\tilde{\theta}_T$ and hence to construct confidence intervals and hypothesis tests for the linear trend.

7.2.2 Estimation of b and d

For the estimation of the long-range dependence parameters b and d, there are by now a number of spectral-based approaches. Since $I_T(\lambda)$ is an approximately unbiased estimator of $f(\lambda) \sim b\lambda^{-2d}$ for λ near 0, a natural approach is to perform an ordinary least squares regression of $\log I_T(\lambda)$ on $\log \lambda$ for a sequence of Fourier frequencies $\lambda_j,\ j = 1, 2, ..., n_c$ where n_c is some cutoff much smaller than $T/2$; the reason for the last restriction is to ensure that the estimation really is confined to a small range of frequencies near 0. This procedure is very nearly the same as that proposed by Geweke and Porter-Hudak (1983) and was placed on a rigorous mathematical footing, with some modifications, by Robinson (1995a).

A second procedure has been variously called Gaussian, maximum likelihood or Whittle estimation (because of its similarity to the Whittle method of estimating parametric time series models from the spectral density) and consists of choosing b and d to minimize

$$\sum_{j=1}^{n_c} \left\{ \log f(\lambda_j; b, d) + \frac{I_T(\lambda_j)}{f(\lambda_j; b, d)} \right\} \tag{7.22}$$

where $f(\lambda; b, d)$ is approximated by $b\lambda^{-2d}$ and we again restrict attention to the first n_c Fourier frequencies. This method was advocated by Künsch (1987) and Smith (1993), and given rigorous mathematical justification by Robinson (1995b). Asymptotic theory shows that for a fixed n_c, the variance of the estimator of b is smaller (by a factor $\pi^2/6$) compared with the Geweke–Porter-Hudak estimator. From now on, we use only this estimator.

7.2.3 Joint estimation of trend and long-range dependence parameters

The approach discussed so far assumes that the trend and long-range dependence parameters are estimated separately, with the trend first estimated by ordinary least squares and substracted from the data to form residuals, which are then analyzed as a stationary time series to estimate the parameters b and d. There are some possible objections to that: for example, OLS estimation of the trend is theoretically inferior to generalized least squares (GLS) estimation, with an asymptotic efficiency (for the case of a linear trend when $0 < d < \frac{1}{2}$) between 0.889 and 1 (quoted by Smith (1993), based on theoretical results due to Yajima (1988, 1991)). There could also be some bias in estimating b and d from a residual series in which the initial estimation and removal of a linear trend is ignored, though this source of bias is usually ignored in theoretical studies. The considerations led Smith and Chen (1996) to propose an alternative method in which the trend and long-range dependence parameters were estimated simultaneously.

Consider the model

$$y_t = \sum_{k=1}^{p} \beta_k x_{k,t} + \nu_t, \quad 1 \le t \le T, \tag{7.23}$$

in which $\{y_t\}$ is an observed time series, $\{x_{k,t},\ 1 \le k \le p\}$ is a vector of p covariates whose values at time t are known, $\beta_1, ..., \beta_p$ are unknown regression coefficients, and $\{\nu_t\}$ is a stationary time series. The linear trend model

$$y_t = \beta_1 + \beta_2 t + \nu_t, \quad 1 \le t \le T, \tag{7.24}$$

is a special case of (7.23).

Define the discrete Fourier transform (DFT) of y_t,

$$D_{y,T}(\lambda) = C_{y,T}(\lambda) + iS_{y,T}(\lambda), \tag{7.25}$$

where $C_{y,T}$ and $S_{y,T}$ are the discrete cosine and sine transforms defined by

$$\begin{aligned} C_{y,T}(\lambda) &= \sqrt{\frac{1}{2\pi T}} \sum_{t=1}^{T} y_t \cos \lambda t, \\ S_{y,T}(\lambda) &= \sqrt{\frac{1}{2\pi T}} \sum_{t=1}^{T} y_t \sin \lambda t. \end{aligned} \tag{7.26}$$

Note that $I_{y,T}(\lambda) = |D_{y,T}(\lambda)|^2 = C_{y,T}(\lambda)^2 + S_{y,T}(\lambda)^2$. By taking the DFT of each side in (7.23), we have

$$D_{y,T}(\lambda) = \sum_{k=1}^{p} \beta_k D_{x_k,T}(\lambda) + D_{u,T}(\lambda). \tag{7.27}$$

The model is therefore of the following structure: if we write the cosine and sine transforms of y_t for the first n_c Fourier frequencies as a $(2n_c)$-dimensional vector, and similarly for each of the covariates $x_{k,t}$, (7.27) allows us to write this $(2n_c)$-dimensional response vector as a linear function of p known covariates (i.e. the DFTs of the original covariates $x_{k,t}$) plus a random error term $D_{u,T}$ of mean 0 and a covariance function which may be assumed to be a known function of d and λ. This creates the possibility of either maximium likelihood or REML estimation of the model based on (7.27).

In approximating the covariance matrix in (7.27), note that if $\tilde{y}_k = \sum_s b_{ks} y_s$ and $\tilde{y}_\ell = \sum_t b_{\ell t} y_t$ are two of the components of $D_{y,T}$, then we may write

$$\begin{aligned} \text{Cov}\,(\tilde{y}_k, \tilde{y}_\ell) &= \sum_s \sum_t b_{ks} b_{\ell t} \int_{-\pi}^{\pi} e^{i(s-t)\lambda} f(\lambda) d\lambda \\ &= \int_{-\pi}^{\pi} B_j(\lambda) B_k^*(\lambda) f(\lambda) d\lambda \end{aligned} \tag{7.28}$$

where $B_j(\lambda) = \sum_s b_{js} e^{i\lambda s}$ and the asterisk denotes complex conjugate. In calculating the likelihood function, (7.28) has been evaluated by direct numerical integration based on 1,000 sampling points. Unfortunately, in this setting it is not valid to apply approximations which effectively assume that the components of $D_{u,T}$ are independent; for further discussion of this point we refer to Smith and Chen (1996).

7.2.4 Application: Central England series

Fig. 7.2(a) gives 342 annual means of the Central England data set, 1659–2000. This series, originally compiled from three observing stations in the center of England, is one of the longest continuously collected direct records of temperature; it was originally collected by Manley (1974) and is today available from the Hadley Center website (http://www.metoffice.gov.uk/research/hadleycentre/CR_data/Monthly/HadCET_act.txt). For the present analysis, we do not (initially) subtract any trend but estimate b and d from the full series — then we test for the significance of a linear trend in several recent sections of the series. The rationale for this approach is that the null hypothesis is that there is no trend but simply long-range dependence; therefore we assess the spectral density from the undetrended series and use that to assess the significance of observed trends. The following analysis differs in a number of features from the analysis of the same series in Smith (1993).

Fig. 7.2(b,c) gives the autocorrelation (ACF) and partial autocorrelation function (PCF) of the full series. Clearly, the ACFs decay slowly but the PCF decay more rapidly except for a (possibly spurious) significant coefficient at lag 15, so on the basis of this, it may be acceptable to approximate the process by a low-order autoregressive model, but we shall continue to pursue this as a long-range dependent series to examine the results which ensue from such an assumption.

Fig. 7.3(a) shows an example of a (smoothed and tapered) spectral density estimate, which confirms that most of the power is in low frequencies. Fig. 7.3(b,c) shows the raw periodogram both on the scale which is usually plotted (log periodogram against frequency) and an alternative log-log scaling which is shown because this is the scale on which we would expect a linear relationship according to the long-range dependence model.

Fig. 7.4 shows Fig. 7.3(c) again, with the estimated straight line $(\log b - 2d \log \lambda)$ superimposed for a number of values of n_c. Although a few of the lines for very small n_c are clearly different from the rest, overall these fits are very similar, confirming the good fit of the long-range dependence model. Fig. 7.5 shows estimates of d and associated 95% confidence intervals for a variety of values of n_c. After initial variability up to $n_c = 20$, the rest of the plot is very stable and shows clear evidence of long-range dependence. For example, for $n_c = 40$, the estimate $\hat{d}$ is .44 with a standard error of .10. Although not shown on the plot, estimates of b and their standard errors are of course also available.

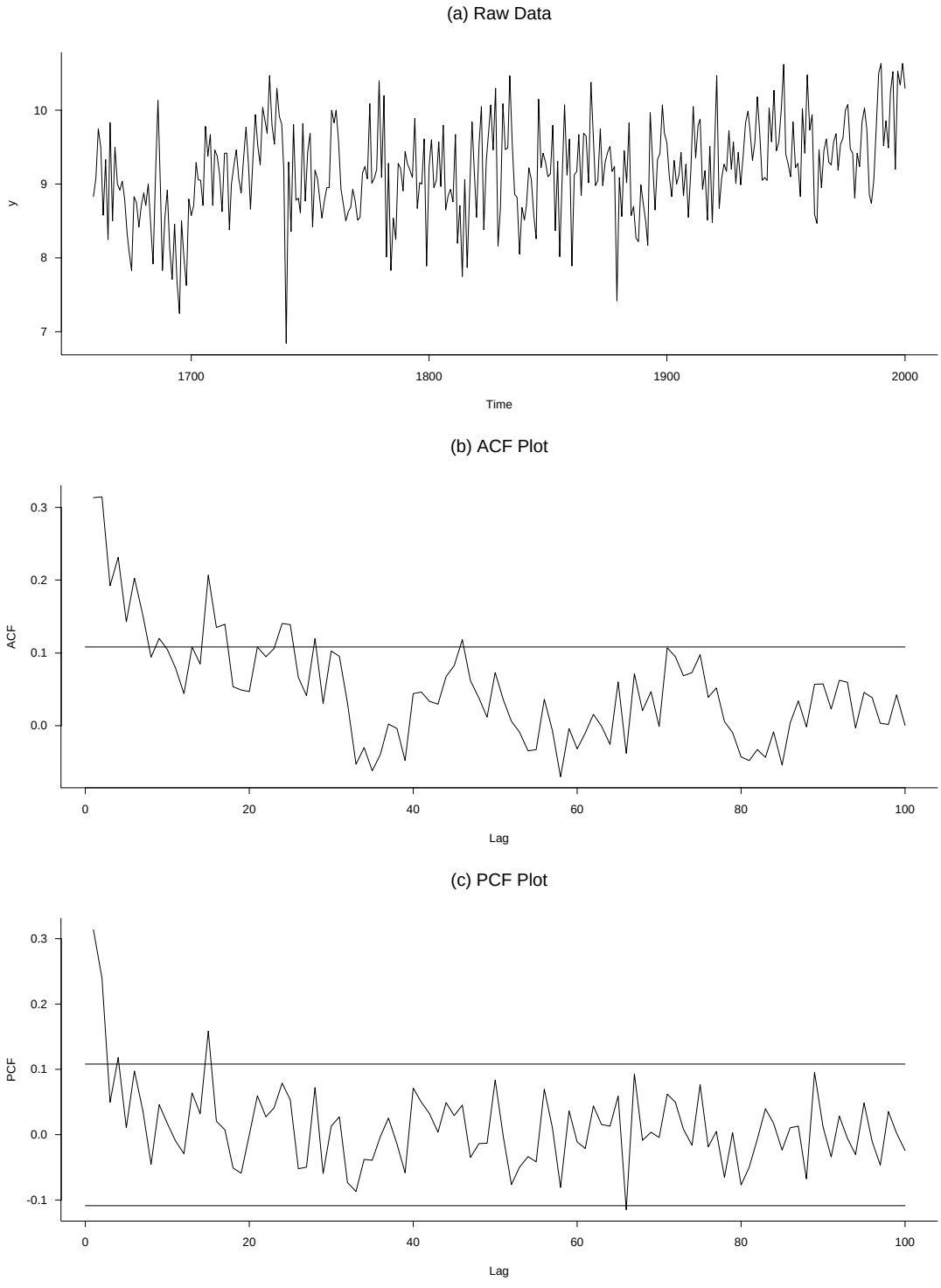


Fig. 7.2. Central England data set: Raw data, AFC and PACF plots.

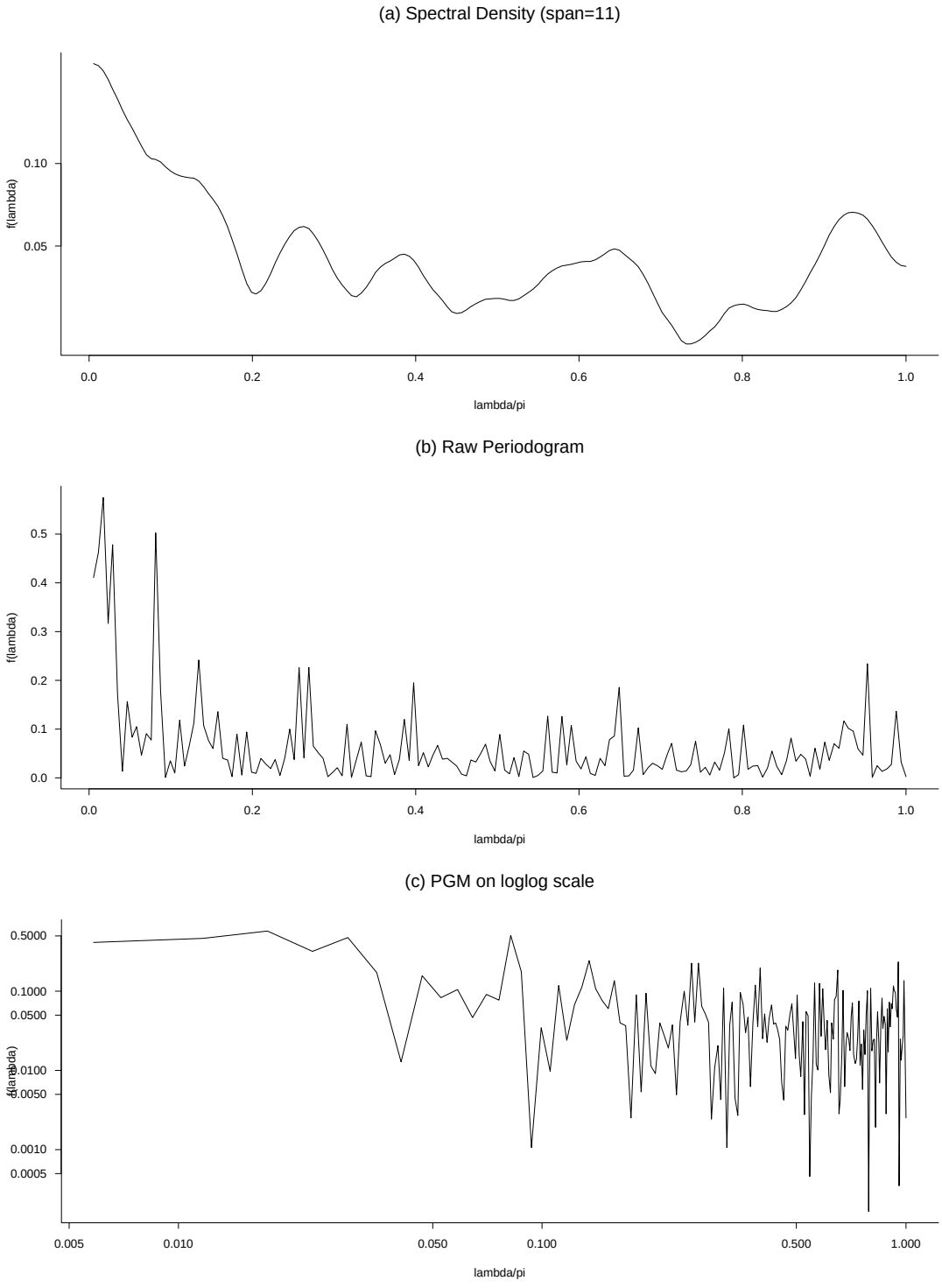


Fig. 7.3. Central England data set: Smoothed spectral density and raw periodogram.

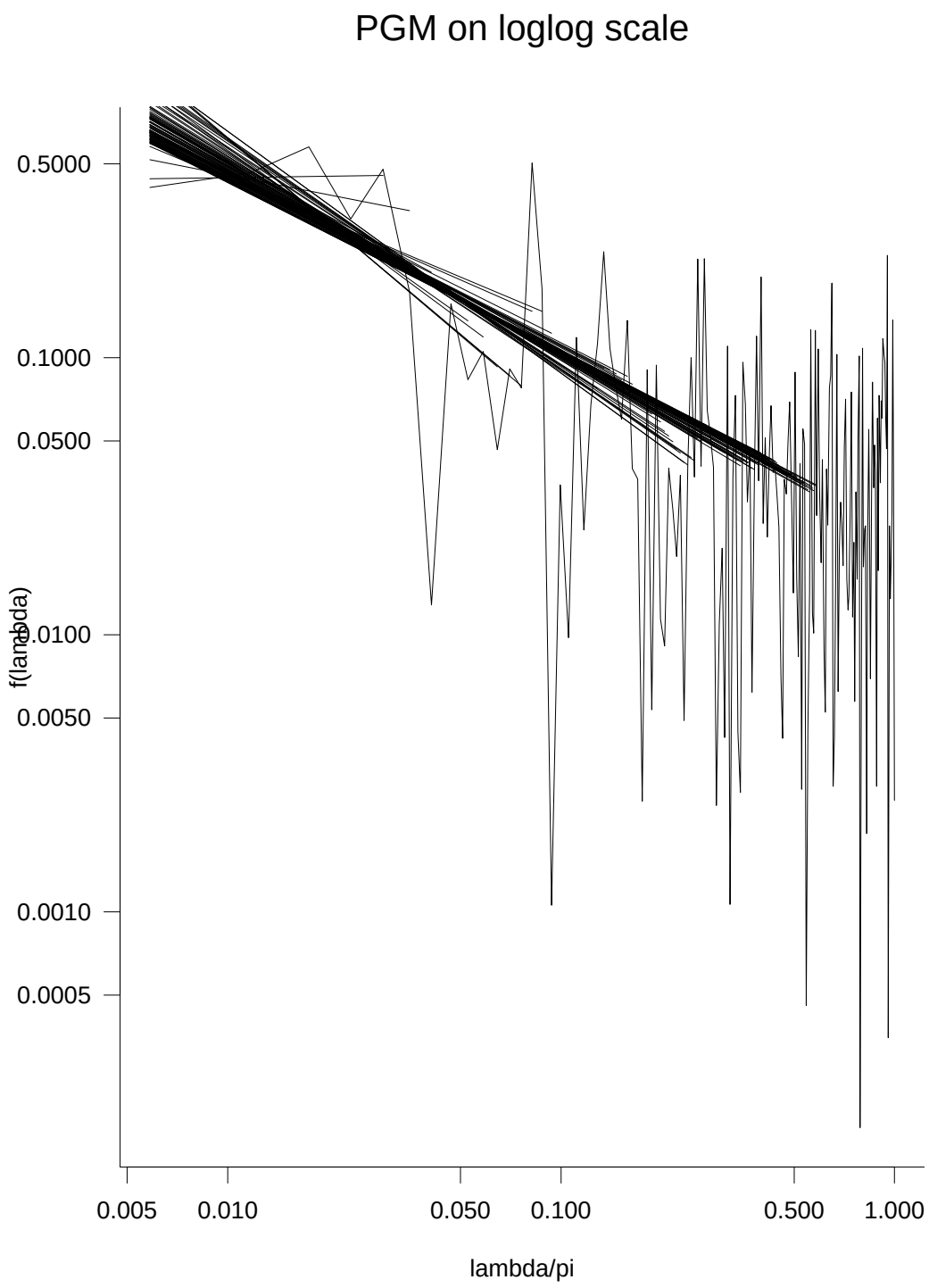


Fig. 7.4. Central England data set: Periodogram on log-log scale and fitted straight lines for various values of n_c.

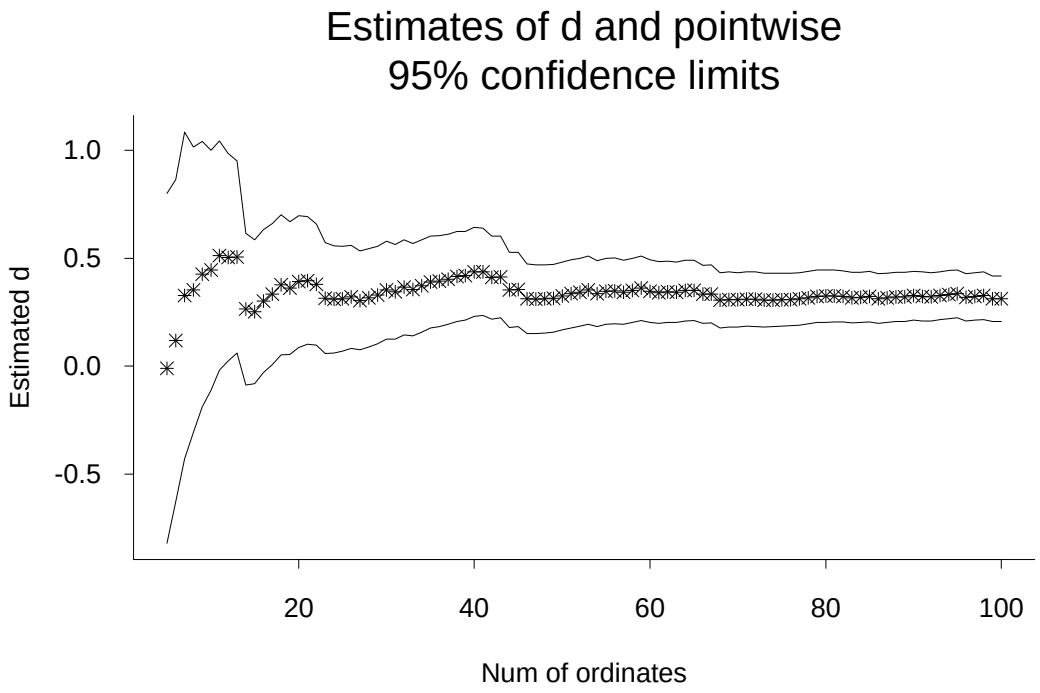


Fig. 7.5. Central England data set: $\hat{d}$ and 95% confidence band by Whittle method for various n_c.

m	Trend	SE	Ratio
5	.23	.17	1.4
10	.098	.07	1.4
15	.077	.046	1.7
20	.054	.034	1.6
25	.042	.027	1.6
30	.029	.022	1.3
35	.025	.019	1.3
40	.026	.016	1.6
45	.018	.014	1.3
50	.016	.013	1.2

Table 7.2. Linear trend estimates, standard errors and t ratios for Central England series over the past m years, for various valus of m.

Table 7.2 shows estimates of the linear trend and associated standard errors (calculated from (7.21), using the estimates of b and d calculated for $n_c = 40$) over the last m years,

for various values of m from 5 to 50. The results show that although there have been some dramatically large trends in recent years, these are not statistically significant. In fact none of the trends shown in the table is significant, though the most nearly so are for m in the range 20–40 years.

Although this is a negative result, it should be borne in mind that this is only a single series and most of the series for which significant trends have been claimed are based on data aggregated from many stations — in particular, this is true of the global and hemispheric temperature series. From this point of view, the fact that several of the trends are "nearly significant" on their own (i.e. without combining with information from other stations) is worthy of note. On the other hand, the standard errors also show that apparent very large trends in some recent sections of the series (e.g. $m = 5, 10$) are probably not significant when compared with the correspondingly large standard errors.

7.2.5 Application: global average temperature series

Figs. 7.6–7.9 show a similar analysis for the current (1856–2000) series of global average temperature available from the University of East Anglia website (http://www.cru.uea.ac.uk/cru/data/temperature; series tavegl.dat). In this case, since there appears to be little dispute over the existence of a trend, we subtract a linear trend (estimated by OLS) at the beginning of the analysis and calculate the time series features from the detrended series. Fig. 7.6(a) shows the raw series, (b) and (c) the ACF and PCF for the residuals, from which we see that there is again a very slow decay in the ACFs but much more rapid decay in the PCFs, suggesting that a low-order AR(p) would again be an adequate fit to the series (recall that Bloomfield (1992) took $p = 4$ in this context). Plots of the smoothed and tapered spectral density, and of the raw periodogram, again show evidence of power concentrated on low frequencies (Fig. 7.7), and the cloud of straight-line fits on log-log scale suggest a good fit except at very low values of n_c (Fig. 7.8). Plots of $\hat{d}$ and associated 95% confidence intervals show clear evidence of long-range dependence (Fig. 7.9). For example, at $n_c = 30$, we have $\hat{d} = .42$ with a standard error .10. Some of the estimates of d are even bigger than $\frac{1}{2}$ — for example, at $n_c = 20$, we have $\hat{d} = .51$ with standard error .12 — suggesting that it is even possible that we are in the region of nonstationarity (see e.g. Hurvich and Ray (1995) for discussion of estimating d in this case). A table of trend estimates and associated standard errors, for various time lengths m including $m - 145$, the full series, is shown in Table 7.3.

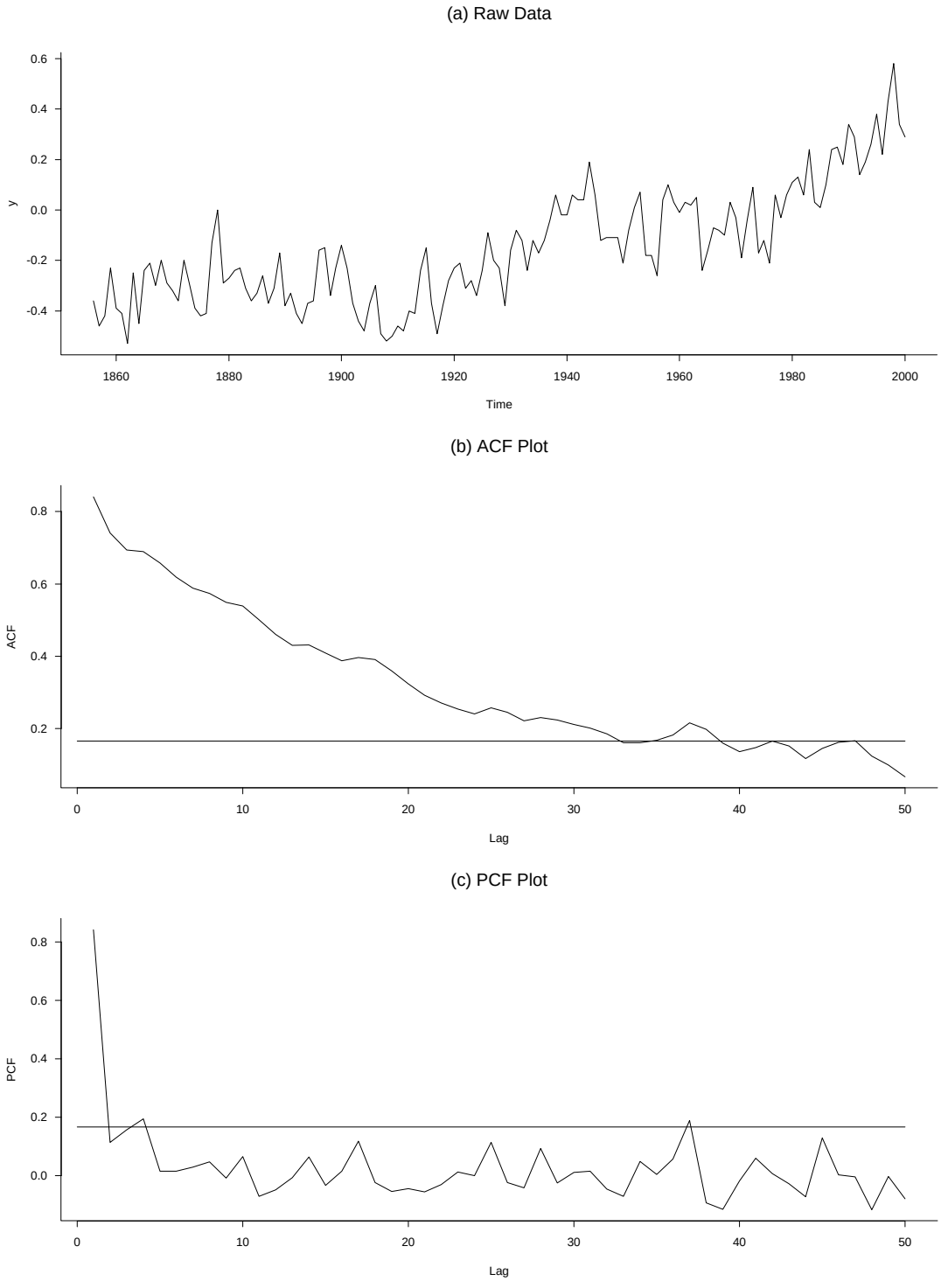


Fig. 7.6. IPCC data set: Raw data, AFC and PACF plots.

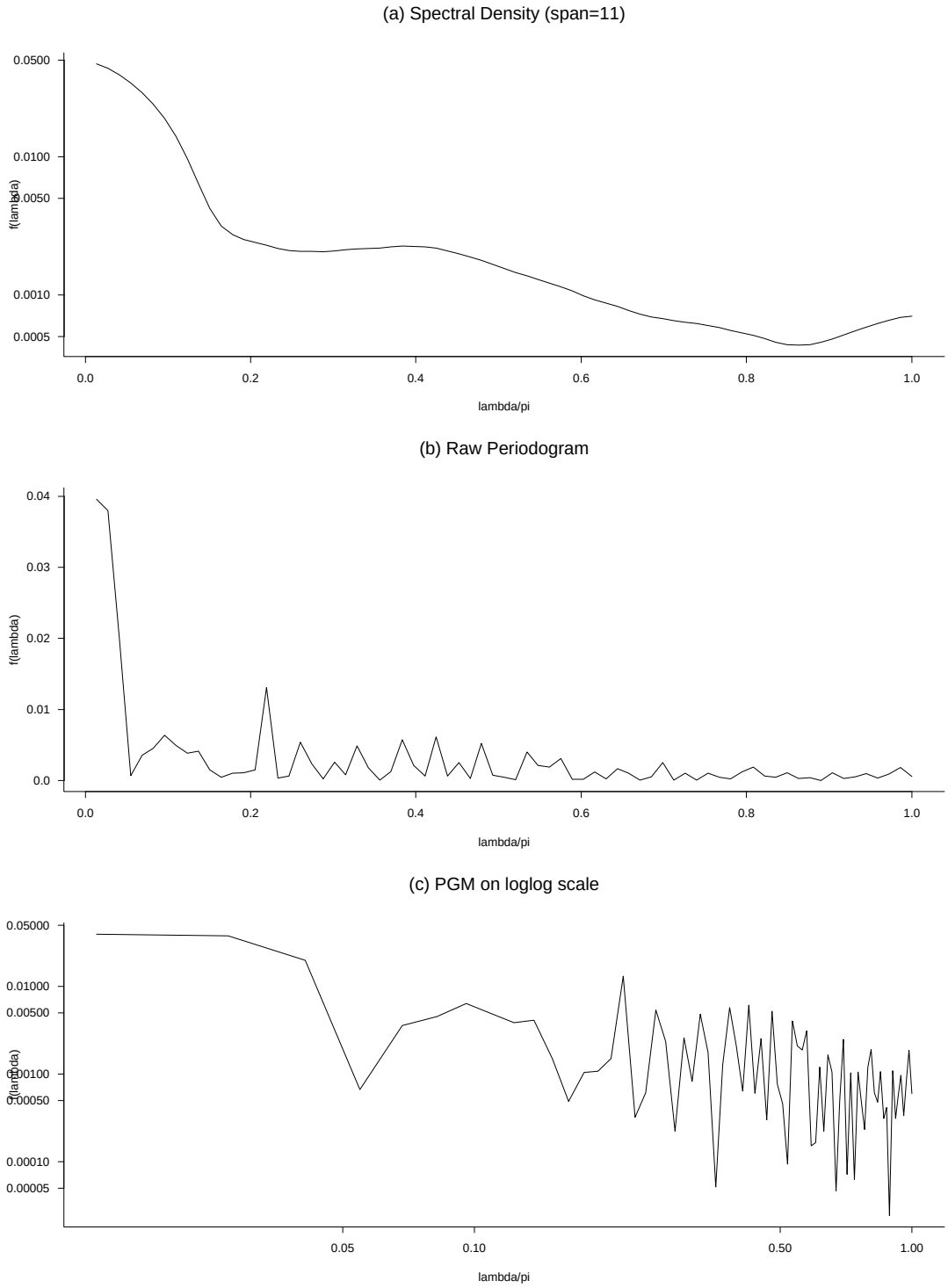


Fig. 7.7. IPCC data set: Smoothed spectral density and raw periodogram.

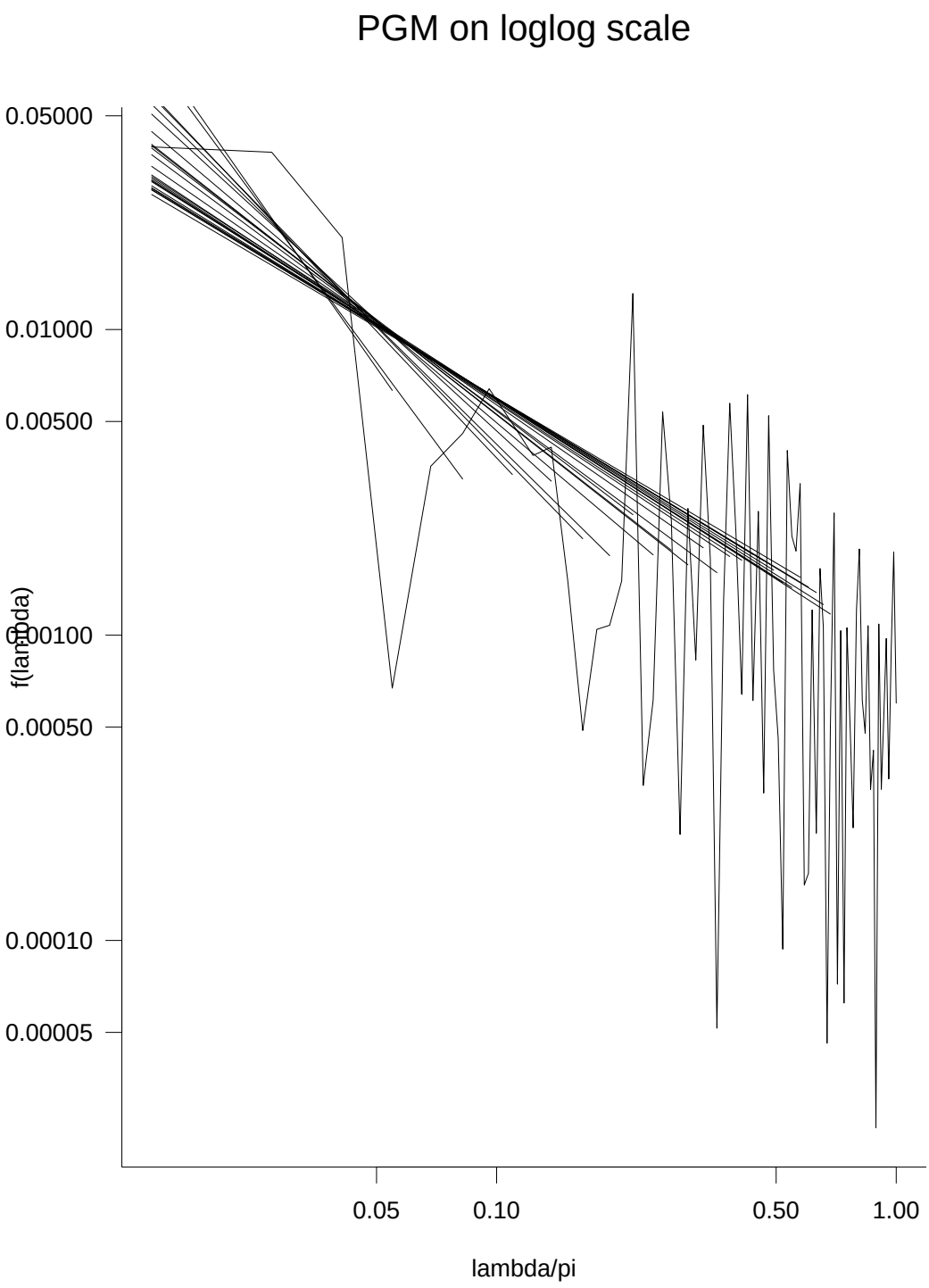


Fig. 7.8. IPCC data set: Periodogram on log-log scale and fitted straight lines for various values of n_c.

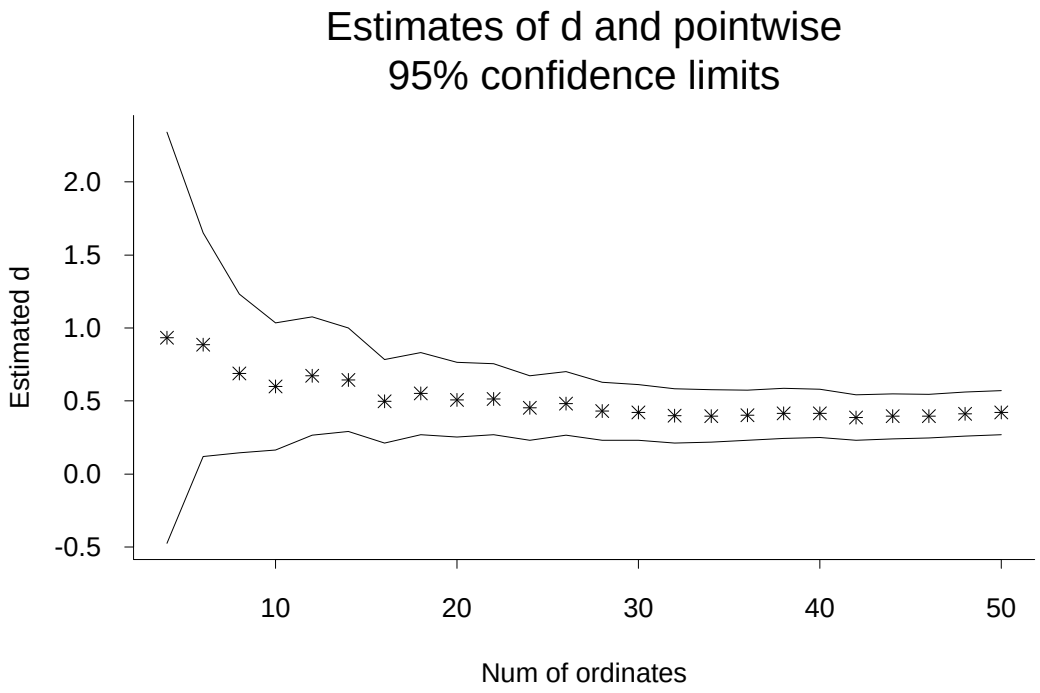


Fig. 7.9. IPCC data set: $\hat{d}$ and 95% confidence band by Whittle method for various n_c.

m	Trend	SE	Ratio
10	.022	.017	1.3
20	.017	.0087	2.0
30	.018	.0056	3.2
40	.013	.0044	3.0
50	.009	.0035	2.8
145	.0043	.0012	3.6

Table 7.3. Linear trend estimates, standard errors and t ratios for global temperature series over the past m years, for various valus of m.

Finally, the procedure of Smith and Chen (1996) has been applied to the full (145-year) series, using $n_c = 20$ for the number of spectral ordinates. Using the maximum likelihood fits to model (7.27), we find $\hat{d} = .46$ (standard error .12) and $\hat{\theta} = .0046$ (.0010). The corresponding REML estimates are $\hat{d} = .50$ (.12) and $\hat{\theta} = .0047$ (.0011). The fact that these estimates based on the joint estimation of d and θ are little different from the earlier estimates, based on OLS trend estimation and time series analysis of the residuals, is some confirmation of the robustness of the methodology, in the sense that different ways of estimating the model seem to lead to similar conclusions. However we look at it, the

conclusion seems to be that there is both clear evidence of a trend and clear evidence of long-range dependence in this series.

7.3 Bivariate time series

This section reviews a recent paper by Smith, Wigley and Santer (2001), in which trend effects associated with a variety of climate models are fitted to bivariate temperature series correponding to northern hemisphere (NH) and southern hemisphere (SH) averages. For preliminary discussion, we refer back to section 1.2; in particular, Fig. 1.1 plots the data.

An earlier analysis of hemispheric data by Kaufmann and Stern (1997) was based on models of the form

$$\begin{aligned} N_t &= \alpha_1 + \sum_{j=1}^{k} \beta_{1j} x_{tj} + \sum_{j=1}^{p_1} \gamma_{1j} N_{t-j} + \sum_{j=1}^{q_1} \delta_{1j} S_{t-j} + \epsilon_{1t}, \\ S_t &= \alpha_2 + \sum_{j=1}^{k} \beta_{2j} y_{tj} + \sum_{j=1}^{p_2} \gamma_{2j} N_{t-j} + \sum_{j=1}^{q_2} \delta_{2j} S_{t-j} + \epsilon_{2t} \end{aligned} \tag{7.29}$$

where N_t and S_t are observed NH and SH temperature averages in year t, and $\{x_{tj},\ y_{tj},\ 1 \leq j \leq k\}$ are covariates which may either represent simple linear time trends or more complex covariates such as greenhouse gases. The coefficients $\gamma_{ij},\ \delta_{ij}$ on various lagged NH and SH terms represent serial correlation both within and between the two hemispheres, while ϵ_{1t} and ϵ_{2t} are error terms which are assumed normally distributed with mean 0, independent from one value of t to another and with variances σ_1^2, σ_2^2 say. We may allow $\mathrm{Corr}(\epsilon_{1t}, \epsilon_{2t}) = \rho$, where ρ may be any number between –1 and 1, though Kaufmann and Stern implicitly assumed $\rho = 0$.

An alternative form of the model is

$$\begin{aligned} N_t &= \alpha_1 + \sum_{j=1}^{k} \beta_{1j} x_{tj} + W_t, \\ S_t &= \alpha_2 + \sum_{j=1}^{k} \beta_{2j} y_{tj} + Z_t, \\ W_t &= \sum_{j=1}^{p_1} \gamma_{1j} W_{t-j} + \sum_{j=1}^{q_1} \delta_{1j} Z_{t-j} + \epsilon_{1t}, \\ Z_t &= \sum_{j=1}^{p_2} \gamma_{2j} W_{t-j} + \sum_{j=1}^{q_2} \delta_{2j} Z_{t-j} + \epsilon_{2t}, \end{aligned} \tag{7.30}$$

which differs from (7.29) by first forming detrended series, W_t and Z_t, by subtracting the deterministic trend terms from N_t and S_t respectively, and then modeling (W_t, Z_t) as a stationary bivariate autoregressive series.

It is possible to fit either form of model (7.29) or (7.30) to data including alternative lags and trend terms, and to compare the results by direct comparisons of fitted likelihoods or through automatic model selection criteria such as AIC or BIC.

Initial analysis of the data by Smith *et al.* (2001) led to the following conclusions:

1. In models which included a linear trend term ($x_{t1} = y_{t1} = t$) and a component representing the El Niño–Southern Oscillation effect (the so-called Southern Oscillation Index or SOI; http://www.cru.uea.ac.uk/cru/data/soi.htm) both the latter terms were found to be significant. The SOI term was lagged six months, i.e. the value taken as a covariate for year t was based on the average of the last six months for year $t-1$ and the first six months for year t.

2. Model (7.30) was a better fit than model (7.29).

3. In comparisons of various orders for the autoregressive terms in (7.30), it was found that an adequate representation was obtained with $p_1 = q_2 = 1$, $p_2 = q_1 = 0$, i.e. simple AR(1) models with no cross-correlation terms. This contrasted with the results of Kaufmann and Stern (1997) who found, for a model based on (7.29) and with no SOI component, that an appropriate model required fourth-order autoregressive terms including a south-to-north dependence (i.e. $p_1 = q_1 = q_2 = 4$; only p_2 was 0).

The conclusion of simple AR(1) models when SOI is included as a covariate might raise speculation that the SOI term entirely explains the long-range dependence, this rendering the detailed discussion of Section 7.2 unnecessary. The paper by Smith *et al.* (2001) did not examine this aspect of the problem, but for direct comparison with the results of Section 7.2, the residuals from the model just discussed have been computed, together with their periodograms, separately for the NH and SH series, and fitted to a long-range dependence model using the same methods as Section 7.2. Fig. 7.10 shows resulting Whittle estimates of the parameter d with 95% confidence limits, analogous to Figs. 7.5 and 7.9 for the Central England and IPCC data. As can be seen, there is similar evidence that d lies between 0.3 and 0.5, suggesting a need to consider long-range dependence in this series as well. At the time of writing, no such bivariate analysis involving long-range dependence has been carried out.

The actual analysis by Smith *et al.* (2001) henceforth assumed a bivariate AR(1) model with $p_1 = q_2 = 1$, $p_2 = q_1 = 0$, and including SOI as a covariate, but instead of a simple linear trend, substituted trends generated by a variety of climate models of a form due to Wigley and Raper (1992) and modified by Raper *et al.* (1996). A total of 24 model-generated trends was included, based on six different combinations of forcing factors and 4 values of the climate sensitivity $\Delta T_{2\times}$; recall that the latter parameter represents the nominal change in the earth's mean temperature corresponding to a doubling of CO_2 from pre-industrial times.

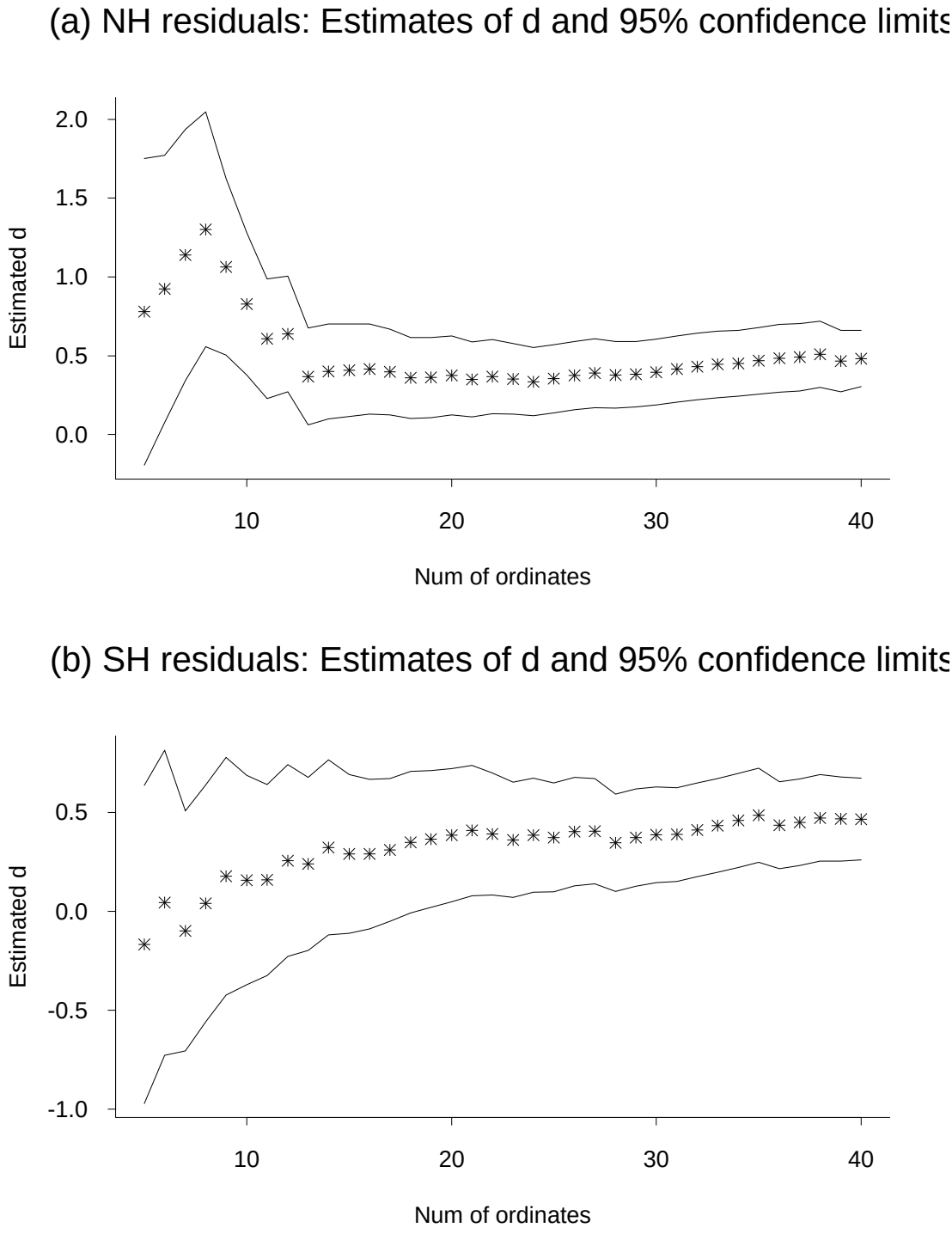


Fig. 7.10. $\hat{d}$ and 95% confidence band by Whittle method for various n_c, for (a) NH data, (b) SH data.

Forcing case case	GHGs and biomass aerosols	Sulfate aerosols	Solar	Optimum $\Delta T_{2\times}$
A: GHGs	yes	no	no	1.36
B: Anthropogenic	yes	high	no	11.93
C: Low SO_4	yes	low	no	3.19
D: Anth. + Sun	yes	high	yes	4.16
E: Low SO_4 + Sun	yes	low	yes	2.32
F: Solar alone	no	no	yes	15.44

Table 7.4. Summary of models used for trend comparison. The "low" and "high" values of sulfate forcing correspond to two different scenarios described by Kattenberg *et al.* (1996). The optimal values given for $\Delta T_{2\times}$ use similar methodology (though a different time period) to Wigley *et al.* (1998).

The six climate models are listed in Table 7.4. The four values of $\Delta T_{2\times}$ were taken as (1) 1.5, (2) 3.0 and (3) 4.5 °C to correspond to the currently accepted range of reasonable values of this parameter, and (4) an "optimum" value calculated separately for each model, on the basis of previous fits of the model output to observational data (so $\Delta T_{2\times}$ should itself be regarded as an unknown parameter, but since the estimation of that is not taken as part of the current regression procedure, we are treating model runs corresponding to different $\Delta T_{2\times}$ as distinct models to be compared by our likelihood-based procedures).

As a means of making comparisons among a large class of models, we use Bayes factors in the manner described by Kass and Raftery (1995). Suppose we have M different models to choose from, and the m'th model ($1 \le m \le M$) defines a probability density function $f_m(y;\theta_m)$ for observed data y in terms of model parameter θ_m. If the prior probability that model m is correct is $\Pi(m)$, and conditional on model m being correct, the prior density of θ_m is $\pi_m(\theta_m)$, then the posterior probability that model m is correct, given the data y, is

$$\Pi(m \mid y) = \frac{\Pi(m) \int f_m(y;\theta_m)\pi_m(\theta_m)d\theta_m}{\sum_{m'} \Pi(m') \int f_{m'}(y;\theta_{m'})\pi_{m'}(\theta_{m'})d\theta_{m'}}. \tag{7.31}$$

An alternative formulation is to rewrite (7.31) in the form

$$\frac{\Pi(m \mid y)}{\Pi(m' \mid y)} = \frac{\Pi(m)}{\Pi(m')} \cdot \frac{\int f_m(y;\theta_m)\pi_m(\theta_m)d\theta_m}{\int f_{m'}(y;\theta_{m'})\pi_{m'}(\theta_{m'})d\theta_{m'}} \tag{7.32}$$

for the comparison of two given models m and m'. The advantage of this approach is that (7.32) separates out the influence of the prior probabilities of the models (the Π factors in the right hand side) from the likelihood components (the ratio of integrals). If we ignore

the Π components in (7.32), the ratio of integrals, which we shall denote by $B(m;m')$, is called the *Bayes factor* of model m relative to model m' and represents the relative "weight of evidence" of the two models. This therefore represents a direct means of comparing two models without any prior assumption that one of them must be correct.

Bayes factors were first popularized in the classical treatise of Jeffreys (1961), who gave the interpretation in Table 7.5, as slightly modified by Kass and Raftery (1995).

Value of $B(m;m')$	Value of $\log_{10} B(m;m')$	Strength of evidence against model m'
1 to 3.2	0 to 0.5	Barely worth a mention
3.2 to 10	0.5 to 1	Substantial
10 to 100	1 to 2	Strong
Greater than 100	> 2	Decisive

Table 7.5. Jeffreys' table of interpretation of Bayes factors, adapted from Kass and Raftery (1995).

In practice, we have assumed the prior densities $\pi_m(\theta_m)$ to be constant, and have evaluated the integrals in (7.32) using Laplace's integral approximation (Kass and Raftery 1995),

$$\int f_m(y;\theta_m)\pi_m(\theta_m)d\theta_m \approx (2\pi)^{p_m/2}|V_m|^{1/2} f_m(y;\hat{\theta}_m) \tag{7.33}$$

where $\hat{\theta}_m$ is the MLE under model m, p_m is the dimension of the model, and V_m is the inverse observed information matrix under model m.

Raw comparisons based on negative log likelihood suggest that model D4 (i.e. row D of Table 7.4 and $\Delta T_{2\times} = 4.16°C$ corresponding to the optimum value for this model) is the best-fitting model to the data, and using this as a reference model, the Bayes factor for model D4 relative to each of the other models is plotted in Fig. 7.11. In this model, we are assuming the regression coefficients β_{11} and β_{21}, corresponding to the model trend terms in (7.30), are both 1, since this is the case when the model exactly corresponds to the data. For comparison among different time series models, Fig. 7.11 shows the Bayes factors both for the cases $p_1 = q_2 = 1$ (i.e. the joint AR(1) model we have described) and for an alternative model with $p_1 = q_2 = 4$. The Bayes factors for the AR(4) models are always below those for AR(1), indicating less strong discrimination among models, but the overall conclusions from the two sets of Bayes factors are comparable. In both cases, model D4 is the best with model D3 very close behind, and most of the other models "decisively worse" according to the Jeffreys criterion.

One particular comparison of interest is model F, which has solar forcing only. For climate sensitivities of 1.5, 3.0 and 4.5, this model is clearly much worse than model D.

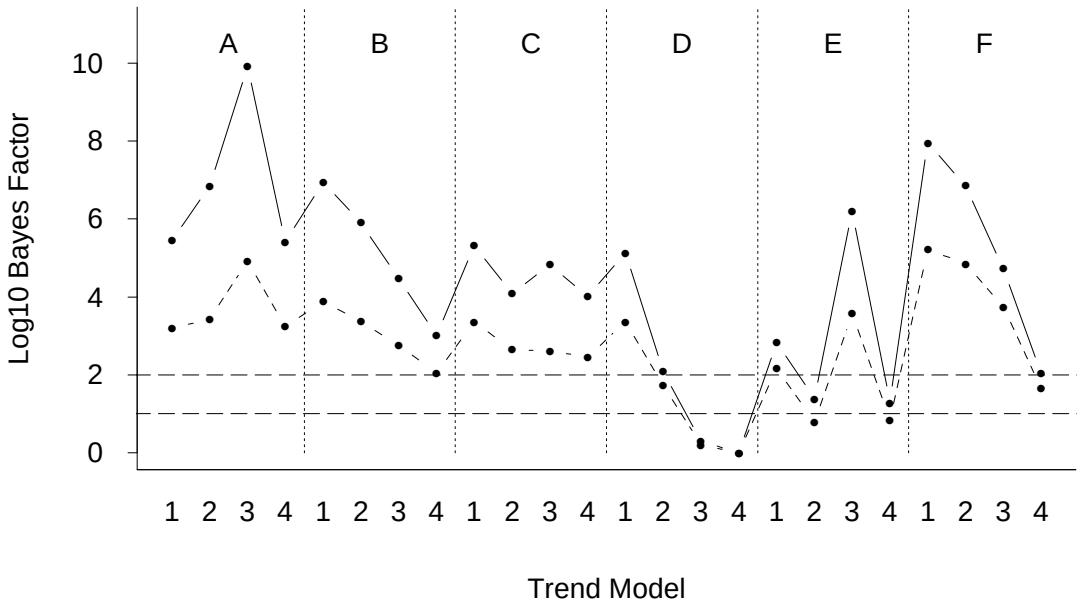


Fig. 7.11. Approximate $\log_{10}$ Bayes factors computed for 24 combinations of climate model and climate sensitivity (same as in Fig. 2), all computed relative to model D4 (for which the $\log_{10}$ Bayes factor is fixed at 0). All fits are for model (2) including SOI, with $p_2 = q_1 = 0$ and $\beta_{11} = \beta_{21} = 0$. Dashed lines: $p_1 = q_2 = 4$. Solid lines: $p_1 = q_2 = 1$. The horizontal dashed lines represent Bayes factor 10 (above which there is strong evidence against the alternative model, compared with D4, according to the Jeffreys interpretion) and Bayes factor 100 (the lower bound for decisive evidence against). Based on Smith *et al.* (2001).

On the other hand, Wigley *et al.* (1998) argued that for the optimal climate sensitivity, a solar-forcing-only model could achieve almost as good results as one involving greenhouse gases; by their argument, the principal objection to this model was the unrealistic climate sensitivity that was required to achieve such agreement. In fact, as shown in Fig. 7.11, even model F4 is clearly a worse fit than model D4 (Bayes factor around 100), while the climate sensitivity required to achieve it, of 15.4 °C, is completely unrealistic according to current theories of climate change.

Fig. 7.12 shows a number of diagnostic plots related to the model fits, plots (a) and (b) showing the raw data plotted on the same graph as the signals, where the latter include the SOI component, and the remaining plots showing various forms of residual plot. These generally confirm the good fit of the model to the data. Fig. 7.13 tests the assumption of a six-month lag in SOI by computing log likelihoods for the full model assuming various lags in SOI — as can be seen, the fit for lags of between 4 and 7 months are virtually indistinguishable, while lags outside that range are clearly inferior.

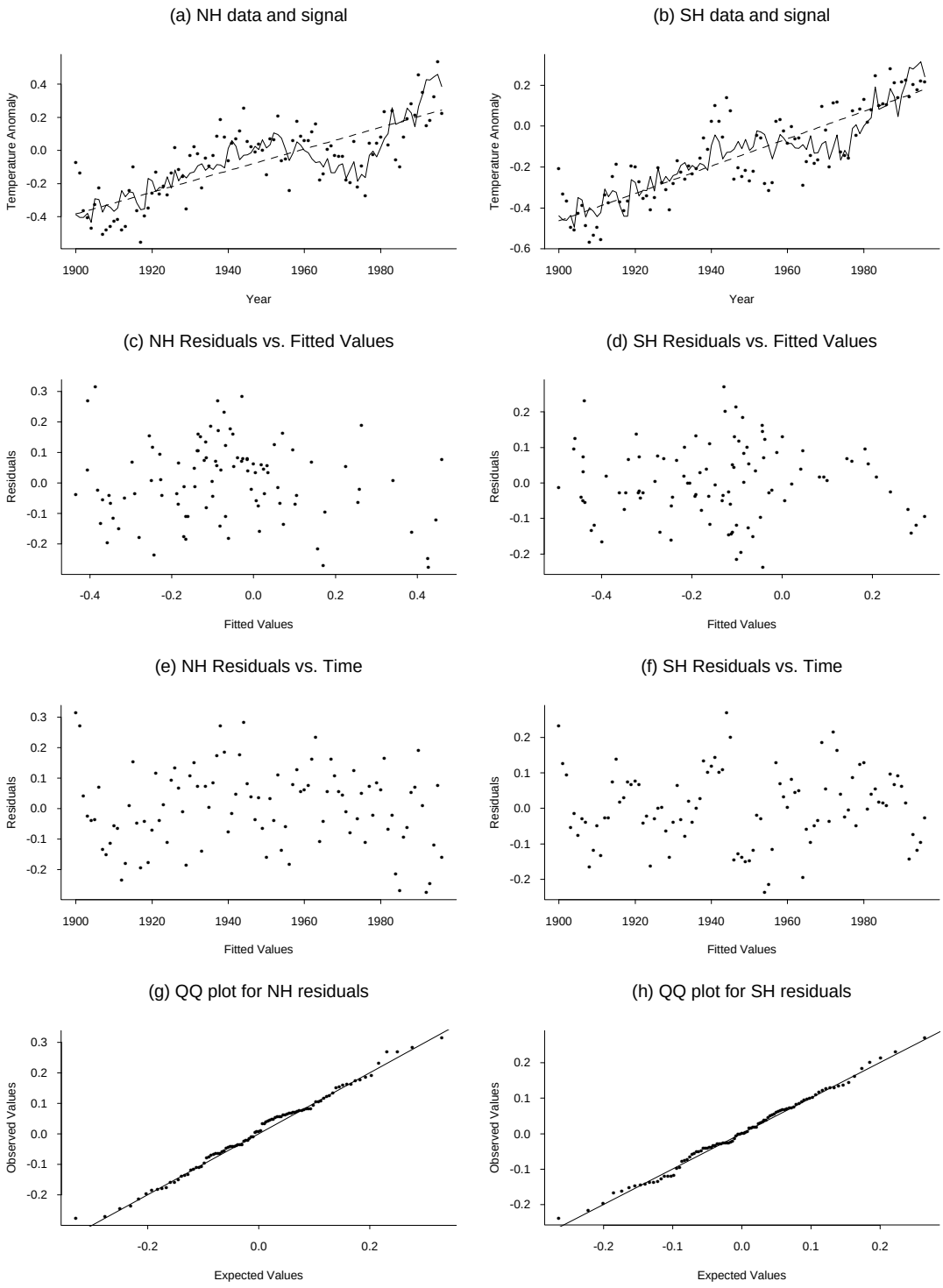


Fig. 7.12. Diagnostic plots. (a,b) Raw data with best-fitting straight-line and model-based trends. (c,d) Residuals (from regression on anthropogenic+solar model) vs. fitted values. (e,f) Residuals vs. time. (g,h) QQ plots of residuals. Based on Smith *et al.* (2001).

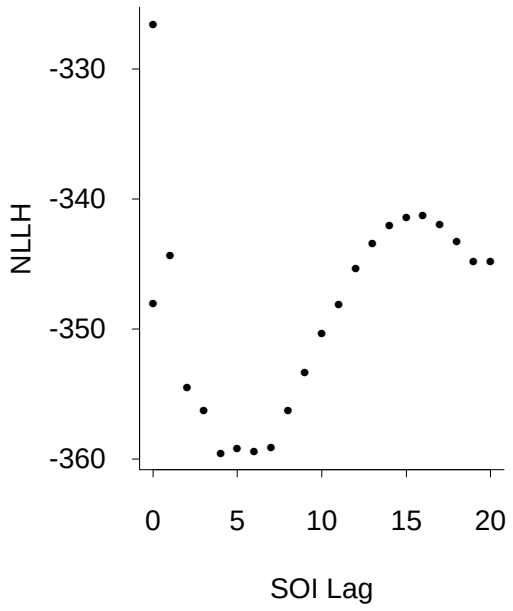


Fig. 7.13. NLLH values for model with various lags of SOI, versus the lag in months. Results are for model (2) with $p_1 = q_2 = 1$, $p_2 = q_1 = 0$, based on model D4 for the forcing component. Based on Smith *et al.* (2001).

The conclusion is therefore that the best model fit, among the range of models considered, is achieved when all three of the most commonly adopted forcing factors — greenhouse gases, sulfate aerosols and solar fluctuations — are included in the model, and with a climate sensitivity of between 4 and 4.5°C. However, apart from the possibility of including other climatic effects (such as volcanoes) in the model, the analysis still leaves open the question of whether a bivariate AR(1) model is an adequate representation of the time series effects, given that Fig. 7.10 implies there is still evidence to support long-range dependence.

CHAPTER 8

Extreme Value Statistics in Meteorology and the Environment

In this chapter, we present an overview of extreme value statistics, with particular attention to applications in meteorology. The extreme value distributions are reviewed, both in their conventional form (Fisher-Tippett, Gumbel, etc.) and in the more modern "threshold" form based on the generalized Pareto distribution. After some initial discussion of graphical techniques and simple summary statistics, an overview is given of the two principal methods used to fit statistical models: the method of maximum likelihood, and Bayesian statistics. This is followed by a number of examples. The latter part of the chapter is concerned with diagnostics for extreme value models, and some extensions to spatial data which mirror some of the spatial techniques discussed in earlier chapters.

8.1. Introduction

Much of conventional statistics is concerned with problems of the following types:

(a) Finding the probability distribution most appropriate to describe a set of data — for example, whether the positive values from a rainfall series should be modelled by a two-parameter gamma distribution, which is the most common choice, or some other distribution such as lognormal or Weibull.

(b) Estimating, or testing hypotheses about, key parameters — for example, the warming or cooling trend in a set of temperature data.

(c) Studying the relationships among two or more variables — for example, one might want to study the relationship between temperature at a single station or a group of stations, and a circulation index such as SOI.

(d) Time series methods, in which the evolution of some quantity over time is studied, taking into account correlations among successive time points.

Most statistical methods are concerned primarily with what goes on in the center of a statistical distribution, and do not pay particular attention to the tails of a distribution, or in other words, the most extreme values at either the high or low end. Indeed, the whole philosophy of one highly studied area of statistics, that concerned with *robust methods* (Hampel *et al.* 1986), is that it is a bad thing for statistical methods to be too much affected by extreme values.

There are some situations, however, in which the extreme values are the most important part of the problem. Hydrologists often want to compute the *N-year return level* [1], where, for instance, $N = 100$, of a variable such as river height or sea level. In the literature on atmospheric pollution, interest is often focussed on trends in the level of some pollutant, but the estimated trend may be different at the more extreme levels of the process than in the center of the distribution. [2] And in climatological studies, much recent attention has been given to whether global climate change can be observed in the more extreme values of observational series. For example, Karl *et al.* (1996) defined a *climate extremes index* (CEI) based on a combination of variables concerning extreme high or low temperatures, high rainfall amounts, and droughts, and concluded that "the climate of the United States has become more extreme in recent decades". Easterling *et al.* (1997) concluded that the observed warming in *global* temperatures is due primarily to a decrease in the diurnal temperature range (DTR), a statement which though not directly connected with long-term extremes of temperatures, nevertheless raises many questions along the lines "is it true that the frequency of very cold days is decreasing?".

As a more concrete example of the problems we shall be discussing, see Fig. 8.1. This plot shows all daily peak gusts over 40 knots over a 20-year period at three stations in North Carolina, Charlotte, Greensboro and Raleigh, indexed C, G and R respectively. This raises a number of questions for consideration:

1 What kind of distribution is appropriate for these data?

2 Should a separate distribution be fitted to hurricane or near-hurricane conditions (greater than 65 knots say), or can these be subsumed within the overall distribution?

3 Is there any tendency for the frequency of extreme winds to increase or decrease during this period?

4 What about dependence between the stations, e.g. is there a tendency for extreme values to occur simultaneously at two or more of the stations?

[1] There is no universally agreed definition of this quantity, but one definition is that it is the level exceeded in any *one* year with probability $1/N$. The less precise definition as the level which is exceeded "once in N years" leaves open all sorts of questions about what this means if there is some trend in the data.

[2] This phenomenon has, for instance, been observed in studies of U.S. tropospheric ozone data (Smith 1989, Huang and Smith 1997), in which there is often no trend at all in the middle of the distribution, but a clear downward trend at the more extreme levels, at the levels which EPA regulations specifically try to control.

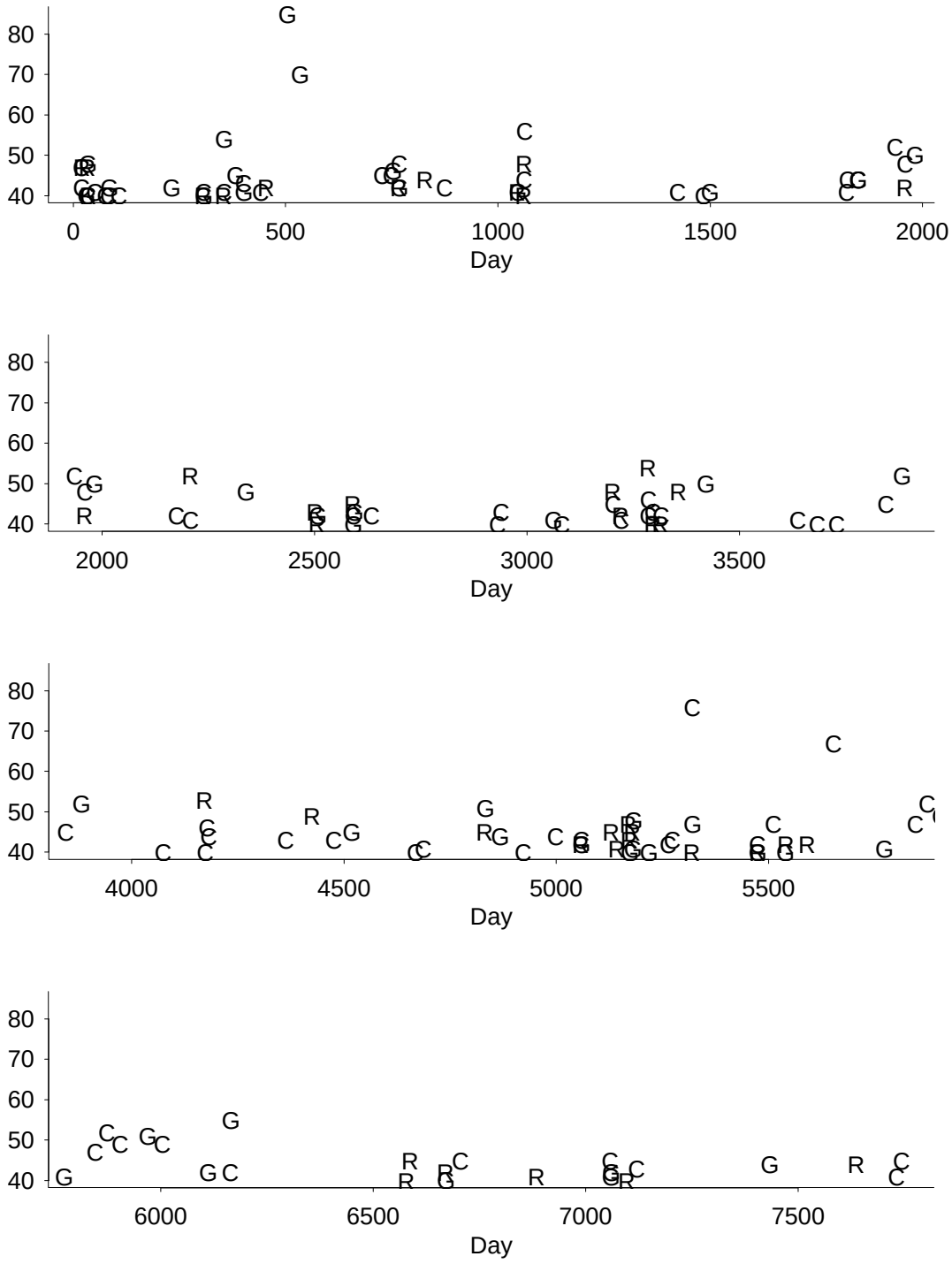


Fig. 8.1. Windspeeds over 40 knots plotted against day of occurrence for Charlotte (C), Greensboro (G) and Raleigh (R).

The purpose of this chapter is to describe a body of methods specifically designed to answer questions of this nature. The cornerstone of the theory is the "three types theorem" first stated by Fisher and Tippett (1928), and given rigorous mathematical justification by Gnedenko (1943), to the effect that there are only three "types" of distributions which can arise as limiting distributions of extremes in random samples. The precise meanings of these statements will be given in section 8.2. This theory led Gumbel, in a long series of papers culminating with his book (Gumbel 1958), to propose a statistical methodology for extreme values based on fitting the extreme value distributions to data consisting of maxima or minima of some random process over a fixed time intervals. For instance, in hydrology Gumbel's methods were often applied to the *annual maxima* of a series of river flows. From a more modern computer-aided perspective, such methods lead to fitting annual maxima or minima with the *generalized extreme value distribution*, which combines the three types of Fisher-Tippett and Gnedenko into a single three-parameter distribution (Jenkinson 1955, NERC 1975, Prescott and Walden 1980, Smith 1985, Hosking *et al* 1985).

In recent years, however, the emphasis of the methodology has changed towards methods based on *exceedances over thresholds* rather than annual maxima. There is an analog of the "three types theorem" in this context, but it leads to a different distribution, the *generalized Pareto distribution*. This is described in section 8.3.

Threshold methods are more flexible than annual maximum methods, for a number of reasons. First, by taking *all* exceedances over a suitably high threshold into account, they use the data more efficiently. Second, they are easily extended to situations where one wants to study how the extreme levels of a variable Y depend on some other variable X —for instance, Y may be the level of tropospheric ozone on a particular day and X a vector of meteorological variables for that day (Smith and Shively 1995). This kind of problem is almost impossible to handle through the annual maximum method.

Because of their greater scope and flexibility, the main emphasis in the present review is given to threshold methods, though the annual maximum approach is also covered, more briefly. There is an intermediate approach, the *r-largest order statistics method*, in which an appropriate joint distribution is fitted to the largest r order statistics in each year. Here, $r = 1$ is the usual annual maximum method, but by allowing $r > 1$ it is possible to take other large values of the series into account, so permitting more efficient estimation. However, as an all-purpose statistical strategy, this method is nowhere near as general and flexible as the threshold method.

Practical implementation of these methods requires methods for estimating their parameters. "Estimating" here means not just finding point estimators for the unknown parameters, but also providing interval estimates or standard errors so that their accuracy may be assessed, and testing hypotheses. There are two "general purpose" methods for estimating parameters of arbitrary distributions:

(1) Maximum likelihood,

(2) Bayesian methods.

Both of these require numerical computation, and one disadvantage is that that the computations are not easily performed with standard statistical packages. There are some programs available which are specifically tailored to extreme values, and a brief review of these is given at the end of this section. Rather than concentrate on packages, my intention in this chapter is to describe the kind of computations that are required at a level of detail that an experienced programmer could write suitable code for him/herself. The main requirement for maximum likelihood estimation is a good subroutine for unconstrained nonlinear optimization. This is something in virtually every package of mathematical subroutines — for example, many of the examples here have been derived using the DFPMIN subroutine of Press *et al.* (1986), which also has implementations in Pascal and C — but readers who already have their own favorite subroutine should have little difficulty adapting it for the purpose in hand. Bayesian methods are rather more complicated, since the main computational technique required is numerical integration, which in most cases, but especially in high dimensions, is a significantly more difficult problem than numerical optimization. In recent years, however, Bayesian methods have been transformed by the use of simulation techniques known as *Markov chain Monte Carlo* (MCMC) methods, which are so easy to apply that the beginner can easily try them out with very little prior experience. A number of recent books have described these methods, including Gilks *et al.* (1996) and Carlin and Louis (1996). It needs to be pointed out, however, that these are not "automatic" methods and that it takes some experience to apply them efficiently and correctly, especially in high-dimensional problems.

The present chapter concentrates on maximum likelihood methods as a principal tool of statistical inference (section 8.5). However, I also cover the subject of Bayesian methods (section 8.6) since these are becoming so popular and because many of the more advanced problems are more satisfactorily solved using Bayesian rather than maximum likelihood methods.

It is not my intention here to dwell on philosophical differences between maximum likelihood (frequentist) and Bayesian approaches, any more than their mathematical foundations, since my main purpose is to focus on practicalities of their implementation. Readers interested in a general discussion of statistical methods and of differences between frequentist and Bayesian approaches are referred to such books as Cox and Hinkley (1974), Berger (1985) and Bernardo and Smith (1994).

Software

Although extreme value methods have yet to be incorporated into standard packages such as S-Plus and SAS, a number of specific packages have been produced.

A package written by Stuart Coles and Mark Dixon is available together with accompanying lecture notes (Coles 1996) from Stuart Coles' home page. Many of the features

described in the present chapter are in this package, which also produces extensive graphical output. The package is written in the statistical programing language SPlus. The web address is http://www.maths.lancs.ac.uk/ ~coless/.

A second package which has many similar features, but which is more geared towards insurance and financial applications, is the EVIS package written by Alexander McNeil of ETH (Zürich). It is also written in SPlus, and available from http://www.math.ethz.ch/ ~mcneil/software.html.

Another reference is the book by Reiss and Thomas (1997), which comes together with its own CD of computer programs, forming the Xtremes package. This has excellent graphics and display features.

Other literature on extremes

Gumbel's (1958) book is still regarded as a classic, though since the techniques described belong to the pre-computer age, nobody uses Gumbel's actual methods any more (except the Gumbel plot — see section 8.9.1).

Books concentrating primarily on the probabilistic theory of extremes are Galambos (1987), Leadbetter *et al.* (1983) and Resnick (1987). Galambos's book covers a wide range but not in the same depth as the other two. Resnick is particularly strong on multivariate extreme value theory. Leadbetter *et al.* give an excellent introduction to the whole subject and are particularly strong on extreme values in stationary stochastic procesess.

A recent addition to this literature is Embrechts *et al.* (1997), which is particularly strong on applications in insurance and finance.

A more applied book is Castillo (1988), with particular focus on engineering applications.

Two very good collections of edited chapters, both resulting from major conferences in extreme value theory, are by Tiago de Oliveira (1984) and Galambos *et al.* (1994).

Finally there are a number of review articles, including one of mine (Smith 1990). A number of the following examples are taken from this.

8.2. The extreme value distributions

The extreme value distributions formally arise as limiting distributions for maxima or minima of a sequence of random variables. Throughout the chapter we shall concentrate on maxima rather than minima, though the results are easily translated from one to the other by replacing random variables $X_1,\ X_2, ...$, by their negatives $-X_1,\ -X_2, ...$

A formal definition is as follows. Suppose $X_1, X_2, ...$ are independent random variables with a common distribution function F; in other words

$$F(x) = \Pr\{X_j \le x\} \tag{8.1}$$

for each j and x. The distribution function of the maximum $M_n = \max\{X_1, ..., X_n\}$ is given by the n'th power of F:

$$\begin{aligned} \Pr\{M_n \le x\} &= \Pr\{X_1 \le x, X_2 \le x, ..., X_n \le x\} \\ &= \Pr\{X_1 \le x\} \Pr\{X_2 \le x\} ... \Pr\{X_n \le x\} \\ &= F^n(x). \end{aligned} \tag{8.2}$$

However for each x within the range of the distribution, $0 < F(x) < 1$ and so $F^n(x) \to 0$ as $n \to \infty$; this, although (8.2) states precisely what is the distribution of the maximum, it does not tell us anything interesting or worthwhile about what happens in large samples, in other words, as $n \to \infty$.

It turns out that we can get interesting results if we *renormalize*: define scaling constants $a_n > 0$ and b_n so that

$$\begin{aligned} \Pr\left\{ \frac{M_n - b_n}{a_n} \le x \right\} &= \Pr\{M_n \le a_n x + b_n\} \\ &= F^n(a_n x + b_n) \\ &\to H(x) \quad \text{as } n \to \infty \end{aligned} \tag{8.3}$$

where H is nondegenerate; in other words, a probability distribution function which is not always either 0 or 1.

For our present purposes we are much less interested in the constants a_n and b_n than the form of the limiting distribution H; it turns out that there are only *three types* [3] of limiting distribution and these are given by

- *Gumbel type*:

$$H(x) = \exp\{-\exp(-x)\}, \quad -\infty < x < \infty, \tag{8.4}$$

- *Fréchet type*:

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ \exp(-x^{-\alpha}) & \text{if } 0 < x < \infty, \end{cases} \tag{8.5}$$

[3] Two distribution functions H_1 and H_2 are said to be *of the same type* if one can be transformed into the other through an equation of the form $H_1(x) = H_2(Ax + B)$ where $A > 0$ and B are fixed constants. If (8.3) holds for some H, then it also holds for any other H of the same type, by redefining the constants a_n and b_n. Therefore, in talking about which limits H can arise, all distribution functions of the same type must be treated as equivalent.

- *Weibull type*:

$$H(x) = \begin{cases} \exp\{-(-x)^{\alpha}\} & \text{if } -\infty < x < 0, \\ 1 & \text{if } x > 0. \end{cases} \tag{8.6}$$

In both (8.5) and (8.6), $\alpha > 0$ is some fixed constant.

The three types theorem and the related *domain of attraction problem* — in other words, which distribution functions F converge to H through an operation of the form (8.3) for suitable a_n and b_n — are classical problems of the mathematical theory of extreme values, but I shall not attempt to treat them in any detail here. There are a number of excellent textbook treatments, for example Leadbetter *et al.* (1983). A brief summary of the main results is

- Any F whose tail is of power law form,

$$1 - F(x) \sim cx^{-\alpha}, \quad x \to \infty \tag{8.7}$$

for constants $c > 0$ and $\alpha > 0$, is in the domain of attraction of the Fréchet type (with the same α). Examples include the Pareto and t distributions.

- Any F with a finite endpoint ω_F, such that $F(\omega_F) = 1$ but $F(x) < 1$ for any $x < \omega_F$, but with power law behavior as $x \uparrow \omega_F$, so that

$$1 - F(\omega_F - y) \sim cy^{\alpha}, \quad y \downarrow 0, \tag{8.8}$$

with constants $c > 0$ and $\alpha > 0$, is in the domain of attraction of the Weibull type. In most textbook treatments, this is applied to minima rather than maxima, and in situations where it is obvious that there is a finite lower bound, e.g. strength of materials, for which there is the natural lower bound 0. Indeed this was the context of Weibull's original derivation of this distribution. In meteorological contexts, the Weibull type may arise in situations where we have good reason to believe that there is that there is a practical upper bound on the random variable being considered, such as temperature. However, the reader should be cautioned against the assumption that the Weibull type *automatically* arises in this context; we still need the polynomial tail assumption (8.8).

- The most common type of distribution in the domain of attraction of the Gumbel type is one for which the endpoint ω_F is infinite but the tail distribution $1 - F(x)$ decays faster than the polynomial case (8.7). A more precise condition is *von Mises' condition*: if $F(x)$ is the distribution function and $f(x) = dF(x)/dx$ is the density, and if

$$\frac{d}{dx}\left\{\frac{1 - F(x)}{f(x)}\right\} \to 0 \quad \text{as } x \to \omega_F, \tag{8.9}$$

then there exist constants $a_n > 0$ and b_n for which (8.3) holds, with H given by (8.4). Examples of F include many common distribution functions, e.g. normal, lognormal,

exponential, Weibull, gamma, etc. All of these have endpoint $\omega_F = \infty$ though (8.9) is valid for both finite and infinite ω_F.

The three types of extreme value distribution may be combined into a single family known as the *generalized extreme value distribution* (abbreviated to GEV) given by

$$H(x;\mu,\psi,\xi) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\psi}\right)^{-1/\xi}\right\} \tag{8.10}$$

defined on the region for which $1+\xi(x-\mu)/\psi > 0$ — elsewhere, H is either 0 or 1. In (8.10), μ is a location parameter, ψ a scale parameter and ξ the all-important *shape parameter* which determines the nature of the tail of the distribution. The case $\xi > 0$ is the Fréchet type with $\alpha = 1/\xi$, the case $\xi < 0$ is of Weibull type with $\alpha = -1/\xi$, while the case $\xi = 0$ depends on the elementary calculus result that

$$\lim_{\xi\to 0} H(x;\mu,\psi,\xi) = \exp\left\{-\exp\left(-\frac{x-\mu}{\psi}\right)\right\}, \tag{8.11}$$

in other words, the Gumbel distribution with arbitrary location and scale parameters.

Here are a few basic properties of the GEV distribution. The mean exists if $\xi < 1$ and the variance if $\xi < \frac{1}{2}$; more generally, the k'th moment exists of $\xi < \frac{1}{k}$. The mean and variance are given by

$$\begin{aligned} \mu_1 &= \mathrm{E}(X) = \mu + \frac{\psi}{\xi}\{\Gamma(1-\xi)-1\}, \qquad (\xi < 1) \\ \mu_2 &= \mathrm{E}\{(X-\mu_1)^2\} = \frac{\psi^2}{\xi^2}\{\Gamma(1-2\xi)-\Gamma^2(1-\xi)\}, \qquad (\xi < \frac{1}{2}) \end{aligned} \tag{8.12}$$

where $\Gamma(\cdot)$ is the Gamma function. In the limiting case $\xi \to 0$, these reduce to

$$\mu_1 = \mu + \sigma\gamma, \qquad \mu_2 = \frac{\psi^2\pi^2}{6}, \tag{8.13}$$

where $\gamma = .5772...$ is Euler's constant.

In most applications to environmental processes, the extreme value distributions are used as approximations to the annual maxima of a process (or maxima over some other time period) without detailed consideration of how they are derived. It is usually implicitly assumed that since the annual maximum can be represented as the maximum of a very large number of daily or hourly values, the approximation represented by (8.3) is reasonable. One objection to this is that environmental processes rarely produce observations that are independent and identically distributed (IID). However, there is an extensive theory of extreme value theory for non-IID processes and it is known that the classical extreme value distributions very often arise in this context as well. One particularly rich theory

along these lines is that of extreme values in stationary processes (Leadbetter *et al.* 1983). A second objection is that sometimes it is argued that alternative distributional families fit the data better — for example, in the 1970s there was a lengthy debate among hydrologists over the use of extreme value distributions as compared with those of log Pearson type III. There is no universal solution to this kind of debate. It is quite possible that, for any particular data set, the log Pearson type III or some other family will be found to fit better than the GEV. However, there are also dangers in making the analysis over-adaptive: simple models with small numbers of parameters generally have better statistical properties than those with more parameters. My own advice is to use the extreme value distributions as the basis for any analysis, but also to examine carefully the goodness of fit of the model, and to be ready to consider alternative forms of model specification, such as including a trend, or the dependence on some other covariate besides time, as an alternative to assuming that the GEV parameters are constant for the whole process.

8.3. Threshold exceedances and the Poisson-GPD model

We now turn to the main alternative approach to extreme value statistics, based on exceedances over high thresholds.

The basic idea is to pick a high threshold u — how exactly this is to be chosen is the subject of considerable discussion later on — and to study all the *exceedances* of u. This means two things: how many exceedances there are over a given time period, and the *excess values*, in other words, the amounts by which the threshold is exceeded. For the latter, we use the *generalized Pareto distribution* (GPD), which is the analog for threshold exceedances of the GEV distribution for annual maxima.

More precisely, suppose X is a random variable whose distribution function is F, and let $Y = X - u$ conditioned on $X > u$. Then

$$\Pr\{Y \le y\} = \Pr\{X \le u + y | X > u\} = F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)}. \tag{8.14}$$

The interest here is as u approaches the (finite or infinite) upper endpoint ω_F. In that case we have an approximation of the form

$$F_u(y) \approx G(y; \sigma_u, \xi) \tag{8.15}$$

where G is the *generalized Pareto distribution* (GPD) given by

$$G(y; \sigma, \xi) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi}. \tag{8.16}$$

The meaning of (8.15) is that for sufficiently high thresholds u, there is some σ_u (which depends on u) and some ξ (which does not) for which the GPD is a very good approximation to the excess distribution function F_u. The connection was made precise by Pickands

(1975), who showed that (8.15) is valid as an approximation whenever (8.3) holds, and that in this case, the ξ that arises in (8.15) is the same as in the GEV representation of H in (8.3).

With the GPD, like the GEV, there are three different cases depending on the sign of ξ:

1. If $\xi > 0$, then (8.16) is valid on $0 < x < \infty$ and the tail distribution function satisfies $1 - G(y; \sigma, \xi) \sim cy^{-1/\xi}$ with $c > 0$; this is a traditional "Pareto tail".

2. If $\xi < 0$, the G has an upper endpoint at $\omega_G = \sigma/|\xi|$, similar to the Weibull type of classical extreme value theory.

3. If $\xi = 0$, then in similar fashion to (8.11), we have

$$G(y; \sigma, 0) = 1 - \exp\left(-\frac{y}{\sigma}\right), \tag{8.17}$$

the exponential distribution with mean σ.

For some further basic properties, see Davison and Smith (1990). As with the GEV distribution, the mean exists if $\xi < 1$, and the variance if $\xi < \frac{1}{2}$, being given by

$$\mathrm{E}(Y) = \frac{\sigma}{1-\xi}, \mathrm{Var}(Y) = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}. \tag{8.18}$$

Another property which comes in useful later (section 8.9.3) is

$$\mathrm{E}(Y - w | y > w) = \frac{\sigma + \xi w}{1 - \xi}, \tag{8.19}$$

valid for any $w > 0$, provided $\xi < 1$.

We next consider how to combine the information on excess values with that on the exceedance times of a fixed threshold u. The simplest case to consider is when the underlying process consists of IID random variables. In that case, the Poisson property of exceedances (Leadbetter *et al.* 1983) suggests the following model which we call the *Poisson–GPD model*:

1. The number, N, of exceedances of the level u in any one year has a Poisson distribution with mean λ,

2. Conditionally on $N \geq 1$, the excess values $Y_1, ..., Y_N$ are IID from the GPD (8.16).

This model is closely related to the GEV distribution for annual maxima, as follows. Suppose $x > u$. The probability that the annual maximum of the process just described is less than x is

$$\begin{aligned}\Pr\{\max_{1\le i\le N} Y_i \le x\} &= \Pr\{N=0\} + \sum_{n=1}^{\infty} \Pr\{N=n,\ Y_1 \le x,\ \ldots\ Y_n \le x\} \\ &= e^{-\lambda} + \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \cdot \left\{1 - \left(1+\xi\frac{x-u}{\sigma}\right)^{-1/\xi}\right\}^n \qquad (8.20) \\ &= \exp\left\{-\lambda\left(1+\xi\frac{x-u}{\sigma}\right)^{-1/\xi}\right\}.\end{aligned}$$

This expression is the same as (8.10) if

$$\sigma = \psi + \xi(u-\mu), \qquad \lambda = \left(1+\xi\frac{u-\mu}{\psi}\right)^{-1/\xi}. \qquad (8.21)$$

Thus the GEV and GPD models are entirely consistent with one another above the GPD threshold, and moreover, (8.21) shows exactly how the Poisson–GPD parameters σ and λ vary with u.

The Poisson–GPD model applies, in its most literal form, only if the underlying process is IID. For dependent processes, a variant long in use among hydrologists is the *peaks over threshold* (POT) method (also known as the partial duration series method), see e.g. NERC (1975), North (1980). The idea behind the method is that high exceedances occur in clusters — in a meteorological context, one cluster might represent a single storm or depression. By separating out the *peaks* within clusters, these will be approximately independent and therefore amenable to the Poisson–GPD model. From a mathematical point of view this is consistent with the modern view of extremes in stationary processes, which shows that under very general conditions, exceedances over high thresholds occur in clusters while the clusters approximately follow a Poisson process (Leadbetter *et al.* 1983, Hsing *et al.* 1988). However in the early hydrological literature, the excess distribution was nearly always assumed to be exponential. This is of course the special case $\xi = 0$ of the GPD, but the GPD allows for a much richer variety of tail behavior. For statistical aspects and examples, see Davison and Smith (1990). The question of how best to define clusters was considered in more detail by Smith and Weissman (1994), and an alternative approach based on modeling dependence directly has been given by Smith *et al.* (1997), but the direct approach, of picking out peaks and then applying the Poisson–GPD model, is certainly easier to handle.

Another way in which the IID assumption may be violated is when the distribution function F is not constant, an elementary example being if the process is seasonal. Approaches to this problem (Davison and Smith 1990) include

(a) Remove seasonal trend from the data before applying the threshold approach. This is the simplest method, but it assumes that the process may be simply decomposed as signal+noise, which may not be valid.

(b) The "separate seasons" approach: subdivide the year into homogeneous seasons and apply the Poisson–GPD model separately within each.

(c) Expand the Poisson–GPD model to include covariates. This is the most flexible approach, and is considered further in section 8.9.

8.3.1. Examples of extreme value distributions

This subsection illustrates the extreme value limit distributions and threshold exceedances via a number of specific examples of the limiting operations. This is more mathematically advanced than the previous material and can be omitted without loss of continuity.

Example 1: The exponential distribution.

Suppose $F(x) = 1 - e^{-x}$. Let $a_n = 1,\ b_n = \log n$. Then

$$\begin{aligned} F^n(a_n x + b_n) &= \left(1 - e^{-x - \log n}\right)^n \\ &= \left(1 - \frac{e^{-x}}{n}\right)^n \\ &\to \exp\left(-e^{-x}\right) \end{aligned}$$

using the well-known limit $(1 + \frac{z}{n})^n \to e^z$ as $n \to \infty$, which is valid for any real or complex z. Thus in the case of the exponential distribution, the appropriate limiting form for sample maxima is the Gumbel distribution.

For the threshold version of the result, set $\sigma_u = 1$. Then

$$\begin{aligned} F_u(\sigma_u z) &= \frac{F(u + \sigma_u z) - F(u)}{1 - F(u)} \\ &= \frac{e^{-u} - e^{-u-z}}{e^{-u}} \\ &= 1 - e^{-z} \end{aligned}$$

so in this case the exponential distribution is the exact distribution for exceedances over a threshold, and therefore is automatically the limiting distribution as $u \to \infty$. Of course, the exponential distribution is a special case of the GPD with $\xi = 0$.

Example 2: Pareto-type tail

Suppose $1-F(x) \sim cx^{-\alpha}$ as $x \to \infty$, with c and α both positive. This form covers the Pareto distribution and also some well-known distributions such as t and F distributions. Let $b_n = 0,\ a_n = (nc)^{1/\alpha}$. Then for $x > 0$,

$$\begin{aligned} F^n(a_n x) &\approx \left\{1 - c\,(a_n x)^{-\alpha}\right\}^n \\ &= \left(1 - \frac{x^{-\alpha}}{n}\right)^n \\ &\to \exp\left(-x^{-\alpha}\right). \end{aligned}$$

So in this case the limiting distribution is Fréchet.

For the threshold form of this result, let $\sigma_u = ub$ where $b > 0$ is to be determined. Then

$$\begin{aligned} F_u(\sigma_u z) &= \frac{F(u + \sigma_u z) - F(u)}{1 - F(u)} \\ &\approx \frac{cu^{-\alpha} - c(u + ubz)^{-\alpha}}{cu^{-\alpha}} \\ &= 1 - (1 + bz)^{-\alpha}. \end{aligned}$$

If we now let $\xi = \frac{1}{\alpha}$ and set $b = \xi$, the limit distribution is exactly as given by (8).

Example 3: Finite upper endpoint

Suppose now $\omega_F = \omega < \infty$ and

$$1 - F(\omega - y) \sim cy^{\alpha} \tag{8.22}$$

as $y \downarrow 0$ for positive c and α. Many, though not all, distributions with finite endpoints are of this form. For example, consider the Beta distribution, with density

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

Then $\omega = 1$ and for $y \downarrow 0$,

$$\begin{aligned} 1 - F(1-y) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{1-y}^{1} x^{a-1}(1-x)^{b-1} dx \\ &\sim \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{y^b}{b} \end{aligned}$$

which is of the form (8.22) with $\alpha = b$ and

$$c = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b+1)}.$$

For a distribution satisfying (8.22), now, set $b_n = \omega$, $a_n = (nc)^{1/\alpha}$. Then for $x < 0$,

$$\begin{aligned} F^n(a_n x + b_n) &= F^n(\omega + a_n x) \\ &\approx \{1 - c\,(-a_n x)^\alpha\}^n \\ &\approx \left\{1 - \frac{(-x)}{n}\right\}^n \\ &\to \exp\{-(-x)^\alpha\}. \end{aligned}$$

The corresponding limit when $x > 0$ is obviously 1. Therefore, this is a case of convergence to the Weibull type.

For the threshold version of this result, let u be very close to ω and consider $\sigma_u = b(\omega - u)$ for $b > 0$ to be determined. Then for $0 < z < \frac{1}{b}$,

$$\begin{aligned} F_u(\sigma_u z) &= \frac{F(u + \sigma_u z) - F(u)}{1 - F(u)} \\ &\approx \frac{c(\omega - u)^\alpha - c(\omega - u - \sigma_u z)^\alpha}{c(\omega - u)^\alpha} \\ &= (1 - bz)^\alpha. \end{aligned}$$

This is of GPD form if we set $\xi = -\frac{1}{\alpha}$ and $b = \frac{1}{\alpha}$.

Example 4: Normal Extremes

Let $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$ denote the standard normal distribution function. There is a well-known asymptotic approximation to the tail of Φ (Feller 1968, page 193), which gives

$$1 - \Phi(x) \sim \frac{1}{x\sqrt{2\pi}} e^{-x^2/2} \quad \text{as } x \to \infty. \tag{8.23}$$

Using (8.23), we can establish the result

$$\begin{aligned} \lim_{u\to\infty} \frac{1 - \Phi(u + x/u)}{1 - \Phi(u)} &= \lim_{u\to\infty} \left(1 + \frac{x}{u^2}\right)^{-1} \exp\left\{-\frac{1}{2}\left(u + \frac{x}{u}\right)^2 + \frac{1}{2}u^2\right\} \\ &= e^{-x}. \end{aligned} \tag{8.24}$$

This result is used in two ways. First, defining $\sigma_u = 1/u$, we see that

$$\frac{\Phi(u + \sigma_u z) - \Phi(u)}{1 - \Phi(u)} \to 1 - e^{-z} \quad \text{as } u \to \infty,$$

showing that in this case the limiting distribution of exceedances over thresholds is exponential.

The second use of (8.24) is for the classical extreme value distribution. Suppose we define b_n by the property

$$\Phi(b_n) = 1 - \frac{1}{n}. \tag{8.25}$$

Let $a_n = 1/b_n$. Then (8.24) shows that

$$n\left\{1 - \Phi\left(a_n x + b_n\right)\right\} = \frac{1 - \Phi\left(a_n x + b_n\right)}{1 - \Phi\left(b_n\right)} \to e^{-x}$$

and hence

$$\Phi^n\left(a_n x + b_n\right) \approx \left(1 - \frac{e^{-x}}{n}\right)^n \to \exp\left(-e^{-x}\right), \tag{8.26}$$

establishing the classical form of extreme value convergence with Gumbel limiting distribution.

We conclude by noting two refinements of this result. First, although (8.25) is not an explicit formula for b_n, for practical purposes b_n is easily calculated numerically. Nevertheless, some people like to see an explicit formula, and one such which has been proposed is

$$b_n = (2\log n)^{1/2} - \frac{1}{2}(2\log n)^{-1/2}\{\log\log n + \log(4\pi)\} \tag{8.27}$$

It is left as an exercise for the reader to show, using (8.23), that with b_n defined by (8.27), both (8.25) and (8.26) are still valid asymptotically. However, Hall (1979) showed that a superior rate of convergence is obtained by using (8.25) directly.

The second refinement is more subtle, but goes back to the original paper by Fisher and Tippett (1928). Although the Gumbel distribution is the correct limiting distribution, one can obtain a better approximation, for any finite value of n, by using a three-parameter Generalized Extreme Value distribution. This is known as the penultimate approximation and leads to results which are accurate at a rate of $O(1/\log^2 n)$, as opposed to $O(1/\log n)$ for (8.26). A rigorous proof of this was given by Cohen (1982a). The practical conclusion is that it is better to use the Generalized Extreme Value type in many cases even when the Gumbel distribution is the true limit — this result holds not just for the normal distribution but for a wide class of alternatives (Cohen 1982b). However, we do not attempt to prove those results here.

A connection with wavelet thresholding

A relatively easy consequence of the preceding results is that if $X_1,\ X_2, ...,$ are independent $N(0,1)$ random variables and $M_n = \max\{X_1, ..., X_n\}$, then

$$\frac{M_n}{(2\log n)^{1/2}} \xrightarrow{p} 1, \tag{8.28}$$

where $\xrightarrow{p}$ denotes convergence in probability.

To see (8.28), first note that it is equivalent to the statement

$$\Pr\left\{\frac{M_n}{(2\log n)^{1/2}} \leq c\right\} \to \begin{cases} 1 & \text{if } c > 1, \\ 0 & \text{if } c < 1. \end{cases} \tag{8.29}$$

To see (8.29), note that the statement

$$\frac{M_n}{(2\log n)^{1/2}} \leq c$$

is equivalent to

$$\frac{M_n - b_n}{a_n} \leq \frac{c(2\log n)^{1/2} - b_n}{a_n}. \tag{8.30}$$

However, the right hand side of (8.30) is asymptotic to

$$(2\log n)^{1/2}\left[(c-1)(2\log n)^{1/2} + \frac{1}{2}(2\log n)^{-1/2}\{\log\log n + \log(4\pi)\}\right]$$
$$\to \begin{cases} +\infty & \text{if } c > 1 \\ -\infty & \text{if } c < 1 \end{cases}.$$

Hence if $c > 1$, for any $K > 0$, for n sufficiently large the right hand side of (8.30) is greater than K, and so the limiting probability in (8.29) is greater than $\exp(-e^{-K})$, which may be made arbitrarily close to 1 by taking K sufficiently large. Hence the limiting probability in (8.29) is 1 when $c > 1$. A similar argument applies when $c < 0$ by taking $K < 0$. This proves (8.29), and hence (8.28).

The connection with wavelet thresholding is that $X_1, ..., X_n$ are a set of coefficients resulting from a wavelet transform, which are standardized so that they are asymptotically independent and normally distributed with variance 1, a typical form of either hard or soft thresholding is to delete all coefficients less than $(2\log n)^{1/2}$, on the grounds that nearly all variables which are not associated with an underlying signal (meaning that the mean of the corresponding X_i is 0) will be eliminated by this procedure. The full justification for this is a little more complicated than (8.28), but nevertheless, (8.28) provides a good intuitive justification of why it works. Full details are in Donoho and Johnstone (1994).

8.4. Alternative probability models

Two other models are discussed here more briefly.

8.4.1 The r largest order statistics model

This is an extension of the GEV model to encompass the r largest order statistics from each year, $r = 1$ being the usual GEV approach.

The theory is based on the fact that for an IID sequence, the joint distribution of the r largest order statistics may be characterized in similar fashion to the GEV itself, see e.g. Leadbetter *et al.* (1983), section 2.3. The key formula is: if $Y_{n,1} \geq Y_{n,2} \geq ... \geq Y_{n,r}$ denote the r largest order statistics of a sample of size n, and if a_n and b_n are the normalizing constants defined in section 8.3, such that $(Y_{n,1} - b_n)/a_n$ converges in distribution to the GEV family (8.10), then

$$\left(\frac{Y_{n,1} - b_n}{a_n}, ..., \frac{Y_{n,r} - b_n}{a_n} \right)$$

converges in distribution to a limiting random vector $(X_1, ..., X_r)$, whose density is

$$h(x_1, ..., x_r) = \psi^{-r} \exp \left\{ - \left(1 + \xi \frac{x_r - \mu}{\psi} \right)^{-1/\xi} - \left(1 + \frac{1}{\xi} \right) \sum_{j=1}^{r} \log \left(1 + \xi \frac{x_j - \mu}{\psi} \right) \right\}.$$

Statistical applications were developed by Smith (1986), Tawn (1988a) and more recently in a novel application by Robinson and Tawn (1995) and Smith (1997a) (see section 8.6).

In general, this is a less flexible method than the threshold approach. The method is only appropriate in this form if the r largest order statistics are essentially independent events, so for a dependent process it is necessary to make a similar restriction to cluster peaks as we have already seen in the case of the threshold method (Tawn 1988a). Extensions of the model to incorporate dependence within clusters as part of the joint distribution may be considered, for example as an application of the formulae in Hsing *et al.* (1988), but these are likely to be too complicated for applications in meteorology and other environmental sciences.

8.4.2 The point process approach

An alternative technique which combines features of both the GEV and threshold models is based on viewing the exceedance times and excess values as part of a two-dimensional point process. Suppose we plot all exceedances of a threshold u in a time period $(0, T)$ on a graph with time as the x-axis and location as the y-axis (Fig. 8.2). Suppose the expected number of excess values within the box $A = (t_1, t_2) \times (y, \infty)$ is given by

$$\Lambda(A) = (t_2 - t_1)\Psi(y; \mu, \psi, \xi) \tag{8.31}$$

where

$$\Psi(y; \mu, \psi, \xi) = \left(1 + \xi \frac{y - \mu}{\psi} \right)^{-1/\xi} \tag{8.32}$$

valid so long as $1 + \xi(y - \mu)/\psi > 0$. We view the function Λ given by (8.31) as the intensity function of a *nonhomogeneous Poisson process*, which means amongst other things that the numbers of excess values within disjoint boxes $A_1, A_2, ...$, are independent Poisson random variables with means $\Lambda(A_1), \Lambda(A_2), ...$

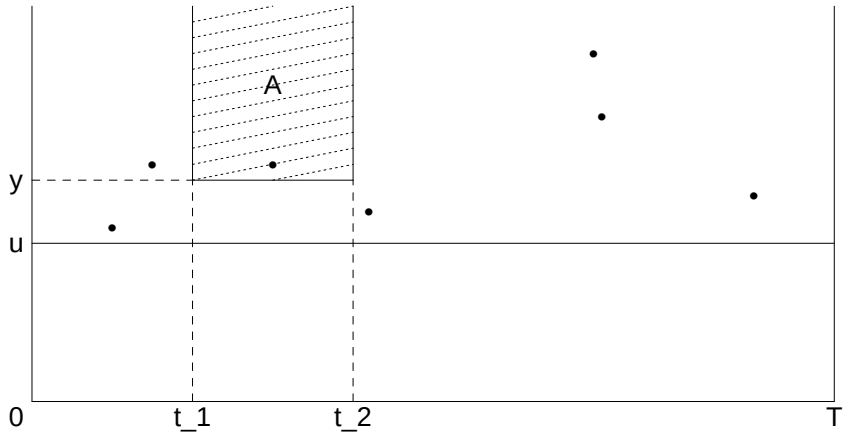


Fig. 8.2. Illustration of point process approach. All exceedances of the process above a level u are noted. The set A consists of all points within the time interval (t_1, t_2) for which the level of the process is above y. The expected number of points in A is given by (8.31).

The motivation for this is based on the fact that the nonhomogeneous Poisson process just described arises in an appropriate limiting sense for exceedances from an IID process, and from a stationary process provided attention is limited to cluster peaks (Leadbetter *et al.* 1983). A corresponding statistical theory was developed by Smith (1989).

The main attraction of this approach is that it allows the GEV and GPD methods to be combined into a single model. The annual maxima of the process indeed follow the GEV distribution (8.10) (with the same parameters μ, ψ and ξ as in (8.32)), while the exceedances over a threshold u can be shown to follow a GPD with scale parameter σ and exceedance rate λ given by (8.21).

Example. Apart from its conceptual value in defining a stochastic process of exceedances, the point process "picture" can also be used to visualize real data. Fig. 8.3 is based on the analysis of a data set by Davison and Smith (1990), of high flow rates from the River Nidd in England. Plot (a) shows all exceedances over the level of 65 cumecs, plotted against the day within the year (1–366). The result shows a clear seasonality in the times of high exceedances, with many fewer exceedances during the summer months. This therefore suggests the need to fit a seasonal model to the data (though for many purposes, where it is only important to consider the total number of exceedances of a threshold rather than the times within the year at which they occur, this feature is actually unimportant for the analysis). Plot (b) shows high exceedances plotted against the total time (number of days) from the start of the series, which was at January 1, 1934. In this case there is some suggestion of an increasing trend in the data, with several high exceedances during

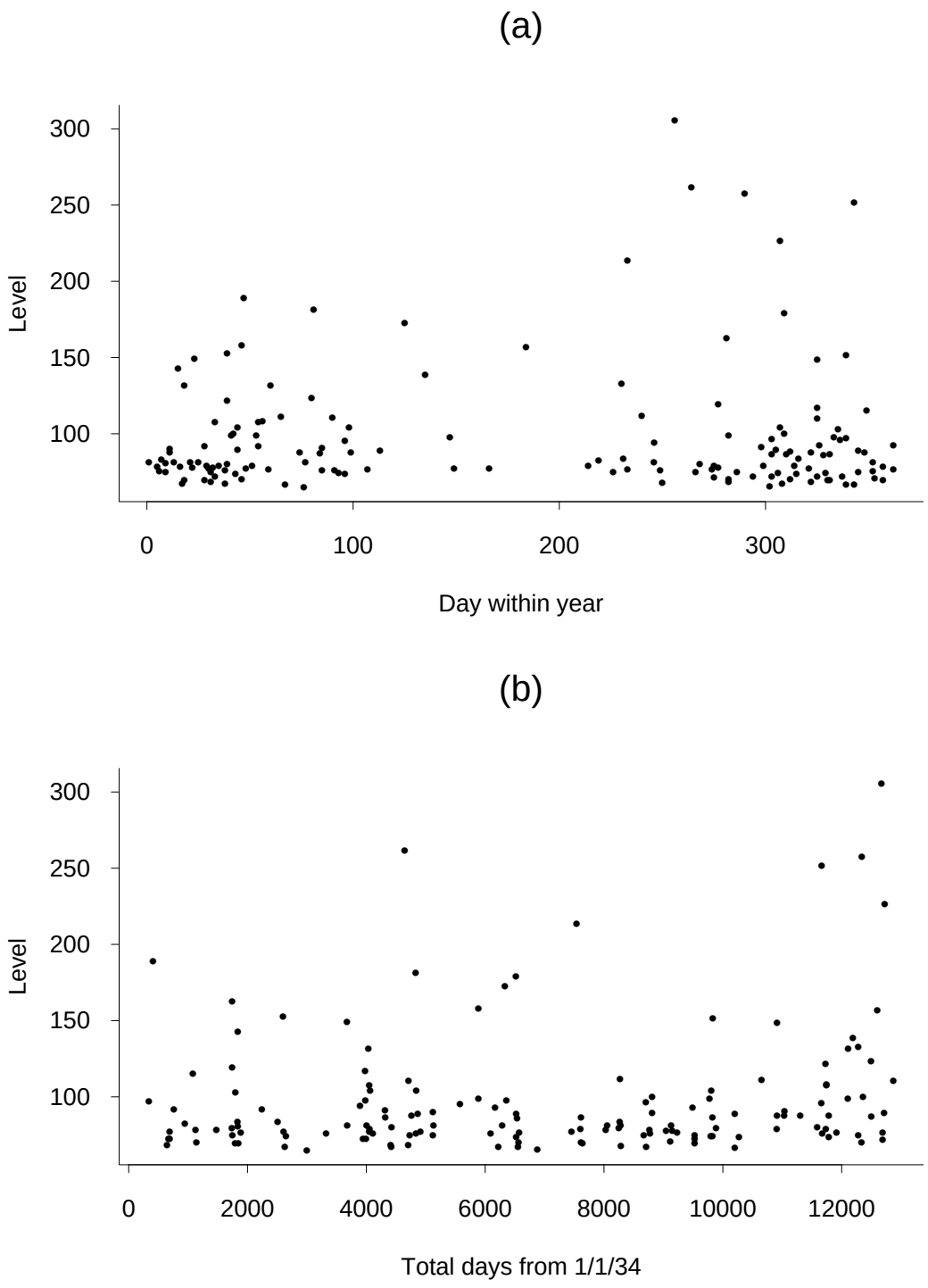


Fig. 8.3. Plots of exceedances of River Nidd, (a) against day within year, (b) against total days from January 1, 1934. Adapted from Davison and Smith (1990).

the last period of the data, though further analysis by Davison and Smith suggested that this did not represent a real increasing trend.

8.5. Statistical methods: maximum likelihood

All the models described so far can be fitted by the method of maximum likelihood (Cox and Hinkley 1974). In this section we give a very brief overview of the main principles behind this approach, with a view towards the GEV and Poisson–GPD models.

Suppose we have data Y whose density is defined by some p-dimensional parametric model with parameter $\theta = (\theta_1, ..., \theta_p)$. Write the density evaluated at $Y = y$ in the form

$$f(y; \theta). \tag{8.33}$$

The *likelihood function* for θ based on data Y is just $f(Y; \theta)$ interpreted as a function of θ. Usually we work with the log likelihood

$$\ell_Y(\theta) = \log f(Y; \theta). \tag{8.34}$$

The *maximum likelihood estimator* (MLE) $\hat{\theta}$ is the value of θ which maximizes $\ell_Y(\theta)$. Usually we assume ℓ_Y is differentiable with a unique interior maximum, so the MLE is given by solving the *likelihood equations*

$$\frac{\partial \ell_Y}{\partial \theta_j} = 0, \qquad j = 1, ..., p. \tag{8.35}$$

The second-derivative or Hessian matrix of $-\ell_Y$, evaluated at $\hat{\theta}$, is called the *observed information matrix*

$$\mathcal{J} = \left[-\frac{\partial^2 \ell_Y}{\partial \theta_i \partial \theta_j}(\hat{\theta}), \qquad i, j = 1, ..., p \right]. \tag{8.36}$$

This is closely related, and asymptotically equivalent, to the *Fisher information matrix*, given by substituting the true value θ for the MLE $\hat{\theta}$ and taking expectations:

$$\mathcal{I} = \left[\mathrm{E}_Y \left\{ -\frac{\partial^2 \ell_Y}{\partial \theta_i \partial \theta_j}(\theta) \right\}, \qquad i, j = 1, ..., p \right]. \tag{8.37}$$

Note. If Y consists of n IID observations then (8.37) is exactly proportional to n, and could therefore be written $\mathcal{I}_n = n\mathcal{I}_1$ where $\mathcal{I}_1$ is the Fisher information based on one observation. In many statistics text books, $\mathcal{I}_1$ is taken as the definition of Fisher information. For the purposes of the present review, it is more convenient to ignore the dependence on n and work directly with (8.37).

The importance of $\mathcal{I}$ or $\mathcal{J}$ rests largely on the following fact: if the sample size n is large, then the distribution of $\hat{\theta}$ is approximately multivariate normal with mean θ and

covariance matrix given by either $\mathcal{I}^{-1}$ or $\mathcal{J}^{-1}$. In particular, the square roots of the diagonal entries of $\mathcal{I}^{-1}$ or $\mathcal{J}^{-1}$ are approximately the standard deviations of $\hat{\theta}_1, ..., \hat{\theta}_p$ and are therefore known as the *standard errors* of the parameter estimates, abbreviated $\mathrm{SE}(\hat{\theta}_j)$ for $j = 1, ..., p$.

Regarding the choice between $\mathcal{I}$ and $\mathcal{J}$, in most cases $\mathcal{J}$ is easier to use, requiring only numerical evaluation of the Hessian matrix, and not computation of an expected value as in (8.37). Moreover, $\mathcal{J}$ usually leads to more accurate SE estimates than $\mathcal{I}$. Therefore in most practical applications, the "information matrix" and standard errors will be assumed to be calculated from $\mathcal{J}$.

Suppose we want to test a hypopthesis $\theta_j = \theta_j^0$ for some given index j and a particular value θ_j^0, often 0. A test may be based on the t-statistics

$$t_j = \frac{\hat{\theta}_j - \theta_j^0}{\mathrm{SE}(\hat{\theta}_j)} \tag{8.38}$$

and the hypothesis rejected if $|t_j|$ is too large. For example, a common criterion is to reject the hypothesis if $|t_j| > 2$, which corresponds to a significance level of about .05.

A more sophisticated test is the *likelihood ratio test*: suppose we are comparing two models M_0 and M_1, where M_0 is nested in M_1, and the difference in dimensionality of the two models is q. This means, in effect, that M_0 is obtained from M_1 by imposing q constraints on the parameters of M_1. For instance, maybe M_1 has p parameters and M_0 corresponds to the hypothesis $\theta_{p-q+1}, ..., \theta_p = 0$. Let $\ell_Y^{(0)}(\theta)$, $\ell_Y^{(1)}(\theta)$ be the log likelihoods under the models M_0, M_1 and suppose the respective MLEs are $\hat{\theta}^{(0)}$, $\hat{\theta}^{(1)}$. Then

$$T = 2\{\ell_Y^{(1)}(\hat{\theta}^{(1)}) - \ell_Y^{(0)}(\hat{\theta}^{(0)})\} \tag{8.39}$$

is called the (log) *likelihood ratio statistic* (LRS). It is also known as the *deviance*. If M_0 is true then, approximately,

$$T \sim \chi_q^2, \tag{8.40}$$

the chi-squared distribution with q degrees of freedom. Thus we reject hypothesis M_0 at significance level α if T is bigger than the upper-α point of the χ_q^2 distribution.

Example. Suppose we want to test the null hypothesis of a Gumbel distribution (8.11) against the GEV alternative (8.10). Since the Gumbel arises from the GEV through the single constraint $\xi = 0$, the distribution of T in this case is χ_1^2. We would reject the Gumbel distribution at 5% significance level if $T > 3.84$. For a discussion of the LRS and some alternatives in this specific context, see Hosking (1984).

Some discussion needs to be given, both of how to put these methods into practice, and of the validity of the various approximations which have been mentioned.

For the GEV, the density $h(x; \mu, \psi, \xi)$ is obtained by differentiating (8.10) with respect to x. The likelihood based on observations $Y_1, ..., Y_N$ is

$$\prod_{i=1}^{N} h(Y_i; \mu, \psi, \xi) \tag{8.41}$$

and so the log likelihood is given by

$$\ell_Y(\mu, \psi, \xi) = -N \log \psi - \left(\frac{1}{\xi} + 1\right) \sum_i \log\left(1 + \xi \frac{Y_i - \mu}{\psi}\right) - \sum_i \left(1 + \xi \frac{Y_i - \mu}{\psi}\right)^{-1/\xi} \tag{8.42}$$

provided $1 + \xi(Y_i - \mu)/\psi > 0$ for each i; otherwise (8.42) is undefined (in effect, $-\infty$).

For the Poisson–GPD model, the likelihood function must be written in two parts, one corresponding to the Poisson component and the other to the GPD. If we observed N exceedances $Y_1, ..., Y_N$ over a T-year period, then the Poisson mean of N is λT, and the joint density of N and $Y_1, ..., Y_N$ is

$$\frac{(\lambda T)^N e^{-\lambda T}}{N!} \prod_{i=1}^{N} g(Y_i; \sigma, \xi) \tag{8.43}$$

where g is obtained from G by differentiating with respect to y in (8.16). Ignoring constants, we find

$$\ell_{N,Y}(\lambda, \sigma, \xi) = N \log \lambda - \lambda T - N \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{N} \log\left(1 + \xi \frac{Y_i}{\sigma}\right) \tag{8.44}$$

provided $1 + \xi Y_i/\sigma > 0$ for all i. Note that (8.44) can be separated into two components, one depending just on λ (with maximum at $\hat{\lambda} = N/T$) and the other just on (σ, ξ), so the maximization can be performed separately for each component. Note, however, that such a separation may not be possible if λ, σ and ξ are written as functions of some covariates, or if the likelihood is rewritten in terms of the GEV parameters (μ, ψ, ξ) through the formula (8.21). This is the reason for stating (8.44) in the form given.

For the maximization of $\ell_Y(\theta)$ for a general model indexed by θ, as mentioned in section 8.1 this may be performed using a packaged nonlinear optimization subroutine, of which several excellent versions are available. A few practical points follow:

1 Although the maximization is unconstrained, there are some practical constraints. For example, (8.42) requires $\psi > 0$ as well as $1 + \xi(Y_i - \mu)/\psi > 0$ for each i. It is advisable to test explicitly for such violations, and to set $-\ell_Y(\theta)$ equal to some very large value such as 10^{10} if the conditions are indeed violated.

2 Efficient use of nonlinear optimization routines requires that the problem be reasonably well scaled: roughly speaking, for each j both $\hat{\theta}_j$ and its standard error should be of the same order of magnitude as 1. In practice, using double precision arithmetic, there is usually little problem if the values are of the order $10^{\pm 3}$, but if the range is of order $10^{\pm 6}$, or worse, there is a problem.

3 Many nonlinear optimization routines require that first-order derivatives be supplied, as well as the values of the function itself. For some maximum likelihood estimation problems, analytic derivatives are easily computed, but for most, this is a computational challenge in itself. However if the problem is reasonably well scaled, it is nearly always possible to get by with approximations of the form

$$\frac{\partial \ell_Y}{\partial \theta_j} \approx \frac{\ell_Y(\theta + \epsilon e_j) - \ell_Y(\theta)}{\epsilon}, \tag{8.45}$$

where for example $\epsilon = 10^{-6}$. Here e_j is the unit vector in the direction of θ_j.

4 Quasi-Newton routines do not use the Hessian matrix directly, but an approximation which is itself updated as the algorithm proceeds. In such cases it is usually adequate to use this approximate Hessian matrix rather than directly evaluate second-order partial derivatives in (8.36). It may be desirable to rerun the algorithm from different starting points to check up on the adequacy of this approximation.

5 All Newton-type routines require the user to supply starting values, but the importance of "good" starting values can be overemphasized. Simple guesses usually suffice, e.g. in (8.42), one might set μ and ψ equal to the sample mean and sample standard deviation respectively, with ξ equal to some crude guess value such as 0.1. However it *is* important to check that the initial conditions are feasible and this can sometimes not be so easy to achieve!

6 In cases of doubt about whether a true maximum has been found, the algorithm may be re-run from different starting values. If the results are highly sensitive to starting values, this is indicative that the problem may have multiple local maxima, or alternatively that a mistake has been made in programming.

A few further comments are necessary regarding the specific application of numerical maximum likelihood estimation to the GEV and Poisson–GPD families. There is a singularity in the likelihood for $\xi < 0$, as $\mu \to Y_{\max}$ in (8.42) or as $\sigma \to -\xi Y_{\max}$ in (8.44). Here $Y_{\max} = \max(Y_1, ..., Y_N)$ and the effect is that $\ell_Y(\theta) \to \infty$. However in most practical cases, there is a local maximum of $\ell_Y(\theta)$ some distance from the singularity, and the presence of the singularity does not interfere with the convergence of the nonlinear optimization algorithm to the local maximum. In this case, the correct procedure is to ignore the singularity and use the local maximum. However, it is possible that no local maximum exists and the singularity dominates. In this case, maximum likelihood estimation fails

and some other method must be sought (Smith 1985). However, this very rarely happens with environmental data.

Another theoretical possibility is that there may be multiple local maxima. In the case of (8.42) or (8.44), such phenomena are certainly extremely rare, and it has been conjectured that they cannot occur at all. However in more complicated problems with many covariates, the possibility of multiple local maxima is real. In this case any quoted "maximum likelihood estimators" need to be treated with extreme caution.

Finally, we should say something about the theoretical status of the approximations involved. The asymptotic theory of maximum likelihood estimation for either the GEV or GPD models is valid provided $\xi > -\frac{1}{2}$ (Smith 1985). Cases with $\xi \leq -\frac{1}{2}$ correspond to an extremely short upper tail and hardly ever occur in environmental applications. A more serious problem is that even when $\xi > -\frac{1}{2}$, the asymptotic theory may give rather poor results with small sample sizes; as an example, see the simulations in Hosking *et al.* (1985). In these situations the rule has to be "caveat emptor": few users will have the time or energy to run their own simulations for every estimation problem they encounter, but they should be aware that asymptotic approximations always need to be treated with caution and especially when the sample sizes are small.

In summary: it is possible that MLEs will fail either numerically or in terms of their asymptotic properties, especially if sample sizes are small. The user should be aware of these possible difficulties but should not be deterred from using these extremely powerful and general methods.

Example. Let us return to the River Nidd example discussed briefly at the end of section 8.4. Ignoring both trend and seasonal factors, Davison and Smith (1990) fitted the Poisson-GPD model to all exceedances over threshold 100. The resulting parameters (standard errors in parentheses) were $\hat{\lambda} = 1.11$ (.18), $\hat{\sigma} = 51$ (14), $\hat{\xi} = 0.00$ (.21). Suppose we wish to find the N-year return level for the process. For known λ, σ, ξ, this is defined as the solution y of the equation $\lambda(1 - \xi y/\sigma)^{-1/\xi} = 1/N$. When λ, σ, ξ are unknown, we may proceed as follows. For each candidate value of y, maximize the likelihood under the constraint $\lambda(1 - \xi y/\sigma)^{-1/\xi} = 1/N$. The resulting expression is called the *profile likelihood* for y. By plotting the resulting expression against y, we can derive both a maximum likelihood estimate and a confidence interval for the return level. These curves for $N = 25$, 50 and 100 years are shown in Fig. 8.4. The dotted line here marks the level 1.92 below the maximum log likelihood. The value 1.92 is chosen because it corresponds to a likelihood ratio statistic of $2 \times 1.92 = 3.84$, which is the 5% upper tail point of the χ^2_1 distribution. Thus, all values for which the profile log likelihood is above the dotted line are within a 95% confidence interval for the N-year return level. For $N = 100$, for instance, this shows a maximum likelihood estimate at $y = 340$, with a 95% confidence interval of (260,925). The extreme asymmetry of this confidence interval reflects the inherent uncertainty of estimation about such extreme quantiles even when the Poisson/GPD model is assumed to hold exactly.

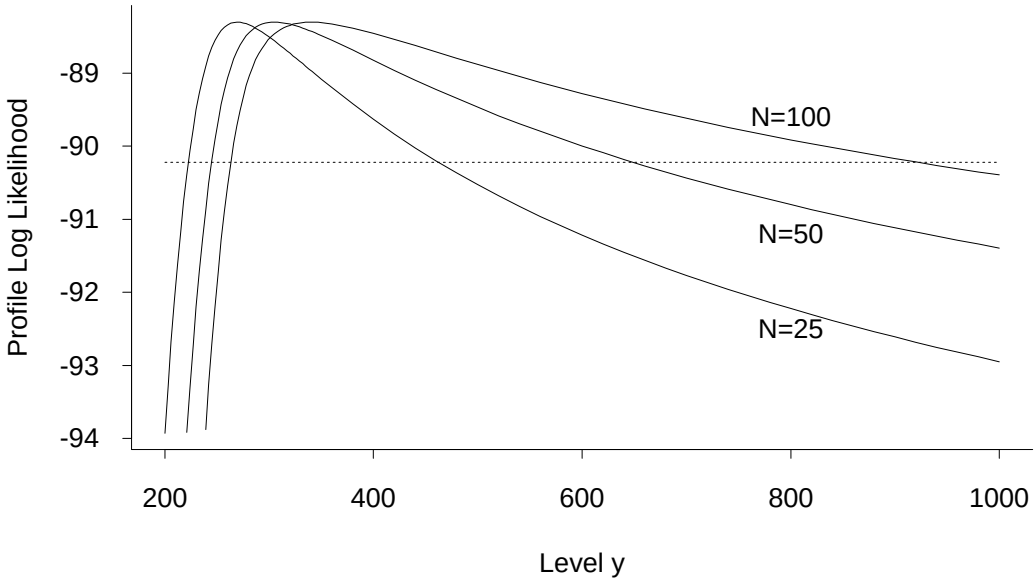


Fig. 8.4. Profile likelihood plots for the N-year return value for the Nidd data, for three values of N. Adapted from Davison and Smith (1990).

8.6. Bayesian methods

Bayesian methods also take the likelihood function as their starting point, but they also require the specification of a *prior density* $\pi(\theta)$ which, according to different interpretations, is supposed either to correspond to the user's subjectively determined initial information about θ, or else be chosen to minimize the amount of implicit information which it imparts. However the prior is specified, the *posterior density* is computed using the formula

$$\pi(\theta|Y) = \frac{\pi(\theta)f(Y;\theta)}{\int \pi(\theta')f(Y;\theta')d\theta'}. \tag{8.46}$$

Very often, we are interested in the *posterior mean* of a scalar quantity $g(\theta)$); for example, $g(\theta) = \theta_j$ for some j; in this case we write

$$\begin{aligned} \mathrm{E}\{g(\theta)|Y\} &= \int g(\theta)\pi(\theta|Y)d\theta \\ &= \frac{\int g(\theta)\pi(\theta)f(Y;\theta)d\theta}{\int \pi(\theta)f(Y;\theta)d\theta}. \end{aligned} \tag{8.47}$$

The principal obstacle to evaluating (8.46) or (8.47) is the need to calculate the integrals involved numerically. For many years, the practical application of Bayesian methods was hindered by the absence of powerful general-purpose algorithms for this, but in recent years

this situation has been transformed by the use of *Markov chain Monte Carlo* (MCMC) methods. The idea of such methods is to generate a random sample of values of θ, say $\theta^{(1)}, ..., \theta^{(M)}$, whose distribution approximates (8.46). Evaluation of posterior expectations, as in (8.47), is then easily carried out by summation.

There are a number of specific examples of MCMC algorithms, the two most important being

• *Gibbs sampling*: given the current value $\theta^{(m)} = (\theta_1^{(m)}, ..., \theta_p^{(m)})$, generate a new value $\theta_1^{(m+1)}$ from the conditional posterior distribution of θ_1, given $\theta_2 = \theta_2^{(m)}, ..., \theta_p = \theta_p^{(m)}$. Then generate a new value $\theta_2^{(m+1)}$ given $\theta_1 = \theta_1^{(m+1)}, \theta_3 = \theta_3^{(m)}, ..., \theta_p = \theta_p^{(m)}$, and so on up to $\theta_p^{(m+1)}$. This defines a new vector $\theta^{(m+1)} = (\theta_1^{(m+1)}, ..., \theta_p^{(m+1)})$ and completes one iteration of the algorithm. The algorithm is iterated many times until the Markov chain $\{\theta^{(m)},\ m = 1, 2, ...,\}$ has apprently converged to its stationary distribution.

• *The Hastings-Metropolis algorithm*: Given $\theta^{(m)}$, generate a *trial value* θ^* from some transition density $q(\theta^*|\theta^{(m)})$. Then perform a second randomization, whereby θ^* is "accepted" with probability

$$\min\left\{1, \frac{\pi(\theta^*|Y)q(\theta^{(m)}|\theta^*)}{\pi(\theta^{(m)}|Y)q(\theta^*|\theta^{(m)})}\right\}; \tag{8.48}$$

otherwise, θ^* is "rejected". If θ^* is accepted then we set $\theta^{(m+1)} = \theta^*$, otherwise $\theta^{(m+1)} = \theta^{(m)}$. Either way, the algorithm is continued from the new $\theta^{(m+1)}$.

In practice, a very common method is to combine the Gibbs sampler and the Hastings-Metropolis algorithm: update one component at a time, as in the Gibbs sampler, but to perform this updating, take one Hastings-Metropolis step. The latter is often taken to be of "random walk" form: for each component j, $\theta_j^* - \theta_j^{(m)}$ are taken IID from some density f_j, which may itself be taken normal with mean and standard deviation s_j. The principal question then is how to choose s_j; it is usually not bad if this value is of the same order of magnitude as the posterior standard deviation of θ_j, which may be chosen by trial and error or updated as the algorithm proceeds.

Space does not permit us to give more than a brief outline of these methods, but they are becoming increasing widely used in statistical practice, especially for high-dimensional parameter spaces, where they may help to overcome some of the difficulties in applying maximum likelihood estimation in high-dimensional situations.

Another situation where Bayesian methods are particularly powerful is in connection with *predictive inference*. Suppose our real interest is not in estimating or testing a hypothesis about a parameter θ, but in predicting a future observation that depends on θ. For example, we may want to ask the question "what is the probability that the city of Miami will experience a wind speed of greater than 90 MPH in the next 10 years?". Such a question may be formulated in terms of the distribution of $Z = \max(Y_1, ..., Y_{10})$ where

Y_1, Y_2, ..., are annual maxima, whose distribution may be taken as GEV with parameters esitmated from past data. The required probability is then

$$\Pr\{Z > z; \mu, \psi, \xi\} = 1 - \exp\left\{-10\left(1 + \xi\frac{z-\mu}{\psi}\right)^{-1/\xi}\right\} \tag{8.49}$$

with $z = 90$. However $\theta = (\mu, \psi, \xi)$ is unknown and simply substituting the MLE $\hat{\theta}$ is not considered a good thing to do because it ignores the uncertainty in estimating θ. A Bayesian solution is based on the posterior predictive distribution

$$\hat{\Pr}\{Z > z\} = \int \Pr\{Z > z; \mu, \psi, \xi\}\pi(\mu, \psi, \xi|Y)d\mu d\psi d\xi \tag{8.50}$$

where $\pi(...|Y)$ denotes the posterior density given past data Y; this may be evaluated similarly to (8.47), by summing over a MCMC sample.

The Bayesian predictive approach has been examined in a number of recent papers, e.g. Coles and Powell (1996), Coles and Tawn (1996a), Smith (1997a). In particular, Coles and Tawn (1996a) gave an elegant discussion of eliciting a prior based on expert judgement in this context. Despite a number of attractive features of this approach, the theoretical properties of such procedures still leave a number of questions to be answered (Smith 1997b).

Example. Fig. 8.5 shows the five best running times by different athletes in the women's 3000 metre track event for each year from 1972 to 1992. Also shown on the plot is the remarkable new world record achieved by the Chinese athlete Wang Junxia in 1993, some 16 seconds faster than the previous record. Many questions were raised about this, including the possibility that Wang might have been taking illegal drugs.

Robinson and Tawn (1995) proposed a statistical test for this, based on fitting an extreme value distribution to the $r = 5$ smallest observations in each year, using the probability model outlined in section 8.4.1 but in the form appropriate for minima rather than maxima. They defined a parameter x_{ult}, in effect the long-term lower boundary point of the distribution, to represent the best possible performance, and obtained confidence intervals for x_{ult}, based on all data up to 1992, using a number of different models. As an example, if we were using the GEV distribution and talking about upper instead of lower extremes, x_{ult} would be given by $\mu - \psi/\xi$ if $\xi < 0$ and $+\infty$ if $\xi \geq 0$, though for this kind of data set it is clear that $\xi < 0$ and so x_{ult} is indeed finite.

To calculate these confidence intervals, Robinson and Tawn used a likelihood ratio testing procedure. Suppose we want to test the hypothesis $x_{ult} = x^*$, for some given x^*, against the alternative $x_{ult} \neq x^*$. This can be carried out using a likelihood ratio testing procedure, as described in section 8.5. A 95% confidence interval for x_{ult} consists of all values of x^* for which this test leads to acceptance of the hypothesis $x_{ult} = x^*$ at significance level .05. The procedures considered by Robinson and Tawn took account of

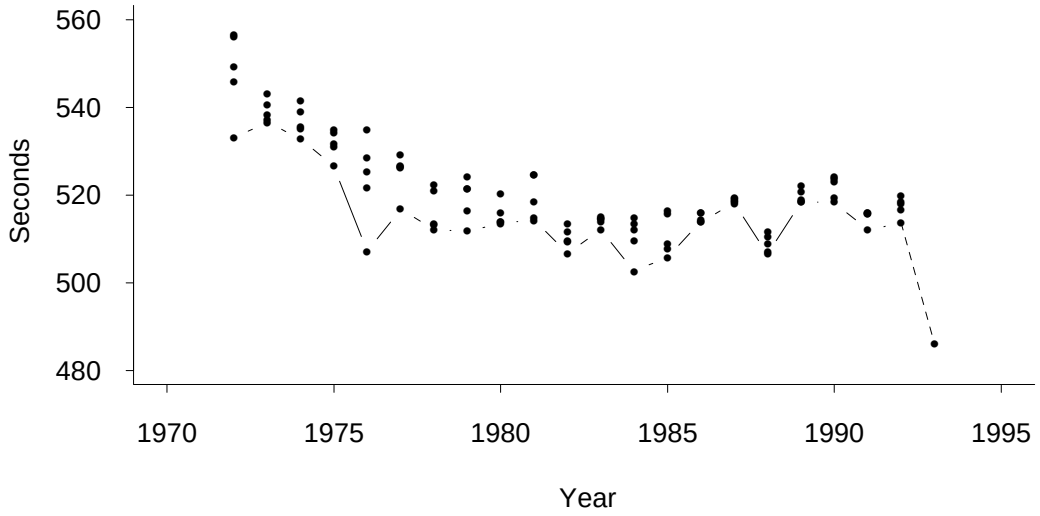


Fig. 8.5. Five best performances per year for the women's 3000 metre event, together with Wang Junxia's performance from 1993.

the trend in the data in various ways, but Smith (1997a) performed this procedure for a simplified model in which the trend is ignored but so are all data prior to 1980. When constructed in this way, Smith's 95% confidence interval for x_{ult} was (481.9, 502.43). This is to be compared with the existing record which stood at 502.62 in 1992. Fig. 8.6 shows a deviance for x_{ult} (similar to Fig. 8.4), together with a corresponding plot for the women's 1500 metre performances (Smith 1997a). Wang's new record was 486.11 and therefore inside this confidence interval. The same is true for the other confidence intervals contructed by Robinson and Tawn. Thus although the analysis strengthens the conclusion that Wang's record was indeed very unusual and perhaps suspicious, it does not provide conclusive evidence that the record was inconsistent with previous data.

Smith (1997a) argued that a superior approach was based on the *predictive distribution* of Z, here defined to be the best performance for 1993, given all observations up to the end of 1992. This may be computed in similar fashion to (8.50). In effect, computing a confidence interval is asking the question "what will be the best performace ever achieved?" (in the indefinite future) while computing a prediction interval is focussing on the much more realistic question of what might happen in a specific year.

The result of this calculation is dramatic. Calculating the Bayesian predictive distribution based on an uninformative prior distribution and using data from 1980 to 1992, the predictive conditional probability of achieving Wang's record given that a new record

is set — in symbols, $\hat{\Pr}\{Z < 486.11 | Z < 502.62\}$ — is about .0006. The fact that this estimated probability is so small creates clear doubts about the authenticity of the record.

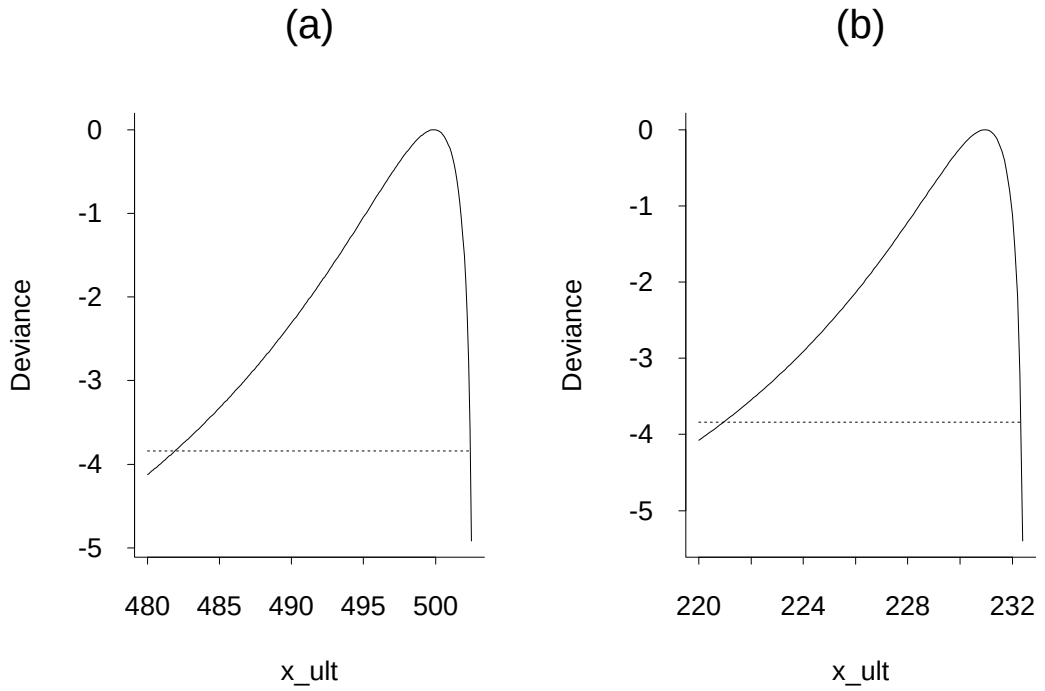


Fig. 8.6. Deviance plot for the "best possible performance" x_{ult}, based on 1972–1992 data. (a) Women's 3000 metre data. (b) Women's 1500 metre data.

8.7. Other methods of estimation

Other estimation techniques have been proposed, and at times strongly advocated, but none has the power and generality of either maximum likelihood or Bayesian methods.

An example is the theory of *probability weighted moments* (PWMs — Hosking *et al.* 1985) and its subsequent development into the theory of *L-moments* (Hosking 1990). These methods are supported by simulations which show, for example, that under certain circumstances, with small sample sizes, the PWM approach to the generalized extreme value distribution produces more efficient estimates of return levels, as measured by mean squared error, than the maximum likelihood approach. I do not dispute the correctness of their simulations, but as a contribution to the development of statistical methods for meteorological and hydrological data, I believe the whole debate over PWM and L-moment methods has been a distraction from far more important issues. Some specific points are

• Often it is not the mean squared error of point estimates which is most relevant to the practical user, but how well the approximate coverage probability of an interval estimate, or nominal P-value of a test, approximates the true coverage probability or P-value. Simulations have shown that likelihood ratio methods (section 8.5) perform very well when assessed from this point of view, in spite of certain theoretical difficulties.

• The PWM methods require certain restrictions on the parameter values which are not present for the method of maximum likelihood. When the maximum likelihood methods are re-evaluated so as to incorporate similar restrictions, the comparative advantage of the PWM method disappears (Coles and Dixon 1997).

• The real potential for improving on standard extreme value techniques comes not in finding estimates which improve slightly on existing ones, but in generalizing the methods to handle richer sources of data. Examples include taking suitable covariates into account, combining data from different series, and incorporating physical information such as that generated by atmospheric and ocean circulation models. Maximum likelihood and Bayesian methods are very general techniques which may often be applied in a routine way to such problems, whereas specialized techniques such as PWMs are limited to the context for which they have been derived.

8.8. Regression models

A key theme of modern statistics is building regression models to show how the variable of interest may depend on other measured covariates. In the present section we illustrate, with examples, two possible ways of bringing such ideas into extreme value analysis.

8.8.1. Ozone exceedances

The (U.S.) Clear Air Act of 1970 required the Environmental Protection Agency (EPA) to establish and periodically revise air quality standards which are "requisite to protect the public health (with) an adequate margin of safety". Ground-level ozone, unlike that in the upper atmosphere, is a human health hazard and has always been one of the pollutants regulated, through the National Ambient Air Quality Standard (NAAQS). The current standard is based on the number of exceedances of daily maximum ozone over the level of 12 parts per hundred million (pphm) [4].

As part of the process of enforcing and reporting on air pollution standards, the national EPA and various state environmental agencies track trends in the levels of pollutants through a rich network of monitors. Since the standard is enforced in terms of exceedances over a threshold, it is natural to ask questions about trends in air pollution levels in the

[4] The revised standard introduced in 1997 reduced the standard level to 8 pphm, but based on the maximum 8-hour average over the course of any 24-hour period, rather than the maximum 1-hour level as in the pre-1997 standard. There were also changes in determining how many exceedances of that level were permitted before the standard is considered to be in violation. At the time of writing, a Court of Appeals ruling has set aside this standard because of various legal and constitutional challenges. Although it seems likely that some form of standard based on 8-hour averages will eventually be approved, the statistical issues discussed in this section are equally valid with the new standard as with the old.

same terms. In other words, are exceedances of the threshold becoming more frequent, or (as we hope) less frequent? This naturally brings us within the realm of extreme value theory.

However, there is more to this than simply counting exceedances of a threshold. Ozone is formed in the atmosphere through a complex sequence of photochemical processes which begin with emissions from vehicle exhausts and from power plants, but which are also strongly affected by meteorology, being most prevalent in hot, still weather in the middle of summer. Changes in the frequency of ozone exceedances could be due to patterns of unusually hot or cool summers as much as to real changes in emissions of ozone-forming chemicals. Although the ozone standard itself takes no account of meteorological conditions, in judging the success of ozone-reduction programs, it is important to take meteorological influences into account. This suggests forming regression models for the frequency of high-level ozone exceedances in which both time and meteorology are regressors.

Smith and Shively (1995) constructed such a model for data from a monitoring station in Houston, Texas, over the period 1983–1992. The analysis generalized earlier analyses due to Smith (1989), who considered trends in ozone exceedances but without taking meteorology into account, and Shively (1991), who modeled the point processes of exceedances of a single level as a function of time and meteorology, but without taking the actual levels of the process into account — in other words, just recording a binary processes according to whether the standard was exceeded or not without taking into account the amount by which the standard was exceeded.

The analysis of Smith and Shively was based on the following key ingredients:

1. The probability of an exceedance of a given level u on day t was represented as $e^{\alpha(t)}$, where

$$\alpha(t) = \alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t), \tag{8.51}$$

where $s(t)$ is the calendar year in which day t falls and $\{w_j(t),\ j = 2, ..., p\}$ are the values of $p - 1$ weather variables on day t.

2. Given that there is an exceedance of the level u on day t, the probability that it exceeds the level $u + x$, where $x > 0$, is represented by the equation

$$\{1 + \xi\beta(t)x\}^{-1/\xi},$$

where

$$\beta(t) = \beta_0 + \beta_1 s(t) + \sum_{j=2}^{p} \beta_j w_j(t). \tag{8.52}$$

Thus the excess values are represent by a GPD, with $\beta(t)$ the reciprocal of the usual GPD scale parameter. (Alternative formulations would allow either $1/\beta(t)$ or $\log \beta(t)$ to be a linear function of covariates, but the form (8.52) appeared to work well in this instance.)

The meteorological variables considered in this analysis were as follows:

TMAX. Maximum hourly temperature between 6am and 6pm.

TRANGE. Difference between maximum and minimum temperature between 6am and 6pm. This is considered to be a proxy for the amount of sunlight.

WSAVG. Average wind speed from 6am to 6pm. Higher windspeeds lead to lower ozone levels because of more rapid dispersion of ozone precursors.

WSRANGE. Difference between maximum and minimum hourly windspeeds between 6am and 6pm.

NW/NE. Percentage of time between 6am and 6pm that the wind direction was between NW and NE.

NE/ESE. Percentage of time between 6am and 6pm that the wind direction was between NE and ESE.

ESE/SSW. Percentage of time between 6am and 6pm that the wind direction was between ESE and SSW.

SSW/NW. Percentage of time between 6am and 6pm that the wind direction was between SSW and NW.

The wind directions are important for Houston because they determine the level of industrial pollution — for instance, there is a lot of industry to the south of Houston, and ozone levels tend to be higher when the wind direction is in the ESE/SSW sector.

In most analyses, the variable SSW/NW was omitted because of the obvious collinearity with the other wind directions.

Standard variable selection procedures were used to determine which variables to include. Note that although the notation in (8.51) and (8.52) implicitly assumes that the same meteorogical variables are included, in practice a separate variable selection is performed for the two equations.

Variable	Coefficient	Standard Error
$s(t)$	–0.149	0.034
TRANGE	0.072	0.016
WSAVG	–0.926	0.080
WSRANGE	0.223	0.051
NW/NE	–0.850	0.408
NE/ESE	1.432	0.398

Table 8.1. Coefficient and standard errors for $\alpha(t)$.

Variable	Coefficient	Standard Error
$s(t)$	0.035	0.011
TRANGE	–0.016	0.005
WSAVG	0.102	0.019
NW/NE	0.400	0.018

Table 8.2. Coefficient and standard errors for $\beta(t)$.

Results of these analyses are shown in Tables 8.1 and 8.2. The model for $\beta(t)$ was based on $\xi = 0$, i.e. the exponential distribution for excess values, but when the same model was repeated assuming $\xi \neq 0$, the result was not statistically significant ($\hat{\xi} = -0.054$ with an standard error of 0.063). Note that where the same variable occurs in both equations, the coefficients are of opposite signs, meaning that those variables which tend to increase $\alpha(t)$ (meaning an increase in the frequency of crossing the threshold) also tend to decrease $\beta(t)$ (which means an increase in the mean excess). In particular, the coefficient of $s(t)$ is negative in $\alpha(t)$ and positive in $\beta(t)$, meaning an overall downward trend with time in the frequency of ozone exceedances and in the mean excess.

As a comparison, Smith and Shively also fitted the same model for trend for trend alone, ignoring meteorological covariates. In this case, the estimated coefficients of $s(t)$ were –0.069 (standard error 0.030) in $\alpha(t)$ and 0.018 (standard error 0.011) in $\beta(t)$. The coefficients are thus much smaller in magnitude if the model is fitted without any meteorology. This confirms the significance of the meteorological component and shows how the failure to take it into account might obscure the real trend.

8.8.2. Wind speeds in North Carolina

An alternative approach is to assume that the model is represented in terms of the GEV parameters (μ, ψ, ξ), either through the point process representation of section 8.4.2, or directly using (8.21) to relate the Poisson-GPD parameters to the equivalent GEV parameters.

We can now extend this as follows: Rewrite (μ, ψ, ξ) as (μ_t, ψ_t, ξ_t) to emphasize the dependence on time t. A typical model is of the form

$$\mu_t = \sum_{j=0}^{q} \beta_j x_{jt}, \qquad \log \psi_t = \sum_{j=0}^{q} \gamma_j x_{jt}, \qquad \xi_t = \sum_{j=0}^{q} \delta_j x_{jt}, \tag{8.53}$$

in terms of covariates $\{x_{jt},\ j = 0, ..., q\}$ where we usually assume $x_{0t} \equiv 1$.

In practice we would probably not adopt the full model (8.53). For example, one simplification is to assume that μ_t varies in the manner described, with $\psi_t \equiv \psi$, $\xi_t \equiv \xi$ constants.

The likelihood for this model may essentially be obtained by rewriting (8.44) as a sum over t. Suppose λ_t, σ_t are derived from (μ_t, ψ_t, ξ_t) using (8.21). Then the log likelihood becomes

$$\sum_t \left\{ \eta_t \log \lambda_t - \frac{\lambda_t}{T_0} - \eta_t \log \sigma_t - \eta_t \left(1 + \frac{1}{\xi_t}\right) \log \left(1 + \xi_t \frac{Y_t}{\sigma_t}\right) \right\} \tag{8.54}$$

where

$\eta_t = 1$ if there is an exceedance on day t, 0 otherwise,

T_0 is a time scaling constant. For example, if the unit of time t is one day and "annual maxima" are defined in the usual way, then $T_0 = 365.25$.

Y_t is the excess over the threshold on day t assuming there is an exceedance (if $\eta_t = 0$ it does not matter how Y_t is defined)

The derivation of (8.54) follows from the point process theory of extreme values, and is given in detail by Smith (1989) and Smith and Shively (1995). In the remainder of this section we present another example, based on the windspeed data from section 8.1.

Consider the data for Raleigh. For a reason to be explained later (section 8.9.3), a threshold 39.5 was chosen for this data set. Fitting the model with no trend produces the parameter estimates and standard errors in Table 8.3.

Parameter	Estimate	Standard Error
μ	42.4	0.9
$\log \psi$	1.49	0.16
ξ	–0.19	0.19

Table 8.3. Parameter estimates and standard errors for Raleigh data set over threshold 39.5; homogeneous model with no covariates.

Perhaps the most important conclusion from these estimates is that although the value of ξ comes out negative, indicating a short-tailed distribution, the standard error shows that the difference with $\xi = 0$ is not statistically significant, so we can't be sure the distribution is short-tailed! The data set does not extend as far as October 1996, when Hurricane Fran produced a maximum wind speed of 79 mph in Raleigh.

The above estimates were produced by maximizing the likelihood function. If we define the neg log likelihood NLLH=–log L, where L is the maximized likelihood, then NLLH=114.833.

If we insert a linear trend in μ_t, for example by writing $\mu_t = \alpha + \beta t$, then NLLH changes only very slightly, to 114.486. Twice the difference in NLLH is the likelihood ratio statistic and has an approximate χ^2_k distribution under the null hypothesis that the simpler model is correct, where k is the difference in the number of parameters between the two models. In this case $k = 1$ and the deviance is only 0.69, which is certainly not significant as a χ^2_1 random variable.

On the other hand, where we can definitely expect to see a significant effect is in looking for seasonal variation. This was modelled by adding terms in $\cos(2\pi t/365)$ and $\sin(2\pi t/365)$ to μ_t, with the results in Table 8.4.

Parameter	Estimate	Standard Error
μ	40.9	0.9
β_1	0.90	1.07
β_2	5.29	1.39
$\log \psi$	1.43	0.13
ξ	–0.12	0.10

Table 8.4. Parameter estimates and standard errors for Raleigh data set over threshold 39.5; model including sinusoidal seasonal term in μ.

Here, β_1 and β_2 are the additional coefficients. In this case NLLH=103.106, for a deviance of 23.6 against the model with no trend, and this is certainly significant as a χ^2_2 variable. Note also a slight, though not statistically significant, change in the value of ξ.

City	Model	ξ	S.E.	NLLH
Greensboro	0	.29	.20	132.448
Greensboro	1	.32	.19	127.200
Greensboro	2	.32	.19	126.922
Charlotte	0	.16	.12	162.190
Charlotte	1	.18	.11	152.775
Charlotte	2	.14	.10	149.340

Table 8.5. Comparison of models for Greensboro and Charlotte.

Table 8.5 shows corresponding results for Greensboro and Charlotte. In this case "Model 0" refers to the model with no seasonal component, while Models 1 and 2 add, respectively, a twelve-month and six-month sinusoidal cycle to μ_t. Since our greatest interest may well be in ξ, the values of this parameter and its standard error are tabulated for each model, as well as the NLLH. Some other models were tried, such as adding seasonal variation to log ψ instead of μ, but these did not produce significant effects.

In both cases the seasonal effect is significant as measured by the deviance between models 0 and 1. For Greensboro, as for Raleigh, there is no significant improvement in passing to model 2. In the case of Charlotte there is an improvement – the LRS is 6.87 with two degrees of freedom, corresponding to a P-value of 0.03 based on the chi-squared test.

Fig. 8.7 shows QQ plots of the residuals. These are based on taking all exceedances above the threshold, transforming to an exponential distribution based on the fitted model, and then plotting the order statistics against their expected values under the assumption that the exponential distribution is the correct model. If this plot stays close to the straight line through the origin with unit slope, then we can conclude that the GPD is a good fit to the data.

The first three plots are all based on Model 1, and show a very good agreement between the observed and expected values. As a comparison, the bottom right hand plot is computed for Charlotte under the original model (Model 0). In this case there is some evidence that the two largest observations are outliers, in the sense that both are substantially above the fitted straight line, reinforcing what we already saw in the mean excess plot. However, the effect almost disappears under Model 1. This neither proves nor disproves the notion that the two largest values are outliers, but it does show that there is at least the possibility of accommodating hurricanes with the rest of the data.

8.9. Testing the fit

An important component of any model identification and estimation procedure is to test whether the resulting model fits the data. Techniques available for this range from simple graphical methods to sophisticated goodness of fit testing. This section consists of a brief review of the methods available, concentrating on those which have been developed specifically for the GEV and GPD models.

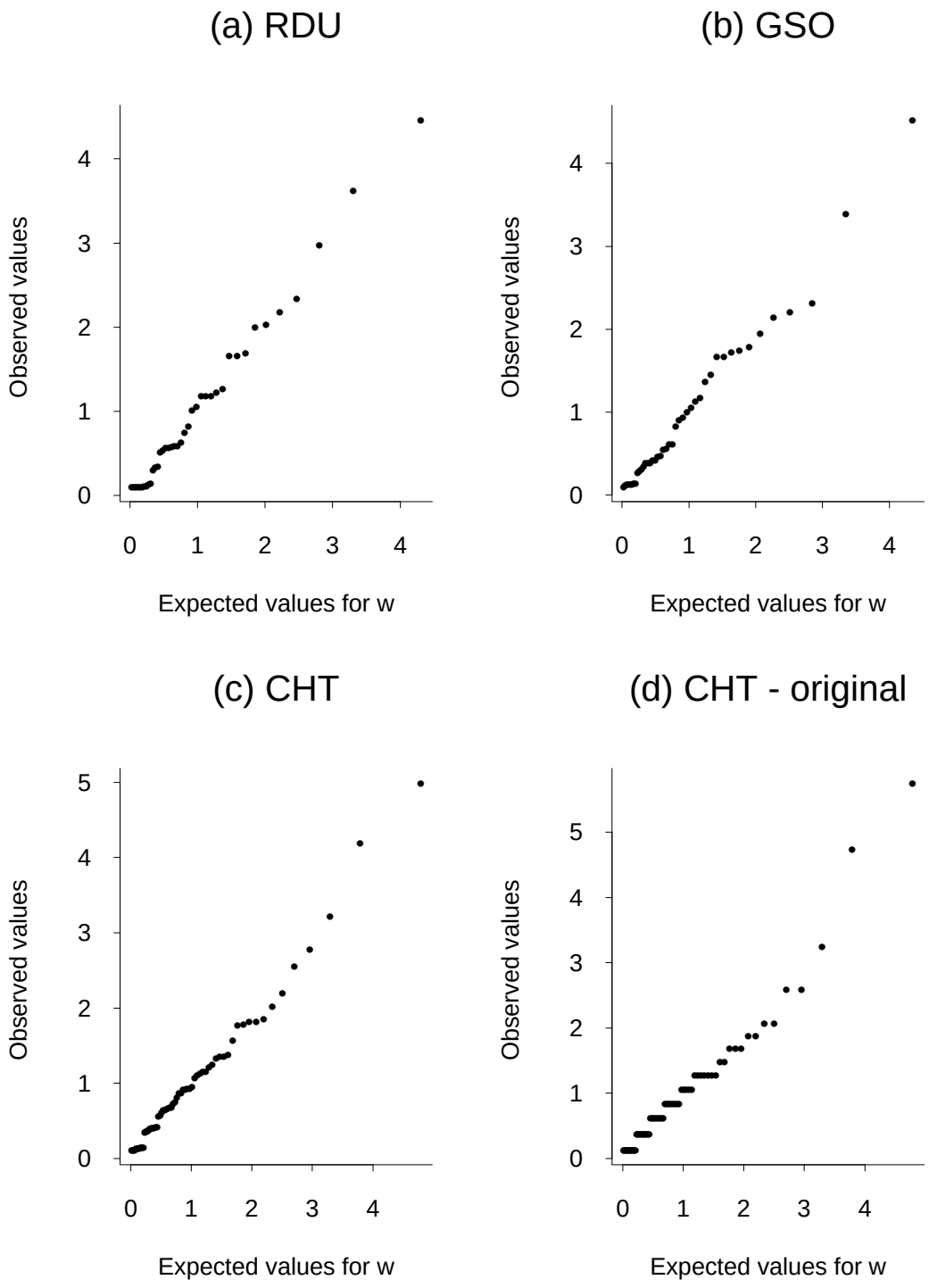


Fig. 8.7. QQ plots of residuals for each of the Raleigh (RDU), Greensboro (GSO) and Charlotte (CHT) data sets. Each is based on Model 1 of the text, which includes seasonal variation. Plot (d) is for Charlotte using Model 0, which is non-seasonal. The largest two residuals do not fit so well in this instance.

8.9.1 Gumbel plots

Suppose $Y_1, ..., Y_N$ represent a sample of size N from the Gumbel distribution (8.11), ordered as $Y_{1:N} \leq ... \leq Y_{N:N}$. The Gumbel probability plot consists of plotting $Y_{i:N}$ against the *reduced value* $x_{i:N}$, defined as

$$x_{i:N} = -\log(-\log p_{i:N}), \tag{8.55}$$

$p_{i:N}$ being the i'th *plotting position*, usually taken to be $(i - \frac{1}{2})/N$.

If the plot looks close to a straight line, then the Gumbel distribution may be presumed a good fit, and in the days before more sophisticated methods such as maximum likelihood became widely available, the intercept and slope of that line were used to define estimates of μ and ψ.

If the plot is not a straight line, for example is there appears to be a systematic curvature up or down, this is often taken as an indication that the data need to be fitted by a three-parameter GEV distribution, with $\xi \neq 0$, rather than the Gumbel distribution. However the plot can also be used to detect other features, such as gross outliers which cannot easily be fitted by any distribution. The appealing thing about this plot is that it can be drawn right away with the raw data, without any preliminary estimation of parameters. However, for the GEV distribution and the GPD, not to mention more complicated situations such as those with covariates, such a simple plotting technique is usually not available.

Example. Fig. 8.8 shows Gumbel plots for two data sets considered by Smith (1990). Plot (a) is based on 35 annual maxima for the River Nidd series discussed in section 8.5. The plot is straight as far as can be seen by eye, suggesting that the Gumbel distribution might be a reasonable fit. Plot (b) is based on annual maximum temperatures at Ivigtut, in Iceland. A characteristic feature of this plot is the influence of the largest order statistic, which seems to be an outlier compared with the other annual maxima. However, when this observation is ignored, the remainder of the plot is clearly curving downwards, suggesting a short-tailed distribution with $\xi < 0$.

8.9.2 QQ plots of residuals

A second type of probability plot is drawn *after* fitting the model, and this is a general technique, not restricted to particular models as is the case with the Gumbel plot. Suppose $Y_1, ..., Y_N$ are IID observations whose common distribution function is $G(y; \theta)$ depending on parameter vector θ. Suppose θ has been estimated by $\hat{\theta}$, and let $G^{-1}(p; \theta)$ denote the inverse distribution function of G, written as a function of θ. A QQ (quantile-quantile) plot consists of first ordering the observations $Y_{1:N} \leq ... \leq Y_{N:N}$, and then plotting $Y_{i:N}$ against the reduced value

$$x_{i:N} = G^{-1}(p_{i:N}; \hat{\theta}), \tag{8.56}$$

where $p_{i:N}$ may be taken as $(i - \frac{1}{2})/N$ as in section 8.9.1. If the model is a good fit, the plot should be roughly a straight line of unit slope through the origin.

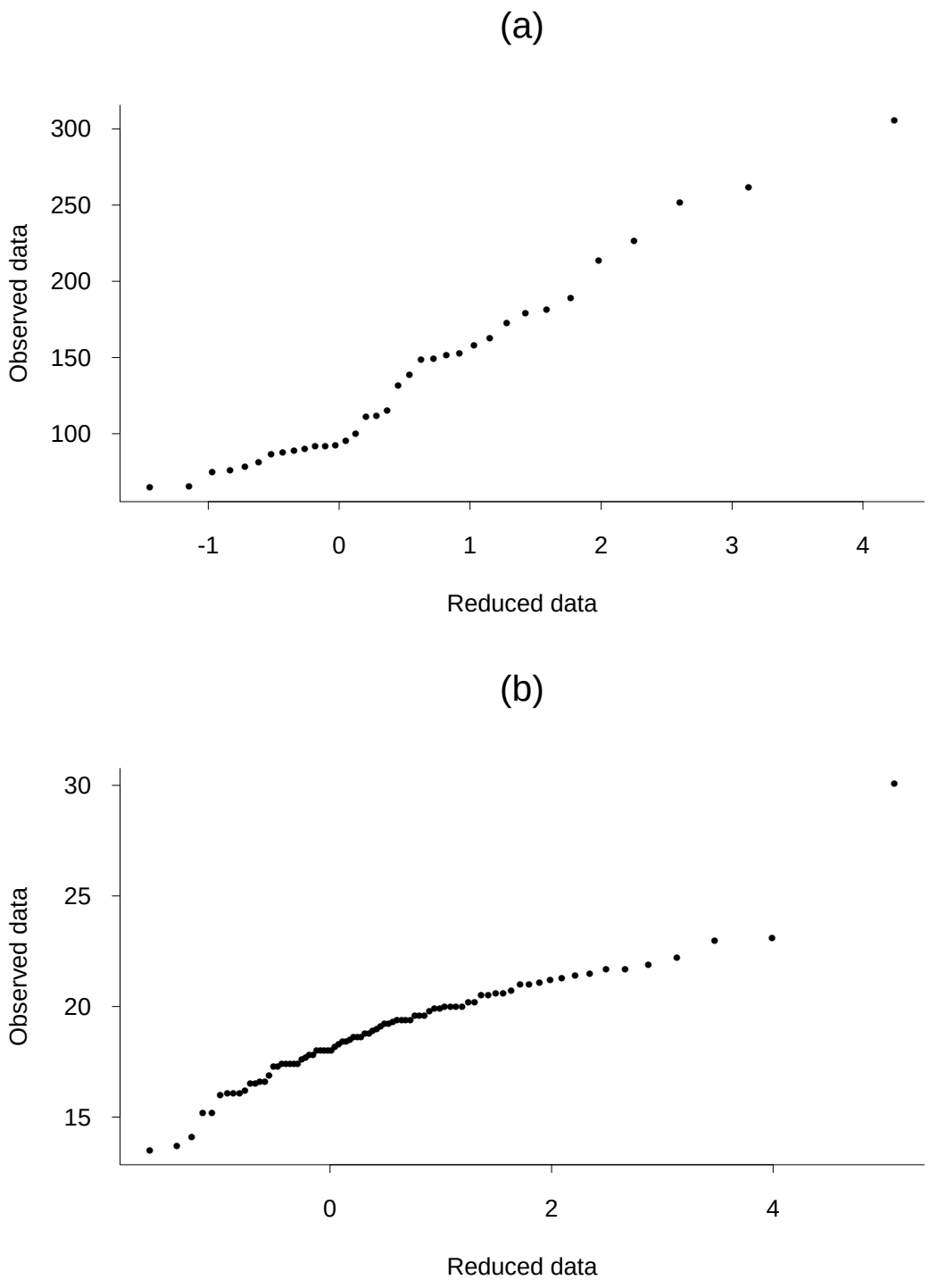


Fig. 8.8. Illustration of Gumbel plots. (a) Based on annual maxima for River Nidd flow series. (b) Based on annual maximum temperatures in Ivigtut, Iceland. Based on Smith (1990).

For example, if $G(y, \theta)$ is the same as $H(y; \mu, \psi, \xi)$ as in (8.10), then the inverse H^{-1} is obtained by solving the equation $H(y; \mu, \psi, \xi) = p$ as a function of y for fixed p; the solution is

$$y = H^{-1}(p; \mu, \psi, \xi) = \mu + \frac{\psi}{\xi}\{(-\log p)^{-\xi} - 1\}. \tag{8.57}$$

Thus the probability plot consists of plotting $Y_{i:N}$ against $H^{-1}(p_{i:N}; \hat{\mu}, \hat{\psi}, \hat{\xi})$ with $\hat{\mu}, \hat{\psi}, \hat{\xi}$ the maximum likelihood estimators.

Similar calculations are easily made for the GPD; Davison and Smith (1990) have several examples.

Residual plotting is best regarded as an informal guide to the fit of the model, rather than a definitive test. Nevertheless such plots form a valuable check on whether the model being fitted is sensible.

Example

Fig. 8.9 shows residual from a GEV model fitted to the Ivigtut data of Fig. 8.8(b). Plot (a) shows the plot done directly. The influence of the outlier is still very clear. Plot (b) was computed as follows: the GEV model was fitted with the outlier removed, but the outlier was included for the purpose of fitting the plot. In other words, the distinction between (a) and (b) is not in the observed values plotted, but in the GEV parameters which define the expected values. The plot in (b) follows a straight line very closely except for the outlier. This shows that if the outlier is deleted, the GEV fits the remaining observations very well, whereas if it is not deleted, it clearly distorts the model fit.

Fig. 8.10 is a similar kind of plot shown after fitting a GPD to high-level exceedances of the River Nidd. Plot (a) is based on threshold $u = 70$, and shows a clear discrepancy in the values of the largest two order statistics. Plot (b) is based on $u = 100$, and shows no such discrepancy. The interpretation is that the model fits better to a threshold $u = 70$ than it does to $u = 100$.

Fig. 8.11 shows in more dramatic fashion the possibilities of a QQ plot to detect outliers. Plot (a) is the time plot of insurance claims incurred by a large company over a 15-year period. Noting that costs are plotted on a logarithmic scale, it can be seen that the data series is dominated by a small number of very large claims, and in fact GPD fits to the data result in estimated ξ values very close to 1 — in other words, an extremely long-tailed distribution. However the QQ plot in (b) shows nothing out of the ordinary — it appears that the very large claims are consistent with the overall data set. In constrast, plots (c) and (d) show a true outlier arising from an analysis of costs of large oil spill. The largest cost arises from the Exxon Valdez disaster of 1989. It is clear that this value is not consistent with an extreme value model.

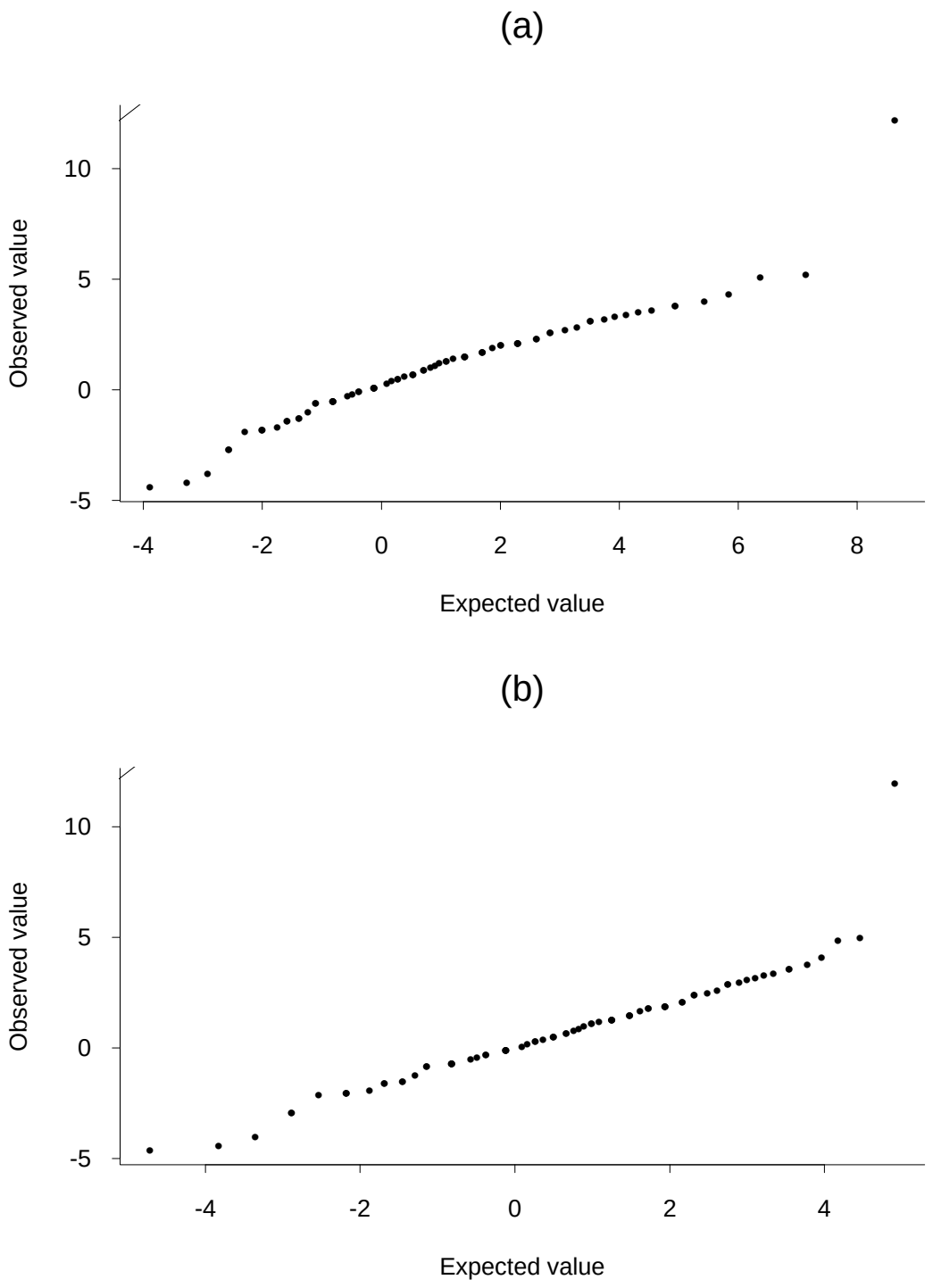


Fig. 8.9. QQ plots of residuals based on GEV model fitted to Ivigtut annual maxima. (a) Based on GEV model fitted to the whole data set. (b) Based on GEV model fitted to data with largest value removed, but the largest value is included in the plot.

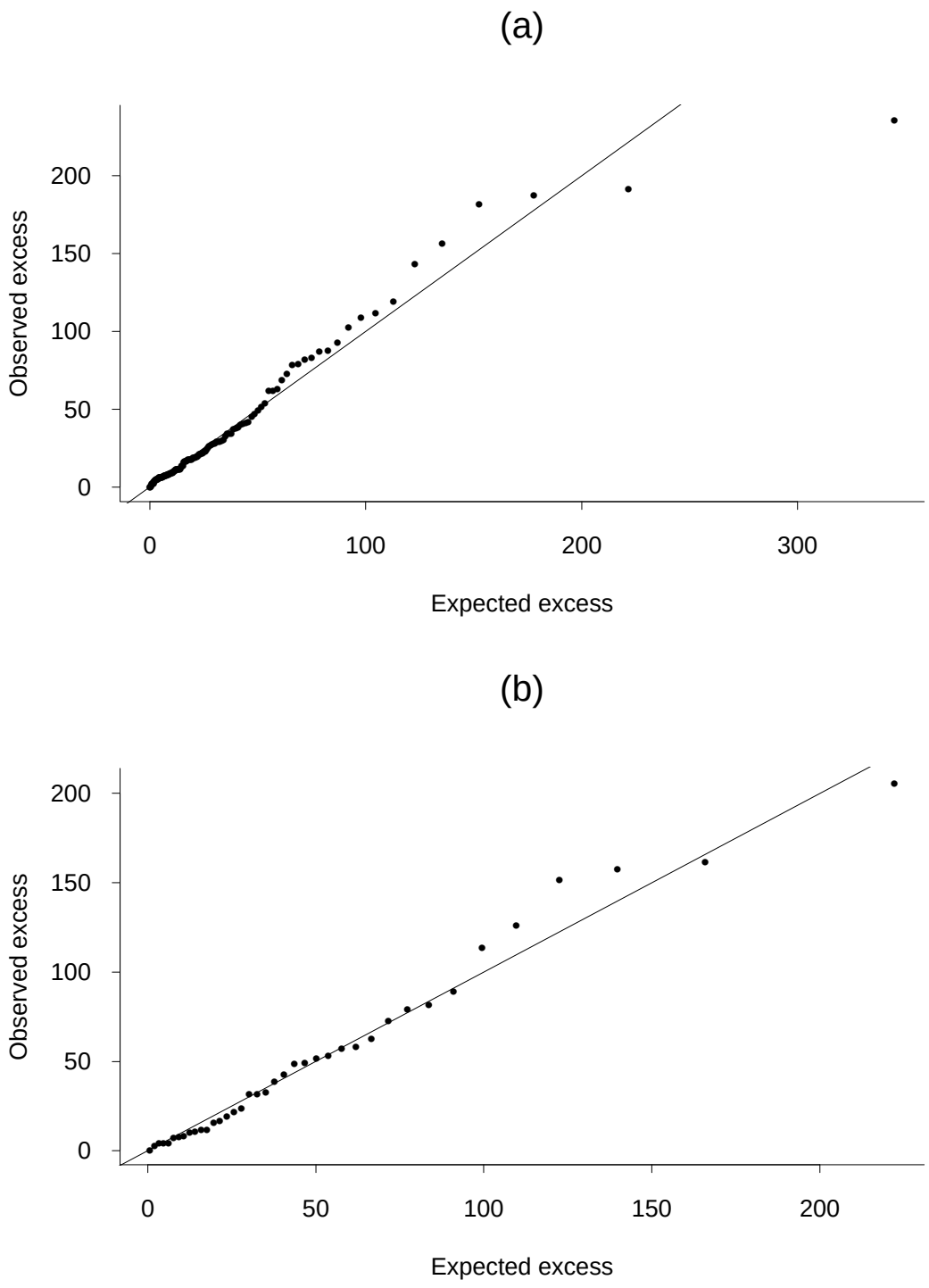


Fig. 8.10. QQ plots for GPD model fitted to high-level exceedances of River Nidd data. (a) Based on threshold $u = 70$. (b) Based on threshold $u = 100$.

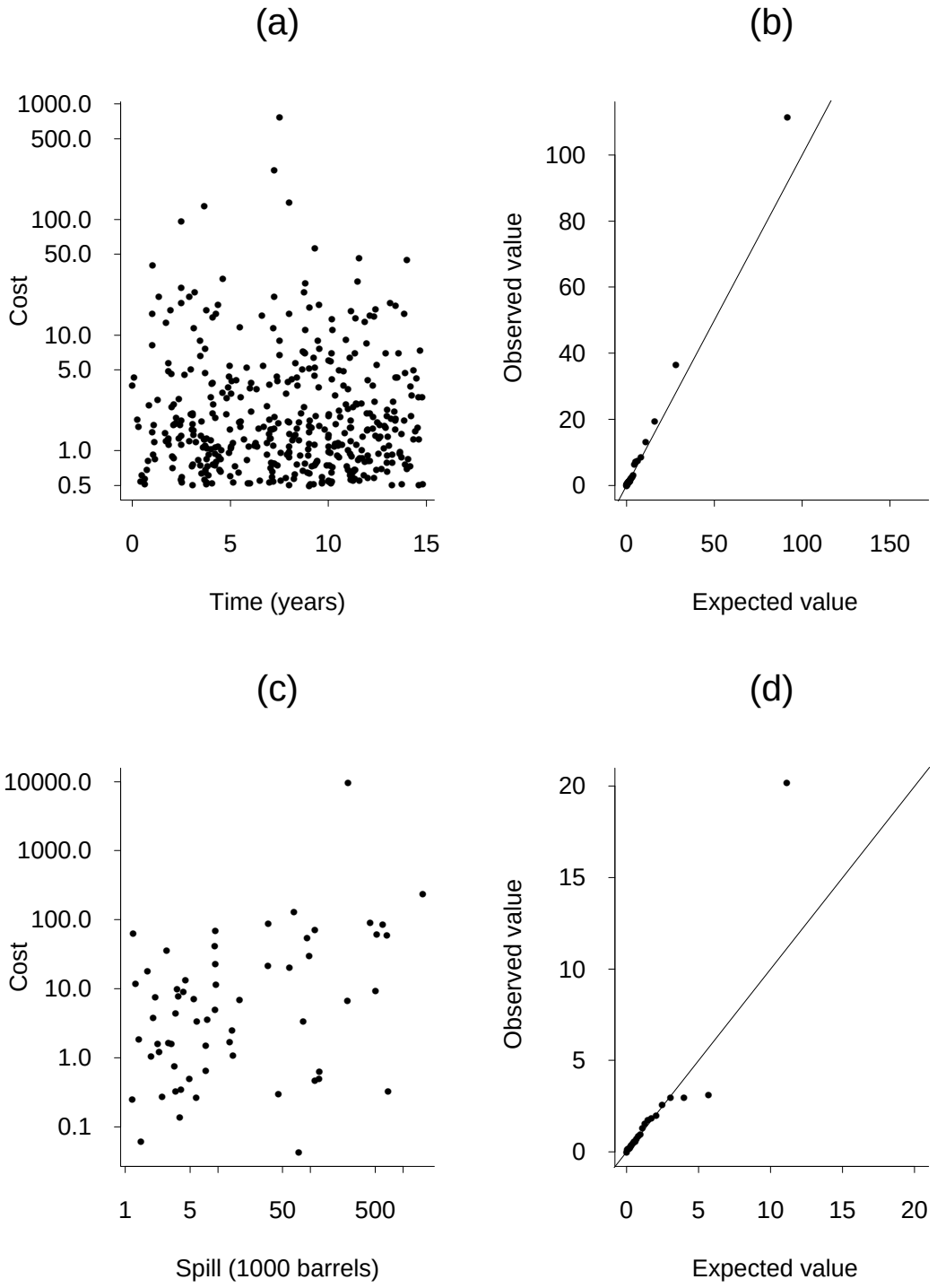


Fig. 8.11. (a) Time plot of large insurance claims incurred by a major company over a 15-year period. (b) QQ plot after fitting a GPD model to exceedances over a high threshold. (c) Plot of costs of major oil spills against size of spill. (d) QQ plot after fitting GPD regression model to exceedances over a high threshold, including spill size as a regressor. The obvious outlier corresponds to the Exxon Valdez spill of 1989. Data supplied by BP Insurance Limited.

8.9.3 The mean-excess plot

A different method was developed for the GPD by Davison and Smith (1990); this can be helpful in deciding on an appropriate threshold for exceedance-based methods.

The idea is based on equation (8.19): if the excesses over a threshold u indeed follow a GPD with parameters (σ, ξ), then the *mean excess* over any level $w > u$ should vary with w according to a straight line of slope $\xi/(1-\xi)$. The mean excess is also called the mean residual life in survival data studies, so another name for the plot is *mean residual life plot.*

A plot is calculated as follows. Given an initial threshold u, for each $w > u$, the actual mean of all excesses over w is calculated. This is then plotted against w. This plot has a discontinuity whenever w crosses an observation of the process, and the discontinuities get larger as w increases. Therefore the plot tends to have rather jagged features. However if the GPD is a good fit, the plot should stay reasonably close to the straight line obtained by substituting the fitted GPD parameters into (8.19).

The judgement may be aided by the following Monte Carlo technique. One hundred simulated samples are generated from the GPD fitted to the threshold u. For each sample, the GPD parameters are re-estimated and, for each $w > u$, the difference between the observed mean excess and the theoretical value obtained from (8.19) is computed. These are then placed in order: the top and bottom 5% points of the Monte Carlo values define the boundaries of a Monte Carlo 90% confidence interval, separately for each w.

Fig. 8.12 shows a mean excess plot, with confidence bands, for the River Nidd data. The confidence bands are based on the fitted GPD for a given threshold level: in (a) these are calculated with respect to threshold 70, and in (b) with respect to threshold 100. In the case of (a), the plot clearly varies outside the confidence bands. The interpretation of this is that the whole of the data over threshold 70 cannot be considered consistent with a single GPD. However, there is no such problem about threshold 100. This further reinforces our earlier decision to use threshold 100 as the basis for the GPD analysis.

Fig. 8.13 shows corresponding plots for the windspeed example from section 8.1. In the case of Raleigh with $u = 20$, the plot shows a jump at around $w = 30$, and the confidence bands show clearly that this *is* a statistically significant jump. In other words, $u = 20$ is too low a threshold for the GPD to form a good fit. However when the plot is recomputed for $u = 39.5$, there is no such problem. Corresponding plots for Greensboro and Charlotte, also based on $u = 39.5$, also show no problem with the GPD fit.

8.9.4. Plots based on the Z and W statistics.

A somewhat more general set of plots was introduced by Smith and Shively (1995) and appears to be of general utility in assessing the fit of extreme value models.

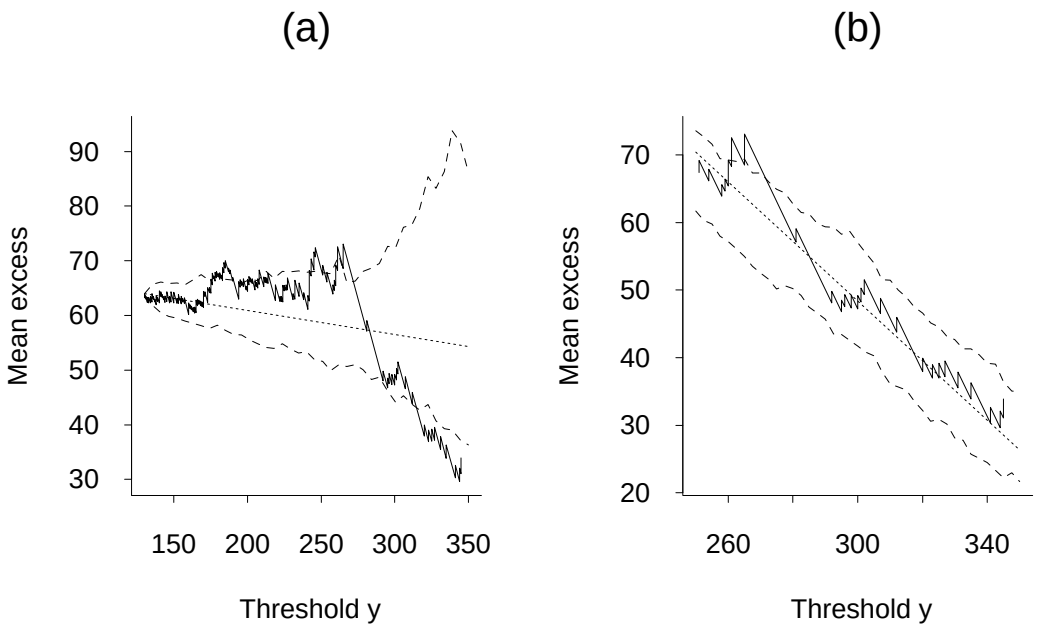


Fig. 8.12. Mean excess over threshold plots for Nidd data, with Monte Carlo confidence bands, relative to threshold 70 (a) and 100 (b).

Consider first the point process of exceedance times of a fixed threshold u. Suppose this is a nonhomogeneous Poisson process with intensity (as a function of time t) given by $\lambda_u(t)$. The model (8.51) is precisely of this form, with $\lambda_u(t) = \{1 + \xi_t(u - \mu_t)/\psi_t)\}^{-1/\xi_t}$ (recall equation (8.21)). Suppose we start observing the process at a time T_0, and observe subsequent exceedances at times $T_1, T_2, ...$ For $k \geq 1$, define

$$Z_k = \int_{T_{k-1}}^{T_k} \lambda_u(s) ds. \tag{8.58}$$

According to standard theory of the nonhomogeneous Poisson process, if the model is indeed nonhomogeneous Poisson with intensity $\lambda_u(\cdot)$, the random variables Z_k are independent, exponentially distributed with mean 1.

In practice, most processes are observed in discrete time rather than continuous time (e.g. daily ozone levels), but in that case it is usually adequate to approximate (8.58) with the obvious discrete approximation to the integral.

A second statistic is denoted W_k, and is defined as follows. Suppose the model is given by (8.53) and the kth exceedance occurs at time T_k. Suppose also that the corresponding excess value, i.e. the amount by which the process exceeds the threshold at time T_k, is Y_k. Then define

$$W_k = \frac{1}{\xi_{T_k}} \log \left\{ 1 + \frac{\xi_{T_k} Y_k}{\psi_{T_k} + \xi_{T_k}(u - \mu_{T_k})} \right\}. \tag{8.59}$$

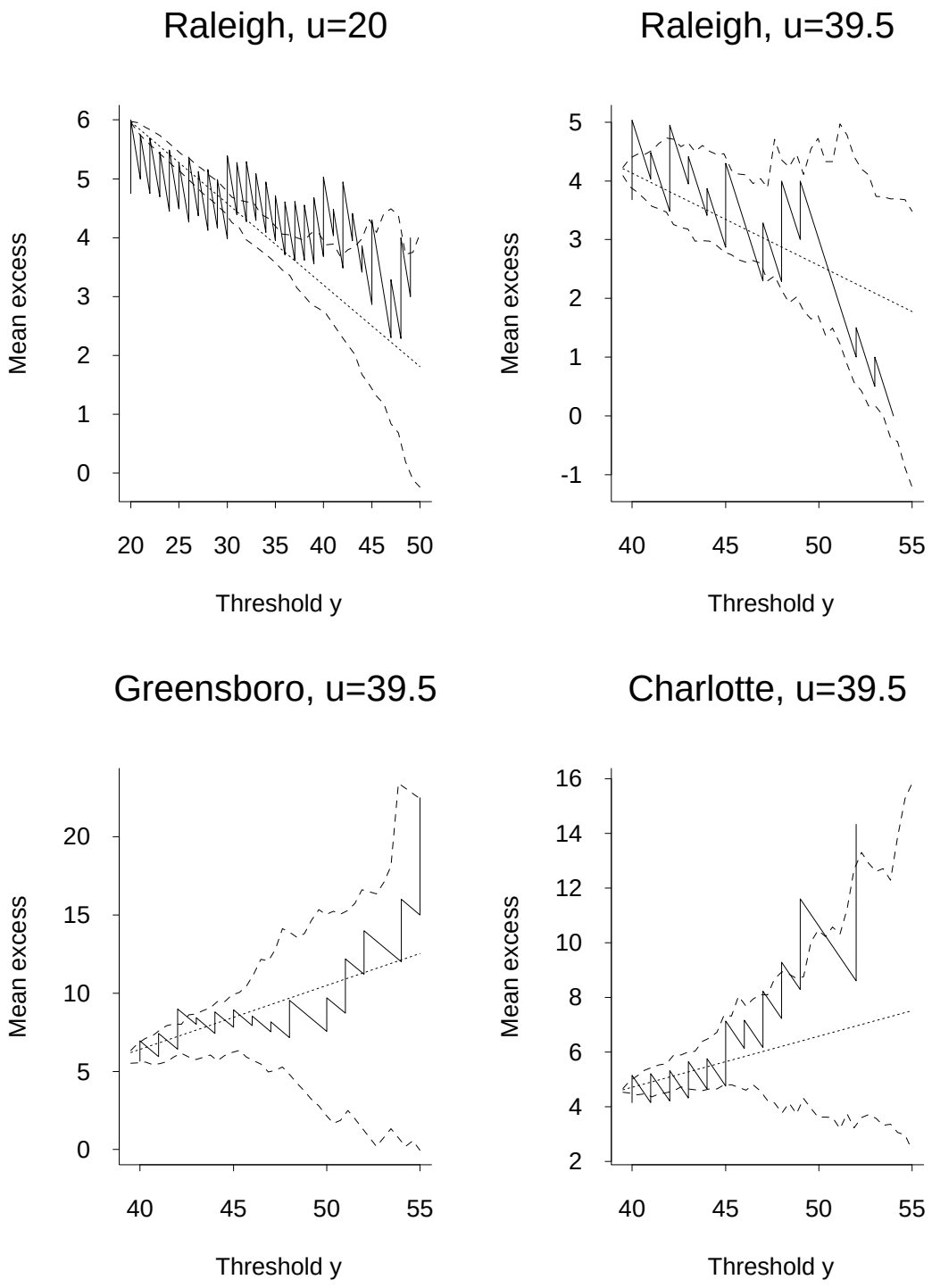


Fig. 8.13. Mean excess plots (with Monte Carlo confidence bands) for windspeed data sets. (a) Raleigh data with threshold 20. (b) Raleigh data with threshold 39.5. (c) Greensboro data with threshold 39.5. (d) Charlotte data with threshold 39.5.

Then, if the assumed model is correct, $W_1, W_2, ...$, are also independent exponentially distributed random variables with mean 1. A derivation of this is as follows. First note that, if we define $\sigma_t = \psi_t + \xi_t(u - \mu_t)$ then, given an exceedance at time t, the excess value follows a GPD with parameters σ_t and ξ_t: this is just (8.21) with suffices t inserted to allow for the time-inhomogeneity of the process. However, in that case, (8.59) is equivalent to

$$W_k = \frac{1}{\xi_{T_k}} \log \left(1 + \frac{\xi_{T_k} Y_k}{\sigma_{T_k}} \right)$$

and this is just a probability integral transformation of the GPD to an exponential distribution with mean 1.

Smith and Shively (1995) constructed QQ plots based on the order statistics of Z_k and W_k showing, in both cases, that for their example there was excellent fit to the assumed distribution. However, there are other plots that one can draw based on the Z_k and W_k statistics —

(i) A plot of either Z_k or W_k against time T_k serves as a check against the presence of residual trends,

(ii) A plot of the autocorrelation of either Z_k or W_k serves as a check on serial dependence.

As an example, Fig. 8.14 shows all six plots for the Charlotte windspeed data from section 8.8. All the plots indicate reasonable agreement with the underlying assumptions.

8.10. Rainfall data over the United States

The following discussion is based on the preprint Smith (1999), and shows how the methods of this chapter may be combined with some of the techniques of earlier chapters to assess spatial patterns in trends in extreme rainfall levels.

The data source is daily rainfall series from 187 stations in the Historical Climatological Network, which covers the continental U.S.A., and are archived by the National Climatic Data Center (www.ncdc.noaa.gov).

The background to this work is a number of papers in the climatological literature, which have examined "indicators of climate change" — in other words, which ones among a large number of climatic variables represent the clearest indication of a possibly anthropogenic climate change. Karl *et al.* (1996) performed time series analysis on a number of climatic series and concluded that those based on the spatial intensity of extreme rainfall events showed a particularly strong temporal trend. A detailed follow-up study by Karl and Knight (1998) classified rainfall data into frequency of occurrence with 20 equiprobable intervals, and demonstrated that the highest interval, representing the top 5% of rainfall

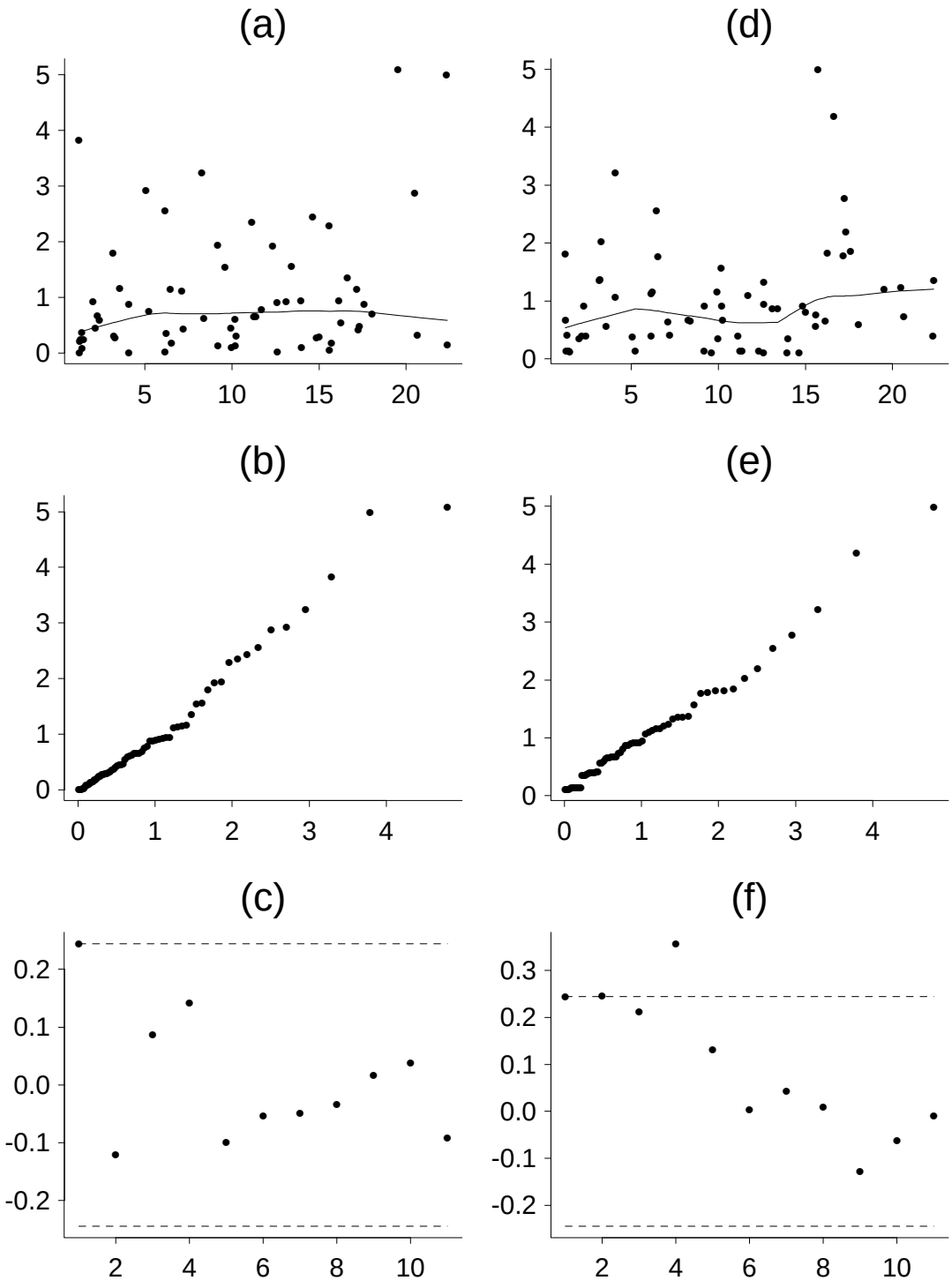


Fig. 8.14. Diagnostic plots based on Z and W statistics for Charlotte seasonal model 1. (a) and (d): plots against time with fitted LOESS curve. (b) and (e): QQ plots. (c) and (f): Autocorrelation plots with approximate 95% confidence bands under the null model of no autocorrelation.

days, displayed the strongest evidence of an increasing trend. The data they used, however, were based on rough spatial aggregations of daily rainfall series, and gave little indication of what was happening in any one such series. Also, their methods did not use any of the techniques of extreme value theory which we have reviewed in this chapter, and therefore missed the opportunity to obtain detailed distributional results.

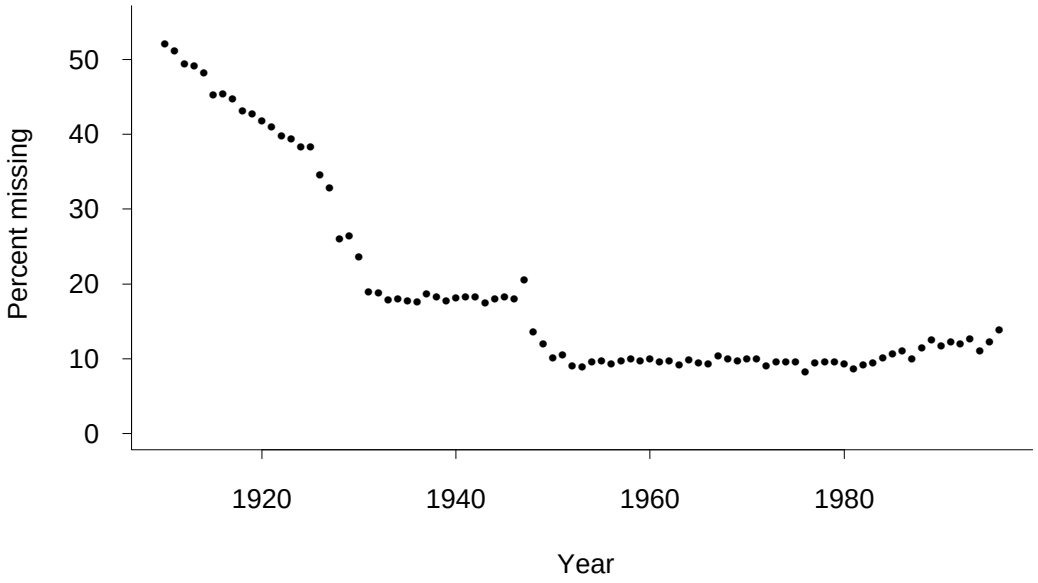


Fig. 8.15. Percentage of missing data for each year.

In the paper Smith (1999), an analysis was pursued based on fitting models of the point process type (section 8.4) to the individual series, followed by spatial aggregation of the results, using the same techniques as in chapter 2 of this book. The data nominally cover the period 1910–1996 for 187 stations, though there are substantial missing data. Fig. 8.15 shows a plot of the total missing data in each year, expressed as a percentage of all possible data points if all stations were open for the entire year. Only after about 1950 is there a consistent level of coverage with around 10% missing data.

The models adopted for the individual stations are similar to those earlier in this chapter. In particular, the main model fitted is similar to (8.53), but written in a different way: with parameters $\mu_t,\ \psi_t,\ \xi_t$ as the time-dependent local parameters of the extreme value point process, we write

$$\mu_t = \mu_0 e^{v_t}, \qquad \psi_t = \psi_0 e^{v_t}, \qquad \xi_t = \xi_0, \tag{8.60}$$

where $\mu_0,\ \psi_0,\ \xi_0$ are constants and

$$v_t = \beta_1 t + \sum_{p=1}^{P} \{\beta_{2p}\cos(\omega_p) + \beta_{2p+1}\sin(\omega_p)\}, \tag{8.61}$$

which represents a linear trend, with coefficient β_1, and P sinusoidal terms representing periodic (seasonal) components — in practice, the analysis used $P = 2$ after a selection process similar to the windspeed examples in section 8.8.2. The motivation for the specific form of model given by (8.60) and (8.61) was primarily ease of interpretation: under this model, the overall rate of increase of the most extreme quantiles of the process is β_1 per year. For example, if $\beta_1 = .001$ (which turns out to be close to the actual value), this corresponds to an increase in the most extreme quantiles of about 0.1% per year. In practice we multiply values of β_1 by 100 so that they have the interpretation of percent increase per year.

Diagnostics

Several of the diagnostic procedures described in section 8.9 were used to inform the choice of models for these data sets. As examples, we apply them to four specific stations, denoted stations 1–4, corresponding to stations in the mountains of North Carolina, western Colorado, the southern California coast and the Atlantic coast of Florida, respectively. The units for daily rainfall are hundredths of an inch, and the scales on Fig. 8.16 highlights the very sharp differences of overall rainfall levels among these four stations.

Fig. 8.16 shows mean excess plots, with Monte Carlo confidence bands. Recall that a linear plot would indicate good fit to the GPD. In stations 1, 3 and 4 there are sharp changes in the slope of the plot near the upper end, but only in station 1 does this appear significant as judged by the Monte Carlo confidence bands.

Figs. 8.17 and 8.18 show probability plots of the Z and W statistics for the same four stations, based on the model (8.60)–(8.61), where the threshold is set at the 98th percentile of the distribution of daily rainfall values, separately for each station. Only the W plot for station 4 shows a really noticeable discrepancy, somewhat contradicting the impression created by Fig. 8.16 (which suggested that the problem lay with station 1, not station 4). However it should be pointed out that Fig. 8.18 is plotted after fitting the seasonal model, and therefore takes account of seasonal variation in the rainfall events, whereas Fig. 8.16 does not take this into account. Therefore, where the two plots give conflicting signals, Fig. 8.18 is probably the more reliable.

Threshold	98%	98%	99%	99%	99.5%	99.5%
	β_1	ξ	β_1	ξ	β_1	ξ
$t > 2$	25	74	22	45	18	34
$t > 1$	73	134	58	114	61	81
$t > 0$	125	162	118	155	109	134
$t < 0$	59	22	66	29	75	50
$t < -1$	21	5	23	8	24	14
$t < -2$	10	1	5	0	5	2

Table 8.6. Summary table of t statistics (estimate divided by standard error) for extreme value model applied to 187 stations and three rules for determining the threshold (top 2%, top 1% and top 0.5%).

Despite these and other discrepancies, the results overall do not show great sensitivity to the choice of threshold. As an example, Table 8.6 classifies the β_1 and ξ parameters according to their t statistics — the parameters divided by their standard errors — using three different thresholds defined by the 98th, 99th and 99.5th percentiles for each station. We are particularly interested to see the number of stations for which $\hat{\beta}_1 > 0$ (indicating an increasing trend) or for which $\hat{\xi} > 0$ (indicating a heavier-than-exponential tail). As a side comment, for rainfall data it is very common to assume a gamma distribution (Stern and Coe 1984, Smith 1994b), which in terms of the asmyptotic theory of sections 8.2 and 8.3, implies limiting extreme value distributions with $\xi = 0$; therefore, a clear indication that $\xi > 0$ in the majority of stations would contradict conventional wisdom about the distribution of rainfall values (though we should point out that other authors have remarked that the gamma distribution does not necessarily fit well in the upper tail of the daily rainfall distribution).

Table 8.6 shows that for each of the three thresholds considered, there is a clear majority of stations for which both $\hat{\beta}_1 > 0$ and $\hat{\xi} > 0$. We therefore see confirming evidence of both an increasing trend and a long-tailed distribution. However, the individual values of $\hat{\beta}_1$ and $\hat{\xi}$ for single stations show huge variability, and consequently are difficult to interpret.

Spatial modeling

We therefore attempt a spatial smoothing, similar to the hierarchical models in chapter 2. The observed $\hat{\beta}_1(s)$ parameters (where s indexes the station) are treated as noisy observations of an underlying smooth random field $\beta_1(s)$, where the standard errors of the noise variables are equated to the standard errors of $\hat{\beta}_1(s)$ in the single-station extreme values analyses. The random field for $\beta_1(s)$ is assumed, after some trial and error, to be of Matérn type with no deterministic spatial trend. Shape parameters of the Matérn field are typically in the range 0–0.3, depending on the precise set of coefficients to which the model is fitted. Such small Matérn shape parameters suggest a field with many local irregularities. After fitting the Matérn model, kriging was used to reconstruct the $\beta_1(s)$ surface. Recall that the values of β_1 have been multiplied by 100 so that they represent annual percentage changes in the extreme rainfall levels.

A first attempt to apply this method resulted in Fig. 8.19, based on a 98% threshold at each station and the whole time period. This shows a highly irregular field and is hard to interpret. A second plot, in Fig. 8.20, is based on the same spatial model fitted to coefficients from an extreme value analysis applied to 95% thresholds, using data from only 1951 onwards. The idea was that by applying the model to a lower threshold, and by restricting the analysis to a time period for which there were relatively few missing values (recall Fig. 8.15), we should obtain more stable estimates. This is confirmed by the plot, though Fig. 8.20 is still hard to interpret in view of the many local irregularities.

From another point of view, however, Fig. 8.20 is easy to interpret. There is very little *large-scale* spatial variation — all the observed variability is very localized, and the

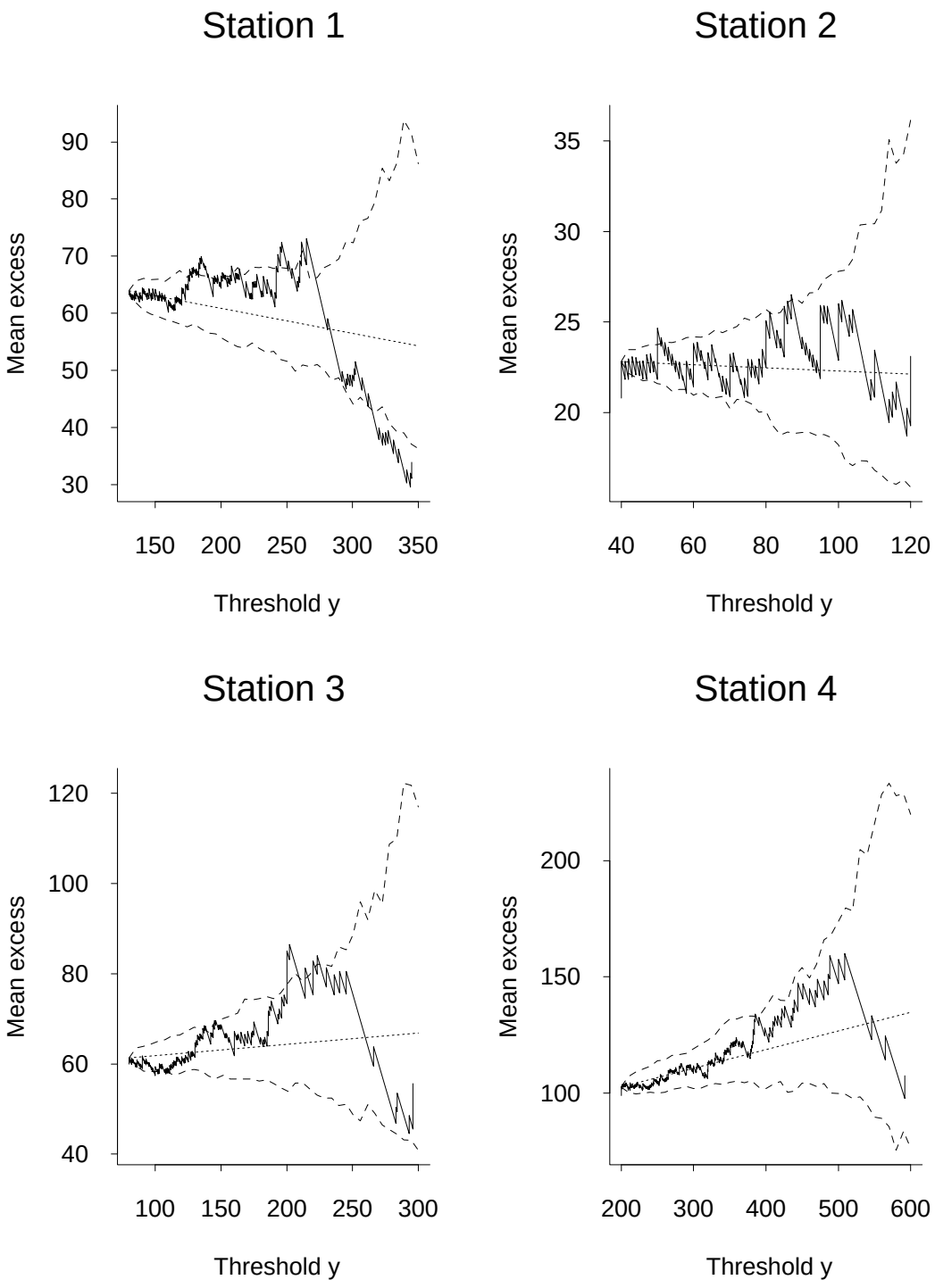


Fig. 8.16. Mean excess plots for four stations.

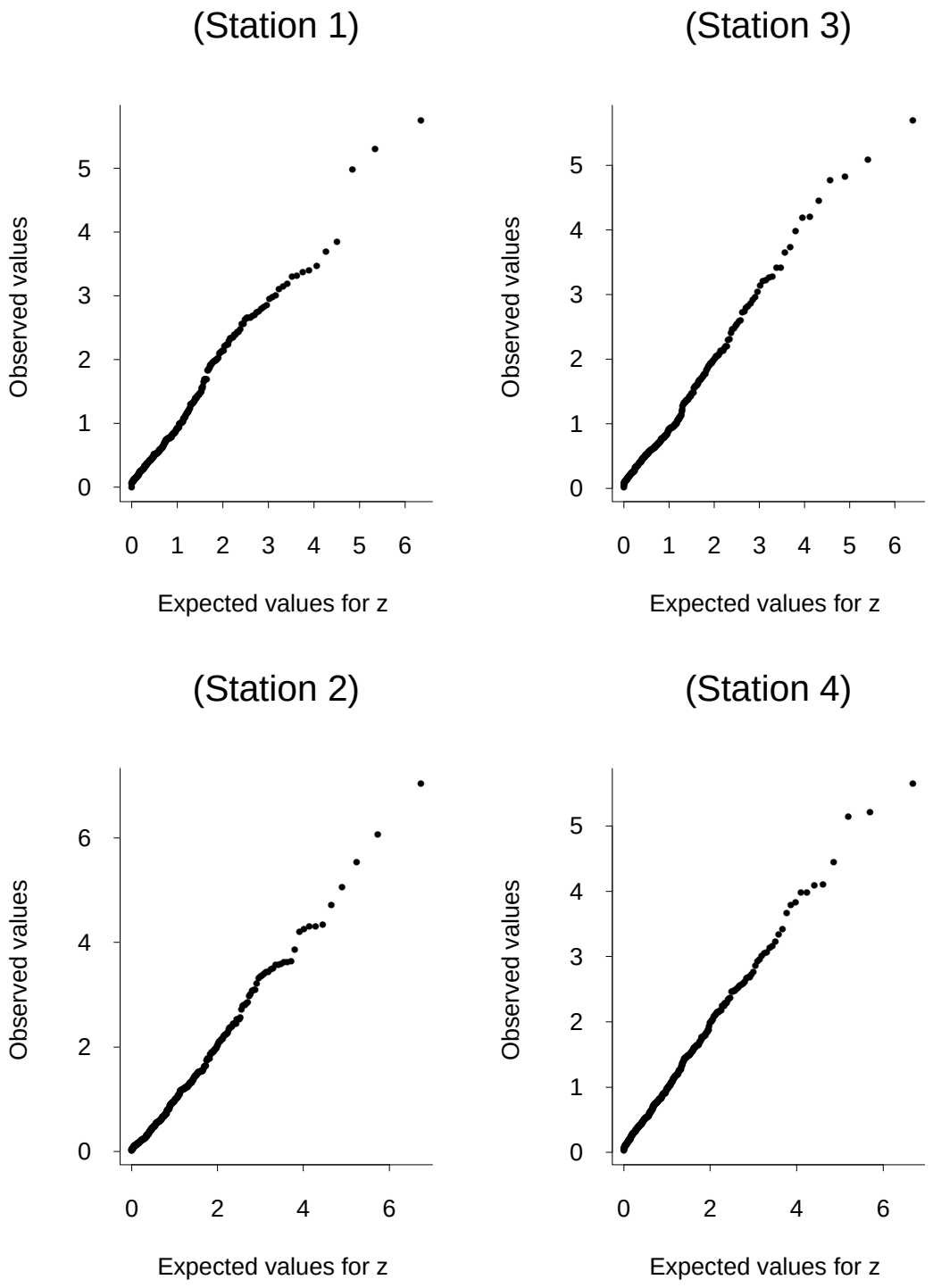


Fig. 8.17. Probability plots for Z statistics.

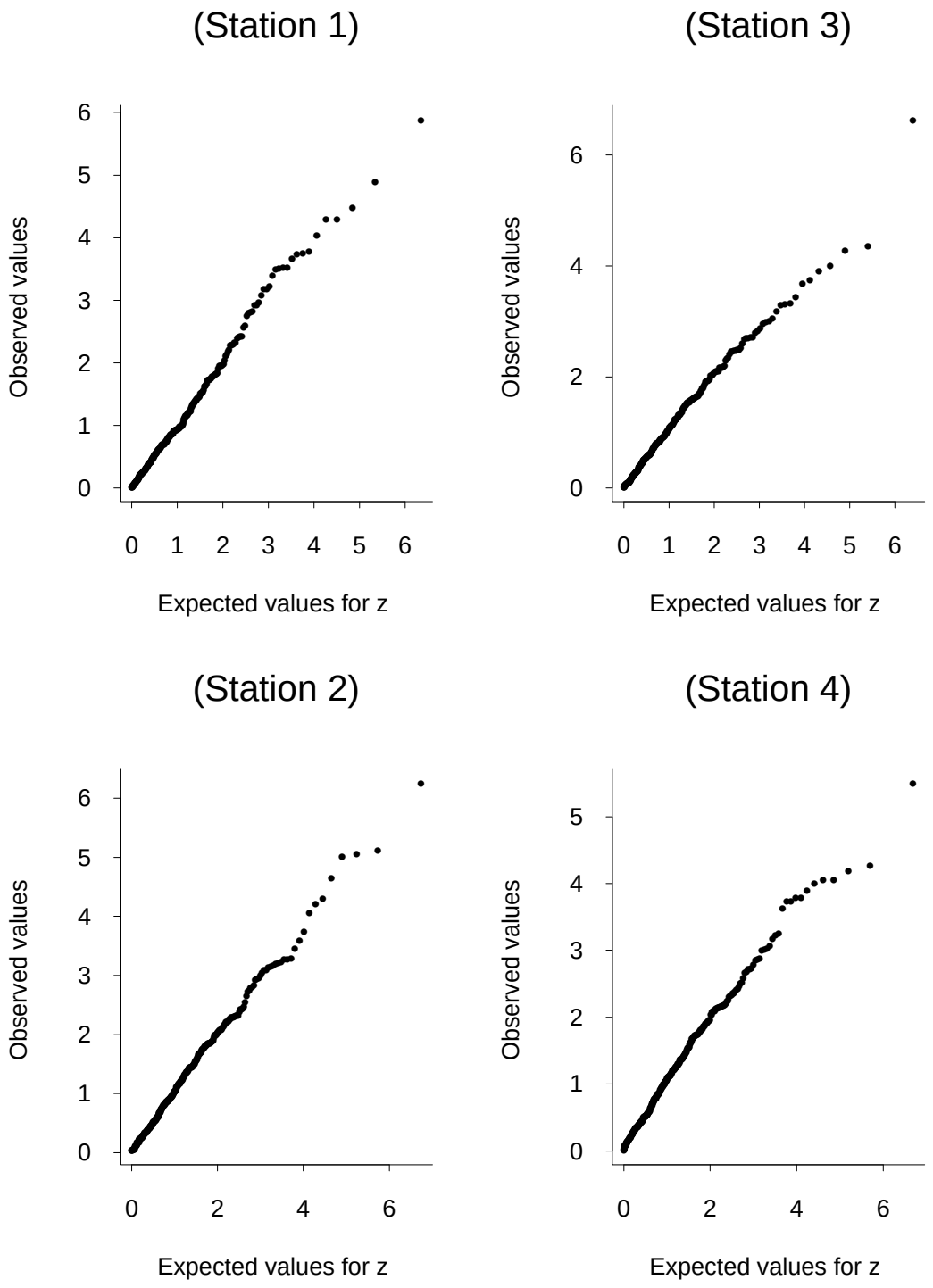


Fig. 8.18. Probability plots for W statistics.

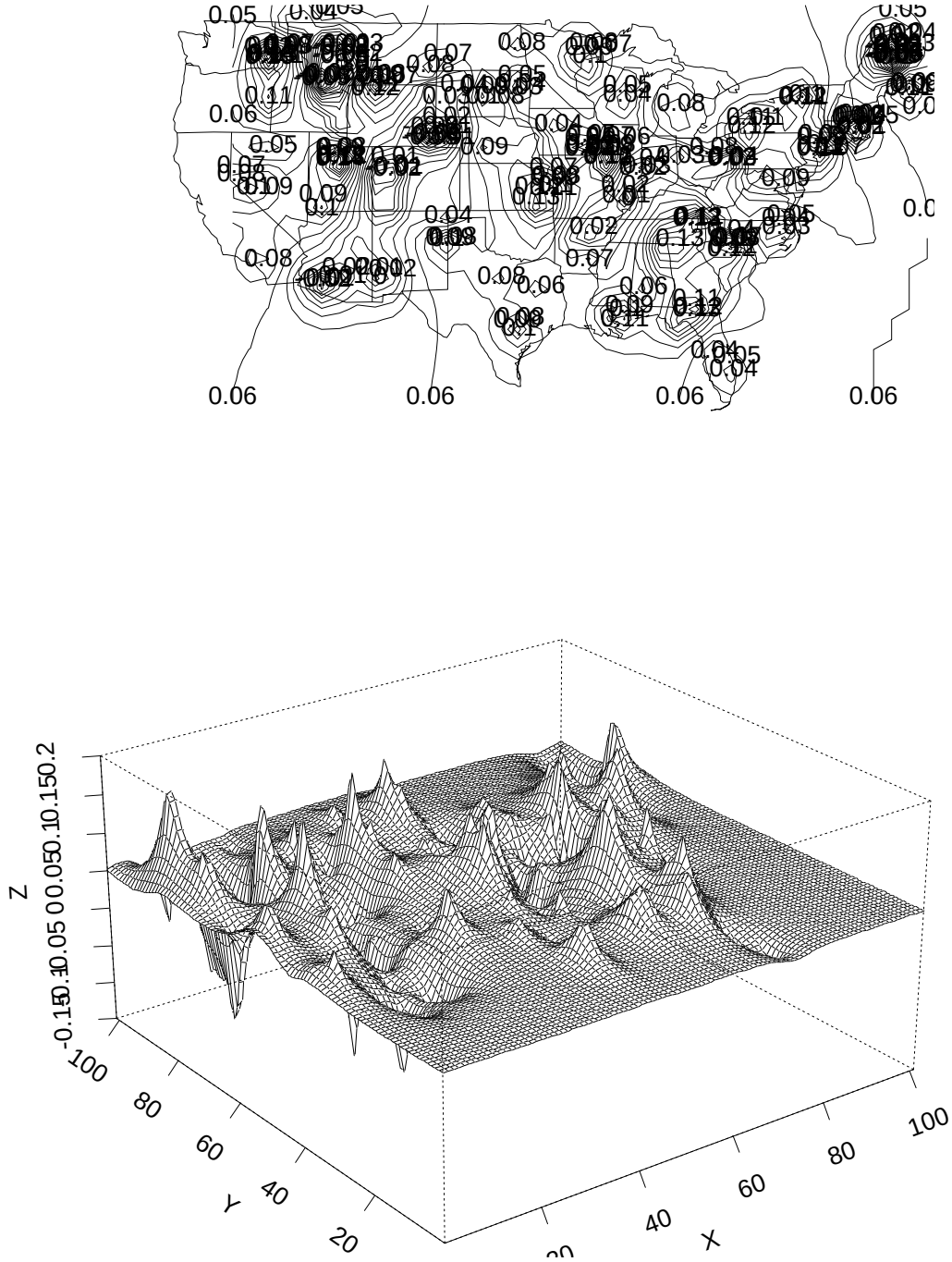


Fig. 8.19. Contour plot of reconstructed trend surface, based on 98% thresholds and time period 1910–1996.

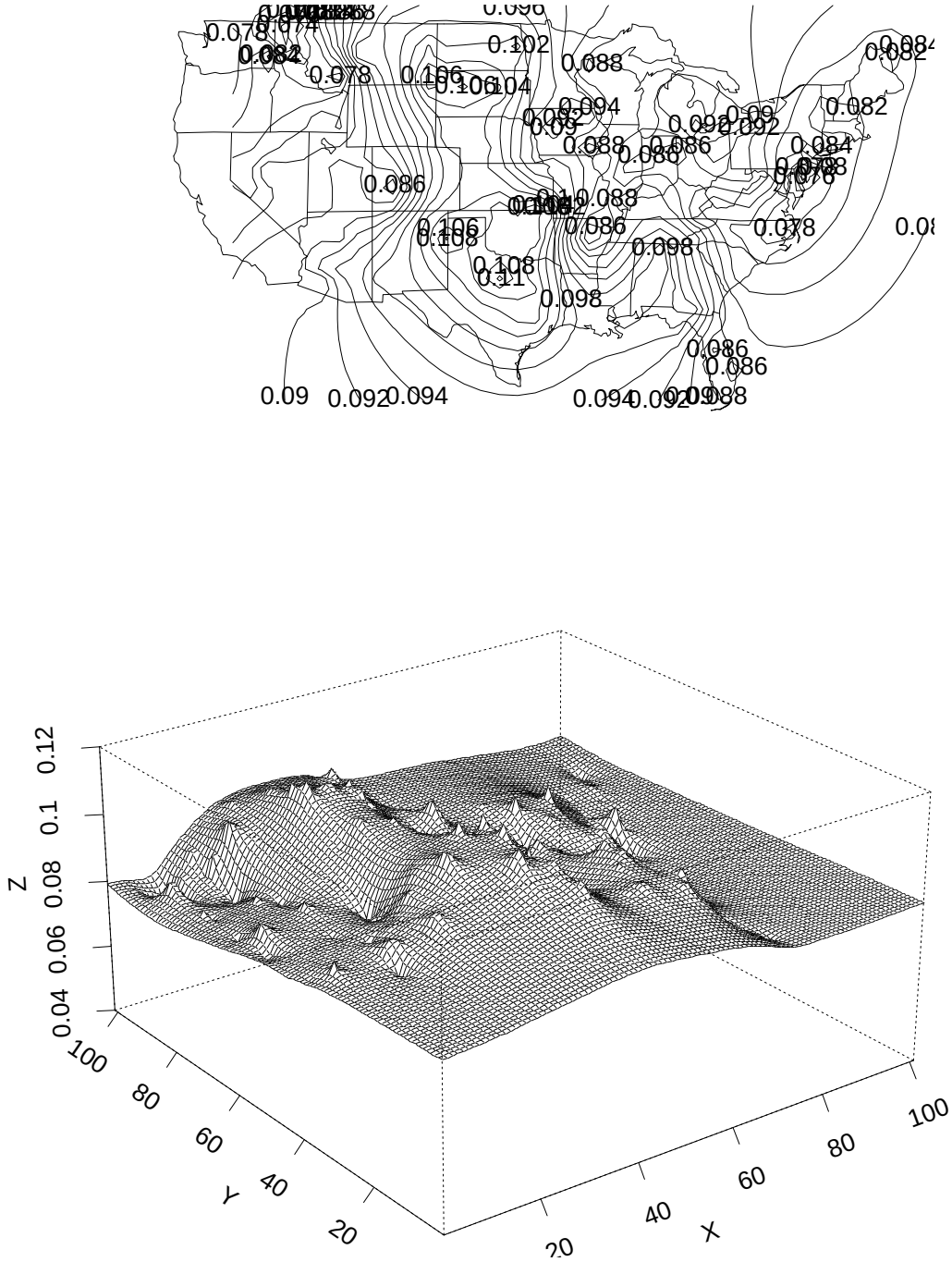


Fig. 8.20. Contour plot of reconstructed trend surface, based on 98% thresholds and time period 1951–1996.

plot does not deviate from an overall value which represents an increasing trend of about 0.09% per year. In Smith (1999), this interpretation was confirmed by computing regional average estimates of $\beta_1(s)$ for a number of regions. There was little variation between the regions, in sharp contrast to similar analyses based on temperature trends in chapter 2.

The implication of this conclusion for climate modeling is that they imply some specific, testable, hypotheses for the response of the climate system to external forcing such as greenhouse gases. It would be of very great interest to see to what extent the conclusions given here could be replicated in numerical climate models under scenarios of increasing greenhouse gases.

REFERENCES

Abramowitz, M. and Stegun, I.A. (1964), *Handbook of Mathematical Functions.* National Bureau of Standards, Washington D.C., reprinted by Dover, New York.

Aitchison, J. (1975), Goodness of prediction fit. *Biometrika* **62**, 547–554.

Allen, M.R., Stott, P.A., Mitchell, J.F.B., Schnur, R. and Delworth, T.L. (2000), Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **407**, 617–620.

Allen, M.R. and Tett, S.F.B. (1999), Checking for model consistency in optimal fingerprinting. *Climate Dynamics* **15**, 419–434.

Almeida, M.P. and Gidas, B. (1993), A variational method for estimating the parameters of MRF from complete or incomplete data. *Annals of Applied Probability* **3**, 103–136.

Altman, N.S. (1990), Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc* **85**, 749–759.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis.* Second edition, Wiley, New York.

Anstreicher, K.M., Fampa, M., Lee, J. and Williams, J. (1996), Using continuous nonlinear relaxations to solve constrained maximum-entropy sampling prolems. Unpublished.

Atkinson, A.C. and Donev, A.N. (1992), *Optimum Experimental Designs.* Oxford University Press.

Baker, J.R., Peck, D.V. and Sutton, D.W. (eds.) (1997), Environmental Monitoring and Assessment Program Surface Waters: Field Operations Manual for Lakes. EPA/620/R-97/001. U.S. Environmental Protection Agency, Washington, D.C. Available online from http://zoology.muohio.edu/oris/Tahoe/Lake_manual/lake_ove.pdf

Barndorff-Nielsen, O.E. and Cox, D.R. (1996), Prediction and asymptotics. *Bernoulli* **2**, 319–340.

Barnett, T.P. (1990), Beware greenhouse confusion (Commentary on Kuo *et al.* (1990)). *Nature* **343**, 696–697.

Barry, R.P. and Ver Hoef, J.M. (1996), Blackbox kriging: spatial prediction without specifying variogram models. *Journal of Agricultural, Biological and Environmental Statistics* **1** 297–322.

Bartlett, M.S. (1938), The approximate recovery of information from field experiments with large blocks. *J. Agric. Sci.* **28**, 418–427.

Bartlett, M.S. (1955), *An Introduction to Stochastic Processes.* Cambridge University Press, Cambridge.

Bartlett, M.S. (1971), Physical nearest-neighbour models and non-linear time series. *J. Appl. Probab.* **8**, 222–232.

Bartlett, M.S. (1971), Physical nearest neighbour models and non-linear time series. *J. Applied Probability* **8**, 222–232.

Bartlett, M.S. (1976), *The Statistical Analysis of Spatial Pattern.* Chapman and Hall, London.

Bartlett, M.S. (1978), Nearest neighbour models in the analysis of field experiments (with discussion). *J.R. Statist. Soc. B* **40**, 147–175.

Bates, R.A., Buck, R.J., Riccomagno, E. and Wynn, H.P. (1996), Experimental design and observation for large systems. *J.R. Statist. Soc. B* **58**, 77–94.

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988), *The New S Language: A Programming Environment for Data Analysis and Graphics*. Wadsworth and Brooks/Cole, Pacific Grove, CA.

Beirlant, J., Teugels, J.J. and Vynicker, P. (1996), *Practical Analysis of Extreme Values*. Leuven University Press, Leuven, Belgium.

Benedetti, R. and Palma, D. (1995), Optimal sampling designs for dependent spatial units. *Environmetrics* **6**, 101–114.

Beran, J. (1994), *Statistics for Long-Memory Processes*. Chapman and Hall, New York; in press.

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis* (second edition). Springer, New York.

Berliner, L.M., Lu, Z.-Q. and Snyder, C. (1999), Statistical design for adaptive weather observations. *J. Atmos. Sci.* **56**, 2536–2552.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. and Songoni, M. (1995), Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine* **14**, 2433–2443.

Bernardinelli, L., Clayton, D. and Montomoli, C. (1995), Bayesian estimates of disease maps: How important are priors? *Statistics in Medicine* **14**, 2411–2431.

Bernardinelli, L. and Montomoli, C. (1992), Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine* **11**, 983–1007.

Bernardinelli, L., Pascutto, C., Best, N.G. and Gilks, W.R. (1997), Disease mapping with errors in covariates. *Statistics in Medicine* **16**, 741–752.

Bernardo, J.M. (1979), Expected information as expected utility. *Annals of Statistics* **7**, 686–690.

Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*. Wiley, New York.

Besag, J.E. (1974), Spatial interaction and the statistical analysis of lattice systems (with discussion). *J.R. Statist. Soc. B* **36**, 192–236.

Besag, J.E. (1975), Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Besag, J. (1986), The statistical analysis of dirty pictures (with discussion). *J.R. Statist. Soc. B* **48**, 259–302.

Besag, J. (1989), A candidate's formula: A curious result in Bayesian prediction. *Biometrika* **76**, 183.

Besag, J.E. and Green, P.J. (1993), Spatial statistics and Bayesian computation. *J.R. Statist. Soc. B* **55**, 25–37.

Besag, J.E., Green, P., Higdon, D. and Mengersen, K. (1995), Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Besag, J.E. and Higdon, D. (1999), Bayesian analysis of agricultural field experiments (with discussion). *J.R. Statist. Soc. B* **61**, 691–746.

Besag, J.E., York, J. and Mollié, A. (1991), Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1–59.

Bielza, C., Müller, P. and Rios Insua, D. (1999), Monte Carlo methods for decision analysis with applications to influence diagrams. *Management Science.*

Bloomfield, P. (1992), Trends in global temperature. *Climatic Change* **21**, 1–16.

Bloomfield, P. and Nychka, D. (1992), Climate spectra and detecting climate change. *Climatic Change* **21**, 275–288.

Bookstein, F.L. (1991), *Morphometric Tools for Landmark Data.* Cambridge University Press.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994), *Time Series Analysis, Forecasting and Control.* Third Edition, Prentice Hall, Englewood Cliffs, N.J.

Box, G.E.P. and Tiao, G.C. (1973), *Bayesian Inference in Statistical Research.* McGraw Hill, New York.

Bras, R.L. and Rodriguez-Iturbe, I. (1976a), Network design for the estimation of areal mean of rainfall events. *Water Resources Research* **12**, 1185–1195.

Bras, R.L. and Rodriguez-Iturbe, I. (1976b), Rainfall network design for runoff prediction. *Water Resources Research* **12**, 1199–1208.

Breslow, N.E. (1984), Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44.

Breslow, N.E. and Clayton, D.G. (1993), Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9–25.

Brook, D. (1964), On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* **51**, 481–483.

Brooks, S.P. and Morgan, B.J.T. (1995), Optimisation using simulated annealing. *The Statistician* **44**, 241–257.

Brown, P.J. (1993), *Measurement, Regression and Calibration.* Oxford University Press, Oxford.

Brown, P.J., Le, N.D. and Zidek, J.V. (1994a), Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics* **22**, 489–509.

Brown, P.J., Le, N.D. and Zidek, J.V. (1994b), Inference for a covariance matrix. In *Aspects of Uncertainty: A Tribute to D.V. Lindley* (eds. A.F.M. Smith and P.R. Freeman). Wiley, Chichester.

Bueso, M.C., Angulo, J.M. and Alonso, F.J. (1998), A state-space model approach to optimum spatial sampling based on entropy. *Environmental and Ecological Statistics* **5**, 29–44.

Cambanis, S. (1985), Sampling designs for time series. In *Handbook of Statistics* (E.J. Hannan, P.R. Krishnaiah, M.M. Rao, eds.), Vol. 5, 337–362. Elsevier, Amsterdam.

Carlin, B.P. and Louis, T.A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, London.

Casdagli, M. (1989), Nonlinear prediction of chaotic time series. *Physica D* **35**, 335-356.

Caselton, W.F. and Zidek, J.V. (1984), Optimal monitoring network designs. *Statistics and Probability Letters* **2**, 223–227.

Caselton, W.F., Kan, L. and Zidek, J.V. (1992), Quality data networks that minimize entropy. Chapter 2 of *Statistics in the Environmental and Earth Sciences*, eds. A. Walden and P. Guttorp, Halsted Press, New York, pp. 10-38.

Castillo, E. (1988), *Extreme Value Theory in Engineering.* Academic Press, Boston.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A., (1983), *Graphical Methods for Data Analysis.* Duxbury, Boston.

Chen, C.F. (1979), Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *J.R. Statist. Soc. B* **41**, 235–248.

Cherry, S., Banfield, J. and Quimby, W.F. (1996), An evaluation of a non-parametric method of estimating semi-variograms of isotropic spatial processes. *Journal of Applied Statistics* **23**, 435–449.

Clayton, D.G.and Kaldor, J. (1987), Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671–681.

Cleveland, W.S. (1979), Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829–836.

Cliff, A.D. and Ord, J.K. (1973), *Spatial Autocorrelation.* Pion, London.

Cliff, A.D. and Ord, J.K. (1981), *Spatial Processes: Models and Applications.* Pion, London.

Clifford, P. (1990), Markov random fields in statistics. In *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, eds. G.R. Grimmett and D.J.A. Welsh, Oxford University Press, pp. 19–32.

Cochrane, D. and Orcutt, G.H. (1949), Application of least squares regression to relationships continaing autocorrelated regression terms. *J. Amer. Statist. Assoc.* **44**, 32–61.

Cohen, A. and Jones, R.H. (1969), Regression on a random field. *J. Amer. Statist. Assoc* **64**, 1172–1182.

Cohen, J.P. (1982a), The penultimate form of approximation to normal extremes. *Adv. Appl. Prob.* **14**, 324–339.

Cohen, J.P. (1982b), Convergence rates for the ultimate and penultimate approximations in extreme value theory. *Adv. Appl. Prob.* **14**, 833–854.

Coles, S. (1996), *Extreme Value Theory and Applications.* Lecture notes available from http://www.maths.lancs.ac.uk/~coless.

Coles, S.G. and Dixon, M.J. (1997), Likelihood-based inference for extreme value models. Unpublished manuscript, Lancaster University.

Coles, S.G. and Powell, E.A. (1996), Bayesian methods in extreme value modelling: a review and new developments. *Internat. Statist. Rev.* **64**, 119–136.

Coles, S.G. and Tawn, J.A. (1991), Modelling extreme multivariate events. *J. R. Statist. Soc. B* **53**, 377-392.

Coles, S.G. and Tawn, J.A. (1994), Statistical methods for multivariate extremes: An applications to structural design (with discussion).

Coles, S.G. and Tawn, J.A. (1996a), A Bayesian analysis of extreme rainfall data. *Applied Statistics* **45**, 463–478.

Coles, S.G. and Tawn, J.A. (1996b), Modelling extremes of the areal rainfall process. *J.R. Statist. Soc. B* **58**, 329–347.

Coles, S., Tawn, J.A.and Smith, R.L. (1994) A seasonal Markov model for extremely low temperatures. *Environmetrics* **5**, 221–239.

Comets, F. (1992), On consistency of a class of estimators for exponential families of Markov random fields on a lattice. *Annals of Statistics* **20**, 455–468.

Comets, F. and Gidas, B. (1991), Asymptotics of maximum likelihood estimators for the Curie-Weiss model. *Annals of Statistics* **19**, 557–578.

Comets, F. and Gidas, B. (1992), Parameter estimation for Gibbs distributions from partially observed data. *Annals of Applied Probability* **2**, 142–170.

Cook, D.G. and Pocock, S.J. (1983), Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics* **39**, 361–371.

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics.* Chapman and Hall, London.

Cox, T.F. and Cox, M.A.A. (1995), *Multidimensional Scaling.* Chapman and Hall, London.

Craigmile, P.F., Percival, D.B. and Guttorp, P. (2000), Wavelet-based parameter estimation for trend contaminated fractionally differenced processes. TRS Number 47, National Research Center for Statistics and the Environment, University of Washington.

Cressie, N. (1985), Fitting variogram models by weighted least squares. *Mathematical Geology* **17**, 563-586.

Cressie, N. (1989), Geostatistics. *The American Statistician* **43**, 197-202.

Cressie, N. (1993), *Statistics for Spatial Data.* Second edition, John Wiley, New York.

Cressie, N. and Chan, N.H. (1989), Spatial modeling of regional variables. *J. Amer. Statist. Assoc.* **84**, 393–401.

Cressie, N. and Glonek, G. (1984), Median based covariogram estimator reduce bias. *Statistics and Probability Letters* **2**, 299-304.

Cressie, N. and Hawkins, D.M. (1980), Robust estimation of the variogram I. *Mathematical Geology* **12**, 115-125.

Cressie, N., Kaiser, M.S., Daniels, M.J., Aldworth, J., Lee, J., Lahiri, S.N. and Cox, L.H. (1999), Spatial analysis of particulate matter in an urban environment. In *geoENV II — Geostatistics for Environmental Applications*, eds. Gómez-Hernández, J., Soares, A. and Froidevaux, R., Kluwer, Dordrecht, 41–52.

Creutin, J.D. and Obled, C. (1982), Objective analysis and mapping techniques for rainfall fields: an objective comparison. *Water Resources Research* **18**, 413–431.

Damian, D., Sampson, P.D. and Guttorp, P. (2001), Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics* **12**, 161–178.

Davis, J.C. (1973), *Statistics and Data Analysis in Geology.* Wiley, New York.

Davis, J.M., Nychka, D. and Bailey, B. (2000), A comparison of regional oxidant model (ROM) output with observed ozone data. *Atmospheric Environment* **34**, 2413–2423.

Davison, A.C. (1984), Modelling excesses over high thresholds, with an application. In *Statistical Extremes and Applications*, J. Tiago de Oliveira (ed.), Reidel, Dordrecht, 461-482.

Davison, A.C. and Smith, R.L. (1990), Models for exceedances over high thresholds (with discussion). *J.R. Statist. Soc.*, **52**, 393-442.

Dawid, A.P. (1981), Some matrix-valued distribution theory. *Biometrika* **68**, 265–274.

Dawid, A.P. and Sebastiani, P. (1999), Coherent dispersion criteria for optimal experimental design. *Ann. Statist.* **27**, 65–81.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion).

Diamond, P. and Armstrong, M. (1984), Robustness of variograms and conditioning of kriging matrices. *Mathematical Geology* **16**, 809-822.

Dickey, D.A. and Fuller, W.A. (1979), Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* **74**, 427–431.

Dickey, D.A. and Fuller, W.A. (1981), Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* **49**, 1057–1072.

Dickey, D.A., Bell, W.R. and Miller, R.B. (1986), Unit roots in time series: Tests and implications. *The American Statistician* **40**, 12–26.

Dickey, J.M. (1967), Matricvariate generalizations of the multivariate t distribution. *Ann. Math. Statist.* **38**, 511–518.

Diggle, P.J., Tawn, J. and Moyeed, R.A. (1998), Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299–350.

Donnelly, C.A. (1995), The spatial analysis of covatriates in a study of environmental epidemiology. *Statistics in Medicine* **14**, 2393–2409.

Donnelly, C.A., Ware, J.H. and Laird, N.M. (1994), Regression analysis fo spatially correlated data: The Kanawha County health study. In *Handbook of Statistics, Vol. 12: Environmental Statistics*, G.P. Patil and C.R. Rao (eds), North Holland Publishing Company, pp. 643–660.

Donoho, D.L. and Johnstone, I.M. (1994), Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Easterling, D.R., Horton, B., Jones, P.D., Peterson, T.C., Karl, T.R., Parker, D.E., Salinger, M.J., Razuvayev, V., Plummer, N., Jamason, P. and Folland, C.K. (1997), Maximum and minimum temperature trends for the globe. *Science* **277**, 364–367.

Efron, B. and Morris, C.M. (1973), Stein's esitmation rule and its competitors — an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117–130.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance.* Springer, New York.

Engle, R.F., Granger, C.W.J., Rice, J. and Weiss, A. (1986), Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc* **81**, 310–320.

Fairfield Smith, H. (1938), An empirical law describing heterogeneity in the yields of agricultural crops. *J. Agric. Sci.* **28**, 1–23.

Fedorov, V.V. (1972), *Theory of Optimal Experiments.* Academic Press, New York.

Fedorov, V. and Müller, W. (1988), Two approaches in optimization of observing networks. in *Optimal Design and Analysis of Experiments*, eds. Y. Dodge, V.V. Fedorov and H.P. Wynn, Elsevier Science, Amsterdam/New York.

Fedorov, V. and Müller, W. (1989), Comparison of two approaches in the optimal design of an observation network. *Statistics* **20**, 339-351.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Vol. I.* (3rd ed.) Wiley, New York.

Fink, A.M. (1988), How to polish off median polish. *SIAM Journal of Scientific and Statistical Computing* **9**, 932-940.

Fisher, R.A. and Tippett, L.H.C. (1928), Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.* **24**, 180–190.

Fuentes, M. (2001), Spectral methods for nonstationary spatial processes. *Biometrika*, to appear.

Fuentes, M. and Smith R.L. (2001), A new class of nonstationary spatial models. In preparation.

Furnival, G.M. and Wilson, R.W. Jr.(1974), Regression by leaps and bounds., *Technometrics* **16**, 499–511.

Galambos, J. (1987), *The Asymptotic Theory of Extreme Order Statistics* (2nd. edn.). Krieger, Melbourne, Fl. (First edn. published 1978 by John Wiley, New York.)

Galambos, J., Lechner, J. and Simiu, E. (eds.) (1994), *Extreme Value Theory and Applications.* Kluwer Academic Publishers, Dordrecht

Gaver, D. and O'Muircheartaigh, I. (1987), Robust empirical Bayes analysis of event rates. *Technometrics* **29**, 1–15.

Gelfand, A.E. and Ecker, M.D. (1997), Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological and Environmental Statistics* **2**, 347–369.

Gelfand, A.E. and Smith, A.F.M. (1990), Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Gelman, A. and Rubin, D.B. (1992), Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.

Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geweke, J. and Porter-Hudak, S. (1983), The estimation and application of long-memory time series models. *J. Time Series Anal.* **4**, 221-238.

Geyer, C.J. (1992), Practical Markov chain Monte carlo. *Statistical Science* **7**, 473–482.

Geyer, C.J. and Thompson, E.A. (1992), Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J.R. Statist. Soc. B 54*, 657–699.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds.) (1996), *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Gnedenko, B.V. (1943), Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44**, 423-453.

Granger, C.W.J. and Joyeux, R. (1980), An introduction to long-memory time series models and fractional differencing. *J. Time Series Anal.* **1**, 15-29.

Green, P.J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Green, P.J. and Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

Grenander, U. (1954), On estimation of regression coefficients in the case of an autocorrelated disturbance. *Ann. Math. Statist.* **25**, 252–272.

Gumbel, E.J. (1958), *Statistics of Extremes*. Columbia University Press, New York.

Guttorp, P., Meiring, W. and Sampson, P. (1994), A space-time analysis of ground-level ozone data. *Environmetrics* **5**, 241–254.

Guttorp, P. and Sampson, P. (1994), Methods for estimating heterogeneous spatial covariance functions with environmental applications. In *Handbook of Statistics 12*, eds. G.P. Patil and C.R. Rao, Elsevier Science B.V., 661–689.

Guttorp, P., Sampson, P.D. and Newman, K. (1992), Nonparametric estimation of spatial covariance with application to monitoring newtork evaluation. Chapter 3 of *Statistics in the Environmental and Earth Sciences*, eds. A. Walden and P. Guttorp, Halsted Press, New York.

Guyon, X. (1982), Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika* **69**, 95–105.

Haan, L. de and Resnick, S.I. (1977), Limit theory for multivariate sample extremes. *Z. Wahrscheinlichkeitstheorie v. Geb.* **40**, 317-337.

Haas, T.C. (1990a), Lognormal and moving-window methods of estimating acid deposition. *J. Amer. Statist. Assoc.* **85**, 950–963.

Haas, T.C. (1990b), Kriging and automated variogram modeling within a moving window, *Atmospheric Environment* **24A**, 1759–1769.

Haas, T.C. (1992), Redesigning continental-scale monitoring networks. *Atmospheric Environment* **26A**, 3323–3333.

Haas, T.C. (1995), Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Amer. Statist. Assoc.* **90**, 1189–1199.

Haas, T.C. (1998), Statistical assessment of spatio-temporal pollutant trends and meteorological transport models. *Atmospheric Environment* **32**, 1865–1879.

Hall, P. (1979), On the rate of convergence of normal extremes. *J. Appl. Prob.* **16**, 433–439.

Hall, P., Fisher, N.I. and Hoffman, B. (1994), On the nonparametric estimation of covariance functions. *Annals of Statistics* **22**, 2115–2134.

Halmos, P.R. (1972), Positive approximants of operators. *Indiana University Mathematics Journal* **21**, 951–960.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Handcock, M.S. and Wallis, J.R. (1994), An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *J. Amer. Statist. Assoc.* **89**, 368–390.

Handcock, M.S. and Stein, M. (1993), A Bayesian analysis of kriging. *Technometrics*, **35**, 403-410.

Hansen, J. and Lebedeff, S. (1987), Global trends of measured surface air temperatures. *J. Geophys. Research* **D92**, 13345–13372.

Hansen, J. and Lebedeff, S. (1988), Global surface air temperatures: update through 1987. *Geophys. Research Letters* **15**, 323–326.

Hartigan, J.A. and Wong, M.A. (1979), A K-means clustering algorithm. *Applied Statistics* **28**, 101–108.

Harville, D.A. (1974), Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

Harville, D.A. (1977), Maximum likelihood approaches to variance components estimation and to related problems. *J. Amer. Statist. Assoc.* **72**, 320–340.

Harville, D.A. and Jeske, D.R. (1992), Mean squared error of estimation or prediction under a general linear model. *J. Amer. Statist. Assoc.* **87**, 724-731.

Haslett, J. and Raftery, A.E. (1989), Space-time modelling with long-memory dependence: assessing Ireland's wind power resource (with discussion). *Applied Statistics* **38**, 1-50.

Hasselmann, K. (1997), Multi-pattern fingerprint method for detection and attribution of climate change. *Climate Dynamics* **13**, 601–611.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*. Chapman and Hall, London.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Healy, M.J.R. (1968), Algorithm AS6: Triangular decomposition of a symmetric matrix. *Applied Statistics* **17**, 195-197.

Hegerl, G.C. and North, G.R. (1997), Comparison of statistical optimal approaches to detecting anthropogenic climate change. *Journal of Climate* **10**, 1125–1133.

Hegerl, G., von Storch, H., Hasselmann, K., Santer, B.D., Cubasch, U. and Jones, P.D. (1996), Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *Journal of Climate* **9**, 2281–2306.

Higdon, D.M. (1998), A process-convolution approach to modeling temperatures in the north Atlantic Ocean. *J. Environ. Ecolo. Statist.* **5**, 173–190.

Higdon, D. (2001), Space and space-time modeling using process convolutions.

Preprint, Duke University.

Higdon, D., Swall, J. and Kern, J. (1999), Non-stationary spatial modeling. In *Bayesian Statistics 6*, eds. J.M. Bernardo *et al.*, Oxford University Press, pp. 761–768.

Higham, N.J. (1988), Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications* **103**, 103–118.

Holland, D.M., De Oliveira, V., Cox, L.H. and Smith, R.L. (2000), Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics*, to appear.

Holland, D., Saltzman, N., Cox, L.H. and Nychka, D. (1999), Spatial prediction of sulfur dioxide in the eastern United States. In *geoENV II — Geostatistics for Environmental Applications*, eds. Gómez-Hernández, J., Soares, A. and Froidevaux, R., Kluwer, Dordrecht, 65–76.

Holmström, I. (1963), On a method for parametric representation of the state of the atmosphere. *Tellus* **XV**, 127–149.

Hosking, J.R.M. (1981), Fractional differencing. *Biometrika* **68**, 165-176.

Hosking, J.R.M. (1984), Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research* **20**, 1898–1908.

Hosking, J.R.M. (1984), Testing whether the shape parameter is zero in the generalized extreme-value distribution. *Biometrika* **71**, 367-374.

Hosking, J.R.M. (1990), L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J.R. Statist. Soc. B* **52**, 105–124.

Hosking, J.R.M. and Wallis, J.R. (1997), *Regional Frequency Analysis: Ann Approach Based on L-Moments.* Cambridge University Press.

Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985), Estimation of the generalised extreme-value distribution by the method of probability-weighted moments. *Technometrics* **27**, 251-261.

Houghton, J.T. et al. (eds.) (1996), *Climate Change 1995 - The Science of Climate Change.* Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge Univ. Press, New York.

Hsing, T. (1988), On the extreme order statistics for a stationary sequence. *Stoch. Proc. Appl.* **29**, 155-169.

Hsing, T., Husler, J. and Leadbetter, M.R. (1988), On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields* **78**, 97-112.

Huang, L.-S. and Smith, R.L. (1997), Meteorologically-dependent trends in urban ozone. Submitted for publication.

Hughes, J.P. and Lettenmaier, D.P. (1981), Data requirements for kriging: Estimation and network design. *Water Resources Research* **17**, 1641–1650.

Hurvich, C.M. and Beltrao, K.I. (1993), Asymptotics for the low frequency ordinates of the periodogram of a long-memory time series. *Journal of Time Series Analysis* **14**, 455-472.

Hurvich, C.M. and Ray, B. (1995), Estimation of the memory parameter for nonstationary or noninvertible fractionally differenced time series models. *Journal of Time Series Analysis* **16**, 17–42.

Ickstadt, K. and Wolpert, R.L. (1999), Spatial regression for marked point processes. In *Bayesian Statistics 6*, eds. J.M. Bernardo *et al.*, Oxford University Press, pp. 323–341.

James, W. and Stein, C. (1961), Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press, 361–379.

Jaynes, E.T. (1963), Information theory and statistical mechanics.
In *Statistical Physics*, Vol. 3, K.W. Ford (ed.), Benjamin, New York, pp. 102–218.

Jenkinson, A.F. (1955), The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Q.J. Roy. Meteor. Soc.* **87**, 158-171.

Jenkinson, A.F. (1969), Statistics of Extremes. *WMO Technical Note No. 98*, Chapter 5, 183–227.

Joe, H., Smith, R.L. and Weissman, I. (1992), Bivariate threshold methods for extremes. *J.R. Statist. Soc. B.*, **54**, 171-183.

Johnson, M.E., Moore, L.M. and Ylvisaker, D. (1990), minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26**, 131–148.

Johnson, N.L. and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York.

Joly, A. *et al.* (1997), The Fronts and Atlantic Storm-Track Experiment (FASTEX): Scientific objectives and experimental design. *Bull. Amer. Meteor. Soc.* **78**, 1917–1940.

Journel, A.G. and Huijbregts, C.J. (1978), *Mining Geostatistics.* Academic Press, London.

Judd, K. and Mees, A.I. (1995), On selecting models for nonlinear time series. *Physica D* **82**, 426–444.

Karhunen, K. (1946), Zur Spektraltheorie stochastischer Prozesse. *Ann. Acad. Sci. Fennicae., Ser. A* **4**, 3–7.

Karhunen, K. (1947), Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae., Ser. A* **37**, 3–79.

Karl, T.R. and Knight, R.W. (1998), Secular trends of precipitation amount, frequency, and intensity in the USA. *Bull. Amer. Meteor. Soc.* **79**, 231–241.

Karl, T.R., Knight, R.W., Easterling, D.R. and Quayle, R.G. (1996), Indices of climate change for the United States. *Bull. Amer. Meteor. Soc.* **77**, 279–292.

Kattenberg, A., Giorgi, F., Grassl, H., Meehl, G.A., Mitchell, J.F.B., Stouffer, R.J., Tokioka, T., Weaver, A.J. and Wigley, T.M.L. (1996), Climate model projections of future climate. *Climate Change 1995: The Science of Climate Change*, J. Houghton, L.G. Meira Filho, B.A. Callander, N. Harris, A. Kattenberg, and K. Maskell, Eds., Cambridge University Press, New York, 285–357.

Kemperman, J.H.B. (1984), Least absolute value and median polish. In *Inequalities in Statistics and Probability*, Y.L. Tong (ed.), IMS monograph series #5.

Kendall, M.G. and Stuart, A. (1979), *The Advanced Theory of Statistics, Volume 2.* Fourth Edition, Griffin, London.

Kiefer, J. and Wolfowitz, J. (1960), The equivalence of two extremum problems. *Canad. J. Math.* **12**, 363–366.

Kitanidis, P.K. (1983), Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research* **19**, 909-921.

Ko, C.-W., Lee, J. and Queyranne, M. (1995), An exact algorithm for maximum entropy sampling. *Oper. Res.* **43**, 684–691.

Komaki, F. (1996), On asymptotic properties of predictive distributions. *Biometrika* **96**, 299–313.

Kosambi, D.D. (1943), Statistics in function space. *J. Indian Math. Soc.* **7**, 76–88.

Krige, D.G. (1951), A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Minining Society of South Africa* **52**, 119–139.

Krzanowski, W.J. and Lai, Y.T. (1988), A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* **44**, 23–34.

Künsch, H. (1986), Discrimination between monotonic trends and long-range dependence. *J. Appl. Probab.* **23**, 1025–1030.

Künsch, H. (1987), Statistical aspects of self-similar processes. in *Proceedings of the First World Congress of the Bernoulli Society*, VNU Science Press, pp. 67–74.

Kuo, C., Lindberg, C. and Thomson, D.J. (1990), Coherence established between atmospheric carbon dioxide and global temperature. *Nature* **343**, 709–714.

Le, N.D. and Zidek, J.V. (1992), Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. *Journal of Multivariate Analysis* **43**, 351–374.

Le, N.D. and Zidek, J.V. (1994), Network designs for monitoring multivariate random spatial fields. In *Recent Advances in Statistics and Probability*, edited by M.L. Puri and J.P. Vilaplana, pp. 191–206 (publisher??)

Le, N.D., Sun, W. and Zidek, J.V. (1997), Bayesian multivariate spatial interpolation with data missing by design. *J.R. Statist. Soc. B* **59**, 501–510.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983), *Extremes and Related Properties of Random Sequences and Series.* Springer Verlag, New York.

Lee, J. (1998), Discussion of Bueso *et al.* (1998). *Environmental and Ecological Statistics* **5**, 45–46. (Rejoinder: page 47.)

Lele, S. (1995), Inner product matrices, kriging, and nonparametric estimation of the variogram. *Math. Geology* **27**, 673–692.

Lewis, T. (1989), Discussion of the paper "Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource", by J. Haslett and A.E. Raftery. *Applied Statistics* **38**, 29.

Lindley, D.V. (1956), On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.

Lindley, D.V. and Smith, A.F.M. (1972), Bayes estimates for the linear model (with discussion). *J.R. Statist. Soc. B* **34**, 1–41.

Loader, C. and Switzer, P. (1989), Spatial covariance estimation for monitoring data. Chapter 4 of *Statistics in the Environmental and Earth Sciences*, eds. A. Walden and P. Guttorp, Halsted Press, New York, pp. 52–70.

Loève, M. (1945, 1946), Sur les fonctions aléatoire de second ordre. *Rev. Sci.* **83**, 297–303, **84**, 195–206.

Lorenz, E.N. and Emanuel, K.A. (1998), Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.* **55**, 399–414.

Lund, R., Hurd, H., Bloomfield, P. and Smith, R.L. (1995), Climatological time series with periodic correlation. *Journal of Climate* **8**, 2787–2809.

Mardia, K.V. and Goodall, C.R. (1993), Spatial-temporal analysis of mulitvariate environmental monitoring data. In *Multivariate Environmental Statistics*, eds. G.P. Patil and C.R. Rao, Elsevier Science Publishers, pp. 347–386.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979), *Multivariate Analysis.* New York: Academic Press.

Mardia, K.V. and Marshall, R.J. (1984), Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-146.

Mardia, K.V. and Watkins, A.J. (1989), On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289-295.

Marshall, A.W. and Olkin, I. (1967), A multivariate extremal distribution. *J. Amer. Statist. Assoc.* **62**, 30-44.

Manley, G. (1974), Central England temperatures: monthly means 1659 to 1973. *Quart. J. R. Met. Soc.* **100**, 389-405.

Matérn, B. (1986), *Spatial Variation.* Lecture Notes in Statistics, Number 36, Springer Varlag, New York. (Second edition: originally published in 1960).

Matheron, G. (1962), Traité de Géostatistique Appliquée, Tome I. *Memoires du Bureau de Recherche Géologiques et Minieres,* No. 14. Editions Technip, Paris.

Matheron, G. (1963a), Traité de Géostatistique Appliquée, Tome II: Le Krigeage. *Memoires du Bureau de Recherche Géologiques et Minieres,* No. 24. Editions Bureau de Recherche Géologiques et Minieres, Paris.

Matheron, G. (1963), Principles of Geostatistics. *Economic Geology* **58**, 1246–1266.

Matheron, G. (1971), *The Theory of Regionalized Variables and Its Applications.* Ecole des Mines, Fontainebleau.

McCoy, E. and Walden, A.T. (1996), Wavelet analysis and synthesis of stationary long-memory processes. *Journal of Computational and Graphical Statistics* **5**, 26–56.

Meinhold, R.J. and Singpurwalla, N.D. (1983), Understanding the Kalman filter. *The American Statistician* **37**, 123-127.

Meiring, W., Guttorp, P. and Sampson, P.D. (1998), Space-time estimation of grid-cell hourly ozone levels for assessment of a determinstic model. *Environmental and Ecological Statistics* **5**, 197–222.

Mercer, W.B. and Hall, M.A. (1911), The experimental error of fields trials. *J. Agric. Sci.* **4**, 54–62.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Monastiez, P. and Switzer, P. (1991), Semiparametric estimaion of nonstationary spatial covariance models by metric multidimensional scaling. *Tech. Rep. 165*, Department of Statistics, Stanford University.

Müller, P. (1999), Simulation based optimal design. In *Bayesian Statistics 6*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, xxx–xxx.

Müller, W.V. (2000), *Collecting Spatial Data: Optimum Design of Experiments for Random Fields.* Second edition, Physica Verlag, Heidelberg.

Müller, W.V. and Pázman, A. (1998), Design measures and approximate information measures for experiments without replications. *J. Statistical Planning and Inference* **71**, 349–362.

Müller, W.V. and Pázman, A. (1999), An algorithm for computation of optimum designs under a given covariance structure. *Computational Statistics* **14**(2), 197–211.

Müller, W.V. and Zimmerman, D.L. (1999), Optimal designs for variogram estimation. *Environmetrics* **10**, 23–37.

NERC (1975), *The Flood Studies Report.* The Natural Environment Research Council, London.

North, G.R. (1984), Empirical orthogonal functions and normal modes. *J. Atmos. Sci.* **41**, 879–887.

North, M. (1980), Time-dependent stochastic models of floods. *J. Hyd. Div. ASCE*, **106**, 649–655.

Nott, D.J. and Dunsmuir, W.T.M. (2000), Analysis of spatial covariance structure for Sydney wind patterns. Preprint, University of New South Wales, Australia.

Nychka, D., Buchberger, R., Wigley, T.M.L., Santer, B.D., Taylor, K.E. and Jones, R.H. (2000), Confidence intervals for trend estimates with autocorrelated observations.

Preprint, Geophysical Statistics Project, National Center of Atmospheric Research, Boulder, CO.

Nychka, D., Piegorsch, W.W. and Cox, L.H. (eds.) (1998), *Case Studies in Environmental Statistics*. Springer Lecture Notes in Statistics, number 132, Springer Verlag, New York.

Nychka, D. and Saltzman, N. (1998), Design of air quality networks. In *Case Studies in Environmental Statistics*, eds. D. Nychka, W. Piegorsch and L.H. Cox, Lecture Notes in Statistics number 132, Springer Verlag, New York, pp. 51–76.

Nychka, D., Wikle, C. and Royle, J.A. (1999), Large spatial prediction problems and nonstationary random fields. Preprint, Geophysical Statistical Program, National Center for Atmospheric Research.

Obukhov, A.M. (1947), Statistical homogeneous random fields on a sphere. *Uspekhi Mat. Nauk.* **2**, 196–198.

Obukhov, A.M. (1954), Statistical description of continuous fields. *Trudy Geofiz. Inst. Akad. Nauk. SSSR*, **24(151)**, 3–42.

Oehlert, G.W. (1993), Regional trends in sulfate wet deposition. *Journal of the American Statistical Association* **88**, 390–399.

Oehlert, G.W. (1995), The ability of wet decomposition networks to detect temporal trends. *Environmetrics* **6**, 327–339.

Oehlert, G.W. (1996), Optimal shrinking of a wet decomposition network. *Atmospheric Environment* **30**, 1347–1357.

Omre, H. (1987), Bayesian kriging — merging observations and qualified guesses in kriging. *Mathematical Geology* **19**, 25–39.

Omre, H. and Halvorsen, K.B. (1989), The Bayesian bridge between simple and universal kriging. *Mathematical Geology* **21**, 767–786.

Omre, H., Halvorsen, K.B. and Bertig, V. (1989), A Bayesian approach to kriging. In *Geostatistics*, Vol. 1, Ed. M. Armstrong. Kluwer Academic, Norwell, MA.

Palmer, T.N., Gelaro, R., Barkmeijer, J. and Buizza, R. (1998), Singular vectors, metrics, and adaptiv observations. *J. Atmos. Sci.* **55**, 633–653.

Papadakis, J.S. (1937), Méthode statistique pour les expériences du champ. *Bull. Inst. Amél. Plantes à Salonique*, No. 23.

Patefield, W.M. (1977), On the maximized likelihood function. *Sankhyā B* **39**, 92–96 (correction p. 409).

Patterson, H.D. and Thompson, R. (1971), Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554.

Pázman, A. and Müller, W.V. (1998), A new interpretation of design measures. In *Model-Oriented Data Analysis 5*, eds. A.C. Atkinson, L. Pronzato and H.P. Wynn. Physica Verlag, Heidelberg.

Pázman, A. and Müller, W.V. (2000), Optimal design of experiments subject to correlated errors. *Statistics and Probability Letters*, to appear.

Penttinen, A. (1984), Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method. *Jy. Stud. Comput. Sci. Econ. Statist.* **7**.

Perrin, O. and Meiring, W. (1999), Identifiability for non-stationary spatial structure. *J. Appl. Prob.* **36**, 1244–1250.

Pesti, G., Kelly, W.E. and Bogardi, I. (1994), Observation network design for selecting locations for water supply wells. *Environmetrics* **5**, 91–110.

Pickands, J. (1975), Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Pickard, D.K. (1987), Inference for discrete Markov fields: the simplest nontrivial case. *J. Amer. Statist. Assoc.* **82**, 90–96.

Pinheiro, J.C. and Bates, D.M. (2000), *Mixed-effects models in S and S-PLUS.* Springer Verlag, New York.

Prasad, N.G.N. and Rao, J.N.K. (1990), The estimation of the mean-squared error of small-area estimators. *J. Amer. Statist. Assoc.* **85**, 163-171.

Prescott, P. and Walden, A.T. (1980), Maximum likelihood estimation of the parameters of the generalized extreme value distribution. *Biometrika* **67**, 723–724.

Press, S.J. (1982), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference.* Krieger, Melborne, FL.

Press, S.J. (1989), *Bayesian Statistics: Principles, Models and Applications.* Wiley, New York.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986), *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge.

Putter, H. and Young, G.A. (2001), On the effect of covariance function estimation on the accuracy of kriging predictors. *Bernoulli* **7**, 421–438.

Rao, C.R. (1979), MINQE theory and its relation to ML and MML estimation of variance components. *Sankhyā B* **41**, 138–153.

Raper, S.C.B., Wigley, T.M.L. and Warrick, R.A. (1996), Global sea level rise: past and future. *Sea-level rise and coastal subsidence: causes, consequences and strategies,* J.D. Milliman, B.U. Haq, Eds., Kluwer Academic Publishers, 11–45.

Reiss, R.-D. and Thomas, M. (1997), *Statistical Analysis of Extreme Values.* Birkhäuser, Basel.

Resnick, S. (1987), *Extreme Values, Point Processes and Regular Variation.* Springer Verlag, New York.

Ripley, B.D. (1981), *Spatial Statistics.* Wiley, New York.

Ripley, B.D. (1988), *Statistical Inference for Spatial Processes.* Cambridge University press, Cambridge, U.K.

Robinson, M.E. and Tawn, J.A. (1995), Statistics for exceptional athletics records. *Applied Statistics* **44**, 499–511.

Robinson, M.E. and Tawn, J.A. (1996), Statistics for extreme sea currents. *Applied Statistics* **45**, 183–205.

Robinson, P.M. (1995a), Log-periodogram regression of time series with long range dependence. *Ann. Statist.* **23**, 1048–1072.

Robinson, P.M. (1995b), Gaussian estimation of long range dependence. *Ann. Statist.* **23**, 1630-1661.

Royle, J.A. and Nychka, D. (1998), An algorithm for the construction of spatial coverage designs with implementation in S-PLUS. *Computers and Geosciences* **24(5)**, 479–488.

Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989), Design and analysis of computer experiments. *Statistical Science* **4**, 409–435.

Sacks, J. and Ylvisaker, D. (1966), Design for regression problems with correlated errors. *Ann. Math. Statist.* **37**, 66–89.

Sacks, J. and Ylvisaker, D. (1968), Design for regression problems with correlated errors; many parameters. *Ann. Math. Statist.* **39**, 46–69.

Sacks, J. and Ylvisaker, D. (1970), Design for regression problems with correlated errors III. *Ann. Math. Statist.* **41**, 2057–2074.

Sampson, P.D. and Guttorp, P. (1992), Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87**, 108-119.

Sampson, P.D., Lewis, P., Guttorp, P., Bookstein, F.L. and Hurley, C. (1991), Computation and interpretation of deformations for landmark data in morphometrics and environmetrics. *Proc. 23d Symposium on the Interface between Computing Science and Statistics.* Interface Foundation, Fairfax Station, 534–541.

Sansó, B. and Guenni, L. (1997), A Bayesian estimation of the parameters of a space-time model for rainfall. Submitted.

Sansó, B. and Müller, P. (1997), Redesigning a network of rainfall stations. ISDS Discussion Paper 97-25, Duke University.

Santer, B.D., Taylor, K.E., Wigley, T.M.L., Johns, T.C., Jones, P.D., Karoly, D.J., Mitchell, J.F.B., Oort, A.H., Penner, J.E., Ramaswamy, V., Schwarzkopf, M.D., Stouffer, R.J. and Tett, S. (1996), A search for human influences on the thermal structure of the atmosphere. *Nature* **382**, 39–46.

Schmidt, A.M. and O'Hagan, A. (2000), Bayesian inference for nonstationary spatial covariance structure via spatial deformations. Preprint, University of Sheffield.

Schumacher, P. and Zidek, J.V. (1993), Using prior information in designing intervention detecting experiments. *Ann. Statist.* **21**, 447–463.

Searle, S.R. (1970), Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrics* **26**, 505–524.

Sebastiani, P. and Wynn, H.P. (2000), Maximum entropy sampling and optimal Bayesian experimental design. *J.R. Statist. Soc. B* **62**, 145–157.

Shapiro, A. and Botha, J.D. (1991), Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics and Data Analysis* **11**, 87–96.

Shewry, M.C. and Wynn, H.P. (1987), Maxiimum entropy sampling. *Applied Statistics* **14**, 165–170.

Shively, T.S. (1991), An analysis of the trend in ground-level ozone using nonhomogeneous Poisson processes. *Atmospheric Environment* **25B**, 387–396.

Silvey, S.D. (1980), *Optimal Design.* Chapman and Hall, London.

Smith, A.F.M. and Roberts, G.O. (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J.R. Statist. Soc. B* **55**, 3–23.

Smith, L.A. (1992), Identification and prediction of low-dimensional dynamics. *Physica D* **58**, 50-76.

Smith, R.L. (1984), Threshold methods for sample extremes. In *Statistical Extremes and Applications*, J. Tiago de Oliveira (ed.), Reidel Dordrecht, 621–638.

Smith, R.L. (1985), Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67-92.

Smith, R.L. (1986), Extreme value theory based on the r largest annual events. *J. Hydrology* **86**, 27-43.

Smith, R.L. (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science* **4**, 367-393.

Smith, R.L. (1990), Extreme value theory. In *Handbook of Applicable Mathematics* **7**, ed. W. Ledermann, John Wiley, Chichester. Chapter 14, pp. 437-471.

Smith, R.L. (1993), Long-range dependence and global warming. In *Statistics for the Environment* (V. Barnett and F. Turkman, eds.), John Wiley, Chichester, 141–161.

Smith, R.L. (1994a), Multivariate threshold methods. In *Extreme Value Theory and Applications*, eds. J. Galambos, J. Lechner and E. Simiu. Kluwer Academic Publishers, Dordrecht, pp. 225–248.

Smith, R.L. (1994b), Spatial modelling of rainfall data. In *Statistics for the Environment*, Volume 2 (V. Barnett & F. Turkman, editors), pp. 19-41. Chichester: John Wiley.

Smith, R.L. (1996), Estimating nonstationary spatial correlations. Preprint, University of North Carolina.

Smith, R.L. (1997a), Statistics for exceptional athletics records: Letter to the editor. *Applied Statistics* **46**, 123–127.

Smith (1997b), R.L. Introduction to the paper by J. Besag (1974), Spatial interaction and the statistical analysis of lattice systems. In *Breakthroughs in Statistics III*, edited by S. Kotz and N.L.Johnson. Springer Verlag, New York, 1997, pp. 285–291.

Smith, R.L. (1997c), Predictive inference, rare events and hierarchical models. Preprint, University of North Carolina.

Smith, R.L. (1999), Trends in rainfall extremes. Preprint, University of North Carolina.

Smith, R.L. and Chen F.-L., (1996), Regression in long-memory time series. In *Athens Conference on Applied Probability and Time Series, Volume II: Time Series Analysis in Memory of E.J. Hannan*, edited by P.M. Robinson and M. Rosenblatt, Springer Lecture Notes in Statistics 115, 378–391.

Smith, R.L. and Shively, T.S. (1995), A point process approach to modeling trends in tropospheric ozone *Atmospheric Environment* **29**, 3489–3499.

Smith, R.L., Tawn, J.A. and Coles, S.G. (1997), Markov chain models for threshold exceedances. *Biometrika* **84**, 249–268.

Smith, R.L., Tawn, J.A. and Yuen, H.K. (1990), Statistics of multivariate extremes. *International Statistical Review*, **58**, 47-58.

Smith, R.L. and Weissman, I. (1994), Estimating the extremal index. *J.R. Statist. Soc. B* **56**, 515–128.

Smith, R.L., Wigley, T.M.L. and Santer, B.D. (2001), A bivariate time series approach to anthropogenic trend detection in hemispheric mean temperatures. Tentatively accepted for *Journal of Climate*.

Snyder, C. (1996), Summary of an informal workshop on adaptive observations and Fastex. *Bull. Amer. Meteor. Soc.* **77**, 953–961.

Sposito, V.A. (1987), On median polish and L_1 estimators. *Computational Statistics and Data Analysis* **5**, 155-162.

Stein, M.L. (1987), Minimum norm quadratic estimation of spatial variograms. *J. Amer. Statist. Assoc.* **82**, 765-772.

Stein, M.L. (1988), Asymptotically efficient spatial interpolation with a misspecified covariance function. *Ann. Statist.* **16**, 55-63.

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging.* Springer Verlag, New York.

Stein, M.L. and Handcock, M.S. (1989), Some asymptotic properties of kriging when the covariance function is misspecified. *Mathematical Geology* **21**, 171-190.

Stern, R.D. and Coe, R. (1984), A model fitting analysis of daily rainfall data (with discussion). *J.R. Statist. Soc. A* **147**, 1-34.

Su, Y. and Cambanis, S. (1994), Sampling designs for regression coefficient estimation with correlated errors. *Ann. Inst. Statist. Math.* **46**, 707–722.

Sun, L., Zidek, J.V., Le, N.D. and Özkaynak, H. (2000), Interpolating Vancouver's daily ambient PM_{10} field. *Environmetrics* **11**, 651–663.

Swendsen, R.H. and Wang, J.-S. (1987), Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Letters* **58**, 86–88.

Switzer, P. (1989), Non-stationary spatial covariances estimated from monitoring data. In *Geostatistics*, Vol. 1, Ed. M. Armstrong. Kluwer Academic, Norwell, MA.

Tawn, J.A. (1988a), An extreme-value theory model for dependent observations. *J. Hydrology* **101**, 227–250.

Tawn, J.A. (1988b), Bivariate extreme value theory - models and estimation. *Biometrika* **75**, 397-415.

Tawn, J.A. (1990), Modelling multivariate extreme value distributions. *Biometrika* **77**, 245-253.

Tett, S.F.B., Stott, P.A., Allen, M.R., Ingram, W.J. and Mitchell, J.F.B. (1999), Causes of twentieth-century temperature change near the Earth's surface. *Nature* **399**, 569–572 (June 10, 1999).

Thiébaux, H.J. and Pedder, M.A. (1987), *Spatial Objective Analysis: With Applications in Atmospheric Sciences.* Academic Press, San Diego.

Tiago de Oliveira, J. (ed.) (1984), *Statistical Extremes and Applications*, Reidel, Dordrecht.

Tibshirani, R. (1995), Regression selection and shrinkage via the *lasso. J.R. Statist. Soc. B* **58**, 267–288.

Tierney, L. (1994), Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–1728.

Trujillo-Ventura, A. and Ellis, J.H. (1991), Multiobjective air pollution monitoring network design. *Atmospheric Environment* **25**, 469–479.

Upton, G.J.G. and Fingleton, B. (1985), *Spatial Data Analysis by Example, Volume I: Point Pattern and Quantitative Data.* Wiley, Chichester.

Upton, G.J.G. and Fingleton, B. (1989), *Spatial Data Analysis by Example, Volume II: Categorical and Directional Data.* Wiley, Chichester.

Ver Hoef, J.M. and Barry, R.P. (1999), Constructing and fitting models for cokriging and multivariable spatial prediction. *J. Statist. Plann. Inference* **69**, 275–294.

Ver Hoef, J.M., Cressie, N. and Barry, R. (2000), Flaxible spatial models based on the fast Fourier transform (FFT) for cokriging. Technical report, Department of Statistics, Ohio State University.

Waller, L.A., Carlin, B.P., Xia, H. and Gelfand, A.E. (1997), Hierarchical spatio-temporal mapping of disease rates. *J. Amer. Statist. Assoc.* **92**, 607–617.

Warnes, J.J. (1986), A sensitivity analysis for universal kriging. *Mathematical Geology* **18**, 653-676.

Warnes, J.J. and Ripley, B.D. (1987), Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika* **74**, 640-642.

Warrick, A.W. and Myers, D.E. (1987), Optimization of sampling locations for variogram calculations. *Water Resources Research* **23**, 496–500.

Whittle, P. (1954), On stationary processes in the plane. *Biometrika* **41**, 434–439.

Wigley, T.M.L. and Raper, S.C.B. (1990), Natural variability of the climate system and detection of the greenhouse effect. *Nature* **344**, 324–327.

Wigley, T.M.L. and Raper, S.C.B. (1991), Internally generated variability of global-mean temperatures. In M.E. Schlesinger (ed.), *Greenhouse-Gas-Induced Climatic Change: A Critical Appraisal of Simulations and Observations*, Elsevier Science Publishers, Amsterdam, The Netherlands, pp. 471–482.

Wigley, T.M.L., and Raper, S.C.B. (1992), Implications for climate and sea level of revised IPCC emissions scenarios. *Nature*, **357**, 293–300.

Wigley, T.M.L., Smith, R.L. and Santer, B.D. (1998), Anthropogenic influence on the autocorrelation function of hemispheric-mean temperatures. *Science*, **282**, 1676–1679.

Woodward, W.A., and Gray, H.L. (1993), Global warming and the problem of testing for a trend in time series analysis. *Journal of Climate*, **6**, 953–962.

Woodward, W.A. and Gray, H.L. (1995), Selecting a model for detecting the presence of a trend. *Journal of Climate*, **8**, 1929–1937.

Wu, S. and Zidek, J.V. (1992), An entropy-based analysis of data from selected NADP/NTN network sites for 1983–1986. *Atmospheric Environment* **26A**, 2089–2103.

Yaglom, A.M. (1987), *Correlation Theory of Stationary and Related Random Functions, I and II.* Springer Verlag, New York.

Yajima, Y. (1988), On estimation of a regression model with long-memory stationary errors. *Ann. Statist.* **16**, 791–807.

Yajima, Y. (1991), Asymptotic properties of the LSE in a regression model with long-memory stationary errors. *Ann. Statist.* **19**, 158–177.

Zhu, L., Carlin, B.P., English, P. and Scalf, R. (2000), Hierarchical modeling of spatio-temporally misaligned data: Relating traffic density to pediatric asthma hospitalizations. *Environmetrics* **11**, 43–61.

Zidek, J.V. (1997), Interpolating air pollution for health impact assessment. In *Statistics for the Environment 3: Pollution Assessment and Control*, edited by V. Barnett and K.F. Turkman. Wiley, Chichester, pp. 251–268.

Zidek, J.V., Sun, W. and Le, N.D. (2000), Designing and integarting composite networks for monitoring multivariate Gaussian pollution fields. *Applied Statistics* **49**, 63–79.

Zimmerman, D.L. and Cressie, N. (1992), Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.* **44**, 27-43.

Zimmerman, D.L. and Zimmerman, M.B. (1991), A comparison of spatial variogram estimators and corresponding ordinary kriging predictors. *Technometrics* **33**, 77-91.

Zwiers, F. and von Storch, H. (1995), Taking serial correlation into account in tests of the mean. *Journal of Climate* **8**, 336–351.

Zygmund, A. (1959), *Trignometric Series.* Second Edition, Cambridge University Press, Cambridge.