NOTE TO USERS

DISSERTATION

STATISTICAL MODELS FOR QUANTIFYING THE SPATIAL DISTRIBUTION OF SEASONALLY DERIVED OZONE STANDARDS

Submitted by

Eric Gilleland

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

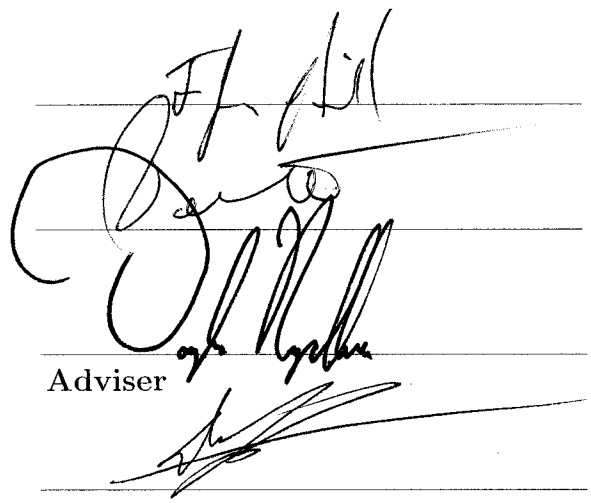
Spring 2005

COLORADO STATE UNIVERSITY

November 5, 2004

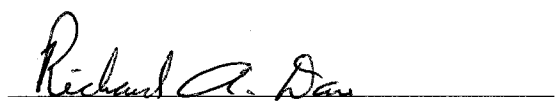
WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY ERIC GILLELAND ENTITLED STATISTICAL MODELS FOR QUANTIFYING THE SPATIAL DISTRIBUTION OF SEASONALLY DERIVED OZONE STANDARDS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work



Adviser

Co-Adviser



Department Chair

ABSTRACT OF DISSERTATION

STATISTICAL MODELS FOR QUANTIFYING THE SPATIAL DISTRIBUTION OF SEASONALLY DERIVED OZONE STANDARDS

The U.S. Environmental Protection Agency's (EPA) National Ambient Air Quality Standard (NAAQS) for ground-level ozone is now based on the fourth-highest daily maximum 8-hour average ozone level (FHDA). Standard geostatistical models may not be appropriate for interpolating such a statistic off of a network of monitoring sites. The performance of different statistical models in predicting this standard at locations where monitors are not located is compared. Special attention is given to two models: a daily model that uses a spatial autoregression to account for spatial and temporal dependence, and a seasonal model that assumes the FHDA field is Gaussian and employs spatial statistical techniques. Based on five seasons of ozone data collected in and around North Carolina, cross-validation shows a preference to the daily model over the seasonal model. In addition to the above models, a spatial extreme value model is also compared to the daily model. Results show that the two vastly different methods give remarkably similar results.

Eric Gilleland
Statistics Department
Colorado State University
Fort Collins, CO 80523
Spring 2005

ACKNOWLEDGMENTS

The work presented here was funded through the National Science Foundation under grants DMS-9312686 and DMS-9815344. I would like to give special thanks to my parents, Karen and Paul Gilleland, and friends for their encouragement, support, and for keeping me focused; with special thanks to Mike Chornack, Rudy Horne and Mike Spowart for all of their advice and for continually motivating, distracting, and pushing me to succeed. I would also like to thank all of the following good people, in no particular order, who helped me to persevere by way of distraction (Ik wol allegearre betankje dy't my genôch ôflaat hawwe om troch te setten): Chad Franzen, Chance McCarty, Maputo Mensah and Logo Ligi, Piter Wilkens, Jans Ingber and The Motet, Lara Maykovich, It Willehúske, la table francophone, and everyone at the CU recreation center. More directly related, I would like to thank Duane Boes who helped prepare me early on for the candidacy exams. I would also like to thank Barb Brown, and the rest of the forecast verification group at the Research Applications Laboratory of the National Center for Atmospheric Research (NCAR) for motivating me to finish. Special thanks to everyone associated with the Geophysical Statistics Project at NCAR, and especially Tim Hoar for all of his computer help from the beginning to the end. I cannot imagine finishing this degree without all of the help and support from so many of these people. Finally, I could not have done this without help and advice from my adviser, Doug Nychka, and everyone on my committee.

Contents

Chapter 1

Introduction

Statistical theory for the analysis of extreme values is well established for the univariate case, but for the multivariate case, it is still an area of active research. Most work on multivariate extremes has been focused on the maximums, or equivalently, the minimums. The work here is focused on making spatial inferences of a fourth-highest order statistic. The motivation for looking at such a seemingly esoteric statistic comes from a practical application; the U.S. Environmental Protection Agency's (EPA) regulatory standard for ground-level ozone. The standard states that if the three-year average fourth-highest daily maximum 8-hour ozone level (FHDA), monitored during the "ozone season" from about May through October, exceeds 84 parts per billion (ppb) for a particular consolidated/metropolitan statistical area (C/MSA), then the region included within this C/MSA is designated as out of attainment.

Three approaches for making spatial inferences of the fourth-highest order statistic are considered. The first, which will be referred to here as the daily model, uses an autoregression with spatially correlated shocks to simulate a sample from the fourth-highest multivariate distribution by way of Monte Carlo simulations. The second approach, referred to here as the seasonal model, assumes that the multivariate distribution for the fourth-highest order statistic, from a sample of size 184, is approximately multivariate normal; and standard geostatistical techniques are used to predict the regulatory standard at locations without monitoring stations. Finally, I will introduce a new method for modeling multivariate exceedances over the threshold using the usual univariate generalized Pareto distribution while including a spatial component that links the distribution over space.

Chapter 2 provides some background on the U.S. EPA's air quality standard for ground level ozone, giving motivation to the work set out here. Chapter 1 will then provide a literature review of previous relevant work including: spatial (geostatistical) models, thin plate splines, first order autoregressive time series models, spatiotemporal models, extreme value theory, previous findings about ozone data, and finally a section on work relating to the new 8-hour ozone standard. Chapter 4 introduces the daily and seasonal models, and compares results of predicting fourth-highest values spatially. The following chapter considers a different analysis based on extreme value theory.

Chapter 2

Motivation

In this chapter I give details about the motivation for looking at the multivariate distribution of the fourth-highest order statistic. Section 2.1 introduces the U.S. EPA's role in trying to improve air quality. Section 2.2 gives health and environmental motivation of the importance of reducing ground-level ozone pollution. Section 2.3 discusses how the monitoring and regulation program works. Section 2.4 discusses the new regulatory standard for ground-level ozone, and finally, section 2.5 discusses the specific statistical problem arising from this new regulatory standard.

2.1 Ground-level Ozone

As required by the Clean Air Act (CAA) of 1971, the EPA has established standards, known as the National Ambient Air Quality Standards (NAAQS),

for six principal air pollutants (also referred to as criteria pollutants): carbon monoxide (CO), lead (Pb), nitrogen dioxide (NO_2), ground-level ozone (O_3), particulate matter (PM) and sulfur dioxide (SO_2): to monitor and control their ambient concentrations. Of these six, ground-level ozone has a new standard based on an order statistic that will be explored in this thesis.

Although ground-level ozone (O_3) is the primary constituent of smog, it is not directly emitted into the atmosphere; rather it is formed when nitrogen oxides (NO_x) and volatile organic compounds (VOCs) react in the presence of sunlight. This catalysis creates an ozone pollution problem particularly in the hot summer months. VOCs are emitted from a variety of sources including cars, factories, refineries, chemical plants, consumer products and many other industrial sources. NO_x are emitted from motor vehicles, power plants and other sources of internal combustion. Most surface ozone is locally produced from these precursors through photochemical processes. A typical diurnal pattern of surface ozone concentration during the ozone season consists of a mid-afternoon maximum, followed by a decay to an early morning minimum and then a rise through the late morning and the middle part of the day [12]. O_3 and its precursor pollutants can easily be transported hundreds of miles from pollution sources. Subsequently, changing weather patterns can attribute much variability in yearly ozone concentrations in a given region [44]. One context for this thesis research is the interest in quantifying ozone pollution on spatial scales beyond the size of an urban area.

2.2 Health and Environmental Effects of Ground-level Ozone

Exposure to ambient ozone for both short (1-3 hours) and long (6-8 hours) amounts of time has been linked to various negative health effects including increased hospital admissions and respiratory problems. Specifically, it can make people more susceptible to respiratory infection, result in lung inflammation, cause a decrease in lung function, increase respiratory symptoms such as chest pain and cough as well as aggravate any pre-existing respiratory diseases such as asthma or chronic lung disease [44].

These problems typically occur for people who are active outdoors, and is especially a concern for children because it puts them at particularly high risk during the summer months when ozone levels are at their highest and children are more likely to be playing outside. Longer-term exposure to moderate levels of ozone can cause irreversible changes in the lungs, leading to premature aging of the lungs, and worsen or cause chronic respiratory illnesses [44].

Vegetation and ecosystems are also affected by ground-level ozone leading to reductions in agriculture and commercial forest yields, reduced growth and survivability of tree seedlings and increased plant susceptibility to disease, pests and other environmental stresses. Over a long period of time, this can have a significant impact on ecosystems and habitats for wildlife, particularly on endangered species [44].

2.3 Monitoring and Regulating Ground-level Ozone

The EPA is involved in both monitoring and regulating emissions of the six criteria pollutants. In part, this means setting up monitoring stations and using the observations obtained from them to apply the NAAQS for each pollutant. Another function of the EPA in this context involves investigating ways of reducing emissions by studying where and how the emissions are made, and how they are transported, and either react or dissipate. Initially, regions referred to as consolidated/metropolitan statistical areas (C/MSA)–in which monitoring and regulation is to take place–are defined by the Office of Management and Budget. States and tribes provide the EPA with their attainment/nonattainment designation recommendations for each C/MSA area, and the EPA can either approve or disapprove them, and then promulgate the designation. Designations are based on monitoring data as well as data indicating whether an area (possibly outside a C/MSA) contributes to a violation. As part of this process, the EPA may consider various factors, such as emissions, traffic and commuting patterns, population density, and expected growth. In particular, if an area is not part of a C/MSA, it may still be found to be out of attainment of the NAAQS if it is believed to contribute to the ozone problem of a C/MSA region found to be out of attainment. However, there is currently no means for the EPA to provide the public with ozone coverage levels for such an area.

Once a C/MSA is decided upon, monitors are placed in the region, and any C/MSA found to be out of attainment must create an implementation plan, called a State Implementation Plan (SIP), with the intent of reducing ozone pollution to meet the standard.

2.3.1 Air Quality Monitoring

The EPA's ambient air quality monitoring program is carried out by local, state and national agencies. The program consists of three major categories of monitoring networks: State and Local Air Monitoring Stations (SLAMS), National Air Monitoring Stations (NAMS) and Special Purpose Monitoring Stations (SPMS): that measure the criteria pollutants, as well as a fourth network of Photochemical Assessment Monitoring Stations (PAMS) that measure ozone precursors. The SLAMS network consists of roughly 4,000 monitoring stations whose size and placement is largely determined by state and local air pollution control agencies to meet their respective SIP requirements. The 1,000-plus NAMS are a subset of the SLAMS network and are key sites emphasizing areas of maximum concentrations and high population density. The SPMS are used for special studies by state and local agencies to support SIPs and other air program activities. These monitors are not permanently established and may be easily changed as needed. As mandated in a 1990 amendment to the CAA, PAMS networks are required in each ozone non-attainment area that is designated serious, severe or extreme. Here, data from a network of 513

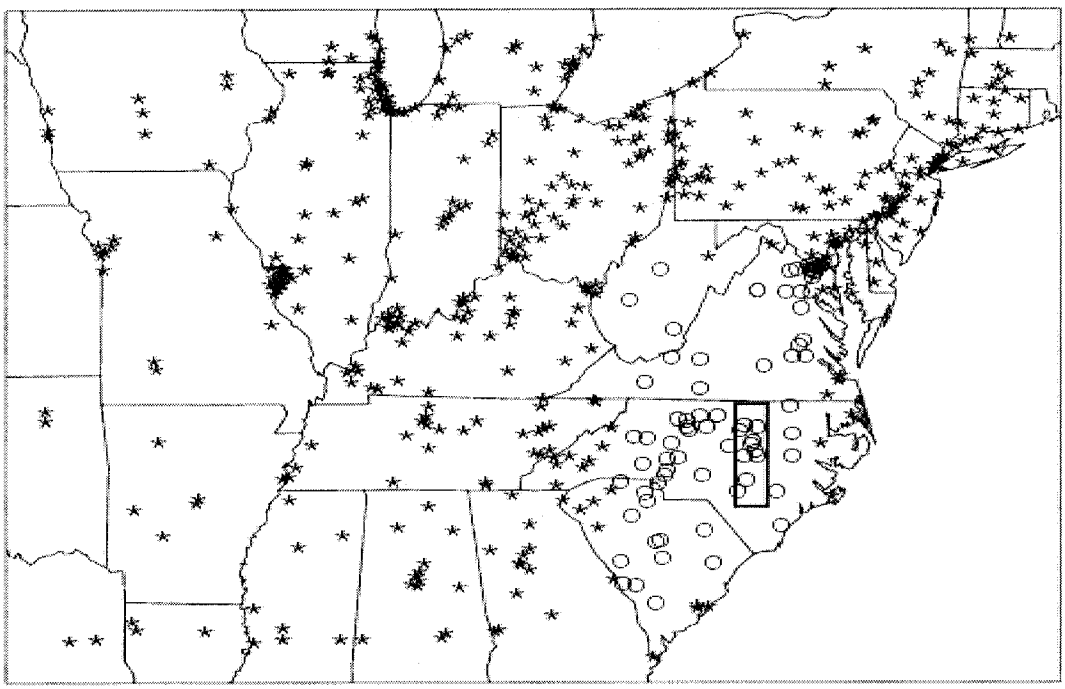


Figure 2.1: 513 ozone monitoring stations in the eastern United States. The rectangle in North Carolina represents a grid area in the Research Triangle Park (RTP) in which interpolations will be made for comparison of the daily and seasonal models for the subset of 72 stations (circled) around North Carolina.

SLAMS stations in the eastern United States (Figure 2.1) are analyzed, with special attention given to a 72 station subset in and around North Carolina. These data are posted at `http://www.cgd.ucar.edu/stats/Data/O3.shtml` along with R programs for their analysis.

Because of the high cost of operating these stations, ozone monitoring occurs only during the hotter months when weather conditions are most conducive to forming ozone. This "ozone season" generally goes from about April to October. A great deal of care is placed in the design of the network in order

to optimize, in some way, the results with as few monitoring stations as possible (see Nychka *et. al* [32]). Strict criteria have been established by the EPA regarding the operation of these monitoring networks including calibrations, independent audits, data validation and a rigorous quality assurance program (see Davis *et al.* [7]).

The instruments used are either chemiluminescence analyzers or ultraviolet photometers and the brand and model must be approved before use. The measurements can be taken continuously over time, but the output is usually sampled at short time intervals and, as is the case for ozone, these discrete measurements are averaged over blocks of time in each day.

Placement of monitors is typically determined by physical and political means. For example, Guttorp *et al.* [17] note that local NO_x sources, such as power plants, cement plants and traffic can hinder the ability of monitors to accurately measure ozone because its representativeness is destroyed by such sources or sinks that dominate the ozone measurements. Additionally, because ozone can be transported easily by wind, monitors placed downwind of an ozone source will perhaps obtain a better measurement of actual ozone presence than those placed upwind.

2.3.2 Numerical Models for Simulating and Predicting Air Quality

The primary tool used by the EPA to investigate solutions to pollution problems is by numerical models based on physical and chemical principles. The Community Multi-scale Air Quality (CMAQ) modeling system has been designed to approach air quality as a whole by including advanced capabilities for modeling multiple air quality issues. Tropospheric (ground-level) ozone, fine particles, toxics, acid deposition and visibility degradation are all incorporated into the CMAQ. Additionally, CMAQ was designed to have multi-scale capabilities negating the need for separate urban and regional scale air quality modeling. Target grid resolutions and domain sizes for CMAQ range both spatially and temporally over several orders of magnitude allowing simulations to be performed to evaluate long term pollutant climatologies as well as short term transport from localized sources [43].

Implementation of multi-scale capabilities in CMAQ requires that several issues be resolved; such as scalable atmospheric dynamics and generalized coordinates depending on the desired model resolution. One example of a difference in assumptions for urban and regional scales is in hydrostatic conditions, which may be assumed to be balanced over vertical pressure and gravitational force with no net vertical acceleration for large scales, but not for small scales. Thus, CMAQ's governing equations are expressed in a generalized coordinate

system, ensuring consistency between CMAQ and the meteorological modeling system. The CMAQ modeling system contains three types of modeling components: a meteorological modeling system for describing atmospheric states and motions, emission models for man-made and natural emissions injected into the atmosphere, and a chemistry-transport modeling system for simulating the chemical reactions and fate [43].

2.4 The NAAQS for Ground-level Ozone

The NAAQS for ozone from 1978 until 1997 was based on daily 1-hour blocks of time where the 1-hour maximum daily ozone measurement among the network of stations monitoring a given area could not exceed 120 ppb more than once per year on average. It was found that there were numerous cases where regions were in attainment of this standard, but out of attainment of another proposed standard based on 8-hour averages; the longer 8-hour averages protect against *exposure* instead of just *concentration*, and it is the level of exposure that is of primary concern for health and the environment (Lefohn and Altshuller [26]). In 1997, the EPA established the 8-hour O_3 standard to protect against longer exposure periods, and the level of the NAAQS for ozone is now 84 ppb, with respect to the fourth-highest daily maximum 8-hour average (FHDA) of an ozone season averaged over three years. To break this down for each day, all possible 8-hour blocks of time are considered, and an

average ozone concentration for each block is taken. The maximum of each of these averages is then recorded for that day; for day d call this value y_d. There are 184 days in an ozone season, giving measurements $y_1, \ldots, y_{184}$ from a particular station. Let $y^{[k]}$ denote the k-th largest order statistic, for this ozone season at this station. Let $y_1^{[4]}, y_2^{[4]}$ and $y_3^{[4]}$ be the FHDA for three consecutive seasons, and a C/MSA region is in compliance if the average $\frac{1}{3}\sum_{i=1}^{3} y_i^{[4]}$ is less than 84 ppb. Note that if more than one station exists in a region, then the highest value from all stations in the region is used. Regions not included in the C/MSA can still be designated out of attainment by the U.S. EPA if it is determined that the region is contributing to a C/MSA being out of attainment. In this work, we take the perspective that air quality on an annual basis is of interest, and rather than aggregating over three years, we consider the FHDA for each year individually.

2.5 The Problem

Currently, the EPA does not consider a spatial analysis of the monitoring data for regulatory purposes. Figure 2.2 shows the county boundaries for North Carolina along with ozone monitoring station locations. Clearly, there are many counties that do not have any monitoring stations, and while a region may be larger than a county, it is reasonable to think of the regions as counties in order to glean an idea as to how many regions are not monitored. Figure 2.1

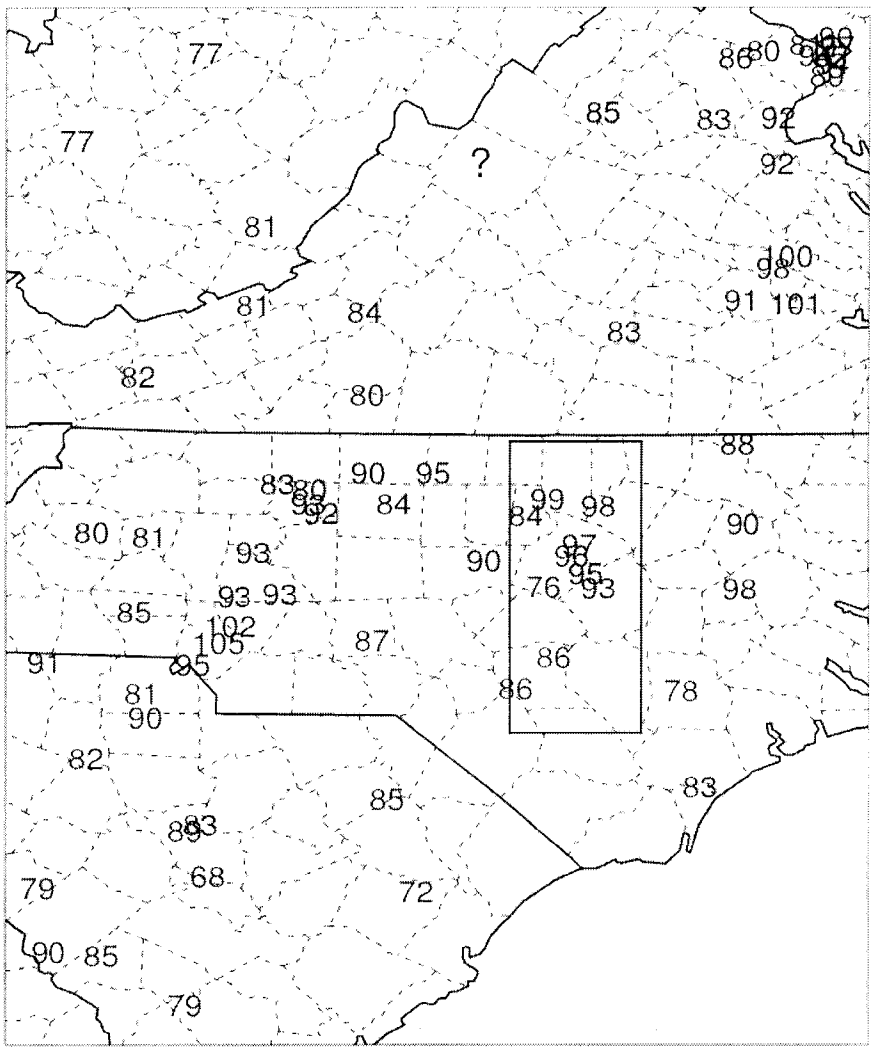


Figure 2.2: County boundaries (dashed) in and around North Carolina with ozone monitoring locations. Numbers are observed fourth-highest daily maximum 8-hour average ozone (FHDA) in parts per billion (ppb) for 1997. Also shown is an area (rectangular box) around the Research Triangle Park (RTP) in North Carolina, where spatial prediction will be compared in chapter 4.

is the study region analyzed in chapter 4. The rectangle shows the boundaries of a gridded region in the Research Triangle Park (RTP) in North Carolina to be used for testing the predictive performance of various models.

Building spatial statistical models for the daily ozone field is straight forward, but the extension to the fourth-highest measurement is difficult because both the covariance and the distribution for the FHDA standard are unknown. Some strategies to sample from the FHDA distribution given the data are:

- Build a space-time model for the daily ozone data and simulate several seasons from it. The fourth-highest value from each season can then be used as an approximate sample from the FHDA distribution.

- Approximate the FHDA distribution by a normal distribution and use geostatistical tools to predict FHDA onto the gridded region.

- Extend methods from extreme value theory (see section 3.2) to approximate the FHDA distribution.

Based on this work, it is also of interest to look at the probabilities of exceeding a threshold at a location, $\mathbf{x}$, not located inside a C/MSA.

Chapter 3

Literature Review

This literature review will begin with a review of space-time modeling in section 3.1. This review will first explore spatial models and associated topics in sections 3.1.1 through 3.1.4, followed by a brief discussion of first order autoregressive models in section 3.1.5. Section 3.1.6 will then give a general overview of space-time models, with discussion on separable and non-separable space-time covariance functions in sections 3.1.8 and 3.1.9. Section 3.2 discusses theoretical results from extreme value theory. Section 3.4 discusses some previous work on ozone modeling with special emphasis on work related to the new 8-hour standard.

3.1 Space-Time Models

Space-time models, or spatiotemporal models, model phenomena that, as the name suggests, have both spatial and temporal components. Statistical models for time series have been well explored. Spatial modeling still has much room for new research particularly in the area of non-stationary covariance modeling. Spatiotemporal statistical models usually try to combine methods from each of these two fields. The emphasis here will be on the spatial component, but will involve a temporal component as well.

3.1.1 Spatial Models

Spatial modeling is a broad term referring to any analysis of data that are spatially correlated. In the case of analyzing ground-level ozone, where the monitor locations are decided upon by physical and political means, best linear unbiased prediction (BLUP), usually called kriging, is generally used for predicting values at unobserved locations. The simplest form of kriging is called simple kriging. For some finite domain D in space, the set-up is as follows.

Let $Z(\mathbf{s})$ be a spatial process with

$$E(Z(\mathbf{s})) = 0$$

for all $\mathbf{s} \in D$ and

$$\text{Cov}(Z(\mathbf{x}), Z(\mathbf{y})) = C(\mathbf{x}, \mathbf{y}) \tag{3.1}$$

Then, for n locations, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and a fixed location, say $\mathbf{x}_0$, it is desired to find the "best" linear predictor of $Z(\mathbf{x}_0)$. To do this, it is necessary to seek weights, $\mathbf{w}' = w_1, \ldots, w_n$, such that

$$\hat{Z}(\mathbf{x}_0) = w_1 Z(\mathbf{x}_1) + \cdots + w_n Z(\mathbf{x}_n) = \mathbf{w}'\mathbf{Z} \tag{3.2}$$

(where $\mathbf{Z} = Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)$) minimizes the mean squared error of prediction. Namely,

$$E(Z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0))^2 = E(Z(\mathbf{x}_0) - \mathbf{w}'\mathbf{Z})^2 \tag{3.3}$$

The right hand side of (3.3) is

$$E(Z^2(\mathbf{x}_0)) - 2\mathbf{w}'E(Z(\mathbf{x}_0)\mathbf{Z}) + \mathbf{w}'K\mathbf{w} \tag{3.4}$$

where $K = E(\mathbf{Z}\mathbf{Z}') = C(\mathbf{x}, \mathbf{y})$ from (3.1).

Minimizing (3.4) with respect to $\mathbf{w}$, $E(Z(\mathbf{x}_0)\mathbf{Z}) = K\mathbf{w}$ or

$$\mathbf{w} = K^{-1}E(Z(\mathbf{x}_0)\mathbf{Z})$$

Furthermore, we have that $E(Z(\mathbf{x}_0)\mathbf{Z}) = (C(\mathbf{x}_0, \mathbf{x}_1), \ldots, C(\mathbf{x}_0, \mathbf{x}_n))'$, which shall be denoted here by $\mathbf{c}'(\mathbf{x}_0)$.

Thus the best linear predictor of $Z(\mathbf{x}_0)$ is given by

$$\hat{Z}(\mathbf{x}_0) = \mathbf{w}'\mathbf{Z} = \mathbf{c}'(\mathbf{x}_0)K^{-1}\mathbf{Z}$$

Rewriting (3.3), the mean squared prediction error (MPSE) is given by

$$E(Z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0))^2 = C(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{c}'(\mathbf{x}_0)K^{-1}\mathbf{c}(\mathbf{x}_0)$$

A more general case of model (3.2); often called universal kriging; allows for a trend function model with spatially correlated noise. That is,

$$Z(\mathbf{x}) = \sum_{i=0}^{p} f_i(\mathbf{x})\beta_i + \varepsilon(\mathbf{x}) \tag{3.5}$$

where $E(\varepsilon(\mathbf{x})) = 0$ for all $\mathbf{x}$ and $\mathrm{Cov}(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{y})) = C(\mathbf{x}, \mathbf{y})$ is assumed known.

Again, it is desired to find the "best" linear unbiased estimate of $Z(\mathbf{x}_0)$ for a new location, $\mathbf{x}_0$, which again involves minimizing the mean squared error, $E(Z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0))^2$, subject to the constraint

$$E(\hat{Z}(\mathbf{x}_0)) = E(Z(\mathbf{x}_0)) = \sum_{i=0}^{p} f_i(\mathbf{x}_0)\beta_i \tag{3.6}$$

Perhaps the simplest method for finding the solution to this problem is to compute the generalized least squares (GLS) estimate of $\underline{\beta}$, yielding

$$\hat{\beta}_{GLS} = (F'K^{-1}F)^{-1}F'K^{-1}\mathbf{Z}$$

where F and K are the matrices $[f_i(\mathbf{x}_j)]$ with $i = 1, \ldots, p$, $j = 1, \ldots, n$ and $[C(\mathbf{x}_k, \mathbf{x}_l)]_{k,l=1}^{n}$.

Once these parameters have been estimated, simple kriging can be performed on the residuals to obtain estimates of the ε terms from (3.5). The mean squared prediction error is given by

$$C(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{c}'(\mathbf{x}_0)K^{-1}\mathbf{c}(\mathbf{x}_0)$$

$$+(\mathbf{f}(\mathbf{x}_0) - F'K^{-1}\mathbf{c}(\mathbf{x}_0))'(F'K^{-1}F)^{-1}(\mathbf{f}(\mathbf{x}_0) - F'K^{-1}\mathbf{c}(\mathbf{x}_0))$$

where the first two terms together are the prediction error variance in spatially correlated noise and the last term is the prediction error variance from the trend.

One major difficulty in fitting a spatial model is in finding an appropriate covariance function, $C(\mathbf{x}, \mathbf{y})$. The general practice is to plot the empirical correlations or variogram–depending on whether or not replicated data is available and then decide on a family of functions that seems appropriate. Finally, a function is fit from this family using nonlinear least squares estimators or other means.

Many families of covariance functions have been proposed for spatial models. For now, we will concentrate on covariances that are isotropic and stationary. Some of the more popular ones include the exponential, Gaussian, spherical and Matérn.

The exponential, which is actually a special case of the Matérn family, has the form

$$C(|h|) = \begin{cases} \sigma^2, & 0 \leq h \leq \varepsilon \\ \sigma^2(1-\alpha)e^{-\frac{h}{\theta}}, & h > \varepsilon. \end{cases} \tag{3.7}$$

where σ^2 is the standard deviation at distance $h = 0$, θ is the range parameter (range=3θ), α is the proportion of nugget effect and the nugget effect is $\sigma^2\alpha$ (see Reich and Davis [35]).

The Gaussian model is a limiting case of the Matérn family, and has the

form

$$C(|h|) = \begin{cases} \sigma^2, & 0 \leq h \leq \varepsilon \\ \sigma^2(1-\alpha)e^{-(\frac{h}{\theta})^2}, & h > \varepsilon. \end{cases} \tag{3.8}$$

where the parameters are as in covariance 3.7 except the range is now 2θ (see Reich and Davis [35]).

As described in Stein [42], covariance (3.8), without a nugget, is infinitely differentiable. Subsequently, all moments of its corresponding spectral density (see section 3.1.7), $f(\omega) = \frac{1}{2}\sigma^2(1-\alpha)(\pi\frac{1}{\theta^2})^{1/2}e^{-(\theta\omega)^2/4}$, are finite so that the field Z has mean square derivatives of all orders; see section 3.1.3 for more on mean square derivatives. A stronger result is that $Z(\mathbf{s})$ can be predicted perfectly for all s based on observing $Z(s')$ for all $s' \in (-\varepsilon, 0]$ for any $\varepsilon > 0$. This property is not generally realistic for modeling physical processes.

The Matérn covariance has perhaps become the most popular choice of covariances because of its flexibility in that it covers a wide variety of covariance forms. For example, the above exponential and Gaussian covariances.

$$C(|h|) = \frac{\sigma}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\nu^{1/2}|h|}{\rho}\right)^{\nu} K_\nu\left(\frac{2\nu^{1/2}|h|}{\rho}\right) \tag{3.9}$$

where K_ν is a modified Bessel function and ν represents the smoothness, σ the variance of the spatial field (also referred to as the sill) and ρ is the range.

The spectral density (section 3.1.7) for the Matérn covariance (3.9) is of the form $f(\omega) = \sigma(\alpha^2 + \omega^2)^{-\nu-1/2}$ for $\nu > 0$, $\sigma > 0$ and $\alpha = \frac{2\nu^{1/2}}{\rho} > 0$, (see Stein [42]). The larger the value of ν, the smoother Z because Z is m times mean square differentiable (see section 3.1.3) if and only if $\nu > m$, since

$\int_{-\infty}^{\infty} \omega^{2m} f(\omega) d\omega < \infty$ if and only if $\nu > m$ (Stein [42]).

3.1.2 Spatial Stationarity and Isotropy

A common simplifying assumption in spatial analyses is to assume that the probabilistic structure of the field is homogeneous across the spatial domain, $D = \Re^d$. Two types of stationarity are used: strict stationarity and weak stationarity. Most of the time, weak stationarity is assumed, and here this type of stationarity will be referred to simply as stationarity. A spatial field is strictly stationary if for all finite n, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$, $\mathbf{h} \in D$ and $c_1, \ldots, c_n \in \Re$,

$$\Pr\{Z(\mathbf{x}_1 + \mathbf{h}) \leq c_1, \ldots, Z(\mathbf{x}_n + \mathbf{h}) \leq c_n\} = \Pr\{Z(\mathbf{x}_1) \leq c_1, \ldots, Z(\mathbf{x}_n) \leq c_n\}.$$

See, for example, Stein [42], Cressie [4], or Reich and Davis [35].

Weak stationarity is similar to strict stationarity, but instead of requiring the entire distribution to be the same regardless of location, weak stationarity simply makes requirements on the first two moments. Specifically, a process is weakly stationary if

- $E\{Z(\mathbf{x})\} = \mu$ (i.e., constant mean)
- $E\{Z(\mathbf{x})Z(\mathbf{y})\} < \infty$ for all $\mathbf{x}, \mathbf{y} \in D$
- $\text{Cov}\{Z(\mathbf{x}), Z(\mathbf{y})\} = K(\mathbf{x} - \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in D$

The function, K, is referred to as the autocovariance function for the spatial field, Z. Strength of association between random variables is better described

using the autocorrelation function, $C(\mathbf{x}) = K(\mathbf{x})/K(\mathbf{0})$; assuming $K(\mathbf{0}) > 0$ (Stein [42]). Clearly, a field that is strictly stationary is also weakly stationary, assuming finite second moments.

While stationarity simplifies spatial analyses by assuming invariance under the translation group of transformations of the coordinates, it is also useful to consider invariance under rotations and reflections. That is, to assume that there is no reason to distinguish from one direction to another for Z. Again, it is possible to consider strict isotropy and weak isotropy. Stein [42] defines strict isotropy in the following way. A random field is strictly isotropic if, for any orthogonal $d \times d$ matrix $\mathbf{\Phi}$ and any $\mathbf{h} \in D$,

$$\Pr\{Z(\mathbf{\Phi}\mathbf{x}_1+\mathbf{h}) \leq c_1, \ldots, Z(\mathbf{\Phi}\mathbf{x}_n+\mathbf{h}) \leq c_n\} = \Pr\{Z(\mathbf{x}_1) \leq c_1, \ldots, Z(\mathbf{x}_n) \leq c_n\}$$

for all finite n, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$ and $c_1, \ldots, c_n \in \Re$.

A random field is weakly isotropic if $E\{Z(\mathbf{x})\} = m$, a constant, and a function K on $[0, \infty)$ such that $\mathrm{Cov}\{Z(\mathbf{x}), Z(\mathbf{y})\} = K(\|\mathbf{x} - \mathbf{y}\|)$ for all $\mathbf{x}, \mathbf{y} \in D$. It is also may be possible to transform the coordinates of a spatial field so that the field is isotropic on the new coordinates. This form of isotropy is called geometric isotropy. Formally, Z is geometric isotropic if there exists an invertible matrix $\mathbf{V}$ such that $Z(\mathbf{V}\mathbf{x})$ is isotropic (Stein [42]).

For Z weakly stationary on D, the autocovariance function, K, must satisfy the following properties.

$$K(\mathbf{0}) \geq 0$$

$$K(\mathbf{x}) = K(-\mathbf{x})$$

$$|K(\mathbf{x})| \leq K(\mathbf{0})$$

Of course, the autocovariance function must be positive definite. Some useful properties of positive definite functions include the following.

1. If K_1 and K_2 are positive definite, then, for all $a_1, a_2 \geq 0$, $a_1K_1 + a_2K_2$ is positive definite.

2. If $K_1, K_2, \ldots$ are positive definite, and $\lim_{n\longrightarrow\infty} K_n(\mathbf{x}) = K(\mathbf{x})$ for all $\mathbf{x} \in D$, then K is positive definite.

3. If K_1 and K_2 are positive definite, then $K(\mathbf{x}) = K_1(\mathbf{x})K_2(\mathbf{x})$ is positive definite.

Property 1 is easily shown as follows.

$$\sum_{j,k}^{n} c_j c_k \{a_1 K_1 + a_2 K_2\} =$$

$$\sum_{j,k}^{n} c_j c_k a_1 K_1 + \sum_{j,k}^{n} c_j c_k a_2 K_2 \geq 0.$$

Property 2 is also easy to show as follows.

$$\sum_{j,k}^{n} c_j c_k K(\mathbf{x}) = \sum_{j,k}^{n} c_j c_k \lim_{n\longrightarrow\infty} K_n(\mathbf{x}) =$$

$$\lim_{n\longrightarrow\infty} \sum_{j,k}^{n} c_j c_k K_n(\mathbf{x}) \geq 0$$

If K_θ is a positive definite autocovariance function on D for all $\theta \in \Re$ and is continuous in θ for all $\mathbf{x}$, and μ is a positive finite measure on $\Re$ with $\int_{\Re} K_\theta(\mathbf{0})\mu(d\theta) < \infty$, then $\int_{\Re} K_\theta(\mathbf{x})\mu(d\theta)$ is positive definite.

3.1.3 Mean Square Continuity and Differentiability

Because there is no simple relationship between the autocovariance function of a random field and the smoothness of its realizations, it is useful to instead relate the autocovariance function to mean square properties of a random field. Specifically, mean square continuity and mean square differentiability.

Definition 3.1.1: $Z(\cdot)$ is *mean square continuous at* $\mathbf{x}$ if $E(Z(\mathbf{y})-Z(\mathbf{x}))^2 \longrightarrow 0$ as $\mathbf{y} \longrightarrow \mathbf{x}$.

A covariance function, K, is continuous everywhere if and only if K is continuous at zero. Specifically,

$$|K(\mathbf{x})-K(\mathbf{y})| = |\mathrm{Cov}(Z(\mathbf{x}),Z(0))-\mathrm{Cov}(Z(\mathbf{y}),Z(0))| = |\mathrm{Cov}(Z(\mathbf{x})-Z(\mathbf{y}),Z(0))|$$

$$\leq [\mathrm{var}(Z(\mathbf{x})-Z(\mathbf{y}))\mathrm{var}(Z(0))]^{1/2} \text{ (Cauchy-Schwartz inequality)}$$

$$= K^{1/2}(0)(2K(0)-2K(\mathbf{x}-\mathbf{y}))^{1/2} \longrightarrow 0 \text{ as } \mathbf{x} \longrightarrow \mathbf{y} \text{ and subsequently}$$

$$|K(\mathbf{x})-K(\mathbf{y})| = |\mathrm{Cov}(Z(\mathbf{x}),Z(0))-\mathrm{Cov}(Z(\mathbf{y}),Z(0))| \longrightarrow 0.$$

Definition 3.1.2: $Z(\cdot)$ is *mean square differentiable* if for all $\mathbf{x}$

$$Z_{\mathbf{h}}(\mathbf{x}) = \frac{Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})}{\mathbf{h}}$$

converges in mean square as $\mathbf{h} \longrightarrow 0$. Call the limiting process $Z'(\mathbf{x})$.

3.1.4 Thin Plate Splines

Thin plate splines were introduced to statistical data analysis in the early 1980s, and are used frequently here. Green and Silverman [16] provide a

good review of thin plate splines. Hastie and Tibshirani [19] also give a good discussion of smoothers generally with some attention to the thin plate spline smoother.

The general model for a thin plate spline is given by

$$\mathbf{Y} = f(\mathbf{X}) + \underline{\varepsilon} \tag{3.10}$$

where f is a d-dimensional surface and $\underline{\varepsilon}$ are zero-mean uncorrelated random errors with variances $\sigma^2\mathbf{W}^{-1}$ with $\mathbf{W}$ a weight matrix.

The estimator of $f(\mathbf{X})$ in (3.10) minimizes the penalized weighted residual sum of squares

$$\frac{1}{n}\sum_i (Y_i - f(\mathbf{x}_i))^2 w_i + \lambda J_m(f)$$

where again w_i are weights, $\lambda > 0$ is a parameter controlling the amount of smoothing and $J_m(f)$ is a roughness penalty based on m-th order derivatives. Larger values of λ yield smoother surfaces for the spatial field. Values of λ near zero produce rougher surfaces, and the estimate interpolates the data.

For one dimension, $d = 1$, and $m = 2$, this becomes the usual cubic spline problem. In this case, a cubic polynomial is fit to interval pieces joined by knots. The requirement at these knots is that f and its first and second derivatives are continuous at each knot and subsequently over the entire interval.

As delineated by Green and Silverman [16], desirable properties for J_m include

- J_m measures rapid variation in f and departure from local linearity or flatness.
- Translation or rotation of the coordinates (in $\Re^2$) does not affect the value of $J_m(f)$.
- $J_m \geq 0$ for all f.
- The problem of finding the surface f that minimizes $J_m(f)$ subject to some constraints is a tractable one.

To satisfy these properties, the roughness is defined to be the integral of the squared m-th order derivatives of f. For example, for $d = 2$ and $m = 2$ and the spatial locations, $\mathbf{x}$, represented as (x_1, x_2), the roughness penalty, $J_m(f)$, is given by

$$J_m(f) = \int\int_{\Re^2} \{(\frac{\delta^2 f}{\delta x_1^2})^2 + 2(\frac{\delta f}{\delta x_1 \delta x_2})^2 + (\frac{\delta^2 f}{\delta x_2^2})^2\} dx_1 dx_2$$

3.1.5 First Order Autoregressive Time Series Model

Similar to the topics discussed above, time series have been studied relatively thoroughly. I discuss some basic properties of time series processes here with special emphasis on the first order autoregressive process because of its role in the daily model approach of chapter 4 of this manuscript. Brockwell and Davis [2] give a very thorough treatment of time series processes. Time series are similar to spatial processes in many ways. The primary difference, of

course, is that a time series implies a natural ordering that does not readily make sense for spatial processes. Analogous to spatial processes, temporal processes have definitions for both strict and weak stationarity. In the time series context, the random time series $\mathbf{Z}(t)$ is said to be strictly *stationary* if the joint distributions of $(Z(t_1), \ldots, Z(t_M))'$ and $(Z(t_1+\tau), \ldots, Z(t_M+\tau))'$ are the same for all positive integers M and all integers, τ. Again, it is common practice to refer to weak (or second order) stationarity simply as stationarity, which is defined as follows. Given a time series, $\mathbf{Z}(t)$ with t an integer, weak stationarity occurs when all of the following are satisfied.

- $E|Z(t)|^2 < \infty$ for all integers t
- $EZ(t) = m$ for all integers t
- $\text{Cov}(Z(t_r), Z(t_s)) = Cov(Z(t_r+\tau), Z(t_s+\tau))$ for all integers r, s and t.

Also analogous to spatial processes, for stationary time series there is an autocovariance function. For the time series $\{Z(t)\}$, this is defined to be

$$\gamma_Z(h) = \text{Cov}(Z(t+h), Z(t))$$

with the autocorrelation function, $\psi_Z(h) = \frac{\gamma_Z(h)}{\gamma_Z(0)}$.

A first order autoregressive process, or AR(1) process, for a time series $Z(t)$ is given by the following.

$$Z(t) = \rho Z(t-1) + \varepsilon(t), \tag{3.11}$$

where $\varepsilon(t) \sim N(0, \sigma^2)$, $|\rho| < 1$ and $\varepsilon(t)$ is independent over time.

Re-writing (3.11) as the infinite sum,

$$Z(t) = \sum_{k=0}^{\infty} \rho^k \varepsilon(t-k), \tag{3.12}$$

makes it easy to see that $E\{Z(t)\} = 0$, $\gamma_Z(h) = \rho^h \gamma_Z(0)$, and $\psi_Z(h) = \rho^{|h|}$ for $h = 0, \pm 1, \ldots$. Additionally, $\text{Cov}(Z(t), \varepsilon(t)) = \sigma^2$ and $\gamma_Z(0) = \frac{\sigma^2}{1-\rho^2}$ so that $\gamma_Z(h) = \rho^h \frac{\sigma^2}{1-\rho^2}$ for $h \geq 0$. Finally, the solution (3.12) is the unique stationary solution to (3.11).

3.1.6 General Framework for Space-Time Models

Both space and time processes are involved with the ozone data, and the previous sections have discussed these processes separately. It is also of interest to consider space-time, or spatiotemporal, processes. Kyriakidis and Journel (1999) [24] provide a good review of geostatistical space-time models, and Cressie [4] also goes into some detail about spatiotemporal modeling. The analyses here considers only a first order autoregressive model with spatially correlated shocks because the interest is in predicting an order statistic spatially (not temporally), but it is worth covering a few topics related to these models.

First, consider finite domains D in space and T in time, with $D \subseteq \Re^d$. $Z(\mathbf{u}, t)$ with $\mathbf{u} \subseteq D$ and $t \subseteq T$ is a spatiotemporal random variable that can take a series of realizations at any location in space and instant in time

according to a probability distribution,

$$F(\mathbf{u}, t; z) = Pr(Z(\mathbf{u}, t) \leq z) \text{ for all } z, (\mathbf{u}, t) \in D \times T$$

A spatiotemporal random function $\{Z(\mathbf{u}, t), (\mathbf{u}, t) \in D \times T\}$ is defined to be a set of typically dependent random variables, $Z(\mathbf{u}, t)$, indexed by location in space and instant in time.

Consider N points in $D \times T$ where $\{Z(\mathbf{u}_1, t_1), \ldots, Z(\mathbf{u}_N, t_N)\}$ has cumulative distribution function

$$F(\mathbf{u}_1, t_1, \ldots, \mathbf{u}_N, t_M; z_{11}, \ldots, z_{NM}) =$$

$$Pr(Z(\mathbf{u}_1, t_1) \leq z_{11}, \ldots, Z(\mathbf{u}_N, t_M) \leq z_{NM}).$$

A statistical analysis generally attempts to find the optimal prediction of an unobserved part of the space-time process. Assuming that the process has finite variance and the mean and covariance of the process at two spatial points and two different time points exist, the simple kriging predictor is the linear combination

$$Z^*(\mathbf{u}_0, t_0) = \mu(\mathbf{u}_0, t_0) + \sum_{i=1}^{k} a_i(Z(\mathbf{u}_i, t_i) - \mu(\mathbf{u}_i, t_i))$$

of the observations that minimize the mean squared prediction error (MPSE) [5], [14]. The MPSE is given by the simple kriging predictor [5]:

$$Z^*(\mathbf{u}_0, t_0) = m(\mathbf{u}_0, t_0) + \mathbf{c}(\mathbf{u}_0, t_0)'\Sigma^{-1}(\mathbf{Z} - \mu),$$

where $\Sigma \equiv Cov(\mathbf{Z}(\mathbf{u},t))$, $\mathbf{c}(\mathbf{u}_0,t_0)' \equiv Cov(Z(\mathbf{u}_0,t_0),Z(\mathbf{u},t))$ and $\mu \equiv E\mathbf{Z}(\mathbf{u},t)$. The corresponding MPSE is then given by $\mathbf{c}(\mathbf{u}_0,t_0)'\Sigma^{-1}\mathbf{c}(\mathbf{u}_0,t_0)$.

Analogous to the definitions in sections 3.1.2 and 3.1.5 for strict and second order stationarity for temporal and spatial processes, the random function $Z(\mathbf{u},t)$ is said to be strictly stationary within $D \times T$ if its spatiotemporal law is invariant by translation $(\mathbf{h},\tau) \in D \times T$. This implies that any two vectors of random variables $(Z(\mathbf{u}_1,t_1),\ldots,Z(\mathbf{u}_N,t_M))'$ and $(Z(\mathbf{u}_1+\mathbf{h},t_1+\tau),\ldots,Z(\mathbf{u}_N+\mathbf{h},t_M+\tau))'$ have the same multivariate cdf regardless of the translated vector $(\mathbf{h},\tau) \in D \times T$. Symbolically, we have that $F(\mathbf{u}_1,t_1,\ldots,\mathbf{u}_N,t_M) = F(\mathbf{u}_1+\mathbf{h},t_1+\tau,\ldots,\mathbf{u}_N+\mathbf{h},t_M+\tau)$ for all $\mathbf{u}_1,t_1,\ldots,\mathbf{u}_N,t_M$ and $(\mathbf{h},\tau) \in D \times T$.

The random function, $Z(\mathbf{u},t)$, has second order stationarity, which will subsequently be referred to simply as stationarity, if

- $EZ(\mathbf{u},t) = m$, for all $(\mathbf{u},t) \in D \times T$ and
- $E[Z(\mathbf{u},t)-m][Z(\mathbf{u}',t')-m] = C_Z(\mathbf{h},\tau)$, where $C_Z(\cdot,\cdot)$ is the space-time covariance function. In words, $\mathrm{Cov}(Z(\mathbf{u},t)) = C_Z(\mathbf{u},t;\mathbf{u}',t')$ depends only on the spatial and temporal lags $\mathbf{h} = \mathbf{u}-\mathbf{u}'$ and $\tau = t-t'$.

3.1.7 Spectral Methods

It is often useful to use spectral methods when studying the autocovariance structure of weakly stationary random processes (spatial or temporal), $\{Z(u)\}$. Essentially, the spectral representation of a stationary random process de-

composes $\{Z(u)\}$ into a sum of sinusoidal components with uncorrelated random coefficients. Brockwell and Davis [2] and Stein [42] both provide a good overview of such representations.

Definition 3.1.3: Z is a *complex-valued random field* if

$$Z(\mathbf{x}) = U(\mathbf{x}) + iV(\mathbf{x})$$

where U and V are real-valued random fields.

If $(U(\mathbf{x}), V(\mathbf{x}))^T$ is jointly weakly stationary (i.e., U and V are each weakly stationary and $\mathrm{Cov}(U(\mathbf{x}), V(\mathbf{y}))$ depends only on the distance between $\mathbf{x}$ and $\mathbf{y}$), then Z is weakly stationary. Here,

$$K(\mathbf{y}) = \mathrm{Cov}(Z(\mathbf{x}+\mathbf{y}), Z(\mathbf{x})) =$$

$$E\{Z(\mathbf{x}+\mathbf{y})\overline{Z(\mathbf{x})}\} \text{ (provided } EZ(\mathbf{x}) = 0) =$$

$$E\{(U(\mathbf{x}+\mathbf{y}) + iV(\mathbf{x}+\mathbf{y}))(U(\mathbf{x}) - iV(\mathbf{x}))\} =$$

$$= K_U(\mathbf{y}) + K_V(\mathbf{y}) + i\{\mathrm{Cov}(V(\mathbf{x}+\mathbf{y}), U(\mathbf{x})) - \mathrm{Cov}(U(\mathbf{x}+\mathbf{y}), V(\mathbf{x}))\},$$

Some properties of $K(\mathbf{y})$ are as follows.

1. $K(\mathbf{0}) \geq 0$

2. $K(-\mathbf{y}) = \overline{K(\mathbf{y})}$

3. $K(\cdot)$ is nonnegative definite (i.e., $\sum_{j,k=1}^{n} c_j \overline{c_k} K(\mathbf{x}_i - \mathbf{x}_j) \geq 0$).

Next, let $Z_1, \ldots, Z_n$ be mean zero uncorrelated complex-valued random variables, and consider

$$Z(\mathbf{x}) = \sum_{k=1}^{n} Z_k e^{i\omega_k^T \mathbf{x}}.$$

Here, $\omega_1, \ldots, \omega_n$ are fixed *frequencies* in $\Re^d$. Clearly, $EZ(\mathbf{x}) = 0$ and

$$E(Z(\mathbf{x}+\mathbf{y})\overline{Z(\mathbf{x})}) = \sum_{k=1}^{n}\sum_{j=1}^{n} e^{i\omega_j^T(\mathbf{x}+\mathbf{y})} e^{-i\omega_k^T \mathbf{x}} E(Z_j \overline{Z_k})$$

$$= \sum_{k=1}^{n} e^{i\omega_k^T \mathbf{y}} f_k, \tag{3.13}$$

where $f_k = E|Z_k|^2$. Because (3.13) depends only on $\mathbf{y}$, $Z(\cdot)$ is weakly stationary with autocovariance function $K(\mathbf{y}) = \sum_{k=1}^{n} e^{i\omega_k^T \mathbf{y}} f_k$.

From the above representation, it is possible to derive the following representations for $K(\mathbf{y})$ and $Z(\mathbf{x})$.

$$K(\mathbf{y}) = \int_{\Re^d} e^{i\omega^T \mathbf{y}} dF(\omega), \tag{3.14}$$

where $F(\cdot)$ is a distribution function that puts mass f_i at frequency ω_i, $i = 1, \ldots, n$.

$$Z(\mathbf{x}) = \int_{\Re^d} e^{i\omega^T \mathbf{x}} M(\omega), \tag{3.15}$$

where $M(\cdot)$ is an orthogonal-increment process that has "jumps" at frequencies ω_i of size Z_i, $i = 1, \ldots, n$.

In fact, an important result is that every mean-zero weakly stationary random process, $Z(\cdot)$, that is mean square continuous has representation (3.15). Specifically,

Theorem 3.1.1: (Bochner's Theorem [1]). A complex-valued function K on $\Re^d$ is the autocovariance function for a weakly stationary mean square continuous complex-valued random process on $\Re^d$ if and only if it can be represented as in (3.14) where F is a positive finite measure.

3.1.8 Space-Time Separable Covariance Functions

Because of fundamental differences in space and time dimensions, the covariance function, $C(\mathbf{h}, \tau)$, taking both spatial and temporal variability into account can often be simplified by separating the space and time components. Usually, this decomposition is performed by using a sum of two components: $C(\mathbf{h}, \tau) = C_1(\mathbf{h}) + C_2(\tau)$, but it can also be decomposed into a product as $C(\mathbf{h}, \tau) = C_1(\mathbf{h})C_2(\tau)$, where in both cases, $C_1(\mathbf{h})$ represents a purely spatial covariance function and $C_2(\tau)$ is a purely temporal covariance function.

Under the assumption of space-time separability the spatial behavior of $Z(\mathbf{u}, t)$ is considered to be the same at all time points. Similarly for the temporal behavior. Thus, no change of the spatial pattern from one time point to another can be accounted for, nor can changes in the temporal pattern from one spatial location to another. Additionally, there are no guidelines for inferring the two component structures, $C_1(\mathbf{h})$ and $C_2(\tau)$. Another problem is that covariance models built from a sum of one-dimensional structures may not be positive definite in higher dimensional spaces; and if no covariance exists,

then there is no model. Nevertheless, the form of separable covariances are mathematically congenial in that it is easier to formulate parametric families that satisfy the definiteness property (Cressie and Huang [5]).

3.1.9 Non-separable Space-Time Covariance Functions

Non-separable space-time covariance functions have been an area of active research, and some new classes of such functions have been derived (see, for example, Cressie and Huang [5], Gneiting [14], and Ma [28]). Cressie and Huang [5] give a clever and simple methodology for developing whole classes of non-separable spatiotemporal stationary covariance functions in closed form, based on Bochner's Theorem [1] (see section 3.1.7).

The results of Cressie and Huang [5] were best described to me by Breidt (personal communication). The following outlines his comments. Let $Z(\underline{s}, t)$ be a stationary spatiotemporal process with continuous covariance function $C(\underline{h}, u)$ and spectral density $g(\underline{\omega}, \tau) \geq 0$ then

$$Z(\underline{s}, t) = \int e^{i\underline{s}'\underline{\omega}} e^{it\tau} g^{\frac{1}{2}}(\underline{\omega}, \tau) dZ_H(\underline{\omega}) dZ_T(\tau) \tag{3.16}$$

where Z_H and Z_T are two uncorrelated processes of orthogonal increments with $E|dZ_H(\underline{\omega})|^2 = d\underline{\omega}$ and $E|dZ_T(\tau)|^2 = d\tau$.

Note that by Bochner's theorem

$$C(\underline{h}, u) = \int \int e^{i\underline{h}'\omega + iu\tau} g(\underline{\omega}, \tau) d\underline{\omega} d\tau \tag{3.17}$$

It is possible to rewrite equation (3.16) as

$$Z(\underline{s},t) = \int e^{i\underline{s}'\underline{\omega}} X_{\underline{\omega}}(t) K^{\frac{1}{2}}(\underline{\omega}) dZ_H(\underline{\omega})$$

where

$$X_{\underline{\omega}}(t) = \frac{\int e^{it\tau} g^{\frac{1}{2}}(\underline{\omega},\tau) dZ_T(\tau)}{K^{\frac{1}{2}}(\underline{\omega})}$$

and

$$K(\underline{\omega}) = \int g(\underline{\omega},\tau) d\tau \tag{3.18}$$

It follows that

$$\mathrm{Cov}(X_{\underline{\omega}}(t+u), X_{\underline{\nu}}(t)) =$$

$$\frac{1}{K(\underline{\omega})} \int e^{i(t+u-t)\tau} g(\underline{\omega},\tau) d\tau = \frac{h(\underline{\omega},u)}{K(\underline{\omega})} = \rho(\underline{\omega},u) \tag{3.19}$$

For each $\underline{\omega}$, $\rho(\underline{\omega},u)$ is an autocorrelation function (ACF) in time, so $X_{\underline{\omega}}(t)$ is a stationary time series with unit variance. This $h(\cdot;\cdot)$ corresponds with the $h(\cdot;\cdot)$ from (6) in Cressie and Huang [5]. That is, they rewrite the Fourier transform of (3.17),

$$g(\underline{\omega};\tau) = (2\pi)^{-d-1} \int\int e^{-i\underline{h}'\underline{\omega} - iu\tau} C(\underline{h};u) d\underline{h} du,$$

in terms of the new function $h(\cdot;\cdot)$ from (3.19). Specifically,

$$g(\underline{\omega};\tau) = (2\pi)^{-1} \int e^{-iu\tau} h(\underline{\omega};u) du,$$

where

$$h(\underline{\omega};u) \equiv (2\pi)^{-d} \int e^{-i\underline{h}'\underline{\omega}} C(\underline{h};u) d\underline{h} = \int e^{iu\tau} g(\underline{\omega};\tau) d\tau$$

Further, Breidt rewrites equation (3.19) as

$$Cov(X_{\underline{\omega}}(t+u), X_{\underline{\nu}}(t)) = \frac{1}{K^{\frac{1}{2}}(\underline{\omega})K^{\frac{1}{2}}(\underline{\nu})} \int e^{iu\tau} g^{\frac{1}{2}}(\underline{\omega}, \tau) g^{\frac{1}{2}}(\underline{\nu}, \tau) d\tau \qquad (3.20)$$

From (3.20) one obtains that

$$\mathrm{Cov}(Z(\underline{s}+\underline{h}, t+u), Z(\underline{s}, t)) =$$

$$\int \int e^{i(\underline{s}+\underline{h})'\underline{\omega}} e^{-i\underline{s}'\underline{\nu}} K^{\frac{1}{2}}(\underline{\omega}) K^{\frac{1}{2}}(\underline{\nu}) E[X_{\underline{\omega}}(t+u) X_{\underline{\nu}}(t)] E[dZ_H(\underline{\omega}) \overline{dZ_H(\underline{\nu})}] =$$

$$\int e^{i\underline{h}'\underline{\omega}} K(\underline{\omega}) \rho(\underline{\omega}, u) d\underline{\omega} = C(\underline{h}, u)$$

as in (8) of Cressie and Huang [5], which states that

$$C(\underline{h}; u) \equiv \int e^{i\underline{h}'\underline{\omega}} \rho(\underline{\omega}; u) K(\underline{\omega}) d\underline{\omega} \qquad (3.21)$$

Note that with $K(\underline{\omega})$ as in (3.18) and $g(\underline{\omega}, \tau)$ a spectral density, we have that $\int K(\underline{\omega}) d\underline{\omega} < \infty$.

Cressie and Huang [5] proceed to construct classes of valid spatiotemporal non-separable covariance functions by specifying functions for $\rho(\underline{\omega}, t)$ and $K(\underline{\omega})$ such that $\rho(\underline{\omega}, \cdot)$ is a continuous ACF for each $\underline{\omega} \in \Re^d$ with $\int \rho(\underline{\omega}, t) dt < \infty$ and $K(\underline{\omega}) > 0$ as defined by equation (3.18). One example of such a parametric family of valid covariance functions, and subsequently spectral density functions is derived by letting

$$\rho(\omega, t) = \exp(-\frac{\|\omega\|^2 t^2}{4}) \exp(-\delta t^2) \qquad (3.22)$$

with $\delta > 0$ and $k(\omega) = \exp(-\frac{c_0 \|\omega\|^2}{4})$, $c_0 > 0$.

These give rise to the covariance function

$$C(\mathbf{h}, t) \propto \frac{1}{(t^2+c_0)^{d/2}} \exp(-\frac{\|\mathbf{h}\|^2}{t^2+c_0}) \exp(-\delta t^2),\ \delta > 0,$$

which is a continuous spatiotemporal covariance function in $\Re^d \times \Re$. Because the limit of a sequence of spatiotemporal stationary covariance functions is still valid if the limit exists [30], as $\delta \longrightarrow 0$, a three-parameter non-separable spatiotemporal stationary covariance family is

$$C(\mathbf{h}, t|\theta) = \frac{\sigma^2}{(a^2t^2+1)^{d/2}} \exp(-\frac{b^2\|\mathbf{h}\|^2}{a^2t^2+1}),$$

where $\theta = (a, b, \sigma^2)'$, $a \geq 0$ is the scaling parameter of time, $b \geq 0$ is the scaling parameter of space and $\sigma^2 = C(\mathbf{0}, 0|\theta) > 0$. Because of redundancy in the parameters a, b and c_0, they set $c_0 = 1$.

The above method for constructing valid spatiotemporal covariance functions depends on finding closed-form Fourier transform pairs. Gneiting [14] proposes another class of non-separable covariance functions that do not depend on closed form Fourier transform pairs. Specifically, let $\phi(t)$, $t \geq 0$, be a completely monotone function and let $\psi(t)$, $t \geq 0$, be a positive function with a completely monotone derivative. Then

$$C(\mathbf{h}, \tau) = \frac{\sigma^2}{\psi(|\tau|^2)^{d/2}} \phi(\frac{\|\mathbf{h}\|^2}{\psi(|\tau|^2)}),\ (\mathbf{h}; \tau) \in \Re^d \times \Re$$

is a space-time covariance function.

Ma [28] arrives at a related result to Cressie and Huang [5] for obtaining families of spatiotemporal-temporal stationary covariances from known purely

spatial and purely temporal covariances. His results can be summarized by the following theorems and corollaries.

His first theorem provides a simple way of deriving space-time covariances from purely spatial or purely temporal ones. Specifically,

Theorem 3.1.2: (Theorem 1 of Ma [28])

Let $\underline{\theta}$ and $\underline{\theta}_0$ be constant vectors on $\Re^d$. If $C_0(\mathbf{s};t)$ is a spatiotemporal-temporal stationary covariance on $\Re^d \times \Re$ then

$$C(\mathbf{s};t) = C_0(\mathbf{s} + \underline{\theta}t; t + \underline{\theta}_0'\mathbf{s}),\ (\mathbf{s};t) \in \Re^d \times \Re$$

is a spatiotemporal-temporal stationary covariance on $\Re^d \times \Re$.

From this theorem, two corollaries result, which simply state that if C_M is a stationary covariance on $\Re^k$ for k a positive integer, then if C_M is a purely spatial covariance function, $C(\mathbf{s};t) = C_M(\mathbf{s} + \underline{\theta}t)$ is a valid stationary covariance function and if C_M is a purely temporal covariance function, $C(\mathbf{s};t) = C_M(t + \underline{\theta}_0'\mathbf{s})$ is a valid stationary covariance function. In this way, it is easy to construct valid spatiotemporal-temporal covariances from known purely spatial or temporal covariances. This result can also be extended so that C_M can vary off of the same hyperplane along which, for example, $\mathbf{s} + \underline{\theta}t$ is constant by allowing $\underline{\theta}$ to be a random vector.

Ma [28] then constructs new families of spatiotemporal-temporal station-

ary covariances from known purely spatial and purely temporal covariances using his second theorem.

Theorem 3.1.3: (Theorem 2 of Ma [28])

Let d_0 be a positive integer and $\mu(\underline{\omega})$ be a nonnegative bounded measure on $\Re_+^{d_0}$. If $C_S(\mathbf{s};\underline{\omega})$ is a stationary covariance of $\mathbf{s} \in \Re^d$ and a measurable function of $\underline{\omega} \in \Re_+^{d_0}$ for every $\mathbf{s} \in \Re^d$ and $C_T(t;\underline{\omega})$ is a stationary covariance of $t \in T$ for every $\underline{\omega} \in \Re_+^d$ and a measurable function of $\underline{\omega} \in \Re_+^{d_0}$ for every $t \in T$, then

$$C(\mathbf{s};t) = \int_{\Re_+^{d_0}} C_S(\mathbf{s};\underline{\omega}) C_T(t;\underline{\omega}) d\mu(\underline{\omega}), (\mathbf{s};t) \in \Re^d \times T \qquad (3.23)$$

is a spatiotemporal-temporal stationary covariance on $\Re^d \times T$, provided that the integral exists for all $(\mathbf{s};t) \in \Re^d \times T$.

Covariance (3.23) is a mixture of separable covariances because for each fixed $\underline{\omega} \in \Re_+^{d_0}$, the integrand of (3.23) is simply a product of the purely spatial and temporal covariances.

It is as a special case of equation (3.23) that Ma [28] derives a method for finding classes of non-separable space-time covariances that are related to the results of Cressie and Huang [5]. Specifically, for each fixed $\underline{\omega} \in \Re_+^d$, $\cos(\underline{\omega}'\mathbf{s})$ is a purely spatial correlation function corresponding to the process

$$Z_{\underline{\omega}}(\mathbf{s}) = A\cos(\underline{\omega}'\mathbf{s}) + B\sin(t;\underline{\omega}'\mathbf{s}),\ \mathbf{s} \in \Re^d,$$

where A and B are uncorrelated random variables with means of zero and variances of 1. Thus substituting $d_0 = d$ and $C_S = \cos(\underline{\omega}'\mathbf{s})$ yields the transform method of corollary 2.1 in Ma [28]. Namely,

Corollary 3.1.1: (Corollary 2.1 of Ma [28])

Let $\mu(\underline{\omega})$ be a nonnegative bounded measure on $\Re_+^d$. If $C_T(t;\underline{\omega})$ is a stationary covariance of $t \in T$ for every $\underline{\omega} \in \Re_+^d$ and a measurable function of $\underline{\omega} \in \Re_+^d$ for every $t \in T$, then

$$C(\mathbf{s};t) = \int_{\Re_+^d} \cos(\underline{\omega}'\mathbf{s}) C_T(t;\underline{\omega}) d\mu(\underline{\omega}), (\mathbf{s};t) \in \Re^d \times T$$

is a spatiotemporal-temporal stationary covariance on $\Re^d \times T$.

Ma [29] continues with this idea of mixture models to introduce two new classes of valid covariances based on mixture models. Specifically, for $T \in \Re^1$ or the set of positive integers, one type is of the form:

$$C(\mathbf{s};t) = \int_{\Re_+^1 \times T} S(u\mathbf{s}) T(vt) dW(u,v), \tag{3.24}$$

where $S(\mathbf{s})$ and $T(t)$ are purely spatial and temporal covariance functions on $\Re^d$ and T and $W(u,v)$ is a nonnegative bounded measure on $\Re_+^1 \times T$. Equation (3.24) reduces to a separable covariance function if $W(u,v)$ is separable.

The second type is of the form:

$$C(\mathbf{s};t) = \int_{\Re_+^2 \times T} S^u(\mathbf{s}) T^v(t) dW(u,v), \tag{3.25}$$

where $S(\mathbf{s}) \geq 0$ and $T(t) \geq 0$ are again purely spatial and temporal covariances on $\Re^d$ and T and $W(u, v)$ is a nonnegative bounded measure on $\Re_+^2 \times T$. To ensure positive definiteness in (3.25), Ma [29] assumes that $S(\mathbf{s}) = \exp\{-\gamma_1(\mathbf{s})\}$ and $T(t) = \exp\{-\gamma_2(t)\}$, where $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ are purely spatial and temporal variograms.

Ma [29] summarizes (3.24) and (3.25) with the following theorem.

Theorem 3.1.4: (Theorem 3 of Ma [29])

Let $L(\theta_1, \theta_2)$ be the Laplace transform of a nonnegative random vector (X_1, X_2). If $\gamma_1(\mathbf{s})$ is a purely spatial variogram on $\Re^d$ and $\gamma_2(t)$ is a purely temporal variogram on T, then

$$C(\mathbf{s}, t) = L(\gamma_1(\mathbf{s}), \gamma_2(t)) \tag{3.26}$$

is a spatiotemporal-temporal covariance function on $\Re^d \times T$.

3.1.10 Spatial AR(1) Models

Space-time models may incorporate both space and time dimensions simultaneously or separately as spatially varying time series or temporally varying spatial processes. Considered here, is a specific model that may be employed; and is used in chapter 4 of this manuscript. One simple way to model space-time data is to use an AR(1) model with spatial shocks. Let $Z(\mathbf{x}, t)$ denote a space-time process with mean 0 and variance 1. For each spatial site, $\mathbf{x}$,

consider an AR(1) model over time for $Z(\mathbf{x}, t)$. Namely,

$$Z(\mathbf{x}, t) = \rho(\mathbf{x})Z(\mathbf{x}, t-1) + \varepsilon(\mathbf{x}, t) \tag{3.27}$$

where $|\rho(\mathbf{x})| < 1$ for all $\mathbf{x} \in D$, the spatial shocks, $\varepsilon(\mathbf{x}, t)$, are independent over time but spatially correlated with covariance function $\mathrm{Cov}(\varepsilon(\mathbf{x}, t), \varepsilon(\mathbf{x}', t))$ given by

$$\sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x}')}\psi(\mathbf{x}, \mathbf{x}').$$

Here, if $\psi(\cdot)$ is a correlation function, and if it is solely a function of a distance between $\mathbf{x}$ and $\mathbf{x}'$, then ε is stationary a stationary spatial process. From (3.27), $Z(\mathbf{x}, t)$ can be written as an infinite sum over time as

$$Z(\mathbf{x}, t) = \sum_{j=0}^{\infty} \rho(\mathbf{x})^j \varepsilon(\mathbf{x}, t-j)$$

From the above expression, it is straightforward to compute the covariance function for $Z(\mathbf{x}, t)$ between two spatial locations at different times. Namely,

$$\mathrm{Cov}(Z(\mathbf{x}, t), Z(\mathbf{x}', t-\tau)) =$$

$$\mathrm{Cov}(\sum_{j=0}^{\infty} \rho(\mathbf{x})^j \varepsilon(\mathbf{x}, t-j), \sum_{k=0}^{\infty} \rho(\mathbf{x})^k \varepsilon(\mathbf{x}', t-\tau-k)) =$$

$$\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \rho(\mathbf{x})^j \rho(\mathbf{x})^k \mathrm{Cov}(\varepsilon(\mathbf{x}, t-j), \varepsilon(\mathbf{x}', t-\tau-k)),$$

where $\mathrm{Cov}(\varepsilon(\mathbf{x}, t-j), \varepsilon(\mathbf{x}', t-\tau-k)) = 0$ whenever $t-j \neq t-\tau-k$ so that it is only nonzero when $j = \tau + k$, and the above expression is equivalent to the following.

$$(\rho(\mathbf{x}))^{\tau} \sum_{k=0}^{\infty} (\rho(\mathbf{x})\rho(\mathbf{x}'))^{k} \mathrm{Cov}(\varepsilon(\mathbf{x}, t-\tau-k), \varepsilon(\mathbf{x}', t-\tau-k)) =$$

$$\frac{(\rho(\mathbf{x}))^{\tau}\sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x}')}\psi(\mathbf{x},\mathbf{x}')}{1-\rho(\mathbf{x})\rho(\mathbf{x}')} \text{ for } \tau \geq 0 \tag{3.28}$$

Note: If $\rho(\mathbf{x}) \neq \rho$ then

- $Z(\mathbf{x}, t)$ is not stationary in space even if $\varepsilon(\mathbf{x}, t)$ is stationary in space.
- $\mathrm{Cov}(Z(\mathbf{x}, t), Z(\mathbf{x}', t-\tau))$ is not space-time separable.

Wikle [45] takes an alternative approach to this type of model. First, he assume an observable and spatially continuous spatial process $Z(\mathbf{x}; t)$, where $\mathbf{x} \in D$, and discrete index of times $t \in \{1, 2, \ldots\}$, and then supposes that the observable process has a component of measurement error expressed through the measurement equation

$$Z(\mathbf{x}; t) = Y(\mathbf{x}; t) + \varepsilon(\mathbf{x}; t), \tag{3.29}$$

where $Y(\mathbf{x}; t)$ is a "smoother" process than $Z(\mathbf{x}; t)$. The goal is to predict the process $Y(\cdot; \cdot)$. In so doing, Wikle [45] assumes that $Y(\mathbf{x}; t)$ from (3.29) can be written

$$Y(\mathbf{x}; t) = Y_K(\mathbf{x}; t) + \nu(\mathbf{x}; t), \tag{3.30}$$

where $\nu(\mathbf{x}; t)$ is a component of variance representing small-scale spatial variation. The component $Y_K(\mathbf{x}; t)$ is assumed to evolve according to the state equation

$$Y_K(\mathbf{x}; t) = \int_D w_{\mathbf{x}}(\mathbf{u}) Y_K(\mathbf{u}; t-1) d\mathbf{u} + \eta(\mathbf{x}; t), \tag{3.31}$$

where $\eta(\mathbf{x};t)$ is a spatially colored noise process and $w_{\mathbf{x}}(\mathbf{u})$ is a function representing the interaction between the state process $Y_K(\mathbf{u};t-1)$ and $Y_K(\mathbf{x};t)$; that is, the temporally dynamic component.

3.2 Extreme Value Statistics

Although space-time models are useful for analyzing daily ozone data, the NAAQS problem for ozone involves an order statistic, which suggests the use of methods from extreme value theory. Characterizing multivariate extreme value order statistics is a relatively new field. Currently, there are two main approaches: classical models and threshold models.

3.2.1 Classical Models

Let X be a continuous random variable with probability density function $f_X(x)$, and cumulative distribution function $F_X(x)$. If a random sample of size n is drawn from $f_X(x)$, the marginal probability density function for the r-th largest order statistic, $X_{(n-r):n}$, is given by

$$f_{X_{(n-r):n}}(u) = \frac{n!}{(r-1)!(n-r)!}[F_X(u)]^{n-r}[1-F_X(u)]^r f_X(u) \tag{3.32}$$

where $1 \leq r \leq n$. For the maximum, or equivalently the minimum, order statistic (3.32) reduces to

$$f_{X_{n:n}}(u) = n(1-F_X(u))^n f_X(u) \tag{3.33}$$

Although models exist for the r-th largest order statistic, most work has been focused on the maximum (or, equivalently, the minimum) order statistic (see, for example, Coles [3] and Leadbetter *et al.* [25]). If F is unknown, distributions (3.32) and (3.33) are not helpful. In such cases, one can look at what happens asymptotically. Because $F^n \longrightarrow 0$ as $n \longrightarrow \infty$, it is necessary to stabilize the location and scale of $M_n = \max_{i=1,\ldots,n} X_i$ as n increases by finding sequences of constants $\{a_n\}$ and $\{b_n\}$ such that for $\frac{M_n - b_n}{a_n} = M_n^*$, $Pr\{M_n^* \leq z\} = F(a_n z + b_n)$ converges in distribution to a non-degenerate distribution function as $n \longrightarrow \infty$. If such sequences can be found, then the extremal types theorem (see, for example, Coles [3]) yields an important result. If there exist sequences of constants $\{a_n\}$ and $\{b_n\}$ such that $P[\frac{M_n - b_n}{a_n} \leq z] \longrightarrow G(z)$ as $n \longrightarrow \infty$ where G is a non-degenerate distribution function, then, regardless of F, G belongs to one of three families of distributions. Namely,

- I (Gumbel): $G(z) = \exp\{-\exp\{-(\frac{z-b}{a})\}\}$, $-\infty < z < \infty$.
- II (Fréchet): $G(z) = \begin{cases} 0, & z \leq b \\ \exp\{-(\frac{z-b}{a})^{-\alpha}\}, & z > b \end{cases}$
- III (Weibull): $G(z) = \begin{cases} \exp\{-[(-\frac{z-b}{a})^{\alpha}]\}, & z < b \\ 1, & z \geq b \end{cases}$

for parameters $a > 0$, b and $\alpha > 0$. The above convergence occurs if and only if $n(1 - F(a_n z + b_n)) \longrightarrow -\log G(z)$ [40].

These three families of distributions can be combined into a single family

of models having distribution functions of the form:

$$G(z) = \exp\{-[1 + \xi(\frac{z-\mu}{\sigma})]^{-1/\xi}\} \tag{3.34}$$

for $\{z : 1 + \xi(\frac{z-\mu}{\sigma}) > 0\}$, $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. Here, $\xi \longrightarrow 0$ corresponds to the Gumbel distribution, $\xi > 0$ to the Fréchet and $\xi < 0$ to the Weibull distribution. The family of models (3.34) is referred to as the *generalized extreme value* (GEV) family of distributions.

Note:

- For n large, $P[M_n \leq z] \approx G(\frac{z-b_n}{a_n}) = G^*(z)$.
- If for all $n \in 2, 3, \ldots$ there exist constants $\alpha_n > 0$ and β_n such that $G^n(\alpha_n z + \beta_n) = G(z)$, then G is said to be *max-stable*. A distribution is max-stable if and only if it is a GEV [3].

There is an analogous extension of the above univariate asymptotic distribution functions to the bivariate case. Take $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent vectors of correlated pairs of random variables and let $M_{x,n} = \max_{i=1,\ldots,n}\{X_i\}$ and $M_{y,n} = \max_{i=1,\ldots,n}\{Y_i\}$ so that $\underline{M}_n = (M_{x,n}, M_{y,n})$. The index i for which the maximum of the X_i sequence occurs need not be the same as that of the Y_i sequence. A similar theorem to that of the univariate case is also given by Coles [3]. Namely, let $\underline{M}_n^* = (\frac{1}{n}M_{x,n}, \frac{1}{n}M_{y,n}) = (M_{x,n}^*, M_{y,n}^*)$ where (X_i, Y_i) are independent vectors each with independent GEV marginal distributions. Specifically, $\tilde{x} = [1 + \xi_x(\frac{x-\mu_x}{\sigma_x})]^{1/\xi_x}$ and $\tilde{y} = [1 + \xi_y(\frac{y-\mu_y}{\sigma_y})]^{1/\xi_y}$.

Then, if $P[M^*_{x,n} \leq x, M^*_{y,n} \leq y] \longrightarrow G(x,y)$ in distribution, where G is a non-degenerate distribution function, then G has the form

$$G(x,y) = \exp\{-V(x,y)\},\ x > 0,\ y > 0$$

where $V(x,y) = 2\int_0^1 \max(\frac{\omega}{x}, \frac{1-\omega}{y})dH(\omega)$ and H is a distribution function on $[0,1]$ such that $\int_0^1 \omega H(\omega) = \frac{1}{2}$.

Note:

- This theorem does not apply if the marginal distributions for the X_i and Y_i sequences are normal or from any other non-GEV distribution.
- $G^n(x,y) = G(\frac{1}{n}x, \frac{1}{n}y)$ for $n = 2,3,\ldots$. This is analogous to the idea of *max-stability* in the univariate case.

3.2.2 Modeling Threshold Exceedances

An extension of this approach is to look at all values above a certain *threshold*. The family of generalized Pareto distributions is a class of limiting distributions for exceedances over a threshold, u, given by:

$$\Pr\{X > x | X > u\} = G(x;\sigma,\xi,u) = 1 - \max\{(1 + \xi\frac{x-u}{\sigma})^{-1/\xi}, 0\}$$

valid on $0 < x < \infty$ for $\xi \leq 0$ or on $0 < x < \frac{\sigma}{\xi}$ for $\xi > 0$ where $x > 0$ is the excess over the threshold, σ is a scale parameter and ξ is a shape parameter with $\xi = 0$ being the exponential distribution. Both σ and ξ can depend on covariates.

Smith [40] advocates viewing the high-level exceedances as points of a Poisson process. That is, for $X_1, \ldots, X_n$, suppose there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that $F^n(a_n x + b_n) \longrightarrow H(x)$ in distribution, where H is non-degenerate and let $Y_{n,i} = (X_i - b_n)/a_n$, $i = 1, \ldots, n$ and denote P_n the point process on $\Re^2$ with points $(\frac{i}{n+1}, Y_{n,i})$, $i = 1, \ldots, n$. The ordinates of P_n will tend to cluster near the lower endpoint of the (rescaled) distribution, but away from the boundary the process will look like a non-homogeneous Poisson process, the intensity of which is given by

$$\Lambda\{(t_1, t_2) \times (z, \infty)\} = (t_2 - t_1)[1 - \xi\tfrac{z-\mu}{\sigma}]^{1/\xi}$$

where $0 \leq t_1 \leq t_2 \leq 1$ and $1 - \xi\frac{z-\mu}{\sigma} > 0$.

The above has been extended to extreme values of dependent stochastic processes [40] (see section 3.2.3 below).

Smith and Huang [41] applied exceedance modeling to several data sets extracted from the Chicago ozone study. One goal was to model the probability that ozone on a given day exceeded a certain threshold as a function of a set of covariates.

They selected three stations where ozone was high and also considered daily maxima across the network. Separate days were assumed independent given the likelihood for the data by:

$$L = \textstyle\prod_i (p_i)^{\delta_i}(1 - p_i)^{1-\delta_i}$$

where p_i gives the probability that the threshold is exceeded on day i and δ_i

is an indicator of whether the threshold is, in fact, exceeded on day i. A logit model was used for p_i. Namely,

$$\log(\tfrac{p_i}{1-p_i}) = \textstyle\sum_j x_{ij}\beta_j$$

where x_{ij} is the value of the jth covariate on day i and β_j is the corresponding coefficient. They also considered the excesses over a threshold using the generalized Pareto distribution, which provided a good fit to the data [7].

3.2.3 Extremes of Dependent Sequences

The above analyses require the data to be independent. However, ground-level ozone data is dependent both in space and time making it necessary to incorporate such dependency into the models. Coles [3] discusses approaches to this problem for both stationary and non-stationary time dependent data. Davis [9] extends on ideas of LePage *et al.* [27] to establish nonnormal stable limits for normalized partial sums of dependent random variables.

Let $X_1, X_2, \ldots$ be a series of stationary time-dependent random variables. Define the following condition for asymptotic "*near-dependence*". For any $i_1 < \cdots < i_p < j_1 < \cdots < j_q$ with $j_1 - i_p > l$, if

$$|Pr\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\} -$$

$$Pr\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\} Pr\{X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\}| \leq \alpha(n, l), \quad (3.35)$$

where $\alpha(n, l) \longrightarrow 0$ for some sequence $\{l_n\}$ such that $\frac{l_n}{n} \longrightarrow 0$ as $n \longrightarrow \infty$,

then this condition of asymptotic "near-dependence" is met.

If condition (3.35) is met, then for $M_n = \max_{i=1,\ldots,n} X_i$ the following theorem holds.

Theorem 3.2.1

If there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{(M_n - b_n)/a_n \leq z\} \longrightarrow G(z)$$

where G is a non-degenerate distribution function, then for $u_n = a_n z + b_n$ for all $z \in \Re$, and assuming condition (3.35) is met, then G is a member of the GEV family of distribution functions (see, for example, Coles [3]).

Furthermore, another theorem gives more specifics as to the nature of G.

Theorem 3.2.2

If $X_1, X_2, \ldots$ are a stationary series with marginal distribution function, F, and $X_1^*, X_2^*, \ldots$ are a sequence of independent random variables also with marginal distribution, F, with $M_n^* = \max_{i=1,\ldots,n} X_i^*$ then if condition (3.35) holds, and under suitable regularity conditions

$$Pr\{(M_n^* - b_n)/a_n \leq z\} \longrightarrow G_1(z) \text{ as } n \longrightarrow \infty$$

with G_1 non-degenerate if and only if

$$Pr\{(M_n - b_n)/a_n \leq z\} \longrightarrow G_2(z),$$

where $G_2(z) = G_1^{\theta}(z)$ for a constant θ such that $0 < \theta \leq 1$ (see, for example, Coles [3]).

Subsequently, if $G_1(z) = \exp\{-[1 + \xi(\frac{z-\mu}{\sigma}]^{-\frac{1}{\xi}}\}$ then $G_2(z) = \exp\{-[1 + \xi(\frac{z-\mu'}{\sigma'})]^{-\frac{1}{\xi}}\}$ where for $\xi \neq 0$, $\mu' = \mu - \frac{\sigma}{\xi}(1 - \theta^{-\xi})$ and $\sigma' = \sigma\theta^{\xi}$. For $\xi = 0$, $\mu' = \mu + \sigma \log \theta$ and $\sigma' = \sigma$. The quantity θ is called the *extremal index*. Essentially, $\theta =$ (limiting mean cluster size)$^{-1}$. For independent series, $\theta = 1$, but if $\theta = 1$ it is not necessarily true that the series is independent.

In modeling the distribution of block maxima, it is appropriate to use the GEV family even if the data is dependent (provided it is stationary), but the validity of such a model may be questionable as the dependence in the series increases. For threshold exceedance models, extra care must be taken because of the tendency for extremes in a stationary series to cluster. One suggestion for handling this problem is to *decluster* the data in order to obtain a set of maxima data that is independent. There are many algorithms for declustering data, but mostly these algorithms rely on human intuition and cannot be performed automatically. One of the more popular methods, called runs declustering, creates clusters based on run lengths between threshold exceedances. Ferro and Segers [10] propose an automatic method for declustering that uses the extremal index to select an appropriate run length.

For non-stationary sequences it is not possible to establish a general theory similar to that for stationary processes. Instead, one can use the stan-

dard extreme value models as basic templates and enhance them by statistical modeling. For example, suppose $Z_t \sim \text{GEV}(\mu(t), \sigma, \xi)$ where $\mu(t) = \beta_0 + \beta_1 t$ for parameters β_0 and β_1. This allows variations through time in the observed process to be modeled as a linear trend in the location parameter of the appropriate extreme value model. Naturally, one can use more complicated models than a simple linear trend. Another option is to include covariates in the model for $\mu(t)$.

It is also possible to allow σ to depend on time, however, it is important to maintain the positivity of σ for all values of t. A useful model employs the exponential link function, $\sigma(t) = \exp(\beta_0 + \beta_1 t)$.

3.2.4 Multivariate Extreme Values

It should first be noted that there are a great many recent papers concerning multivariate extremes. However, because the emphasis in this work is on approaching the seasonal statistic from spatiotemporal methods, only a very cursory literature review is given here. The problem of characterizing the FHDA field is a multivariate problem so that it is important to investigate the joint distributions of random variables instead of simply looking at the marginal distributions conditional on other locations as described in the previous section. Multivariate extreme value theory is an active area of research, as is demonstrated in a very recent paper by Heffernan and Tawn [20], where they take multivariate extremes in an entirely different direction than any previous

work. The fundamental difference in their approach from previous approaches is that their method considers multivariate distributions where possibly only one component of a random vector is extreme–as opposed to limiting results where all of the components must be extreme. That is, previous work has predominantly been concerned with the characterization:

$$\lim_{n\longrightarrow\infty} \Pr\left[\frac{\max\{X_{11},\ldots,X_{n1}\}-b_{n1}}{a_{n1}} \leq z_1, \ldots, \frac{\max\{X_{11},\ldots,X_{nd}\}-b_{nd}}{a_{nd}} \leq z_d\right] =$$

$$G(z_1, \ldots, z_d), \tag{3.36}$$

where $(X_{i1}, \ldots, X_{id})$, $i = 1, 2, \ldots$ are independent identically distributed d-dimensional random vectors, a_{ij}, b_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, d$ are normalizing constants, and G is a non-degenerate d-dimensional distribution function. Under weak conditions (Resnick [36] Ch. 5), there exists a real number, θ, such that the normalized maximum of all the variables converge in distribution to the Fréchet. That is,

$$G(z) = \exp(-\theta/z) \tag{3.37}$$

One important result is that representation (3.36) is equivalent to multivariate regular variation (Resnick [36]). This discovery has led to many parametric families for multivariate extreme value distributions, as well as the development of threshold methods for multivariate extremes (see, for example, Richard Smith's discussion on Heffernan and Tawn [20]).

Much work on multivariate extremes has addressed the extremal coefficient,

θ (Schlather and Tawn [38], Ferro and Segers [10], Coles [3]). Specifically, G from (3.36) is given by

$$G(z_1, \ldots, z_d) = \exp\{-V(z_1, \ldots, z_d)\}, \tag{3.38}$$

where $V(z_1, \ldots, z_d) = \int \max_i(\frac{w_i}{z_i}) dH(w_1, \ldots, w_d)$ and H is a measure with all marginal expectations equal to 1. From (3.38) and (3.37), we have that

$$\theta = \int \max_i w_i dH(w_1, \ldots, w_d).$$

3.3 Nonstationarity

The interaction of chemical and physical atmospheric processes that tend to produce data patterns for ground-level ozone with large spatial variability are nonstationary processes, in the sense that the spatial structure varies with location. For small regions stationarity is often a reasonable assumption, however, it is often of interest to look at larger regions. Fuentes [11] proposed modeling a nonstationary process locally as a stationary random field with some parameters that describe the local spatial structure. These parameters are then allowed to vary across space to reflect the lack of stationarity over a large region–such as the eastern United States (Fig. 2.1).

Specifically, Fuentes [11] assumes the model

$$Z(\mathbf{x}) = \sum_{i=1}^{M} Z_{\theta(\mathbf{s}_i)}(\mathbf{x}) K(d(\mathbf{x}, \mathbf{s}_i)) \tag{3.39}$$

where each $Z_{\theta(\mathbf{s}_i)}$ represents a spatial field around one of M nodes and the $Z_{\theta(\mathbf{s}_i)}$ fields are orthogonal with covariance $C_{\theta(\mathbf{s}_i)}$.

A valid nonstationary covariance for Z, the spatial field for the entire region, is given by

$$\text{Cov}(Z(\mathbf{x}), Z(\mathbf{y})) = \sum_{i=1}^{M} K(d(\mathbf{x}, \mathbf{s}_i))K(d(\mathbf{y}, \mathbf{s}_i))C_{\theta(\mathbf{s}_i)}(d(\mathbf{x}, \mathbf{y})) \tag{3.40}$$

Fuentes uses Matérn covariances (3.9) to model the individual covariances, $C_{\theta(\mathbf{s}_i)}$, $i = 1, \ldots, M$, where $\theta(\mathbf{s}_i) = (\nu_i, \sigma_i, \rho_i)$.

Other methods proposed for incorporating nonstationarity into a spatial process include: the deformation approach of Sampson and Guttorp [39], a moving windows approach (Haas [18]), and an extension of the *empirical orthogonal functions* (EOF) (Nychka *et al.* [32]). Higdon *et al.* [21] propose nonstationary spatial covariances based on convolutions of kernels. Specifically, let $\psi(\cdot)$ be a Gaussian white noise process with convolution kernel $K_{\mathbf{x}}$ centered at the point $\mathbf{x}$ with its shape being a function solely of location $\mathbf{x} \in \Re^2$, then the correlation between two points $\mathbf{x}$ and $\mathbf{x}'$ is proportional to

$$\int_{\Re^2} K_{\mathbf{x}_i}(\mathbf{u})K_{\mathbf{x}_j}(\mathbf{u})d\mathbf{u}.$$

Then, the the process

$$Z(\mathbf{x}) = \int_{\Re^2} K_{\mathbf{x}}(\mathbf{u})\psi(\mathbf{u})d\mathbf{u}.$$

is valid provided $\sup \int_{\Re^2} K^2_{\mathbf{x}}(\mathbf{u})d\mathbf{u} < \infty$. Note that a major difference between this approach and that of Fuentes [11] is that, here, each location has its own kernel function. Paciorek [33] generalizes the approach of Higdon *et al.* [21] to form a class of nonstationary correlation functions.

Here, a brief outline of Paciorek's work is given. A fundamental theorem of Schoenberg [37] states that the class of functions positive definite on Hilbert space is identical with the class of functions of the form

$$R(\tau) = \int_0^\infty \exp(-\tau^2 s) dH(s), \tag{3.41}$$

where $H(\cdot)$ is non-decreasing and bounded, and $s \geq 0$. The class of positive definite functions on Hilbert space is identical to the class of functions that are positive definite on $\Re^d$ for every $d \in \{1, 2, \ldots\}$. The result of Paciorek [33] is as follows.

Theorem 3.3.1

If an isotropic correlation function, $R(\tau)$, is positive definite on $\Re^d$ for every $d \in \{1, 2, \ldots\}$, then the function, $R(\cdot, \cdot)$ defined by

$$C(\mathbf{x}_i, \mathbf{x}_j) = \frac{2^{d/2} |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}}}{|\Sigma_i + \Sigma_j|^{\frac{1}{2}}} R(\sqrt{Q_{ij}}), \tag{3.42}$$

is a positive definite nonstationary correlation function, where $R(\cdot)$ is as in (3.41), and Q_{ij} is given by

$$Q_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \left[\frac{\Sigma_i + \Sigma_j}{2} \right]^{-1} (\mathbf{x}_i - \mathbf{x}_j).$$

3.4 Previous Work on Ozone Modeling and the FHDA field

There are numerous statistical papers associated with ground-level ozone. Because the new NAAQS for ozone has only recently been enacted, most of these papers either deal with the old 1-hour standard, regulation in other countries, or are not concerned with regulation. Therefore, there is not much literature on interpolating this new standard off of the network. Fuentes [11] applies the nonstationary model (3.39) to predict the FHDA on a grid for the entire eastern United States. Here, I will go into a little more detail on this approach, and then discuss some other work related to the new standard. Finally, I give a very brief summary of some of the other statistical work related to ozone.

3.4.1 Interpolating NAAQS for Ozone off of Monitoring Network

Fuentes [11] presents a new statistical model for interpolation of nonstationary processes also with the objective of estimating ground-level ozone concentrations on and off of the monitoring network for determination of attainment of the NAAQS. She applies a Bayesian framework for interpolation and calculates the predictive posterior distribution (ppd) in place of finding a point prediction. This provides more information about the distributional characteristics of the FHDA field than a simple point estimate can provide. Simulated val-

ues from the ppd are used to estimate the probability of non-attainment at a particular location by calculating the proportion of simulated values from the ppd that are out of compliance.

Fuentes [11] applies model (3.39) with covariance given by (3.40) using a Bayesian paradigm to make inferences on the FHDA field. The FHDA field was found using 1995 to 1997, 1996 to 1998 and 1997 to 1999 to obtain the three year averages of FHDA. Non-attainment regions were typically near big cities. In the northeast, they included regions in and around Baltimore, Washington, D.C., Philadelphia, New York, Hartford, Providence and Boston. Other regions were in and around Pittsburgh, Atlanta, Memphis, Charlotte, Gary, Saint Louis, Dayton and Cleveland. Fuentes found that 1996 was associated with lower ozone, but that it also had a particularly cool summer, which is likely the reason for the lower ozone. Similarly, 1998 had a hot, dry summer with higher ozone levels at least in Kentucky, West Virginia and Tennessee. Changes found for these areas for 1999 (which uses the average of 1997 to 1999) by Fuentes [11] were most likely due to changes in meteorology rather than emissions of NO_x and volatile organic compounds.

Fuentes [11] also compared the region around the Research Triangle Park (RTP) in North Carolina to Atlanta and Indianapolis. All three, on average, were out of compliance, but each had very different distributions. Indiana and Atlanta both had an average of 96 ppb while RTP had an average of 88 ppb. Atlanta had much more variability (standard deviation of 5.8 ppb) than either

RTP (1.7 ppb) or Indiana (2.2 ppb).

It should be noted that Fuentes [11] does not perform any cross-validation to give an independent validation of the model.

3.4.2 Other Work Related to the NAAQS for Ozone

Davis and Speckman [8] look at predicting both the maximum and the 8-hour average for one day into the future, but not off of the network. Their region of interest was Houston, Texas with 11 stations for 214 days from April through October and for the years from 1981 through 1992. They also investigated the incorporation of numerous meteorological covariates. Ultimately they settled on just a few of the meteorological components. Specifically, averages of hourly variables including: early morning, mid-morning and daytime wind components, daytime opaque cloud cover, maximum one day lagged ozone, daily maximum temperature and morning mixing depth.

Davis and Speckman [8] tested many different models including linear regression, non-linear regression and neural nets, but the best model was found to be loess/generalized additive model approach (GAM). The root mean squared error for 8-hour averages forecasts ranged from 13.2 to 16.3 ppb with an R^2 in the range 0.66 to 0.73 and for the maximum ozone the root mean squared error ranged from 18.5 to 22.0 ppb with 0.61 to 0.68 for the R^2 coefficient of correlation. It was found that the one day lag term removed the need to fit an overall trend. The forecasts did well generally, but forecasts for meteorological

transitions, such as frontal passages, were unreliable and subsequently their model did not attain the predictive capabilities that had been desired.

3.4.3 Other Work Relating to Ground-level Ozone

Guttorp *et al.* [17] examine hourly ozone data collected as part of a model evaluation study for ozone transport in the San Joaquin Valley of California. They looked at a space-time analysis of ground level ozone and performed spatial interpolation off of the network for the hourly ozone data, which had periodic effects during the day in addition to day-to-day cycles. Their data came from the SARMAP study, which is a regional model application project for this area and includes surface records of many pollutants as well as meteorological variables. They looked at 326 monitoring sites in the San Joaquin Valley with particular emphasis on a subset of 17 stations around Sacramento, California.

Among Guttorp *et al.*'s [17] findings is a distinction between stations at high elevations from those at lower elevations. Particularly, they noticed that those at the higher elevations had a flatter 24-hour temporal mean curve implying that the depletion of ozone by chemical reaction is more rapid below the nocturnal inversion layer. In both cases, there were strong temporal correlations in the residuals from subtracting station specific hourly means. They tried several time series models, but found that an AR(2) model was sufficient over more complicated ARIMA or ARMA models.

Chapter 4

Comparison of Daily and Seasonal Models for Ozone

The application of statistical techniques to environmental problems often involves a trade-off between simple methods that are easily implemented and interpreted, and more complicated methods that may have smaller errors. In this chapter, simple and complicated statistical models are compared for interpolating the NAAQS for ground-level ozone off of a network of monitoring sites. It should be noted that the methods discussed here are applicable to other problems besides FHDA. For example, if the standard were to be changed to a different order statistic, such as third-highest daily maximum 8-hour average ozone, the techniques can still be used with only minor modifications. Here, a small homogeneous subset of 72 monitoring sites around North Carolina, where the standardized daily field is assumed to be stationary, are examined.

The new NAAQS (section 2.4) presents a new statistical problem for inferring regions of attainment or non-attainment because it is not clear that the FHDA field (a field of order statistics) is Gaussian–a fundamental assumption of most standard spatial statistical techniques.

Although the use of spatial statistics for interpolating air quality measurements would not be disputed by a statistical audience, surprisingly the use of monitoring data in a regulatory context is often limited to point locations. Accordingly, Holland *et al.* [22] argue for the introduction of modern statistical methods to understand the spatial and temporal extent of pollution fields based on monitoring data. Given the range of statistical backgrounds associated with the regulatory community, it is appropriate to propose statistical methods that are simple and understandable to a broad group when such methods provide an accurate and defensible analysis. In particular, for interpolating the NAAQS, it is useful to ascertain the feasibility of approximate statistical methods that treat the FHDA statistics directly. From this perspective, two statistical models are compared. The first, a fairly complex model, uses a spatial AR(1) model for daily ozone measurements and samples the FHDA field conditional on the data for the entire season. This approach will be referred to as the *daily model*. The second model, referred to the *seasonal* model, is a geostatistical model that predicts the FHDA field from the network values using standard best linear unbiased estimation, or kriging (section 3.1.1). This seasonal model is similar to the model proposed by

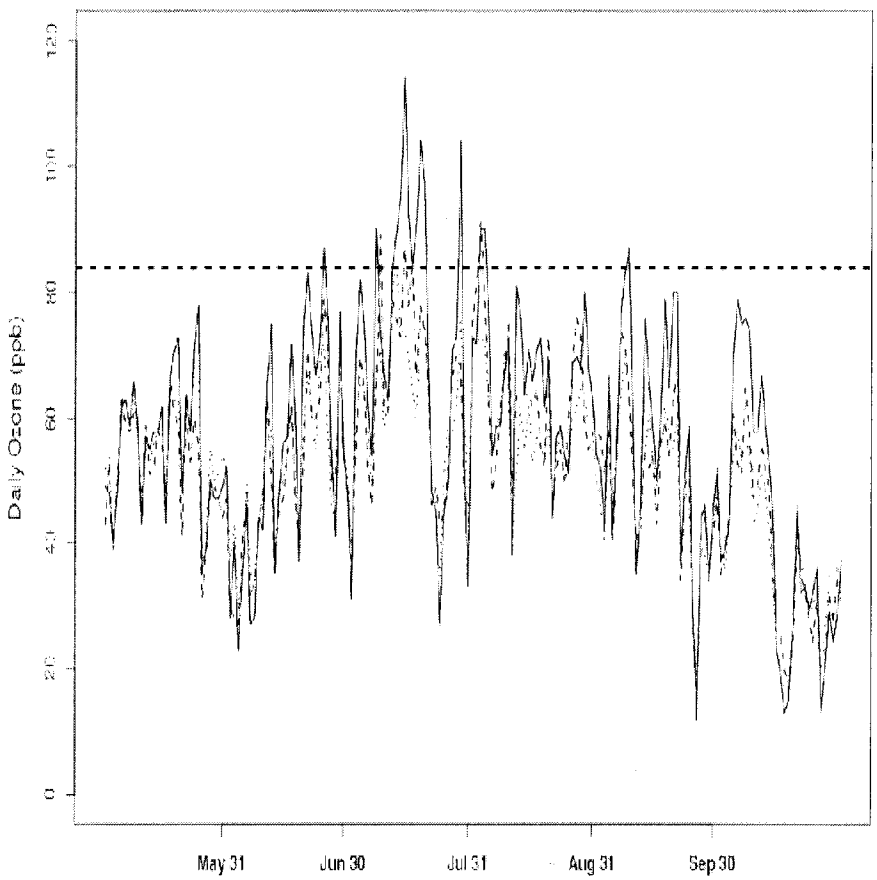


Figure 4.1: Time series of ground-level ozone (ppb) for the three monitoring stations circled in Fig. 2.2 for 1997 ozone season.

Fuentes [11], except that the region of interest here is much smaller and so can be assumed to be spatially and temporally stationary. A third approach that will be used as a benchmark estimates the FHDA field by way of a thin plate spline (section 3.1.4). This last method is generic and uses the least amount of information concerning the actual air quality context.

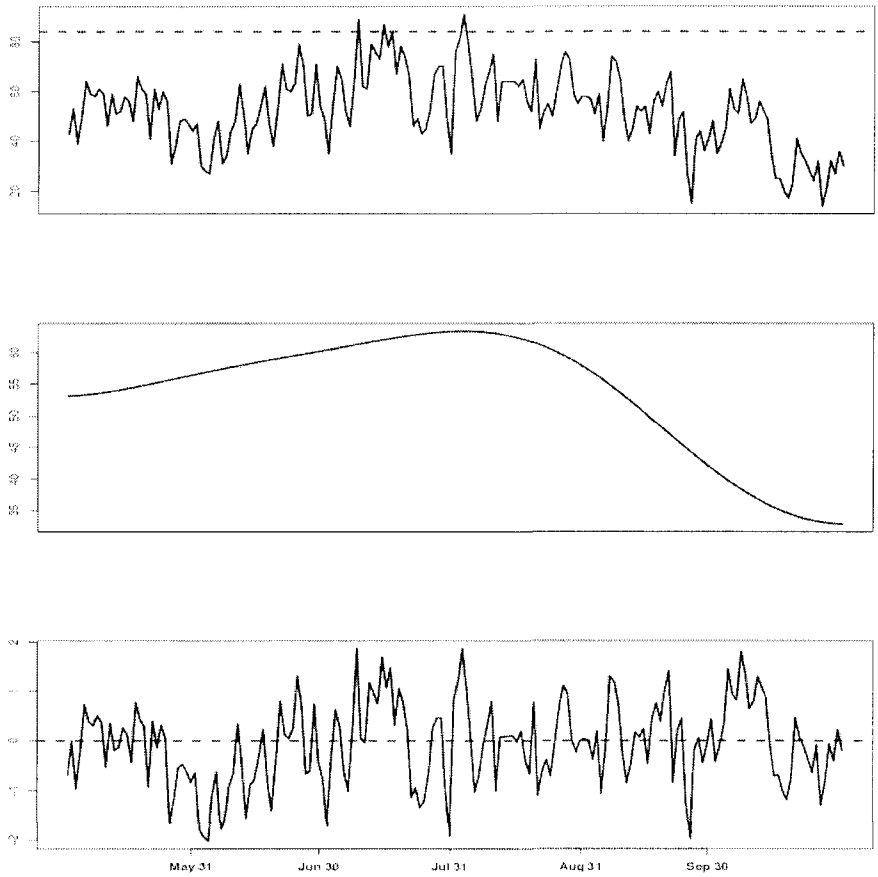


Figure 4.2: Time series of standardized ground-level ozone for 1997 ozone season at one station for (top) daily maximum 8-hour average ozone (ppb), (middle) regression fit and (bottom) residuals (de-seasonalized/standardized daily ozone).

4.1 Fitting an AR(1) Spatiotemporal Model

4.1.1 Standardizing the Data

Ozone has a seasonal effect even during the relatively short ozone season described in section 2.3.1. In fact, inspection of Fig. 4.1 suggests that there is seasonality in the daily data. It is useful to account for this seasonality as a fixed effect before modeling space-time structure.

Recall that for the ozone NAAQS example, daily values are the daily maxi-

mum 8-hour average ozone. Therefore, let $O(\mathbf{x}, t)$ denote the maximum 8-hour average ozone at location $\mathbf{x}$ and day t. The following standardization is used for the daily maximum 8-hour ozone measurement.

$$O(\mathbf{x}, t) = \mu(\mathbf{x}, t) + \sigma(\mathbf{x})u(\mathbf{x}, t) \tag{4.1}$$

and it is assumed that $u(\mathbf{x}, t)$ for any given location and time has mean zero and variance one. Note that μ is a function of both time and space in order to remove any seasonality.

The seasonal means are smoothed over space using a singular value decomposition approach. First, the m individual station time series are regressed on an intercept and three sine and cosine pairs with periods 365, 365/2 and 365/4. Here, three sine and cosine pairs were found to be adequate based on F-statistics from the regression fits, and the periods chosen in order to capture an overall yearly, semi-annual and quarterly trend. However, other periods and numbers of sine and cosine pairs could be used based on the data being analyzed.

Let $\mathbf{B}$ denote the $m \times 7$ matrix of regression coefficients across all the stations, then $\mathbf{B}$ can be decomposed as $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{D}$ is a diagonal matrix of the singular values of $\mathbf{B}$. By setting some of the singular values of $\mathbf{D}$ to zero (call the resulting matrix $\mathbf{D}^*$), the multiplication $\mathbf{B}^* = \mathbf{U}\mathbf{D}^*\mathbf{V}^T$ yields a matrix of a constrained set of the original regression parameters, having reduced the variability across stations. For the analyses here, the first three principle components, which

explain 96% of the variation, are retained (i.e., the last four singular values were set to zero); and, in this case, results smooth the estimated parameters over space. Each of the components include a relatively large loading for the intercept term, and the first and second components give slightly differing weights to each of the periods. The first gives more weight to the annual period, while the second gives more weight to the semi-annual and quarterly periods. Finally, the estimates of μ and σ based on station locations are extrapolated to unobserved locations using thin plate spline interpolation.

The daily model approach uses Monte Carlo simulations to generate a sample from the FHDA distribution. Specifically, a space-time model is used to simulate daily ozone data at arbitrary locations conditional on what is observed for an entire season of ozone, and from this sample the fourth-highest value is taken. This process is repeated many times, say 1000, to achieve the desired sample.

4.1.2 The Daily Model

Given the standardized process, $u(\mathbf{x}, \mathbf{t})$, consider the spatial autoregressive models.

$$u(\mathbf{x}, t) = \rho(\mathbf{x})u(\mathbf{x}, t-1) + \varepsilon(\mathbf{x}, t) \tag{4.2}$$

and

$$u(\mathbf{x}, t) = \rho_1(\mathbf{x})u(\mathbf{x}, t-1) + \rho_2(\mathbf{x})u(\mathbf{x}, t-2) + \varepsilon(\mathbf{x}, t) \tag{4.3}$$

Here, the shocks, $\varepsilon(\mathbf{x}, t)$, are assumed to be independent over time and be a mean zero Gaussian process over space. For the AR(1) process, this covariance is given by

$$\sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x}')}\psi(d(\mathbf{x},\mathbf{x}')) \tag{4.4}$$

Here, the covariance (4.4) is considered to be isotropic and stationary with $d(\mathbf{x},\mathbf{x}')$ the great circle distance transformed to ensure that ψ is positive definite (i.e., $d(\mathbf{x},\mathbf{x}') = 2\sin(h/2)$, where h is the angular great circle distance) (see, for example, Gneiting [15], and Gaspari and Cohn [13]). Equation (4.2) implies a space-time covariance function

$$C(u(\mathbf{x},t), u(\mathbf{x}', t-\tau)) = \frac{(\rho(\mathbf{x}))^\tau\sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x}')}\psi(d(\mathbf{x},\mathbf{x}'))}{1-\rho(\mathbf{x})\rho(\mathbf{x}')}, \tau \geq 0 \tag{4.5}$$

Thus, if the AR(1) parameters are not constant over space, then (i) the spatial process $u(\mathbf{x}, t)$ is not stationary even if the shocks are stationary in space and (ii) covariance (4.5) is not space-time separable. Note that for $\mathbf{x} = \mathbf{x}'$, (4.5) reduces to $(\rho(\mathbf{x}))^\tau$ so that for $\tau = 0$, $Var(u(\mathbf{x}, t) = 1$ and the covariance at one location and two different times is a function of the autoregressive coefficient raised to the time lag. For $\rho(\mathbf{x}) = \rho$ (4.5) reduces to $\rho^\tau\psi(d(\mathbf{x},\mathbf{x}')$. Additionally, covariance (4.5) is not fully symmetric (i.e., $\mathrm{Cov}(u(\mathbf{x}, t), u(\mathbf{x}', t-\tau))$ and $\mathrm{Cov}(u(\mathbf{x}', t), u(\mathbf{x}, t-\tau))$ generally differ). The violation of full symmetry is physically justifiable here because ozone is often transported by wind in only one direction.

4.1.3 Sampling the distribution of FHDA conditioned by the monitoring data

Under the assumption that all the components of the data model are known, there is a straightforward algorithm for sampling the FHDA field conditional on the observed data. This algorithm is quite efficient and uses the autoregressive structure over time to recursively generate the daily process. Let $\mathbf{x}_0$ be a location where ozone is unobserved. A spatial prediction for the FHDA at this location involves two steps. First, a sample of the time series of daily ozone measurements at this location conditional on the observed data (for all locations and all times) is obtained, then the FHDA for this series is found. By elementary probability, the resulting FHDA statistics will be a sample of the FHDA field at $\mathbf{x}_0$ conditional on the data. Repeating these two steps, one can generate a random sample that approximates the FHDA conditional distribution; and, of course, the sample mean is a point estimate for the conditional expectation of the FHDA at $\mathbf{x}_0$. The conditional variance can be used as a measure of uncertainty.

Sampling from the conditional distribution of the ozone is simplified by the autoregressive structure over time and the restriction of spatial dependence to the shocks in the AR(1) innovation. Here, all parameters $(\mu(\mathbf{x}, t), \sigma(\mathbf{x}), \rho(\mathbf{x}), \psi)$ are fixed quantities and assumed known. Also let $\{\mathbf{x}_k, \text{ for } 1 \leq k \leq m\}$ be the station locations. Based on these assumptions it is sufficient to find the

conditional distribution of $\{u(\mathbf{x}_0, t), 1 \leq t \leq T\}$ given $\{u(\mathbf{x}_k, t),\ 1 \leq t \leq T,$ and $1 \leq k \leq m\}$ because the standardized random variables can always be transformed back to the raw scale of the measurements. Moreover, the σ-field of $\{u(\mathbf{x}, t),\ 1 \leq t \leq T\}$, for any $\mathbf{x}$ is equivalent, through the autoregressive relationship, to the field generated by $\{u(\mathbf{x}, 1), \varepsilon(\mathbf{x}, t),\ 2 \leq t \leq T\}$. Recall that the AR shocks are temporally independent so that the conditional distribution for the ozone fields at $\mathbf{x}_0$ can be found based on the much simpler conditional distribution of $\varepsilon(\mathbf{x_0}, t)$ given $\{\varepsilon(\mathbf{x_k}, t),\ 1 \leq k \leq m\}$. Thus, it is easy to generate a conditional ozone field by considering the conditional field of the AR(1) shocks and then transforming these results to the original scale of measurements.

An approximate algorithm for the conditional sampling of the FHDA field is now summarized below.

1. Initialize the time series by interpolating $u(\mathbf{x}_0, 1)$ from $u(\mathbf{x}_k, 1)$ $1 \leq k \leq m$.

2. For t in 2 to T sample the spatial shocks from $[\varepsilon(\mathbf{x_0}, t)|\{\varepsilon(\mathbf{x_k}, t),\ 1 \leq k \leq m\}]$.

3. Accumulate the sampled shocks and initial values using the autoregressive relationship (4.2) to obtain a conditional realization of the standardized process $u(\mathbf{x}_0, t)$.

4. Unstandardize and compute the FHDA at $\mathbf{x}_0$ based on the series simu-

lated for the ozone season.

Note that the shocks at a station location are based on the actual daily observations so that the sample is tied explicitly to the data. If, in fact, $\mathbf{x}_0$ is at a station location, and the spatial process has a zero nugget variance, then the resulting conditional sample will simply be the observed data. Thus the "conditional realization of the FHDA field" will be the FHDA statistic for that station's measurement.

It should be noted that this algorithm works because complete observations at the station locations are assumed. It would be more complicated if observations were sparse over time. For these data there are no instances where there are missing observations at a given time point at every station. Therefore, when shocks are sampled from the conditional distribution, locations that have missing values are simply not used in the calculation for that time point.

Although it is possible to sample in Step 1 exactly, approximate sampling was done from a geostatistical model fit to the standardized fields. Because $\rho(\mathbf{x})$ does not vary greatly (estimates are all within the range of about 0.5 to 0.7, with approximately 70% of them within the range of 0.5 to 0.6), the covariance of the standardized process is stationary. Subsequently, based on data from an entire season, the data will be well estimated by geostatistical methods. Moreover, based on the magnitude of the autoregressive coefficients, the time series becomes nearly independent of the initial value after several

days. For these reasons, the approximation is adequate.

If it is desired to obtain a sample from a different distribution than that of the FHDA, it is a small matter of computing a different statistic in Step 4 of this algorithm. For example, one could take the third-highest value of the simulated season of data, or even the mean.

In this algorithm it is straightforward to replace the conditional sampling of a single location with a vector, or grid of locations. Thus one obtains a conditional field with spatial and temporal dependence among the grid points consistent with the space-time model. In addition, this algorithm can be modified simply to simulate a space-time process that follows this model. In this case one does not condition on observed data, and one substitutes an unconditional sample for the conditional sample of the shocks at Step 2. This unconditional sampling is used in the next section to identify an approximate Gaussian model for the FHDA field.

4.2 Bivariate Fourth-highest Order Statistic Distribution

In this section, I argue that, despite being a field of order statistics, the FHDA field is approximately multivariate Gaussian. A bivariate sample of size 1000 of fourth-highest order statistics from samples of bivariate mean-

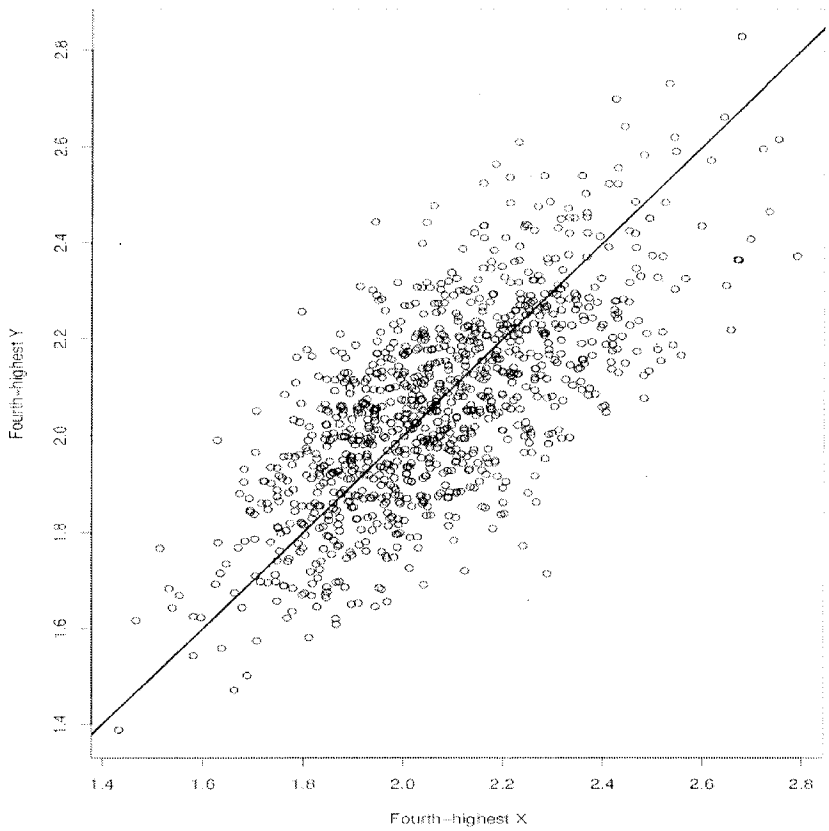


Figure 4.3: Bivariate sample of size 1000 of fourth-highest order statistics taken from samples of independent bivariate mean-zero, variance-one normal random variables, (X, Y), with correlation 0.95 and size 184.

zero, variance-one normal random variables with correlation 0.95 and size 184 was found. Note that resulting pairs of fourth-highest values do not necessarily correspond to pairs from the original bivariate samples, and the resulting correlation is about 0.70. Inspection of this particular sample suggests that for the bivariate case, the assumption of approximate bivariate normality is reasonable (see, for example, Fig. 4.3), so it may be reasonable for the multivariate case. Note that the correlation between pairs dropped by about 25% after taking the fourth-highest value. For smaller correlations, this suggests that multivariate fourth-highest random variables may be approximately independent so that

checking for approximate marginal normality may be sufficient.

Let X and Y be independent identically distributed bivariate random vectors with joint cdf, $F_{X,Y}(x,y)$, and marginal cdf's, $F_X(x)$ and $F_Y(y)$ respectively, and let $X' = X_{(n-3):n}$ and $Y' = Y_{(n-3):n}$ denote the fourth-highest order statistic for each component. The joint cdf for the bivariate fourth-highest order statistics, $P\{X' \leq x, Y' \leq y\}$, is given by the following.

$$\sum_{k=0}^{3}\sum_{j_1=0}^{3}\sum_{j_2=0}^{3}\binom{n}{n-k,j_1,j_2,k-j_1-j_2}F_{X,Y}^{n-k}(x,y)\{F_Y(y)-F_{X,Y}(x,y)\}^{j_1}\times$$

$$\{F_X(x)-F_{X,Y}(x,y)\}^{j_2}\{1-F_X(x)-F_Y(y)+F_{X,Y}(x,y)\}^{k-j_1-j_2} \qquad (4.6)$$

The marginal cdf for the fourth-highest order statistic $X' = X_{(n-3):n}$ is given by

$$\Pr\{X' \leq x\} = \sum_{k=0}^{3}\binom{n}{k}F_X^{n-k}(x)(1-F_X(x))^k,$$

which leads to the pdf

$$f_{X'}(x) = \binom{n}{n-4,1,3}f_X(x)F_X^{n-4}(x)(1-F_X(x))^3 \qquad (4.7)$$

The exact density (4.7) for the fourth-highest order statistic with $n = 184$ is shown in Fig. 4.4 along with a normal density with mean 2.06 and variance 0.0441, and it appears that the exact fourth-highest distribution can be well approximated by a normal density. However, perhaps a better diagnostic are the normal quantile-quantile plots shown in Fig. 4.5 for predicted FHDA (from the daily model) for a single grid point in two different years, where one year the grid point was predicted in attainment, and the other year out

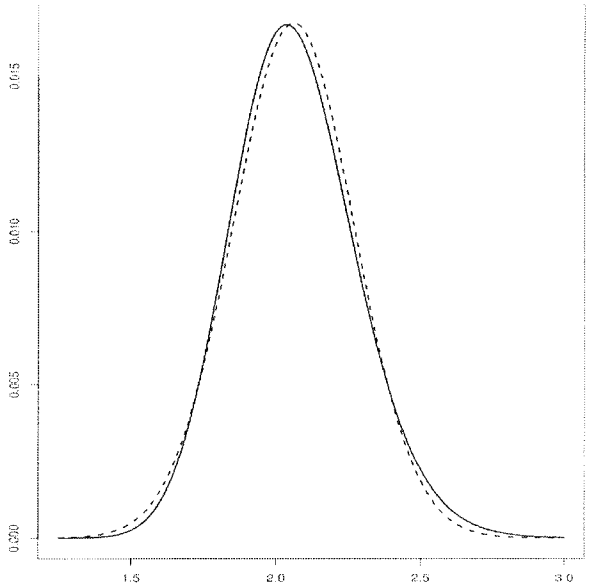


Figure 4.4: Exact marginal fourth-highest order statistic $X' = X_{(n-3):n}$ density function for $n = 184$ with underlying $X \sim \mathrm{N}(0, 1)$ (solid line), and normal density function with $\mu = 2.06$ and $\sigma^2 = 0.0441$ (dashed line).

of attainment. The figure shows that the distribution of FHDA at this grid point does not deviate substantially from normality in either case. Such plots at other grid points showed similar results for, at least, marginal normality.

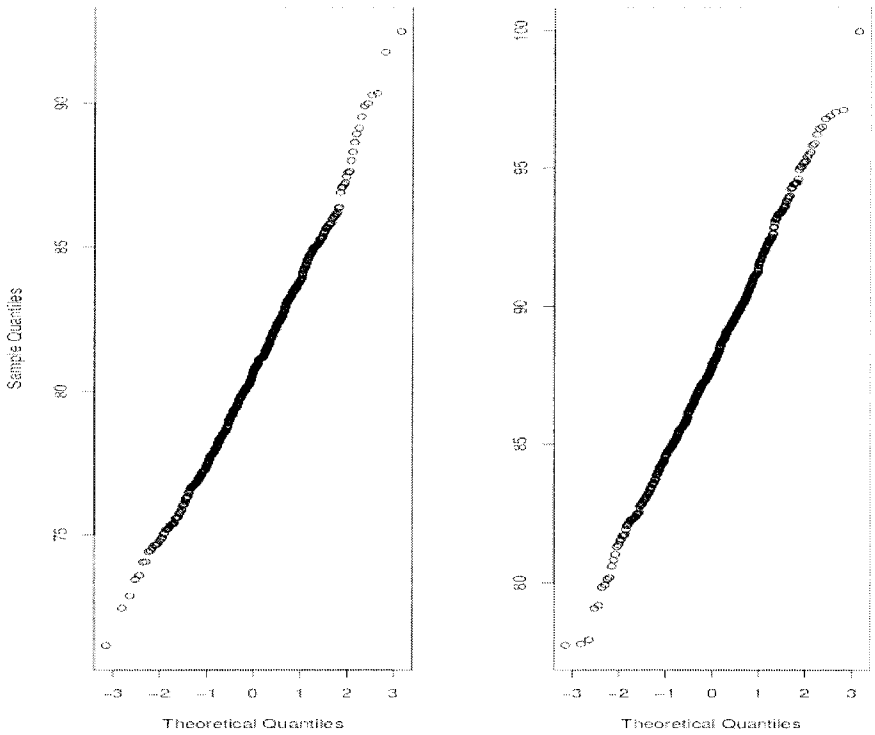


Figure 4.5: Normal quantile plots for simulated (daily model) FHDA values for a grid point in the RTP region in 1997 (left) and 1999 (right). Note that this grid point is predicted by the daily model to be in attainment of the NAAQS for ground-level ozone in 1997, but not in 1999.

4.3 The Seasonal Model

For the seasonal model the supposition is that the FHDA field is approximately Gaussian distributed, so the main modeling issue is to derive a suitable covariance function. Results from section 4.2 suggest that the assumption that the distribution of the fourth-highest order statistic is approximately normal is reasonable.

To estimate the covariance function for FHDA, it is convenient to use Monte Carlo simulations similar to those used for the daily model approach in order to look at the correlogram of the FHDA field and fit a function. The algorithm

used to do this is essentially the same as that of the daily model (section 4.1.2) except that the spatial shocks are sampled from an unconditional distribution in step 2. Specifically,

$$\varepsilon(\mathbf{x}, t) \sim \text{Gau}(\mathbf{0}, \mathbf{\Sigma}_\varepsilon) \tag{4.8}$$

where $\mathbf{\Sigma}_\varepsilon$ is given by (4.4). A covariance function can be derived from empirical correlations of the sample from the unconditional FHDA field. For comparison, a variogram derived from the observed FHDA field is also used.

4.4 Results of the Two Models

4.4.1 Results from the Daily Model

In the daily model approach, the data of interest for spatial prediction is the standardized daily maximum 8-hour average, which is the average ozone reading taken for each 8-hour block of time in a given day; the maximum of each of these blocks of time is the record for that day. Supposing that the ozone readings are normally distributed, taking the averages over blocks of time will preserve the normality. However, maximum values are typically not normally distributed. Despite having many 8-hour intervals in each day, the maximums tend to occur around the same time of day each day; and so is analogous to taking the average of the 8-hour interval centered around 2:00pm (see, for example, Davis *et al.* [6]). For this reason, and that the daily value is an average over this interval, it is possible that the data are, at least

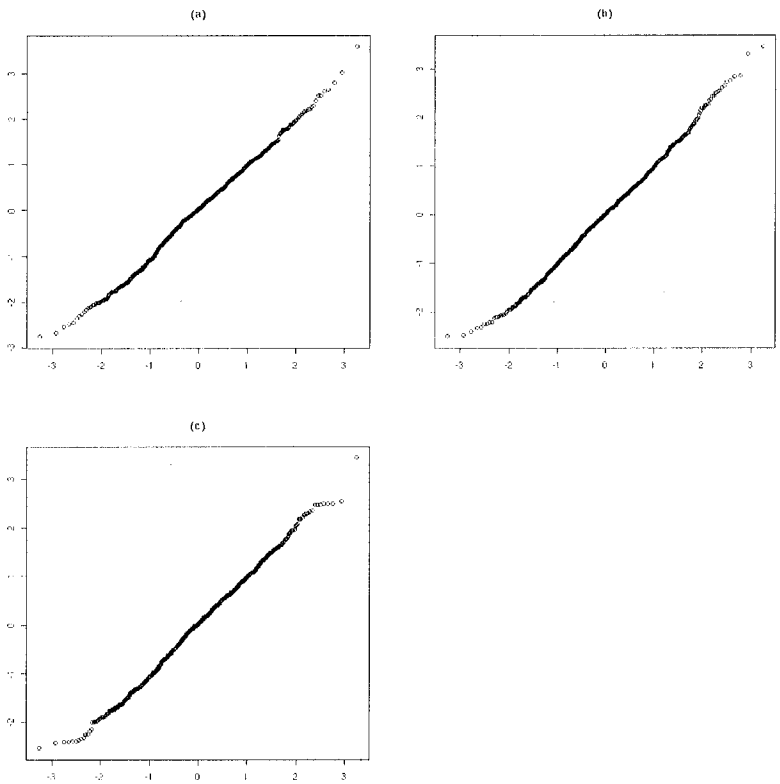


Figure 4.6: Normal QQ-plots for standardized daily maximum 8-hour average ozone levels from the three stations (circled in Fig 2.2).

approximately, normal.

Examination of normal quantile plots and histograms for standardized daily maximum 8-hour averages, u of Eq. (4.1), shows that, at least approximate, normality is a reasonable assumption--Figs. 4.6 and 4.7 show these plots for three stations (circled in Fig. 2.2).

Partial autocorrelation plots (Fig. 4.8) suggest that it is reasonable to use a spatial AR(1) model, as discussed in section 3.1.10, to estimate the daily field.

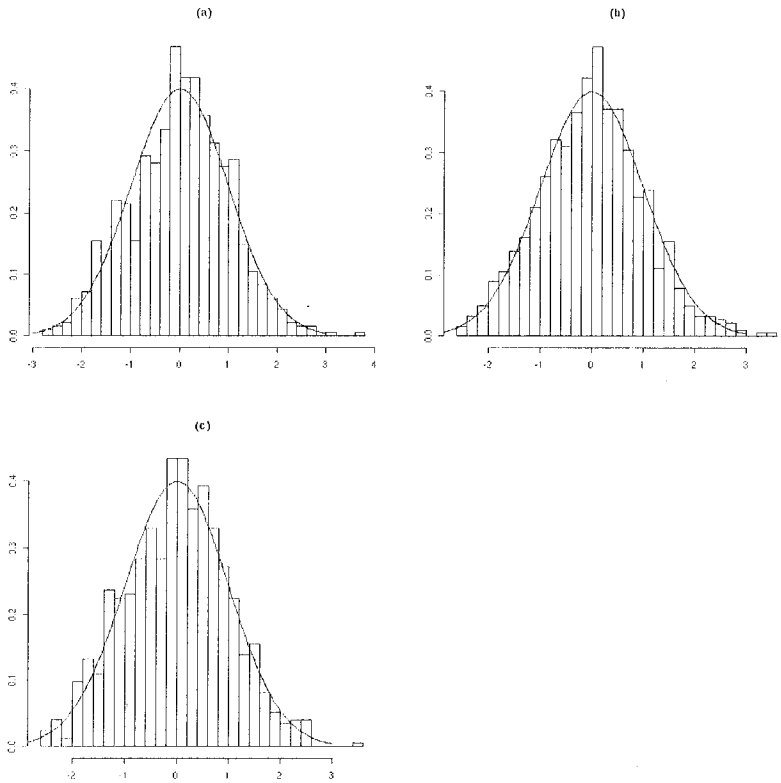


Figure 4.7: Histograms of standardized daily maximum 8-hour average ozone levels from the stations (circled in Fig 2.2) with standard normal density (solid line).

Further, Fig. 4.9 shows box plots of the estimated autoregressive correlation coefficients $\hat{\rho}_1(\mathbf{x})$ and $\hat{\rho}_2(\mathbf{x})$. Clearly, $\hat{\rho}_2$ is near zero for all stations suggesting its contribution is negligible. The plot of $\hat{\rho}_1$ suggests that the estimates are similar to those found in fitting the AR(1) model. Additionally, empirical correlograms for the AR(1) (Fig. 4.10) and AR(2) (Fig. 4.11) shocks also appear to be very similar; providing further evidence that an AR(1) model is sufficient.

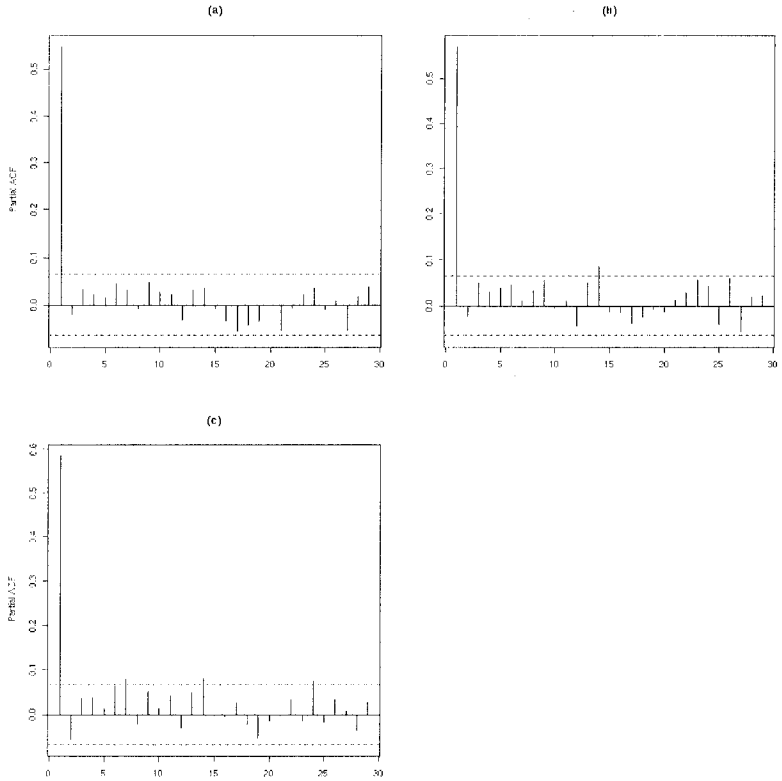


Figure 4.8: Partial Autocorrelation Function (PACF) for stations (a) 11, (b) 32 and (c) 34 (circled in Fig. 2.2).

The AR(1) parameters vary across stations (Fig. 4.12), which is to be expected even under stationarity. To assess stationarity of the autoregressive shocks, a local correlogram is fit for each station location using a *single* exponential covariance function. Standard errors of parameters were found using a parametric bootstrap. The estimated nugget variance and range parameters do not vary significantly across the domain–each having a small range and standard deviation; from 0.83 to 1.03 ppb (0.04 ppb) for the estimated nugget

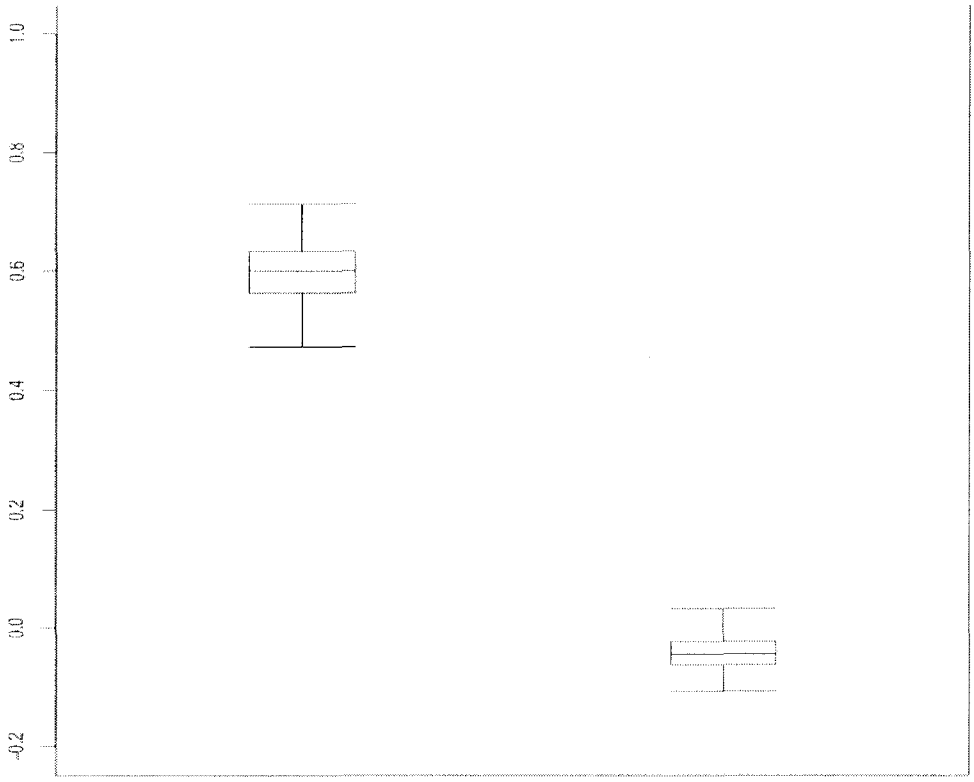


Figure 4.9: Box plots of estimated AR(2) coefficients $\hat{\rho}_1$ (left) and $\hat{\rho}_2$ (right).

and 164 to 328 miles (33 miles) for the estimated range–suggesting that the spatial shocks field can be approximated by a stationary process (Fig. 4.13). Further, the largest differences in the estimates generally occur on the edges to the West and along the southern coast (Fig. 4.14).

The general shape of the empirical correlogram (Fig. 4.10) suggests fitting

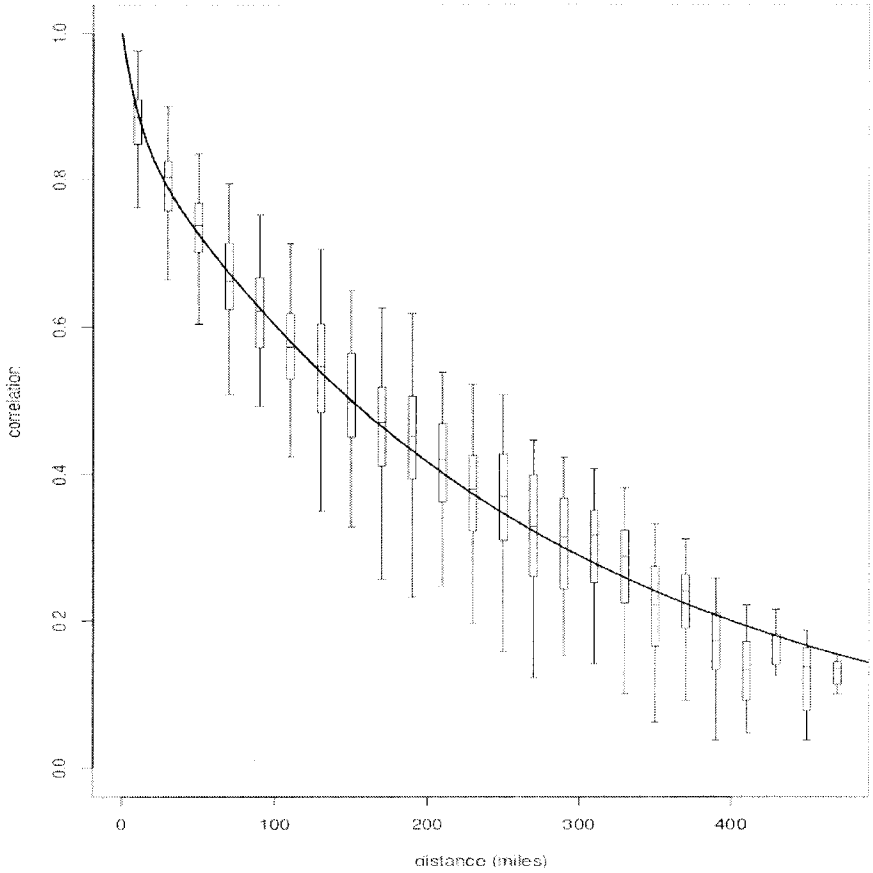


Figure 4.10: Empirical correlogram for AR(1) shocks.

a mixture of exponentials correlation function.

$$\psi(d(\mathbf{x},\mathbf{x}')) = \alpha \exp(-d(\mathbf{x},\mathbf{x}')/\theta_1) + (1-\alpha)\exp(-d(\mathbf{x},\mathbf{x}')/\theta_2) \qquad (4.9)$$

where $d(\mathbf{x},\mathbf{x}')$ is the distance between two locations $\mathbf{x}$ and $\mathbf{x}'$, θ_1 accounts for short range correlation and θ_2 for long range correlation.

Correlation model (4.9) allows the spatial field to be interpreted as the sum of two independent spatial processes with possibly different correlation scales without changing the smoothness of ψ at zero, but the shape will be modified for short distances. The reader should note that unlike a geostatistical analysis for a single field, the correlations associated with the shocks are statistics based

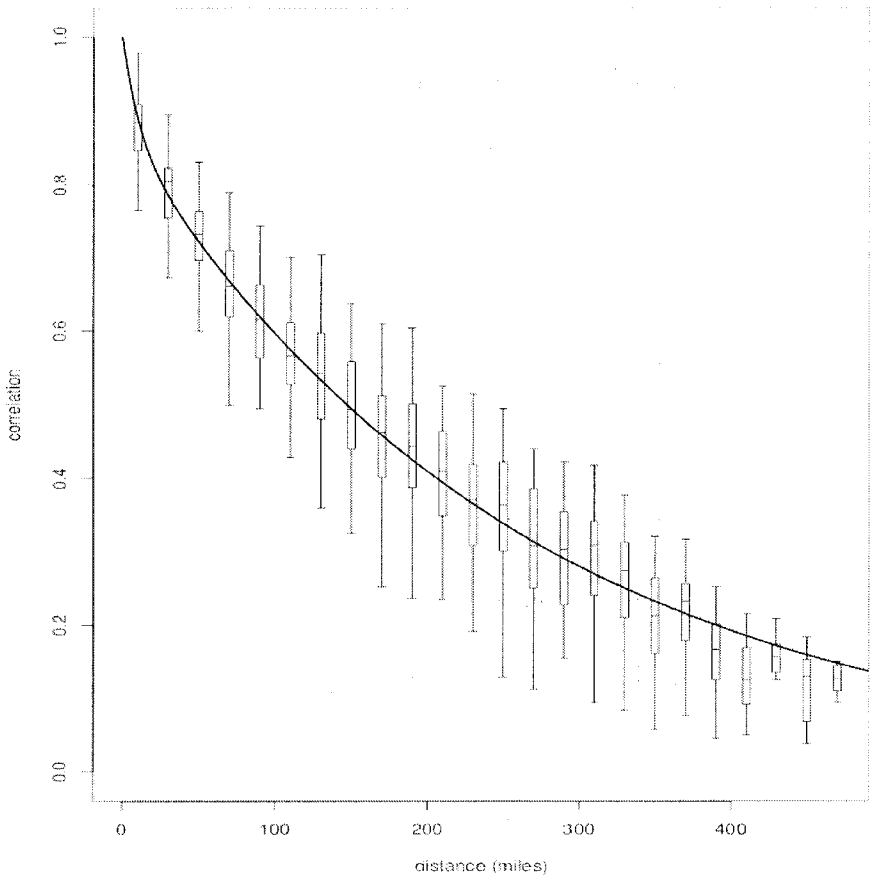


Figure 4.11: Empirical correlogram plot for AR(2) shocks.

on a large ($n > 500$) sample size, which enables enough accuracy to facilitate modeling detailed features such as the mixture component.

The fitted parameters for Equation (4.9) for the AR(1) shocks are (range parameter estimates converted to miles): $\hat{\alpha} \approx 0.13$ (0.02), $\hat{\theta}_1 \approx 11$ miles (3.37 miles) and $\hat{\theta}_2 \approx 272$ miles (16.89 miles) with parametric bootstrap standard errors in parentheses. Results from a parametric bootstrap show that estimates of the AR(1) coefficients are unbiased because bootstrap estimates of the AR coefficients subtracted from the autoregressive coefficients estimated from the data are centered around zero (Fig. 4.15), and the variability in the estimates

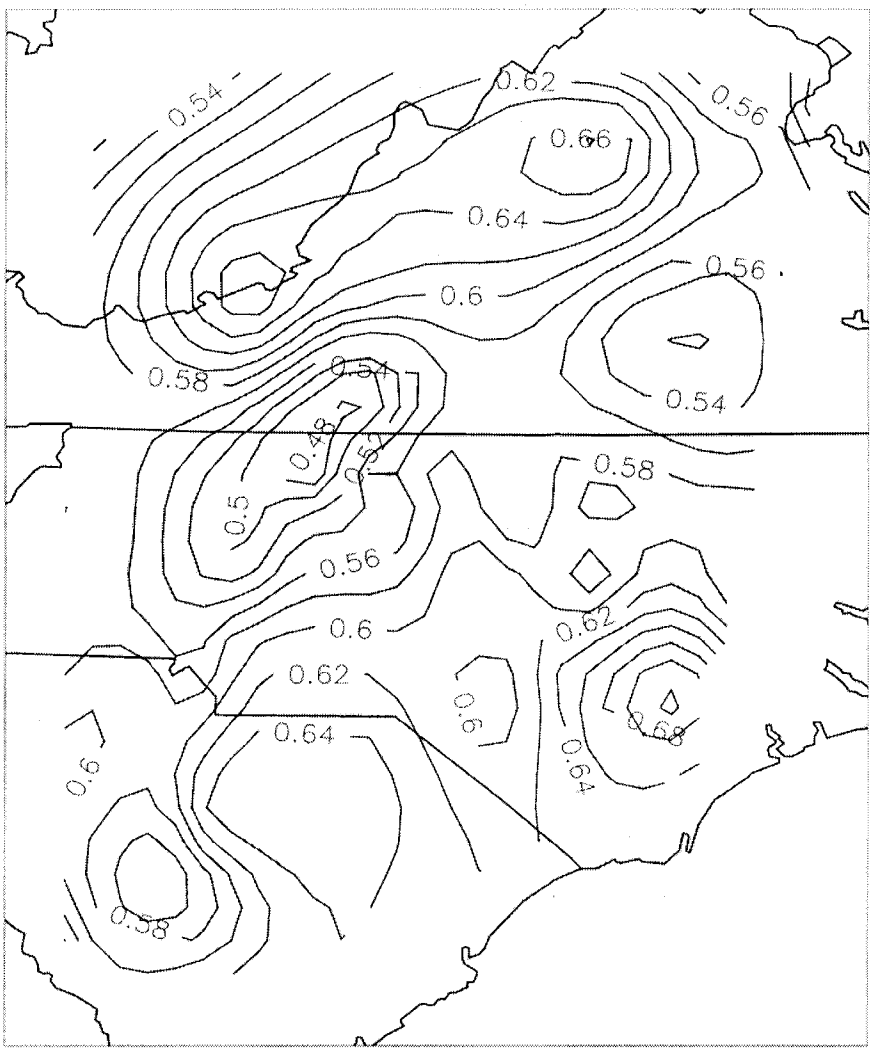


Figure 4.12: AR(1) coefficients estimated (and interpolated by thin plate spline) for each of the 72 locations in and around North Carolina.

is much larger than any potential bias.

The fitted model was used to generate conditional fields for the FHDA for each year and a 15×15 rectangular grid in the RTP subregion (Fig. 2.2). One thousand Monte Carlo realizations were used to approximate the distribution. Leave-one-out cross-validation results for each year are shown in Table 4.1,

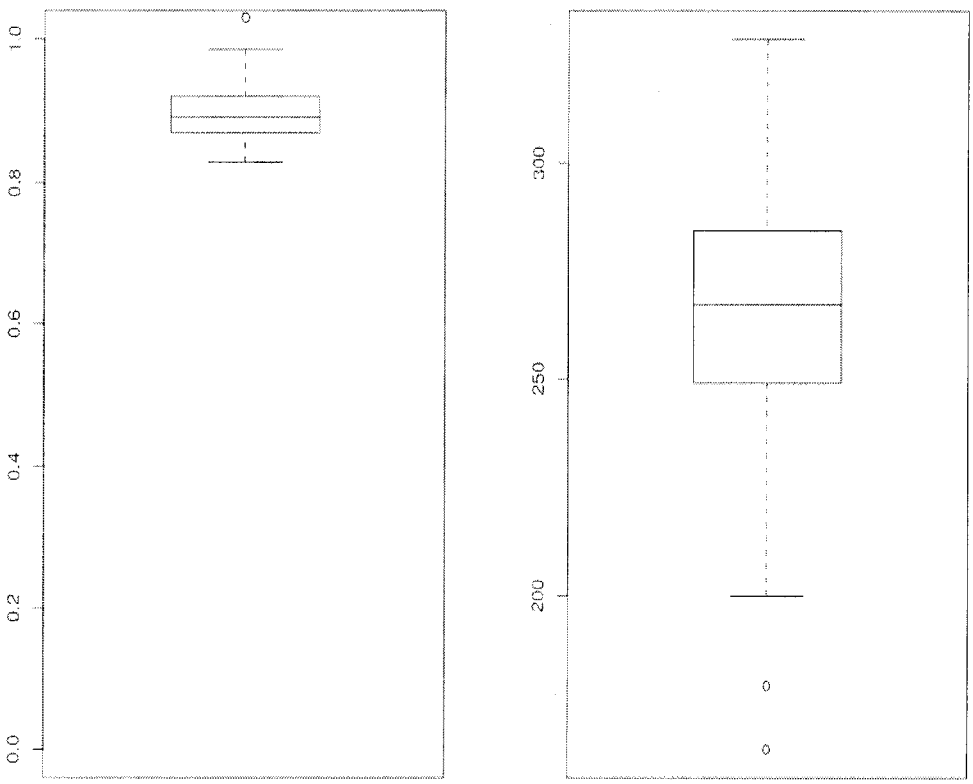


Figure 4.13: Boxplots of estimated nugget (left) and range (right) parameters for single exponential covariances fit at individual stations.

and on average the RMSE of the cross-validation residuals is about 4.46 ppb. Model prediction standard errors (MPSE) are summarized in Table 4.2, and on average these prediction errors are about 3 ppb. The model generated prediction errors underestimate the uncertainty in the predicted values of FHDA. Across the five seasons, the daily model tends to underestimate the standard error by about 34% (compared with cross-validation residuals). One explanation of this bias is that the daily model is not able to account for occasional large ozone values that appear in the data at a daily time scale. Not accounting for this non-Gaussian distribution in the shocks may not accurately capture the variability of the FHDA field.

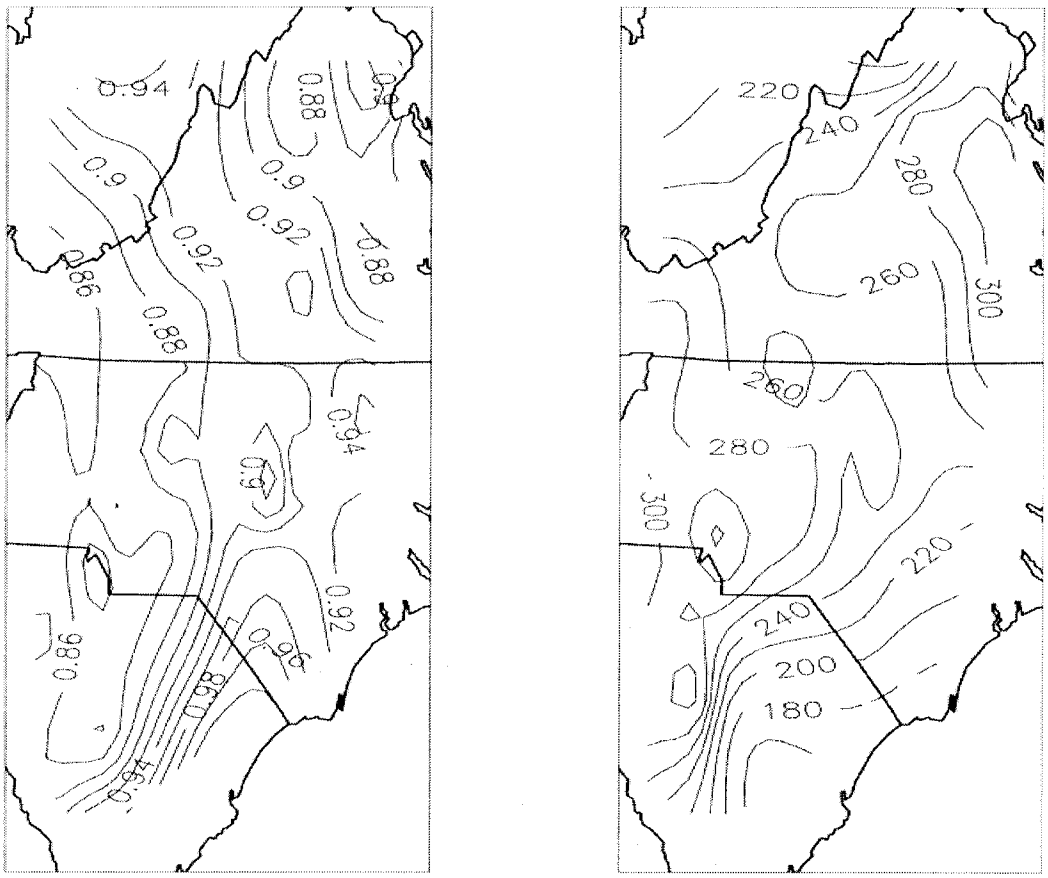


Figure 4.14: Estimated nugget (left) and range (right) parameters for single exponential covariances fit at individual stations. Here, these values are interpolated by way of a thin plate spline.

Table 4.1: Leave-one-out cross-validation RMSE (ppb) for predicting FHDA.

	Thin Plate Spline	Seasonal Model (ψ_v)	Seasonal Model (ψ_m)	Daily Model
1995	5.34	5.19	5.33	4.73
1996	5.61	5.51	5.68	4.84
1997	6.27	6.03	6.05	4.59
1998	5.00	4.98	4.93	3.25
1999	6.25	6.47	6.30	4.91

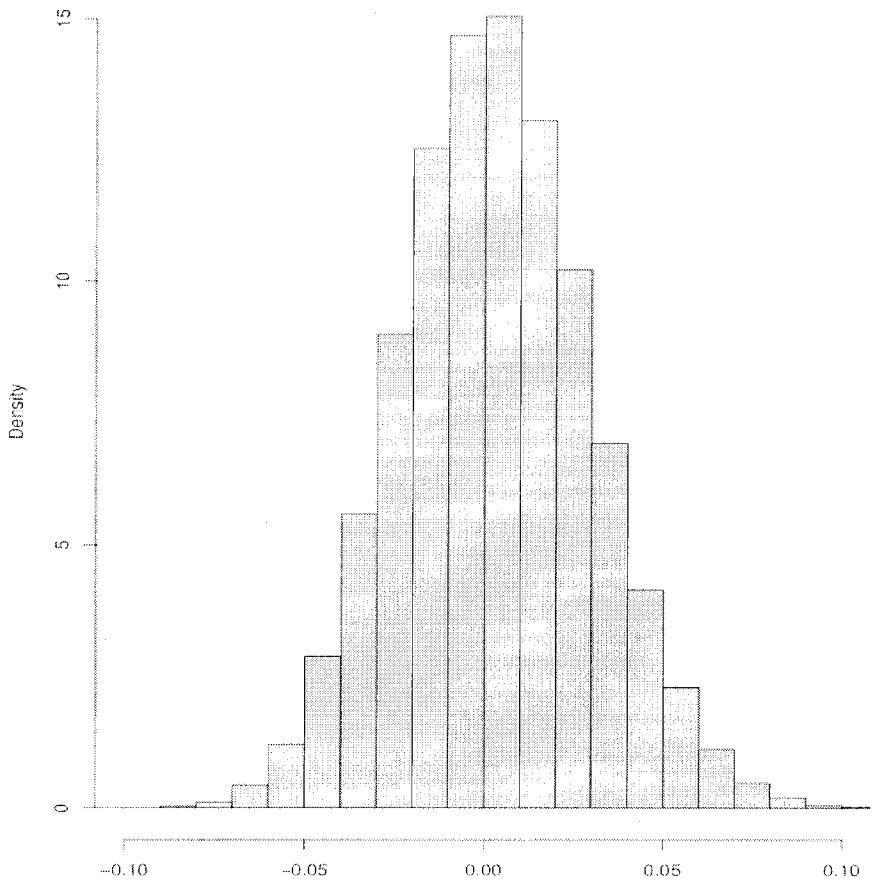


Figure 4.15: AR(1) coefficients estimated from the data ($\hat{\rho}$) minus parametric bootstrap AR(1) estimates ($\tilde{\rho}$).

Table 4.2: Averages of Model Prediction Standard Errors (MPSE) (ppb).

	Thin Plate Spline	Seasonal Model (ψ_v)	Seasonal Model (ψ_m)	Daily Model
1995	2.23	5.68	5.27	2.67
1996	2.49	5.96	5.90	2.85
1997	2.91	6.41	6.02	3.01
1998	2.75	5.35	4.85	2.93
1999	4.34	6.76	6.22	2.94

4.4.2 Seasonal Model Results

The seasonal approach applies a spatial model directly to the FHDA values, so a key step is to estimate a covariance function for this field. Empirical variograms for each of the 5 seasons indicate that almost all of the spatial dependence in the FHDA field appears to be limited to a very short range of less than 100 miles. A mixture of exponentials variogram

$$\gamma(d(\mathbf{x},\mathbf{x}')) = \sigma^2(1-\alpha\exp(-d(\mathbf{x},\mathbf{x}')/\theta_1) - (1-\alpha)\exp(-d(\mathbf{x},\mathbf{x}')/\theta_2))$$

was fit using all five years of data with parameter estimates: $\hat{\sigma} \approx 7.37$ ppb, $\hat{\alpha} \approx 0.38$, $\hat{\theta}_1 \approx 0.62$ miles and $\hat{\theta}_2 \approx 48.61$ miles and subsequently converted to a covariance function, ψ_v. Average MPSEs on the RTP grid are summarized in Table 4.2. On average, these prediction errors are about 6 ppb, slightly greater than that of the daily model approach.

For comparison to estimating the covariance from the FHDA variogram, a covariance function was estimated from unconditional simulations of the daily model. Based on a Monte Carlo sample of 600 FHDA simulated fields, a mixture of exponentials (4.9) was fit to the empirical correlations, call it ψ_m. The estimated parameters are $\hat{\alpha} \approx 0.51$, $\hat{\theta}_1 \approx 8.66$ and $\hat{\theta}_2 \approx 128.76$. The spatial prediction errors using this covariance are summarized in Table 4.2 and are, on average, 5.6 ppb, which is comparable to the seasonal model prediction errors using ψ_v. The seasonal analysis is done using the fields package [31] in R.

The thin plate spline model was fit (using `Tps` from the R package `fields` (Nychka *et al.* [31])) with a linear drift and the smoothing parameter chosen by generalized cross-validation.

MPSE (Table 4.2) are, on average, about 3 ppb, and are generally close to those of the daily model MPSEs.

4.4.3 Model Comparison

Daily model MPSE (Table 4.2) are generally smaller than the seasonal models; particularly away from station locations. The spline method tends to have similar prediction standard errors as the daily model, but there is less prediction precision away from the monitoring network than the seasonal model.

Model-based standard errors can either be reliable or misleading depending on the adequacy of the spatial model. It is also of interest to use cross-validation (CV) to evaluate the average prediction error of these methods. The standard leave-one-out procedure was applied to each monitoring location and method, and Table 4.1 reports for each year the CV RMSE for the differences between the predicted FHDA and the actual station values. The seasonal model CV RMSE for either choice of covariance and thin plate spline are very similar for each year. The daily model CV RMSE is consistently lower than the other models, but only slightly.

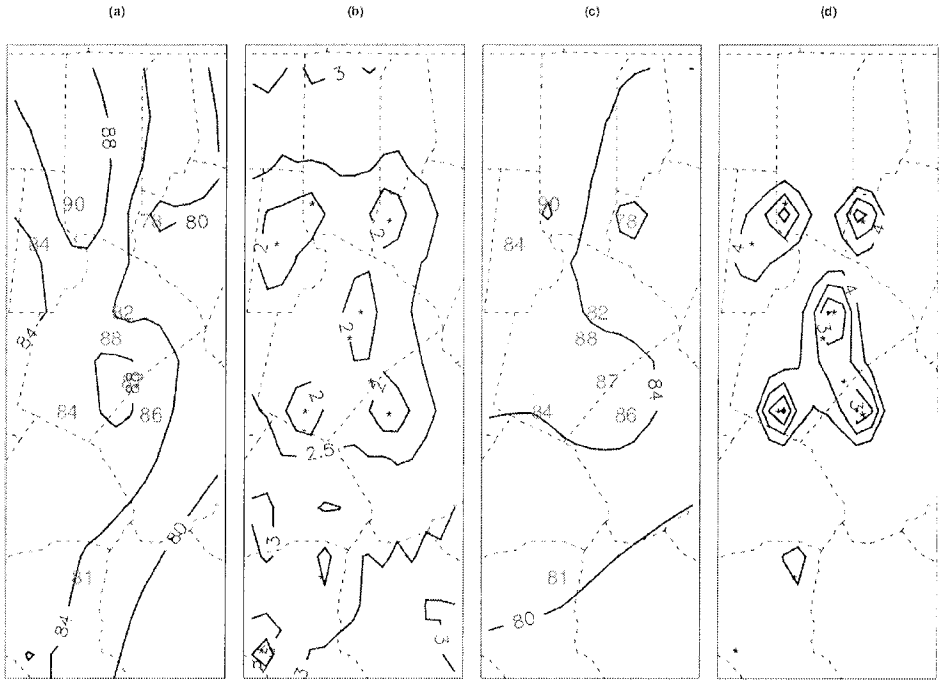


Figure 4.16: Predicted FHDA from (a) daily model and (c) seasonal model (faded numbers are observed FHDA), and model prediction standard errors (MPSE) for (b) daily model and (d) seasonal model for 1995.

Both models predict FHDA similarly (Figs. 4.16 through 4.20); the daily model having a tendency to predict a bit higher away from station locations than the seasonal model. The MPSE for the daily model is generally lower away from station locations, but of course, because the daily model is not able to account for occasional large ozone values that appear in the data based on a Gaussian assumption at a daily time scale. Therefore, the consistently low MPSEs for the daily model are not believable.

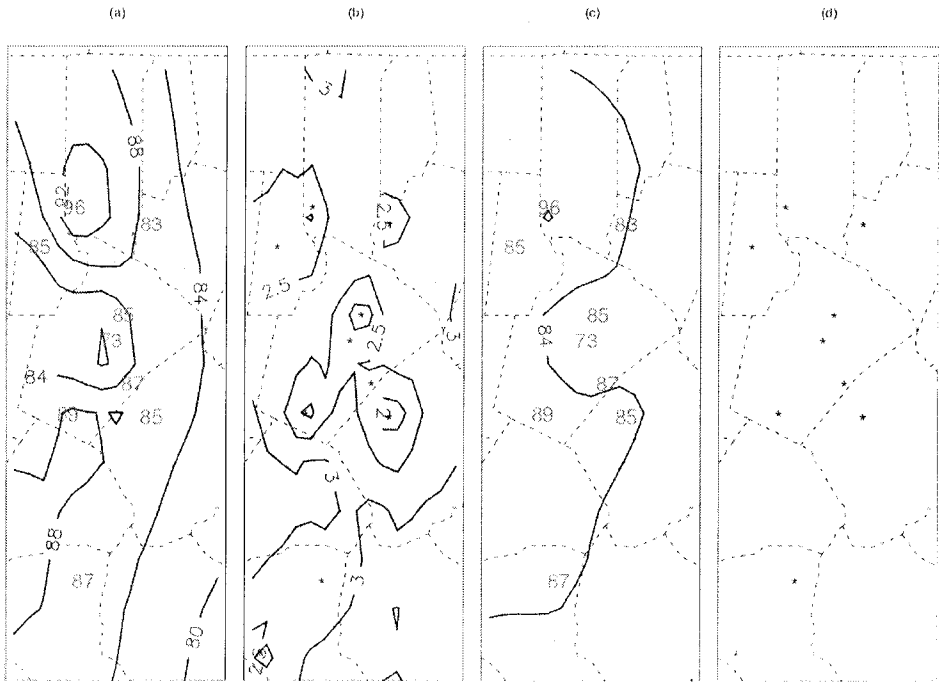


Figure 4.17: Predicted FHDA from (a) daily model and (c) seasonal model (faded numbers are observed FHDA), and model prediction standard errors (MPSE) for (b) daily model and (d) seasonal model for 1996.

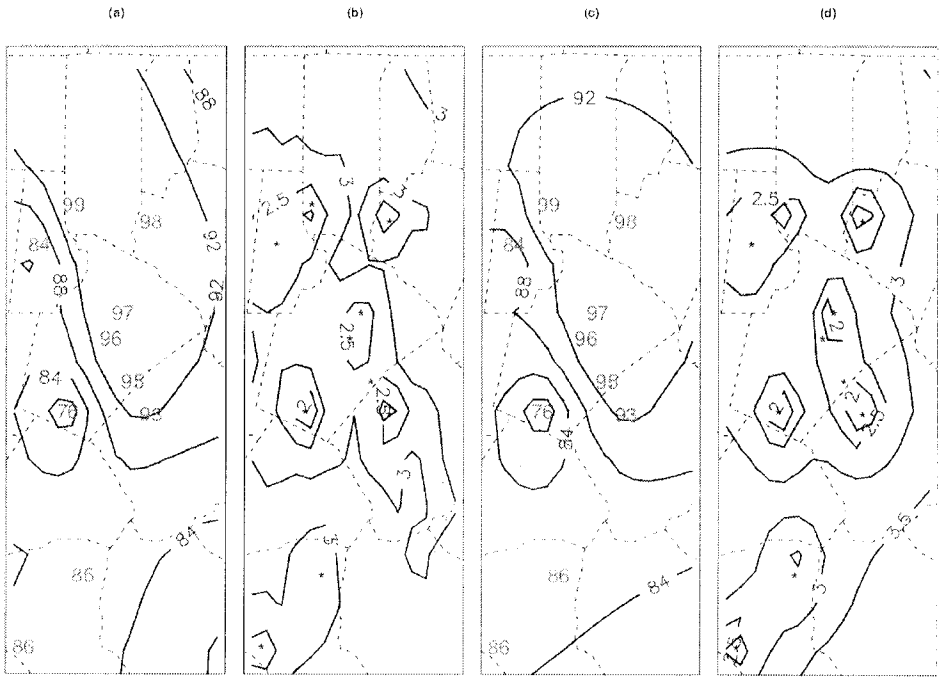


Figure 4.18: Predicted FHDA from (a) daily model and (c) seasonal model (faded numbers are observed FHDA), and model prediction standard errors (MPSE) for (b) daily model and (d) seasonal model for 1997.

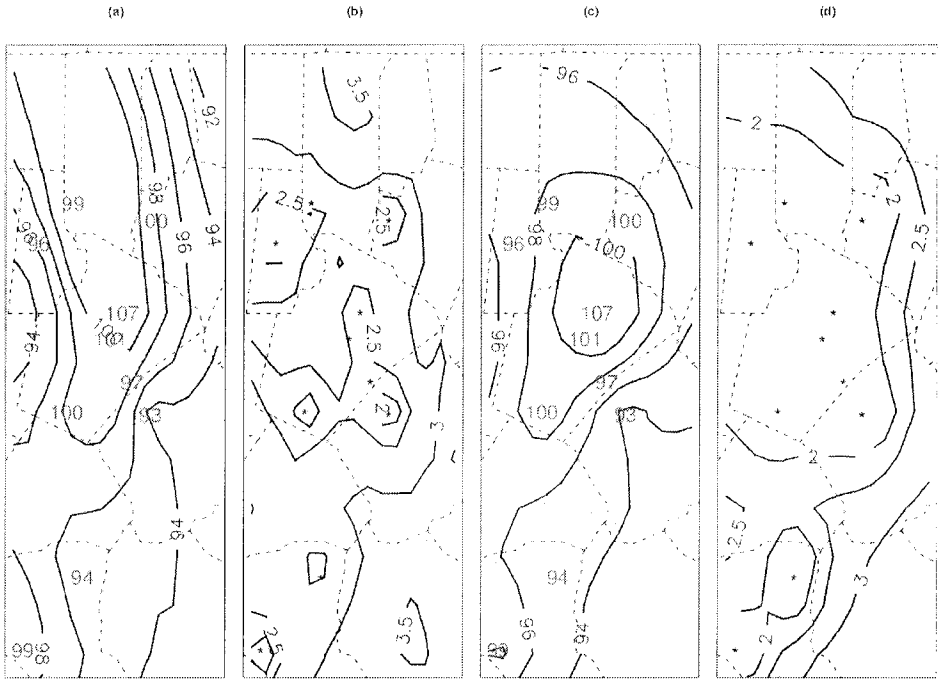


Figure 4.19: Predicted FHDA from (a) daily model and (c) seasonal model (faded numbers are observed FHDA), and model prediction standard errors (MPSE) for (b) daily model and (d) seasonal model for 1998.

Although care is needed in generalizing results from a specific data set to other cases, this work has shown a preference to analyze the FHDA standard using a daily model for ozone and then aggregating over the season to infer the FHDA field. The results for the North Carolina study region show that the seasonal model is reasonable, but the daily model is generally more accurate, based on the CV measures of RMSE in addition to having lower model standard errors of prediction.

Conceptually, the daily model has advantages in using fairly simple sta-

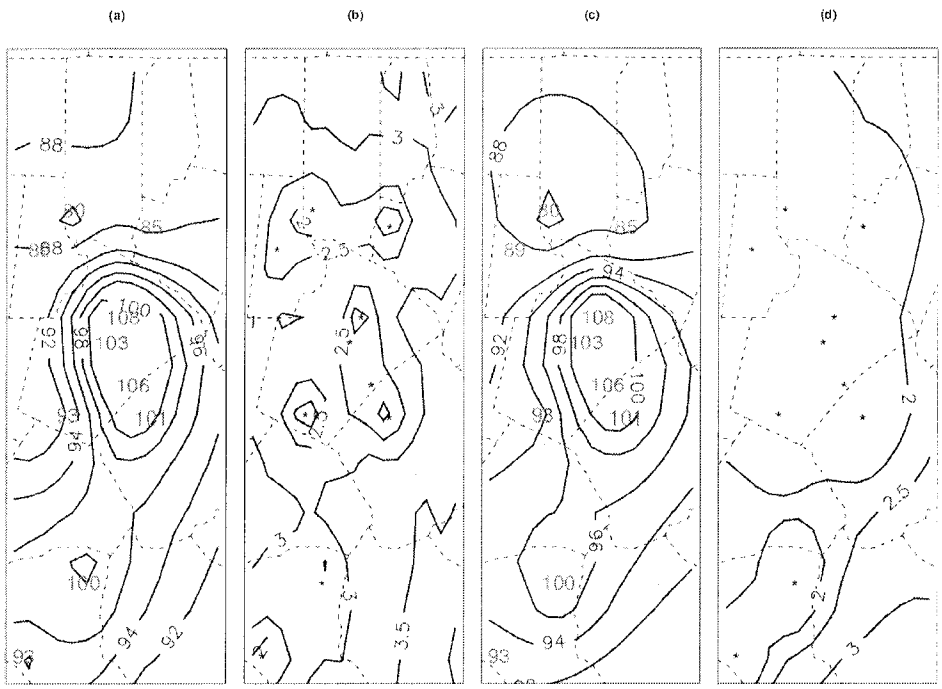


Figure 4.20: Predicted FHDA from (a) daily model and (c) seasonal model (faded numbers are observed FHDA), and model prediction standard errors (MPSE) for (b) daily model and (d) seasonal model for 1999.

tistical components on a daily scale that can produce relatively complicated seasonal statistics. For example, as long as the AR(1) shocks are stationary over space, the entire daily model can be fit using standard geostatistical and regression methods even if the original field (in this case standardized maximum 8-hour ozone levels) is nonstationary. Part of the success of the daily model may be because much of the spatial correlation and the nonstationarity of the raw measurements can be accounted for by standardizing the process and building in a temporal evolution. While the seasonal model is much simpler and easier to employ in general, it can actually be more complicated if the FHDA field is not stationary.

The lack of long-range correlation structure in the FHDA field simulated

by the daily model approach (conditional on the data) and reaffirmed by empirical variograms of the observed FHDA field suggest that standard spatial techniques may not be very effective at predicting the FHDA at locations relatively far from any monitoring station. Fig. 4.21 contrasts the different correlation scales among different transformations of the ozone field. Note the marked difference between daily fields and the seasonal FHDA. This is further justified by the greater standard errors of prediction found by both the seasonal model and the thin plate spline at locations away from the monitoring network. Additionally, the apparent correlation structure in the FHDA field, found from using the daily model approach without conditioning on the data, may be an artifact of the model. As hypothesized in section 4.4.1, one source of model bias may be the lack of a more heavy tail distribution associated with the AR shocks.

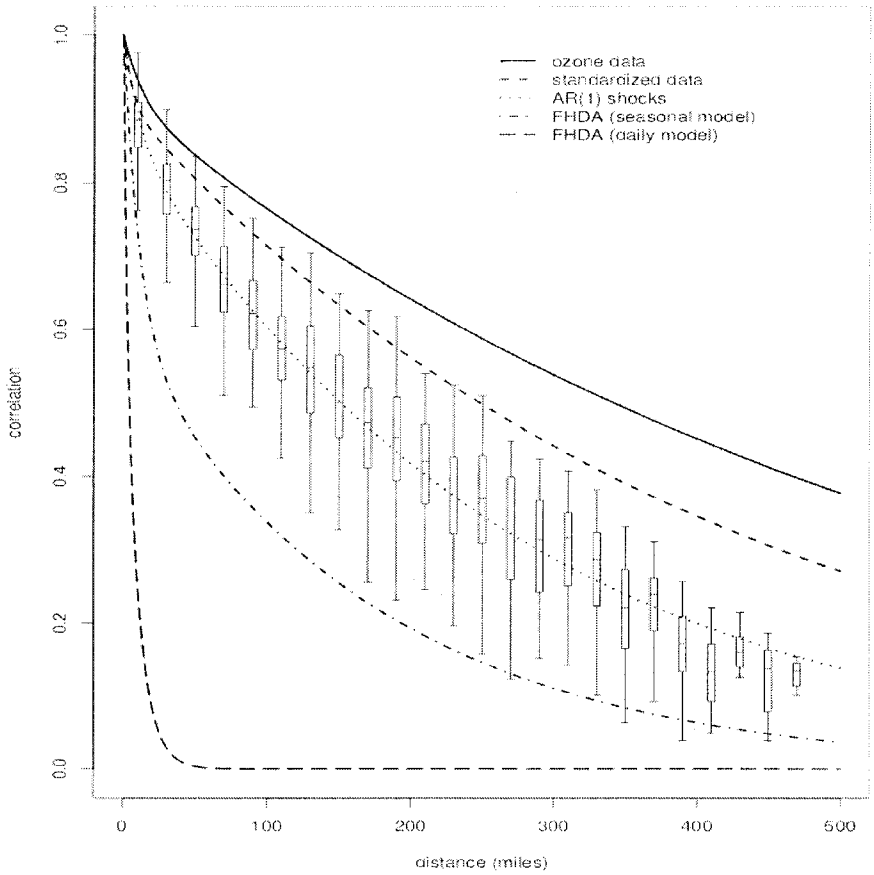


Figure 4.21: Fitted empirical correlation functions for original daily maximum 8-hour average ozone measurements, the standardized daily values, the spatial AR(1) shocks and unconditional (seasonal model) and conditional (daily model) simulations of the FHDA field.

Chapter 5

Modeling the Air Quality Standard Using Extreme Value Theory

5.1 Spatial models for extremes

The analyses from Chapter 4 arrive at estimates of extreme properties of the ozone monitoring data by a space-time model based on daily ozone measurements. Most of the fitting is focused on mean and variance properties and the distribution of the tails is partly constrained by the multivariate assumptions made for the daily model. An alternative approach is to model the tail behavior of the station measurements directly. This can be achieved by fitting an exceedance over threshold model (section 3.2.2), such as a GPD, for each

station. However, the number of observations is small for any one station, and one would expect significant uncertainty in the estimates because of too few observations (exceeding the threshold). Also, without additional spatial structure, the individual station models provide no obvious way to extrapolate to locations where ozone is not measured.

One strategy to improve the accuracy and provide for spatial prediction is to include a spatial component that links the distribution for different stations. In this section, a hierarchical component is added that treats the parameters of the GPD as a smooth surface. This device is not only reasonable given the spatial dependence of the surface ozone, but also combines strength across the stations to give a more stable estimate of the tail parameters. It should be noted that in working through this example, there are several places where simplifying assumptions in the model have been made that may have dubious justification. Many of these could be avoided at the cost of more complex models, but some amount of simplicity is desired for comparison with the methods presented in chapter 4. Here, the concern is in modeling the distribution of $Z(\mathbf{x})$, and not the joint distribution of $Z(\mathbf{x})$ and $Z(\mathbf{x}'$–the spatial component is on the parameters of the GPD distribution.

5.1.1 Elements of a hierarchical model

Let $\mathbf{x}_k$ $k = 1, \ldots, n$ represent the locations for the ozone stations, and $y(\mathbf{x}, t)$ denotes the daily ozone at an arbitrary location $\mathbf{x}$ and day t. The goal is to

estimate the surfaces $\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$ that describe how the GPD parameters change as a function of location. Based on these surfaces, the probability of an extreme ozone event can be evaluated at any point in the region.

In terms of a hierarchical model, we assume that; conditional on the values of $\sigma(\mathbf{x})$, $\xi(\mathbf{x})$ and threshold u; the exceedances of ozone at location $\mathbf{x}$ follow a GPD. Denote the probability density function (pdf) of this conditional distribution as

$$[y(\mathbf{x},t)|\sigma(\mathbf{x}),\xi(\mathbf{x}),u]. \tag{5.1}$$

The next level in the hierarchy is a statistical model for $\sigma(\mathbf{x})$, $\xi(\mathbf{x})$ and u. Denote the pdf for these components as

$$[\sigma(\mathbf{x}),\xi(\mathbf{x}),u|\boldsymbol{\theta}] \tag{5.2}$$

In general, $\boldsymbol{\theta}$ is a vector of hyperparameters controlling the distributions for $\sigma(\mathbf{x})$, $\xi(\mathbf{x})$ and u, and the final stage in the hierarchy is a prior distribution on $\boldsymbol{\theta}$ (denoted $[\boldsymbol{\theta}]$).

Multiplying these pieces together gives the joint pdf

$$[y(\mathbf{x}_i,t)|\sigma(\mathbf{x}),\xi(\mathbf{x}),u] \quad [\sigma(\mathbf{x}),\xi(\mathbf{x}),u|\boldsymbol{\theta}] \quad [\boldsymbol{\theta}]. \tag{5.3}$$

Here t $(1 \leq t \leq T)$ indexes the T days and i $(1 \leq i \leq M)$ indexes the M station locations.

For a formal Bayesian analysis, the specification of (5.3) is a complete recipe for inference on the parameters. Using Bayes Theorem, the posterior for $\sigma(\mathbf{x})$ and $\xi(\mathbf{x})$ given the data ($[\sigma(\mathbf{x}),\xi(\mathbf{x})|y(\mathbf{x},t),\theta]$) can, in principle, be computed.

In particular, a useful summary of the posterior distribution is the combination of parameters that has the highest probability given the observed data. This combination is known as the posterior mode. It is an elementary fact that the posterior mode can be found by maximizing the joint density in (5.3), and this equivalence is used to find estimates of the surfaces of the parameters. Although the basic outline of the Bayes analysis is clear, the details of the model are important. Most practical applications require a balance among the full richness of the hierarchical model,the limitations of the data and a lack of detailed prior knowledge concerning hyperparameters. This is also true of the analysis of the ozone data given in the next section.

5.1.2 Modeling assumptions for the ozone application

Under the assumption that the observations are conditionally independent over both time and space, the joint distribution of parameters and data is

$$\prod_{t=1,i=1}^{T,M} [y(\mathbf{x}_i,t)|\sigma(\mathbf{x}_i),\xi(\mathbf{x}_i),u]\ [\sigma(\mathbf{x}_i),\xi(\mathbf{x}_i),u|\theta]\ [\theta]. \tag{5.4}$$

The assumption of conditional independence is a strong one, but can be justified because extreme values tend to be less correlated than more central parts of a distribution. In particular, the results from simulating bivariate distributions of the fourth-highest order statistic (section 4.4.3) suggest that the spatial correlation of the FHDA is much weaker than the correlation among daily ozone measurements.

Finally, in order to give the specific form for the model in (5.3), it is helpful to make several additional assumptions. Assume that u is specified, $\xi(\mathbf{x}) \equiv \xi$ is a constant, and $\sigma(\mathbf{x})$ is assumed to be a Gaussian random field with the form

$$\sigma(\mathbf{x}) = P(\mathbf{x}) + e(\mathbf{x}); \tag{5.5}$$

where P is a fixed linear function, and $e(\mathbf{x})$ is a mean zero spatial process related to a Matérn covariance (Stein (1999)). P is known as the spatial drift; and as a linear function, has three parameters that will be denoted by the vector $\boldsymbol{\beta}$. Creating a matrix with the constant and linear terms for the observed locations, $\boldsymbol{X}$, the spatial drift contribution to the scale parameter at the stations is the vector $\boldsymbol{X\beta}$.

Recall from section 3.1.1 that the Matérn family of covariance functions (3.9) has three parameters: σ, ν and ρ. The full set of parameters would be difficult to identify with the ozone data set, however, because there is little prior knowledge of their values, and the data set is small. Given these constraints, this analysis restricts ν to 2, and estimates the combination of the scale and range parameters that describes how spatial correlations vary for small distances. This function is referred to as the principle irregular term (Stein [42] page 32), and the coefficient for this term is a combination of σ^2 and ρ. Here, denote this coefficient by λ. Note that λ is also the smoothing parameter commonly used in penalized likelihood problems and for splines. This approximation matches the spatial process model associated with

a second order thin plate spline, and there is both heuristic and theoretical support that the approximation provided with just this single parameter is adequate.

The last component of the model is the specification of prior distributions for ξ and the hyperparameters $\boldsymbol{\beta}$, the spatial drift, and λ (i.e., $\boldsymbol{\theta} = (\xi, \beta, \lambda)$ in the previous section). A prior for these hyperparameters would lead to another level of the hierarchy that is only indirectly related to the observed data. Again, some simplifying assumptions based on practicality and the limitations of the data are useful. Specifically, take an empirical Bayes approach by not specifying priors; or, equivalently, assuming them to be improper and constant. With this simplification, finding the posterior mode can be interpreted as applying maximum likelihood to determine these parameters.

With all the assumptions included, the logarithm of the likelihood derived from the joint distribution from (5.4) is given by

$$\sum_{i=1}^{M} \ell_{GPD}(\boldsymbol{Y}_i, \sigma(\mathbf{x}_i), \xi) - \lambda(\boldsymbol{\sigma} - \boldsymbol{X\beta})^T (K^{-1})(\boldsymbol{\sigma} - \boldsymbol{X\beta})/2 - \log(|\lambda K|) + C. \quad (5.6)$$

Note that conspicuously absent are priors for ξ, $\boldsymbol{\beta}$ and λ. Here, the log-likelihood, ℓ_{GPD}, is exactly the GPD log-likelihood with a fixed threshold of $u = 60$ ppb described in section 3.2.2. $\boldsymbol{Y}_i$ is the vector of ozone measurements for the i^{th} station, $\boldsymbol{\sigma}$ is the vector of scale parameters with i^{th} element $\sigma(\boldsymbol{x}_i)$, K is the covariance for the scale parameters among the station locations and C is a constant independent of the parameters. Because the data is conditioned on $\boldsymbol{\sigma}$, it is sufficient to find the maximum over this vector of parameters. The

posterior mode for $\sigma(\mathbf{x})$ at an arbitrary location can be approximated as the conditional expectation of $\sigma(\mathbf{x})$ given $\boldsymbol{\sigma}$ at the observed locations, and based on the Matérn covariance for this surface. This estimate is not exact because this simplification fixes the parameters of the covariance at their mode values. For multivariate normal distributions, this conditional expectation is also the well known kriging estimate from geostatistics, and it is common practice to condition on the covariance parameters when forming a spatial estimate.

5.1.3 Posterior modes for the GPD

As an initial analysis and benchmark, the individual MLEs for each station were found under the constraint that the shape is constant, but the scale parameter can vary. For this case, $\hat{\xi} = -0.343$ and the posterior mode surface for the scale is plotted in Fig. 5.1 (a). The surface for the scale can be recovered using a spatial statistics estimate that extrapolates from the estimates of σ at the observed locations. Here, the spatial model assumed for σ is equivalent to a thin plate spline, so the estimate of the surface simply involves an interpolation using standard spline algorithms. These results can be interpreted as the limiting case when the parameter λ in (5.6) becomes very small. Although the surface in Figure 5.1 (a) is a useful visual benchmark, it is not believable because the surface interpolation assumes that each station's GPD parameters are known without error. In fact, it is known that there is substantial uncertainty in these individual estimates because of the small number

of exceedances measured for each station.

The spatial analysis based on maximizing (5.6) combines information about the GPD scale parameters across stations. Specifically, the combination depends on the value of the smoothing parameter, λ. Because the mode is sensitive to the value of λ, examination of the estimates for some fixed choices of this parameter is prudent; and to be precise, let $\hat{\sigma}_\lambda(\mathbf{x})$ and $\hat{\xi}_\lambda$ be the parameter values that maximize (5.6) for a fixed value of λ. The plots in Fig. 5.1 (b)-(d) show the estimates of the $\hat{\sigma}_\lambda(\mathbf{x})$ surface for different values of λ–$\hat{\xi}_\lambda$ does not vary significantly as a function of lambda. The sequence of surfaces illustrate why this parameter controls the smoothing. As λ increases the surface tends to be smoother; with fewer sharp features and less resolution. Because of the linear spatial drift in the model, as λ increases, the surface will simplify to a linear function; a plane.

Fig. 5.2 is a plot of the profile likelihood of $\log \lambda$, and might be used to draw inferences about values for this parameter. The increasing profile indicates that the posterior is maximized for very large values of λ; and, in the limit, describes a surface for σ that is simply a plane. The fitted plane in this case has a small gradient and little variability over the data region.

Surprisingly, this result suggests that there is little evidence for spatial

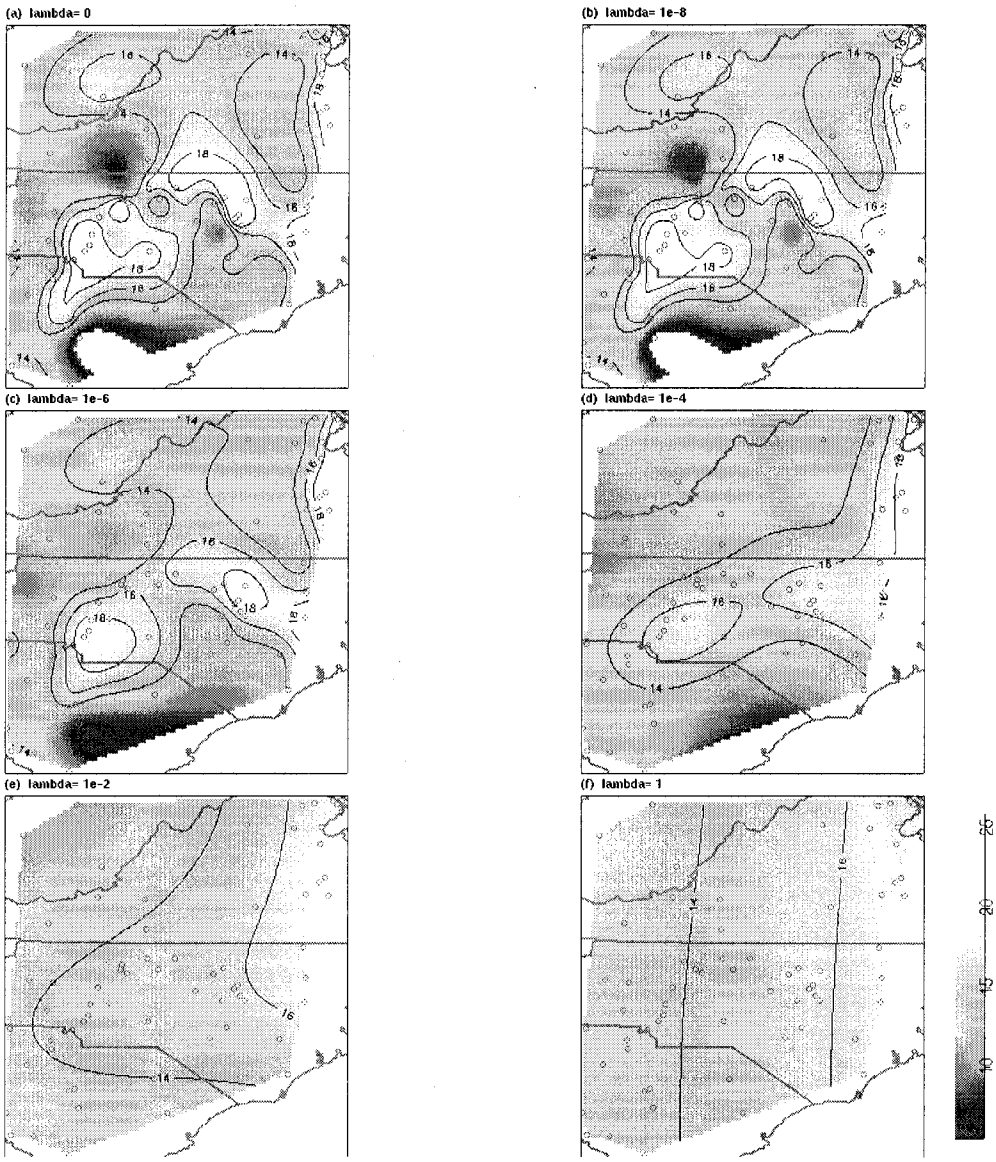


Figure 5.1: Estimated surfaces of the GPD scale parameter, $\hat{\sigma}(\mathbf{x})$, for different values of the smoothing parameter, λ.

structure in the scale parameter surface. These results can be contrasted with an *ad hoc* approach of smoothing the GPD MLEs directly. A simpler approach, though lacking a rigorous statistical model, is to smooth the individual estimates of the scale parameters at the locations using a thin plate spline. Generalized cross-validation, a frequentist criterion for estimating λ,

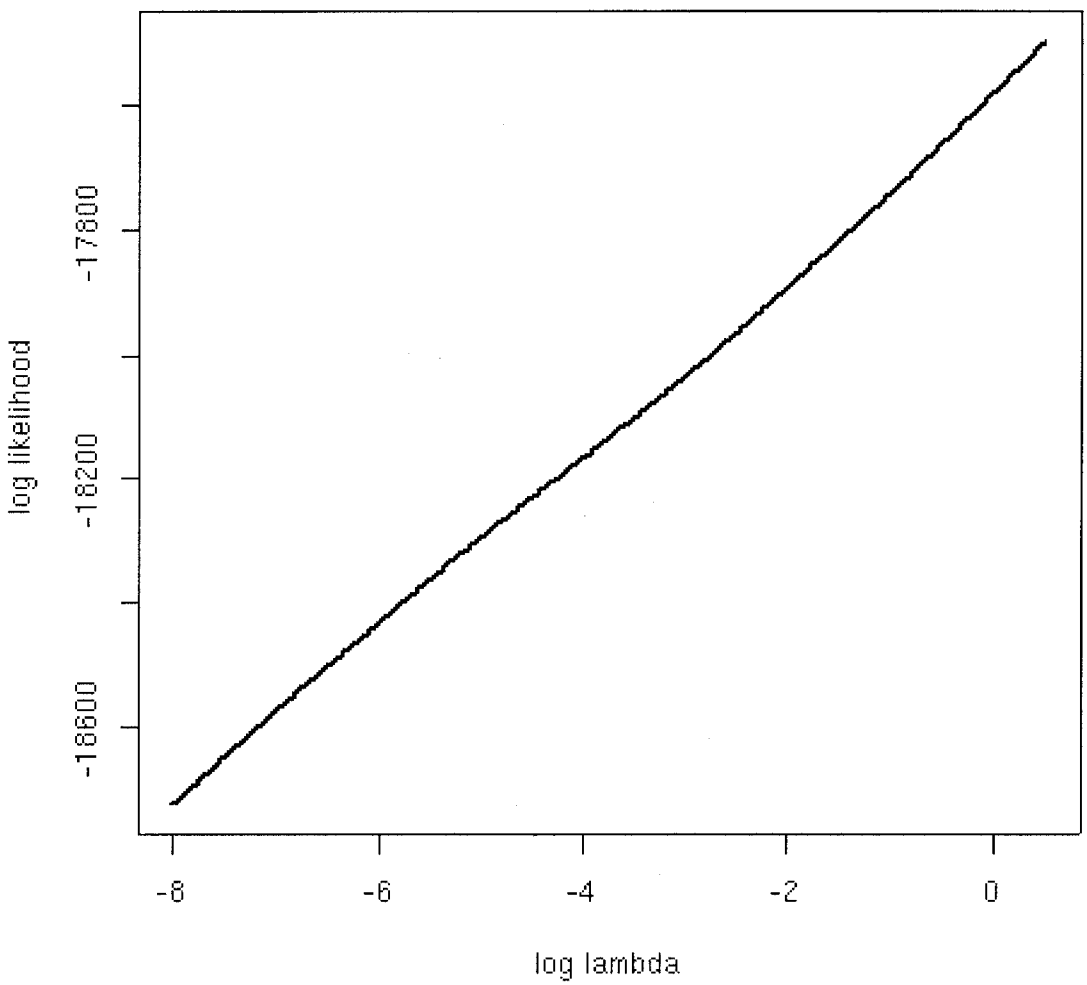


Figure 5.2: Profile likelihood for λ, the smoothing parameter for the surface of scale parameters.

is used to select a value for smoothing that gives a surface similar to Fig. 5.1 (d). Note that the surface in (d) exhibits higher levels along the urban corridor through North Carolina and Virginia, and levels are lower in more rural areas such as Western Virginia. Although this interpretation is reasonable, it is unclear how to reconcile this with the profile likelihood that suggest that σ

has little spatial dependence.

Assuming that the intermediate value of λ depicted in Figure 5.1 (d) gives a useful summary of the monitoring data, consider a more interpretable functional of the GPD distribution. Recall that for meeting proposed EPA air quality standards it is important that the FHDA fall below 84 ppb. Figure 5.3 (b) is the probability of the FHDA exceeding 84 ppb estimated from the spatial GPD model. Here, the probability of a location exceeding the threshold of 60 ppb is estimated from a thin plate spline fit to the empirical probabilities from each station. This quantity is ζ_u from Coles [3] section 4.3.2. The surface of probabilities, $\zeta_u(\mathbf{x})$, is combined with the surface of scale and shape parameter estimates to estimate the probability that daily ozone exceeds 84 ppb. Under the assumption of independence between daily exceedances, the binomial distribution is used to calculate the probability that the FHDA exceeds 84 ppb (i.e., four or more events out of 184). The surfaces in Figs. 5.3 through 5.7 show surprisingly similar results between (a) and (b), where (a) is the probability of FHDA exceeding 84 ppb using the daily model of section 4.1.2, and (b) is the probability of daily ozone exceeding 84 ppb using model (5.4). In each case, the surfaces indicate a broad region across the Southeastern United States where there is high probability that the FHDA will exceed 84 ppb. Areas of lower probability tend to be in more rural areas or at the edges of this region. The 1998 and 1999 seasons show particularly high probabilities across the entire region, except for the southern coast; this may be explained

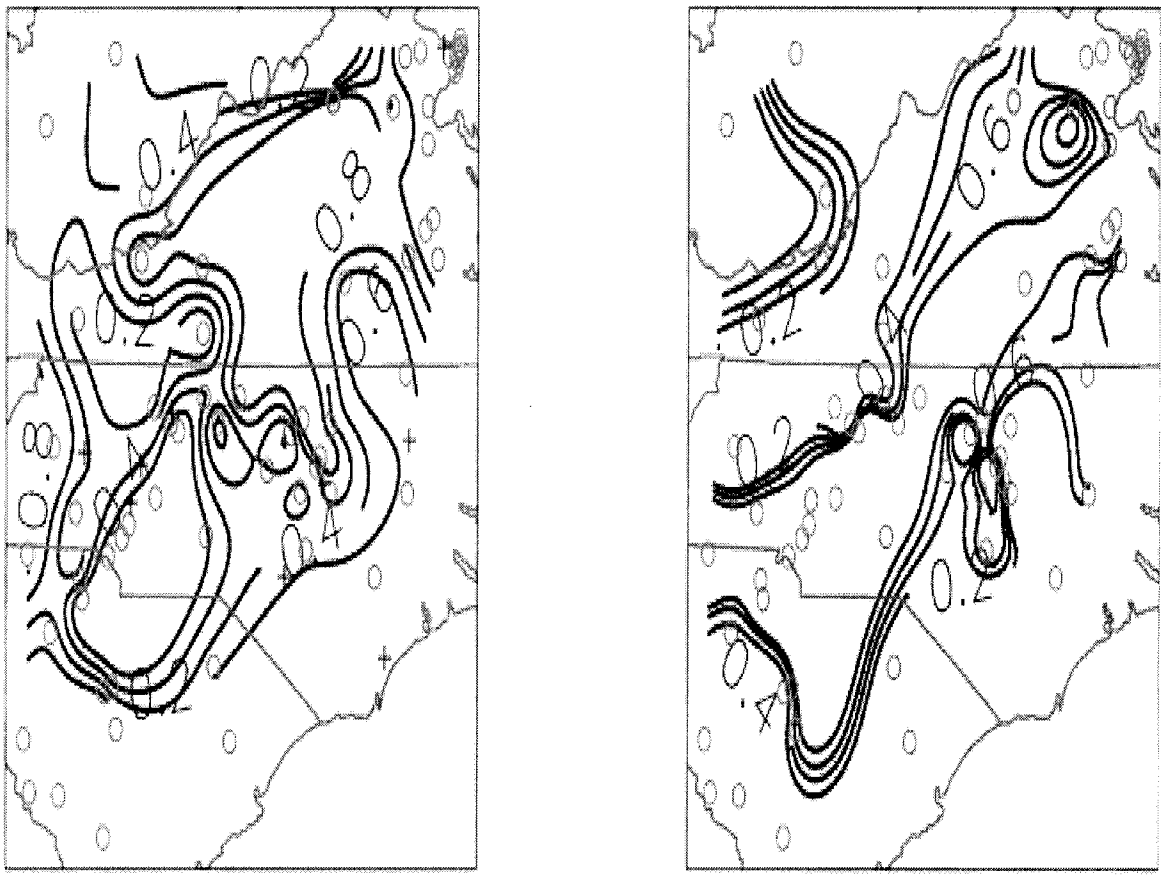


Figure 5.3: Probability of FHDA exceeding 84 ppb using (left) daily model of section 4.1.2 and (right) daily ozone using model (5.4) for 1995.

by higher average temperatures in these summers for this region.

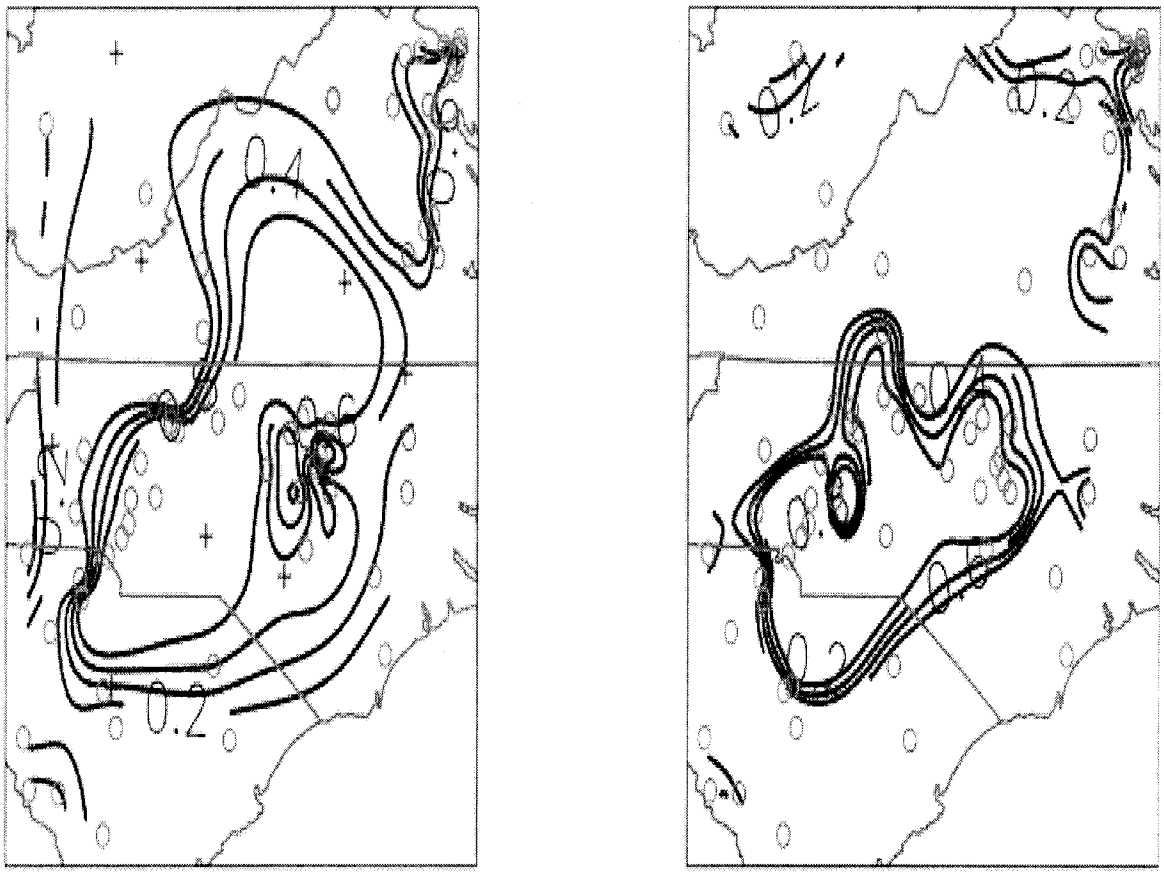


Figure 5.4: Probability of FHDA exceeding 84 ppb using (left) daily model of section 4.1.2 and (right) daily ozone using model (5.4) for 1996.

5.1.4 Extensions and discussion

There are several extensions to this work that could improve the model, or give more accurate estimates. With more data, one could consider a more flexible covariance model for σ, and add a spatial component for ξ. One way to accumulate more observations is to extend the analysis over multiple years; there are five years of data considered here for the daily model. One difficulty with multiple years is that ozone levels would need to be adjusted by covariates such as meteorology and time trends. Another extension is to include a link

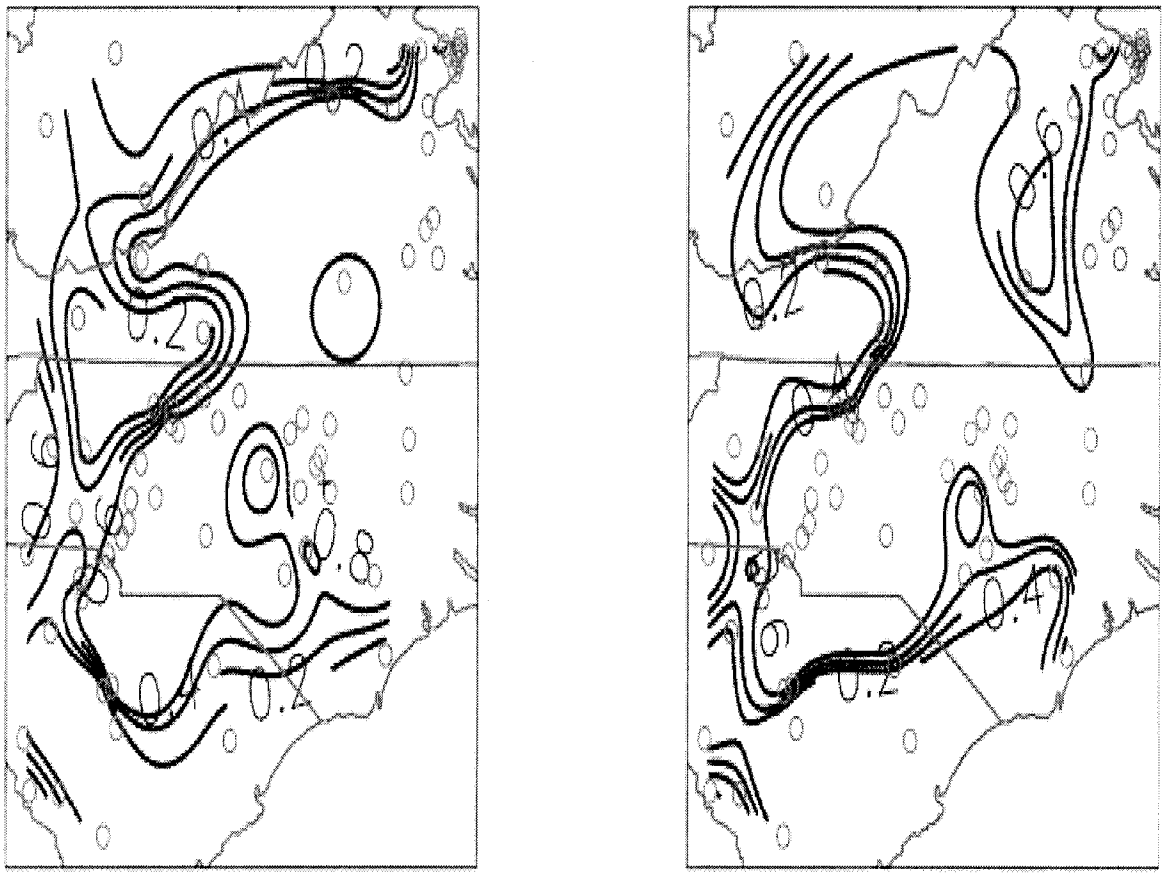


Figure 5.5: Probability of FHDA exceeding 84 ppb using (left) daily model of section 4.1.2 and (right) daily ozone using model (5.4) for 1997.

function for σ, such as the exponential, in order to preserve positivity of σ, and possibly to give a better approximation of a Gaussian field. Finally, by adding proper priors to this analysis, it may be possible to sample from the posterior to obtain a Monte Carlo approximation to the posterior distribution.

It should be noted that this method solves a much more general problem than simply inferring properties of the NAAQS for ground-level ozone. There are numerous datasets that could be analyzed with these methods for various purposes. For example, for areas where flooding is of particular concern, this model could be used to infer probabilities of extreme precipation events

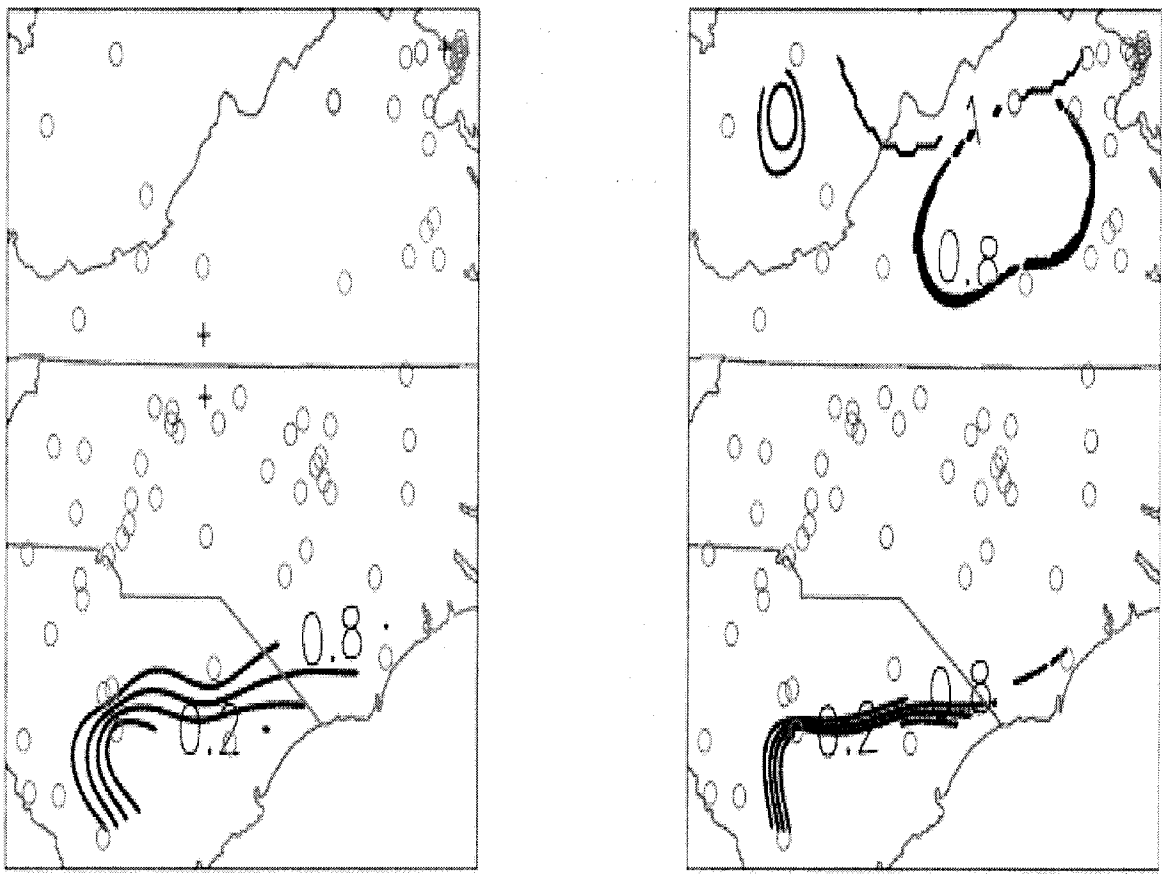


Figure 5.6: Probability of FHDA exceeding 84 ppb using (left) daily model of section 4.1.2 and (right) daily ozone using model (5.4) for 1998.

in a way that incorporates spatial information. Of course, there are many other problems where extreme events are of concern, and inclusion of spatial information could be important.

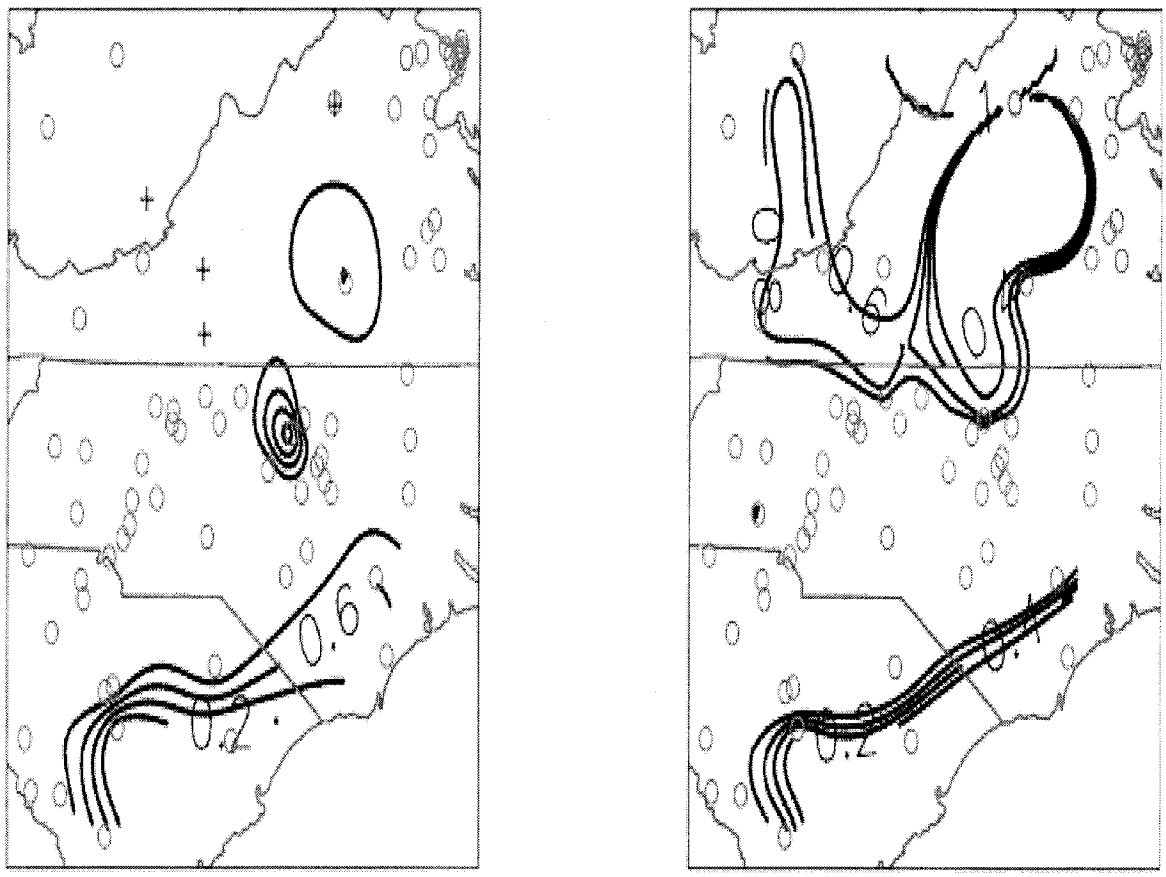


Figure 5.7: Probability of FHDA exceeding 84 ppb using (left) daily model of section 4.1.2 and (right) daily ozone using model (5.4) for 1999.

Chapter 6

Conclusions

This thesis has addressed the statistical analysis of the ozone seasonal standard based on a space-time model for daily daily data. This new model is given by

$$Y(\mathbf{x}, t) = \mu(\mathbf{x}, t) + \sigma(\mathbf{x})u(\mathbf{x}, t) \tag{6.1}$$

with $\mu(\mathbf{x}, t)$ and $\sigma(\mathbf{x})$ as in (4.1), and $u(\mathbf{x}, t)$ as in (4.2).

Section 6.1 discusses details discovered in using this approach. Section 6.2 discusses the more direct model (chapter 5) that uses distributions found to be appropriate for modeling extreme behavior, at least asymptotically. Finally, section 6.4 gives some ideas for future work.

6.1 Daily Model Discussion

Although care is needed in generalizing results from a specific data set to other cases, this work has shown a preference to analyze the FHDA standard using

a daily space-time model for ozone and then aggregating over the season to infer the FHDA field. The results for the North Carolina study region show that model (6.1) is competitive with the seasonal model in that, based on the CV measures of RMSE, it is slightly more accurate. The next section details the benefits of this approach.

Conceptually, the daily model has advantages in using fairly simple statistical components on a daily scale that can produce relatively complicated seasonal statistics. For example, as long as a standard correlation function is reasonable for $\psi(d(\mathbf{x}, \mathbf{x}'))$ from (4.4), the entire daily model can be fit using standard geostatistical and regression methods even if the observed FHDA field is nonstationary. I believe that part of the success of the daily model is that much of the spatial correlation and the nonstationarity of the raw measurements can be accounted for by standardizing the process and building in a temporal evolution. Additionally, this standardization probably accounts for the spatial trend better than the seasonal approach, and may explain the better performance of the daily model. While the seasonal model is much simpler and easier to employ in general, it can actually be more complicated if the FHDA field is not stationary.

The lack of long-range correlation structure in the FHDA field simulated by the daily model approach (conditional on the data) and reaffirmed by empirical variograms of the observed FHDA field suggest that standard spatial techniques may not be very effective at predicting the FHDA at locations rel-

atively far from any monitoring station. Correlation scales contrasted among different transformations of the ozone field (Fig. 4.21) displayed marked differences between daily fields and the seasonal FHDA. Specifically, at about 100 miles the daily model FHDA fitted correlogram function shows nearly zero correlation, whereas that of the seasonal model shows a correlation above 0.3. This is further justified by the greater MPSE found by both the seasonal model (Fig. 4.16 (d) through 4.20 (d)) and the thin plate spline (not shown) at locations away from the monitoring network. Additionally, the apparent correlation structure in the FHDA field, implied by the daily model may be biased because of the lack of a heavy tail distribution for the autoregressive shocks.

One interesting and perhaps disappointing, finding in this work is that the daily model prediction standard errors are generally too optimistic. I believe this is because the daily is not able to account for occassional large ozone values that appear in the data based on a Gaussian assumption at a daily time scale.

Although this problem suggests ample areas of new research, I also believe the daily model provides a substantial improvement in interpolating monitoring data with respect to the regulatory standard. Moreover, this method is easily implemented with supporting packages in the R [34] environment, and so can be used by a broad group of scientists beyond statisticians.

6.2 The Direct Extreme Value Model

Chapter 5 provided an introduction to incorporation of spatial dependence of extreme value statistics, and then applied these ideas to the spatial example drawn from the ozone dataset. Part of the interest here is to compare the results with the daily model that focuses on the central part of the distribution of standardized ozone measurements. For interest only in the regulatory statistic, the probability that the FHDA exceeds 84 ppb in a year is similar using either the extremes or space-time modeling approach–at least for this study region in 1995 through 1999. One interesting feature of this correspondence is that the approaches are very different in character, and involve very different assumptions. The extremes approach largely ignores temporal and spatial dependence conditional on the parameters of the GPD, but is more flexible in representing the larger values of observed ozone. The space-time approach is a hierarchical model that represents the daily dependences of ozone over time and its correlation over space, but it relies on normal distributions for the daily distributions. The agreement between the surfaces in Fig. 5.3 suggests that for both approaches the assumptions are reasonable.

6.3 Other Applications

It is important to note that while all of these models have been motivated by the NAAQS for ground-level ozone, the problem solved for this standard is

much more generally applicable. For example, the extreme value model could be used to characterize extreme events for numerous other types of data, such as extreme precipitation events, convective weather events, or occurrences of health problems such as west nile virus to name only a few. There are also other criteria pollutants that could benefit from either the extreme value model or the space-time daily model. For example, particulate matter (PM) has a similar standard to that of ozone that is based on a high percentile.

Additionally, both the space-time daily model and the extreme value model can easily be modified to account for different standards. For example, if the U.S. EPA decided to change the ozone standard to one based on a third-highest value instead of the current fourth-highest value, it is a small matter to take the third-highest value from Step 4 of the daily model algorithm. Of course, any order statistic or quantile could be sampled in Step 4 providing a great deal of flexibility in the daily model approach. This is also true for the extreme value approach because the approach does not specifically rely on the particular order statistic.

For the seasonal model approach, care should be taken to check that the field is still multivariate Gaussian for other order statistics. It was found here that the assumption that the field is, at least approximately, Gaussian is reasonable for the fourth-highest order statistic field. However, this may not be the case for other order statistics; especially for higher order statistics than the fourth-highest.

6.4 Future Work

Here, covariates have been only surreptitiously included in the daily model. On a daily level, it is known that certain meteorological covariates are important in the creation and transport of ozone–particularly temperature, cloud cover and wind. Generally, meteorological data, such as temperature, are difficult to use for the daily model approach because at any time point the temperatures or wind vectors at two locations can vary greatly, and meteorological measurements are generally gathered only on a coarse spatial scale. Therefore, incorporation of such covariates becomes problematic. Two possible covariates, however, that could be incorporated more explicitly into the daily model are elevation and aspect. In particular, it might be useful to use these covariates when interpolating the autoregressive parameters spatially.

For the daily model, I have considered some parameter uncertainty in parts of the model (the autoregressive parameters and parameters of the spatial shocks covariance function), but have not propagated the uncertainty into the FHDA fields. A fully Bayesian model could perhaps synthesize covariates, model parameters, and any uncertainty associated with them elegantly, but at the cost of much greater complexity and computational intensity. Note that by varying the model parameters in the algorithm, one can include uncertainty into the daily model analysis resulting from uncertainty in the parameters. Although a fully Bayesian approach might provide a very elegant solution,

bootstrapping is a good compromise in terms of less demands for new software and computing resources; and scales to a larger area without further complication. For example, one could use a parameteric bootstrap to generate a sample of parameters that reflect the uncertainty (in a frequentist sense!) with respect to the MLEs. These values would then be used to generate the conditional FHDA fields.

There are many areas of extreme value statistics that need more statistical research; including algorithms to compute (or sample) posteriors from a Bayesian analysis. A key step would be the ability to sample the surface of GPD parameters from a posterior. This would allow for quantifying the uncertainty in the estimated parameters and the subsequent quantities based on the GPD; such as return times and exceedance probabilities. Despite many open methodological questions, there is much benefit from an extremes perspective. In particular, if one is interested in extreme events, it may be possible to avoid some of the complexity of the spatial and temporal dependence that is ordinarily associated with the majority of the measurements.

A primary direction for future work is to extend model (6.1) to the entire Eastern United States. This project would produce an analysis comparable to Fuentes' [11] work. Initial results indicate that the deseasonalization process, together with the autoregressive model, leaves a spatial shock field that is approximately stationary even over this relatively large region. The nonstationary model proposed by Fuentes [11] does not seem appropriate to use

with model (6.1). Particularly, there are large regions where the spatial shocks field appears to be homogeneous, with only small pockets (usually at higher altitudes) where the field appears to be nonstationary. Because covariance (3.40) fits covariances across entire regions, it is impossible to attend to only small variations in the field. Additionally, this Fuentes covariance mixture model assumes that fields around each node are orthogonal, or independent, of each other; but if the field is actually stationary, then two fields would not be independent. Therefore, this model does not handle the stationary case. Furthermore, there are many parameters to be estimated; from the placement and number of nodes, to the regional covariance models; which adds a new level of complexity to the model that might not be worth the return. A better choice would be one that fits covariances to individual stations as in the Higdon [21] model, or the associated Paciorek [33] model. It should be noted that covariance (3.42) is only defined for Euclidean distance. Although a similar technique could be employed for spherical distances, it is not clear that a closed form of the model could be obtained. A practical solution would be to project coordinates to a 2 dimensional plane with the assumption that a relatively small region on the Earth is well represented by a rectangular surface patch.

Bibliography

[1] Bochner, S. *Harmonic Analysis and the Theory of Probability.* University of California Press, Berkeley and Los Angeles, 1955.

[2] Brockwell, Peter J., Davis, Richard A. *Time Series: Theory and Methods (Second Edition).* Springer-Verlag, N.Y., 175 Fifth Avenue, New York, N.Y. 10010, 1991.

[3] Coles, Stuart. *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, London, London, England, 2001.

[4] Cressie, Noel A.C. *Statistics for Spatial Data (Revised Edition).* Wiley Interscience, NY, 1993.

[5] Cressie, Noel and Huang, Hsin-Cheng. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94:1330–1340, 1999.

[6] Davis, Jerry M., Eder, Brian K. and Bloomfield, Peter. Regional and temporal models for ozone along the Gulf Coast. *Lecture Notes in Statis-*

tics: Case Studies in Environmental Statistics, volume 132, pages 5–24. Edited by Douglas Nychka, Walter W. Piegorsch and Lawrence H. Cox. Springer, 175 Fifth Avenue, New York, NY 10010, 1998.

[7] Davis, Jerry M., Eder, Brian K. and Bloomfield, Peter. Modeling Ozone in the Chicago Urban Area. *Lecture Notes in Statistics: Case Studies in Environmental Statistics*, volume 132, pages 5–24. Edited by Douglas Nychka, Walter W. Piegorsch and Lawrence H. Cox. Springer, 175 Fifth Avenue, New York, NY 10010, 1998.

[8] Davis, J.M. and Speckman, P. A model for predicting maximum and 8h average ozone in houston. *Atmospheric Environment*, 33:2487–2500, 1999.

[9] Davis, R.A.. Stable limits for partial sums of dependent random variables. *Annals of Probability* 11:262–269, 1983.

[10] Ferro, Christopher A. T. and Segers, Johan. Inference for clusters of extreme values. *Journal of the Royal Statistical Society B*, 65(2):545–556, 2003.

[11] Fuentes, Montserrat. Statistical assessment of geographic areas of compliance with air quality. *Journal of Geophysical Research*, 108(D24), 2003.

[12] Gao, Feng, Sacks, Jerome and Welch, William J. Predicting urban ozone levels and trends with semiparametric modeling. *Journal of Agricultural, Biological and Environmental Statistics*, 1:404–425, 1996.

[13] Gaspari, Gregory and Cohn, Stephen E. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125:723–757.

[14] Gneiting, Tilmann. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002.

[15] Gneiting, Tilmann. Correlation functions for atmospheric data analysis. *Quarterly Journal of the Royal Meteorological Society*, 125:2449–2464, 1999.

[16] Green, P.J. and Silverman, B.W. *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, 2-6 Boundary Row, London SE1 8HN, UK, 1994.

[17] Guttorp, Peter, Meiring, Wendy and Sampson, Paul D. A space-time analysis of ground-level ozone data. *Environmetrics*, 5:241–254, 1994.

[18] Haas, T.C. Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association*, 90:1189–1199, 1995.

[19] Hastie, T.J. and Tibshirani, R.J. *Monographs on Statistics and Applied Probability 43: Generalized additive models.* Chapman and Hall, First CRC reprint 1999, CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, FL 33431.

[20] Heffernan, Janet E. and Tawn, Jonathan A. A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society B*, 66(3):497–546, 2004.

[21] Higdon, D., Swall, J., and Kern, J. Non-stationary spatial modeling. *Bayesian Statistics* 6, eds. J. Bernardo, J. Berger, A. Dawid, and A. Smith, Oxford, U.K.: Oxford University Press, 761–768, 1999.

[22] Holland, David M., Cox, William M., Scheffe, Rich, Cimorelli, Alan J., Nychka, Douglas and Hopke, Philip K. Spatial prediction of air quality data. *EM A White Paper*, pages 31–35, 2003.

[23] Johnson, M.E., Moore, L.M. and Ylvisaker, D. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148, 1990.

[24] Kyriakis, P.C. and Journel, A.G. Geostatistical space-time models: A review. *Mathematical Geology*, 31:651–684, 1999.

[25] Leadbetter, M.R., Lindgren, G., and Rootzén, H. *Extremes and Related Properties of Random Sequences and Series.* Springer-Verlag, New York, 1983.

[26] Lefohn, Allen S. and Altshuller, A. Paul Environmental concentrations, patterns, and exposure estimates. `http://www.epa.gov/ttn/oarpg/naaqsfin/ria.html`. *Air Quality Criteria for Ozone and Related Photochemical Oxidants* chapter 4, 1996.

[27] LePage, R., Woodroofe, M., and Zinn, J. Convergence to a stable distribution via order statistics. *Annals of Probability* 9:624–632, 1981.

[28] Ma, Chunsheng. Families of spatio-temporal stationary covariance models. *Journal of Statistical Planning and Inference*, in press, 2002.

[29] Ma, Chunsheng. Stationary spatio-temporal covariance models. *Journal of Multivariate Analysis*, in press, 2002.

[30] Matern, B. *Lecture Notes in Statistics: Spatial Variation (2nd Edition).*, volume 36. Springer, NY, 1960, 1986.

[31] Nychka, D., Meiring, W., Royle, J.A., Fuentes, M. and Gilleland, E. Fields: R tools for spatial data. `http://www.cgd.ucar.edu/stats/software`, 2002.

[32] Nychka, Douglas and Saltzman, Nancy. Design of Air Quality Monitoring Networks. *Lecture Notes in Statistics: Case Studies in Environmental*

Statistics, volume 132, pages 51–75. Edited by Douglas Nychka, Walter W. Piegorsch and Lawrence H. Cox. Springer, 175 Fifth Avenue, New York, NY 10010, 1998.

[33] Paciorek, Chris, Schervish, Mark J. (advisor). *Ph.D. dissertation:* Nonstationary Gaussian processes for regression and spatial modelling. Carnegie Mellon University, Pittsburgh, PA, 2003.

[34] R Development Core Team R: A language and environment for statistical computing. *R Foundation for Statistical Computing* Vienna, Austria, ISBN 3-900051-00-3. `http://www.R-project.org`, 2003.

[35] Reich, Robin M. and Davis, Richard. *Quantitative Spatial Analysis.* Course Notes for NR/ST 523, Colorado State University, 2000.

[36] Resnick, S.I. *Extreme Values, Regular Variation, and Point Processes.* Springer, New York, 1987.

[37] Schoenberg, I. Metric spaces and completely monotone functions. *Annals of Mathematics* 39:811–841, 1938.

[38] Schlather, Martin and Tawn, Jonathan A. A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika* 90(1):139–156, 2003.

[39] Sampson, P.D. and Guttorp, P. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87:108–119, 1992.

[40] Smith, R.L. Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science*, 4:367–393, 1989.

[41] Smith, R.L. and Huang, L.-S. Modeling high threshold exceedances of urban ozone. Technical Report 6, National Institute of Statistical Sciences, Research Triangle Park, NC, 1993.

[42] Stein, Michael L. *Interpolation of spatial data: some theory for kriging.* Springer-Verlag, 175 Fifth Ave., New York, N.Y. 10010, 1999.

[43] U.S. Environmental Protection Agency. Community multi-scale air quality (cmaq). `http://www.epa.gov/AMD/models3/cmaq.html`.

[44] U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Emissions Monitoring and Analysis Group. National air quality and emissions trends report, 1997. `http://www.epa.gov/oar/aqtrnd97`.

[45] Wikle, C.K., Cressie, N. (Major Professor) and Chen, T.C. (Major Professor). *Ph.D. dissertation: Spatio-temporal statistical models with applications to atmospheric processes*, chapter A spatially descriptive, tem-

porally dynamic statistical model with applications to atmospheric processes. Iowa State University, Ames Iowa, 1996.