ABSTRACT

CHOI, JUNGSOON. Multivariate Spatial-Temporal Modeling of Environmental-Health Processes. (Under the direction of Professor Montserrat Fuentes.)

In many applications in environmental sciences and epidemiology, data are often collected over space and time. In some cases, the spatial-temporal data of interest are multivariate, and these multivariate spatial-temporal processes often have a complicated dependency structure. Hence, multivariate spatial-temporal modeling is a very challenging task. In this study, we develop statistical models to effectively account for multivariate spatial-temporal dependency structures of air pollution concentrations and human health outcomes.

Fine particulate matter ($PM_{2.5}$) is an atmospheric pollutant that has been linked to serious health problems, including mortality. $PM_{2.5}$ has five main components: sulfate, nitrate, total carbonaceous mass, ammonium, and crustal material. These components have complex spatial-temporal dependency and cross dependency structures. It is important to gain better understanding about the spatial-temporal distribution of each component of the total $PM_{2.5}$ mass, and also to estimate how the composition of $PM_{2.5}$ changes with space and time. We introduce a multivariate spatial-temporal model for speciated $PM_{2.5}$. Our hierarchical framework combines different sources of data and accounts for potential bias. In addition, a spatiotemporal extension of the linear model of coregionalization is developed to account for spatial and temporal dependency structures for each component as well as the associations among the components. We apply our framework to speciated $PM_{2.5}$ data in the United States for the year 2004.

In addition, the chemical composition of $PM_{2.5}$ varies across space and time so

the association between $PM_{2.5}$ and mortality could change with space and season. Thus, we develop and implement a multi-stage Bayesian framework that provides a very broad and flexible approach to studying the spatial-temporal associations between mortality and population exposure to daily $PM_{2.5}$ mass, while accounting for different sources of uncertainty. In the first stage, we map ambient $PM_{2.5}$ air concentrations using all available monitoring data and an air quality model (CMAQ) at different spatial and temporal scales. In the second stage, we examine the spatial-temporal relationships between the health end-points and the exposures to $PM_{2.5}$ by introducing a spatial-temporal generalized Poisson regression model. We adjust for time-varying confounders, such as seasonal trends. A common seasonal trends model uses a fixed number of basis functions to account for these confounders, but the results can be sensitive to the number of basis functions. Thus, instead the number of the basis functions is treated as an unknown parameter in our Bayesian model, and we use a space-time stochastic search variable selection method. The framework is illustrated using a data set in North Carolina for the year 2001.

Multivariate Spatial-Temporal Modeling of Environmental-Health Processes

by
Jungsoon Choi

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fullfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2008

APPROVED BY:

Dr. Montserrat Fuentes Chair of Advisory Committee	Dr. Jerry M. Davis
Dr. Sujit K. Ghosh	Dr. John F. Monahan
Dr. Sujit K. Ghosh	Dr. John F. Monahan

DEDICATION

To my family and friends

BIOGRAPHY

Jungsoon Choi was born in Busan, Republic of Korea in 1978. She entered the department of Statistics at Ewha Womans University in Seoul in 1997 and earned her B.S. in 2000 and M.S. in 2002 in statistics. In the fall of 2003, she joined the department of Statistics at North Carolina State University in Raleigh. She received her Ph.D. in Statistics in 2008.

ACKNOWLEDGMENTS

I could not complete the dissertation without the help of so many people. I would like to thank all the people who supported me to accomplish this dissertation.

First, I would like to express my deeply gratitude to my advisor, Dr. Montserrat Fuentes, who taught me how to do research, supported me financially, and provided insightful feedback and continuing encouragement. Under her direction, the past several years have been some of the most important and formative of my life. I am tremendously grateful for her mentoring and instruction throughout that time.

I am deeply indebted to Dr. Jerry Davis for his valuable comments and support. Dr. Davis has been very supportive and understanding. He helped me understand chemistry and meteorology things. I also would like to express my deepest gratitude to Dr. Reich. He has been of great help for me to complete this thesis. The members of my dissertation committee, Dr. Ghosh, Dr. Monahan, and Dr. Zhang have generously given their time and expertise to better my work. I thank them for their helpful suggestions and comments. I also would like to thank Dr. Pantula, Dr. Arroway, Dr. Wang, all my professors, Terry Byron, and Adrian Blue for all their help and encouragement in during my course of studies. I am also grateful to Dr. Holland and Prakash for their insightful comments about the real applications.

On a more personal level, I would like to express my deepest gratitude and biggest appreciation to my family and friends for their unconditionally support and encouragement. Without their encouragement and support, I would never have pursued graduate studies. Lastly, my boyfriend, Sangwon, has my deepest gratitude for his love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In diverse fields such as environmental sciences and epidemiology, data are collected over space and time. For example, temperature data are measured at many monitoring stations every day in the United States. Also, mortality data in the state of North Carolina are collected every day. These data often have spatial-temporal dependency, and such dependency structure should be considered when analyzing the data. Thus, developing statistical models to account for spatial and temporal dependency is not only very challenging but also essential. In some cases, the spatial-temporal data of interest are multivariate. For example, levels of several pollutants (e.g., sulfate, nitrate, and ammonium) are measured at monitoring stations over time. In this situation, multivariate statistical model is needed to explain cross dependency as well as spatial-temporal dependency. In this thesis, we develop statistical frameworks for analyses of multivariate spatial-temporal data on air pollution concentrations and human health outcomes.

In order to model spatial-temporal data, we first discuss different types of data. Spatial data are generally categorized into one of three basic types: point-referenced data, areal data, and point pattern data. *Point-referenced* data are called when a location $\mathbf{s} \in \mathbb{R}^d$ varies continuously over a fixed region, $D_1 \subset \mathbb{R}^d$ and often referred to as *geostatistical* data. For instance, weekly ozone concentrations are measured at fixed monitoring stations over the eastern United States. One of main interests in point-referenced data is to model the spatial distribution of measurements for taking into account spatial correlation and for predicting values at new locations. *Areal* data arise when a region D_1 is a fixed domain of regular or irregular shape and divided into a finite number of subregions with well defined boundaries (e.g., zipcodes, counties, and districts). The data are often referred to as *lattice* data, and statistical models for such data provide adjacency information of the areal units. For example, the number of deaths due to air pollution is collected in the counties of North Carolina. *Point pattern* data arise when a region D_1 is itself random, and an event of interest (e.g., an outbreak of a disease) occurs at random locations. That is, the response variable is often fixed, and only the locations are random. In some cases, this information might be supplemented by other covariate information (e.g., age) at the event locations. Such data are often interested in studying whether the pattern is exhibiting complete spatial randomness, clustering, or regularity. Residences of persons suffering from lung cancer could be one of examples. See the books by Banerjee et al. (2004) and Schabenberger and Gotway (2004) for many examples and additional discussion.

Time series data like spatial data are classified into one of two types: continuous or discrete. *Continuous* time series data arise when observations are made at every

instant of time (e.g., lie detectors and electrocardiograms). *Discrete* time series data arise when observations are made at equi-spaced intervals (e.g., weekly rainfall). Note that in practice discrete time series can approximate continuous time series.

In this thesis, we consider point-referenced data or areal data for space and discrete time series data. In this chapter, we provide some general concepts in modeling spatial-temporal data in Section 1.1, and then we give some background for the spatial-temporal modeling of air pollution and its health effects in Sections 1.2 and 1.3, respectively.

1.1 Introduction of spatial-temporal process

We suppose that the data

$$\{Z(\mathbf{s}_i, t) : i = 1, \ldots, N,\ t = 1, \ldots, T\} \tag{1.1}$$

are a finite sample of the stochastic process $\{Z(\mathbf{s}, t) : \mathbf{s} \in D_1, t \in D_2\}$, where $D_1 \subset \mathbb{R}^d$ and $D_2 \subset \mathbb{R}$. The process $Z(\mathbf{s}, t)$ is called *weakly stationary* (or *stationary*) if its mean function is constant and

$$\mathrm{Cov}(Z(\mathbf{s} + \mathbf{h_1}, t + h_2), Z(\mathbf{s}, t)) = C(\mathbf{h_1}, h_2) < \infty, \tag{1.2}$$

where $\mathbf{h_1} \in \mathbb{R}^d$ and $h_2 \in \mathbb{R}$. Thus, a stationary spatial-temporal process has the covariance depending only on the separation vector $(\mathbf{h_1}, h_2)$.

A stationary process is called *isotropic* if

$$C(\mathbf{h_1}, h_2) = C(||\mathbf{h_1}||, |h_2|). \tag{1.3}$$

Thus, the covariance function depends on vector $(\mathbf{h_1}, h_2)$ only through Euclidean distance between them. A process which is not isotropic is called *anisotropic*. Isotropic processes are simple and interpretable, and they have several simple parametric forms for the covariance, so they are very popular.

The process $Z(\mathbf{s}, t)$ is called *separable* in space and time if

$$\text{Cov}(Z(\mathbf{s}, t), Z(\mathbf{s}', t')) = C^{(1)}(\mathbf{s}, \mathbf{s}')C^{(2)}(t, t'), \tag{1.4}$$

where $C^{(1)}$ is a spatial covariance and $C^{(2)}$ is a temporal covariance. The covariance matrix of a separable spatial-temporal process can be written as the Kronecker product of a covariance matrix for time with a covariance matrix for space. Thus, it is easy to compute the covariance matrix. If a separable process is stationary, then

$$\text{Cov}(Z(\mathbf{s}, t), Z(\mathbf{s}', t')) = C^{(1)}(\mathbf{s} - \mathbf{s}')C^{(2)}(t - t'). \tag{1.5}$$

In addition, if a separable stationary process is isotropic, the covariance can be written as

$$\text{Cov}(Z(\mathbf{s}, t), Z(\mathbf{s}', t')) = C^{(1)}(||\mathbf{s} - \mathbf{s}'||)C^{(2)}(|t - t'|). \tag{1.6}$$

A common choice for the functions $C^{(1)}(\cdot)$ and $C^{(2)}(\cdot)$ is the Matern covariance function (Matern, 1986) and the ARMA covariance function (Box and Jenkins, 1976), respectively.

There is a large amount of literature on spatial-temporal modeling. Early approaches include the STARMA (Pfeifer and Deutsch, 1980a,b) and STARMAX (Stoffer, 1986) models which add spatial covariance structure to standard time series models. Recently, Carroll et al. (1997) used a spatially homogeneous and temporally

stationary Gaussian process models, assuming a separable spatial-temporal covariance function to study ground level ozone. Brown et al. (2000) modeled rainfall data by using a stochastic differential equation approach, which is non-separable spatial-temporal model. They addressed that non-separable modeling is more appropriate for realistic problems.

To capture non-separable and/or non-stationary structure when analyzing spatial-temporal data, Cressie and Huang (1999) and Gneiting (2002) introduced non-separable stationary covariance functions using a spectral approach. Stein (2005) also provided a class of spectral densities extending the Matern form. These densities allow spatial-temporal covariance functions to be non-separable. Guttorp et al. (1994) proposed a non-stationary spatial-temporal model for ozone levels via the spatial deformation method of Sampson and Guttorp (1992).

Researchers have also worked with dynamic models, or state-space models for analyzing spatial-temporal data. Examples include Huang and Cressie (1996), Mardia et al. (1998), and Wikle and Cressie (1999). Huang and Cressie (1996) provided a separable dynamic model to predict snow water, and Mardia et al. (1998) proposed a kriged Kalman filter.

In recent years, Bayesian approaches are becoming very popular for spatial-temporal modeling. Since spatial-temporal data in environmental sciences and epidemiology have variability over space and time, the statistical characterization of such complicated processes using traditional spatial-temporal models and methods can be limited. Moreover, there can often be very different spatial behavior at different points in time as well as different temporal variability at different locations in space, and the space-

time variability can be more complicated. However, Bayesian hierarchical modeling provides a natural framework in which scientific knowledge is included and accounts for all sources of uncertainty. Also, it offers not only a technique for combining data from different sources but also posterior distributions on quantities of interest that can be used for scientific inference.

There are various examples of hierarchical Bayesian modeling for spatial-temporal data. For example, Wikle et al. (1998) analyzed monthly averaged maximum temperature data by introducing a Bayesian non-stationary spatial-temporal model. Golam Kibria et al. (2002) developed a Gaussian generalized inverted Wishart model to map air pollution in Philadelphia, allowing the covariance structure to be non-stationary. Schmidt and O'Hagan (2003) used a Bayesian approach to construct a non-stationary spatial-temporal covariance structure using spatial deformations. Fuentes et al. (2005) proposed a generalized class of non-stationary and also non-separable spatial-temporal covariance models. In addition, a Markov random field structure in the form of conditionally autoregressive (CAR) specifications has been employed on spatial-temporal modeling for areal data. Using Markov random field structures, Waller et al. (1997) analyzed Ohio lung cancer rates during the period 1968-1988, and Gelfand et al. (1998) focused single family home sales.

Similarly, Bayesian approaches have been reported in dynamic modeling. In West and Harrison (1997), Markov Chain Monte Carlo methods for dynamic modeling were explained in detail. Tonellato (1997) developed a state-space model for fixed observation stations, incorporating both stationary and non-stationary components, then Tonellato (1998) applied the model to an Irish wind power prediction problem.

The dynamic model proposed by Stroud et al. (2001) is applied to any data set that is discrete in time and continuous in space and allows quick and straightforward computation using Kalman filtering. Gelfand et al. (2005) explained univariate space-time dynamic models and multivariate spatial dynamic models.

We briefly introduce the hierarchical Bayesian model proposed by Wikle et al. (2001). The strategy for the basic hierarchical spatial-temporal model is based on the formulation of three stages:

- Stage 1: Data model: $[data|process, \theta_1]$
- Stage 2: Process model: $[process|\theta_2]$
- Stage 3: Prior on parameters: $[\theta_1, \theta_2]$

Here $[\cdot]$ represents the probability density function, $[\cdot|\cdot]$ represents the conditional density function (e.g., Gelfand and Smith, 1990), and θ_1 and θ_2 are parameters introduced in the modeling. The basic idea is to approach complex problems by breaking them into subproblems.

1. *First stage: Data Model*

   Stage 1 models only measurement errors. Let $\widehat{Z}$ denote observed data. The model is specified as

   $$[\widehat{Z}|Z, \theta_1], \tag{1.7}$$

   where Z is the true process, and θ_1 is a collection of parameters considered in this stage. The fundamental assumptions are that, conditional on Z and

θ_1, $\widehat{Z}(\mathbf{s},t)$ for all $\mathbf{s}$ and t are independent. For example, we assume normally distributed errors,

$$\widehat{Z}(\mathbf{s},t) \sim N(Z(\mathbf{s},t), \sigma^2_{s,t}), \tag{1.8}$$

where $\theta_1 = \{\sigma^2_{s,t}\}$ is the set of measurement error variances.

2. *Second stage: Process model*

   In this stage, the true process Z is decomposed into several meaningful components. We assume

$$Z(\mathbf{s},t) = \mu(\mathbf{s},t) + w_1(\mathbf{s};\theta_{w1}) + w_2(t;\theta_{w2}), \tag{1.9}$$

   where $\mu(\mathbf{s},t) = X(\mathbf{s},t)\beta(\mathbf{s},t)$ is the mean structure with the vector of covariates $X(\mathbf{s},t)$ and the vector of coefficients $\beta(\mathbf{s},t)$, w_1 is a spatial random effect with the parameter vector θ_{w1}, and w_2 is a temporal random effect with the parameter vector θ_{w2}. Under the assumption that the components μ, w_1, w_2 are mutually independent,

$$[Z|\boldsymbol{\beta}, \theta_{w1}, \theta_{w2}] = [\mu|\boldsymbol{\beta}][w_1|\theta_{w1}][w_2|\theta_{w2}], \tag{1.10}$$

   where $\boldsymbol{\beta} = \{\beta(\mathbf{s},t)\}$ and $\theta_2 = \{\boldsymbol{\beta}, \theta_{w1}, \theta_{w2}\}$.

3. *Third stage: Prior on parameters*

   This stage specifies the priors for all parameters in the model. The standard assumption is that

$$[\theta_1, \theta_2] = [\theta_1][\theta_2]. \tag{1.11}$$

   In this case, it can be written as

$$[\theta_1, \theta_2] = [\sigma^2][\boldsymbol{\beta}][\theta_{w1}][\theta_{w2}]. \tag{1.12}$$

1.2 Statistical models for air pollution

Particulate matter (PM) is a complex mixture of small solid and liquid particles found in the ambient air. Generally, PM consists of many components including dust, dirt, smoke, acids (such as nitrates and sulfates), and organic chemicals. PM has been linked to a range of cardiovascular, respiratory, and other serious health problems including mortality. PM is classified by its size, aerodynamic diameter, because the size of particles is directly linked to their potential for causing adverse health outcomes. In particular, particles of interest are PM_{10} and $PM_{2.5}$. PM_{10} includes particles less than or equal to $10\mu m$ in diameter, and $PM_{2.5}$ (known as fine PM) includes particles less than or equal to $2.5\mu m$ in diameter. Exposure to these smaller particles leads to a variety of serious health effects because of their ability to penetrate into the respiratory tract. Moreover, the U.S. Environmental Protection Agency (EPA) set standards for $PM_{2.5}$ in 1997 because fine particles are more closely associated with serious health effects.

$PM_{2.5}$ has five main components by U.S. EPA (2003): sulfate, nitrate, total carbonaceous mass, ammonium, and crustal material (including calcium, iron, silicon, aluminum, and titanium). The different components come from specific sources and are often formed in the atmosphere. Sulfates form from sulfur dioxide emissions from power plants and industrial facilities, and nitrates form from nitrogen oxide emissions from cars, trucks, and power plants. Total carbonaceous mass is a mixture of elemental and organic carbon. Elemental carbon is emitted directly from fossil fuel combustion sources, and organic carbon is from combustion, geological processes,

road dusts, and photochemistry. Ammonium sulfate and ammonium nitrate are the most common compounds containing ammonium. Crustal material is estimated by summing the elements predominantly associated with soil.

These $PM_{2.5}$ components vary by location and by time of year. For example, fine particles in the eastern United States contain more sulfates than those in the western United States, while fine particles in southern California contain more nitrates than other areas of the country. Total carbonaceous mass is a substantial component of $PM_{2.5}$. Thus, the components have complex spatial-temporal dependency and cross dependency structures. In addition, some of the toxicology studies indicated that some components of $PM_{2.5}$ were more closely linked with human health effects than other components (U.S. EPA, 2004). Fuentes et al. (2006) introduced a statistical model to study the spatial association between speciated $PM_{2.5}$ and mortality and provided that each component of speciated $PM_{2.5}$ has different impact on mortality across the United States. In order to investigate the health effects associated to speciated fine PM across space and time together, we need to have first the speciated $PM_{2.5}$ information. However, daily speciated $PM_{2.5}$ measurements are available at a limited number of monitoring stations, and missing values may occur at a given time point. Rao et al. (2003) and Malm et al. (2004) investigated the spatial and temporal patterns of speciated $PM_{2.5}$, but they only conducted an exploratory analysis of speciated $PM_{2.5}$. Thus, it is very challenging to develop a spatial-temporal model for speciated $PM_{2.5}$ to study its spatial-temporal patterns and predict speciated $PM_{2.5}$.

There exists considerable literature on the spatial-temporal modeling of air pollution data. In an early Bayesian approach, Guttorp et al. (1994) developed a spatial

covariance function for hourly ozone levels using the deformation approach of Sampson and Guttorp (1992). The parameters of the model depended on a function of the time of day. Carroll et al. (1997) considered stationary Gaussian processes based on a separable spatial-temporal covariance function to study ground level ozone in Texas. In their model, the correlation in the residuals was a nonlinear function of space and time, and in particular the spatial structure was a function of the lag between observations.

In recent years, hierarchical Bayesian approaches for air pollution modeling have been developed. Zidek et al. (2002) developed predictive distributions of hourly PM_{10} at unmonitored sites in Vancouver, Canada. Shaddick and Wakefield (2002) used a dynamic linear modeling framework for four pollutants at eight monitoring sites in London to provide predictions of the pollutants for investigating the health effects of air pollution. Huerta et al. (2004) modeled hourly ozone levels in Mexico City and estimated ozone values using a dynamic model. Sahu et al. (2006) developed a hierarchical non-stationary spatial-temporal model for $PM_{2.5}$ in 2001. Their model introduced two spatial-temporal processes: one for rural or background effects and the other for additional risks for urban areas. The daily $PM_{2.5}$ was modeled using a weighted combination of the two processes.

Smith et al. (2003) developed a spatial-temporal model for $PM_{2.5}$, which was used for spatial prediction within three southeastern states. The $PM_{2.5}$ field was represented as the sum of semi-parametric spatial-temporal functions with a random component that was spatially but not temporally correlated. They applied a variant of the expectation-maximization algorithm to explain missing data. Sahu and Mardia

(2005) used the Bayesian kernel convolution approach discussed by Ver Hoef and Barry (1998) for short-term forecasting of $PM_{2.5}$ data.

In Chapter 2 of this thesis, we propose a multivariate spatial-temporal model for speciated $PM_{2.5}$ using a hierarchical Bayesian approach in order to explore the spatial-temporal patterns of speciated $PM_{2.5}$ and predict speciated $PM_{2.5}$ at all locations and times of interest. In addition, we introduce a statistical framework in order to combine all available information for speciated $PM_{2.5}$.

1.3 Statistical models for the health effects of air pollution

The health impact of air pollution has received much attention in public health over the past few years, and numerous epidemiological studies investigated the association between exposure to PM and adverse human health outcomes in a number of U.S. cities. Several epidemiologic studies have showed that exposures to PM may result in tens of thousands of excess deaths per year and many more cases of illness among the U.S. population (e.g., Bates et al., 1990; Ostro et al., 1991; Dockery et al., 1992; Schwartz, 1994; Pope et al., 1995a; American Thoracic Society and Bascom, 1996a,b). Most of the epidemiological studies of air pollution have been conducted based on two types: time series studies and cohort studies. Time series studies assess the association between short-term exposure to PM and adverse health outcomes, and they estimate acute air pollution effects on health outcomes (e.g., Dockery et

al., 1992; Schwartz, 1994; Dominici et al., 2000). On the other hand, cohort studies investigate the association between long-term exposure to PM and adverse health outcomes over several years, so they are suitable for the study of chronic air pollution effects on health outcomes (e.g., Dockery et al., 1993; Pope et al., 1995b). However, cohort studies are limited by the lack of available data, and we focus on time series studies in this thesis.

By the early 1990's, most of time series studies were conducted at a single site (e.g., Bates et al., 1990; Schwartz, 1995) and showed health effects associated with exposure to air pollution. However, these studies obtained results from a specific site, and the statistical approaches vary with each study. In addition, sites have different characteristics. Thus, the results can not be generalized to other sites. Due to the limitations of single-site time series studies, multi-site time series studies were introduced, which data on air pollution and health for each site are assembled under a common structure and analyzed using the same statistical approach (Burnett and Krewski, 1994; Katsouyanni et al., 1997; Dominici et al., 2000). Hierarchical modeling is an appropriate approach for combining information on air pollution and health across multi sites, and recently, it has been applied to analysis of multi-site time series data. Dominici et al. (2000) estimated the association between exposure to PM_{10} and daily mortality for the 20 largest cities in U.S. using a two-stage linear regression model. Daniels et al. (2000) then extended this study to estimate the shape of the PM_{10} mortality dose-response curve. The National Morbidity, Mortality and Air Pollution Study (NMMAPS) (Samet et al., 2000a,b) is the largest multi-site time series study analyzing data for the 90 largest U.S. cities. The NMMAPS

estimated city-specific, regional, and national effects of PM_{10} on mortality. Dominici et al. (2002a) developed a three-stage Bayesian hierarchical model to analyze the NMMAPS data base. The results obtained for the 88 largest cities in U.S. from 1987 to 1994 showed positive associations between previous-day PM_{10} concentration and mortality in most of the cities.

Most previous studies on the statistical association between PM and health effects have been done only for PM_{10} because until recently, daily $PM_{2.5}$ concentrations were available at only a limited number of monitoring stations. More recently, Fuentes et al. (2006) combined several sources of data about fine PM and showed the spatial association between $PM_{2.5}$ and mortality in the entire United States. But, they only used monthly data and did not consider temporal association. Thus, it is suggested that more studies are needed to investigate the spatial and temporal association between $PM_{2.5}$ and daily mortality.

An important issue when studying the association between ambient PM concentrations and daily mortality counts is whether the increased mortality associated with higher PM levels is restricted to very frail people for whom life expectancy is short even in the absence of PM exposure. This possibility is called the "harvesting hypothesis" (also known as mortality displacement). If the effect exists, then higher $PM_{2.5}$ levels would lead to an increased risk of mortality for frail people and decrease the pool of at-risk people. On subsequent days, the number of death counts would be reduced, and the association between PM and daily mortality could be negative. Recently, some of researchers have proposed statistical methods to investigate this issue (e.g., lag models and time-scale models). Almon (1965) first proposed the dis-

tributed lag model to estimate the effect of air pollution exposure over a few days on daily health outcome. The model restricts the coefficients to being a low-degree polynomial in the lags. Pope and Schwartz (1996) used parametric distributed lag models, and Zanobetti et al. (2000) used non-parametric smoothing functions for the coefficients. Time-scale models (e.g., Zeger et al., 1999; Schwartz, 2000; Dominici et al., 2003) are used to estimate associations between smooth variations of air pollution and daily health outcomes using orthogonal predictors obtained by applying a Fourier decomposition. Zeger et al. (1999) used a frequency domain log-linear regression to estimate the association between air pollution and mortality at different timescales. Schwartz (2000) examined the association between daily mortality and different timescales of air pollution using the filtering method proposed by Cleveland et al. (1990). Based on the work by Zeger et al. (1999) and Schwartz (2000), Dominici et al. (2003) decomposed the time series of air pollution into orthogonal components using a discrete Fourier transform and estimated a relative risk of mortality at each component. This decomposition technique is simple and it provides computational advantages.

In Chapter 3, we explore the spatial-temporal association between exposure to $PM_{2.5}$ and daily mortality that is resistant to short-term harvesting by using time-scale models. We develop a Bayesian hierarchical framework to estimate the effects of $PM_{2.5}$ on mortality over space and time. In order to overcome the lack of fine PM data, we use numerical model output as well as observed $PM_{2.5}$ data, which improve estimates of the effects of $PM_{2.5}$ on mortality.

Chapter 2

Multivariate spatial-temporal modeling and prediction of speciated fine particles

2.1 Introduction

The study of the association between ambient particulate matter (PM) and human health has received much attention in epidemiological studies in the last several years. Özkaynak and Thurston (1987) conducted an analysis of the association between several particle measures and mortality using available data in 1980. Their results showed the importance of considering particle size, composition, and source information when modeling particle pollution health effects. In particular, fine PM,

$PM_{2.5}$, is an atmospheric pollutant that has been linked to numerous adverse health effects (e.g., respiratory and cardiovascular diseases). As presented in the previous chapter, $PM_{2.5}$ is a mixture of pollutants which the U.S. EPA (2003) classified into five main components: sulfate, nitrate, total carbonaceous mass (TCM), ammonium, and crustal material (including calcium, iron, silicon, aluminum, and titanium). These $PM_{2.5}$ components have complex spatial-temporal dependency as well as cross dependency structures and each component of speciated $PM_{2.5}$ has different impact on mortality. In order to study the association between speciated $PM_{2.5}$ and adverse health outcomes across space and time, we need to interpolate speciated $PM_{2.5}$ at the locations and times of interest. However, previous studies just carried out an exploratory analysis of speciated $PM_{2.5}$ (Rao et al., 2003; Malm et al., 2004). Our goal here is to develop a statistical framework using all available sources of data about speciated $PM_{2.5}$ to investigate the spatial-temporal patterns of speciated $PM_{2.5}$ and then predict speciated $PM_{2.5}$ at all locations and times of interest. We also study the spatial-temporal patterns of the so called "unknown components" which are not the main components of $PM_{2.5}$.

In this chapter, we introduce a new statistical framework to combine information for speciated $PM_{2.5}$ from two monitoring networks, while accounting for potential bias. Daily speciated $PM_{2.5}$ measurements are available at a limited number of monitoring sites and missing values are common. Therefore, we supplement these observations with measurements of the $PM_{2.5}$ mass. These observations are indirectly informative about the individual components and greatly expand our spatial and temporal coverage. Incorporating total $PM_{2.5}$ measurements poses a challenging data fusion

problem. In our Bayesian approach, speciated $PM_{2.5}$ data are represented in terms of the underlying true sum of the five main components and the true proportions of each speciated component relative to the total. We develop a spatial-temporal multinomial logistic model which allows the proportions to vary smoothly across space and time. Also, we extend the linear model of coregionalization (Grzebyk and Wackernagel, 1994; Wackernagel, 1998; Gelfand et al., 2004) to the spatiotemporal setting to account for the complex dependency structures of the speciated $PM_{2.5}$.

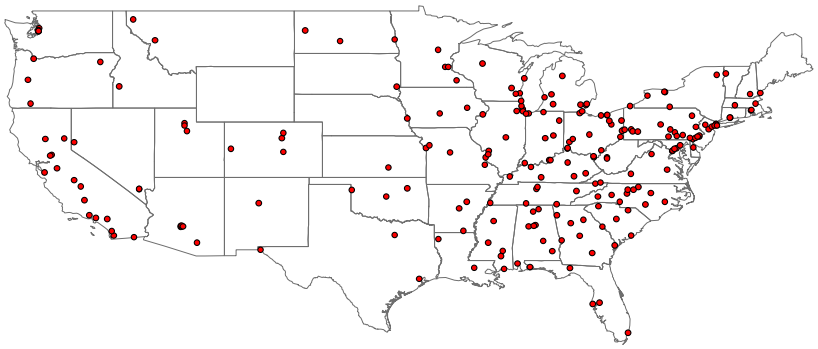
We use a speciated $PM_{2.5}$ data set that has not previously been analyzed. To our knowledge, this is the first time that a statistical framework has been used to analyze speciated $PM_{2.5}$ across the entire United States. A Bayesian hierarchical framework is used to study different random effects of interest. We show that the total $PM_{2.5}$ measurements are generally positively biased relative to the sum of the speciated components and that magnitude of the bias varies across the U.S. We also show that the proportions of each component vary considerably across space and time and that accounting for the cross-dependency in the speciated components dramatically improves prediction. In addition, we present a new approach to combine different sources of $PM_{2.5}$ information, which improves the prediction of the sum of the five components of $PM_{2.5}$. Wikle et al. (2001) presented a similar approach to combine different sources of information of ocean surface winds, but they treated one of the data sources as a prior process. In our approach, all of the data sources are simultaneously represented in terms of the underlying truth, and we also model the potential bias of the different sources of information as spatial-temporal processes.

This chapter is organized as follows. Section 2.2 describes the data used in this

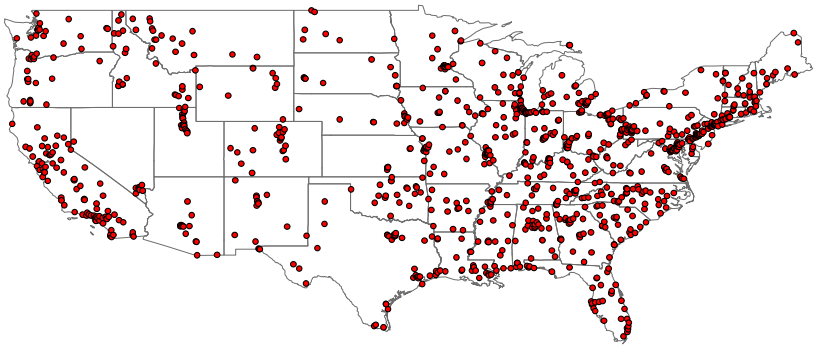
study. In Section 2.3, we present a Bayesian hierarchical multivariate spatial-temporal model for speciated $PM_{2.5}$. In Section 2.4, we present the results, and in Section 2.5, we offer a general discussion.

2.2 Data Description

$PM_{2.5}$ data from two monitoring networks and meteorological data in the conterminous United States for the year 2004 are used in this study. The first source of $PM_{2.5}$ data is the Speciated Trends Network (STN) established by the U.S. EPA in 1999. The STN measures the speciated $PM_{2.5}$ either every day, every third day, or every sixth day. It included about 200 monitoring sites in 2004, which are mostly in urban areas. Figure 2.1 (a) shows the STN monitoring stations around the nation in 2004. Even though the STN collects numerous trace elements, elemental carbon, organic carbon, and ions (sulfate, nitrate, sodium, potassium, and ammonium), we only consider the five main components of $PM_{2.5}$ presented in the previous section. In the STN, sulfate, nitrate, and ammonium are measured independently. Total carbonaceous mass is the sum of elemental carbon mass and estimated organic carbon mass which is $1.4 \times ([OC] - 1.53)$, where $[OC]$ is the measured organic carbon value, 1.4 is the factor to correct organic carbon mass for other elements (Rao et al., 2003), and 1.53 is the blank correction factor to adjust for sampling artifacts (Flanagan et al., 2003). Elemental carbon is also measured at the STN monitoring sites. Crustal material is computed using the IMPROVE equation (Malm et al., 2004) for the five most prevalent trace elements.



(a) STN



(b) FRM

Figure 2.1: (a) STN monitoring stations in 2004 (b) FRM monitoring stations in 2004.

Since the $PM_{2.5}$ data at the STN monitoring stations provide sparse spatial coverage, using only the STN monitoring data might be insufficient for modeling the speciated $PM_{2.5}$ over the entire United States. Therefore, we use the total $PM_{2.5}$ data from the Federal Reference Method (FRM) monitoring network which includes rural and urban sites and measures $PM_{2.5}$ samples either every day, every third day, or every sixth day. While the STN is a smaller network, the FRM network is a large national network, which consisted of about 1000 monitoring sites in 2004. Figure 2.1 (b) presents the monitoring stations of the FRM network in 2004.

Meteorological data for 2004 have been provided from the U.S. National Climate Data Center. We use five daily meteorological variables: minimum temperature (℃), maximum temperature (℃), dew point temperature (℃), wind speed (m/s), and pressure (hPa).

2.3 Statistical Models

While speciated $PM_{2.5}$ is only available at STN monitoring stations, information about the sum of the five components comes from both STN and FRM networks. Therefore we expect the sum of the five main components to be better identified than the individual components. Thus, we develop a statistical two-stage hierarchical framework using empirical Bayes approach. In the first stage, we model and estimate the sum of the five components using STN and FRM data across the entire United States by introducing a spatial-temporal framework. The posterior estimates for the true sum of the five components are used as the inputs in the next stage. In the second

stage, we introduce a multivariate statistical model for speciated $PM_{2.5}$ in terms of the estimate of the true sum of the five components and the relative proportion of each component to the sum. Thus, we can predict speciated $PM_{2.5}$ at locations and times of interest. As a consequence on this two-stage approach, our posterior estimates of variability for the speciated components will reflect our uncertainty in the proportions, but not our uncertainty in the sum of the five components in the first stage. However, our calibration analysis in Section 2.4 shows that our prediction intervals maintain the proper coverage probability.

2.3.1 Stage 1: Model for the sum of the five main components

We introduce a spatial-temporal model for the sum of the five main components using FRM and STN data (see Figure 2.2). We define the reconstructed fine mass (RCFM), $\widehat{Z}_R(\mathbf{s}, t)$, as the sum of the five main components from the STN data at location $\mathbf{s} \in D_1$ and at time $t \in D_2$, where $D_1 = \{\mathbf{s} : \mathbf{s}_1, \ldots, \mathbf{s}_{N_s}\} \subset \mathbb{R}^2$ and $D_2 = \{t : 1 = t_1, \ldots, T = t_T\} \subset \mathbb{R}$. We assume

$$\widehat{Z}_R(\mathbf{s}, t) = Z(\mathbf{s}, t) + e_R(\mathbf{s}, t), \tag{2.1}$$

where $Z(\mathbf{s}, t)$ is the true sum of the five components, and $e_R(\mathbf{s}, t) \sim N(0, \sigma_R^2)$ is the measurement error at location $\mathbf{s}$ and time t, which is independent of the true underlying process, $Z(\mathbf{s}, t)$.

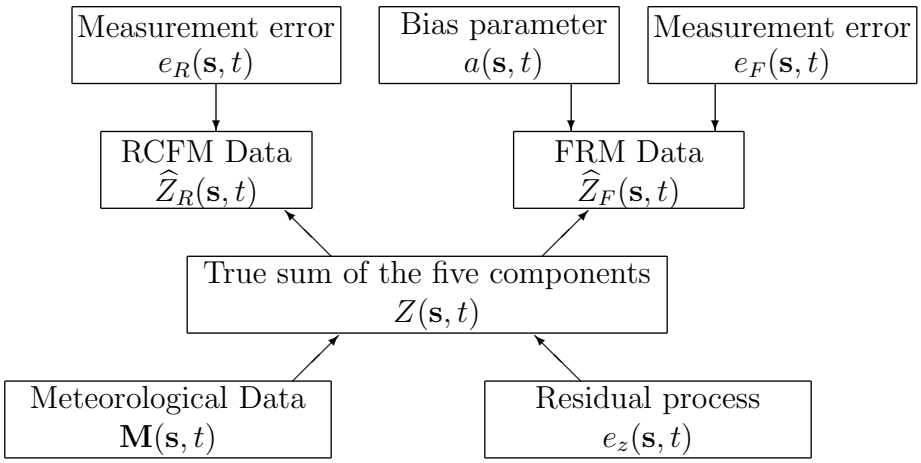


Figure 2.2: Model framework for the sum of the five main components.

A second estimate of the sum of the five main components is the observed total $PM_{2.5}$ mass from the FRM network, $\widehat{Z}_F(\mathbf{s},t)$. We model $\widehat{Z}_F(\mathbf{s},t)$ as

$$\widehat{Z}_F(\mathbf{s},t) = a(\mathbf{s},t) + Z(\mathbf{s},t) + e_F(\mathbf{s},t), \tag{2.2}$$

where $e_F(\mathbf{s},t) \sim N(0,\sigma_F^2)$ is the measurement error at location $\mathbf{s}$ and at time t, which is also independent of processes e_R and Z. Since total $PM_{2.5}$ mass from the FRM network consists of more pollutants than the five main components, the additive bias term $a(\mathbf{s},t)$ is needed. The bias term $a(\mathbf{s},t)$ accounts for systematic differences between the two networks. In addition, the bias term can be represented as "unknown components" which are not the main components of $PM_{2.5}$.

Exploratory analysis suggests that the additive bias varies over space and time, and we model $a(\mathbf{s},t)$ using a hierarchical framework,

$$a(\mathbf{s},t) = a_1(t) + a_2(\mathbf{s},t), \tag{2.3}$$

$$a_1(t) = h(t) + e_1(t), \tag{2.4}$$

$$a_2(\mathbf{s},t) = \delta_{a_2} a_2(\mathbf{s},t-1) + e_2(\mathbf{s},t), \tag{2.5}$$

where $a_1(t)$ represents the overall temporal trend in the bias of the FRM data, and $h(t)$ is a smoothing function of time to explain seasonality in the additive bias term. The process $a_2(\mathbf{s},t)$ accounts for the spatial-temporal structure which is not captured by the overall temporal trend. We assume the process $a_2(\mathbf{s},t)$ is an AR(1) with coefficient δ_{a_2}, and e_1 and e_2 are independent white noise processes and independent of the process Z. In this study, based on exploratory analysis, we assume that $h(t)$ is a linear combination of one sine and one cosine function with respect to one-year period. We use a normal prior, N(0,0.1) (0.1 is the precision), for δ_{a_2}.

The true sum of the five components, $Z(\mathbf{s},t)$, is modeled using a dynamic spatiotemporal linear model (Gelfand et al., 2005). We assume

$$Z(\mathbf{s},t) = \mathbf{M}^T(\mathbf{s},t)\boldsymbol{\beta}(\mathbf{s},t) + e_z(\mathbf{s},t), \tag{2.6}$$

where $\mathbf{M}(\mathbf{s},t)$ is a vector of meteorological variables (minimum temperature, maximum temperature, dew point temperature, wind speed, and pressure) and $e_z(\cdot,t) = (e_z(\mathbf{s}_1,t),...,e_z(\mathbf{s}_{N_s},t))$ is normal with mean $\psi_z e_z(\cdot,t-1)$ and, based on exploratory analysis, exponential covariance $\sigma_z^2 \exp(-h_1/\phi_z)$, where $h_1 = ||\mathbf{s}-\mathbf{s}'||$ (in km). In general, the regression coefficients $\boldsymbol{\beta}(\mathbf{s},t)$ may vary across space and time. In Section 2.4's application, $\boldsymbol{\beta}(\mathbf{s},t)$ is constant over time and constant within nine geographic

regions. The meteorological data might not exist at all the sites of interest. We thus interpolate the weather data at those locations using a spatial model for each time point (as part of the hierarchical framework). We use a normal prior, N(0,0.1) (0.1 is the precision), for ψ_z and uniform priors, Unif(1,1000) and Unif(0,100), for ϕ_z and σ_z, respectively.

Since the results are somewhat sensitive to the priors of the standard deviations σ_F and σ_R, σ_F and σ_R are fixed based on prior information regarding the precision of the FRM monitoring devices and the STN monitoring devices, respectively. U.S. EPA (1997, 2000) suggested that the coefficient of variation (CV) is 15% for the FRM data and 13% for the RCFM data. From the CV, the standard deviation can be calculated as sd = CV*mean/100, giving $\sigma_F = 1.796$ and $\sigma_R = 1.376$.

We seek to predict values of Z at location $\mathbf{s}_0$ and time t_0 given the data, $\widehat{Z}_F$, $\widehat{Z}_R$, and $\mathbf{M}$. Therefore, the posterior predictive distribution of $Z(\mathbf{s}_0, t_0)$ given the observations $\widehat{Z} = (\widehat{Z}_F, \widehat{Z}_R)$ and $\mathbf{M}$ is

$$p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}) \propto \int p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}, \mathbf{\Theta}_Z) p(\mathbf{\Theta}_Z|\widehat{Z}, \mathbf{M}) d\mathbf{\Theta}_Z, \tag{2.7}$$

where $\mathbf{\Theta}_Z$ is a collection of all parameters considered in this stage. The posterior predictive distribution given the data is approximated using Markov Chain Monte Carlo (MCMC) algorithms. We use a blocking Gibbs sampling algorithm to simulate values from the posterior distribution of the parameters $\mathbf{\Theta}_Z$ (using WinBUGS). Our Gibbs sampling algorithm has three steps. We alternate between the coefficients for the weather covariates and the covariance parameters of the spatial-temporal process $e_z(\mathbf{s}, t)$ (Step 1), the parameters for the bias components of the FRM data (Step 2),

and the values of Z at all monitoring sites and time points (Step 3). After simulating N_1 values from the posterior distribution of the parameters $\mathbf{\Theta}_Z$, the estimator for the predictive distribution is

$$p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}) = \frac{1}{N_1} \sum_{n_1=1}^{N_1} p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}, \mathbf{\Theta}_Z^{(n_1)}), \tag{2.8}$$

where $\mathbf{\Theta}_Z^{(n_1)}$ is the n_1^{th} draw from the posterior distribution. These estimates are used as the inputs in the second stage.

2.3.2 Stage 2: Model for speciated $PM_{2.5}$

In this stage, we parameterize our statistical model for speciated $PM_{2.5}$ in terms of the true sum of the five components from the first stage and the relative proportion of each component to the total. The proportion of each component to the sum varies over space and time, and we use a hierarchical framework to account for the spatial-temporal associations of the proportions. Even though the spatial-temporal dependency structures of the proportions are considered, it could be insufficient to capture the spatial-temporal dependency and the cross dependency structures of speciated $PM_{2.5}$. We thus include a mean-zero spatial-temporal process in the model, which explains the dependency structures of speciated $PM_{2.5}$. This approach allows us to estimate both the speciated $PM_{2.5}$ in terms of the sum of the five components and the cross-covariance between $PM_{2.5}$ components. Figure 2.3 shows the framework for the speciated fine PM.

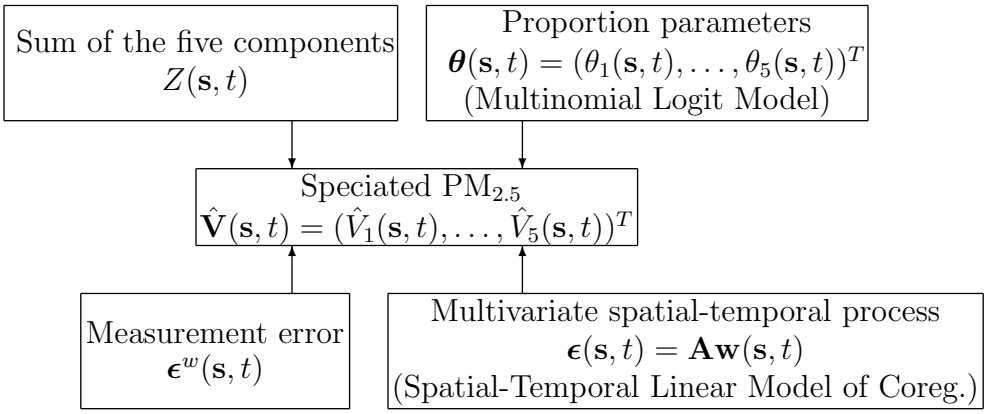


Figure 2.3: Model framework for speciated $PM_{2.5}$.

Let $\hat{\mathbf{V}}(\mathbf{s},t) = (\hat{V}_1(\mathbf{s},t),\ldots,\hat{V}_5(\mathbf{s},t))^T$ be a vector of the observed speciated $PM_{2.5}$ at location $\mathbf{s}$ and at time t from the STN. The parameter $\theta_k(\mathbf{s},t)$ denotes the proportion of the sum attributed to component k, and $Z(\mathbf{s},t)$ is the estimate for the sum of the five components from the first stage. Our model for speciated $PM_{2.5}$ is

$$\hat{V}_k(\mathbf{s},t) = \theta_k(\mathbf{s},t)Z(\mathbf{s},t) + \epsilon_k(\mathbf{s},t) + \epsilon_k^w(\mathbf{s},t), \qquad \text{for} \quad k = 1,\ldots,5, \tag{2.9}$$

where $\epsilon_k^w(\mathbf{s},t)$ is the pure measurement error process which is assumed to be normal and independent of $\epsilon_k(\mathbf{s},t)$. The spatial-temporal process $\boldsymbol{\epsilon}(\mathbf{s},t) = (\epsilon_1(\mathbf{s},t),\ldots,\epsilon_5(\mathbf{s},t))^T$ is assumed to be a Gaussian process with mean zero and a covariance matrix which changes with space and time.

To ensure that the proportions add to one at each site and time, we extend the multinomial logit model (McFadden, 1974) to the spatiotemporal setting. Let

$$\theta_k(\mathbf{s},t) = \frac{\exp(\delta_k(\mathbf{s},t))}{\sum_{j=1}^{5}\exp(\delta_j(\mathbf{s},t))}, \qquad \text{for} \quad k = 1,\ldots,5, \tag{2.10}$$

where $\delta_j(\mathbf{s},t)$ are independent across j. To guarantee that the multinomial logit model is identifiable, we fix $\delta_5(\mathbf{s},t)=0$ for all $\mathbf{s}$ and t. In our study, crustal material is taken to be the 5^{th} component because it is the most stable component. We use a dynamic hierarchical framework for $\delta_j(\mathbf{s},t)$ as

$$\begin{aligned}
\delta_k(\mathbf{s},t) &= \eta_k(t)+\gamma_k(\mathbf{s},t), \qquad \text{for } \ k=1,\ldots,4, && (2.11)\\
\eta_k(t) &= f_k(t)+e_{\eta_k}(t), && (2.12)\\
\gamma_k(\mathbf{s},t) &= \delta_{\gamma_k}\gamma_k(\mathbf{s},t-1)+e_{\gamma_k}(\mathbf{s},t). && (2.13)
\end{aligned}$$

The function $\eta_k(t)$ denotes the overall temporal trend of the k^{th} logit component, which is expressed by the smoothing function of time $f_k(t)$ to explain seasonality of the k^{th} logit component. The process e_{η_k} is assumed to be a white noise Gaussian process. We model the process $\gamma_k(\mathbf{s},t)$ using an AR(1) with coefficient δ_{γ_k} and we assume the process $e_{\gamma_k}(\cdot,t)$ is a Gaussian process with mean zero and a spatial covariance function to explain the spatial dependency structure. Thus, the process $\gamma_k(\mathbf{s},t)$ accounts for the spatial-temporal structure of the k^{th} logit component not explained by the overall temporal trend. In this study, based on exploratory analysis, the function $f_k(t)$ is assumed to be a linear combination of one sine and one cosine function with respect to one-year period. The spatial covariance of e_{γ_k} is assumed to be an exponential covariance $\sigma^2_{\gamma_k}\exp\left(-h_1/\phi_{\gamma_k}\right)$. We use a normal prior, N(0,0.1), for δ_{γ_k}. We use uniform priors, Unif$(0,100)$ and Unif$(1,1000)$, for σ_{γ_k} and ϕ_{γ_k}, respectively.

Spatial-temporal linear model of coregionalization (STLMC)

We account for spatial-temporal dependency and cross dependency of speciated $PM_{2.5}$ in the errors $\boldsymbol{\epsilon}(\mathbf{s},t) = (\epsilon_1(\mathbf{s},t),\ldots,\epsilon_5(\mathbf{s},t))^T$ by introducing a spatial-temporal linear model of coregionalization (STLMC). The STLMC is an extension of the linear model of coregionalization (LMC) used in multivariate spatial analysis (Grzebyk and Wackernagel, 1994; Wackernagel, 1998; Gelfand et al., 2004). The basic idea of the STLMC is that dependent spatial-temporal processes are expressed as linear combination of uncorrelated spatial-temporal processes. The STLMC provides a very rich class of multivariate spatial-temporal processes with simple specification and interpretation. The STLMC like the LMC could be used as a dimension reduction method, which means that the given multivariate processes are represented as lower dimensional processes. Recently, Schmidt and Gelfand (2003), Banerjee et al. (2004), and Gelfand et al. (2004) used the LMC approach to construct valid cross-covariance functions in the multivariate spatial modeling. In this study, we also use the STLMC to construct a valid cross-covariance function of multivariate spatial-temporal processes, i.e.,

$$\boldsymbol{\epsilon}(\mathbf{s},t) = \mathbf{A}\mathbf{w}(\mathbf{s},t), \tag{2.14}$$

where $\mathbf{w}^T(\mathbf{s},t) = (w_1(\mathbf{s},t),\ldots,w_5(\mathbf{s},t))$, and $\mathbf{A}$ is a 5×5 weight matrix explaining the association among the five variables. Without loss of generality, we assume $\mathbf{A}$ is a lower triangular matrix. For computational convenience, we adopt a simple approach to model the spatial-temporal process $\mathbf{w}(\mathbf{s},t)$. We assume that $w_i(\mathbf{s},t)$, $i = 1,\ldots,5$, are independent Gaussian spatial-temporal processes with mean zero and separable

spatial-temporal covariance,

$$\text{Cov}(w_i(\mathbf{s}_l, t_j), w_i(\mathbf{s}_{l'}, t_{j'})) = C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \boldsymbol{\phi}_i) C_i^{(2)}(t_j, t_{j'}; \boldsymbol{\psi}_i), \tag{2.15}$$

where $C_i^{(1)}$ is a spatial covariance with the parameter vector $\boldsymbol{\phi}_i$, and $C_i^{(2)}$ is a temporal autocovariance with the parameter vector $\boldsymbol{\psi}_i$. The STLMC in (2.14) implies $E(\boldsymbol{\epsilon}(\mathbf{s}, t)) = 0$ and

$$\text{Cov}(\boldsymbol{\epsilon}(\mathbf{s}_l, t_j), \boldsymbol{\epsilon}(\mathbf{s}_{l'}, t_{j'})) = \sum_{i=1}^{5} C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \boldsymbol{\phi}_i) C_i^{(2)}(t_j, t_{j'}; \boldsymbol{\psi}_i) \mathbf{T}_i, \tag{2.16}$$

where $\mathbf{T}_i = \mathbf{a}_i \mathbf{a}_i^T$ and $\mathbf{a}_i$ is the i^{th} column vector of $\mathbf{A}$. Under this model, the covariance matrix of $\boldsymbol{\epsilon}$ at any site $\mathbf{s}$ and time t is $\mathbf{T} = \sum_{i=1}^{5} \mathbf{T}_i$.

We form $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_5^T)^T$ and $\boldsymbol{\epsilon}_i^T = (\boldsymbol{\epsilon}_i^T(t_1), \ldots, \boldsymbol{\epsilon}_i^T(t_T))$ for $i = 1, \ldots, 5$, where $\boldsymbol{\epsilon}_i^T(t_j) = (\epsilon_i(s_1, t_j), \ldots, \epsilon_i(s_{N_s}, t_j))$ for $j = 1, \ldots, T$. Then, the covariance matrix of $\boldsymbol{\epsilon}$ is

$$\boldsymbol{\Sigma}^{\boldsymbol{\epsilon}} = \sum_{i=1}^{5} \mathbf{T}_i \otimes \mathbf{U}_i \otimes \mathbf{R}_i, \tag{2.17}$$

where $\otimes$ denotes the Kronecker product. Each $\mathbf{R}_i$ is a $N_s \times N_s$ matrix with $(R_i)_{ll'} = C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \boldsymbol{\phi}_i)$, which accounts for spatial associations. Each $\mathbf{U}_i$ is a $T \times T$ matrix with $(U_i)_{jj'} = C_i^{(2)}(t_j, t_{j'}; \boldsymbol{\psi}_i)$, which explains temporal associations. This covariance matrix, $\boldsymbol{\Sigma}^{\boldsymbol{\epsilon}}$, is nonseparable, except in the special case of the STLMC where $C_i^{(1)} = C_{i'}^{(1)} = C^{(1)}$ and $C_i^{(2)} = C_{i'}^{(2)} = C^{(2)}$ for all $i, i' = 1. \ldots, 5$. In this case, $\boldsymbol{\Sigma}^{\boldsymbol{\epsilon}} = \mathbf{T} \otimes \mathbf{U} \otimes \mathbf{R}$ for $(R)_{ll'} = C^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \boldsymbol{\phi})$ and $(U)_{jj'} = C^{(2)}(t_j, t_{j'}; \boldsymbol{\psi})$.

In our study, we use a stationary exponential covariance function $C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \sigma_i^2, \phi_i) = \sigma_i^2 \exp(-||\mathbf{s}_l - \mathbf{s}_{l'}||/\phi_i)$, $i = 1, \ldots, 5$ and the autocovariance function of the AR(1) $C_i^{(2)}(t_j, t_{j'}; \psi_i) = \psi_i^{|t_j - t_{j'}|}/(1 - \psi_i^2)$. We use uniform hyperpriors Unif(1,1000) and

Unif(0,100), for ϕ_i and σ_i, respectively. For the temporal parameter ψ_i, we use a normal hyperprior, N(0,0.1) (0.1 is the precision). Since the matrix $\mathbf{A}$ is a lower triangular matrix, we assign inverse gamma priors to the diagonal elements and normal priors to the off-diagonal elements.

Algorithm for Estimation and Prediction

We now discuss estimation and prediction of the speciated $PM_{2.5}$ using a Bayesian approach. In order to predict the speciated $PM_{2.5}$ at location $\mathbf{s}_0$ and at time t given the data $\hat{\mathbf{V}}$ and $\mathbf{Z}$ (all available speciated $PM_{2.5}$ data and estimates for the true sum of the five components from the first stage, respectively) and an estimate $Z(\mathbf{s}_0, t)$ from the first stage, we need the posterior predictive distribution of $\mathbf{V}(\mathbf{s}_0, t)$:

$$p(\mathbf{V}(\mathbf{s}_0,t)|\hat{\mathbf{V}},\mathbf{Z},Z(\mathbf{s}_0,t)) \propto \int p(\mathbf{V}(\mathbf{s}_0,t)|\hat{\mathbf{V}},\mathbf{Z},\boldsymbol{\Theta},Z(\mathbf{s}_0,t))p(\boldsymbol{\Theta}|\hat{\mathbf{V}},\mathbf{Z})d\boldsymbol{\Theta}, \quad (2.18)$$

where $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ is a collection of all of the unknown parameters in the second stage. The vector $\boldsymbol{\Theta}_1$ includes parameters used to model the vector of the proportions, $\boldsymbol{\theta}$, and the vector $\boldsymbol{\Theta}_2$ includes the STLMC parameters. We use a Blocking Gibbs Sampling algorithm to sample N_2 values from the posterior distribution of the parameter $\boldsymbol{\Theta}$ (within the software WinBUGS). Like the previous stage, our Gibbs sampling algorithm has three steps. We alternate between the proportion parameters $\boldsymbol{\Theta}_1$ given the data (Step 1), the covariance parameters $\boldsymbol{\Theta}_2$ given the data and the values of $\boldsymbol{\Theta}_1$ updated (Step 2), and the unobserved true values of $\mathbf{V}$ at all sites and time points (Step 3). We obtain the conditional posterior distribution of the parameters $\boldsymbol{\Theta}$ given the data updated in Step 3.

The conditional distribution of $\mathbf{V}$ at location $\mathbf{s}_0$ and time t is

$$p(\mathbf{V}(\mathbf{s}_0,t)|\hat{\mathbf{V}},\mathbf{Z},\boldsymbol{\Theta},Z(\mathbf{s}_0,t)) \sim N(\boldsymbol{\mu}(\mathbf{V}(\mathbf{s}_0,t)),\mathbf{T}-\Sigma_{12}{\Sigma_{\hat{\mathbf{V}}}}^{-1}(\Sigma_{12})^T), \tag{2.19}$$

where $\boldsymbol{\mu}(\mathbf{V}(\mathbf{s}_0,t)) = \boldsymbol{\theta}(\mathbf{s}_0,t)Z(\mathbf{s}_0,t) + \Sigma_{12}{\Sigma_{\hat{\mathbf{V}}}}^{-1}(\hat{\mathbf{V}}-\boldsymbol{\theta}\mathbf{Z})$, $\Sigma_{12} = \mathrm{Cov}(\mathbf{V}(\mathbf{s}_0,t),\hat{\mathbf{V}})$ is a $5 \times (5TN_s)$ matrix, and $\Sigma_{\hat{\mathbf{V}}} = \mathrm{Cov}(\hat{\mathbf{V}},\hat{\mathbf{V}})$. Using the Rao-Blackwellized estimator (Gelfand and Smith, 1990), the predictive distribution is approximated by

$$p(\mathbf{V}(\mathbf{s}_0,t)|\hat{\mathbf{V}},\mathbf{Z},Z(\mathbf{s}_0,t)) = \frac{1}{N_2}\sum_{n_2=1}^{N_2} p(\mathbf{V}(\mathbf{s}_0,t)|\hat{\mathbf{V}},\mathbf{Z},\boldsymbol{\Theta}^{(n_2)},Z(\mathbf{s}_0,t)), \tag{2.20}$$

where $\boldsymbol{\Theta}^{(n_2)}$ is the n_2^{th} draw from the posterior distribution for the parameters.

2.4 Application

We apply our statistical framework to the daily speciated $\mathrm{PM}_{2.5}$ data in the United States in 2004. In the first stage, we study the spatial-temporal patterns of the additive bias term of the FRM process. Preliminary exploratory analysis suggests that the coefficients of the weather covariates are different in different regions. Therefore we implement our framework for the nine geographic regions as defined by the United States Census: New England; Middle Atlantic; East North Central; Midwest; South Atlantic; East South Central; West South Central; Mountain; Pacific. We assume that the regression parameters $\boldsymbol{\beta}(\mathbf{s},t)$ vary across regions but are constant (over space) within these regions and have vague normal priors, $N(0,0.01^2)$ (0.01^2 is the precision).

In the second stage, we study the spatial-temporal association for each component as well as the associations among the components. Due to computational costs,

we implement our entire spatial-temporal framework for speciated $PM_{2.5}$ using only California data, but we work with our framework over the entire United States at fixed times (June 14 and December 14) and at fixed locations (Los Angeles, Phoenix, and New York City).

All MCMC sampling is carried out in the freely-available software WinBUGS (`http://www.mrc-bsu.cam.ac.uk/bugs/`). For all MCMC sequences, we conducted a MCMC convergence diagnosis using the Gelman and Rubin (1992) convergence diagnostics, autocorrelation functions, and trace plots. All the hyperpriors chosen here ensure acceptable MCMC convergence.

Figure 2.4 shows boxplots of the posterior mean of the monthly average of the additive bias term of the FRM, $a(\mathbf{s}, t)$, for July 2004 and December 2004 in nine geographic regions as defined by the U.S. Census Bureau. In July, the bias' posterior mean is positive for all nine geographic regions except for in the Pacific region. The mean is the largest in the South Atlantic region. The negative bias (the mean is -0.94) in the Pacific region is not surprising because in California during summer about $60 - 90\%$ of the nitrate is lost due to evaporation in the FRM's total $PM_{2.5}$ mass measurement (Frank, 2006). In December, the posterior mean bias is positive in all regions. Overall, the bias is higher in July than in December because the sulfate concentrations are high in the summer and the FRM total $PM_{2.5}$ mass includes a large amount of water during the summer (Frank, 2006).

Figure 2.5 shows maps of the posterior mean for the sulfate proportion and the nitrate proportion on June 14 and on December 14. Overall, the sulfate proportion is high in the eastern United States and the northwestern United States in June. The

sulfate proportion is higher than the nitrate proportion across the entire United States in June. The sulfate proportion is higher in June than in December. In contrast, the nitrate proportion is lower in June than in December.

Figure 2.6 and 2.7 present maps of the estimated concentrations of speciated $PM_{2.5}$ for the spatial analysis on June 14 and December 14, respectively. Overall, the sulfate concentration is high in the Eastern U.S. in June, and the nitrate concentration is high in Southern California in June. The sulfate concentration is higher than the nitrate concentration in June. However, sulfate concentrations decrease and nitrate concentrations increase in December. The nitrate concentration is high in the Western U.S. in December. TCM is the highest component among the components in June and December.

The time series plots of the estimated concentrations of speciated $PM_{2.5}$ in Los Angeles, Phoenix, and New York City are presented in Figure 2.8. These three cities have markedly different temporal patterns, illustrating the difficulty in modeling speciated $PM_{2.5}$ across the entire United States. For all three cities, ammonium and crustal concentrations are relatively constant over time. In Los Angeles, the most abundant components are sulfate in the summer and nitrate and TCM in the winter, and all of the components are high during March. Sulfate is also high in the New York City in the summer, but unlike Los Angeles the other components are fairly stable over time. TCM is the dominant component throughout the year in Phoenix.

Many of these patterns are also apparent in Figure 2.9's map of the estimated speciated $PM_{2.5}$ composition by region and by season in 2004. In this figure circle size corresponds to the sum of the five components, and we can clearly see their

Table 2.1: Posterior quantiles of the correlations between components, (i.e., $T_{ij}/\sqrt{T_{ii}T_{jj}}$) in California in 2004.

Parameters	2.5%	Mean	97.5%
Sulfate/Nitrate	-0.497	-0.151	-0.058
Sulfate/Ammonium	-0.131	0.448	0.607
Sulfate/TCM	-0.934	0.042	0.101
Sulfate/Crustal	-0.095	0.869	0.943
Nitrate/Ammonium	0.434	0.836	0.994
Nitrate/TCM	-0.095	0.038	0.286
Nitrate/Crustal	-0.741	-0.223	-0.104
Ammonium/TCM	-0.699	-0.558	-0.156
Ammonium/Crustal	-0.604	0.024	0.279
TCM/Crustal	-0.767	0.674	0.991

spatial-temporal pattern. In the Pacific region, the sum of the five components is high over all seasons. During the summer (April - September), the sum of the five components is high in the eastern United States, while during the winter (January - March), the sum of the five components is high in the western United States.

TCM has the highest proportion among the components over the entire U.S. Sulfate concentrations are highest during the summer in most of the Eastern U.S. because increased photochemical reactions in the atmosphere increase sulfate formation (Baumgardner et al., 1999). Nitrate concentrations are highest during the winter (January - March) because high ammonia availability, low temperature, and

high relative humidity favor ammonium nitrate condensation. On average, nitrate concentration during the winter is $2.98\mu g/m^3$ over the United States (vs. $1.70\mu g/m^3$ for the year 2004). Overall, ammonium concentrations are roughly constant over all seasons. TCM concentrations are high during the summer and fall because of high fire-related activities. Crustal material concentrations are high over the entire United States during the spring and summer because of low soil moisture and high wind speeds. In particular, the western United States and some of the eastern regions are also impacted by Asian dust, Saharan dust, and North African dust (Malm et al., 2004).

Table 2.1 summarizes the posterior correlations between components using the data from California in 2004. Several 95% posterior intervals do not cover zero. The strongest correlation (posterior mean 0.836) is between nitrate and ammonium. This relationship can also be seen in Los Angeles in Figure 2.8 as both of these components have strong peeks in March and October.

Finally, to illustrate the need for our hierarchical framework, we present model diagnostics using our entire spatial-temporal framework using only data from California in 2004. We use the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) to compare models. Deviance is defined as $D(V) = -2\log(f(\hat{V}|V))$. The DIC statistic is $DIC = \overline{D} + p_D$, where $\overline{D} = E(D(\mathbf{V})|\hat{V})$ measures fit and $p_D = \overline{D} - D(E(\mathbf{V}|\hat{V}))$, the effective number of parameters, measures complexity. Models with smaller DIC are preferred. In addition to DIC, we also compare models using the root mean square prediction error (RMSPE). In Table 2.2, we compare three models. Model 1 is the statistical framework proposed in Section 2.3. Model

Table 2.2: Model comparisons using DIC (Deviance Information Criterion) and RMSPE (Root Mean Squared Prediction Error).

Model	DIC (p_D)	RMSPE
Model 1	950 (3444)	0.075
Model 2	8235 (1334)	1.462
Model 3	15219 (10)	6.082

2 ignores the STLMC process $\mathbf{Aw}(\mathbf{s}, t)$. Model 3 removes both the STLMC and the hierarchical framework of the proportion parameters, i.e., the proportion parameters are constant over space and time. The DIC is 950 ($p_D = 3444$) for Model 1, 8235 ($p_D = 1334$) for Model 2, and 15219 ($p_D = 10$) for Model 3. The RMSPE value for Model 1 is 0.075, for Model 2 it is 1.462, and for Model 3 it is 6.082. Thus, our statistical framework has the lowest DIC and RMSPE values among the three models. These results confirm the need for the multivariate spatial-temporal model.

In addition, we did conduct a calibration analysis for the speciated $PM_{2.5}$ in Phoenix to test the performance of our framework. We selected Phoenix because it is one of the locations with most complete data. We randomly selected 30 observations in 2004, and we obtained 95% prediction intervals for the t^{th} time given the data, not using data from the t^{th} time we are predicting. Figure 2.10 plots the actual and predicted values. The percentages of the observed values that are outside the interval are between 0% and 3.3%. It appears that the model is well-calibrated. We also did calibration analyses for the other locations and the results were similar.

2.5 Conclusion

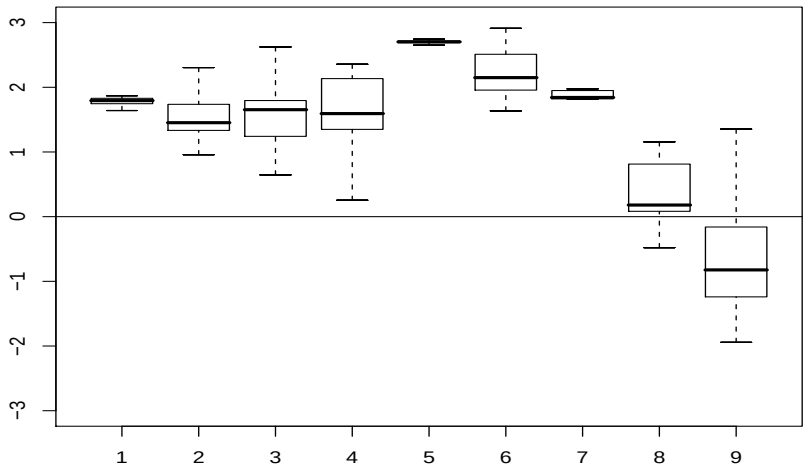
In this chapter we present a flexible hierarchical framework to study speciated $PM_{2.5}$. The multivariate spatial-temporal model proposed here allows for spatial-temporal dependency for each component and cross dependency structures among the components. A hierarchical framework provides a natural way to investigate the spatiotemporally varying contribution of each component to the sum. Using our framework, we can estimate speciated $PM_{2.5}$ at unobserved locations of interest in the United States. We also introduce a new statistical framework to incorporate $PM_{2.5}$ data from different sources, which takes into account bias over space and time. Diagnostics verify the performance of our model.

We find that the additive bias term of the FRM network is generally positive. That is, the FRM's total $PM_{2.5}$ mass is higher than the sum of the five main components measured by the STN. However, in the Pacific region, we see different results during the summer season because of nitrate losses. In the eastern United States, the contribution of sulfate to the sum tends to be higher during the summer. In almost all regions, sulfate concentrations are high during the summer. Also, the spatial differences in the sulfate concentrations are the largest during the summer. Nitrate concentrations are high during the winter, and they are also high in urban areas because of high nitrogen oxide (NO_x) emissions from automobiles. During the summer, nitrate concentrations are low over the entire United States. TCM concentrations explain most of total $PM_{2.5}$ mass. It is found that TCM concentrations are high in the summer and fall seasons. During the spring and summer seasons, crustal material

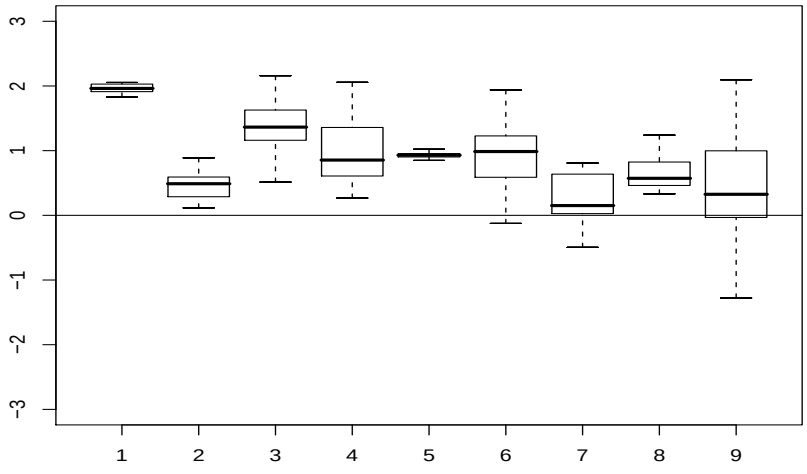
concentrations are high over the entire United States. Our results for the speciated $PM_{2.5}$ are consistent with previous analyses (Malm et al., 2004).

Our approach has some limitations. The computational burden prohibits a fully-Bayesian analysis. We employ a two-stage algorithm that first estimates the sum of the five components and then estimates the individual components. It is not clear how much uncertainty is ignored by this approach. Also, the spatial-temporal process for the true sum of the five components $Z(\mathbf{s}, t)$ and $\mathbf{w}(\mathbf{s}, t)$ in the STLMC could be modeled as a nonstationary and/or nonseparable spatial-temporal process. However, the computational burden is exacerbated in these cases so we use simple spatial-temporal models.

The multivariate spatial-temporal model could also be applied in other areas, such as meteorology, ecological modeling, and exposure analysis. The framework and results presented here will be essential for the health analysis.

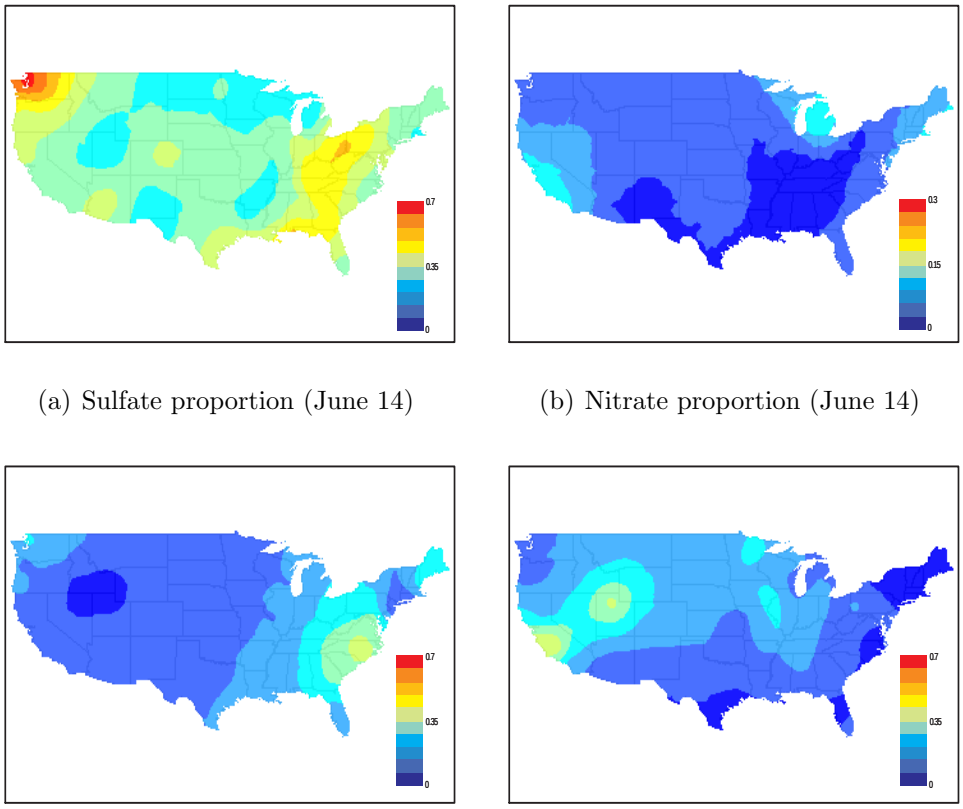


(a) July



(b) December

Figure 2.4: Boxplots of the posterior mean of the monthly average of the additive bias of the FRM, $a(\mathbf{s}, t)$, for (a) July 2004 and (b) December 2004 by geographic region (as defined by the U.S. Census). Region (1): Northeast (New England); (2): Northeast (Middle Atlantic); (3): Midwest (East North Central); (4): Midwest (West North Central); (5): South (South Atlantic); (6): South (East South Central); (7): South (West South Central); (8): West (Mountain); (9): West (Pacific).



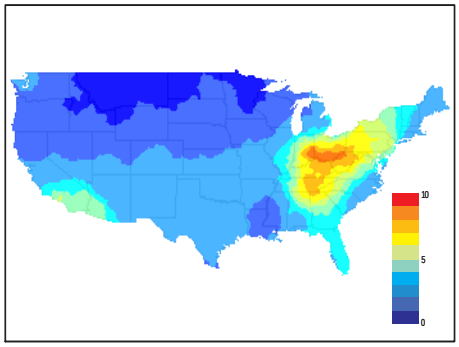
(a) Sulfate proportion (June 14)

(b) Nitrate proportion (June 14)

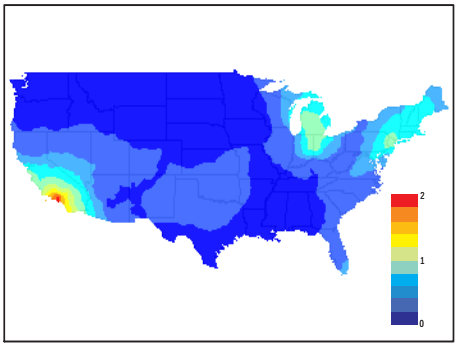
(c) Sulfate proportion (December 14)

(d) Nitrate proportion (December 14)

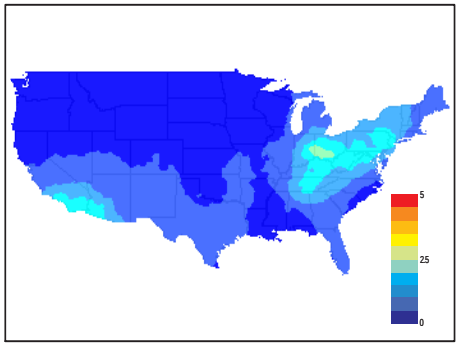
Figure 2.5: Maps of the posterior mean of the sulfate proportions to the sum of the five components and the nitrate proportions to the sum on June 14, 2004 and on December 14, 2004, respectively.



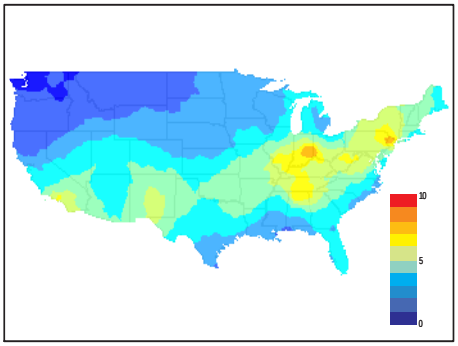
(a) Sulfate



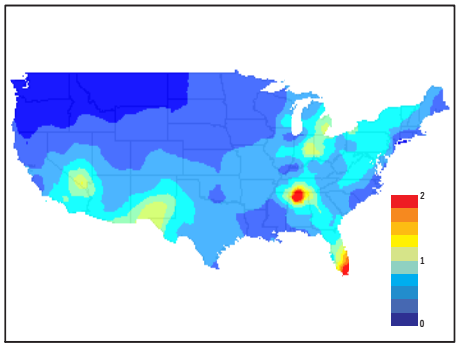
(b) Nitrate



(c) Ammonium

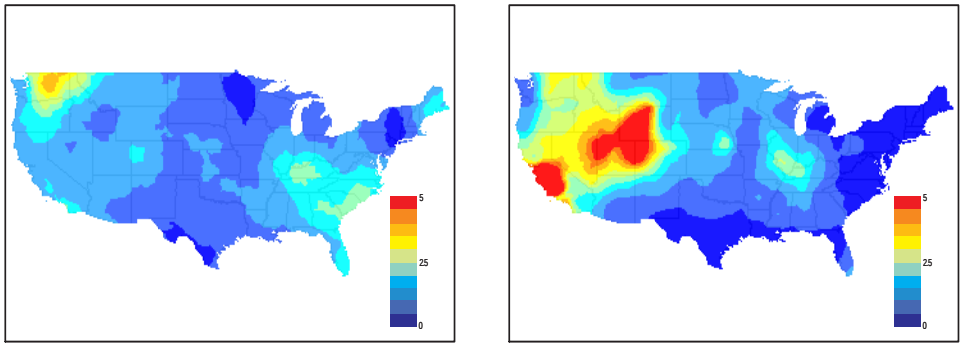


(d) TCM



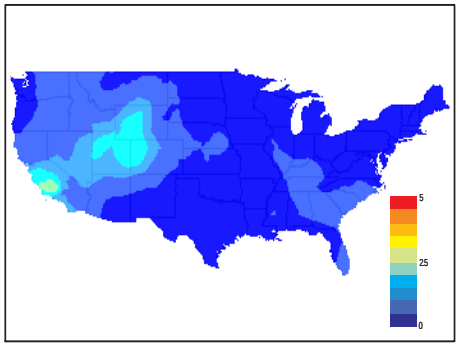
(e) Crustal

Figure 2.6: Maps of the posterior mean of speciated $PM_{2.5}$ ($\mu g/m^3$) on June 14, 2004.

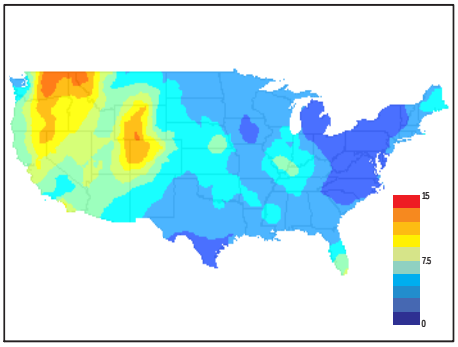


(a) Sulfate

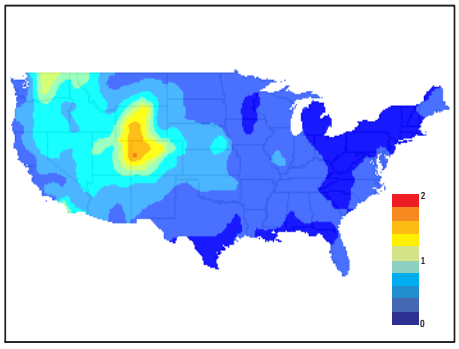
(b) Nitrate



(c) Ammonium



(d) TCM



(e) Crustal

Figure 2.7: Maps of the posterior mean of speciated $PM_{2.5}$ ($\mu g/m^3$) on December 14, 2004.

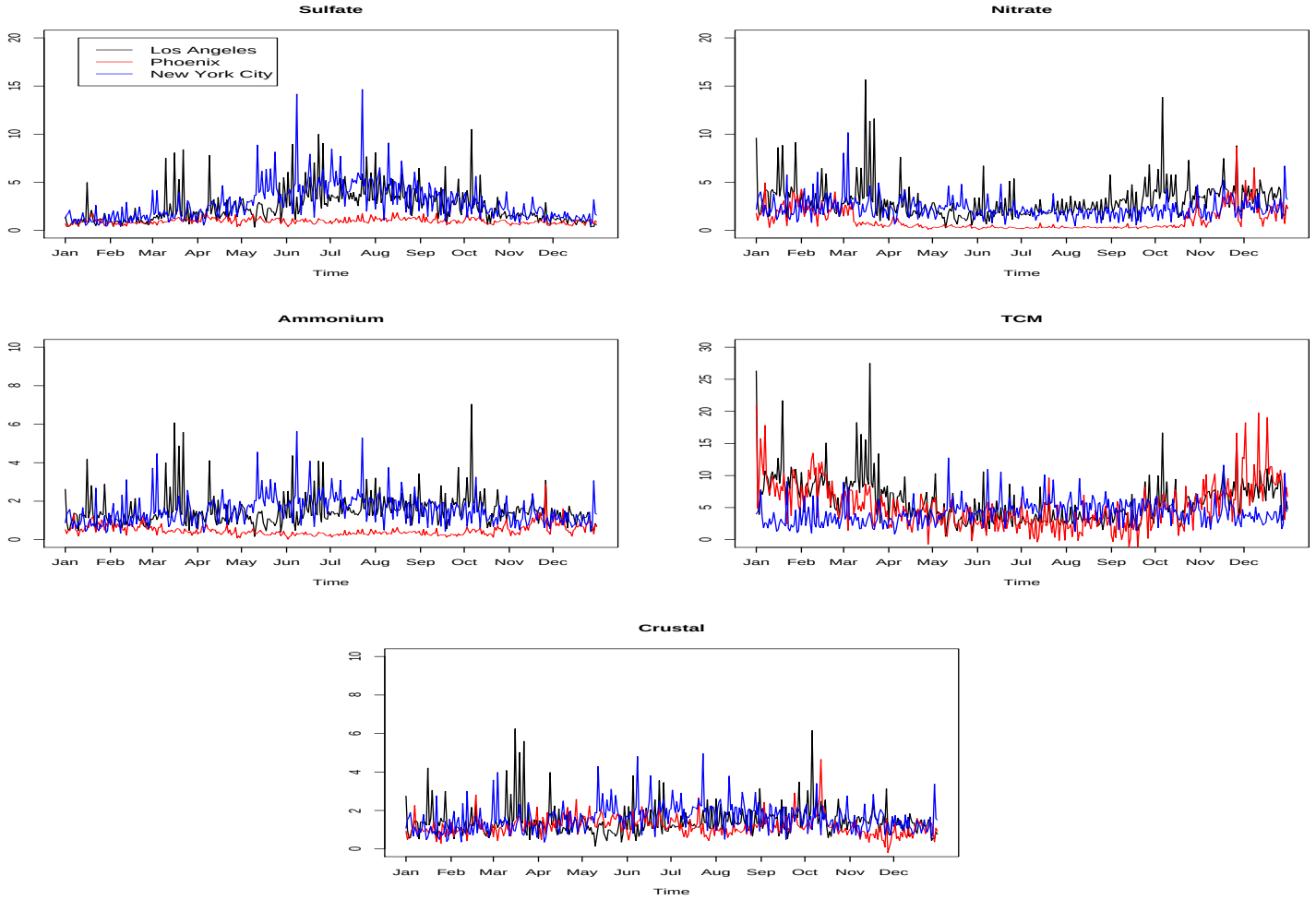
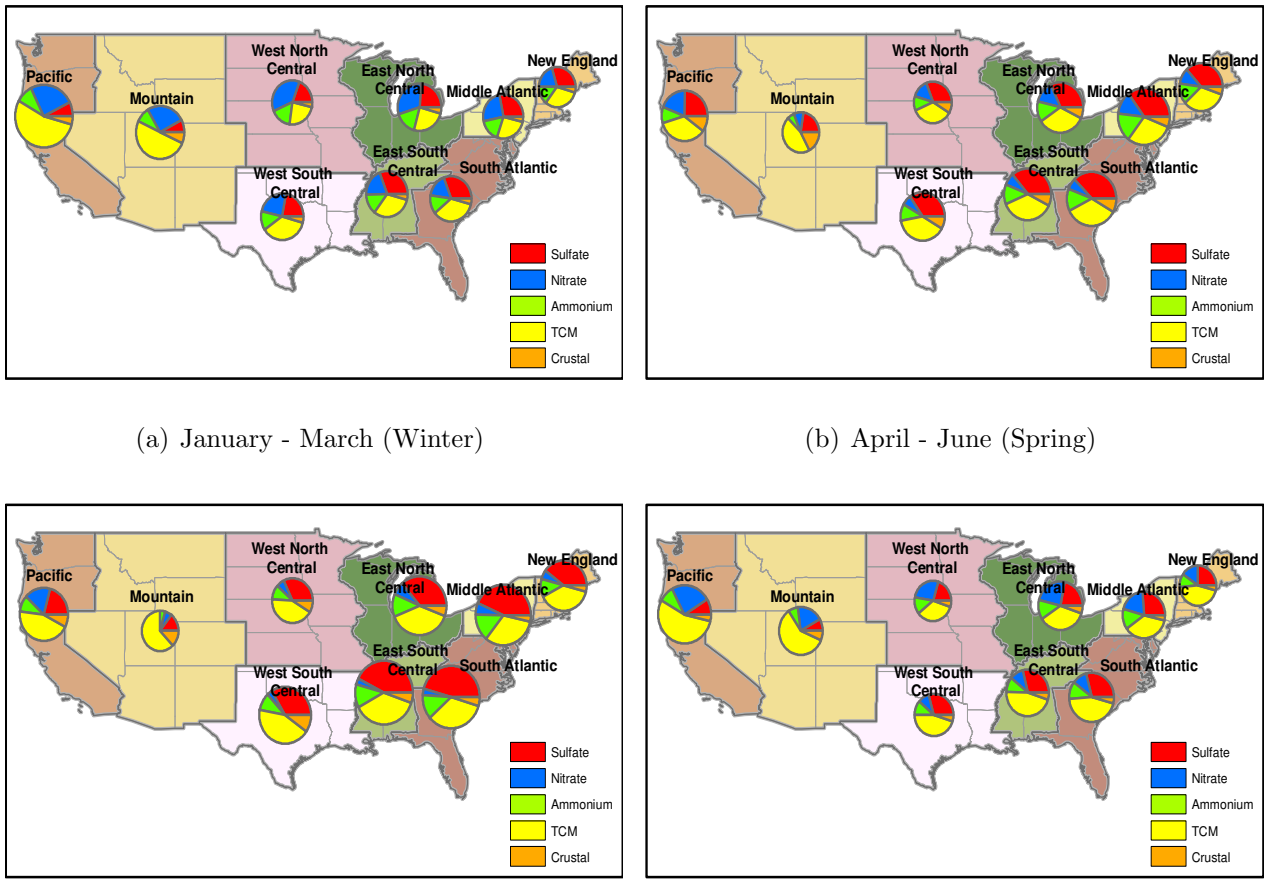


Figure 2.8: Time series plots of the estimated speciated $PM_{2.5}$ ($\mu g/m^3$) for three cities (Los Angeles, Phoenix, and New York City) in 2004.



(a) January - March (Winter)

(b) April - June (Spring)

(c) July - September (Summer)

(d) October - December (Fall)

Figure 2.9: Maps of the estimated speciated $PM_{2.5}$ composition by region and by season in 2004.

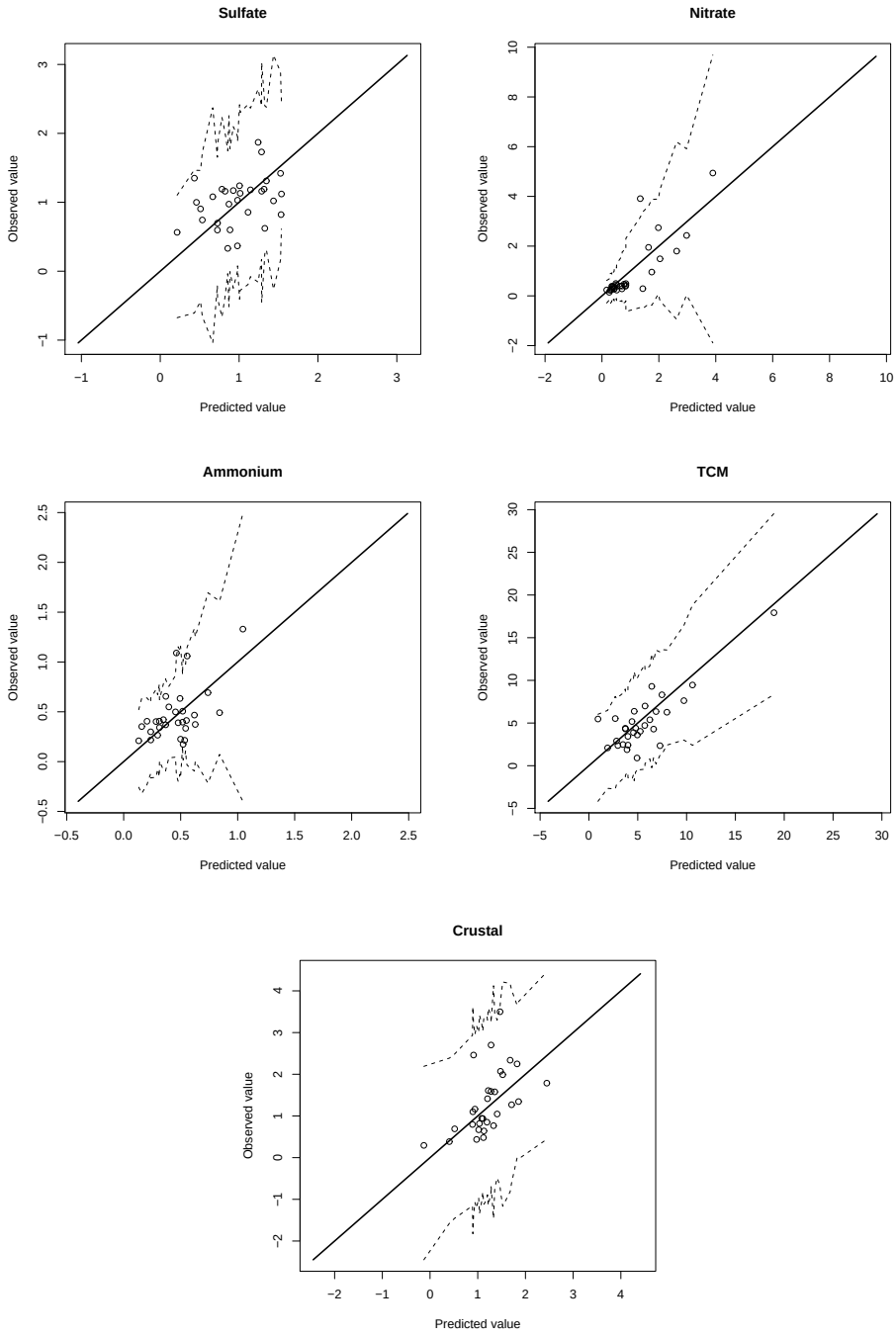


Figure 2.10: Model diagnostics for speciated $PM_{2.5}$ in Phoenix: STN values of each component versus the means of the predictive posterior distribution of the component at time t eliminating the t^{th} observation. The dotted lines show the 95% prediction intervals.

Chapter 3

Spatial-temporal association between fine particulate matter and daily mortality

3.1 Introduction

Many epidemiological studies have investigated the association between PM exposure and adverse human health outcomes (e.g., Dockery et al., 1992; Schwartz, 1994; American Thoracic Society and Bascom, 1996a,b; Dominici et al., 2000). However, the work by Smith et al. (2000) on fine particles, $PM_{2.5}$, provided evidence of lack of significant association between fine PM and mortality. This seems to suggest that more studies are needed to understand the effects of $PM_{2.5}$ on mortality.

Most of the previous analyses of PM health effects have been conducted in urban areas; very little is known about the rural PM-related health effects. One reason for this is that monitoring data are not only sparse across space but also time since most stations only measure $PM_{2.5}$ every third or sixth day. We overcome this limitation by supplementing monitoring data with atmospheric deterministic models (e.g., Community Multiscale Air Quality (CMAQ) models). CMAQ predicts air pollution levels at any given location and time. However, these numerical models could have a significant bias that needs to be quantified. Also, numerical models provide areal pollution estimates, rather than spatial point estimates. Thus, we have a change of support problem (see e.g., Gotway and Young, 2002), since monitoring data and numerical model output do not have the same spatial resolution. In this chapter, we develop a multi-stage spatial-temporal modeling approach which allows us to address these knowledge gaps, the change of support problem, and related uncertainties in assessing fine PM concentrations and health effects.

Recently, rigorous statistical time series modeling approaches have been used to better control for potential confounders in the epidemiological analysis of mortality associated with elevated ambient air pollutant levels. Furthermore, sophisticated analytical techniques have been introduced to adjust for seasonal trends in the data, culminating in the introduction of the generalized additive models (GAM). Although temporal trends can be explicitly included in the model, non parametric local smoothing methods (LOESS) based on GAM were widely used to take into account such trends in the analysis. Dominici et al. (2002b) suggested another approach using parametric natural cubic splines in the GAM model instead of the LOESS. One of

the main limitations of this type of time series modelling approach is that it is necessary to choose the time span in the LOESS smoothing process, or the degrees of freedom of the cubic splines, and the results can be very sensitive to how that is done. In our framework, we use an alternative approach which does not involve the selection of the number of basis functions or the degrees of freedom. We estimate the shape of time-varying confounders by introducing a stochastic search variable selection (SSVS) method (George and McCulloch, 1993) in a space-time context, while characterizing the spatial association of the time-varying confounders. SSVS was originally introduced for linear regression models and has been adopted for generalized linear models (George and McCulloch, 1997), log-linear models (Ntzoufras et al., 1997), and multivariate regression models (Brown et al., 1998). Smith and Kohn (1996) used Bayesian variable selection in a nonparametric regression model. The work presented here is the first attempt to extend Smith and Kohn's idea to model spatial-temporal data by randomly including/excluding basis functions from the model.

As presented in the previous chapter, $PM_{2.5}$ and its chemistry change with space and time, so the effects of $PM_{2.5}$ on mortality could change across space and time. Dominici et al. (2002a) showed that different cities have different relative risks of mortality due to PM exposure. Fuentes et al. (2006) smoothed the relative risk spatially. Lee and Shaddick (2007) smoothed the relative risk temporally. This is the first study to combine these two approaches. In our framework we allow the relative risk of mortality due to exposure to $PM_{2.5}$ to vary across space and time, taking into account spatial dependencies of the mortality data and the pollution data.

In this work we introduce an innovative hierarchical framework for spatial-temporal

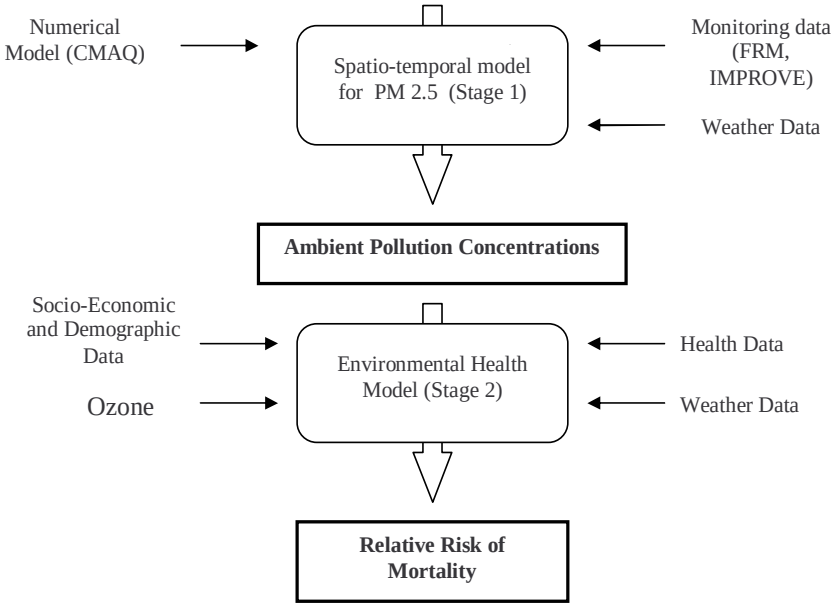


Figure 3.1: Hierarchical Bayesian framework to study the spatial and temporal association between fine particulate matter and mortality.

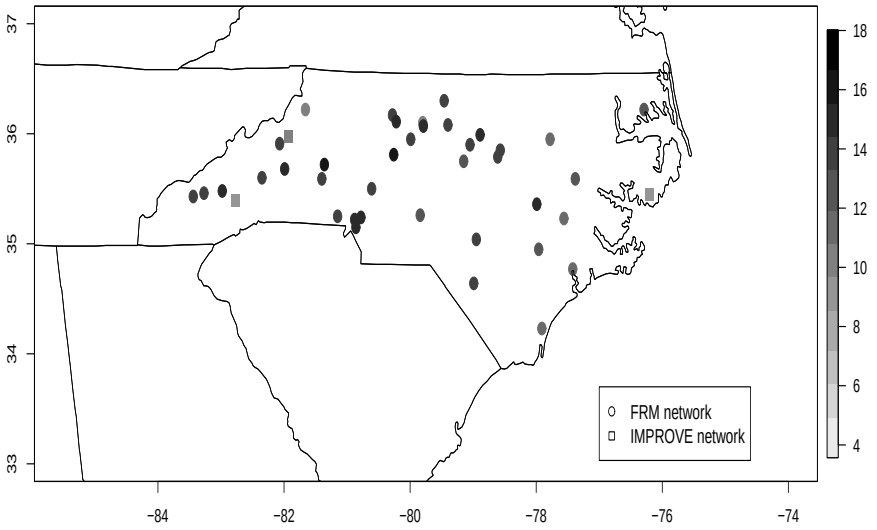
prediction and modelling of fine PM integrating atmospheric numerical model output with monitoring data, and we investigate the adverse health outcomes associated with population exposure to fine particulate matter (see Figure 3.1). We characterize geographic differences in the $PM_{2.5}$ health effects across the state of North Carolina for the year 2001. In the first stage we incorporate multi-source and multi-level information and knowledge (monitoring network [FRM, IMPROVE], air quality numerical model, meteorological data) about ambient environment into a flexible Bayesian space-time modeling framework for estimating ambient fine PM concentrations. These refined exposure indices of $PM_{2.5}$ mass from stage 1 are incorporated in a likelihood-based version of Poisson regression models (stage 2) to estimate the relative risks and to characterize the population susceptibility for $PM_{2.5}$ associated in-

creases in mortality. The hierarchical framework introduced here to combine different sources of spatial-temporal data, while characterizing uncertainty and bias associated to them, is adopted to obtain more reliable estimates of air pollution levels and to reduce the variability of the relative risk parameter, which explains the association between pollution and mortality. To the best of our knowledge, this is the first study to use numerical model output in studying the association between $PM_{2.5}$ and mortality. However, this framework is flexible enough that can be adopted and implemented in many other situations where we have spatial (or spatial-temporal) information from different sources.

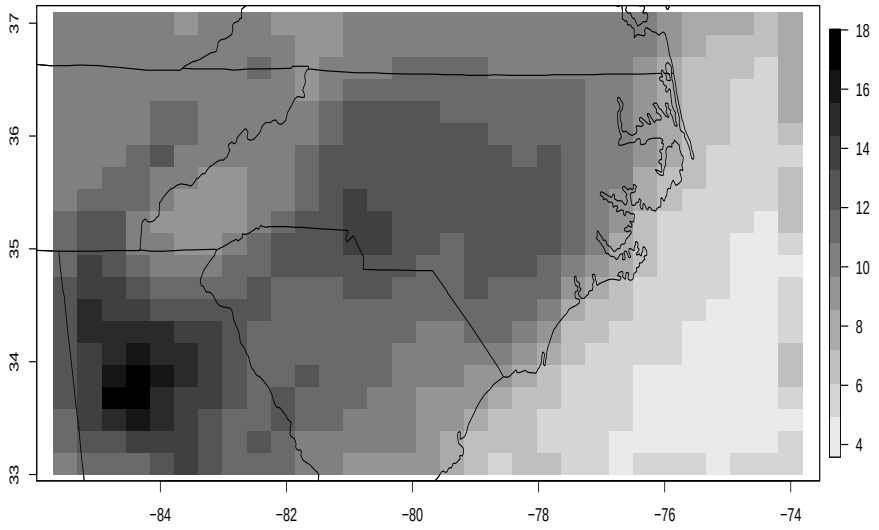
This chapter is organized as follows. In Section 3.2, we describe the different sources of data used in this study. In Section 3.3, we present a statistical framework to study the spatial-temporal association between $PM_{2.5}$ and mortality. We also describe a spatial-temporal model for $PM_{2.5}$. In Section 3.4, we present the results of this study. Finally, we provide a general discussion in Section 3.5.

3.2 Data Description

In this study we use the available $PM_{2.5}$ data in North Carolina for the year 2001. The data were provided by the U.S. EPA. The first source of $PM_{2.5}$ data has been obtained from the Federal Reference Method (FRM) monitoring network, which includes rural and urban sites and collects $PM_{2.5}$ samples either every day, every third day, or every sixth day. The second source of information for $PM_{2.5}$ is from the Interagency Monitoring of Protected Visual Environments (IMPROVE) network.



(a) FRM and IMPROVE networks



(b) CMAQ model

Figure 3.2: Yearly average of total $PM_{2.5}$ mass ($\mu g/m^3$) from (a) FRM network and IMPROVE network and (b) CMAQ model for 2001.

The IMPROVE network stations are located at national parks and wilderness areas, and they collect $PM_{2.5}$ samples either every day, every third day, or every sixth day. Figure 3.2 (a) presents the yearly average of total $PM_{2.5}$ mass ($\mu g/m^3$) at the 38 FRM monitoring stations and 3 IMPROVE monitoring stations in North Carolina for the year 2001.

Another important source of $PM_{2.5}$ over large areas can be obtained from three-dimensional (3-D) regional scale air quality models such as the U.S. EPA CMAQ modeling system (Binkowski and Roselle, 2003; Byun and Schere, 2006). CMAQ simulations over an airshed of interest provide gridded hourly concentrations and dry/wet deposition fluxes of major air pollutants such as $PM_{2.5}$. In this study we use CMAQ output from the surface layer. Figure 3.2 (b) presents the yearly average of CMAQ's total gridded $PM_{2.5}$ mass ($\mu g/m^3$) for the year 2001. The CMAQ resolution used in this study is 36km $\times$ 36km, and each CMAQ value represents the averaged pollution levels within each grid cell.

Ozone measurements (O_3) are monitored (on the hourly basis) through the State and Local Air Monitoring Stations (SLAMS), National Air Monitoring Stations (NAMS), and Clean Air Status and Trends Network (CASTNET). We have access to the SLAMS/NAMS measurements (`http://www.epa.gov/oar/oaqps/qa/monprog.html`) and CASTNET measurements (`http://www.epa.gov/castnet/`), and we use them to study the influence of the ozone as possible causative factor of adverse health effects. We determine the effects of ozone measurements and fine particles jointly.

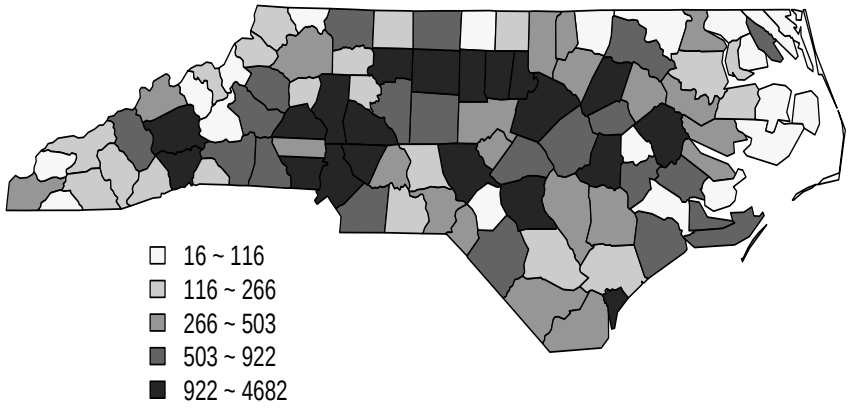
Daily meteorological data in North Carolina have been obtained from the U.S. National Climate Data Center. We use the following weather variables: minimum

temperature (℃), maximum temperature (℃), dew point temperature (℃), wind speed (m/s), and pressure (hPa).

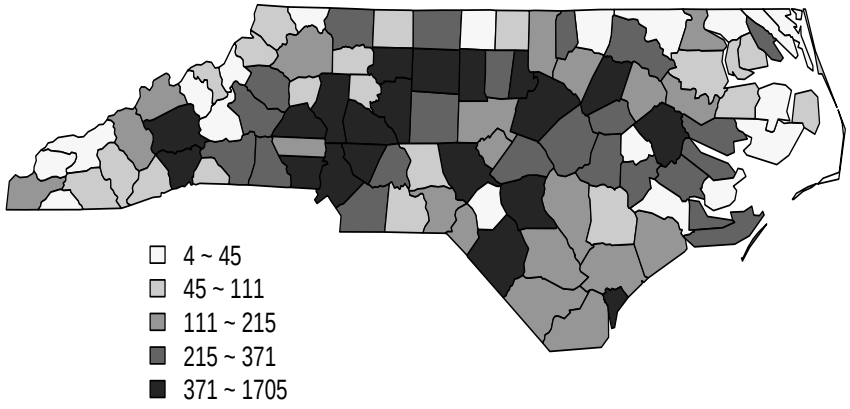
We obtained daily mortality data in North Carolina from the Odum Institute at the University of North Carolina (`http://www.irss.unc.edu`). These data include daily deaths from natural and cardiovascular causes by county in North Carolina for the year 2001. Figure 3.3 shows total counts of natural deaths and cardiovascular deaths over the entire state of North Carolina for 2001.

3.3 Statistical Models

Our hierarchical framework has two main stages (see flowchart in Figure 3.1). In the first stage we model and estimate the $PM_{2.5}$ concentrations, that are used in the health model proposed in stage 2. Fitting this complex hierarchical framework is done stage-by-stage, and we take the $PM_{2.5}$ estimates from the first stage as the inputs for the next stage. Within each stage we use a fully Bayesian approach to get the posterior distributions. As the implementation is based on the sequential version of the Bayesian theorem, the corresponding model uncertainties are captured at the final stage of our hierarchical model. Gelman (2004) has described the benefits of this type of directional Bayesian approach. This approach does not only offer computational benefits, but also in some settings, like the one presented here, the lack of an iteration between stages 1 and 2 might be desired. For example, we would not want the health data (Stage 2) help us to explain the pollution variables (Stage 1).



(a) Total counts of natural deaths



(b) Total counts of cardiovascular deaths

Figure 3.3: (a) Total counts of natural deaths in North Carolina for 2001 (b) Total counts of cardiovascular deaths in North Carolina for 2001.

3.3.1 Stage 1: Model for fine particulate matter

We introduce a spatial-temporal model for $PM_{2.5}$ using both observed data and numerical model output; this is an extension of the approach presented by Fuentes and Raftery (2005) in a purely spatial setting. We do not consider FRM measurements to be the "true" values because they are measured with error. Thus, we denote the observed total $PM_{2.5}$ mass at location $\mathbf{s} \in D_1$ on day $t \in D_2$ from the FRM network by $\widehat{Z}_F(\mathbf{s}, t)$, where $D_1 = \{\mathbf{s} : \mathbf{s}_1, \ldots, \mathbf{s}_{N_s}\} \subset \mathbb{R}^2$ and $D_2 = \{t : 1, \ldots, T\} \subset \mathbb{R}$, and it is modeled as

$$\widehat{Z}_F(\mathbf{s}, t) = Z(\mathbf{s}, t) + e_F(\mathbf{s}, t), \tag{3.1}$$

where $Z(\mathbf{s}, t)$ is the unobserved "true" underlying spatial-temporal process at location $\mathbf{s}$ on day t. The measurement error $e_F(\mathbf{s}, t) \sim N(0, \sigma_F^2)$ is assumed to be independent of the true underlying process.

We use a similar representation for the observed $PM_{2.5}$ measurements from the IMPROVE network, which is denoted by $\widehat{Z}_I$. We have

$$\widehat{Z}_I(\mathbf{s}, t) = Z(\mathbf{s}, t) + e_I(\mathbf{s}, t), \tag{3.2}$$

where $e_I(\mathbf{s}, t) \sim N(0, \sigma_I^2)$ is the measurement error and assumed to be independent of the processes $Z(\mathbf{s}, t)$ and $e_F(\mathbf{s}, t)$.

Since the CMAQ values are averages over grid squares, not point measurements, we model the $PM_{2.5}$ CMAQ values, $\widetilde{Z}(B_b, t)$, where subregions $B_1, \ldots, B_B$ cover the spatial domain B, as follows:

$$\widetilde{Z}(B_b, t) = a(B_b) + \frac{1}{|B_b|} \int_{B_b} Z(\mathbf{s}, t) d\mathbf{s} + e_N(B_b, t), \tag{3.3}$$

where $a(B_b)$ is the additive bias of the CMAQ output in subregion B_b and assumed to be a polynomial function of the centroid of the subregion, $\mathbf{s}_b$, with a vector of coefficients, $\mathbf{a_0}$. The process $e_N(B_b, t) \sim N(0, \sigma_N^2)$ accounts for the random deviation with respect to the underlying true process and is independent of $e_F(\mathbf{s}, t)$, $e_I(\mathbf{s}, t)$, and $Z(\mathbf{s}, t)$.

The true underlying process Z is modeled as a function of the weather covariates:

$$Z(\mathbf{s}, t) = \mathbf{M}^T(\mathbf{s}, t)\boldsymbol{\zeta} + e_z(\mathbf{s}, t), \tag{3.4}$$

where $\mathbf{M}(\mathbf{s}, t)$ is a vector of meteorological variables (minimum temperature, maximum temperature, dew point temperature, wind speed, and pressure) with a coefficient vector $\boldsymbol{\zeta}$. The weather information is obtained from weather stations, but they are not necessarily located at the same locations at which we have air pollution data, thus we have a spatial misalignment problem. To deal with this problem, we add in our hierarchical framework another level, stage 0, in which we use a spatial model for the weather variables for each day and predict these variables at the locations of interest for stages 1 and 2.

In order to predict $Z(\mathbf{s}_0, t_0)$, the true $PM_{2.5}$ value at space $\mathbf{s}_0$ and time t_0, given the data, $\widehat{Z} = (\widehat{Z}_F, \widehat{Z}_I, \widetilde{Z})$ and $\mathbf{M}$, we need the posterior predictive distribution of $Z(\mathbf{s}_0, t_0)$,

$$p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}) \propto \int p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}, \boldsymbol{\Theta}_Z) p(\boldsymbol{\Theta}_Z|\widehat{Z}, \mathbf{M}) d\boldsymbol{\Theta}_Z, \tag{3.5}$$

where $\boldsymbol{\Theta}_Z$ is a collection of all parameters considered in this stage. The posterior predictive distribution (3.5) given the data is approximated using Markov Chain Monte

Carlo (MCMC) algorithms. We use a blocking Gibbs sampling algorithm to simulate values from the posterior distribution of the parameters $\mathbf{\Theta}_Z$ (using WinBUGS). Like the algorithm in Section 2.3.1, our Gibbs sampling algorithm has three steps. We alternate between the coefficients for the weather covariates and the covariance parameters of the spatial-temporal process $e_z(\mathbf{s}, t)$ (Step 1), the parameters for the measurement errors and bias components of the $PM_{2.5}$ data (Step 2), and the values of Z at all monitoring sites (Step 3). The predictive distribution is obtained using the Rao-Blackwellized estimator (Gelfand and Smith, 1990)

$$p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}) = \frac{1}{N_1} \sum_{n_1=1}^{N_1} p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}, \mathbf{\Theta}_Z^{(n_1)}), \tag{3.6}$$

where $\mathbf{\Theta}_Z^{(n_1)}$ is the n_1^{th} draw from the posterior distribution.

The quantities of interest are the true total $PM_{2.5}$ mass averaged over a spatial domain C_j within county j on day t denoted by $Z_j(t)$,

$$Z_j(t) = \frac{1}{|C_j|} \int_{C_j} Z(\mathbf{s}, t) d\mathbf{s}. \tag{3.7}$$

The estimate of $Z_j(t)$ is obtained by averaging estimates of true $PM_{2.5}$ values at several locations randomly chosen within county j on day t. These estimates are used in the second stage.

Spatial priors

We use uniform priors, Unif(0,5), for σ_F and σ_I. We set these priors based on the information provided by EPA (U.S. EPA, 1997, `http://vista.cira.colostate.edu/improve/`) regarding the precision of the instrumentation used in these networks. Based on analysis of other similar datasets, we impose a uniform prior, Unif(0,5), for

σ_N. Based on exploratory analysis, $e_z(\cdot, t) = (e_z(\mathbf{s}_1, t), \ldots, e_z(\mathbf{s}_{N_s}, t))$ is normal with mean $\psi_z e_z(\cdot, t-1)$ and exponential covariance $\sigma_z^2 \exp(-h_1/\phi_z)$, where $h_1 = ||\mathbf{s} - \mathbf{s}'||$ (in km). We use a normal prior, N(0,0.1) (0.1 is the precision), for ψ_z and uniform priors, Unif(1,500) and Unif(0,100), for ϕ_z and σ_z, respectively.

3.3.2 Stage 2: Environmental Health Model

There are various statistical methods for modeling mortality data in the literature (e.g., Dominici et al., 2002a). The commonly-used model to study the association between air pollution and human health outcomes is a standard Poisson regression model with the independence assumption for the counts. However, an assumption of the Poisson model is that the mean and variance of the response variable are equal for each observation. This may be too restrictive. For example, the variance of the count data can be either smaller (under-dispersion) or larger (over-dispersion) than the mean. As described in Section 3.2, many counties in NC have very few counts at each day so a standard Poisson regression model might not be reasonable in this case.

We use a generalized Poisson regression model (Famoye, 1993; Fuentes et al., 2006) to characterize the potential over-dispersion or under-dispersion of the mortality data. Let $Y_j(t)$ be the number of natural deaths of county j for day t, for $j = 1, \ldots, J$ and $t = 1, \ldots, T$. We assume that $Y_j(t)$ follows a generalized Poisson distribution. The

probability function of a generalized Poisson distribution is defined by

$$f(y_j(t)) = Pr[Y_j(t) = y_j(t)] = \Big(\frac{\mu_j(t)}{1+\alpha\mu_j(t)}\Big)^{y_j(t)} \frac{\{1+\alpha y_j(t)\}^{y_j(t)-1}}{y_j(t)!} \times \exp\Big\{-\frac{\mu_j(t)(1+\alpha y_j(t))}{1+\alpha\mu_j(t)}\Big\}, \quad (3.8)$$

where $y_j(t) = 0, 1, \ldots, \alpha_-$, and $\alpha_- = -\alpha I(\alpha < 0) = -min(\alpha, 0)$. This distribution is denoted as $Y_j(t) \sim \text{GPoi}(\alpha, \mu_j(t))$, where α is a dispersion parameter, and $\mu_j(t) \geq 0$ is a mean parameter. The mean and variance are $E[Y_j(t)] = \mu_j(t)$ and $Var[Y_j(t)] = \mu_j(t)\{1+\alpha\mu_j(t)\}^2$, respectively. If $\alpha > 0$, then the model represents the over-dispersion ($Var[Y_j(t)] > E[Y_j(t)]$), and if $\alpha < 0$, then it represents the under-dispersion ($Var[Y_j(t)] < E[Y_j(t)]$). If $\alpha = 0$, then it becomes a standard Poisson distribution. Based on a generalized Poisson distribution for mortality, we develop a hierarchical regression model for estimating the spatial-temporal association between $PM_{2.5}$ and mortality.

As discussed in Chapter 1, one of important issues when investigating the association between $PM_{2.5}$ and daily mortality is "harvesting hypothesis" (or mortality displacement). We introduce a space-time model to estimate the association between $PM_{2.5}$ and mortality that is resistant to short-term harvesting. The method is a spatial adaption of the approach by Dominici et al. (2003) in a purely temporal context, and it is based on the assumption that harvesting alone creates associations only at shorter time scales. We use a spectral approach for the log-linear regression to decompose the information about the pollution-mortality association into distinct time scales taking into account the spatial dependency structure of the mortality and pollution data, and our relative risk estimates are harvesting-resistant because we ex-

clude the short-term information that is affected by harvesting. Thus, we decompose the daily time series of $PM_{2.5}$ estimates for county j, $Z_j(t)$, into L orthogonal different timescales components, $Z_{j1}(t), \ldots, Z_{jL}(t)$, using a discrete Fourier transform method (see Appendix A).

The effect of each orthogonal decomposition of the $PM_{2.5}$ time series is allowed to vary by county and by season. The index k refers to the seasons; we set $k = 1$ for the winter season (January-March), $k = 2$ for the spring season (April-June), $k = 3$ for the summer season (July-September), and $k = 4$ for the fall season (October-December). The parameter β_{jlk} represents the effect of air pollution for county j on timescale l for season k; the log relative risk (RR) parameter is defined as $\beta_{jlk} * 10^3$. We assume

$$\begin{aligned} Y_j(t) &\sim \text{GPoi}(\alpha, \mu_j(t)), \\ \log(\mu_j(t)) &= \gamma_j + \sum_{l=1}^{L} \beta_{jlk} Z_{jl}(t) + f_j(t) + O_j(t)\gamma_o \\ &\quad + S_1(temp_j(t), df_1) + S_2(dew_j(t), df_2) + S_3(wind_j(t), df_3). \end{aligned} \tag{3.9}$$

The function $f_j(t)$ adjusts for the seasonality of mortality, which varies with county j. In addition to the orthogonal $PM_{2.5}$ predictions, we also consider the co-pollutant $O_j(t)$, the daily ozone concentration for county j and day t, imputed using a similar spatial-temporal model as in Section 3.3.1 (see Appendix B). The S_i's are smooth functions of the weather covariates (temperature, dew point temperature, and wind speed) with the degrees of freedom (df) per year (df_i's). These weather variables are important to affect health outcomes.

Confounders

We consider the following confounders: age, gender, race, and hispanic/non-hispanic. Each confounder is treated as a categorical variable in our health model. We study the potential impact of these confounders on the RR by allowing an interaction between our estimated $PM_{2.5}$ component and the different confounders. In this study the groups for each confounder are:

- Age: $0-14$ years old (children), $15-64$ (adults), ≥ 65 (senior adults).
- Gender: male, female.
- Race: white, black, American Indian, Other.
- Hispanic: Non-hispanic, hispanic.

Spatial priors

Since the number of deaths for each county may depend on its population size, we assume that the intercept parameter γ_j is a spatial random effect, representing the baseline log relative risk of mortality for each county j. We use a conditional autoregressive (CAR) prior (Besag et al., 1991) for $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^T$,

$$\boldsymbol{\gamma} \sim N(\mu_{\boldsymbol{\gamma}}, \sigma^2_{\boldsymbol{\gamma}}(\mathbf{B}_+ - \rho\mathbf{B})^{-1}), \tag{3.10}$$

where $\sigma^2_{\boldsymbol{\gamma}}$ is the overall variance parameter and ρ is the spatial association parameter. The matrix $\mathbf{B} = (B_{jj'})$ includes the neighboring information, where $B_{jj'} = 1$ if county j is adjacent to county j', and $B_{jj'} = 0$ otherwise. The matrix $\mathbf{B}_+$ is a $J \times J$ diagonal matrix with elements $m_j = \sum_{j'} B_{jj'}$, $j = 1, \ldots, J$. Thus, m_j is the number of "neighbors"(adjacent counties) of county j. The mean parameter $\mu_{\boldsymbol{\gamma}}$ has

a normal prior, $N(0, 0.01)$ (0.01 is the precision). The parameter $\sigma^2_{\boldsymbol{\gamma}}$ has an inverse gamma prior, IG(0.5,0.0005), as recommended by Kelsall and Wakefield (1999), and the parameter ρ has a uniform prior with bounds which are determined in order to guarantee that the variance matrix of $\boldsymbol{\gamma}$ is symmetric positive definite (Banerjee et al., 2004).

To account for the spatial and temporal similarity of the effect of $PM_{2.5}$ for each timescale l, the multivariate CAR prior for $\boldsymbol{\beta}_l = (\boldsymbol{\beta}_{1l}, \ldots, \boldsymbol{\beta}_{Jl})^T$ with $\boldsymbol{\beta}_{jl} = (\beta_{jl1}, \ldots, \beta_{jl4})^T$ would be proper. Jin et al. (2007) introduced a general approach for multivariate spatial modelling, offering different alternatives to model the prior process for $\boldsymbol{\beta}_l$. In this study, we use a particular case of a multivariate CAR, called a multivariate intrinsic autoregressive (MIAR) prior (Gelfand and Vounatsou, 2002), that corresponds to a relatively smooth spatial process ($\rho = 1$),

$$\boldsymbol{\beta}_{jl} | \boldsymbol{\beta}_{j'l \ j' \neq j}, \sim N\Big(\frac{1}{m_j} \sum_{j' \neq j} B_{jj'} \boldsymbol{\beta}_{j'l}, \frac{1}{m_j} \Sigma_{\boldsymbol{\beta}_l}\Big), \tag{3.11}$$

where the positive definite 4×4 matrix $\Sigma_{\boldsymbol{\beta}_l}$ accounts for the conditional variability as well as cross-covariance relationships between the different seasons given the neighboring sites for each time scale l. Even though the MIAR is improper, the posterior will be proper under some regulatory conditions (see e.g., Sun et al., 1999). To guarantee that this prior is identified, we include 4 centering constraints $\sum_j \beta_{jlk} = 0$ for $k = 1, \ldots, 4$ to identify separate intercept terms in the model. We use a Wishart prior for the 4×4 precision matrix of the MIAR, Wishart$((0.01 I_4)^{-1}, 4)$, where I_4 is the identity matrix of size 4.

Seasonality of mortality

Selecting the number of basis functions for adjusting for the seasonal trend of mortality is always problematic. Here, we propose an approach that avoids fixing the number of basis functions. We write the seasonal trend for county j, $f_j(t)$, using a Fourier basis (same for all counties), $C_q(t)$, $q = 1, \ldots, Q$,

$$f_j(t) = \sum_{q=1}^{Q} c_{jq} C_q(t), \tag{3.12}$$

where Q is the number of basis functions and c_{jq}'s are unknown regression parameters that control the shape of the seasonal trend at each county j. Instead of selecting the number of basis functions, we assume that Q is large enough to capture the true model, and we use a Bayesian variable selection technique to stochastically include/exclude terms from the seasonal trend. We introduce a binary variable, w_{jq}, and a continuous spatial variable, r_{jq}, and express c_{jq} as

$$\begin{aligned} c_{jq}|w_{jq}, r_{jq} &= w_{jq} r_{jq}, \\ w_{jq} &\sim Bernoulli(0.5), \end{aligned}$$

where the vectors of coefficients $\mathbf{r}_q = (r_{1q}, \ldots, r_{Jq})$, for $q = 1, \ldots, Q$, follow independent CAR priors. If $w_{jq} = 0$, then $c_{jq} = 0$, and the corresponding basis function is not included in the model. If $w_{jq} = 1$, then $c_{jq} = r_{jq}$, and c_{jq} is non-zero, thus, the corresponding basis function is included in the model. We summarize the model complexity using the posterior of $W_j = \sum_{q=1}^{Q} w_{jq}$, which is the number of basis functions included in the model for county j.

3.4 Application

We apply our statistical framework to data in North Carolina for the year 2001 to study the spatial-temporal association between exposure to $PM_{2.5}$ and daily natural and cardiovascular deaths. We compare seasonal patterns in the effects of $PM_{2.5}$ and its different timescales on mortality. We study the effects of ozone on mortality. Here, we decompose the daily time series of $PM_{2.5}$ into five orthogonal components: < 3.5 days, $3.5 - 6$ days, $7 - 13$ days, $14 - 29$ days, and ≥ 30 days (Dominici et al., 2003).

The prior distributions of the spatial models in stages 1 and 2 are described in Sections 3.3.1 and 3.3.2. In the mortality model, we use natural cubic splines for the smooth functions S_i's with B-spline basis functions (Eilers and Marx, 1996). To select the degrees of freedom (df_i's), we considered up to 10 df per year for each smooth function. This value seemed to be large enough based on preliminary analysis. We found that 6 df per year for temperature and 3 df per year for dew point temperature and wind speed seemed appropriate using the deviance information criterion (DIC) of Spiegelhalter et al. (2002). Since we use 1-year data, we set the number of basis functions $Q = 30$. We use normal distribution, N(0,0.01), for the hyperpriors of the polynomial coefficients. We obtain the results using WinBUGS (`http://www.mrc-bsu.cam.ac.uk/bugs`) and R (`http://www.r-project.org/`). For all MCMC sequences, we conducted a MCMC convergence diagnosis using the Gelman and Rubin (1992) convergence diagnostics, autocorrelation functions, and trace plots.

Figure 3.4 maps the posterior mean of the monthly average of the $PM_{2.5}$ concentrations for January 2001 and August 2001. The estimated $PM_{2.5}$ values in January

and August were the highest in the central part of NC. Overall, the estimated $PM_{2.5}$ concentrations in January were lower than in August. On average, the $PM_{2.5}$ concentration was $13.76\mu g/m^3$ for January and $16.26\mu g/m^3$ for August.

Figure 3.5 (a) presents the time series of the estimated $PM_{2.5}$ and its different timescales for Wake County. As expected, the plots of the short-term timescales vary rapidly from day to day, while the time series plots for the long-term timescales are fairly smooth. The $PM_{2.5}$ value for each day is the same as the value obtained by adding the values of the five timescales for that day. Figure 3.5 (b) shows the daily time series of mortality (total and cardiovascular disease), ozone, temperature, dew point temperature, and wind speed for Wake County.

The estimated RRs at different timescales for 4 counties are presented in Figure 3.6. We found that the estimated RR values at longer timescale variations (>= 14 days) are larger than those at shorter timescale variations (< 14 days) in winter and summer, with few exceptions. The standard deviation (SD) of the RR is the highest for the longest timescale (>= 30 days), due to the potential correlation with the seasonal trend term. We also obtained estimated RR values of current day mortality using nondecomposed $PM_{2.5}$ time series. The effects of $PM_{2.5}$ on mortality in the winter and the summer seem to be similar. The RR values of mortality by season for Wake County are summarized in Table 3.1. For all seasons, the RR at timescales greater than 1 month was larger than those at timescales less than 3.5 days. The effect of $PM_{2.5}$ on current day mortality in the spring was the smallest among all seasons, while the effect in the winter was the largest.

We also studied the RR parameter of cardiovascular mortality by season. Table 3.2

Table 3.1: Posterior mean (SD) of log relative rates of mortality (percent increase in mortality per increase of $10\mu g/m^3$ of $PM_{2.5}$ concentrations) for Wake County by season.

	Winter	Spring	Summer	Fall
≥ 30	18.0 (15.3)	6.5 (21.1)	33.8 (14.1)	9.9 (13.7)
$14-29$	17.8 (13.9)	0.9 (12.2)	6.9 (9.8)	-13.3 (7.7)
$7-13$	1.0 (7.4)	-2.7 (8.0)	6.3 (10.3)	10.1 (7.6)
$3.5-6$	1.6 (6.6)	1.2 (8.5)	8.0 (8.7)	-7.4 (7.6)
< 3.5	4.4 (6.5)	-3.2 (7.8)	3.3 (11.4)	-3.1 (8.2)
overall	6.5 (5.5)	0.3 (6.1)	5.1 (3.5)	3.5 (5.4)

presents the estimated RRs of cardiovascular mortality by season for Wake County. We found a similar pattern for all seasons, greater effects at timescales greater than 1 month than at timescales less than 3.5 days. The spatial pattern of the RR for cardiovascular mortality due to $PM_{2.5}$ exposure was similar to that of the RR for natural mortality.

We studied the impact of ozone on the association between $PM_{2.5}$ and mortality. Figure 3.7 shows the differences in the RR parameter in the winter and the summer when ozone is included in the model and when ozone is not included. The differences in the RR in the summer seemed to be higher than those in the winter. However, the differences were small relative to the SD of the RR parameter so ozone did not have a significant effect. The 95% posterior interval for the parameter γ_o was $(-0.0009, 0.0061)$.

Table 3.2: Posterior mean (SD) of log relative rates of cardiovascular mortality (percent increase in mortality per increase of $10\mu g/m^3$ of $PM_{2.5}$ concentrations) for Wake County by season.

	Winter	Spring	Summer	Fall
≥ 30	9.6 (20.9)	2.6 (11.0)	12.3 (13.6)	8.4 (17.8)
$14-29$	0.7 (20.1)	3.0 (19.0)	3.1 (12.0)	-2.9 (13.1)
$7-13$	-2.3 (12.0)	1.4 (3.2)	7.6 (12.0)	27.3 (15.9)
$3.5-6$	0.9 (8.0)	0.6 (12.7)	-2.5 (13.2)	-3.1 (14.0)
< 3.5	-0.6 (5.9)	1.2 (13.6)	-3.4 (10.8)	-4.5 (14.0)
overall	0.7 (4.9)	-0.2 (5.8)	4.9 (6.3)	8.3 (9.5)

We examined the model complexity using the estimated W_j for each county j. This index is based on the adjustment for the seasonal trend of mortality. As shown in Figure 3.8 the posterior mean of the number of basis functions varied considerably by county. On average, the estimated number of basis functions included in the model across all counties was 10, and its SD was 2.3.

The estimated values for the intercept in the model are presented in Figure 3.9. As expected, higher estimated values for the intercept γ_j were obtained in counties with larger population sizes.

None of the confounders appeared to have a significant impact on the RR. The interaction term between the estimated $PM_{2.5}$ for the 5 timescales and the confounders was not significant across space. We conducted another study to examine the significance of the interaction term between same-day $PM_{2.5}$ exposure and the confounders,

Table 3.3: Model comparisons using DIC (Deviance Information Criterion) and RMSPE (Root Mean Squared Prediction Error).

Model	DIC (p_D)	RMSPE
Model 1	96327 (1038)	2.749
Model 2	96974 (1009)	2.781
Model 3		6.998

and it was not significant either.

CMAQ

In order to examine the contribution of CMAQ to the relative risk, we repeated the analysis without the CMAQ output for $PM_{2.5}$. The posterior means of the RR parameter when the CMAQ output was not used in our model were similar to those from the full model (Figure 3.10 (a)). However, Figure 3.10 (b) shows that including the CMAQ output substantially reduces the posterior SDs of the RR. Thus, it seems that including the numerical model output improves our estimate of the effect of $PM_{2.5}$ on mortality.

Model Diagnostics and Calibration

In our generalized Poisson model, the posterior mean of the dispersion parameter α was 0.049, and the 95% posterior interval was $(0.040, 0.057)$. This provides some evidence that the data might overdispersed and that a generalized Poisson model is needed in this application. Table 3.3 compares three different statistical models using the DIC and the root mean squared prediction error (RMSPE). The RMSPE

is defined as $\sqrt{\frac{1}{N}\sum_1^N (O_i - P_i)^2}$, where O_i are the observed mortality values for each county, and P_i are the predicted mortality values (using the mean of the predictive posterior distribution). The DIC for our full model (Model 1) was 96327 ($p_D = 1038$) and the DIC for the model with a constant RR across space (Model 2) was 96974 ($p_D = 1009$). The RMSPE value was also smaller for the full model (2.749) compared to the model with a constant RR across space (2.781). This justifies the need of a model that allows for spatial temporal variation in the RR, even within the relatively small geographic domain of this study. In addition, we considered a generalized linear model (GLM) in order to assess the need for our more complex Bayesian space-time framework. We fit a traditional GLM with a Poisson model for the number of deaths (Model 3) and allowed the regression coefficients to be independent over space and time. The RMSPE value of this model was 6.998. The fact that the RMSPE was almost 3 times the value obtained using our space-time model justifies the importance and relevance of taking into consideration the spatial-temporal structure of the data and uncertainties associated to them.

In addition, we did calibration analysis. In Figure 3.11, we present at a couple of randomly selected counties (Catawba and Durham) calibration plots for the mortality analysis during the summer and the fall seasons. The percentage of the observed values that are outside the interval is 10% for the summer and 6% for the fall. Similar results were obtained at other locations. We conclude our model is well calibrated.

We conducted sensitivity analysis to study the sensitivity of the estimated RR with respect to degrees of freedom used to explain the role of the weather variables. We fit several models using 3 and 9 dfs per year for temperature and using 6 and 9

Table 3.4: DIC (Deviance Information Criterion) for the selection of the degrees of freedom (df_i's) in the smooth functions of the weather variables (temperature (temp), dew point temperature (dew), and wind speed (wind)).

		df for wind		
df for temp	df for dew	3	6	9
3	3	96694 (1037)	96399 (1039)	96622 (1043)
	6	96802 (1040)	96723 (1045)	96793 (1051)
	9	96631 (1049)	96867 (1055)	96865 (1059)
6	3	96327 (1038)	96723 (1041)	96613 (1042)
	6	96471 (1043)	96606 (1047)	96820 (1056)
	9	96345 (1045)	96714 (1053)	96694 (1054)
9	3	96740 (1050)	96876 (1052)	96748 (1057)
	6	96632 (1058)	96642 (1053)	96766 (1058)
	9	96820 (1061)	96696 (1057)	96892 (1068)

dfs per year for the dew point temperature and wind speed. When we fit each model, we used the same functions for the other weather variables. Figure 3.12 shows the sensitivity of the RR by using different degrees of freedom per year in the smooth functions of the weather variables for Wake County in winter. The effects at the shorter timescales were similar in all cases, while the effects at the longer timescales were slightly different. Overall, there was not a significant impact on the RR by using different dfs per year. In table 3.4, we present the DIC for models with different degrees of freedom per year in the smooth functions of the weather variables, and our final model, with 6 df per year for temperature and 3 df per year for dew point temperature and wind speed, is the one with the smallest DIC. Thus, the selection of degrees of freedom in the smooth functions of the weather variables in our model is reasonable.

3.5 Discussion

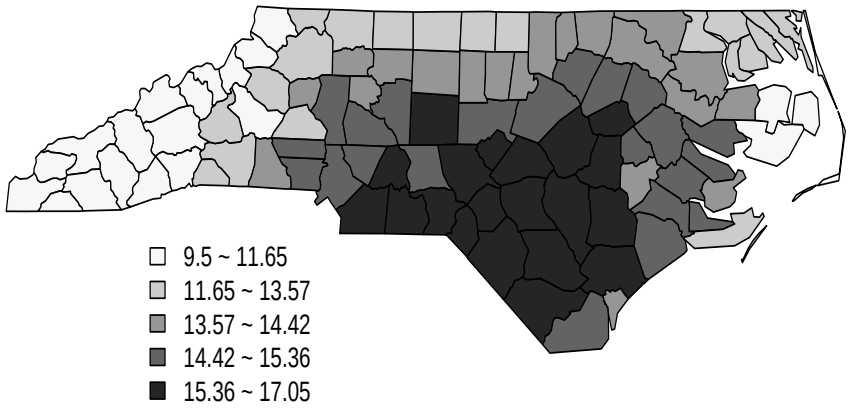
This chapter presents a Bayesian framework to investigate the spatial-temporal association between $PM_{2.5}$ exposure and daily mortality. We introduce a spatial-temporal model for $PM_{2.5}$ to obtain daily $PM_{2.5}$ concentrations by combining observed $PM_{2.5}$ data and numerical model output for $PM_{2.5}$. We estimate the association between daily mortality and different timescales of $PM_{2.5}$ to investigate the harvesting effect. Our approach to adjust for time-varying confounders does not require the selection of the number of basis functions. Our hierarchical framework takes into account the spatial and temporal dependency in the pollution and mortality data,

and different sources of uncertainty about them.

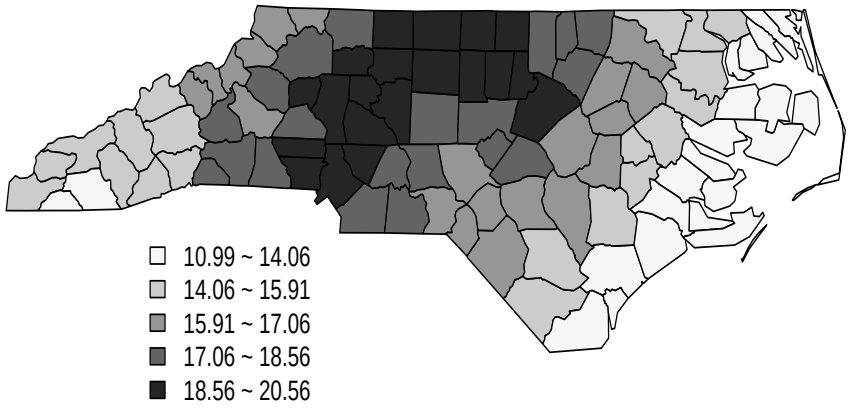
The $PM_{2.5}$ and mortality association in NC is inconsistent with the harvesting-only hypothesis, and our harvesting resistant estimates of the relative risk are actually larger, not smaller, than the ordinary estimates. Our results are consistent with some other harvesting analyses (Zeger et al., 1999; Schwartz, 2000; Dominici et al., 2003). We found a similar association between different timescales and mortality for all seasons in NC. However, the association of $PM_{2.5}$ and the current day mortality in the winter is higher than in the spring in NC.

In this study, we used sparse monitoring $PM_{2.5}$ data (across space and time) and the CMAQ output for $PM_{2.5}$. Our results show that adding the CMAQ output reduces the amount of uncertainty in our estimated relative risk parameter.

The framework introduced here is the first step to illustrate the benefits of combining different sources of information using a hierarchical framework that allows for a space-time varying risk assessment. This approach could easily be implemented for other geographic domains, including data for the conterminous United States and for longer time windows.

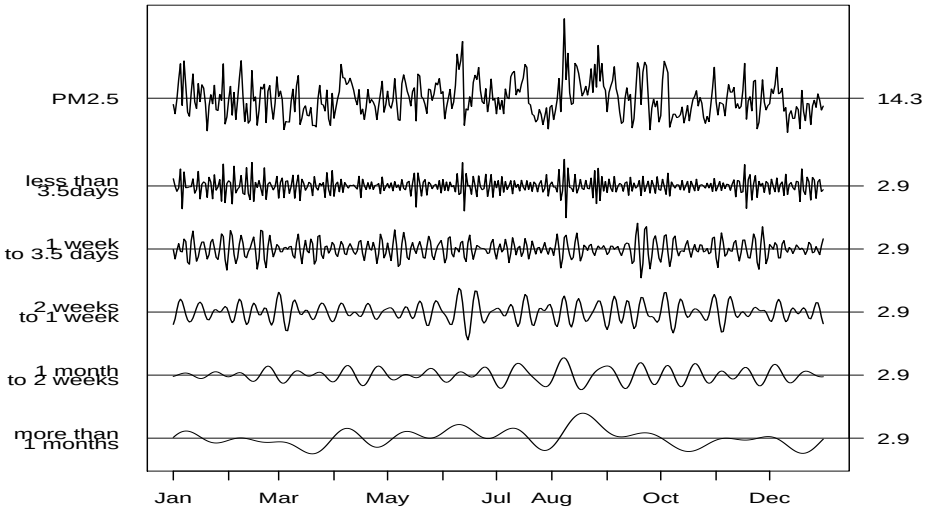


(a) January 2001

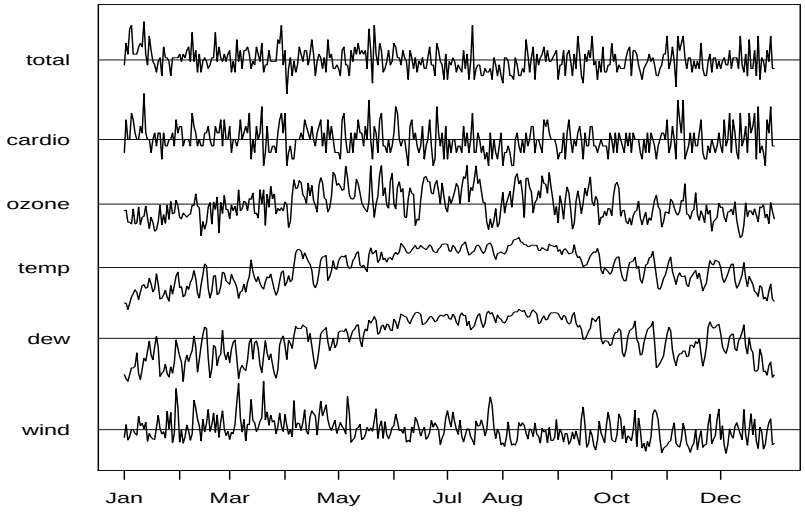


(b) August 2001

Figure 3.4: Maps of the estimated monthly average of the $PM_{2.5}$ concentrations for (a) January 2001 and (b) August 2001.

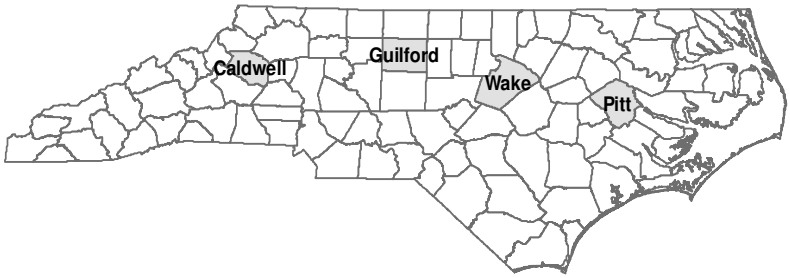


(a) Orthogonal decomposition of the $PM_{2.5}$ time series

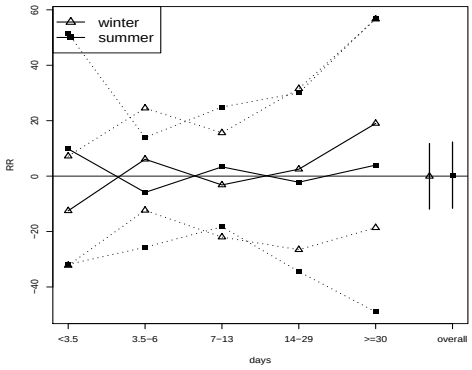


(b) Time series of mortality, ozone, and weather variables

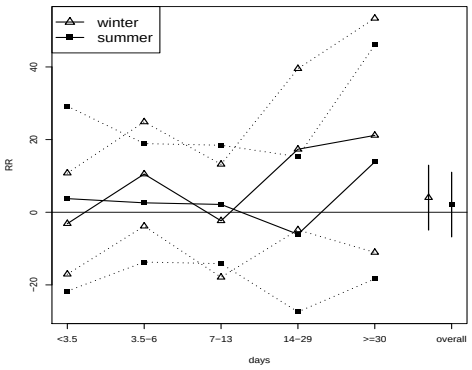
Figure 3.5: (a) Orthogonal decomposition of the $PM_{2.5}$ time series and (b) time series of total natural deaths (total), cardiovascular deaths (cardio), ozone, temperature (temp), dew point (dew), and wind speed (wind) for Wake County in the year 2001. Horizontal lines show the mean value. For Wake County, the mean of the estimated $PM_{2.5}$ is $14.3\mu g/m^3$ and the mean of each timescale is $2.9\mu g/m^3$.



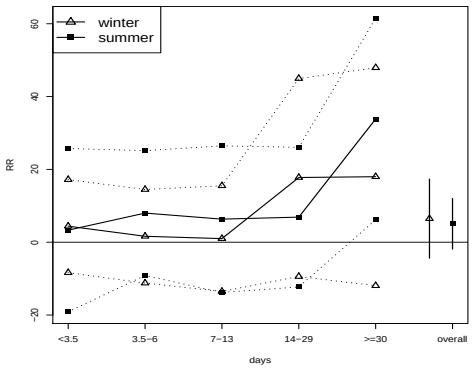
(a) Map



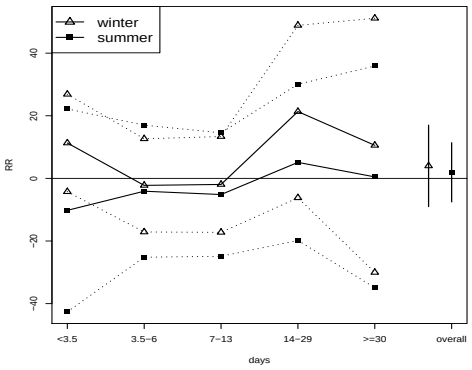
(b) Caldwell County



(c) Guiford County

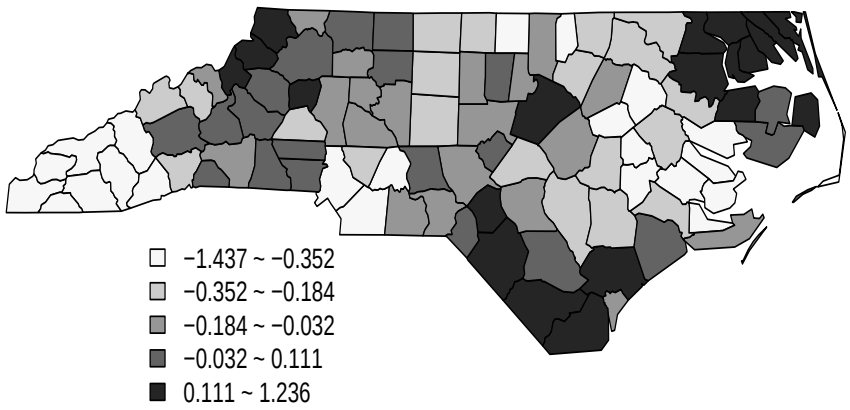


(d) Wake County

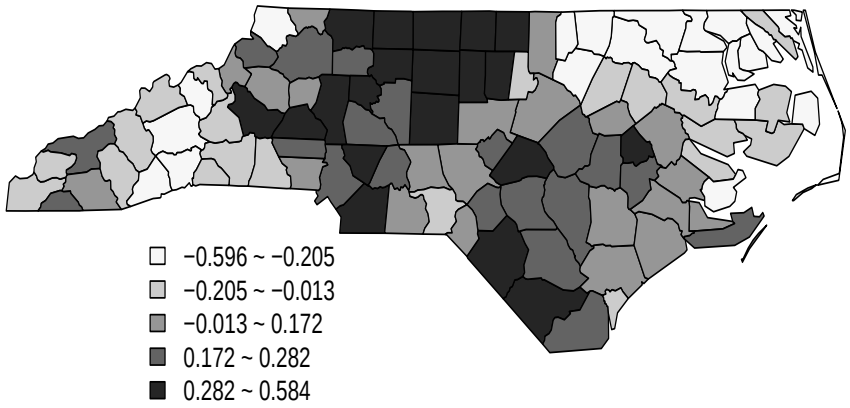


(e) Pitt County

Figure 3.6: Map shows the location of 4 counties in NC. Mean of the posterior distribution and 95% prediction intervals for the log relative rates of mortality at different timescales (percent increase in mortality per increase of $10\mu g/m^3$ of $PM_{2.5}$ concentrations) in winter and summer. The values presented at "overall" are the estimates of log relative rates of mortality due to same-day $PM_{2.5}$ exposure.



(a) Winter



(b) Summer

Figure 3.7: Impact of ozone. Change in the log relative risk of mortality by adding ozone (a) in the winter and (b) in the summer (RR when ozone is not in the model minus RR when ozone is in the model).

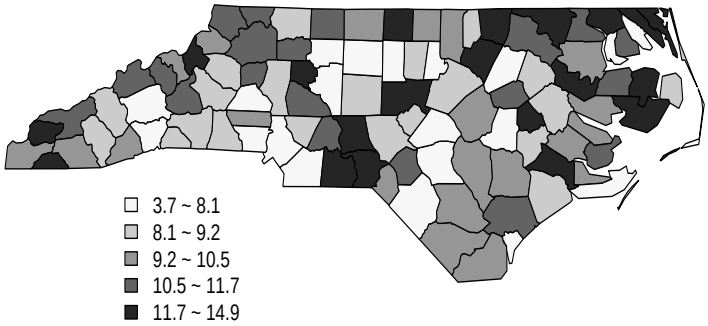


Figure 3.8: Bayesian estimates (mean of posterior distribution) of the number of basis functions included in the model, W_j.

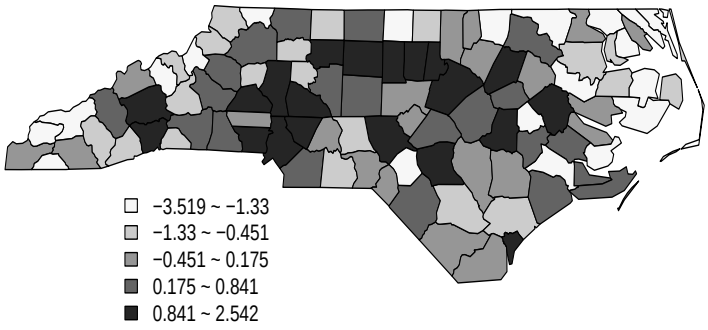
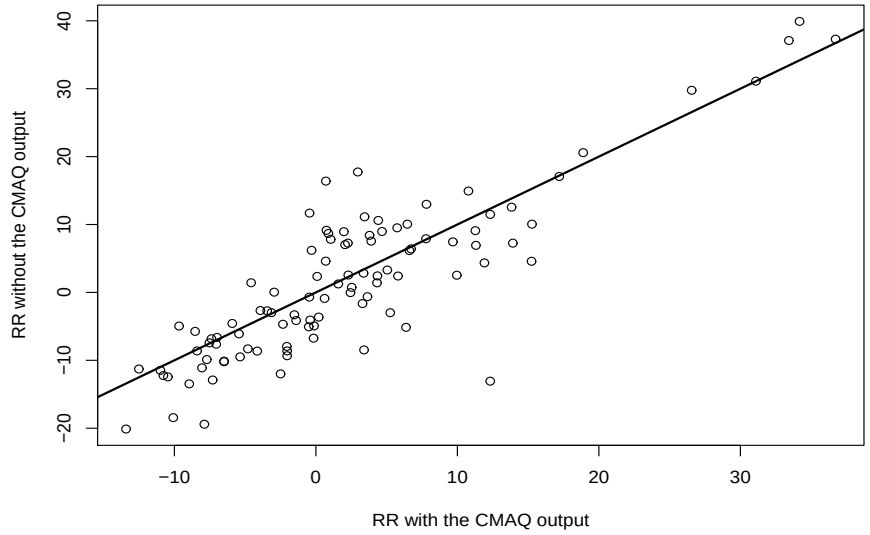
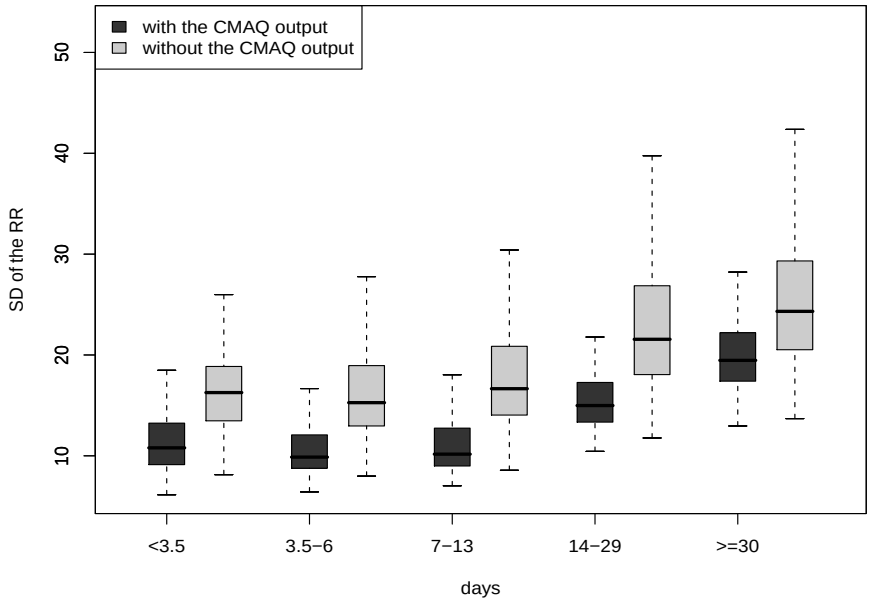


Figure 3.9: Bayesian estimates (mean of posterior distribution) of the intercept, γ_j.

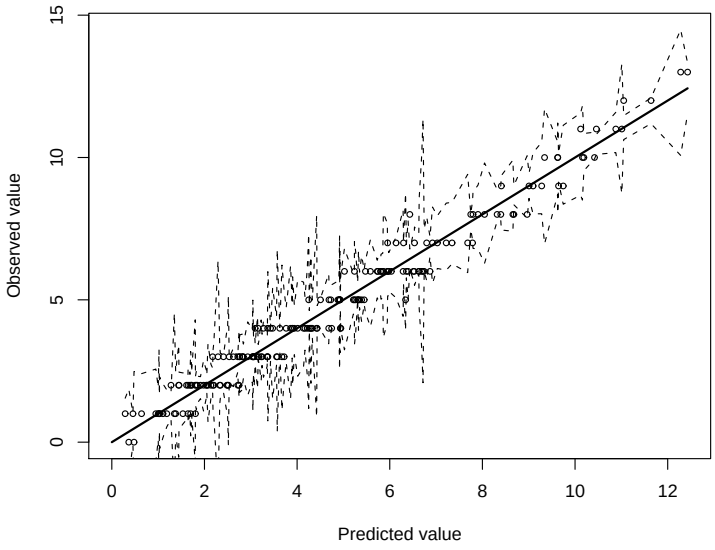


(a) Estimated RR values on the shortest timescale with and without CMAQ

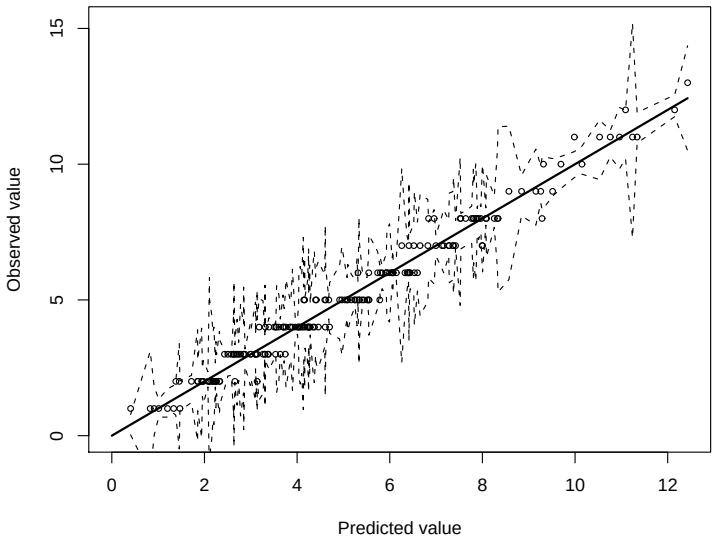


(b) Estimated SDs with and without CMAQ

Figure 3.10: (a) Estimated RR values on the shortest timescale in the winter with and without using CMAQ output in our model and (b) Standard deviations of the estimated RR in the winter when the CMAQ output was used in the model and when the CMAQ output were not used. The solid line in (a) shows $y = x$.

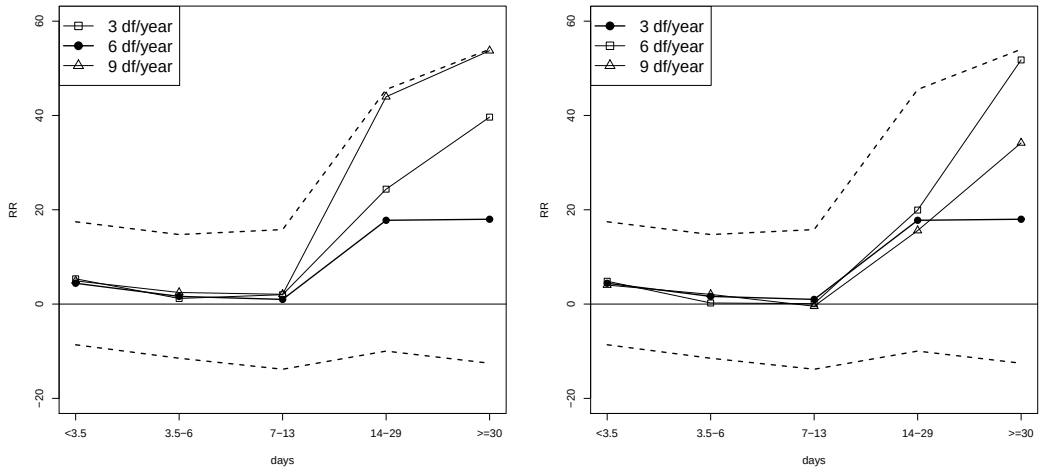


(a) Summer



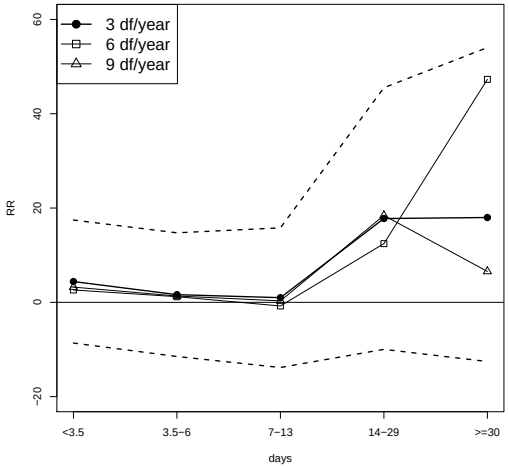
(b) Fall

Figure 3.11: Model diagnostics for mortality (a) during the summer and (b) during the fall: The dotted lines show the 95% prediction intervals.



(a) Temperature

(b) Dew point temperature



(c) Wind speed

Figure 3.12: Sensitivity analysis using different numbers of degrees of freedom (df) for (a) temperature, (b) dew point temperature, and (c) wind speed but the same functions for the other weather variables. The estimates of log relative rates of mortality are plotted at different timescales for Wake County in winter (percent increase in mortality per increase of $10\mu g/m^3$ of $PM_{2.5}$ concentrations). The dashed lines indicate 95% prediction intervals for the estimates when we used (a) 6 df per year for temperature, (b) 3 df per year for dew point temperature, and (c) 3 df per year for wind speed.

Chapter 4

Conclusion

In this thesis we developed statistical models for speciated $PM_{2.5}$ and the effects of $PM_{2.5}$ on mortality. This chapter presents a general discussion of the results in this study and the future work.

We developed a multivariate spatial-temporal model for speciated $PM_{2.5}$ using all available sources of data about speciated $PM_{2.5}$ in order to investigate the spatial-temporal patterns of speciated $PM_{2.5}$ and predict speciated $PM_{2.5}$. A linear model of coregionalization was extended to the spatiotemporal setting to account for spatial-temporal dependency structures for each component and the associations among the components. We also introduced a statistical framework to combine different sources of $PM_{2.5}$ data, while accounting for potential bias over space and time. Our statistical framework was applied to data in the United States for the year 2004. We found that sulfate concentrations were high during the summer while nitrate concentrations were high during the winter. Total carbonaceous mass concentrations were high during the

summer and fall seasons. We also found some spatial patterns of speciated $PM_{2.5}$. Sulfate concentrations were high in the eastern United States during the summer. Nitrate concentrations were high in urban areas, and crustal material concentrations were high in the eastern United State during the spring.

In the second part of the thesis, we explored the spatial-temporal association between exposure to $PM_{2.5}$ and daily mortality because $PM_{2.5}$ and its chemical components change with space and season and the effects of $PM_{2.5}$ on mortality may vary across space and season. However, observed $PM_{2.5}$ data are sparse across space and time so we combined both observed $PM_{2.5}$ data and the CMAQ output for $PM_{2.5}$ by introducing a spatial-temporal model which predicts $PM_{2.5}$ concentrations. We introduced a Bayesian hierarchical regression model to examine the association between mortality and different timescales of $PM_{2.5}$ across space and season. We also proposed an approach to adjust for time-varying confounders using a stochastic search variable selection approach, which did not need the selection of the number of basis functions. The hierarchical framework takes into account the spatial and temporal dependency in the mortality data and the pollution data. By adding the CMAQ output, we improved estimates of the relative risk for $PM_{2.5}$.

As part of our further research, we could investigate the association between speciated $PM_{2.5}$ and mortality over the entire United States. Some of the toxicology studies indicated that some chemical components of $PM_{2.5}$ were more closely linked with health problems than other components. However, there are few studies on the association between speciated $PM_{2.5}$ and mortality because of the lack of availability of data. Fuentes et al. (2006) studied the health effects of speciated $PM_{2.5}$ over space

using monthly data, but they did not consider a temporal association. We could estimate speciated $PM_{2.5}$ at all locations and times of interest using the statistical framework introduced in Chapter 2, which would be used as the inputs in the health model. Our health model for $PM_{2.5}$ presented in Chapter 3 could be extended to the speciated $PM_{2.5}$ to take into account the effects of speciated $PM_{2.5}$ on mortality over space and time.

Bibliography

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica* **33**, 178-196.

American Thoracic Society, and Bascom, R. (1996a). Health effects of outdoor air pollution. Part 1. *American Journal of Respiratory and Critical Care Medicine* **153**, 3-50.

American Thoracic Society, and Bascom, R. (1996b). Health effects of outdoor air pollution. Part 2. *American Journal of Respiratory and Critical Care Medicine* **153**, 477-498.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall, New York.

Bates, D.V., Baker-Anderson, M., and Sizto, R. (1990). Asthma attack periodicity: A study of hospital emergency visits in Vancouver. *Environmental Research* **51**, 51-70.

Baumgardner, R.E., Isil, S.S., Bowser, J.J., and Fitzgerald, K.M. (1999). Measurements of rural sulfur dioxide and particle sulfate: Analysis of CASTNet data, 1987 through 1996. *Journal of Air Waste Management Association* **49**, 1266-1279.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1-59.

Binkowski, F.S., and Roselle, S.J. (2003). Models-3 community multiscale air quality (CMAQ) model aerosol component, 1. Model description. *Journal of Geophysical Research* **108**, 4183, doi:10.1029/2001JD001409.

Box, G. and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, revised ed.

Brown, P.E., Karesen, K.F., Roberts, G.O., and Tonellato, S. (2000). Blurgenerated nonseparable space-time models. *Journal of the Royal Statistical Society, Series B* **62**, 847-860.

Brown, P.J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian selection and prediction. *Journal of the Royal Statistical Society, Series B* **60**, 627-641.

Burnett, R., and Krewski, D. (1994). Air pollution effects of hospital admission rates: A random effects modelling approach. *The Canadian Journal of Statistics* **22**, 441-458.

Byun, D.W., and Schere, K.L. (2006). Review of the governing equations, computational algorithms and other components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. *Applied Mechanics Reviews* **59**, 51-77.

Carroll, R., Chen, R., George, E., Li, T., Newton, H., Schmiediche, H., and Wang, N. (1997). Ozone exposure and population density in Harris County, Texas. *Journal of the American Statistical Association* **92**, 392-415.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., and Terpenning, I. (1990). Seasonal-trend decomposition procedure based on LOESS. *Journal of Official Statistics* **6**, 3-73.

Cressie, N., and Huang, H.C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* **94**, 1330-1340.

Daniels, M., Dominici, F., Samet, J.M., and Zeger, S.L. (2000). Estimating PM_{10}-mortality dose-response curves and threshold levels: An analysis of daily timeseries for the 20 largest U.S. cities. *American Journal of Epidemiology* **152**, 397-412.

Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., and Speizer, F.E. (1993). An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* **329**, 1753-1759.

Dockery, D.W., Schwartz, J., and Spengler, J.D. (1992). Air pollution and daily mortality: Associations with particulates and acid aerosols. *Environmental Research* **59**, 362-373.

Dominici, F., Daniels, M., Zeger, S.L, and Samet, J.M. (2002a). Air pollution and mortality: Estimating regional and national dose-response relationships. *Journal of the American Statistical Association* **97**, 100-111.

Dominici, F., McDermott, A., Zeger, S.L., and Samet, J.M. (2002b). On the use of generalized additive models in time series of air pollution and health. *American Journal of Epidemiology* **156**, 193-203.

Dominici, F., McDermott, A., Zeger, S.L., and Samet, J.M. (2003). Airborne particulate matter and mortality: Timescale effects in four US cities. *American Journal of Epidemiology* **157**, 1055-1065.

Dominici, F., Samet, J.M., and Zeger, S.L. (2000). Combining evidence on air pollution and daily mortality from the twenty largest U.S. cities: A hierarchical modeling strategy, (with discussion). *Journal of the Royal Statistical Society, Series A* **163**, 263-302.

Eilers, P., and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89-121.

Famoye, F. (1993). Restricted generalized Poisson regression model. *Communication in Statistics-Theory and Methods* **22**, 1335-1354.

Flanagan, J.B., Peterson, M.R., Jayanty, R.K.M., and Rickman, E.E. (2003). *Analysis of Speciation Network Carbon Blank Data*. RTI International. RTP, NC.

Frank N.H. (2006). Retained nitrate, hydrated sulfates, and carbonaceous mass in federal reference method fine particulate matter for six eastern U.S. cities. *Journal of Air Waste Management Association* **56**, 500-511.

Fuentes, M., Chen, L., Davis, J.M., and Lackmann, G.M. (2005). Modeling and predicting complex space-time structures and patterns of coastal wind fields. *Environmetrics* **16**, 449-464.

Fuentes, M., and Raftery, A.E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61**, 36-45.

Fuentes, M., Song, H., Ghosh, S.K., Holland, D.M., and Davis, J.M. (2006). Spatial association between speciated fine particles and mortality. *Biometrics* **62**, 855-863.

Gelfand, A.E., Banerjee, S., and Gamerman, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* **16**, 465-479.

Gelfand, A.E., Ghosh, S.K., Knight, J.R., and Sirmans, C.F. (1998). Spatio-temporal modeling of residential sales data. *Journal of Business and Economic Statistics* **16**, 312-321.

Gelfand, A.E., Schmidt, A.M., Banerjee, S., and Sirmans, C.F. (2004). Nonstationary multivariate process modelling through spatially varying coregionalization (with discussion). *Test* **13**, 1-50.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Gelfand, A.E., and Vounatsou, P. (2002). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**, 11-25.

Gelman, A. (2004). Parameterization and Bayesian modelling. *Journal of the American Statistical Association* **99**, 537-545.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457-72.

George, E.I., and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.

George, E.I., and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-373.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97**, 590-600.

Golam Kibria, B.M., Sun, L., Nhu, L., and Zidek, V. (2002). Bayesian spatial prediction of random space-time fields with application to mapping PM 2.5 exposure. *Journal of the American Statistical Association* **97**, 112-124.

Gotway, C.A., and Young, L.J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association* **97**, 632-648.

Grzebyk, M., and Wackernagel, H. (1994). Multivariate analysis and spatial/temporal scales: real and complex models. *Proceedings of the XVIIth International Biometrics Conference*, 19-33.

Guttorp, P., Meiring, W., and Sampson, P. (1994). A space-time analysis of ground level ozone data. *Environmetrics* **5**, 241-254.

Huang, H.C., and Cressie, N.A.C. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis* **22**, 159-175.

Huerta, G., Sanso, B., and Stroud, J.R. (2004). A spatiotemporal model for Mexico city ozone levels. *Applied Statistics* **53**, 1-18.

Jin, X., Banerjee, S., and Carlin, B.P. (2007). Order-free coregionalized lattice models with application to multiple disease mapping. *Journal of the Royal Statistical Society, Series B* **69**, 817-838.

Katsouyanni, K., Touloumi, G., Spix, C., Schwartz, J., Balducci, F., Medina, S., Rossi, G., Wojtniak, B., Sunyer, J., Bacharova, L., Schouten, J., Ponka, A., and Anderson, H. (1997). Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities. *British Medical Journal* **314**, 1658-1663.

Kelsall, J.E., and Wakefield, J.C. (1999). Discussion of "Bayesian models for spatially correlated disease and exposure data", in: Bayesian Statistics 6, Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M. (eds.), Oxford: Oxford University Press, p. 151.

Lee, D., and Shaddick, G. (2007). Time-varying coefficient models for the analysis of air pollution and health outcome data. *Biometrics* **63**, 1253-1261.

Malm, W.C., Schichtel, B.A., Ames, R.B., and Gebhart, K.A. (2002). A 10-year spatial and temporal trend of sulfate across the United States. *Journal of Geophysical Research* **107**, D224627, doi:10.1029/2002JD002107.

Malm, W.C., Schichtel, B.A., Pitchford, M.L., Ashbaugh, L.L., and Eldred, R.A. (2004). Spatial and monthly trends in speciated fine particle concentration in the United States. *Journal of Geophysical Research* **109**, D03306, doi:10.1029/2003JD003739.

Mardia, K.V., Goodall, C., Redfern, E.J., and Alonso, F.J. (1998). The kriged Kalman filter (with discussion). *Test* **7**, 217-285.

Matern, B. (1986). *Spatial variation*. Springer-Verlag, Berlin.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. in *Frontiers of Econometrics*, ed. by P. Zarembka. Academic Press, New York. p.105-142.

Ntzoufras, I., Forster, J.J., and Dellaportas, P. (1997). Stochastic search variable selection for log-linear models. *Technical Report at Faculty of Mathematics, Southampton University, Southampton, UK.*

Ostro, B.D., Lipsett, M.J., Wiener, M.B., and Selner, J.C. (1991). Asthmatic responses to airborne acid aerosols. *American Journal of Public Health* **81**, 694-702.

Özkaynak, H., and Thurston, G.D. (1987). Associations between 1980 U.S. mortality rates and alternative measures of airborne particle concentration. *Risk Analysis* **7**, 449-461.

Pfeifer, P.E., and Deutsch, S.J. (1980a). Independence and sphericity tests for the residuals of space-time ARMA models. *Communications in Statistics, B* **9**, 533-549.

Pfeifer, P.E., and Deutsch, S.J. (1980b). Stationarity and invertibility regions for low order STARMA models. *Communications in Statistics, B* **9**, 551-562.

Pope, C.A., Dockery, D., and Schwartz, J. (1995a). Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation Toxicology* **47**, 1-18.

Pope, C.A., and Schwartz, J. (1996). Time series for the analysis of pulmonary health data. *American Journal of Respiratory and Critical Care Medicine* **154**, S229-S233.

Pope, C.A., Thun, M.J., Namboodiri, M.M., Dockery, D., Evans, J., Speizer, F., and Health, C.M. (1995b). Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal Respiratory and Critical Care Medicine* **151**, 669-674.

Rao, V., Frank, N., Rush, A., and Dimmick, F. (2003). Chemical speciation of $PM_{2.5}$ in urban and rural areas. *National Air Quality and Emissions Trends Report*, 13-23.

Sahu, S.K., Gelfand, A.E., and Holland, D.M. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 61-86.

Sahu, S.K., and Mardia, K.V. (2005). A Bayesian kriged-Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society, Series C* **54**, 223-244.

Samet, J., Zeger, S., Dominici, F., Curriero, F., Coursac, I., Dockery, D., Schwartz, J., and Zanobetti, A. (2000a). *The National Morbidity, Mortality, and Air Pollution Study Part II: Morbidity and Mortality from Air Pollution in the United States.* Health Effects Institute, Cambridge, MA.

Samet, J., Zeger, S., Dominici, F., Dockery, D., and Schwartz, J. (2000b). *The National Morbidity, Mortality, and Air Pollution Study Part I: Methods and Methodological Issues.* Health Effects Institute, Cambridge, MA.

Sampson, P., and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association* **87**, 108-119.

Schabenberger, O., and Gotway, C.A. (2004) *Statistical Methods for Spatial Data Analysis.* Chapman and Hall, Florida.

Schmidt, A.M., and Gelfand, A.E. (2003). A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research* **108**, 8783.

Schmidt, A.M., and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society, Series B* **65**, 745-758.

Schwartz, J. (1994). Air pollution and daily mortality: A review and meta analysis. *Environmental research* **64**, 36-52.

Schwartz, J. (1995). Air pollution and daily mortality in Birmingham, Alabama. *American Journal of Epidemiology* **137**, 1136-1147.

Schwartz, J. (2000). Harvesting and long-term between exposure effects in the relationship between air pollution and mortality. *American Journal of Epidemiology* **151**, 440-448.

Shaddick, G., and Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Applied Statistics* **51**, 351-372.

Smith, M., and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317-343.

Smith, R.L., Kim, Y., Fuentes, M., and Spitzner, D. (2000). Threshold dependence of mortality effects for fine and coarse particles in Phoenix, Arizona. *Journal of the Air and Waste Management Association* **50**, 1367-1379.

Smith, R.L., Kolenikov, S., and Cox, L.H. (2003). Spatio-temporal modelling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research-Atmospheres* **108**(D24), 9004, doi:10.1029/2002JD002914.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 583-639.

Stein, M.L. (2005). Space-time covariance functions. *Journal of the American Statistical Association* **100**, 310-321.

Stoffer, D. (1986). Estimation and identification of space-time ARMAX modelsin the presence of missing data. *Journal of the American Statistical Association* **81**, 762-772.

Stroud, J.R., Müller, P., and Sanso, B. (2001). Dynamic models for spatiotemporal data. *Journal of Royal Statistical Society, Series B* **63**, 673-689.

Sun, D., Tsutakawa, R.K., and Speckman, P. (1999). Posterior distribution of hierarchical models using CAR (1) distributions. *Biometrika* **86**, 341-390.

Tonellato, S. (1997). Bayesian dynamic linear models for spatial time series. *Technical report at Dipartimento di Statistica, Universita CaFoscari di Venezia, Venice, Italy.*

Tonellato, S. (1998). Spatial prediction with space-time models. *Technical report at Dipartimento di Statistica, Universita CaFoscari di Venezia, Venice, Italy.*

U.S. Environmental Protection Agency (1997). *National ambient air quality standards for particulate matter; Final Rule, Part II.* Federal Register **40**, CFR Part 50.

U.S. Environmental Protection Agency (1998). *Quality assurance requirements for state and local air monitoring stations (SLAMS).* Federal Register **40**, CFR Part 68.

U.S. Environmental Protection Agency (2000). *Quality assurance guidance document, quality assurance project plan: $PM_{2.5}$ speciation trends network field sampling.* EPA 454/R-01-001, RTP, NC. `http://www.epa.gov/ttn/amtic/files/ambient/pm25/spec/1025sqap.pdf`.

U.S. Environmental Protection Agency (2003). *National air quality and emissions trends report, 2003, special studies edition.* EPA 454/R-03-005, RTP, NC. `http://www.epa.gov/air/airtrends/aqtrnd03/`.

U.S. Environmental Protection Agency (2004). *Air quality criteria for particulate matter.* National Center for Environmental Assessment-RTP Office, RTP, NC. `http://cfpub.epa.gov/ncea/[9November,2004]`.

Ver Hoef, J.M., and Barry, R.D. (1998). Constructing and fitting models for cokriging and multivariate spatial prediction. *Journal of Statistical Planning and Inference* **69**, 275-294.

Wackernagel, H. (1998). *Multivariate Geostatistics-An Introduction with applications* (2nd ed.). Springer-Verlag, New York.

Waller, L., Carlin, B.P., Xia, H., and Gelfand, A.E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* **92**, 607-617.

West M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). Springer, New York.

Wikle, C., Berliner, M., and Cressie, N. (1998). Hierarchical Bayesian spacetime models. *Environmental and Ecological Statistics* **5**, 117-154.

Wikle, C., and Cressie, N. (1999). A dimension reduced approach to space-time Kalman filtering. *Biometrika* **86**, 815-829.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association* **95**, 1076-1987.

Zanobetti, A., Schwartz, J., and Ryan, L.M. (2000). Generalized additive distributed lag models: Quantifying mortality displacement. *Biostatistics* **1**, 279-292.

Zeger, S.L., Dominici, F., and Samet, J. (1999). Harvesting-resistant estimates of air pollution effects on mortality. *Epidemiology* **10**, 171-175.

Zidek, J.V., Sun, L., Le, N., and Ozkaynak, H. (2002). Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM_{10} field. *Environmetrics* **13**, 595-613.

APPENDICES

Appendix A

Discrete Fourier Transformation

The daily time series of $PM_{2.5}$ for county j, $Z_j(t)$, $t = 0, \ldots, T-1$, is decomposed into L orthogonal timescale components, $Z_{j1}(t), Z_{j2}(t), \ldots, Z_{jL}(t)$, where $\sum_{l=1}^{L} Z_{jl}(t) = Z_j(t)$. For each county j, the discrete Fourier transform is defined as

$$d_j(\omega_m) = \frac{1}{T} \sum_{t=0}^{T-1} Z_j(t) \exp(-i\omega_m t),$$

where $1 \leq m \leq T-1$, i is the imaginary unit ($i^2 = -1$), and T is the length of the time series $Z_j(t)$. The m^{th} Fourier frequency is $\omega_m = 2\pi m/T$, where $0 \leq \omega_m \leq 2\pi$, and it has m cycles in the length of the data. Note that for $m \geq T/2$, $d_j(\omega_{T-m}) = \overline{d_j(\omega_m)}$, where $\overline{d_j(\omega_m)}$ is the complex conjugate of $d_j(\omega_m)$.

The inverse discrete Fourier transform is given by

$$Z_j(t) = \sum_{m=0}^{T-1} d_j(\omega_m) \exp(i\omega_m t).$$

Let $[0 = \omega_0, \omega_1, \ldots, \omega_l, \ldots, \omega_L, \pi]$ be a partition of the interval $[0, \pi]$, and we set

$I_l = (\omega_{l-1}, \omega_l] \cup [\omega_{T-l}, \omega_{T-l+1})$. Then, the previous equation is represented as

$$\begin{aligned} Z_j(t) &= \sum_{l=1}^{L} \Big\{ \sum_{\omega_m \in I_l} d_j(\omega_m) \exp(i\omega_m t) \Big\} \\ &= \sum_{l=1}^{L} Z_{jl}(t). \end{aligned}$$

Thus, $Z_j(t)$ can be decomposed into Z_{jl}'s using the following algorithm, for $l = 1, \ldots, L$,

(i) Compute the discrete Fourier transform of $Z_j(t)$ and obtain $d_j(\omega_m)$.

(ii) Let $d_j^*(\omega_m) = d_j(\omega_m)$, if $\omega_m \in I_l$, and $d_j^*(\omega_m) = 0$, if $\omega_m \notin I_l$.

(iii) Obtain Z_{jl} by the inverse of the discrete Fourier transform using $d_j^*(\omega_m)$, $m = 1, \ldots, T/2$.

Appendix B

Ozone Model

We introduce a spatial-temporal model for ozone to obtain daily predicted values of ozone concentrations for all counties, that are used as inputs in the health model in Section 3.3.2. Let $\widehat{O}(\mathbf{s},t)$ be the observed ozone concentration at location $\mathbf{s}$ on day t. We denote "true" ozone concentration at location $\mathbf{s}$ on day t by $O(\mathbf{s},t)$. The model for ozone is

$$\widehat{O}(\mathbf{s},t) \;\; = \;\; O(\mathbf{s},t) + e_{\hat{o}}(\mathbf{s},t),$$

where $e_{\hat{o}}(\mathbf{s},t) \sim N(0,\sigma^2_{\hat{o}})$ is the measurement error at location $\mathbf{s}$ on day t, which is independent of the true underlying process. Following the guidance for the precision provided by the U.S. EPA (1998), we assign an informative uniform prior, Unif(0,5), for $\sigma_{\hat{o}}$.

Since ozone is affected by weather covariates and has seasonal trends, the true ozone concentration, $O(\mathbf{s},t)$, is modeled as

$$O(\mathbf{s},t) = \mathbf{M}^T(\mathbf{s},t)\boldsymbol{\zeta}_o + S_o(t) + e_o(\mathbf{s},t),$$

where $\boldsymbol{\zeta}_o$ is a coefficient vector corresponding to the weather variables $\mathbf{M}$ (minimum temperature, maximum temperature, dew point temperature, wind speed, and pressure). The function $S_o(t)$ explains the seasonal trend in ozone. The error process $e_o(\mathbf{s}, t)$ is assumed to be normal with mean zero and a spatial-temporal covariance function. In the application, from the results of the periodogram analysis, we assume that $S_o(t)$ is a linear combination of two sine and two cosine functions with respect to 3-month and 12-month periods. Based on exploratory analysis, $e_o(\cdot, t) = (e_o(\mathbf{s}_1, t), \ldots, e_o(\mathbf{s}_{N_s}, t))$ is normal with mean $\psi_o e_o(\cdot, t-1)$ and exponential covariance $\sigma_o^2 \exp(-h_1/\phi_o)$, where $h_1 = ||\mathbf{s} - \mathbf{s}'||$ (in km). We use a normal prior, N(0,0.1) (0.1 is the precision), for ψ_o and we use uniform priors, Unif(1,500) and Unif(0,100), for ϕ_o and σ_o, respectively.

We predict $O(\mathbf{s}_0, t_0)$, the true ozone value at space $\mathbf{s}_0$ and time t_0, given the data, $\widehat{O}$ and $\mathbf{M}$. The posterior predictive distribution of $O(\mathbf{s}_0, t_0)$ given $\widehat{O}$ and $\mathbf{M}$ is

$$p(O(\mathbf{s}_0, t_0)|\widehat{O}, \mathbf{M}) \propto \int p(O(\mathbf{s}_0, t_0)|\widehat{O}, \mathbf{M}, \boldsymbol{\Theta}_O) p(\boldsymbol{\Theta}_O|\widehat{O}, \mathbf{M}) d\boldsymbol{\Theta}_O,$$

where $\boldsymbol{\Theta}_O$ is a collection of all parameters considered in the ozone model. After simulating N_3 values from the posterior distribution of the parameters $\boldsymbol{\Theta}_O$, the estimator for the predictive distribution is

$$p(O(\mathbf{s}_0, t_0)|\widehat{O}, \mathbf{M}) = \frac{1}{N_3} \sum_{n_3=1}^{N_3} p(O(\mathbf{s}_0, t_0)|\widehat{O}, \mathbf{M}, \boldsymbol{\Theta}_O^{(n_3)}),$$

where $\boldsymbol{\Theta}_O^{(n_3)}$ is the n_3^{th} draw from the posterior distribution.

To obtain daily ozone values for each county, the underlying ozone value for county j on day t, $O_j(t)$, averages true ozone values over a spatial domain C_j within county

j on day t

$$O_j(t) = \frac{1}{|C_j|} \int_{C_j} O(\mathbf{s}, t) d\mathbf{s}.$$

The ozone estimate $O_j(t)$ is obtained by averaging estimates of true values at several locations randomly selected within county j on day t. We use these estimates as inputs in the health model in Section 3.3.2.