SONG, HAE-RYOUNG. Associations between Gaussian Markov Random Fields and Gaussian Geostatistical Models with an Application to Model the Impact of Air Pollution on Human Health. (Under the direction of Montserrat Fuentes and Sujit Ghosh.)

In many applications in environmental sciences and in epidemiology, data concerning a spatial process of interest are often observed at different spatial resolutions. For example, in studies of the association between air pollution exposure and adverse health effects, the pollution data and the health outcomes may be available as point-referenced data and as areal data, respectively. One common solution to this spatial misalignment problem is to aggregate the point-referenced data to the area level and create a common support for both variables. Once the point-referenced data are aggregated to the areal level, the process representing the aggregated data is modeled using integrals of spatial continuous process. Modeling aggregated data using these spatial integrals is computationally very expensive. Hence, because of the computational benefits it is becoming more common to use Gaussian Markov random fields (GMRFs) in modeling aggregated point-referenced data.

The modeling of aggregated point data using GMRFs serves as one of our motivations to investigate the connection between GMRFs and GGMs (Gaussian geostatistical models) which are generally used to model areal and point data, respectively. GGMs are used in modeling an underlying continuous process observed as point data. On the other hand, GMRFs are widely used for modeling areal data such as counts or spatial averages of a quantity over subregions of a larger region. In this thesis, the
relations between GMRFs and GGMs are explored based on approximations of GM-
RFs by GGMs, and vice versa. The proposed framework for the comparison of GGMs
and GMRFs is based on minimizing the distance between the corresponding spectral
density functions. In particular, the Kullback-Leibler discrepancy of spectral densi-
ties and the chi-squared distance between spectral densities are used as the metrics
for the approximation. The performance of our proposed methodology is compared
to that of the approaches using covariances in terms of mean squared prediction error
(MSPE) and parameter estimation.

As a part of application, we model associations between speciated fine particulate
matter (PM) and mortality. Since mortality counts and PM are obtained at county
and point levels, respectively, we aggregate PM to the county level to combine the
variables with different support areas. The aggregated PM are modeled using GM-
RFs, and associations between PM and mortality are investigated based on Bayesian
hierarchical spatio-temporal framework. This model is applied to speciated PM_{2.5}
and monthly mortality counts over the entire U.S. region for 1999-2000. We obtain
high relative risks of mortality associated to PM_{2.5} in the Eastern and Southern Cal-
ifornia area. Particularly, NO_3 and crustal materials have greater health effects in
the Western U.S., while SO_4 and NH_4 have more of an impact in the Eastern U.S.
We show that the average risk associated with PM_{2.5} is approximately twice what we
obtained for PM_{10}
Associations between Gaussian Markov random fields and Gaussian geostatistical models with an application to model the impact of air pollution on human health

by

Hae-Ryoung Song

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

STATISTICS

Raleigh

2005

APPROVED BY:

________________________________________  ________________________________________
Dr. Montserrat Fuentes  Dr. Sujit Ghosh
Co-Chair of Advisory Committee  Co-Chair of Advisory Committee

________________________________________  ________________________________________
Dr. David Dickey  Dr. Peter Bloomfield

________________________________________
Dr. Jerry Davis
Biography

Hae-Ryoung Song was born in Seoul, Republic of Korea in 1978. She received her B.S. in the department of Applied Statistics from Yonsei University in Republic of Korea in 2000. She joined the Department of Statistics at North Carolina State University in the fall of 2001. She earned her Master of Statistics in 2003 and her Ph.D. in 2006. As of January 3, 2006, she is a research associate at the Department of Epidemiology and Biostatistics at University of South Carolina.
Acknowledgements

This dissertation would not have been possible to complete 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 advisors, Dr. Montserrat Fuentes and Dr. Sujit Ghosh for their patience and encouragement during my work on my dissertation. They provided me wonderful guidance and direction throughout this dissertation research. Dr. Jerry Davis has my gratitude for his careful reading of my drafts and many helpful comments. I would also like to thank Dr. David Holland for his contribution and good-natured support. The members of my dissertation committee, Dr. David Dickey and Dr. Peter Bloomfield have generously given their time and expertise to better my work. I thank them for their helpful suggestions and comments. I must acknowledge as well many friends, colleagues and faculty members who assisted, advised, and supported my research. Finally, I am grateful to my family for always being at my side, listening to me and giving me support.
List of Tables

2.1 Approximations of GMRFs with the 5th neighborhood order and various values of \( \alpha \) using GGMs. This table presents the estimated parameters of the GGM. .................................................. 22
2.2 Approximations of GMRFs with the 5th order neighborhood and various values of \( \alpha \) using GGMs. Predicting GMRFs using GGMs. .................. 23
2.3 Approximations of GMRFs with \( \alpha = 0.95 \) and various orders of neighborhood using GGMs. This table shows the estimated parameters of a GGM. .................................................. 23
2.4 Approximations of GMRFs with \( \alpha = 0.95 \) and various orders of neighborhood by GGMs. Predicting a GMRF using a GGM. ............................. 24
2.5 Approximations of a GGM with sill=1 and with various values of the range parameter using a GMRF with a 5th order neighborhood. Parameter estimates of the GMRF. ........................................ 24
2.6 Approximations of a GGM with sill=1 and with various values of the range parameter using a GMRF with a 5th order neighborhood. Predicting a GMRF. .................................................. 25
2.7 Parameter estimates of the GMRF using the first and second order neighborhoods .................................................. 27
2.8 MSPE in the approximations of aggregated PM\(_{2.5}\) using GMRFs with a first and second order neighborhood structures. .................. 27

3.1 Summary of the posterior distributions for the coefficients of the weather and pollution covariates for June 2000 ........................................ 51
3.2 Summary of the proportion of speciated PM for June 2000 .................. 51
3.3 Summary of the posterior distributions for the relative risk parameters for June 2000 .................................................. 56
3.4 Model comparisons are done using DIC (Deviance Information Criterion) and RMSPE (Root Mean Squared Prediction Error) .......................... 56
List of Figures

1.1 Isotropic covariance functions .................................................. 7
1.2 (a) First order (b) Second order (c) Third order neighborhood .......... 9
2.1 the 5th order neighborhood ......................................................... 22
2.2 county averaged PM\textsubscript{2.5} values ...................................... 26
3.1 (a): Monitoring stations for PM\textsubscript{2.5} (b): Weather stations from NCDC. 34
3.2 Natural death counts in June 2000 ............................................. 35
3.3 Total PM\textsubscript{2.5} mass (µg/m\textsuperscript{3}) (June 2000). .................. 52
3.4 Posterior mean of log relative risk of mortality (β\textsubscript{j}): percent increase in mortality per increase in 10 µg/m\textsuperscript{3} of PM\textsubscript{2.5} concentrations. (June 2000). .......................................................... 52
3.5 Standard deviation of the log relative risk. ................................... 53
3.6 Total PM\textsubscript{10} mass (µg/m\textsuperscript{3}) (June 2000). ...................... 53
3.7 Posterior mean of log relative risk of mortality (β\textsubscript{j}): percent increase in mortality per increase in 10 µg/m\textsuperscript{3} of PM\textsubscript{10} concentrations. (June 2000). .......................................................... 54
3.8 Posterior mean of log relative risk of mortality (β\textsubscript{j}): percent increase in cardiovascular mortality per increase in 10 µg/m\textsuperscript{3} of PM\textsubscript{2.5} concentrations. (June 2000). .......................................................... 54
3.9 Posterior mean of log relative risk of mortality (β\textsubscript{j}) with the 2nd order neighborhood: percent increase in mortality per increase in 10 µg/m\textsuperscript{3} of PM\textsubscript{2.5} concentrations. (June 2000). .......................................................... 55
3.10 Posterior mean of log relative risk of mortality (β\textsubscript{j}) for December 2000: percent increase in natural mortality per increase in 10 µg/m\textsuperscript{3} of PM\textsubscript{2.5} concentrations. .......................................................... 55
3.11 Impact of ozone. Change in the log relative risk of mortality (β\textsubscript{j}) by adding ozone, (beta parameter when ozone is NOT in the model minus beta parameter with ozone). .......................................................... 56
3.12 Posterior mean of relative risk of mortality due to total carbon: percent increase in mortality per increase in $10\,\mu g/m^3$ of nitrate concentrations. (June 2000). .......................... 57

3.13 Posterior mean of relative risk of mortality due to ammonium: percent increase in mortality per increase in $10\,\mu g/m^3$ of ammonium concentrations. (June 2000). .......................... 57

3.14 Posterior mean of relative risk of mortality due to crustal material: percent increase in mortality per increase in $10\,\mu g/m^3$ of crustal material concentrations. (June 2000). .......................... 58

3.15 Relative risk for sulfate by regions. Regions: 1 New England, 2 Middle Atlantic, 3 MIDWEST (East North Central), 4 MIDWEST (West North Central), 5 SOUTH (South Atlantic), 6 SOUTH (South Central), 7 WEST (Mountain), 8 WEST (Pacific), 9 WEST (Southern California). .................................................. 58

3.16 Relative risk for nitrate by regions. Regions: 1 New England, 2 Middle Atlantic, 3 MIDWEST (East North Central), 4 MIDWEST (West North Central), 5 SOUTH (South Atlantic), 6 SOUTH (South Central), 7 WEST (Mountain), 8 WEST (Pacific), 9 WEST (Southern California). .................................................. 59

3.17 Expected mortality ($\mu_i$). (June 2000). .......................... 59

3.18 Relative risk by ethnic groups. Groups, a) the Caucasian, b) the African-American and c) the Hispanic .......................... 60

3.19 Differences of RR between different ethnic groups. a)RR for Caucasian-RR for African-American b)RR for Caucasian-RR for Hispanic .... 60

3.20 Relative risk by age. Groups: (0-14, 15-64, over 65) .......................... 61

3.21 Selected locations for cross validation and model diagnostics .......................... 61

3.22 Model diagnostics for mortality based on the 95% prediction intervals for June 2000 .......................... 62

3.23 Model diagnostics for PM: 95% prediction intervals at location $i^{th}$, eliminating observation $i^{th}$. Total number of observations is 909. Good calibration: only 4% of the time the truth does not lie in the interval. 62

3.24 Model diagnostics for mortality for the months June, September and December. The solid line represents the true mortality counts, the dashed-line represents the predicted values, and the dotted-line represents 95% prediction intervals. .......................... 63
Chapter 1

Introduction to GGMs and GMRFs

This chapter introduces GGMs (Gaussian geostatistical models) and GMRFs (Gaussian Markov random fields). GGMs and GMRFs are two distinct approaches of modeling spatial phenomena which are used for different types of data. GGMs are used for point observations of a continuous process associated with exact locations and GMRFs are based on areal data whose underlying state space of the process is discrete. Section 1 contains the basic assumptions of GGMs and the parametric covariance models of GGMs. Section 2 deals with an overview of GMRFs and Section 3 reviews previous researches on the connections between the two models.

1.1 Gaussian geostatistical models (GGMs)

Geostatistical models are used in modeling point referenced spatial data which are collected at geographical locations over a fixed continuous space. Some common
examples of point referenced data include ozone measurements collected at fixed monitoring sites, elevation measured at various locations and nitrogen concentration from a soil sample.

In geostatistical models, it is often assumed that observations arise from a Gaussian spatial stochastic process. The Gaussian assumption is useful in that the specification of the Gaussian finite-dimensional distributions is simple and reasonable for many natural phenomena.

**Definition 1.1.1** Consider a spatial process \( \{Z(s), s \in D \subset R^d\} \) where \( s \) is a spatial location over a fixed domain \( D \), and \( R^d \) refers to a \( d \)-dimensional Euclidean space. The process \( Z(s) \) is defined as a Gaussian process if the joint distribution of \( Z = \{Z(s_1), \ldots, Z(s_n)\} \) is multivariate Gaussian for any collection of locations \( s_1, \ldots, s_n \), with each \( s_i \in R^D \).

The Gaussian process is completely specified by the mean function \( \mu(s) = E(Z(s)) \) and the covariance function \( C(s_i, s_j) = \text{cov}(Z(s_i), Z(s_j)) \). The mean function \( \mu(s) \) explains the global trend using explanatory variables, and the covariance function \( C(s_i, s_j) \) represents small scale spatial correlation of the process. The mean function can be chosen arbitrarily, but the covariance function must be positive definite, which is a sufficient and necessary condition for establishing consistent finite-dimensional distributions.

**Definition 1.1.2** The function \( C(\cdot, \cdot) \) is positive definite in \( R^d \), and hence a valid covariogram if
\[
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j C(s_i, s_j) > 0,
\]
for any choice of $s_1, \cdots, s_n$ and $a_1, \cdots, a_n$ such that $\sum_1^n a_i^2 > 0$.

GGMs usually assume stationarity which provides the replication of the process in space which is necessary for statistical estimation. If a process is stationary, observations from any area within the region can be used to make inference about the overall underlying structure. There are two different types of stationarity: strict stationarity and weak stationarity. Strict stationarity is more restrictive compared to weak stationarity, which implies that the joint probability distribution of the process depends on the relative positions of locations.

Definition 1.1.3 The spatial process $\{ Z(s), s \in D \}$ is said to be strictly stationary if for any given $n \geq 1$, any set of $n$ sites $\{ s_1, \cdots, s_n \}$, and any $h \in R^d$,

$$F(z(s_1), \cdots, z(s_n)) = F(z(s_1 + h), \cdots, z(s_n + h)),$$

where $F(\cdot)$ represents the joint distribution function, i.e. $F(z(s_1), \cdots, z(s_n)) = P(Z(s_1) \leq z(s_1) \cdots, Z(s_n) \leq z(s_n))$.

Since the strict stationarity condition is rarely satisfied in real data, a less restrictive condition is given by weak stationarity (also called second-order stationarity) which assumes that the mean and variance of the spatial process do not depend on the exact locations. That is, the mean is assumed constant and the variability depends only on the separation vector.

Definition 1.1.4 The spatial process $\{ Z(s), s \in D \}$ is said to be weakly stationary if the mean and covariance processes satisfy the following conditions,

$$(i) \ E(Z(s)) = \mu \ \forall s \in D$$
\[(ii) \quad \text{Cov}(Z(s_i), Z(s_j)) = C(s_i - s_j), \forall s_i, s_j \in D.\]

The function \(C(s_i - s_j)\) is called the covariogram, which is used to describe the spatial variability of the random field. If the second moments exist, the strict stationarity implies the weak stationarity.

The second moment structure of the weakly stationary random field does not depend on only the spatial distance, but also on the direction. If the function depends only on the Euclidean distance between the sites, the model is said to be \textit{isotropic}. If the relationship between distance and covariance differs depending on the direction of the difference between the locations, the spatial structure is \textit{anisotropic}.

\textbf{Definition 1.1.5} A stationary Gaussian field is called \textit{isotropic} if the covariance function depends only on the Euclidean distance between the sites. In other words,

\[C(s_i - s_j) = C(||s_i - s_j||), \forall s_i, s_j \in D,\]

where \(|| \cdot ||\) denotes Euclidean distance.

For most stationary and isotropic processes, it is assumed that the covariance functions can be represented in terms of three key parameters; the range, nugget and sill. The \textit{range} measures the distance beyond which the observations are uncorrelated, or at least approximately uncorrelated. The \textit{sill} is the variance of the process and the \textit{nugget} is a discontinuity at the origin due to microscale variations and measurement errors. The partial sill is the difference between the sill and nugget, which indicates the amount of the change of the covariance at a very short distance.

Using these parameters, isotropic covariance functions are defined as follows. Here,
is the Euclidean distance between sites, and the range, nugget and partial sill are represented by the parameters, \( \phi > 0 \), \( \tau^2 \geq 0 \) and \( \sigma^2 \geq 0 \), respectively.

1) Spherical Model

\[
C(h; \theta) = \begin{cases} 
0 & \text{if } h \geq \phi \\
\sigma^2(1 - \frac{3}{2}(h/\phi) + \frac{1}{2}(h/\phi)^3) & \text{if } 0 < h \leq \phi \\
\tau^2 + \sigma^2 & \text{if } h = 0
\end{cases}
\]

2) Exponential Model

\[
C(h) = \begin{cases} 
\sigma^2 \exp(-3h/\phi) & \text{if } h > 0 \\
\tau^2 + \sigma^2 & \text{if } h = 0
\end{cases}
\]

3) Powered exponential Model

\[
C(h) = \begin{cases} 
\sigma^2 \exp(-3|h/\phi|^p) & \text{if } h > 0 \\
\tau^2 + \sigma^2 & \text{if } h = 0
\end{cases}
\]

4) Gaussian Model

\[
C(h) = \begin{cases} 
\sigma^2 \exp(-3h^2/\phi^2) & \text{if } h > 0 \\
\tau^2 + \sigma^2 & \text{if } h = 0
\end{cases}
\]

5) Rational quadratic Model

\[
C(h) = \begin{cases} 
\sigma^2(1 - \frac{h^2}{1+\phi^2}) & \text{if } h > 0 \\
\tau^2 + \sigma^2 & \text{if } h = 0
\end{cases}
\]

6) Wave

\[
C(h) = \begin{cases} 
\sigma^2(\frac{\phi}{h} \sin(\frac{h}{\phi})) & \text{if } h > 0 \\
\tau^2 + \sigma^2 & \text{if } h = 0
\end{cases}
\]

7) Matérn Model

\[
C(h) = \begin{cases} 
\sigma^2 \frac{1}{2^{v-1} \Gamma(v)}(\frac{2\sqrt{\pi h}}{\phi})^v 2K_v(\frac{2\sqrt{\pi h}}{\phi}) & \text{if } h > 0 \\
\tau^2 + \sigma^2 & \text{if } h = 0
\end{cases}
\]
where $K_v$ is the modified Bessel function of the second kind with order $v > 0$ which controls the smoothness of the function. The Matérn covariance is widely used in a variety of spatial applications due to its flexibility. When $v = 1/2$, the Matérn covariance reduces to an exponential model and when $v \to \infty$; it approaches to a Gaussian model.

1.2 Gaussian Markov Random Fields (GMRFs)

GMRFs are widely used for areal data such as counts or spatial averages of a quantity over subregions of a larger study region. Areal data arises from the attributes associated with ZIP code data, census tract data or remotely sensed data observed at pixel levels. Examples of areal data include the number of deaths due to lung cancer in counties, the number of hospital admissions in a zip code, and the number of elk per square hectare.

A GMRF is a subfield of a Gaussian random field, which satisfies the Markov property,

$$P(Z(s_i) = z_i | \{ Z(s_j) = z_j : j \neq i \}) = P(Z(s_i) = z_i | \{ Z(s_j) = z_j : j \in \partial_i \}),$$

where $\partial_i$ is a neighbor of site $s_i$. The Markov property in space implies that $Z(s_i)$ depends on $Z(s_j)$ only if site $s_j$ is in the neighbor of site $s_i$.

As a special case of the Markov property, the GMRF represents the data at site $s_i$ as a linear combination of neighboring values, and is specified through the full
(a) Spherical

(b) Exponential

(c) Powered exponential with $p=1.5$

(d) Gaussian

(e) Rational quadratic

(f) Wave

(g) Matérn with $v=0.5$

(h) Matérn with $v=1.5$

Figure 1.1: Isotropic covariance functions
conditionals \{Z(s_i)|Z(s_j) : j \in \partial_i\},

\[ Z(s_i)|Z(s_j) : j \in \partial_i \sim N(\mu_i + \alpha \sum_{j=1}^{n} b_{ij}(z_j - \mu_j), \tau_i), \]

where \( \mu_i \) is the mean of \( Z(s_i) \) and \( \tau_i \) is the conditional variance of \( Z(s_i) \). The parameter \( b_{ij} \) is a weight that measures the degree to which \( Z(s_i) \) depends on \( Z(s_j) \). If site \( s_j \) is a neighborhood of site \( s_i \), then \( b_{ij} \) is a non-zero value and otherwise \( b_{ij} = 0 \). By convention, we set \( b_{ii} = 0 \) so that no region is a neighborhood of itself. The parameter \( \alpha \) is a spatial dependency parameter that controls the strength of associations between the components of \( Z(s) \). When \( \alpha = 0 \), we obtain an independent model, but it should not be interpreted as the correlation parameter (Wall, 2004).

Once the conditional distribution is given, the joint distribution of \( Z(s_1), \ldots, Z(s_n) \) can be derived using Brook’s Lemma (Brook, 1964).

**Brook’s lemma** Let \( P(z_1, \ldots, z_n) \) be the density for \((z_1, \ldots, z_n) \in \mathbb{R}^d \) then

\[
P(z_1, \ldots, z_n) = \frac{P(z_1|z_2, \ldots, z_n)}{P(z_2|z_1, z_3, \ldots, z_n)} \cdot \frac{P(z_2|z_{10}, z_3, \ldots, z_n)}{P(z_{10}|z_2, z_3, \ldots, z_n)} \ldots \frac{P(z_n|z_{10}, \ldots, z_{n-1,0})}{P(z_{n0}|z_{10}, \ldots, z_{n-1,0})} \cdot P(z_{10}, \ldots, z_{n0}),
\]

where \((z_{10}, \ldots, z_{n0})\) is any fixed point in the support of \( P(z_1, \ldots, z_n) \).

Brook’s Lemma states that if the positivity condition is satisfied, the joint distribution is uniquely determined by the full conditionals. By Brook’s Lemma and Hammersley-Clifford Theorem (Besag 1974), the joint distribution \( \mathbf{Z} = (Z(s_1), \ldots, Z(s_n))^T \) is uniquely determined as,

\[ \mathbf{Z} \sim N(\mathbf{\mu}, (\mathbf{I} - \alpha \mathbf{B})^{-1} \mathbf{M}) \]
where \( \mathbf{\mu} \) is an \( n \times 1 \) vector of means with elements \( \mu_i \), and \( \mathbf{M} \) is an \( n \times n \) diagonal matrix with elements \( \tau_i \).

The matrix \( \mathbf{B} \) is an \( n \times n \) neighborhood matrix with elements \( b_{ij} \), which satisfy

\[
b_{ii} = 0, \quad b_{ij} \tau_j = b_{ji} \tau_i \quad \text{and} \quad b_{ij} = 0,
\]

unless \( i \) and \( j \) are neighbors. The requirement that \((\mathbf{I} - \alpha \mathbf{B})^{-1} \mathbf{M}\) should be symmetric yields the condition, \( b_{ij} \tau_j = b_{ji} \tau_i \) for all \( i, j \). The matrix \( \mathbf{B} \) allows the modeling of different patterns of spatial association by the specification of different neighborhood structures which are determined by neighborhoods and weights to the neighboring values.

In the case of a regular lattice, some popular neighborhoods are defined as orders depending on the touching borders between grid cells (Figure 1.2). For the first order neighborhood (Figure 1.2.(a)), \( P(Z_{ij}|Z_{kl}; kl \neq ij) \) depends only upon \( Z_{i-1j}, Z_{i+1j}, Z_{ij-1} \) and \( Z_{ij+1} \) where \((ij)\) indicate the location of site at the \( i\)th row and \( j\)th column. For irregular lattices, two sites are considered neighbors if they are within some specified distance of one another.

The weights of neighborhoods are defined in several ways. One simple way is to
use a binary weight which gives 1 if two sites are neighbors and 0 otherwise. A scaled weight is to scale the weights by its row sum, $\tilde{B} = Diag(1/b_{i+})B$ where $b_{i+}$ is the sum of the $i$th row. Weights also can be defined as a function of distance between two sites such as $b_{ij} = \frac{1}{||s_i - s_j||^m}$ where $m$ is a parameter which controls the degree of dependency between two locations.

For a well-defined model, we require $(I - \alpha B)$ to be non-singular which is determined by the value of the spatial dependency parameter $\alpha$. For the non-scaled neighborhood matrix, the range of $\alpha$ for a non-singular matrix is $(1/\lambda_{\text{min}}, 1/\lambda_{\text{max}})$, where $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ are the minimum and maximum eigenvalues of the neighborhood matrix $B$, respectively. For the scaled neighborhood matrix $\tilde{B} = Diag(1/b_{i+})B$, the range of $\alpha$ is simplified to $(-1, 1)$. When $\alpha = 1$, the covariance becomes singular and yields an improper joint distribution, but it has still the proper full conditional distributions. The distribution with $\alpha = 1$ is known as an intrinsic GMRF (Besag et al. 1991; Besag and Kooperberg, 1995) which has been extensively used in spatial statistics as a prior for random effects (Sun et al. 1999; Banerjee and Carlin, 2003) in Bayesian hierarchical models.

### 1.3 Relations between GGMs and GMRFs

Although GGMs and GMRFs have been developed in parallel for modeling different types of data, several researchers have investigated relations between GGMs and GMRFs based on the approximations of GGMs by GMRFs and GMRFs by GGMs. Besag (1981), Griffith and Csillag(1993), and Hrafneksson and Cressie (2003) inves-
tigated the approximations of GMRFs by GGMs, while Rue and Tjelmeland (2002) researched the approximations of GGMs by GMRFs. In particular, approximations of GGMs by GMRFs have been motivated by the computational advantages of GMRFs over Gaussian fields. GMRFs can be computed and simulated very quickly relative to Gaussian fields, and GMRFs are also computationally convenient for Gibbs sampling used in Bayesian inference due to the conditional form. From these advantages, GMRFs have been regarded as being more convenient than GGMs in dealing with large data sets.

Besag (1981) showed that correlation functions $\rho_{st}$ of GMRFs with the first order neighborhood defined on an infinite rectangular lattice, could be approximately represented in terms of a modified Bessel function of the second kind with order zero. Besag considered the following recurrence equations,

$$
\rho_{st} = \delta_{st} + \beta(\rho_{s-1,t} + \rho_{s+1,t} + \rho_{s,t-1} + \rho_{s,t+1}),
$$

where $\rho_{st}$ is the correlation of site $(s,t)$ $s, t = 0, \pm 1, \pm 2, \cdots$, $\delta_{st} = 1$ if $s = t = 0$ and $\delta_{st} = 0$ otherwise, and $|\beta| < \frac{1}{4}$. The inverse Fourier transform of $\rho_{st}$ can be expressed as,

$$
\rho_{st} = \frac{1}{\pi^2} \int_0^\pi \int_0^\pi \frac{\cos(ws)\cos(\eta t)dw \, d\eta}{1 - 2\beta\cos(w) - 2\beta\cos(\eta)}.
$$

From the above equation, Besag derived the modified Bessel function,

$$
\rho_{st} = \begin{cases} 
1 & \text{if} (s,t) = (0,0) \\
\frac{1}{2\pi^2\beta\rho_{00}} K_0((\frac{1-4\beta}{\beta})^{1/2}\sqrt{(s^2 + t^2)}) & \text{if} (s,t) \neq (0,0)
\end{cases}
$$

where $K_0$ denotes the modified Bessel function of the second kind with order zero. This approximation equation shows that the correlation of the GMRF can be closely approximated by a monotonic decreasing function of distance. 

11
Griffith and Csillag (1993) investigated the approximations of GMRFs using several geostatistical models. Based on the theoretical autocorrelation of the GMRF $\rho(s, t)$ over an infinite lattice, Griffith and Csillag estimated the parameters of a GGM via the minimization of the squared differences between correlations of GMRFs and GGMs for lags $(0,0)$ through $(10,10)$. The best fitting model of the GGM was an exponential model which is a monotonic decreasing function of distance. However, for a spatial dependency parameter close to its upper limit, even the exponential model did not fit well to the GMRF.

Hrafnkelsson and Cressie (2003) calibrated a Matérn covariance model to a specific class of the GMRF. Hrafnkelsson and Cressie established some parametric functions which represent the relationships between the parameters of a Matérn geostatistical model and those of a specific class of the GMRF. The parameters of a geostatistical model were estimated by non-linear least squares. The simulation results showed that the Matérn model with scale parameter 0.5 fitted well to the GMRF, and the fitting became gradually poorer as the spatial parameter approached its boundary.

While previous research focused on the approximation of GMRFs by GGMs, Rue and Tjelmeland (2002) researched the approximation of GGMs by GMRFs. They defined the GMRF with a square neighborhood on a torus $T_n$ instead of a finite lattice for a stationary process, and parametrized the precision matrix of the GMRF. For approximations of GGMs by GMRFs, they suggested two methods, one using Kullback-Leibler (KL) discrepancy and the other one by matching the correlation methods. The KL method estimates the parameters of the GMRF by minimizing the
KL distance between the GGM and the GMRF,

$$KL(f, \tilde{f}) = \int f(x) \log \left( \frac{f(x)}{\tilde{f}(x)} \right) dx = -\frac{1}{2} \sum \sum (\log \lambda_{ij} q_{ij}(\theta) - \lambda_{ij} q_{ij}(\theta) + 1)$$

where $f$ is a zero mean stationary GGM, $\tilde{f}$ is a zero mean stationary GMRF. $\lambda_{ij}$ and $q_{ij}$ are the eigenvalues of the covariance of the GGM and the precision matrix of the GMRF, respectively. The matching correlation method minimizes the weighted differences between the correlations of the GMRF $\rho(\theta)$ and the GGM $\rho$,

$$||\rho - \rho(\theta)||^2 = \sum (\rho_{ij} - \rho_{ij}(\theta))^2 w_{ij}$$

The weight of the differences $w_{ij}$ is defined as,

$$w_{ij} \propto \begin{cases} 1 & \text{if } d(i, j) = 0 \\ \frac{1}{2\pi d(i,j)} & \text{o.w.} \end{cases}$$

where $d(i, j)$ is the Euclidean distance on the torus between $(i, j)$ and $(0, 0)$. Rue and Tjelmelnd approximated several geostatistical models: exponential, powered exponential, Matérn and Gaussian covariance models by the GMRF defined on a large torus. Using the KL method in the approximation of GMRFs by GGMs, the fitting performed well only when the range was within the neighborhood, but it gave a poor fit when the range was outside the neighborhood. On the other hand, using the matching correlation method showed a good fitting both inside and outside of the range. Among the several geostatistical models, the exponential covariance was fitted quite accurately, and the powered exponential model and Matérn covariance model were also good. However, the Gaussian covariance function was more difficult to fit than the other covariance functions.
Chapter 2

Associations between GMRFs and GGMs

In this chapter, we investigate the association between GMRFs and GGMs based on new metrics using spectral densities on a finite regular lattice. Spectral densities and covariance functions are related through Fourier transforms. While spectral densities and covariance functions are both used to explore the second order properties of a spatial process, they provide different ways of analyzing the process. Also, there are several advantages of analysis using the spectral density over using the covariance function. Spectral densities evaluated at different frequencies are approximately independent, and hence the inference based on the spectral density is usually easier to derive than the covariance function. In addition, the spectral density function can highlight subtle differences in the second order structure more easily than the covariance function. Considering these advantages of the spectral approaches, we introduce approximation approaches using spectral densities.
Section 1 contains the introduction to spectral densities. New approximation methods using spectral densities are proposed in Section 2. Section 3 presents approximation methods using covariances in the spatial domain and a comparison criterion of these methodologies. In Section 4, we present the empirical results of approximations of GGMs by GMRFs and GMRFs by GGMs. In Section 5, we apply the methods to a dataset on PM\textsubscript{2.5} in California to study the relation between GMRFs and GGMs.

### 2.1 Introduction to spectral densities

Let $Z(s)$ denote a stationary spatial process measured at location $s \in D$ where $D \subset \mathbb{R}^2$. The process $Z(s)$ can be represented as a process by Fourier-Stieltjes integrals,

$$Z(s) = \int_{\mathbb{R}^2} \exp(iw^Ts) dX(w),$$

where \{\(X(w), w = (w_1, w_2) \in \mathbb{R}^2\)\} is an orthogonal process with zero mean and orthogonal increments, i.e.

$$E(X(w)) = 0$$

$$E(X(w_3) - X(w_2), X(w_1) - X(w_0)) = 0,$$

for disjoint rectangles $(w_3, w_2)$ and $(w_1, w_0)$. The process $X$ is called the spectral process associated with a stationary process $Z$ (Yaglom, 1987; Priestley, 1991).

Using the spectral process $X$, we define a monotonically increasing function $F(w)$ as

$$E(|dX(w)|^2) = dF(w),$$
where \( dX(w) = \lim_{h \to 0} \frac{X(w+h)-X(w)}{h} \) and \( f(w) = \lim_{h \to 0} \frac{F(w+h)-F(w)}{h} \). The function \( F(w) \) represents the contribution to the variance accounted for by frequencies in the range \((0, w)\). The derivative of \( F(w) \) denoted by \( f(w) \), is called a spectral density function.

The covariance function \( C(h) \) of the weakly stationary process and the spectral density function \( f(w) \) form a Fourier transform pair,

\[
C(h) = \int_{R^2} \exp(iw^T h)f(w)dw \\
\]
\[
f(w) = \frac{1}{(2\pi)^2} \int_{R^2} C(h)\exp(-iw^T h)dh,
\]
where \( h = ||s_1 - s_2||, s_1, s_2 \in D \subset R^2 \). If a stochastic process is defined on a lattice \( \Lambda \), integration in the expression for \( f(w) \) is replaced by summation, and the frequency domain is restricted to \( A = (-\pi, \pi]^2 \),

\[
C(h) = \int_{A} \exp(iw^T h)f(w)dw \\
\]
\[
f(w) = \frac{1}{(2\pi)^2} \sum_{h \in \Lambda} C(h)\exp(-iw^T h).
\]

When the underlying process is continuous, but observed only on an integer lattice \( \Lambda \) with an interval between neighboring observations \( \delta \), we can not distinguish between the frequency components \( w \) and \( w + 2\pi l/\delta \), \( l = (l_1, l_2), l_1, l_2 = \pm 1, \cdots \). Hence, the power in the sampled process from the set of frequencies \( w \in (-\pi/\delta, \pi/\delta) \) is the accumulation of power in the original process from all sets of frequencies \( w + (2\pi l/\delta) \).

The spectral density of the sampled process is defined as follows,

\[
f_\delta(w) = \sum_{Q \in Z^2} f(w + \frac{2\pi l}{\delta}),
\]
for \( w \in (-\pi, \pi]^2 \).

For a nonstationary process, the process can also be represented by sinusoidal functions, but the decomposed process is no longer orthogonal but correlated. If a process is defined on a lattice \( \Lambda \), the Fourier transform pairs are

\[
C(z_1, z_2) = \int_{A^2} \exp(i(z_1 w_1 - z_2 w_2)) f(w_1, w_2) dw_1 dw_2
\]

\[
f(w_1, w_2) = \frac{1}{(2\pi)^2} \sum_{z_1, z_2} C(z_1, z_2) \exp(-i(z_1 w_1 - z_2 w_2)),
\]

where \( z_1, z_2 \in \Lambda \) and \( w_1, w_2 \in (-\pi, \pi]^2 \). For the special case of \( w_1 = w_2 \), the spectral density of a nonstationary process becomes equivalent to that of a stationary process.

### 2.2 Approximation methods using spectral densities

We assume that we observe a spatial process on a lattice. We define a GGM and a GMRF on a \( r \times c \) finite lattice \( \Lambda \) with an interval \( \delta \) between neighboring points. The GGM represents a weakly stationary and isotropic process, and the GMRF is a nonstationary process due to the finiteness of the lattice and the edge effect. For the GGM, the underlying process is continuous, but observed only on a discrete lattice \( \Lambda \). Hence, the spectral density of the sampled GGMs \( f_\delta \) is defined as,

\[
g_\delta(w) = \sum_{l \in Z^2} f \left( w + \frac{2\pi l}{\delta} \right),
\]

where \( g(w) = \frac{1}{(2\pi)^2} \int_{R^2} C(h) \exp(-iw^T h) dh \)

for \( w \in (-\pi, \pi]^2 \) and \( Z = \{0, \pm 1, \pm 2, \cdots\} \).
The nonstationary process of the GMRF defined on a lattice is not represented as a function with orthogonal increments; rather, it has correlated increments. The spectral density of a nonstationary process is

\[ f(w_1, w_2) = \frac{1}{(2\pi)^4} \sum_{z_1, z_2} C(z_1, z_2) \exp(-i(z_1 w_1 - z_2 w_2)), \]

where \( z_1, z_2 \in \Lambda \) and \( w_1, w_2 \in (\pi, \pi]^2 \).

Using a spectral framework, we suggest two approximation criteria: the KL discrepancy between spectral densities and the chi-squared distance (CSD) between spectral densities. We use these criteria to approximate the spectral density of a stationary geostatistical process with the spectral density of a nonstationary GMRF process or vice versa by minimizing the differences of the spectral densities of stationary and nonstationary processes when \( w_1 = w_2 = w \). This corresponds to the stationary component of the GMRF, and we basically ignore the nonstationary contribution due to the edge effect.

The KL discrepancy between two spectral densities is defined as,

\[ KL(f, \tilde{f}(\theta)) = \int_{-\pi}^{\pi} f(w) \log \frac{f(w)}{\tilde{f}(w; \theta)} dw, \]

where \( f \) and \( \tilde{f}(\theta) \) are the spectral densities of the true and approximated model, respectively. Through the numerical minimization of the criterion, we estimate the parameters of the approximated model,

\[ \hat{\theta} = \arg\min_{\theta} KL(f, \tilde{f}(\theta)). \]

The CSD between two spectral densities is expressed as,

\[ G(f, \tilde{f}(\theta)) = \int_{-\pi}^{\pi} \frac{(f(w) - \tilde{f}(w; \theta))^2}{f(w)} dw. \]
where \( f \) and \( \tilde{f}(\theta) \) are the spectral densities of the true and approximated model, respectively. The best fitting model is obtained using the optimal parameters estimated by minimizing the CSD between two spectral densities,

\[
\hat{\theta} = \arg\min_{\theta} G(f, \tilde{f}(\theta)).
\]

### 2.3 Comparisons of spectral density approximation methods with covariance function approximation methods

We compare our approximation approaches using spectral densities in the frequency domain to those using covariance functions in the spatial domain. The KL approach in the spatial domain is to estimate the parameters of the approximated model by minimizing the KL discrepancy between the distribution functions of two models,

\[
\hat{\theta} = \arg\min_{\theta} KL(g, \tilde{g}(\theta)) = \arg\min_{\theta} \int_{\mathbb{R}^2} g(x) \log \frac{g(x)}{\tilde{g}(x; \theta)} dx,
\]

where \( g(x) \) and \( \tilde{g}(x; \theta) \) are probability densities of the true and approximation model, respectively. The squared relative differences (SRD) approach in the spatial domain is to minimize the squared relative differences between the covariances of two models,

\[
\hat{\theta} = \arg\min_{\theta} \int_{\mathbb{R}^2} \left( \frac{C(s_1, s_2) - \tilde{C}(s_1, s_2; \theta)}{C(s_1, s_2)} \right)^2 ds_1 ds_2,
\]

where \( C(s_1, s_2) \) and \( \tilde{C}(s_1, s_2; \theta) \) are the true and approximated covariances at \( s_1, s_2 \in \mathbb{R}^2 \). We compare the four approximation methods based on the mean squared pre-
diction error (MSPE). The MSPE is defined as,

\[ MSPE(p(Z_{-i}; s_i); Z(s_i)) = E(Z(s_i) - p(Z_{-i}; s_i))^2, \]

where \( Z(s_i) \) denotes the true value at site \( s_i \) and \( p(Z_{-i}; s_i) \) is the predicted value at site \( s_i \) given other data \( Z_{-i} = (Z(s_1), \cdots, Z(s_{i-1}), Z(s_{i+1}), \cdots, Z(s_N)) \). We express the MSPE as a function of the covariances of GGMs and GMRFs. We assume a process with mean zero. In the approximations of GGMs by GMRFs, the predicted value at site \( s_i \) given other values \( p(Z_{-i}; s_i) \) is expressed as \(-\sum_{j=1, j\neq i}^{N} Q_{ij} Q^{-1}_{ii} Z(s_j)\) where \( Q_{ij} \) is the element of the precision matrix of the approximated GMRF. Then, the MSPE of a GGM is represented as

\[ MSPE_{GGM} = E(Z(s_i) - \left(-\sum_{j=1, j\neq i}^{N} Q_{ij} Q^{-1}_{ii} Z(s_j)\right))^2 = E(C'Z)^2 = C'\Sigma_{GGM}C, \]

where \( C \) is an \( N \times 1 \) vector with \( c_i = 1 \) and \( c_j = Q_{ij} Q^{-1}_{ii}, Z = (Z(s_1), \cdots, Z(s_N))' \) and \( \Sigma_{GGM} \) is the covariance matrix of a GGM. In the approximation of GMRFs by GGMs, \( p(Z_{-i}; s_i) \) is expressed as \( \sigma_{AB}\sigma^{-1}_{BB}Z_{-i} \) where \( \sigma_{AB} \) is an \( 1 \times (N - 1) \) matrix with the \( j \)th element, \( \text{cov}(Z(s_i), Z(s_j)) \), and \( \sigma_{BB} \) is a \( (N - 1) \times (N - 1) \) matrix with the element, \( \text{cov}(Z(s_j), Z(s_k)) \) where \( j, k = 1, \cdots, i - 1, i + 1, \cdots, N \). The MSPE of a GMRF is

\[ MSPE_{GMRF} = E(Z(s_i) - \sigma_{AB}\sigma^{-1}_{BB}Z_{-i})^2 = E(D'Z)^2 = D'\Sigma_{GMRF}D, \]

where \( D \) is an \( N \times 1 \) vector with \( d_i = 1, d_j = -\sigma_{AB}\sigma^{-1}_{BB} j \) for \( j < i \), \( d_j = -\sigma_{AB}\sigma^{-1}_{BB} j - 1 \) for \( j > i \) and \( \Sigma_{GMRF} \) is the covariance matrix of a GMRF.
2.4 An empirical study

In this Section, an empirical study is conducted to compare the spectral density approximation approaches to the covariance approximation methods. Let us define a Gaussian Geostatistical process (GGP) and a GMRF on a $10 \times 10$ lattice. To reduce the nonstationarity problem of the GMRF which arises due to the edge effect, we extend the lattice to a $12 \times 12$ grid to cover a larger domain, and we consider only the covariance of the original lattice within the covariance of the extended lattice. We consider a Matérn model with smoothness parameter 0.5 and nugget 0 for the GGP and a binary scaled weight neighborhood structure for the GMRF. The neighborhood of the GMRF is defined depending on the touching border and the 2nd, 3rd and 5th neighborhood orders are considered. The 5th neighborhood is displayed in Figure 2.1. This study is composed of two steps, the estimation of parameters and the prediction of data. In the estimation step, the parameters of the approximated models are estimated by minimizing the criteria suggested in the previous Section: CSDS (chi-squared differences between spectral densities), KLS (Kullback-Leibler discrepancy between spectral densities), SRDC (squared relative differences between covariance functions), and KLP (Kullback-Leibler discrepancy between probability density functions). We estimate the range ($\phi$) and sill ($\sigma^2$) parameters of the Matérn model and the spatial dependency ($\alpha$) and scale parameters of the GMRF using the R software (http://www.r-project.org/). Based on the approximated model, we obtain the MSPE.

Table 2.1 shows the estimated parameters of a GGM, when we approximate a
Figure 2.1: the 5th order neighborhood

Table 2.1: Approximations of GMRFs with the 5th neighborhood order and various values of $\alpha$ using GGMs. This table presents the estimated parameters of the GGM.

<table>
<thead>
<tr>
<th>method</th>
<th>$\alpha=0.9$</th>
<th>$\alpha=0.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sill (SE)</td>
<td>range (SE)</td>
</tr>
<tr>
<td>CSDS</td>
<td>2.56 (0.68)</td>
<td>0.55 (0.31)</td>
</tr>
<tr>
<td>KLS</td>
<td>2.57 (0.43)</td>
<td>0.46 (0.17)</td>
</tr>
<tr>
<td>SRDC</td>
<td>0.22 (0.004)</td>
<td>3.68 (0.043)</td>
</tr>
<tr>
<td>KLP</td>
<td>1.02 (0.14)</td>
<td>0.38 (0.17)</td>
</tr>
</tbody>
</table>

$\alpha=0.98$

<table>
<thead>
<tr>
<th>method</th>
<th>sill (SE)</th>
<th>range (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSDS</td>
<td>1.99 (0.60)</td>
<td>2.60 (0.53)</td>
</tr>
<tr>
<td>KLS</td>
<td>2.72 (0.47)</td>
<td>0.54 (0.19)</td>
</tr>
<tr>
<td>SRDC</td>
<td>0.60 (0.01)</td>
<td>10.69 (0.28)</td>
</tr>
<tr>
<td>KLP</td>
<td>1.03 (0.15)</td>
<td>0.41 (0.17)</td>
</tr>
</tbody>
</table>

GMRF (with different values for $\alpha$ and the 5th order neighborhood) using a GGM. As $\alpha$ increases, the estimated range parameter increases.

Using the estimated parameters of the GGM, we predict the observations generated from the GMRF, and then we obtain the mean squared prediction error (MSPE) (see Table 2.2). As $\alpha$ increases, the MSPE increases using all criteria. The smallest MSPE is obtained with the KLS, and the SRDC gives the largest MSPE.

In Table 2.3 we present the estimation results for GGMs, in the approximation.
Table 2.2: Approximations of GMRFs with the 5th order neighborhood and various values of $\alpha$ using GGMs. Predicting GMRFs using GGMs

<table>
<thead>
<tr>
<th>method</th>
<th>$\alpha=0.9$</th>
<th>$\alpha=0.95$</th>
<th>$\alpha=0.98$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSDS</td>
<td>1.16</td>
<td>1.22</td>
<td>1.38</td>
</tr>
<tr>
<td>KLS</td>
<td>1.16</td>
<td>1.22</td>
<td>1.31</td>
</tr>
<tr>
<td>SRDC</td>
<td>1.34</td>
<td>1.37</td>
<td>1.41</td>
</tr>
<tr>
<td>KLP</td>
<td>1.17</td>
<td>1.26</td>
<td>1.41</td>
</tr>
</tbody>
</table>

Table 2.3: Approximations of GMRFs with $\alpha=0.95$ and various orders of neighborhood using GGMs. This table shows the estimated parameters of a GGM.

<table>
<thead>
<tr>
<th>method</th>
<th>2nd order</th>
<th>3rd order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sill (SE)</td>
<td>range (SE)</td>
</tr>
<tr>
<td>CSDS</td>
<td>3.59 (0.76)</td>
<td>1.13 (0.33)</td>
</tr>
<tr>
<td>KLS</td>
<td>3.63 (0.76)</td>
<td>0.92 (0.27)</td>
</tr>
<tr>
<td>SRDC</td>
<td>0.65 (0.013)</td>
<td>2.67 (0.025)</td>
</tr>
<tr>
<td>KLP</td>
<td>1.42 (0.28)</td>
<td>0.94 (0.30)</td>
</tr>
</tbody>
</table>

of a GMRF with $\alpha=0.95$ and various orders for the neighborhood structure using a GGM. For the CSDS, KLS and KLP, the estimated sill and range parameters tend to decrease as the neighborhood order increases. However, the SRDC gives an estimated range parameter that increases with the neighborhood order. Table 2.4 shows the prediction results using the estimated parameters. As the neighborhood order increases, the MSPE decreases and the SRDC gives the largest MSPE for all neighborhood orders.

Table 2.5 shows the parameter estimates of a GMRF corresponding to a GGM.
Table 2.4: Approximations of GMRFs with $\alpha=0.95$ and various orders of neighborhood by GGMs. Predicting a GMRF using a GGM.

<table>
<thead>
<tr>
<th>method</th>
<th>2nd order</th>
<th>3rd order</th>
<th>5th order</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSDS</td>
<td>1.34</td>
<td>1.27</td>
<td>1.22</td>
</tr>
<tr>
<td>KLS</td>
<td>1.32</td>
<td>1.22</td>
<td>1.22</td>
</tr>
<tr>
<td>SRDC</td>
<td>1.40</td>
<td>1.36</td>
<td>1.37</td>
</tr>
<tr>
<td>KLP</td>
<td>1.32</td>
<td>1.22</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Table 2.5: Approximations of a GGM with sill=1 and with various values of the range parameter using a GMRF with a 5th order neighborhood. Parameter estimates of the GMRF.

<table>
<thead>
<tr>
<th>method</th>
<th>range=2 (range&lt;MDN)</th>
<th>range=3 (range=MDN)</th>
<th>range=5 (range&gt;MDN)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$ (SE)</td>
<td>$\tau$ (SE)</td>
<td>$\alpha$ (SE)</td>
</tr>
<tr>
<td>CSDS</td>
<td>0.99 (0.092)</td>
<td>0.175 (1.054)</td>
<td>0.99 (0.016)</td>
</tr>
<tr>
<td>KLS</td>
<td>0.99 (0.009)</td>
<td>0.153 (0.025)</td>
<td>0.99 (0.002)</td>
</tr>
<tr>
<td>SRDC</td>
<td>0.81 (0.001)</td>
<td>2.818 (0.043)</td>
<td>0.90 (0.001)</td>
</tr>
<tr>
<td>KLP</td>
<td>0.98 (0.019)</td>
<td>0.391 (0.055)</td>
<td>0.99 (0.009)</td>
</tr>
</tbody>
</table>

with sill=1 and with various values for the range parameter. The estimated $\alpha$ is always close to one. This might be a justification for the popular intrinsic GMRF models. The MSPE becomes smaller as the range parameter increases (Table 2.6).

2.5 Application

In this section, we apply the approximation methods described in the previous sections to aggregated PM$_{2.5}$ data at the county level in California. Particulate matter
Table 2.6: Approximations of a GGM with sill = 1 and with various values of the range parameter using a GMRF with a 5th order neighborhood. Predicting a GMRF.

<table>
<thead>
<tr>
<th>method</th>
<th>range=2</th>
<th>range=3</th>
<th>range=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSDS</td>
<td>0.77</td>
<td>0.65</td>
<td>0.50</td>
</tr>
<tr>
<td>KLS</td>
<td>0.77</td>
<td>0.65</td>
<td>0.50</td>
</tr>
<tr>
<td>SRDC</td>
<td>0.78</td>
<td>0.66</td>
<td>0.50</td>
</tr>
<tr>
<td>KLC</td>
<td>0.77</td>
<td>0.65</td>
<td>0.50</td>
</tr>
</tbody>
</table>

(PM) is the general term used for a complex mixture of solid particles and liquid droplets suspended in the air. It is an air pollutant that represents a broad class of chemically diverse particles whose size ranges from 0.005 \( \mu m \) to 100 \( \mu m \) in diameter. PM\(_{2.5}\), also known as fine PM, includes the fine particles that are less than or equal to 2.5 \( \mu m \) in diameter. Our PM\(_{2.5}\) observations are point-referenced data collected at monitoring stations. This type of data can be aggregated to a level of interest. For example, in the study of associations between PM\(_{2.5}\) and adverse health effects, health outcomes are generally collected over geographic regions and point-referenced PM\(_{2.5}\) values are aggregated to the level of health outcomes (Fuentes et al., 2005). Let \( Y(s) \) be a continuous Gaussian process with the PM\(_{2.5}\) values, with mean function \( \mu(s) \) and covariance function \( c(s_i, s_j) \) for \( s_i, s_j \subset D \subset R^d \) and let \( |B| = \int_B ds \) denote the area of county \( B \). The averaged process \( Y(B) = \frac{1}{|B|} \int_B Y(s)ds \) has a multivariate normal distribution with the mean function \( \mu(B) \) and the covariance function \( \Sigma(B_i, B_j) \),

\[
\mu(B) = E(Y(B)) = |B|^{-1} \int_B \mu(s)ds
\]

\[
\Sigma(B_i, B_j) = cov(Y(B_i), Y(B_j)) = |B_i|^{-1}|B_j|^{-1} \int_{B_i} \int_{B_j} c(||s_i - s_j||)ds_i ds_j.
\]

To obtain the aggregated PM\(_{2.5}\) data, \( Y(B) \), we predict PM\(_{2.5}\) at 20 locations within each county and then we get the average of the predicted values. Figure 2.2 displays
the county level PM$_{2.5}$ values, showing high concentrations in the Southern California area.

Table 2.7 is the parameter estimates of a GMRF using the first and second order neighborhoods. The estimated $\alpha$ is close to 1 for all methods and the estimated $\tau$ increases as the neighborhood order increases. After estimating the parameters of the GMRF, the performance of the different methods are evaluated through the MSPE (Table 2.8). The smallest MSPE is obtained by the KLS method. Comparing the MSPE of different neighborhood orders, the first order neighborhood gives smaller MSPE than the second order neighborhood.
Table 2.7: Parameter estimates of the GMRF using the first and second order neighborhoods

<table>
<thead>
<tr>
<th>method</th>
<th>first order neighborhood</th>
<th>second order neighborhood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$ (SE)</td>
<td>$\tau$ (SE)</td>
</tr>
<tr>
<td>CSDS</td>
<td>0.99 (0.005)</td>
<td>4.13 (3.70)</td>
</tr>
<tr>
<td>KLS</td>
<td>0.99 (0.001)</td>
<td>3.73 (1.02)</td>
</tr>
<tr>
<td>SRDC</td>
<td>0.99 (0.04)</td>
<td>1.25 (0.86)</td>
</tr>
<tr>
<td>KLP</td>
<td>0.99 (0.003)</td>
<td>5.68 (0.22)</td>
</tr>
</tbody>
</table>

Table 2.8: MSPE in the approximations of aggregated PM$_{2.5}$ using GMRFs with a first and second order neighborhood structures.

<table>
<thead>
<tr>
<th>method</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>first order neighborhood</td>
</tr>
<tr>
<td>CSIDS</td>
<td>0.664</td>
</tr>
<tr>
<td>KLS</td>
<td>0.663</td>
</tr>
<tr>
<td>SRDC</td>
<td>0.668</td>
</tr>
<tr>
<td>KLP</td>
<td>0.668</td>
</tr>
</tbody>
</table>
In this chapter, we consider modeling the association between fine particulate matter (PM) and mortality. PM data are available at point level and mortality data are obtained at county-level. To solve the mismatch of support areas between two variables, we aggregate PM to the county level and model using a GMRF rather than using an averaged continuous process due to computational efficiency. We introduce some background to the study in Section 3.1. In Section 3.2, we describe the data used in this study. Section 3.3 presents the statistical model which estimates the impacts of fine PM on mortality. Section 3.4 shows the spatial modeling of PM$_{2.5}$. The results of this study are presented in Section 3.5.

3.1 Background

PM represents a complex mixture of organic and inorganic substances, which vary in size, composition and origin. PM is usually classified by its aerodynamic
diameter because it is related to the deposition within the respiratory system and is also associated with the chemical composition and sources of particles. In particular, particles of special concern are smaller particles, PM$_{10}$ and PM$_{2.5}$. PM$_{10}$ refers to particles less than or equal to 10 $\mu m$ in diameter, and PM$_{2.5}$ which is also known as fine PM, includes the fine particles that are less than or equal to 2.5 $\mu m$ in diameter. These smaller particles cause most of the adverse health effects because of their ability to penetrate deeply into the lungs. Exposure to smaller particles is associated with a variety of serious health effects such as respiratory and cardiovascular problems, increased hospital admissions, and even premature death. In particular, PM$_{2.5}$ is considered to be more serious than PM$_{10}$ due to the smaller size. PM$_{2.5}$ is mainly composed of several chemical components such as carbon, sulfate, nitrate and crustal materials, which varies by time, location, and weather conditions.

The effects of PM on health is a concern in public health, and numerous epidemiological studies have been conducted to investigate associations between PM and adverse health effects in a variety of U.S. cities. The health effects of air pollution have been investigated based on two types of epidemiologic studies, time series studies and cohort studies. Time series studies investigate the association between short-term exposure to PM and human health on either the same day or proceeding days (e.g, Schwartz 1994; Dockery et al. 1992; Dominici et al. 2000). On the other hand, cohort studies are used for assessing associations between longer-term PM exposure and public health (e.g, Pope et al. 1995; Dockery et al. 1993) over several years. The findings from these two types of study are complementary, which show acute and chronic air pollution effects on health, respectively.
Due to the lack of availability of data collection for cohort studies, time series studies have been done showing the effects of air pollution on human health. Many studies have identified an association between changes in the concentration of ambient PM and the number of deaths, as well as other harmful human effects (e.g., Bates, Baker-Anderson, and Sizto 1990; Dockery, Schwartz, and Spengler, 1992, Ostro et al., 1991; Schwartz 1994; Pope, Dockery, and Schwartz, 1995a, American Thoracic Society and Bascom 1996a, 1996b). However, the work by Smith et al. (2000) on fine particles, \( PM_{2.5} \) (< 2.5\( \mu \)m in diameter), provided evidence to the contrary, suggesting that more studies are needed to understand the association between fine particles and mortality. Clyde (2000) reanalyzed data from Birmingham (Alabama) using Bayesian Model Averaging with carefully selected prior distributions to assess the impact of model uncertainty on estimates of relative risk due to changes in \( PM_{10} \) (< 10\( \mu \)m in diameter). Özkaynak and Thurston (1987) explored the association between several particulate measures and excess mortality incidence using available data in 1980. Their analysis indicated the importance of considering particle size, composition, and source information in modeling particulate pollution health effects.

By the early 1990’s, most studies were conducted on a single city (e.g. Bates et al. 1990; Schwartz 1994) and showed statistically significant negative associations between PM and mortality and morbidity. These studies have limitations in that the results obtained from a particular city can not be generalized to other areas because of different characteristics of cities. Due to those limitations, the single-city studies have been replaced by multi-city studies, which assemble city-specific data under a common framework (Burnett and Krewski, 1994; Katsouyanni et al., 1997; Dominici
et al., 2000). Dominici et al. (2000) estimated the effects of PM$_{10}$ on daily mortality from the 20 largest cities in U.S. Daniels et al. (2000) extended this analysis to estimate the PM mortality dose-response curve. The National Morbidity, Mortality and Air Pollution Study (NMMAPS) (Samet et al. 2000a; 2000b) is the largest multi-city time series study based on 90 cities. NMMAPS reported strong evidence linking daily increases in particulate pollution to increases in death. On average, overall mortality increased by 0.5 percent for every 10 microgram per cubic meter increase in PM$_{10}$ measured the day before death.

Previous city-wide epidemiological studies have produced varying conclusions, which are highly sensitive to the choice of covariates, mortality data, and definition of human exposure. Also, most previous studies have mainly focused on the adverse health effects for both mortality and morbidity due to PM$_{10}$ rather than PM$_{2.5}$ because until recently, PM$_{2.5}$ data were available at only a few locations. In this study, we investigate the associations between PM$_{2.5}$ and mortality not at a few locations but over all counties in the U.S. The work proposed here attempts to model the spatial distribution of PM over the U.S., and to develop a framework to relate PM and mortality in a spatially explicit manner. The spatial modeling of PM$_{2.5}$ is needed for several reasons. One of the main reason is to consider the measurement error. Since PM$_{2.5}$ values collected at monitoring stations are not, but contain some measurement errors, we can obtain better estimates of PM$_{2.5}$ using spatial modeling. Also, we can interpolate PM$_{2.5}$ values at the sites where no observations are available from the spatial modeling of PM$_{2.5}$ and assess regional trends of PM$_{2.5}$. Several studies have investigated spatial modeling of PM. Cressie et al. (1999) explored the
spatial interpolation of PM$_{10}$ monitored at 27 monitoring stations in the Pittsburgh area, using a geostatistical method and a Markov random field approach. Sun et al. (2000) interpolated PM$_{10}$ using Bayesian methods and Kibria et al. (2002) suggested a Bayesian framework to map PM$_{2.5}$ in Philadelphia.

In this study, we develop a Bayesian hierarchical spatio-temporal model to relate PM$_{2.5}$ and mortality based on the best available PM$_{2.5}$ information. A Bayesian hierarchical model is adequate in this analysis in that it combines data across regions and quantifies the sources of variability. We adjust for confounding factors such as meteorology and socio-economic conditions for a better estimation of the effects of PM$_{2.5}$ on mortality. Our Bayesian hierarchical model also handles the change of support problem which arises when combining point level PM$_{2.5}$ data and county level mortality data. Our PM$_{2.5}$ data are collected from point monitoring networks and mortality data reported by public health sources are aggregated data over regions rather than at individuals. Hence, when combining these two variables together, the support areas of these variables are not matched, which is often referred to as the change of support problem. The new hierarchical framework proposed here combines methods for point-level misaligned data and change of support regression. This framework provides a better estimation of the hypothesized increased rate of mortality with increased PM$_{2.5}$ levels across all counties within in the U.S. The temporal dimension is not considered in the application. The primary objectives of our research are (1) to quantify the impact of PM$_{2.5}$ on mortality using a Bayesian hierarchical model, (2) to explore the heterogeneity of effects of mortality due to PM$_{2.5}$ across geographic regions, and (3) to identify the components of PM$_{2.5}$ which are the most significant
in causing mortality.

### 3.2 Data description

In this study, we combine data from three different sources; air pollution, mortality, and meteorological data for year 2000. PM$_{2.5}$ data have been collected from two monitoring networks, the Federal Reference Method (FRM) monitoring network and the Interagency Monitoring of Protected Visual Environments (IMPROVE) network. In 1997, the U.S. Environmental Protection Agency introduced a new PM standard based on PM$_{2.5}$ to supplement an earlier standard based on PM$_{10}$. The new standard provides more protection to the groups who are most sensitive to the exposure of PM$_{2.5}$. This new annual standard for PM$_{2.5}$ is currently set at 15$\mu$g/m$^3$. In conjunction with the new PM standard, a new monitoring network, known as the FRM network has been established. The FRM network consists of about 1000 monitoring sites including rural and urban sites. Samples of PM$_{2.5}$ are collected everyday, every third day or every sixth day. While the FRM network is a big national network located widely over the nation, the IMPROVE network is a smaller network, including 156 sites located at the national parks and wilderness areas. The IMPROVE network has been established to meet the requirement of the Clean Air Act to monitor the status and trends of air quality in the national parks. Samples of PM$_{2.5}$ from IMPROVE are monitored every day, every third day or every sixth day. Figure 3.1 (a) presents the monitoring stations of the FRM and IMPROVE networks. The IMPROVE and FRM networks include measurements of both the composition and the concentration of PM$_{2.5}$.
In addition to PM$_{2.5}$, we also include PM$_{10}$ and ozone in the model to study the effects of these co-pollutants and PM$_{2.5}$ jointly. PM$_{10}$ and ozone data are monitored through the State and Local Air Monitoring Stations (SLAMS) and the National Air Monitoring Stations (NAMS).

Daily meteorological data for 1999 and 2000 have been obtained from the U.S. National Climate Data Center. We use daily minimum temperature (°C), daily maximum temperature (°C), daily wind speed (m/s), daily dew point temperature (°C) and daily pressure (hPa). Figure 3.1(b) shows the weather stations.

Monthly mortality data have been obtained from the National Center for Health Statistics at the level of counties in the U.S. We consider natural mortality and cardio-
vascular mortality for all counties in U.S. for 1999 and 2000, and classify the mortality counts by several racial groups (Caucasian, African-American and Hispanic) and age groups (between 0 and 14, between 15 and 64, over 65). The spatial and temporal resolutions of our analysis are limited by the resolution at which the mortality data are reported. Here, we use monthly mortality data summarized over counties in the U.S. Figure 3.2 shows the total mortality counts for June 2000 across the U.S.

### 3.3 Statistical Models

County level mortality data have been modeled in various ways in the literature (e.g., Dominici et al, 2002, Short et al., 2002, Zhu et al., 2003). In modeling the association between air pollution and human health, the most common model is the Poisson regression model. For example, Best et al. (2000) studied the relationships between different spatially varying quantities to an underlying unobservable random
field for a Poisson regression analysis of health and exposure data. In many empirical studies of mortality, the mortality count is modeled as a function of other social and economic variables such as patients education level and family income. Such count variables are often either temporally or spatially correlated and such correlations should be considered when fitting the regression models. The commonly used models include the standard Poisson and negative binomial regression models with an independence assumption. These models account for the fact that the mortality count within a county is non-negative. Although the Poisson model is a natural choice for modeling count data, it requires the strict assumption that the mean and variance of the response variable are equal. In practice, this assumption is too restrict in many count data, and the variance of count data can be either larger (over dispersion) or smaller (under dispersion) than the mean. Therefore, inference based on the estimated standard errors are no longer valid, and we need to consider the over and under dispersion in modeling count data. To accommodate the over and under dispersion of count data, we model mortality counts using a generalized Poisson regression model rather than a standard Poisson model. In our research, we develop a Bayesian hierarchical spatio-temporal model based on a generalized Poisson regression model and estimate the effect of exposure to PM$_{2.5}$ on mortality.

3.3.1 Generalized Poisson regression models

Let $Y$ follow a generalized Poisson distribution with parameters $\mu$ and $\alpha$. The probability function of a generalized Poisson (Gpoi) random variable is defined in the
following way (Famoye, 1993);
\[ f(y) = \Pr[Y = y] = \left( \frac{\mu}{1 + \alpha \mu} \right)^y \frac{(1 + \alpha y)^{y-1}}{y!} \exp \left\{ -\frac{\mu (1 + \alpha y)}{1 + \alpha \mu} \right\}, \quad y = 0, 1, \ldots, \frac{1}{\alpha_-}, \]
where \( \mu \) and \( \alpha \) are the parameters of the distribution and \( \alpha_- = -\alpha I(\alpha < 0) = \min(\alpha, 0) \). This generalized Poisson distribution is denoted as \( Y \sim GP(\alpha, \mu) \). The parameter \( \mu \geq 0 \) represents the mean of the process \( Y \), and \( \alpha \) controls the dispersion of the process. When \( \alpha < 0 \), the GPoi distribution is truncated at \( \left[ -\frac{1}{\alpha} \right] \). The GPoi model is a generalization of the standard Poisson model, which becomes the standard Poisson distribution for \( \alpha = 0 \). In contrast with the standard Poisson distribution, the GPoi model does not require the assumption that the mean and variance are equal, but accommodates the over and under dispersion using the dispersion parameter \( \alpha \).

The mean and variance of the Gpoi model are given by \( E(Y) = \mu \) and \( \text{Var}(Y) = \mu(1 + \alpha \mu)^2 \), respectively. When \( \alpha > 0 \), the GPoi model represents the over dispersion \( (\text{Var}[Y] > E[Y]) \), while \( \alpha < 0 \) represents the under dispersion \( (\text{Var}[Y] < E[Y]) \).

Based on the GPoi distribution for mortality counts, we develop a regression model to study the association between PM$_{2.5}$ and mortality. Regression models generally involve stating the relationship between the explanatory and response variables. Regression models have been widely used in epidemiological research, especially when dealing with large numbers of covariates. The main reason for using a regression model in epidemiology is to obtain estimates of the coefficients associated with the variables of interest. For example, in the regression model with mortality and PM$_{2.5}$ variables which are used as response and explanatory variables, respectively, we can estimate the effects of an increases in PM$_{2.5}$ on the risk of death from the coefficients of PM$_{2.5}$. Hence, regression models are performed widely used in identifying the vari-
ables of interest and the strength of the effect. Also we can investigate the change of the strength of the effect due to other variables in the regression model. If several variables are highly correlated, including these variables together in the model, may change the estimates obtained from the model in which only one variable is included. This is likely to be the case when we estimate the effects of several different pollutants. If we are interested in the effect of ozone on the association between PM$_{2.5}$ and mortality, we may include these pollutants together and study the changes in the coefficients of PM$_{2.5}$ before and after adding the ozone variable in the model.

Based on the Generalized Poisson regression model, we develop a hierarchical Bayesian framework in the next section to investigate the effects of PM$_{2.5}$ on mortality. We call our developed model a Generalized Poisson Hierarchical Bayesian (GPoiHB) model.

### 3.3.2 Generalized Poisson Hierarchical Bayesian models

Hierarchical models have been used in the studies of air pollution and health to obtain more stable estimates of the regression coefficients. Hierarchical models offer a convenient way of handling the different levels of correlation. For example, mortality counts and covariates are measured both at the county level and on individual patients within the county (e.g. ethnicity, age). In the hierarchical model, a random effect can be incorporated, referring to the fact that the regression coefficients can vary by individual. Hierarchical models may be fitted using likelihood methods, but are naturally viewed from a Bayesian perspective. The Bayesian approach provides a natural framework for dealing with hierarchical models, incorporating the uncertainty
Based on a hierarchical Bayesian model, we investigate the association between PM$_{2.5}$ and mortality. In our model, we do not include the population since the population density is not a good estimate of the population at risk which we can not estimate. Hence, instead of modeling mortality data standardized with respect to the population at risk in a region, we model the raw mortality counts using a Generalized Poisson model.

Let $Y_{jk}(t)$ be the mortality count for county $j$ strata $k$ and time $t$. We assume $Y_{jk}(t)$ follows a Generalized Poisson distribution,

$$Y_{jk}(t) \sim GPoi(\alpha, \mu_{jk}(t)),$$

where $\alpha$ is the dispersion parameter, and $\mu$ is the mean parameter, called an expected mortality. The index $t$ denotes the temporal units which are months in this analysis. The index $j$ represents spatial units which are county, and the index $k$ denotes strata of the category of socio-economic factors (i.e., a stratum could be hispanic females over 65 years old).

At the second stage, we model $\mu$ with confounding and exposure covariates,

$$\log(\mu_{jk}(t)) = U_j^T(t)\gamma_1 + X_k^T(t)\gamma_2 + V_j^T(t)\beta_j(t),$$

where $V_j^T(t) = (V_{1j}(t), \ldots, V_{lj}(t))$. The vector $U_j(t)$ represents a vector of confounders at time $t$ defined at the county level (e.g., ozone). For each person in our database, we have the information about the gender, age and ethnic group. In our analysis, we treat these variables as categorical variables. $X_k(t)$ is a vector of
confounders at time $t$ including age, ethnicity, and gender. The exposure covariate represented by $V_j(t)$ is the vector or exposure covariates such as speciated PM$_{2.5}$ that varies with both time $t$ and county $j$. $I$ is the dimensionality of the exposure covariate (i.e. the number of components of PM$_{2.5}$). The parameter $\beta$ explains the effect of the exposure variables on mortality in county $j$ at month $t$ and $\beta \times 10^3$ is defined as the log relative risk (RR) parameter.

Within a county $j$, there are different sources of information about PM$_{2.5}$ (e.g., output of numerical models, monitoring networks) that may have different spatial resolution and different sources of uncertainty and bias. To combine different information together, we need to interpolate the “true” (total mass) PM$_{2.5}$ process $Z(s,t)$ for site $s$ and time $t$. We explain the spatial modeling of the $Z(\cdot, \cdot)$ process in Section 3.4. Once we model the true PM$_{2.5}$, we use it in modeling the speciated PM$_{2.5}$. In this analysis, we consider 5 different components of PM$_{2.5}$, i.e. carbon, sulfate, nitrate, ammonium and crustal materials. The composition of PM$_{2.5}$ varies by time, location, and weather conditions. Speciated PM$_{2.5}$ data are recently monitored and only available at a few monitoring stations. Also there are periods of missing data due to speciated PM$_{2.5}$ not being measured at a site and a monitor being inactive by design. Thus, there are not enough speciated PM$_{2.5}$ data to interpolate values over the entire U.S and we need to interpolate speciated PM$_{2.5}$ using other information. Speciated PM$_{2.5}$ is a portion of total PM$_{2.5}$ mass. Considering this, we can express speciated PM$_{2.5}$ as a proportion of the total PM$_{2.5}$ mass. This is the way we solve the lack of information of speciated PM$_{2.5}$. Rather than modeling each $PM_{2.5}$ component directly using its own data, we model speciated PM$_{2.5}$ in terms of the portion
of the total PM$_{2.5}$ mass. It is easier to estimate the speciated PM$_{2.5}$ in terms of the total PM$_{2.5}$ mass than using the monitored speciated PM$_{2.5}$ data, since total PM$_{2.5}$ is measured more often and more accurately than speciated PM$_{2.5}$. We model speciated PM$_{2.5}$ data in the following way,

$$V_{ij}(t) = \xi_i(t) \int_{C_j} Z(s,t)ds + \phi_{ij}(t),$$

The parameter $\xi_i$ denotes the proportion of the true total PM$_{2.5}$ $Z(s,t)$ explained by the $ith$ speciated PM$_{2.5}$ components. We estimate $\xi_i(t)$ at each time $t$, independently of $\xi_i(t-1)$. Alternatively, we could use an autoregressive structure for $\xi_i(t)$. The parameter $\phi_{ij}(t)$ is a spatially structured random effect for the $ith$ component of PM$_{2.5}$ in county $j$ at time $t$.

The modeled speciated PM are obtained at a county level to solve the change of support problem which arises in combining point level PM$_{2.5}$ and county level mortality data. Since PM$_{2.5}$ data are available at point monitoring stations and mortality data are collected at the county level, we need to adjust the support levels of variables to a common level. Since mortality data can not be disaggregated to a finer resolution, we aggregate PM$_{2.5}$ to the county level using an integration of the true PM$_{2.5}$ over a county area $C_j$, (i.e. $\int_{C_j} Z(s,t)ds$) and match the support area of the two variables.

In modeling speciated PM$_{2.5}$, we also consider the spatially structured random effects $\phi_{ij}(t)$ for the $ith$ PM$_{2.5}$ component in county $j$ at time $t$. The random effects are contained in the model to capture the error terms which are spatially correlated. To estimate the random effects of 5 different components of PM$_{2.5}$ simultaneously
for county $j$ and time $t$, we consider a multivariate GMRF prior for the spatially structured random effects.

### 3.3.3 Multivariate GMRF

A multivariate GMRF (MGMRF), also known as a multivariate conditional autoregressive (MCAR) model, have been initially proposed by Mardia (1988). Later, Gelfand and Vounatsou (2003) proposed the condition for the properness of GMRFs and employed the proper GMRF in the analysis of child growth data. Carlin and Banerjee (2003) and Jin and Carlin (2003) have used the MGMRF for cancer studies.

We adopt the MGMRF prior for the spatially structured random effects, $\phi_j(t) = (\phi_{1j}(t), \cdots, \phi_{Jj}(t))^T$ in modeling PM$_{2.5}$. Our MGMRF is different from that proposed by Banerjee and Carlin (2003) in that we extend the MGMRF model by allowing it to change over time using temporally varying dynamic parameters (e.g. $\Sigma_j(t)$),

$$
\phi_j(t) | \phi_{-j}(t) \sim N \left( \frac{1}{m_j} \sum_{j' \neq j} B_{jj'}(t)\phi_j(t), \frac{1}{m_j} \Sigma_j(t) \right),
$$

where $\phi_{-j}(t) \equiv (\phi_1(t), \ldots, \phi_{j-1}(t), \phi_{j+1}(t), \ldots, \phi_J(t))$, and $m_j$ is the number of neighbors (adjacent counties) of county $j$. The component $B(t)_{jj'}$ is an $I \times I$ adjacency matrix which can be simplified as $b_{jj'}I$ under the assumption that each component has the same adjacency matrix. For an adjacency coefficient $b_{jj'}$, binary weight can be used which is 1 if two counties, $j$ and $j'$ are adjacent and 0 otherwise. The matrix $\Sigma_j(t)$ explains the conditional variability and cross-covariance relationships at time $t$ between the different exposure variables (speciated PM$_{2.5}$) given the neighboring sites. This is an improper prior, but the posterior will be proper. Since this prior is
only identified up to an additive constant, for each time unit \( t \), we add \( I \) centering constraints \( \sum_j \phi_{ij}(t) = 0 \) for \( i = 1, \ldots, I \) to identify an intercept term in our model. In our analysis, we simplify the covariance matrix \( \Sigma_{\phi}(t) \) as,

\[
\Sigma_{\phi}(t) = D \otimes \Sigma_j(t)^{-1},
\]

where \( \Sigma_{\phi}(t) \) is a symmetric p.d. \( IJ \times IJ \) matrix, \( D = \text{Diag}(m_i) \) is a \( J \times J \) matrix, and \( \Sigma_j(t) \) is a p.d. and symmetric \( I \times I \) matrix that explains the variation across the PM components.

### 3.3.4 Spatial model for PM

PM\(_{2.5}\) data obtained from the FRM and IMPROVE networks are not true values but include measurement errors. We assume there is an underlying field \( Z(s,t) \) which measures the true values of PM\(_{2.5}\) at location \( s \) and time \( t \). Let \( \hat{Z}_F(s,t) \) denote the observation of PM\(_{2.5}\) at location \( s \) and time \( t \) collected from the FRM. We assume that \( \hat{Z}_F(s,t) \) is represented by the true process and measurement error,

\[
\hat{Z}_F(s,t) = Z(s,t) + e_F(s,t)
\]

where \( e_F(s,t) \sim N(0, \sigma^2_F) \) represents the measurement error at location \( s \). We assume that the error process is independent of the true underlying process.

In a similar way, we represent the IMPROVE measurements with the true process and measurement error,

\[
\hat{Z}_I(s,t) = a(s,t) + Z(s,t) + e_I(s,t),
\]
where $e_I(s, t) \sim N(0, \sigma_I^2)$ is the measurement error at location $s$ and time $t$, which is independent of the true underlying process. The parameter $a(s, t)$ explains some potential instrument bias.

The true underlying process $Z$ is unobservable, and we model it using meteorological variables $W$,

$$Z(s, t) = W^T(s, t)\eta + \epsilon(s, t), \tag{3.3}$$

where $W$ is a vector of meteorological variables (such as temperature, dew point, wind speed and pressure) and $\epsilon(s, t)$ is a white noise. The complex relations between meteorological variables and true PM$_{2.5}$ process are modeled using a generalized additive model (GAM) (Hastie and Tibshirani, 1990). The GAM models assume that the mean of the dependent variable depends on an additive predictor through a nonlinear link function. Therefore the GAMs provide a flexible framework for controlling for non-linear dependence on potential covariates. A number of time series studies have used GAM models to assess the relationship between air pollution and health since GAM models are used to allow for long-term patterns, such as seasonality. For example, Schwartz (1994) analyzed the relationship between PM$_{10}$ and SO$_2$ and mortality in Birmingham using a GAM model. In our analysis, a GAM framework offers a lot of flexibility for modeling meteorological variables to capture and explain the potential seasonality, diurnal cycles, and some other periodicities.

The weather data might not be available at all the locations of interest. Thus, we need to interpolate the weather data (as part of our hierarchical framework) at those locations using the spatial model. The weather and pollution information obtained are daily summaries that were averaged for each month prior to being used in our
GPoiHB framework. However, this change of temporal support could have been incorporated into our GPoiHB framework, and it could have been handled in a fashion similar to how we handled the spatial change of support problem. In the application presented in this paper, the emphasis is on the spatial aspect. Because of this and the computational demand, we limit our work to the averaged monthly values.

**Estimation and prediction of the parameters**

The parameters in the model are estimated from the summary of the posterior distribution of the parameters given all the observed data. The posterior distributions are obtained based on simulating values of $Z$ for all counties.

$$P(Z|\hat{Z}),$$

where $\hat{Z} = (\hat{Z}_F, \hat{Z}_I)$.

For spatial prediction, the quantity of interest is the predictive distribution for $Z(s_0, t)$ (the true PM value at location $s_0$ and time $t$) given all the available data $\hat{Z}$ (IMPROVE, FRM). When the goal is to predict $Z$ at a location $s_0$, the Bayesian approach uses the posterior predictive distribution of $Z(s_0, t)$

$$p(Z(s_0, t)|Z) \propto \int p(Z(s_0, t)|Z, \eta), p(\eta|Z) d\eta, \quad (3.4)$$

We use a Gibbs sampling approach to simulate $m$ values from the posterior of the vector parameter $\eta$. Gibbs sampling consists of sampling from full conditional distributions. The predictive distribution can be approximated by using Markov chain Monte Carlo (MCMC), more specifically by using the *Rao-Blackwellized estimator*
(Gelfand and Smith, 1990):

\[
p(Z(s_0, t)|Z) = \frac{1}{m} \sum_{i=1}^{m} p(Z(s_0, t)|Z, \eta^{(i)})
\]  

(3.5)

For measuring the association between PM and mortality, the parameters $\gamma$ and $\beta$ are estimated from the posterior distribution,

\[
P(\gamma_1, \gamma_2, \beta|X, Z, U, V).
\]

### 3.4 Application

We studied the spatial association between monthly mortality and exposure to monthly averages of daily $PM_{2.5}$ mass and its components based on the GPoiHB model. The model described in the previous section is applied to the monthly mortality counts obtained over county level, daily exposure data and daily meteorological data monitored at the point monitoring stations. Since data come from varying sources with different spatial and temporal resolutions, we obtained average exposure and meteorological data over the month instead of using daily data. The spatial misalignment is solved by aggregating the point level data to the county level.

After we adjust the spatial and temporal misalignment problems, we combine monthly county level mortality data, exposure data and meteorological data. For mortality data, natural mortality and cardiovascular mortality are considered. Following the U.S. E.P.A. guideline, we group $PM_{2.5}$ into five components: sulfate, ammonium, nitrate, total carbon and crustal material that contains calcium, aluminum, silicon, magnesium and iron. These five main components constitute our exposure
variables.

For confounding variables, we used several meteorological covariates (minimum temperature, maximum temperature, dew point temperature, wind speed and pressure) and add elevation as a geographic covariate. Age and ethnicity are included as intercept-based covariates. Age is categorized as under 14 years (children), 15-64 years (adults) and over 65 years (elderly group), and ethnic groups are categorized as Caucasian, African-American and Hispanic. Besides these covariates, we also add ozone to study the effects of ozone on the association between PM$_{2.5}$ and mortality.

For the priors, we adopted a normal prior $N(\mu_\beta, \Gamma_\beta)$ for the log relative risk parameter ($10^3 \times \beta$ parameter) with hyper parameters $\mu_\beta$ and $\Gamma_\beta$. The parameter $\mu_\beta$ has a normal prior, and an uniform prior is used for the standard deviation (S.D.) of the hyperpriors of the standard deviation. The coefficients of other covariates are given normal priors. The MGMRF prior is used for spatial random effects.

The results were obtained using WinBUGS, R and ArcView. We fit our data with a generalized Poisson model, a negative binomial model and a Poisson model and compared the models using the Deviance Information Criterion (DIC). The DIC for the negative binomial model was 622 and the Poisson and the generalized Poisson was 615 and 618, respectively. The dispersion parameter did not appear to be significant (mean: 0.05 and S.D. 0.04). We also compared models using the root mean squared prediction error (RMSPE). The RMSPE of the negative binomial and the Poisson model was 1.13 and 1.11, respectively. From these results, we found that the Poisson model fits better than the other models. Though there is no evidence of over and
under dispersion in this particular application, it is recommended to estimate the dispersion parameter of the generalized Poisson model to investigate over or under dispersion of the data.

Figure 3.3 shows the prediction map for total PM$_{2.5}$ mass on a grid for the entire U.S., with a resolution of 1 degree longitude by 0.5 degrees latitude, given the two monitoring networks (FRM and IMPROVE), and then averaging these gridded values at the county level. Figure 3.4 displays the mean of the posterior distribution for the RR, which explains the percent increase in monthly mortality per increase in one unit of the exposure variable. We can notice that the spatial patterns in the RR, show higher risks in the Eastern U.S. and Southern California area. The S.D. of the RR is displayed in Figure 3.5. The average RR over space was 6.6% for the month June. We compared the risk of PM$_{2.5}$ with PM$_{10}$. Figure 3.6 displays total PM$_{10}$ values and Figure 3.7 shows the RR of total PM$_{10}$. Comparing RR values of PM$_{2.5}$ to those of PM$_{10}$, the risk for PM$_{10}$ is approximately half of the risk for PM$_{2.5}$ except for the areas in Texas which show high concentrations of PM$_{10}$. We also obtained the RR of cardiovascular mortality (Figure 3.8). The spatial patterns and magnitude of the RR for cardiovascular mortality were similar to those of the RR as is natural mortality. The risks were also high in the Eastern and Southern California area, and the average RR for cardiovascular mortality was 6.0% which was a little bit smaller than the average RR for natural mortality (6.6%). We studied the effects of neighborhood on the RR of natural mortality. Figure 3.9 is the RR values using the 2nd order neighborhood. The average RR obtained by the 2nd order neighborhood was 7.0%, which was a little higher than the RR from the 1st order neighborhood.
To investigate the seasonal effects of RR, we obtained the RR in December (Figure 3.10). The magnitude of RR for December was smaller than that of RR for June. The average RR was 4.0%, and the risk in the Southern California area were particularly higher than in the other areas.

We studied the impact of ozone on the association between PM and mortality. Figure 3.11 presents the differences between the RR when ozone is not included in the model and when ozone is included. The impact of ozone does not seem to be significant taking into account the S.D. of the RR parameter.

We investigated the association between speciated PM$_{2.5}$ and mortality and obtained the RR for each component. Figure 3.12, Figure 3.13 and Figure 3.14 show the RR of mortality due to total carbon, ammonium and crustal material. Figure 3.15 and Figure 3.16 are boxplots of the mean of the posterior distribution of the estimated RR of mortality due to sulfate and nitrate for nine geographic regions of interest defined by U.S. Census regions. In the case of nitrate and crustal material, the Eastern U.S and the Southern California area show relatively high RR parameter values compared to other regions, and for sulfate, total carbon and ammonium, the middle Atlantic area shows high risk values.

We obtained spatial estimates of expected mortality. Figure 3.17 displays the mean of the posterior predictive distribution for the $\mu$ parameter, showing high expected mortality in the Eastern U.S. and Southern California.

Environmental justice issues encompass impacts on the natural and physical environment and interrelated social, cultural and economic effects. Under the Clean
Air Act, the U.S. EPA is required to ensure that environmental justice concerns are considered by all federal agencies. Therefore, it is of interest to better understand if there are significant differences in the RR of mortality for different ethnic groups. Ethnicity is used as a surrogate to help identify groups at higher risk. By including an interaction between the ethnicity covariates (categorical variable) and the exposure variable, we can obtain the RR for different groups. Figure 3.18 presents the estimated RR parameter for the Caucasian, African-American and Hispanic groups in the California. To compare the RR between different groups, we obtain the differences between the estimated RR between the Caucasian and African-American groups, and between the Caucasian and the Hispanic groups (Figure 3.19). We found that there are some significant differences in the RR between the Caucasian and African-American groups, and the Caucasian and Hispanic groups in the Northern California taking into account the size of the S.D. We also compared risks of PM$_{2.5}$ by different age groups. Figure 3.20 displays the estimated RR parameter for different age groups; children (0-14), adult (15-64) and elderly group (over 65). As expected, the RR parameter tends to increase for the older group.

Finally, we present some model diagnostics. We conducted model diagnostics at the randomly selected counties. Figure 3.21 shows the location where the model diagnostics are conducted. We eliminate the mortality counts at those locations and predict the values using our GPoiHB framework. We used 95% prediction intervals and had only 4% of the predictions lie outside the interval for the month of June 2000 (Figure 3.22). We also showed the calibration analysis for the PM$_{2.5}$ exposure variable in Figure 3.23. Using a leave one out cross validation method, only 4% of
Table 3.1: Summary of the posterior distributions for the coefficients of the weather and pollution covariates for June 2000

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>wind</td>
<td>-0.48</td>
<td>0.02</td>
<td>-0.53</td>
<td>-0.48</td>
<td>-0.43</td>
</tr>
<tr>
<td>elevation</td>
<td>0.08</td>
<td>0.03</td>
<td>0.01</td>
<td>0.08</td>
<td>0.13</td>
</tr>
<tr>
<td>max. temperature</td>
<td>-1.21</td>
<td>0.04</td>
<td>-1.29</td>
<td>-1.21</td>
<td>-1.12</td>
</tr>
<tr>
<td>min. temperature</td>
<td>2.14</td>
<td>0.06</td>
<td>2.03</td>
<td>2.15</td>
<td>2.25</td>
</tr>
<tr>
<td>dew point</td>
<td>-1.38</td>
<td>0.04</td>
<td>-1.46</td>
<td>-1.30</td>
<td>-1.30</td>
</tr>
<tr>
<td>ozone</td>
<td>-0.36</td>
<td>0.02</td>
<td>-0.41</td>
<td>-0.36</td>
<td>-0.30</td>
</tr>
</tbody>
</table>

Table 3.2: Summary of the proportion of speciated PM for June 2000

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>sulfate</td>
<td>0.256</td>
<td>0.030</td>
<td>0.255</td>
<td>0.266</td>
<td>0.275</td>
</tr>
<tr>
<td>nitrate</td>
<td>0.064</td>
<td>0.019</td>
<td>0.049</td>
<td>0.058</td>
<td>0.079</td>
</tr>
<tr>
<td>ammonium</td>
<td>0.086</td>
<td>0.011</td>
<td>0.087</td>
<td>0.090</td>
<td>0.093</td>
</tr>
<tr>
<td>total carbon</td>
<td>0.508</td>
<td>0.146</td>
<td>0.370</td>
<td>0.484</td>
<td>0.579</td>
</tr>
<tr>
<td>crustal material</td>
<td>0.103</td>
<td>0.049</td>
<td>0.066</td>
<td>0.082</td>
<td>0.139</td>
</tr>
</tbody>
</table>

The predictions do not lie within the 95% prediction interval for June. In Figure 3.24, additional model diagnostics are presented for 3 different months, June, September and December in 2000 at 3 randomly chosen locations. The solid line, dashed-line and dotted-line represent the true mortality counts, predicted mortality values and 95% prediction intervals, respectively. The true mortality values lie within the prediction intervals, showing good calibration results.

The summary results are presented. Table 3.1 shows the the posterior coefficient of weather and other pollutant variables. Table 3.2 presents the summary of the proportion of speciated PM. In Table 3.3 the RR of three different types of exposure variables; total PM$_{2.5}$ mass, PM$_{10}$ and speciated PM$_{2.5}$ are described. The average risk associated with total carbon was larger than those of other components.
Figure 3.3: Total PM$_{2.5}$ mass ($\mu g/m^3$) (June 2000).

Figure 3.4: Posterior mean of log relative risk of mortality ($\beta_j$): percent increase in mortality per increase in 10 $\mu g/m^3$ of PM$_{2.5}$ concentrations. (June 2000).
Figure 3.5: Standard deviation of the log relative risk.

Figure 3.6: Total $PM_{10}$ mass ($\mu g/m^3$) (June 2000).
Figure 3.7: Posterior mean of log relative risk of mortality ($\beta_j$): percent increase in mortality per increase in 10 $\mu g/m^3$ of $PM_{10}$ concentrations. (June 2000).

Figure 3.8: Posterior mean of log relative risk of mortality ($\beta_j$): percent increase in cardiovascular mortality per increase in 10 $\mu g/m^3$ of $PM_{2.5}$ concentrations. (June 2000).
Figure 3.9: Posterior mean of log relative risk of mortality ($\beta_j$) with the 2nd order neighborhood: percent increase in mortality per increase in $10 \mu g/m^3$ of $PM_{2.5}$ concentrations. (June 2000).

Figure 3.10: Posterior mean of log relative risk of mortality ($\beta_j$) for December 2000: percent increase in natural mortality per increase in $10 \mu g/m^3$ of $PM_{2.5}$ concentrations.
Figure 3.11: Impact of ozone. Change in the log relative risk of mortality ($\beta_j$) by adding ozone, (beta parameter when ozone is NOT in the model minus beta parameter with ozone).

Table 3.3: Summary of the posterior distributions for the relative risk parameters for June 2000

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{10}$</td>
<td>3.0</td>
<td>0.4</td>
<td>2.7</td>
<td>3.0</td>
<td>3.2</td>
</tr>
<tr>
<td>PM$_{2.5}$</td>
<td>6.6</td>
<td>0.7</td>
<td>6.3</td>
<td>6.5</td>
<td>6.8</td>
</tr>
<tr>
<td>sulfate</td>
<td>6.5</td>
<td>0.7</td>
<td>6.2</td>
<td>6.5</td>
<td>6.8</td>
</tr>
<tr>
<td>nitrate</td>
<td>7.4</td>
<td>0.7</td>
<td>7.0</td>
<td>7.4</td>
<td>7.6</td>
</tr>
<tr>
<td>ammonium</td>
<td>7.0</td>
<td>0.8</td>
<td>6.7</td>
<td>7.0</td>
<td>7.2</td>
</tr>
<tr>
<td>total carbon</td>
<td>7.7</td>
<td>0.7</td>
<td>7.3</td>
<td>7.6</td>
<td>7.8</td>
</tr>
<tr>
<td>crustal material</td>
<td>4.4</td>
<td>0.6</td>
<td>4.2</td>
<td>4.4</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Table 3.4: Model comparisons are done using DIC (Deviance Information Criterion) and RMSPE (Root Mean Squared Prediction Error)

<table>
<thead>
<tr>
<th>method</th>
<th>pd</th>
<th>DIC</th>
<th>RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>31</td>
<td>618</td>
<td>1.11</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>31</td>
<td>622</td>
<td>1.13</td>
</tr>
<tr>
<td>Generalized Poisson</td>
<td>33</td>
<td>615</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.12: Posterior mean of relative risk of mortality due to total carbon: percent increase in mortality per increase in $10 \mu g/m^3$ of nitrate concentrations. (June 2000).

Figure 3.13: Posterior mean of relative risk of mortality due to ammonium: percent increase in mortality per increase in $10 \mu g/m^3$ of ammonium concentrations. (June 2000).
Figure 3.14: Posterior mean of relative risk of mortality due to crustal material: percent increase in mortality per increase in $10 \mu g/m^3$ of crustal material concentrations. (June 2000).

Figure 3.15: Relative risk for sulfate by regions. Regions: 1 New England, 2 Middle Atlantic, 3 MIDWEST (East North Central), 4 MIDWEST (West North Central), 5 SOUTH (South Atlantic), 6 SOUTH (South Central), 7 WEST (Mountain), 8 WEST (Pacific), 9 WEST (Southern California).
Figure 3.16: Relative risk for nitrate by regions. Regions: 1 New England, 2 Middle Atlantic, 3 MIDWEST (East North Central), 4 MIDWEST (West North Central), 5 SOUTH (South Atlantic), 6 SOUTH (South Central), 7 WEST (Mountain), 8 WEST (Pacific), 9 WEST (Southern California).

Figure 3.17: Expected mortality ($\mu_i$). (June 2000).
Figure 3.18: Relative risk by ethnic groups. Groups, a) the Caucasian, b) the African-American and c) the Hispanic

Figure 3.19: Differences of RR between different ethnic groups. a)RR for Caucasian-RR for African-American b)RR for Caucasian-RR for Hispanic
Figure 3.20: Relative risk by age. Groups: (0-14, 15-64, over 65)

Figure 3.21: Selected locations for cross validation and model diagnostics
Figure 3.22: Model diagnostics for mortality based on the 95% prediction intervals for June 2000

Figure 3.23: Model diagnostics for PM: 95% prediction intervals at location $i^{th}$, eliminating observation $i^{th}$. Total number of observations is 909. Good calibration: only 4% of the time the truth does not lie in the interval.
Figure 3.24: Model diagnostics for mortality for the months June, September and December. The solid line represents the true mortality counts, the dashed-line represents the predicted values, and the dotted-line represents 95% prediction intervals.
Chapter 4

Conclusion

The work presented in this thesis consists of the study of the relations between GMRFs and GGMs, and modeling the effects of air pollution on mortality. This chapter provides an overall discussion of the results in this thesis and possible future developments.

In the first part of the thesis, we explored relations between GMRFs and GGMs through approximations of GMRFs by GGMs, and vice versa. We approximated GMRFs by GGMs and GGMs by GMRFs using four approximation methods based on spectral density functions and covariance functions, which were evaluated in terms of an average MSPE. The empirical studies showed that approximations of GGMs by GMRFs gave smaller average MSPE than those of GMRFs by GGMs. In the approximations of GMRFs by GGMs, the estimated range parameter increased as $\alpha$ increased. This result might be expected since $\alpha$ and the range parameter control the strength of the correlation of the data in GMRFs and GGMs, respectively. We also studied the impact of the neighborhood selection on the estimated range parameter.
in the approximations of GMRFs by GGMs. Although the parameter estimates were
different depending on the criteria used, the results in terms of prediction were not
very different. In most cases, the SRDC method gave the largest AMSPE. In the ap-
proximations of GMRFs by GGMs, as the neighborhood order increased, the AMSPE
became smaller, while the AMSPE became larger as \( \alpha \) increased. The approxima-
tions of GGMs with various values of the range parameter using GMRFs gave similar
results with the four approximation methods. The smallest AMSPE was obtained for
the larger range value.

In the application part, the association between PM\(_{2.5}\) and mortality were inves-
tigated. While many previous studies focused on the effects of PM\(_{10}\), the analyses
presented here estimated the effects of speciated PM\(_{2.5}\) based on a Bayesian hierchical
spatio-temporal model. The special features of our model are taking into account of
spatial random effects of different components of PM\(_{2.5}\) simultaneously and adapta-
tion of the over and under dispersion in mortality data. This analysis was carried
out for all counties over U.S. for 2000 adjusting for socio-economic factors. We found
that the average monthly RR of PM\(_{2.5}\) was 6% which was approximately twice of
that of PM\(_{10}\) (3%). The monthly RR values of PM\(_{2.5}\) were particularly high in the
Northeastern and Southern California area. We also found some spatial patterns in
the RR for each component of PM\(_{2.5}\). In the Eastern U.S., we found that SO\(_4\) and
NH\(_4\) had more impact on mortality than the other components, while in the Western
U.S., NO\(_3\) and Crustal materials explained most of the impact of PM\(_{2.5}\).

Further investigations are being conducted to study the effect of the level of ag-
gregation on the potential bias of the results. Since our analysis is based on the
spatially and temporally aggregated data, the potential bias is large and can not be interpreted as an individual RR. Hence, we need a better design of such studies to improve the reliability of the estimates. Also we plan to study the bias that might occur in estimated parameters when GMRFs are used in modeling aggregated data instead of averaged continuous spatial process. Since GMRFs are used for a discrete process rather than a continuous process, the results obtained by using GMRFs might contain some bias. Comparing the effects of GMRFs and averaged continuous spatial process on estimated parameters might be an interesting research area. Finally, our model can be extended by considering lag effects of exposure and meteorological variables. Due to the large time unit (month) in our model, the lag effects are not considered. When working on daily data, the analysis could then incorporate the lag effects for better estimations.
REFERENCES


