Multivariate spatial-temporal modeling and prediction of speciated fine particles [1]

Jungsoon Choi, Montserrat Fuentes, Brian J. Reich, and Jerry M. Davis

Abstract

Fine particulate matter ($PM_{2.5}$) is an atmospheric pollutant that has been linked to serious health problems, including mortality. $PM_{2.5}$ is a mixture of pollutants, and it has five main components: sulfate, nitrate, total carbonaceous mass, ammonium, and crustal material. These components have complex spatial-temporal dependency and cross dependency structures. It is important to gain insight and better understanding about the spatial-temporal distribution of each component of the total $PM_{2.5}$ mass, and also to estimate how the composition of $PM_{2.5}$ might change with space and time, by spatially interpolating speciated $PM_{2.5}$. This type of analysis is needed to conduct spatial-temporal epidemiological studies of the association of these pollutants and adverse health effects.

We introduce a multivariate spatial-temporal model for speciated $PM_{2.5}$. We propose a Bayesian hierarchical framework with spatiotemporally varying coefficients. In addition, a linear model of coregionalization is developed to account for spatial and temporal dependency structures for each component as well as the associations among the components. We also introduce a statistical framework to combine different sources of data, which accounts for bias and measurement error. We apply our framework to speciated $PM_{2.5}$ data in the United States for the year 2004. Our study shows that sulfate concentrations are the highest during the summer while nitrate concentrations are the highest during the winter. The results also show total carbonaceous mass

[1]J. Choi is a Graduate Student at the Department of Statistics, North Carolina State University. M. Fuentes is a Associate Professor at the Department of Statistics, North Carolina State University. (Email: fuentes@ncsu.edu). B. J. Reich is a Postdoctoral Fellow at the Department of Statistics, North Carolina State University. J. M. Davis is a Professor in the Department of Marine Earth and Atmospheric Sciences, North Carolina State University.

concentrations are high during the summer and fall seasons.

Key words: multivariate spatiotemporal processes; Bayesian inference; linear coregionalization model; air pollution; environmental statistics.

1 Introduction

The study of the association between ambient particulate matter (PM) and human health has received much attention in epidemiological studies over the past few years. Özkaynak and Thurston (1987) conducted an analysis of the association between several particle measures and mortality. Their results showed the importance of considering particle size, composition, and source information when modeling particle pollution health effects. In particular, fine particle matter, $PM_{2.5}$ ($< 2.5\mu m$ in diameter), is an atmospheric pollutant that has been linked to numerous adverse health effects (e.g. respiratory and cardiovascular diseases). $PM_{2.5}$ is a mixture of pollutants, and is classified into five main components (U.S. EPA, 2003): sulfate, nitrate, total carbonaceous mass (TCM), ammonium, and crustal material (that includes calcium, iron, silicon, aluminum, and titanium). These components have complex spatial-temporal dependency and cross dependency structures. Fuentes et al. (2006) studied the health effects of $PM_{2.5}$ and its components using monthly data across the United States. In order to investigate the health effects associated to speciated fine PM across space and time, we need to have first the speciated $PM_{2.5}$ information at the locations and times of interest. However, daily speciated $PM_{2.5}$ measurements are available at limited monitoring sites, and missing values may occur at given time point. Rao et al. (2003) and Malm et al. (2004) showed the spatial and temporal patterns of speciated $PM_{2.5}$, but they only conducted an exploratory analysis of speciated $PM_{2.5}$. The research presented here is part of a larger project to study the association between speciated $PM_{2.5}$ and adverse health outcomes across the entire U.S. Our goal here is to develop a statistical framework using all available sources of data about speciated $PM_{2.5}$ to investigate the spatial-temporal patterns of speciated $PM_{2.5}$ and then be able to predict speciated $PM_{2.5}$ at all locations and times of interest. We also study the spatial-temporal patterns of the so called "unknown components" which are not the main speciated $PM_{2.5}$ components (as defined by EPA).

In this article we introduce a new statistical framework to combine different sources of information of $PM_{2.5}$ taking into account potential bias and measurement error, and we develop a multivariate spatial-temporal model for speciated $PM_{2.5}$ accounting for the

complex dependency structures of the speciated $PM_{2.5}$. We develop a spatial-temporal linear model of coregionalization (STLMC) to account for multivariate spatial-temporal dependency structures. In addition, a hierarchical framework is proposed to investigate the relative contribution of each component to the total $PM_{2.5}$ mass, which changes over space and time.

We use a new speciated $PM_{2.5}$ data set. To our knowledge, this is the first time that a statistical framework has been proposed and used to analyze speciated $PM_{2.5}$ across the entire United States. The proposed framework captures the cross dependency structure among the $PM_{2.5}$ components and explains their spatial-temporal dependency structures. A Bayesian hierarchical framework is used to study different random effects of interest. The STLMC proposed here allows for very general multivariate spatial-temporal covariance structures, and offers some computational advantages. We can accurately predict speciated $PM_{2.5}$ at any location or time point of interest. In addition, we present a new approach to combine different sources of information about $PM_{2.5}$, that improve the prediction of the total $PM_{2.5}$ mass. Wikle et al. (2001) presented a similar approach to combine different sources of information of ocean surface winds, but they treated one of the data sources as a prior process. In our approach, all of the data sources are simultaneously represented in terms of the underlying truth, and we also model the potential bias of the different sources of information as spatial temporal processes.

This article is organized as follows. In Section 2 we describe the data used in this study. In Section 3, we present a Bayesian hierarchical multivariate spatial-temporal model for speciated $PM_{2.5}$, and we also introduce the STLMC. A statistical framework to combine $PM_{2.5}$ data is described in Section 4. In Section 5 we present the results, and in Section 6 we offer a general discussion.

2 Data Description

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

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

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

3 Statistical Models

While speciated $PM_{2.5}$ is only available at about 200 STN stations, total $PM_{2.5}$ mass is available at about 1000 FRM network stations. In addition, we can compute an estimate of the total $PM_{2.5}$ mass from the $PM_{2.5}$ components. Thus, we model the total $PM_{2.5}$ mass using the $PM_{2.5}$ data from both networks, and then express the mean of each $PM_{2.5}$ component as a proportion of the total $PM_{2.5}$ mass. The proportion of each component to the total $PM_{2.5}$ mass varies over space and time, and we use a hierarchical framework to account for the spatial-temporal associations of the proportion. To ensure that at each site and time the sum of the proportions of the components to the total $PM_{2.5}$ mass is one, we use a multinomial logit function (McFadden, 1974) for the proportion parameters. Even though the spatial-temporal dependency structures of the proportions are considered, it could be insufficient to capture the spatial-temporal dependency and the cross dependency structures of speciated $PM_{2.5}$. We thus include a mean-zero spatial-temporal process in the model, which explains the dependency structures. This approach allows us to estimate both the speciated $PM_{2.5}$ in terms of the total $PM_{2.5}$ mass and the cross-covariance between the $PM_{2.5}$ components, which is not captured by the mean function of the speciated $PM_{2.5}$. Figure 1 shows the framework of the speciated fine PM.

We assume there is an underlying (unobserved) field $Z(\mathbf{s}, t)$, where $Z(\mathbf{s}, t)$ represents the true total $PM_{2.5}$ value at location $\mathbf{s} \in D_1$ at time $t \in D_2$, where $\{\mathbf{s} : \mathbf{s}_1, \ldots, \mathbf{s}_{N_s}\} \in D_1 \subset \mathbb{R}^2$ and $\{t : t_1, \ldots, t_{N_t}\} \in D_2 \subset \mathbb{R}$. Let $\mathbf{V}(\mathbf{s}, t) = (V_1(\mathbf{s}, t), V_2(\mathbf{s}, t), \ldots, V_5(\mathbf{s}, t))^T$ be a vector of the speciated $PM_{2.5}$ at location $\mathbf{s}$ and at time t. The parameter $\boldsymbol{\theta}(\mathbf{s}, t)$ is a vector of the proportions of the speciated $PM_{2.5}$ to the total $PM_{2.5}$ mass at location $\mathbf{s}$ and time t. Each parameter $\theta_i(\mathbf{s}, t)$ denotes a proportion of the total $PM_{2.5}$ mass attributed to the component

i, for $i = 1, \ldots, 5$. The model of the speciated $PM_{2.5}$ is defined as:

$$\begin{aligned}
\mathbf{V}(\mathbf{s},t) &= \boldsymbol{\theta}(\mathbf{s},t)Z(\mathbf{s},t) + \boldsymbol{\epsilon}(\mathbf{s},t) + \boldsymbol{\epsilon}^w(\mathbf{s},t), \\
&\quad \text{where } \boldsymbol{\theta}^T(\mathbf{s},t) = (\theta_1(\mathbf{s},t), \ldots, \theta_5(\mathbf{s},t)) \\
\text{logit}[\theta_i(\mathbf{s},t)] &= \eta_i(t) + \gamma_i(\mathbf{s},t), \qquad (1) \\
\eta_i(t) &= f_i(t) + e_{\eta_i}(t), \\
\gamma_i(\mathbf{s},t) &= \gamma_i(\mathbf{s},t-1) + e_{\gamma_i}(\mathbf{s},t), \\
Z(\mathbf{s},t) &= \mathbf{M}^T(\mathbf{s},t)\boldsymbol{\beta}(\mathbf{s},t) + e_z(\mathbf{s},t), \qquad (2) \\
\boldsymbol{\epsilon}(\mathbf{s},t) &= \mathbf{A}\mathbf{w}(\mathbf{s},t), \qquad (3)
\end{aligned}$$

where $\text{logit}[\theta_i(\mathbf{s},t)] = \log[\theta_i(\mathbf{s},t)/\theta_5(\mathbf{s},t)]$ is modeled using a dynamic hierarchical model (Gelfand et al., 2005). The function $\eta_i(t)$ denotes the overall temporal trend of the i^{th} logit component, which is expressed by the function $f_i(t)$ to explain seasonality of the i^{th} logit component. The process e_{η_i} is assumed to be a white noise Gaussian process. We model the process $\gamma_i(\mathbf{s},t)$ using a spatial-temporal first-order Markovian process, which accounts for the spatial-temporal structure of the i^{th} logit component not explained by the overall temporal trend. We assume the process $e_{\gamma_i}(\cdot,t)$ is a Gaussian process with mean zero and a spatial covariance function to explain spatial dependency structure. To guarantee that the multinomial logit model is identifiable, we assume $\eta_5(t) = 0$ for all t, and $\gamma_5(\mathbf{s},t) = 0$, for all $\mathbf{s}$ and t. In our study, crustal material is assumed to be the 5^{th} component because it is the most stable component.

The true total $PM_{2.5}$, $Z(\mathbf{s},t)$, is affected by a vector of meteorological weather variables $\mathbf{M}(\mathbf{s},t)$ (minimum temperature, maximum temperature, dew point temperature, wind speed, and pressure). The meteorological influence could vary in space and time, and we define $\boldsymbol{\beta}(\mathbf{s},t)$ to be meteorological parameters at location $\mathbf{s}$ and time t. The meteorological data might not exist at all the sites of interest. We thus interpolate the weather data at those locations using a spatial model for each time point (as part of the hierarchical framework). The residual process denoted by $e_z(\mathbf{s},t)$ is assumed to be normal. We explain the

spatial-temporal modeling of the $Z(\cdot,\cdot)$ in Section 4.

The spatial-temporal process $\boldsymbol{\epsilon}(\mathbf{s},t) = (\epsilon_1(\mathbf{s},t),\ldots,\epsilon_5(\mathbf{s},t))^T$ is assumed to be a Gaussian process with mean zero and a covariance matrix, which changes with space and time. The covariance matrix of the process $\boldsymbol{\epsilon}(\mathbf{s},t)$ represents the dependency structures of the speciated $PM_{2.5}$ not captured by the mean function of the speciated fine PM. The STLMC is developed to specify the covariance function of the process $\boldsymbol{\epsilon}(\mathbf{s},t)$. The process $\boldsymbol{\epsilon}(\mathbf{s},t)$ is expressed by the weight matrix $\mathbf{A}$ and the vector $\mathbf{w}(\mathbf{s},t)$ having independent Gaussian spatial-temporal processes. We discuss the STLMC in detail in the following subsection. The pure measurement error process, $\boldsymbol{\epsilon}^w(\mathbf{s},t)$, is assumed to be normal and be independent of $\boldsymbol{\epsilon}(\mathbf{s},t)$.

3.1 The Spatial-Temporal Linear Model of Coregionalization

The basic idea of the STLMC is that dependent spatial-temporal processes are expressed using the linear combination of uncorrelated spatial-temporal processes in the spatial-temporal modeling. The STLMC provides a very rich class of multivariate spatial-temporal processes with simple specification and interpretation. The STLMC like the linear model of coregionalization use in multivariate spatial analysis (Grzebyk and Wackernagel, 1994; Wackernagel, 1998) could be used as a dimension reduction method, which means that the given multivariate processes are represented as lower dimensional processes. In this study, we use the STLMC to construct a valid cross-covariance function of multivariate spatial-temporal processes. The STLMC used here is:

$$\boldsymbol{\epsilon}(\mathbf{s},t) = \mathbf{A}\mathbf{w}(\mathbf{s},t), \tag{4}$$

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

independent Gaussian spatial-temporal processes with mean zero and separable spatial-temporal covariance, $Cov(w_i(\mathbf{s}_l, t_j), w_i(\mathbf{s}_{l'}, t_{j'})) = C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \boldsymbol{\phi}_i) C_i^{(2)}(t_j, t_{j'}; \boldsymbol{\psi}_i)$, where $C_i^{(1)}$ is a spatial covariance with the parameter vector $\boldsymbol{\phi}_i$, and $C_i^{(2)}$ is a temporal autocovariance with the parameter vector $\boldsymbol{\psi}_i$. The STLMC in (4) implies $E(\boldsymbol{\epsilon}(\mathbf{s}, t)) = 0$ and

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

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

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

$$\boldsymbol{\Sigma}^{\boldsymbol{\epsilon}} = \sum_{i=1}^{5} \mathbf{T}_i \otimes \mathbf{U}_i \otimes \mathbf{R}_i, \tag{6}$$

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

3.2 Algorithm for Estimation and Prediction

We now discuss estimation and prediction of the speciated $PM_{2.5}$ using a Bayesian approach. In order to predict the speciated $PM_{2.5}$ at location $\mathbf{s}_0$ and time t_0 given the data $\mathbf{V}$ and $\mathbf{Z}$ (all available speciated $PM_{2.5}$ data and total $PM_{2.5}$ mass data, respectively) and a value $Z(\mathbf{s}_0, t_0)$, we need the posterior predictive distribution of $\mathbf{V}(\mathbf{s}_0, t_0)$:

$$p(\mathbf{V}(\mathbf{s}_0, t_0) | \mathbf{V}, \mathbf{Z}, Z(\mathbf{s}_0, t_0)) \propto \int p(\mathbf{V}(\mathbf{s}_0, t_0) | \mathbf{V}, \mathbf{Z}, \boldsymbol{\Theta}, Z(\mathbf{s}_0, t_0)) p(\boldsymbol{\Theta} | \mathbf{V}, \mathbf{Z}) d\boldsymbol{\Theta}, \tag{7}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ is a collection of all of the unknown parameters in our statistical framework. The vector $\boldsymbol{\Theta}_1$ includes parameters used to model the proportions of the speciated $PM_{2.5}$ to the total $PM_{2.5}$ mass, $\boldsymbol{\theta}$, and the vector $\boldsymbol{\Theta}_2$ includes covariance parameters of the speciated $PM_{2.5}$. We use a Markov chain Monte Carlo (MCMC) sampling algorithm to sample N_1 values from the posterior distribution of the parameter $\boldsymbol{\Theta}$ (within the software WinBUGS). Our MCMC sampling algorithm has three stages. We alternate between the proportion parameters $\boldsymbol{\Theta}_1$ given the data (Stage 1), the covariance parameters $\boldsymbol{\Theta}_2$ given the data and the values of $\boldsymbol{\Theta}_1$ updated (Stage 2), and the unobserved true values of $\mathbf{V}$ at all sites and time points (Stage 3). We obtain the conditional posterior distribution of the parameters $\boldsymbol{\Theta}$ given the data updated in Stage 3.

The conditional distribution of $\mathbf{V}$ at location $\mathbf{s}_0$ and time t_0 is:

$$p(\mathbf{V}(\mathbf{s}_0, t_0)|\mathbf{V}, \mathbf{Z}, \boldsymbol{\Theta}, Z(\mathbf{s}_0, t_0)) \sim N(\boldsymbol{\mu}(\mathbf{V}(\mathbf{s}_0, t_0)), \mathbf{T} - \Sigma^{12}\Sigma^{\mathbf{V}}\Sigma^{21}) \tag{8}$$

where $\boldsymbol{\mu}(\mathbf{V}(\mathbf{s}_0, t_0)) = \boldsymbol{\theta}(\mathbf{s}_0, t_0)Z(\mathbf{s}_0, t_0) + \Sigma^{12}\Sigma^{\mathbf{V}}(\mathbf{V} - \boldsymbol{\theta}\mathbf{Z})$, $(\Sigma^{21})^T = \Sigma^{12} = Cov(\mathbf{V}(\mathbf{s}_0, t_0), \mathbf{V})$ is a $5 \times (5N_tN_s)$ matrix and $\Sigma^{\mathbf{V}} = Cov(\mathbf{V}, \mathbf{V})$. Using the Rao-Blackwellized estimator (Gelfand and Smith, 1990), the predictive distribution is approximated by

$$p(\mathbf{V}(\mathbf{s}_0, t_0)|\mathbf{V}, \mathbf{Z}, Z(\mathbf{s}_0, t_0)) = \frac{1}{N_1}\sum_{n_1=1}^{N_1} p(\mathbf{V}(\mathbf{s}_0, t_0)|\mathbf{V}, \mathbf{Z}, \boldsymbol{\Theta}^{(n_1)}, Z(\mathbf{s}_0, t_0)), \tag{9}$$

where $\boldsymbol{\Theta}^{(n_1)}$ is the n_1^{th} draw from the posterior distribution for the parameters.

4 Spatial-Temporal Model for $PM_{2.5}$

This section introduces a spatial-temporal model for total $PM_{2.5}$ mass. We do not consider the $PM_{2.5}$ measurements at the FRM monitoring stations to be the "true" values because of measurement error. We thus denote the observed total $PM_{2.5}$ mass at location $\mathbf{s}$ at time

t from the FRM network by $\widehat{Z}_F(\mathbf{s},t)$, and we assume that

$$\widehat{Z}_F(\mathbf{s},t) = Z(\mathbf{s},t) + e_F(\mathbf{s},t), \tag{10}$$

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

A second estimate of the total $PM_{2.5}$ mass is the reconstructed fine mass (RCFM), $\widehat{Z}_R(\mathbf{s},t)$, defined as a sum of five main $PM_{2.5}$ components measured at the STN monitoring stations. We model $\widehat{Z}_R(\mathbf{s},t)$ as

$$\widehat{Z}_R(\mathbf{s},t) = a(\mathbf{s},t) + Z(\mathbf{s},t) + e_R(\mathbf{s},t), \tag{11}$$

where $e_R(\mathbf{s},t) \sim N(0,\sigma_R^2)$ is the measurement error at location $\mathbf{s}$ and time t and is also independent of processes e_F and Z. Since the "true" total $PM_{2.5}$ mass consists of more pollutants than the five main components, the additive bias $a(\mathbf{s},t)$ is needed. The bias can be represented as the bias of the RCFM data, $\widehat{Z}_R$, with respect to the FRM data, $\widehat{Z}_F$. Exploratory analysis suggests the additive bias varies over space and time, and we model $a(\mathbf{s},t)$ using a hierarchical framework,

$$\begin{aligned} a(\mathbf{s},t) &= a_1(t) + a_2(\mathbf{s},t), & (12)\\ a_1(t) &= h(t) + e_1(t), & (13)\\ a_2(\mathbf{s},t) &= a_2(\mathbf{s},t-1) + e_2(\mathbf{s},t), & (14) \end{aligned}$$

where $a_1(t)$ represents the overall temporal trend in the bias of the RCFM data, and $h(t)$ is a smoothing function of time to explain seasonality in the additive bias term. The process $a_2(\mathbf{s},t)$ accounts for the spatial-temporal structure which is not captured by the overall temporal trend. We assume the process $a_2(\mathbf{s},t)$ is a spatial-temporal first-order Markovian process, and e_1 and e_2 are independent white noise processes and are independent of the process Z.

The true underlying $PM_{2.5}$ process, $Z(\mathbf{s}, t)$, is modeled in terms of meteorological variables,

$$Z(\mathbf{s}, t) = \mathbf{M}^T(\mathbf{s}, t)\boldsymbol{\beta}(\mathbf{s}, t) + e_z(\mathbf{s}, t), \tag{15}$$

where the residual process e_z is assumed to be normal with a spatial-temporal covariance function.

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

$$p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}) \propto \int p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}, \boldsymbol{\Theta}_Z) p(\boldsymbol{\Theta}_Z|\widehat{Z}, \mathbf{M}) d\boldsymbol{\Theta}_Z, \tag{16}$$

where $\boldsymbol{\Theta}_Z$ is a collection of all parameters considered in the $PM_{2.5}$ model. After simulating N_2 values from the posterior distribution of the parameters $\boldsymbol{\Theta}_Z$, the estimator for the predictive distribution is $p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}) = \frac{1}{N_2} \sum_{n_2=1}^{N_2} p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}, \boldsymbol{\Theta}_Z^{(n_2)})$, where $\boldsymbol{\Theta}_Z^{(n_2)}$ is the n_2^{th} draw from the posterior distribution.

5 Application

Our statistical framework is applied to the daily speciated $PM_{2.5}$ data in the United States for the year 2004. In the $PM_{2.5}$ data framework, we study the spatial-temporal patterns of the additive bias term of the RCFM process. In the speciated $PM_{2.5}$ model, we study the spatial-temporal associations for each component as well as the associations among the components, and we also study the spatial-temporal pattern of the proportion of each component of the total $PM_{2.5}$ mass. We work with our spatial-temporal framework using only California data because of the computational costs, but we work with our framework over the entire United States at a fixed time or at a fixed location for year 2004.

In the $PM_{2.5}$ data framework, we observed that the coefficients of the weather covariates are different in different regions from the results of the preliminary exploratory analysis.

We thus implement our framework for the nine geographic regions as defined by the United States Census: Region (1): Northeast (New England); (2): Northeast (Middle Atlantic); (3): Midwest (East North Central); (4): Midwest (West North Central); (5): South (South Atlantic); (6): South (East South Central); (7): South (West South Central); (8): West (Mountain); (9): West (Pacific). We assume that $\boldsymbol{\beta}(\mathbf{s}, t)$ varies across regions but is constant over space and time within each region. For the error term e_z, we use a separable spatial-temporal covariance, a stationary exponential covariance for space and autocovariance of the first-order autoregressive function AR(1) for time. From the exploratory analysis, it appears that the overall temporal pattern of bias is a sine or cosine function. We assume that the overall temporal trend function for the bias, $h(t)$, in (13) is a linear combination of one sine and one cosine function with respect to a 12-month period (1/frequency). Similarly, we conducted a preliminary data analysis to choose the temporal function, $f_i(t)$, in the multinomial logit model of the proportion parameters. We fitted possible smoothing functions, and we found the third-order autoregressive function AR(3) seemed appropriate using AIC and BIC. We use AR(3) for the function $f_i(t)$. For the process $\mathbf{w}$, we use a stationary exponential covariance function $C_i^{(1)}(\mathbf{s}_l, \mathbf{s}_{l'}; \sigma_i^2, \phi_i) = \sigma_i^2 \exp(-h_1/\phi_i)$, $i = 1, \ldots, 5$, for $h_1 = ||\mathbf{s}_l - \mathbf{s}_{l'}||$ (in km) with the sill parameter σ_i^2 and the range parameter ϕ_i and the autocovariance function of the AR(1) $C_i^{(2)}(t_j, t_{j'}; \psi_i) = \psi_i^{h_2}/(1 - \psi_i^2)$ for $h_2 = |t_j - t_{j'}|$. For the spatial covariance function of the process $e_{\gamma_i}(\mathbf{s}, t)$, we also use an exponential covariance function with unit variance.

We now describe the prior distributions used here. We use gamma priors, G(0.01,0.01), for the precision of the error terms, $e_1(t)$ and $e_2(\mathbf{s}, t)$. Following the guidance and general approach used by U.S. EPA (1997, 2000), we use informative uniform priors, Unif(1.778,1.814) and Unif(1.362,1.390) for σ_F and σ_R, respectively. For the coefficients of the weather covariates and the sine and cosine functions, we use vague normal priors, $N(0, 0.01)$ (0.01 is the precision). For the error term e_z, we use normal hyperprior, $N(0, 0.1)$, for the parameter in the temporal covariance, and we use uniform hyperpriors, Unif(1,100) and Unif(0,100), for the range and sill parameters, respectively. We assume the measurement error term

$\boldsymbol{\epsilon}^w(\mathbf{s},t) \sim N(0, \sigma_w^2 I_5)$ where I_5 is an identity matrix. We use a uniform prior, Unif(0,100), for σ_w^2. In terms of the priors for the coefficients of the function $f_i(t)$, we use normal priors, $N(0, 0.01)$. For the error term e_{η_i}, we use a normal prior, $N(0, \sigma_{\eta_i}^2)$, $\sigma_{\eta_i}^2$ being a uniform hyperprior, Unif(0,100). For the process e_{γ_i}, we use a uniform hyperprior, Unif(1,100), for the range parameter. For the process w_i, we use uniform hyperpriors Unif(1,100) and Unif(0,100), for the range and sill parameters, respectively. The parameter ψ_i has normal hyperprior, $N(0, 0.1)$. Since $\mathbf{A}$ is a lower triangular matrix, we need to assign priors for elements $A_{ii'}$, $i, i' = 1, \ldots, 5$ and $i \geq i'$. We set $A_{ii}^2 = 1$, and we use normal priors, $N(0, 0.1)$, for off-diagonal elements.

Figure 2 shows boxplots of the estimated additive bias parameter, $a(\mathbf{s}, t)$, for July 2004 and December 2004 in nine geographic regions as defined by the U.S. Census Bureau. The estimated daily bias values are the medians of the posterior distribution. Here, we used monthly averaged values at each location. In July 2004, the estimated bias values were the lowest in the South Atlantic region and the highest in the Pacific region. In all regions except the Pacific region, the total $PM_{2.5}$ mass measured at the FRM network was larger than the RCFM in July 2004. The difference results in the Pacific region are not surprising because California during summer season is known to have the losses of about $60 - 90\%$ of nitrate due to evaporation when the total $PM_{2.5}$ mass at the FRM network is measured (Frank, 2006). In December 2004, all regions had negative values for the estimated additive bias. Overall, the estimated additive bias values in July was lower than in December because summer season has high sulfate and ammonium concentrations and the FRM total $PM_{2.5}$ mass includes high amount of water during the summer season (Frank, 2006).

Figure 3 presents maps of the estimated concentrations of sulfate and nitrate on June 14 and December 14. Sulfate concentrations were higher than nitrate concentrations across the United States on June 14. On average, sulfate concentration was $3.22 \mu g/m^3$ on June 14, while nitrate concentration was $0.38 \mu g/m^3$. On June 14, sulfate concentrations in the eastern United States were highest over the United States. Nitrate concentrations seemed to be high in Southern California. However, on December 14, sulfate concentrations decreased

and nitrate concentrations increased. The northwestern United States had high sulfate concentrations on December 14, while Southern California and the Mountain region had high nitrate concentrations.

The time series plots of the estimated concentrations of speciated $PM_{2.5}$ in Los Angeles, Phoenix, and New York City in 2004 are presented in Figure 4. The estimated sulfate and ammonium concentrations in Los Angeles and New York City were higher than those in Phoenix. Sulfate concentrations tended to be highest during the summer in Los Angeles and New York City. For these three cities, nitrate concentrations tended to be highest during the winter. In particular, it is interesting that nitrate concentrations in Phoenix peak during the winter. It appears that Los Angeles and Phoenix have high TCM concentrations during the winter. In Los Angeles, all of the components had high concentrations during March, and total $PM_{2.5}$ mass was also high during March.

Figure 5 shows maps of the estimated speciated $PM_{2.5}$ composition by region and by season in 2004. In the maps, circle size corresponds to total $PM_{2.5}$ mass, and we can clearly see the spatial-temporal pattern of the total $PM_{2.5}$ mass. During the summer season (July-September), the total $PM_{2.5}$ mass was highest in the eastern United States. TCM had the highest proportion of the total $PM_{2.5}$ mass among the components over the entire United States. Sulfate concentrations were highest during the summer season in most of the eastern United States and the Pacific region because increased photochemical reactions in the atmosphere increases sulfate formation (Baumgardner et al., 1999). Nitrate concentrations were highest during the winter season (January-March) because higher ammonia availability, the cooler winter temperatures, and higher relative humidities favor ammonium nitrate condensation. On average, nitrate concentration during the winter season was $1.92\mu g/m^3$ over the United States (vs. $1.09\mu g/m^3$ for the year 2004). We also see the seasonal pattern of ammonium concentration. On average, during the winter and spring seasons (January-June), ammonium concentrations were about 3.2 time higher than during the summer and fall seasons. During the summer and fall seasons, TCM concentrations were high because of secondary organic aerosol formation. Crustal material concentrations were higher in the

eastern United States during the spring season because of low soil moisture and high wind speeds. Also, these regions are impacted by North African dust during the spring (Malm et al., 2004).

We present in Table 1 the posterior estimates of the elements (T_{ij}) of the $\mathbf{T}$ matrix using the data from California in 2004. The results show the correlations between a pair of speciated $PM_{2.5}$ not explained by the mean function of the speciated $PM_{2.5}$.

Finally, we present some model diagnostics using the deviance information criterion (DIC) of Speigelhalter et al. (2002), as well as the calibration. We compare three different models using only data from California in 2004. Model 1 is our statistical framework proposed in this article. Model 2 ignores the spatial-temporal process $\boldsymbol{\epsilon}$ in Model 1. Model 3 removes both the spatial-temporal process $\boldsymbol{\epsilon}$ and the hierarchical framework of the proportion parameters. The DIC for Model 1 was -1149. For Model 2 it was 15575, and for Model 3 it was 17604. We also use the root mean squared prediction error (RMSPE). The RMSPE value for Model 1 was 0.4631, for Model 2 it was 2.092, and for Model 3 it was 3.726. Thus, our framework has the lowest DIC and RMSPE values among three models. In addition, we did calibration analysis for the speciated $PM_{2.5}$ in Phoenix to see the performance of our framework. We selected randomly 30 observations in 2004, and we obtained 95% prediction intervals for the i^{th} time given the data, not using data from the i^{th} time we are predicting. In Figure 6, we present the calibration plots for sulfate and nitrate in Phoenix. The percentages of the observed values that are outside the interval are 3.3%. We also did calibration analyses for the other three components in Phoenix. The percentages of the observed values lying outside the interval are between 0% and 6.7%. We obtained very good calibration results.

6 Conclusion

In this article we present a flexible hierarchical framework to study speciated $PM_{2.5}$. The multivariate spatial-temporal model proposed here allows for spatial-temporal dependency for each component and cross dependency structures among the components. A hierarchical

framework provides a natural way to investigate the spatiotemporally varying contribution of each component to the total $PM_{2.5}$ mass. Using our framework, we can estimate speciated $PM_{2.5}$ at unobserved locations of interest in the United States. We also introduce a new statistical framework to incorporate $PM_{2.5}$ data from different sources, which takes into account bias and measurement error over space and time.

We found that the additive bias term of the RCFM process overall seems to be negative and the RCFM is less than the total $PM_{2.5}$ mass observed at the FRM network. However, in the Pacific region, we can clearly see the different results during the summer season because of the losses of nitrate. In the eastern United States, the contribution of sulfate to the total $PM_{2.5}$ mass tends to be higher during the summer. In almost all regions, sulfate concentrations are higher during the summer. Also, the spatial differences in the sulfate concentrations are the largest during the summer. On average, the sulfate proportions and concentrations in the East South Central region are highest where sulfur is emitted from coal-fired sources (Malm et al., 2002). In general, nitrate concentrations are higher during the winter, and they are also higher in urban areas because of high nitrogen oxide (NO_x) emissions from automobiles. During the summer, nitrate and ammonium concentrations in the western United States are low. TCM concentrations explain most of total $PM_{2.5}$ mass. It is found that TCM has high concentrations in the summer and fall seasons because of high fire-related activity. During the spring season, crustal material concentrations are high in the eastern United States. Our results for the speciated $PM_{2.5}$ are consistent with previous analyses (Malm et al., 2004).

The diagnostics show the adequate performance of our model. Some extensions could be considered in our statistical framework. For example, the total $PM_{2.5}$ process, $Z(\mathbf{s}, t)$, could be modeled as a nonstationary spatial-temporal Gaussian process with the meteorological variables. In the STLMC, the separable spatial-temporal process for $\mathbf{w}(\mathbf{s}, t)$ we assumed could be extended to a non-separable process. However, computational burden is exacerbated in these cases and we used simple spatial-temporal models.

We are currently working on association between speciated $PM_{2.5}$ and daily mortality.

The framework and results presented here will be essential for the health analysis.

Acknowledgements

The authors would like to thank Drs. Holland, Tesh, and Prakash at EPA for providing the data as well as helpful and insightful discussions.

References

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

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

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

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

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

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

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

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

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

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

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

Rao, V., Frank, N., Rush, A., and Dimmick, F. (2003), "Chemical Speciation of $PM_{2.5}$ in Urban and Rural Areas," *National Air Quality and Emissions Trends Report*, pp. 13-23.

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

U.S. Environmental Protection Agency (1997), "National Ambient Air Quality Standards for Particulate Matter; Final Rule, Part II," *Federal Register*, 40, CFR Part 50.

U.S. Environmental Protection Agency (2000), *Quality Assurance Guidance Document, Quality assurance Project Plan: $PM_{2.5}$ Speciation Trends Network Field Sampling*, EPA 454/R-01-001, RTP, NC, Available at: http://www.epa.gov/ttn/amtic/files/ambient/pm25/spec/1025sqap.pdf.

U.S. Environmental Protection Agency (2003), *National Air Quality and Emissions Trends Report, 2003*, Special Studies Edition, EPA 454/R-03-005, RTP, NC, Available at: http://www.epa.gov/air/airtrends/aqtrnd03/.

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

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

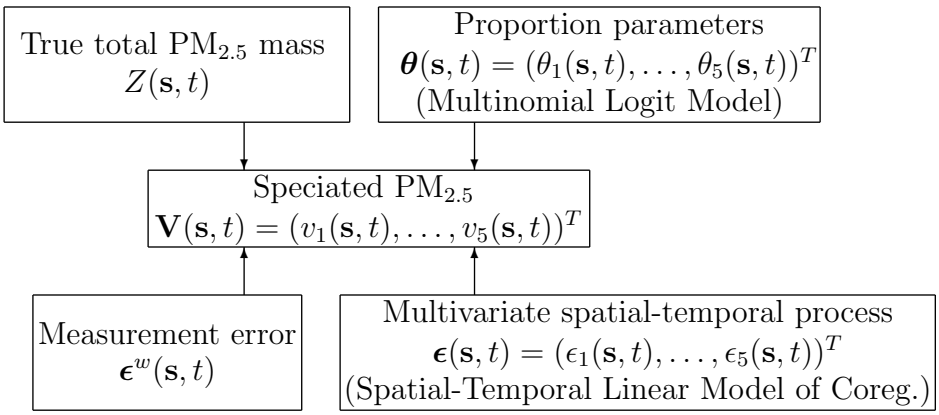
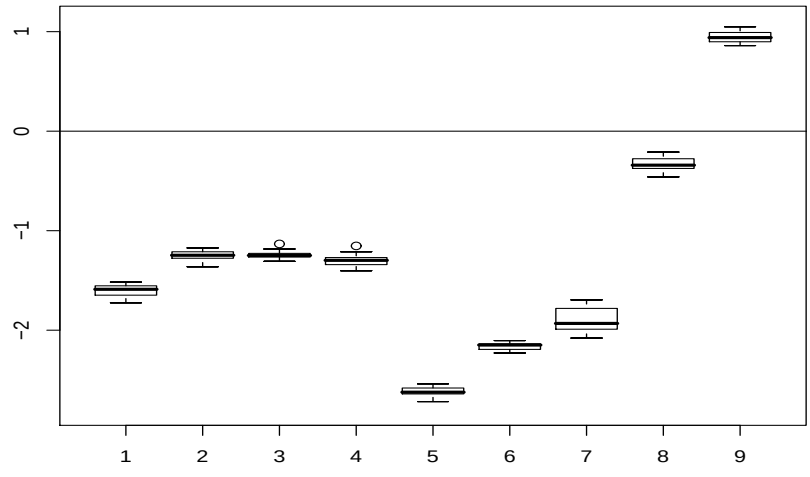
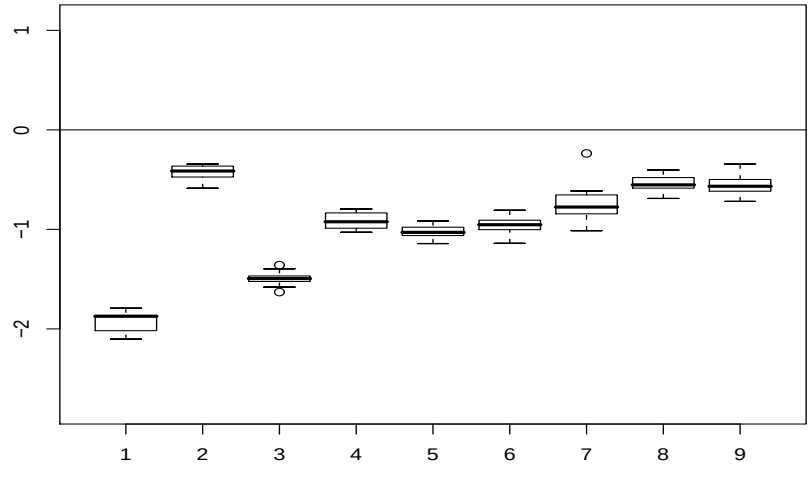


Figure 1: Model framework for speciated $PM_{2.5}$.

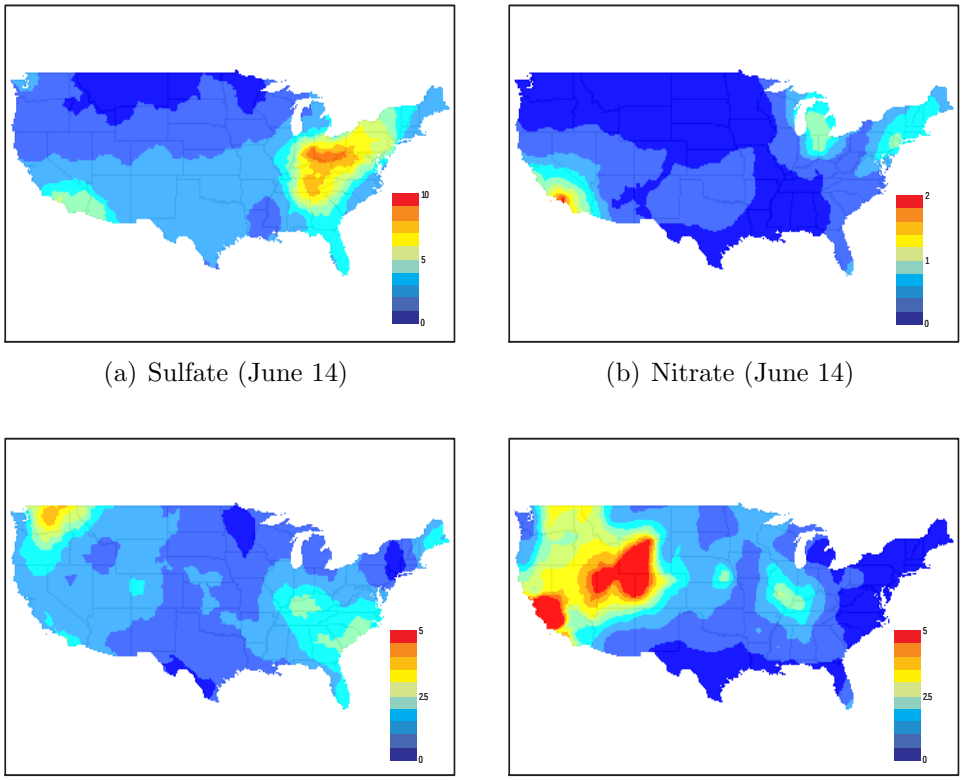


(a) July 2004



(b) December 2004

Figure 2: Boxplots of the additive bias of the RCFM process for (a) July 2004 and (b) December 2004 by geographic region (as defined by the U.S. Census). Region (1): Northeast (New England); (2): Northeast (Middle Atlantic); (3): Midwest (East North Central); (4): Midwest (West North Central); (5): South (South Atlantic); (6): South (East South Central); (7): South (West South Central); (8): West (Mountain); (9): West (Pacific).



(a) Sulfate (June 14)

(b) Nitrate (June 14)

(c) Sulfate (December 14)

(d) Nitrate (December 14)

Figure 3: Maps of the posterior median of sulfate concentration ($\mu g/m^3$) and nitrate concentration ($\mu g/m^3$) on June 14, 2004 and on December 14, 2004, respectively.

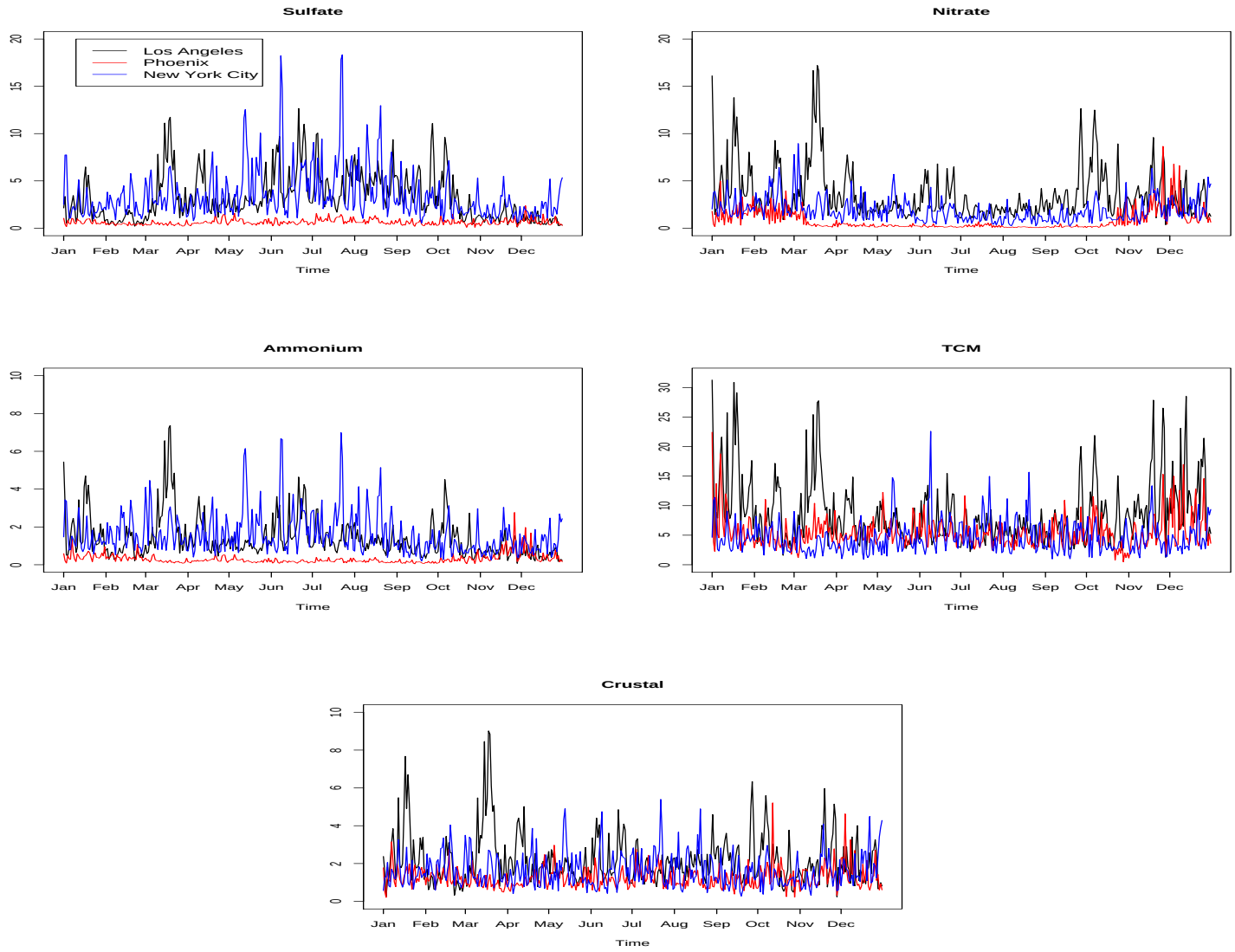
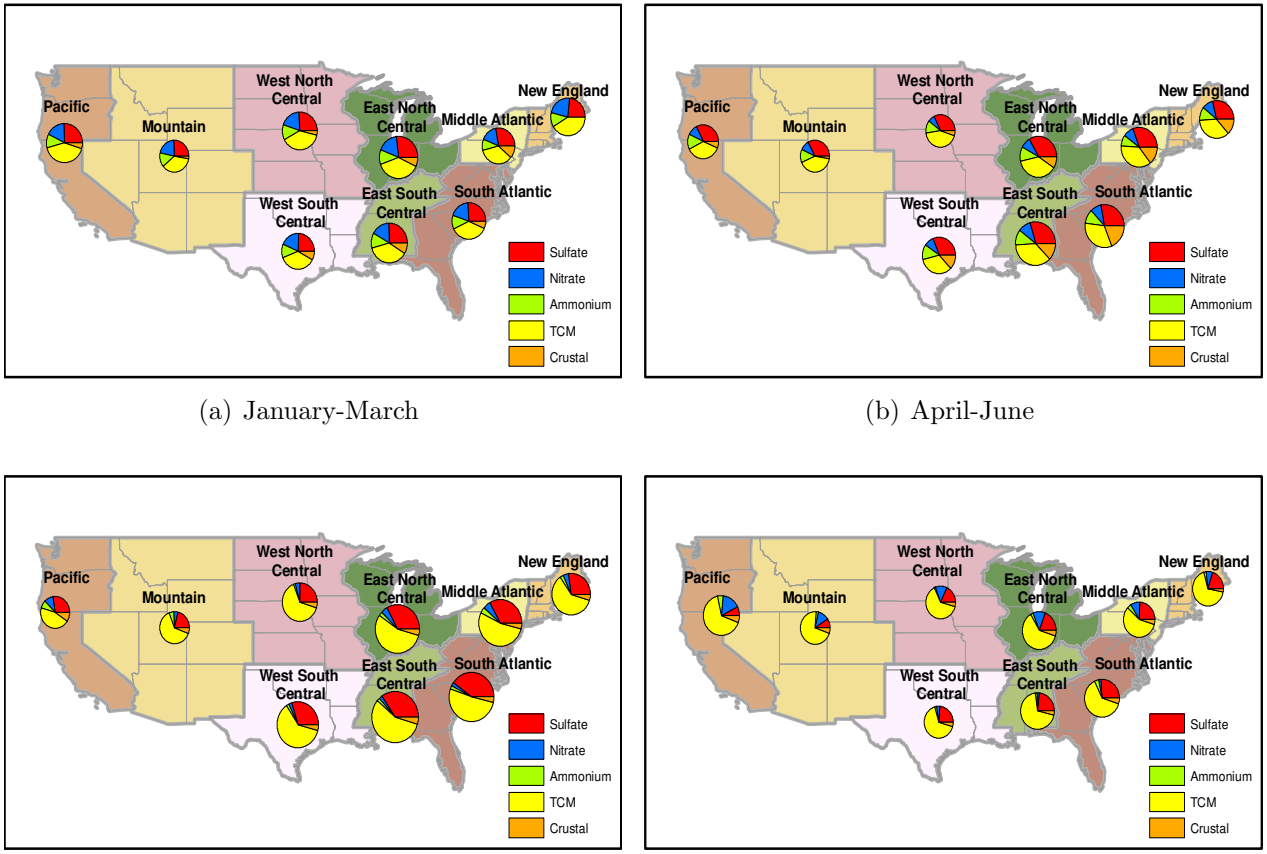


Figure 4: Time series plots of the estimated speciated $PM_{2.5}$ ($\mu g/m^3$) for three cities (Los Angeles, Phoenix, and New York City) in 2004.



(a) January-March

(b) April-June

(c) July-September

(d) October-December

Figure 5: Maps of estimated speciated $PM_{2.5}$ composition by region and by season in 2004.

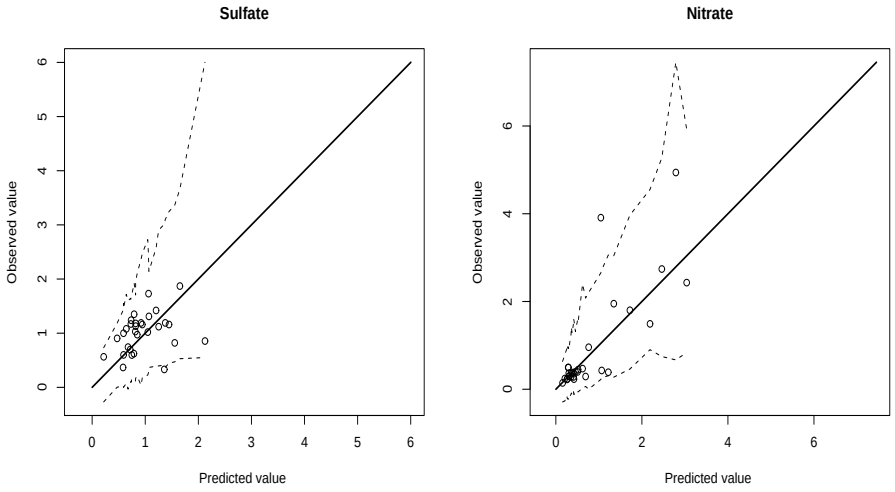


Figure 6: Model diagnostics for sulfate and nitrate in Phoenix: STN values of each component versus the median of the predictive posterior distribution of the component values at time j eliminating the i^{th} observation. The dotted lines show the 95% prediction intervals. The percentages of the observed values that are outside the prediction intervals are 3.3%.

Table 1: Summary of the posterior distributions for the **T** matrix in California in 2004: The labelling is 1, sulfate, 2, nitrate, 3, ammonium, 4, total carbonaceous mass, 5, crustal material.

Parameters	Mean	SD	2.5%	Median	97.5%
$T_{12}/\sqrt{T_{11}T_{22}}$	0.757	0.031	0.718	0.776	0.800
$T_{13}/\sqrt{T_{11}T_{33}}$	0.438	0.167	0.262	0.601	0.609
$T_{14}/\sqrt{T_{11}T_{44}}$	-0.085	0.056	-0.217	-0.055	-0.014
$T_{15}/\sqrt{T_{11}T_{55}}$	0.292	0.250	0.036	0.481	0.585
$T_{23}/\sqrt{T_{22}T_{33}}$	-0.288	0.075	-0.381	-0.230	-0.198
$T_{24}/\sqrt{T_{22}T_{44}}$	0.234	0.336	-0.189	0.550	0.581
$T_{25}/\sqrt{T_{22}T_{55}}$	-0.440	0.049	-0.501	-0.456	-0.345
$T_{34}/\sqrt{T_{33}T_{44}}$	-0.875	0.039	-0.932	-0.880	-0.790
$T_{35}/\sqrt{T_{33}T_{55}}$	0.786	0.115	0.640	0.891	0.905
$T_{45}/\sqrt{T_{44}T_{55}}$	-0.857	0.046	-0.921	-0.845	-0.791