Multivariate Receptor Models for Spatially Correlated Multi-Pollutant Data

Mikyoung Jun[1] and Eun Sug Park[2]

Abstract:

The goal of multivariate receptor modeling is to estimate the profiles of major pollution sources and quantify their impacts based on ambient measurements of pollutants. Traditionally, multivariate receptor modeling has been applied to multiple air pollutant data measured at a single monitoring site or measurements of a single pollutant collected at multiple monitoring sites. Despite the growing availability of multi-pollutant data collected from multiple monitoring sites, there has not yet been any attempt to incorporate spatial dependence that may exist in such data into multivariate receptor modeling. We propose a spatial statistics extension of multivariate receptor models that enables us to incorporate spatial dependence into estimation of source composition profiles and contributions. The proposed method yields more precise estimates of source profiles by accounting for spatial dependence in the estimation. In addition, it enables predictions of source contributions at unmonitored sites as well as monitoring sites when there are missing values. The method is illustrated with the simulated data and real multi-pollutant data collected from 8 monitoring sites in Harris County, Texas.

Key worlds: Multiple air pollutants; Multiple monitoring sites; Source apportionment; Source composition profile; Source contributions; Spatial correlation.

[1]Department of Statistics, Texas A&M University, College Station, TX 77843-3143

[2]Texas Transportation Institute, The Texas A&M University System, College Station, TX 77843-3135, Corresponding author. Tel: +1 979 845 9942; Fax: +1 979 845 6008; E-mail: e-park@tamu.edu

1. INTRODUCTION

Receptor modeling is a collection of methods for identifying major pollution sources and estimating the contribution of each source based on ambient measurements of air pollutants obtained at a given monitoring site, or a receptor. A comprehensive review of the field of receptor modeling can be found in Hopke (1991, 2003). Traditionally, multivariate receptor models have been used to resolve the observed air pollutant mixtures into contributions from individual sources (or source types) based on time series of multiple (or multivariate) air pollutants, such as Volatile Organic Compounds (VOCs) or specific metal constituents of fine particulate matter ($PM_{2.5}$), at a receptor site (see e.g., Hopke 1985; Henry 1997a; Park, Guttorp, and Henry 2001; Wolbers and Stahel 2005; Hopke et al. 2006; Heaton and Christensen 2010).

A basic multivariate receptor model takes the form of

$$Y_{jt} = \sum_{k=1}^{q} P_{jk} G_{kt} + E_{jt}, \tag{1}$$

where Y_{jt} is the mass concentration of pollutant j ($j = 1, \ldots, p$) measured at time t ($t = 1, \ldots, T$), q is the number of major pollution sources, P_{jk} is the relative concentration of pollutant j in source k ($k = 1, \ldots, q$), G_{kt} is the mass concentration (contribution) of source k at time t, and E_{jt} is the error associated with the jth pollutant concentration measured at time t. In matrix terms, model in (1) can be written as

$$\mathbf{Y} = \mathbf{PG} + \mathbf{E}, \tag{2}$$

where $\mathbf{Y}$ is a p by T data matrix containing T concentrations of p pollutants at a receptor, $\mathbf{P}$ is the p by q source composition matrix (where each column, a source composition profile, can be considered as a chemical fingerprint for a source), $\mathbf{G}$ is the q by T source contribution matrix, and $\mathbf{E}$ is an p by T error matrix. In relation to statistical models, this may be viewed as a factor analysis model or latent variable model (see Park, Oh, and Guttorp 2002) in the sense that $\mathbf{Y}$ is the only observable quantity whereas q (number of factors), $\mathbf{P}$ (factor loading matrix), and $\mathbf{G}$ (factor score matrix) are all unknown quantities that need to be estimated (or predicted). The usual challenges in factor analysis such as the unknown number of factors (sources) and non-identifiability of parameters (i.e., there are an infinite number of solutions to (2)) are also encountered in multivariate receptor models.

Various forms of factor analysis or principal component analysis methods have been applied in multivariate receptor modeling for more than three decades. Among several methods, Positive Matrix Factorization (PMF, Paatero and Tapper 1994; Paatero 1997) and UNMIX (Henry and Kim 1990; Kim and Henry 1999, 2000) gained most popularity among environmental engineers and scientists and have been widely used in practice. Until recently, there have been relatively few contributions by statisticians to the field of multivariate receptor modeling. See Pollice (2009) for a review of multivariate receptor modeling from a statistical perspective. Park et al. (2001) proposed time series extension of multivariate receptor models to account for temporal correlation in air pollution data into parameter estimation under a confirmatory factor analysis model. Billheimer (2001) developed compositional receptor modeling assuming that the source contributions and the errors are logistic normally distributed. Christensen and Sain (2002) developed an approach to account for temporal dependence, a nested block bootstrap method, in multivariate receptor modeling. Park, Spiegelman, and Henry (2002) proposed new sets of realistic identifiability conditions for multivariate receptor models and a constrained nonlinear least squares (CNLS) approach for parameter estimation. In Park, Oh, and Guttorp (2002) and Park, Guttorp, and Kim (2004), the unknown number of pollution sources and unknown identifiability conditions have been taken into account in the form of model uncertainty using a Bayesian approach. Gajewski and Spiegelman (2004) developed estimators that are robust to outliers. Wolbers and Stahel (2005) proposed the lognormal structural mixing model assuming a multiplicative error structure. Christensen, Schauer, and Lingwall (2006) developed an iterated confirmatory factor analysis approach to source apportionment. Spiegelman and Park (2007) performed a jackknife evaluation of the uncertainty of the estimates of the source contribution and source composition matrices as a way of incorporating dependence in air pollution data into estimation. Lingwall, Christensen, and Reese (2008) developed Dirichlet based Bayesian multivariate receptor modeling, and Heaton and Christensen (2010) proposed a Dirichlet Process model to incorporate time-varying source profiles in multivariate receptor models. Nikolov, Coull, Catalano, and Godleski (2010) extended the multiplicative factor analysis model proposed by Wolbers and Stahel (2005) by imposing mixed models on the latent source contributions to include the covariate effects and to adjust for temporal correlation in the source contribution.

In all of the previous approaches, however, multivariate receptor models were applied to

multiple air pollutant data measured at a single monitoring site or to single pollutant data (e.g., non-speciated $PM_{2.5}$) collected from multiple monitoring sites (see e.g., Henry 1997b; Park, Spiegelman, and Henry 2002; Park, Oh, and Guttorp 2002; Park et al. 2004). Despite the growing availability of the multi-pollutant data collected from multiple monitoring sites, the method that can jointly analyze such data is lacking in receptor modeling. Previous studies on source identification and apportionment employed a conventional multivariate receptor modeling approach to analyze the multi-pollutant data at each site separately (see, e.g., Buzcu and Fraser 2006) and ignored spatial correlations in the data. Incorporating spatial correlations in the multi-pollutant data collected from multiple monitoring sites into multivariate receptor modeling has been an open problem for many years (Park et al. 2001; Park et al. 2004; Pollice 2009).

In this paper we propose a spatial statistics extension of multivariate receptor models that enables us to incorporate spatial dependence into estimation of source composition profiles and contributions. We not only account for spatial dependence of each source contribution, but also account for the cross covariance of pairs of source contributions.

Recently spatial covariance models for multivariate processes have gotten attention in spatial statistics community and a few approaches for multivariate covariance models have been developed (e.g. Goulard and Voltz 1992; Wackernagel 2003; Gneiting, Kleiber, and Schlather 2010; Apanasovich and Genton 2010). The most traditional method is so called, Linear Model of Coregionalization (LMC) (Goulard and Voltz 1992; Wackernagel 2003). Gneiting et al. (2010) developed a multivariate version of Matérn covariance function. In this paper, we use the multivariate Matérn model for fitting the multivariate receptor model (see Section 2.1 for more details). The LMC model is used to simulate spatially dependent multivariate source contributions in Section 3.

Accounting for spatial dependence of multivariate air pollution data in source identification and apportionment will lead to more efficient estimation of source profiles and contributions. In addition, it will enable prediction of pollutant concentration and source contributions at locations other than the monitoring sites. Section 2 introduces a spatial model in multivariate receptor modeling for multi-pollutant data measured from spatially dispersed monitoring sites. Sections 3 contains a discussion of the performance of the spatial model as compared to a model not accounting for spatial dependence based on several simulated

datasets. Section 4 presents a real application to the Harris County air pollution data. Finally, concluding remarks are made in Section 5.

2. METHOD

We consider an extension of the models in (1) and (2) for the problem of the multiple pollutants over multiple spatial locations and time points. We write $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_p)^{'}$ (a $p \times NT$ matrix) with $\mathbf{Y}_j$ a vector of size $NT \times 1$, when N is the number of spatial locations and T is the number of time points. In particular, we write

$$\mathbf{Y}_j = (Y_j(\mathbf{s}_1, t_1), \ldots, Y_j(\mathbf{s}_N, t_1), Y_j(\mathbf{s}_1, t_2), \ldots, Y_j(\mathbf{s}_N, t_T))^{'}.$$

Note $\mathbf{s}_i$ and t_i give the spatial location and time point, respectively. We also write $\mathbf{G}$ and $\mathbf{E}$ in a similar way as in (2) except that $\mathbf{G}$ is a $q \times NT$ matrix and $\mathbf{E}$ is a $p \times NT$ matrix. The source composition matrix $\mathbf{P}$ is a $p \times q$ matrix. The number of major pollution sources, q, is assumed known throughout the paper. The ordering of spatial and temporal points in the rows of $\mathbf{G}$ and $\mathbf{E}$ are the same as that of $\mathbf{Y}_j$'s. We assume $\mathbf{G}$ is a Gaussian random field and $\mathbf{E}$ is a mean zero Gaussian white noise (except that each row of $\mathbf{E}$ has its own variance). Columns of $\mathbf{G}$ have a common mean vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^{'} \in \mathbb{R}^q$. The challenge here is to model the covariance structure of $\mathbf{G}$. We assume temporal independence of $\mathbf{G}$ (and thus temporal independence of $\mathbf{Y}$) throughout the paper (discussion of extension of temporally correlated model is given in Section 5). We consider $\mathbf{G}$ as a multivariate spatio-temporal process and focus on modeling the spatial dependence within each row as well as cross covariance across rows of $\mathbf{G}$.

We assume that $\mathbf{G}$ and $\mathbf{E}$ are multivariate stochastic processes varying over space and time. At each spatial location $\mathbf{s}$ and time t, we extend the model in (1) as

$$\begin{aligned}
Y_1(\mathbf{s}, t) &= P_{11}G_1(\mathbf{s}, t) + \cdots + P_{1q}G_q(\mathbf{s}, t) + E_1(\mathbf{s}, t), \\
Y_2(\mathbf{s}, t) &= P_{21}G_1(\mathbf{s}, t) + \cdots + P_{2q}G_q(\mathbf{s}, t) + E_2(\mathbf{s}, t), \\
&\cdots \\
Y_p(\mathbf{s}, t) &= P_{p1}G_1(\mathbf{s}, t) + \cdots + P_{pq}G_q(\mathbf{s}, t) + E_p(\mathbf{s}, t).
\end{aligned} \tag{3}$$

Under the above model, the mean of Y_j, $\mu_j = E\{Y_j(\mathbf{s}, t)\} = \sum_{k=1}^{q} P_{jk}\xi_k$, is constant across

space and time for each $j = 1, \ldots, p$. We focus on building joint covariance models for G_i's. The process E_j is modeled as a white noise with $\text{Var}(E_j) = \eta_j^2$.

We estimate parameters using maximum likelihood estimation method through the model,

$$\mathbb{Y} = (\mathbf{Y}_1^{'}, \ldots, \mathbf{Y}_p^{'})^{'} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{V}),$$

with

$\boldsymbol{\mu} = \mathbf{M}\boldsymbol{\xi}$, an $NTp \times 1$ vector,

$$\mathbf{M} = \begin{pmatrix} P_{11}\mathbf{1}_{NT} & P_{12}\mathbf{1}_{NT} & \cdots & P_{1q}\mathbf{1}_{NT} \\ \vdots & \vdots & \vdots & \vdots \\ P_{p1}\mathbf{1}_{NT} & P_{p2}\mathbf{1}_{NT} & \cdots & P_{pq}\mathbf{1}_{NT} \end{pmatrix}, \text{ an } NTp \times q \text{ matrix},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\Sigma}_{p1} & \boldsymbol{\Sigma}_{p2} & \cdots & \boldsymbol{\Sigma}_{pp} \end{pmatrix}, \text{ an } NTp \times NTp \text{ matrix},$$

$$\boldsymbol{\Sigma}_{jh} = \begin{pmatrix} \boldsymbol{\Sigma}_{jh}^{(1)} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{jh}^{(T)} \end{pmatrix}, \text{ an } NT \times NT \text{ block diagonal matrix } (j, h = 1, \ldots, p),$$

$\boldsymbol{\Sigma}_{jh}^{(t)}$, an $N \times N$ matrix, the spatial covariance matrix of $\mathbf{Y}_j^{(t)}$ and $\mathbf{Y}_h^{(t)}$, $\mathbf{Y}_j^{(t)} = (Y_j(\mathbf{s}_1, t), \ldots, Y_j(\mathbf{s}_N, t))^{'}$,

and $\mathbf{V} = \begin{pmatrix} \eta_1\mathbf{I}_{NT} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \eta_p\mathbf{I}_{NT} \end{pmatrix}$, an $NTp \times NTp$ diagonal matrix.

Here, $\mathbf{1}_{NT}$ is a vector of ones with dimension $NT \times 1$ and $\mathbf{I}_{NT}$ is an identity matrix of dimension $NT \times NT$. The matrix $\boldsymbol{\Sigma}_{jh}$ is a block diagonal matrix due to the temporal independence assumption. At time t, the (l,m) element of $\boldsymbol{\Sigma}_{jh}^{(t)}$ is given by

$$\text{Cov}\{Y_j(\mathbf{s}_l, t), Y_h(\mathbf{s}_m, t)\} = \sum_{i,k=1}^{q} P_{ji}P_{hk}\text{Cov}\{G_i(\mathbf{s}_l, t), G_k(\mathbf{s}_m, t)\} + \eta_j^2\mathbf{1}_{(j=h)}, \tag{4}$$

and is free of t. That is, we assume that $\boldsymbol{\Sigma}_{jh}^{(t)}$ is the same across $t = 1, \ldots, T$. Note that for the mean of $\mathbb{Y}$, we estimate $\mathbf{P}$ and $\boldsymbol{\xi}$ and then obtain the $NTp \times 1$ vector, $\boldsymbol{\mu}$.

2.1 Spatial model for multiple source contributions

To model the covariance structure of G_i's (that is, $\text{Cov}\{G_i(\mathbf{s}_l, t), G_k(\mathbf{s}_m, t)\}$), we consider the following two models. The first model ignores the spatial dependence of G_i's but only

considers the cross covariance of G_i's (we call this spatially independent model, SIM) and the second model accounts for the spatial dependence of G_i's as well as their cross covariances (we call this spatially dependent model, SDM). The model SDM not only accounts for the spatial dependence of individual G_i's but also the spatial dependence of cross covariance of pairs of G_i's.

Under the SIM, it is easy to see from (4) that, for each t, $\mathbf{\Sigma}_{jh}^{(t)}$ is a diagonal matrix. Under the SDM, on the other hand, $\mathbf{\Sigma}_{jh}^{(t)}$ is no longer diagonal and we need a spatial covariance model for a multivariate spatial process. We use the multivariate Matérn model developed by Gneiting et al. (2010). At each time point t, we assume the multivariate process $\{G_1(\mathbf{s},t),\ldots,G_q(\mathbf{s},t)\}$ is spatially isotropic and temporally independent, and let

$$C_{ik}(|\mathbf{s}_1-\mathbf{s}_2|)=\mathrm{Cov}\{G_i(\mathbf{s}_1,t),G_k(\mathbf{s}_2,t)\}.$$

The covariance function $C_{ik}(\cdot)$ does not depend on t. The multivariate Matérn model gives $C_{ii}(d)=\sigma_i^2 M(d|\nu_i,\beta)$, $C_{ik}(d)=\sigma_i\sigma_k\rho_{ik}M(d|\nu_{ik},\beta)$ for $1\le i\ne k\le q$ and

$$M(d|\nu,\beta)=\frac{1}{2^{\nu-1}\Gamma(\nu)}(\frac{d}{\beta})^{\nu}\mathcal{K}_{\nu}(\frac{d}{\beta}), \tag{5}$$

with $\mathcal{K}_\nu$ the modified Bessel function of order ν. Here σ_i,β, and ν_i are the covariance parameters to be estimated ($\sigma_i,\beta,\nu_i>0$), $\nu_{ik}=\frac{\nu_i+\nu_k}{2}$, and ρ_{ik} ($-1\le\rho_{ik}\le 1$) is the co-located correlation coefficient. Theorem 1 of Gneiting et al. (2010) shows that if we let

$$\rho_{ik}=\gamma_{ik}\frac{\Gamma(\nu_i+\frac{3}{2})^{1/2}}{\Gamma(\nu_i)^{1/2}}\frac{\Gamma(\nu_k+\frac{3}{2})^{1/2}}{\Gamma(\nu_k)^{1/2}}\frac{\Gamma(\nu_{ik})}{\Gamma(\nu_{ik}+\frac{3}{2})},$$

where the matrix $(\gamma_{ik})_{i,k=1}^q$ (with diagonal elements $\gamma_{ii}=1$ for $i=1,\ldots,q$ and off-diagonal elements γ_{ik} for $1\le i\ne k\le q$) is symmetric and nonnegative definite, then C_{ii} and C_{ik} together give a valid covariance model for the multivariate process G_i's.

The spatial range parameter, β, determines how far the spatial correlation of the multivariate spatial process lasts. Larger β gives longer range of spatial correlation structure. The smoothness parameter, ν_i, controls the smoothness of the multivariate spatial process (the larger ν_i is, the smoother the ith process is). The parameter σ_i controls the covariance level of the ith process. The co-located correlation coefficient determines the strength of cross-correlation of the multivariate process.

One limitation of the above covariance model is that each G_i has the same spatial range parameter, β. However, the number of monitoring sites for pollutants typically used in receptor modeling (e.g., VOCs or PMs) is usually moderate to small (e.g., less than 15) and thus the data may not provide enough information to estimate all of the spatial covariance parameters, in particular the spatial range and smoothness parameters. Moreover, as Zhang (2004) points out, in general not all of the parameters in Matérn covariance model are consistently estimable. Therefore, the above model may not be limited for the receptor modeling problems.

2.2 Constraint on source composition matrix

It is well-known that parameters of models in (1) and (2) are not uniquely defined without imposing additional constraints on them. The same non-identifiability problem continues to be manifest in model (3). To avoid nonidentifiability of multivariate receptor models, we enforce additional constraints on either $\mathbf{P}$ or $\mathbf{G}$ matrix (called 'identifiability conditions'). See Park, Spiegelman, and Henry (2002) for identifiability conditions that are meaningful in multivariate receptor models. Here, we employ identifiability conditions on $\mathbf{P}$ that are often used in receptor modeling. One set of such conditions is:

C1 There are at least $q-1$ zero elements in each column of $\mathbf{P}$.

C2 The rank of $\mathbf{P}^{[k]}$ is $q-1$, where $\mathbf{P}^{[k]}$ is the matrix composed of the rows containing the assigned zeros in the kth column with those assigned zeros deleted.

C3 $\sum_{j=1}^{p} P_{jk} = 1$ for each $k = 1, \ldots, q$.

The conditions C1-C2 imply that some pollutants (corresponding to zeros in $\mathbf{P}$) are not contributed by a particular source type, and no two sources share the exactly same set of zeros. These are the same conditions as those used in confirmatory factor analysis to remove factor indeterminacy problem (see, for example, Anderson (1984), Chapter 14.2.2). Note that pre-specification of zero elements in $\mathbf{P}$ requires of the investigator some prior knowledge on the source types (that might be obtained from previous studies or exploratory analyses). The normalization constraint C3 is enforced to remove the multiplication of a column of $\mathbf{P}$ by a scale constant, which is enough for the purpose of receptor modeling.

2.3 Estimation and spatial prediction of multiple source contributions

We now describe how to estimate (or predict) the source contributions at any spatial location and time (say $\mathbf{s}_0$ and t_0) under the SDM, based on the conditional distribution, $G_i(\mathbf{s}_0, t_0)|\mathbb{Y}$, following Chapter 14.7 of Anderson (1984). These spatial location and time may or may not be where we have the observations.

Since we assume G_i's and E_i's are Gaussian, $\mathbf{G}(\mathbf{s}_0, t_0) \in \mathbb{R}^q$ and $\mathbb{Y} \in \mathbb{R}^{NTp}$ are jointly normally distributed with the mean vector $(\boldsymbol{\xi}', \boldsymbol{\mu}')'$ and the covariance matrix

$$\begin{pmatrix} \boldsymbol{\Phi} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}' & \boldsymbol{\Sigma} + \mathbf{V} \end{pmatrix}.$$

Here, $\boldsymbol{\xi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma} + \mathbf{V}$ are defined in Section 2.1 and $\boldsymbol{\Phi}$ denotes the $q \times q$ covariance matrix of $\mathbf{G}(\mathbf{s}_0, t_0)$. The matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{q \times NTp}$ is the cross covariance matrix of $\mathbf{G}(\mathbf{s}_0, t_0)$ and $\mathbb{Y}$. Then, the conditional distribution, $\mathbf{G}(\mathbf{s}_0, t_0)|\mathbb{Y}$, is Gaussian with mean given by

$$\mathrm{E}\{\mathbf{G}(\mathbf{s}_0, t_0)|\mathbb{Y}\} = \boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \mathbf{V})^{-1}(\mathbb{Y} - \boldsymbol{\mu}) + \boldsymbol{\xi}. \tag{6}$$

This is our estimator or predictor of the source contribution at ($\mathbf{s}_0$,t_0). Under the SIM, the same idea applies for the estimation and we use (6) to estimate the source contribution at the monitoring sites.

When we predict the source contribution at an unmonitored site under the SIM, however, $\boldsymbol{\Lambda}$ becomes a matrix consisting of only zeros due to the spatial independence assumption. Therefore, we take the average of the estimated source contributions at all of the monitoring sites at a given time point instead as a predicted source contribution at the unmonitored site at the given time point. Although we described the method for the situation where we have one spatial location $\mathbf{s}_0$ and time point t_0, it can easily be extended to multiple spatial locations and time points.

3. SIMULATION STUDY

In this section, we compare the performances of the SIM and the SDM in terms of estimation of source composition profiles and other parameters such as the mean of the source contributions and the error variances. We also compare the estimated (and predicted) source contributions at monitored (and unmonitored) sites from both methods. We set $p = 9$,

$q = 3$, $N = 8$, and $T = 100$. The spatial locations are generated at random (uniformly) over unit square (see Figure 1). We first generate the source contributions, G_i, $i = 1, \ldots, q$, and the error process, E_j, $j = 1, \ldots, p$, through Gaussian random fields. The error process, E_j, is iid normal (that is, independent over space and time) and its variances are set $(\eta_1^2, \ldots, \eta_p^2) = (0.03, 0.02, 0.03, 0.02, 0.01, 0.04, 0.02, 0.03, 0.03)$. The true source composition matrix, $\mathbf{P}_{true}$, is set

$$\mathbf{P}'_{true} = \begin{pmatrix} 0.1 & 0.05 & 0.25 & 0.1 & 0 & 0 & 0.3 & 0.1 & 0.1 \\ 0 & 0.4 & 0 & 0.1 & 0.1 & 0.05 & 0.1 & 0.05 & 0.2 \\ 0.1 & 0 & 0.05 & 0 & 0.1 & 0.4 & 0.05 & 0.2 & 0.1 \end{pmatrix}.$$

This matrix satisfies the conditions C1-C3 in Section 2.2. When we estimate the source composition matrix, we give the same constraints C1-C3 and we use the same pre-assigned locations of zeros as those in the true source composition matrix, $\mathbf{P}_{true}$. Therefore, regarding the source composition matrix, there are 21 nonzero elements (parameters) to be estimated.

We generate spatially dependent source contribution G_i's. To make a fair comparison between the SIM and the SDM, we do not use the covariance model for G_i's in the SDM to generate G_i's. Instead, we use the LMC model for the simulation in the following way. Consider q latent mean zero spatio-temporal processes and we denote them as W_i's. Here we assume W_i's are independent of each other but each W_i process has spatial dependence. For each G_i, we let

$$G_i = \xi_i + \alpha_{i,1} W_1 + \alpha_{i,2} W_2 + \cdots + \alpha_{i,q} W_q. \tag{7}$$

Although each W_i's are independent, the resulting G_i's are no longer independent and depending on how we set the coefficients $\alpha_{i,j}$'s, the cross-covariance structure of G_i's can be quite flexible. For the spatial dependence structure of W_i's, we use a Matérn covariance model in (5). That is, we let

$$\mathrm{Cov}\{W_i(\mathbf{s}_1, t_1), W_i(\mathbf{s}_2, t_2)\} = \begin{cases} M(|\mathbf{s}_1 - \mathbf{s}_2||\nu_i, \beta_i), & \text{if } t_1 = t_2 \\ 0, & \text{if } t_1 \neq t_2. \end{cases}$$

Therefore, the variances of W_i's are one. We set $\boldsymbol{\xi} = (4, 6, 10)'$, $\boldsymbol{\alpha} = (\alpha_{i,k}) = \begin{pmatrix} 1 & 0 & 0 \\ 0.1 & 0.995 & 0 \\ 0.5 & -0.553 & 0.667 \end{pmatrix}$

(so that the covariance matrix of G_i's at the same spatial location is $\boldsymbol{\alpha}\boldsymbol{\alpha}^{'} = \begin{pmatrix} 1 & 0.1 & 0.5 \\ 0.1 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{pmatrix}$), $\beta_1 = 0.3$, $\beta_2 = 0.4$, $\beta_3 = 0.5$, $\nu_1 = 0.5$, $\nu_2 = 1$, and $\nu_3 = 1.5$. Note that the covariance model used in the SDM is somewhat limited for the simulated data under this setting since we have distinct β_i's in the simulation while the multivariate Matérn model used for the SDM has common spatial range parameter, β. See the discussion in Section 2.1. We demonstrate later that despite such limitation, the SDM fits the data quite well and is significantly better compared to the SIM.

To assess the prediction performance of the SDM and the SIM for the source contributions at unmonitored sites, we simulate G_i's at 10 spatial locations. We use the data over 8 locations to fit the model and estimate parameters, and then predict the source contributions at the other 2 locations for model validation. We repeat the simulation 100 times and report the estimates and predictions along with their Mean Squared Errors (MSEs) and Mean Squared Prediction Errors (MSPEs) from both models. At every simulation, we make sure the numerical maximization of the loglikelihood for both models converge properly.

Figure 2 gives the biases and square root of MSEs (RMSEs) of the estimates for the 21 non-zero elements of $\mathbf{P}_{true}$. Both the SIM and the SDM models give unbiased estimates, but for several elements the biases of estimates from the SIM are quite large. The RMSEs from the SDM are consistently small (the largest is 0.026) and much smaller than those from the SIM. This demonstrates that the estimates for $\mathbf{P}_{true}$ from the SDM are much more efficient. We also compared the estimates of $\alpha_{i,k}$'s, ξ_i's, and η_i^2's for the two models, and for all of them the estimates from the SDM were much more efficient (results are omitted for the brevity of the paper).

The estimates of the covariance parameters, β and ν_i's, are given in Figure 3. Despite the limitation of the covariance model used for the SDM, the estimates for β are mostly within the range of the true range parameters, β_1, β_2, and β_3. We do not see much differences between the estimates for ν_2 and ν_3, although the true ν_3 is larger than the true ν_2. Note, though, that the smoothness parameters in the multivariate Matérn model and those in the LMC models are not quite comparable from the construction of the covariance models.

Next, we estimate (predict) the source contribution, G_i's, at the 8 sites used for the

modeling fitting as well as at the two unmonitored sites. Figure 4 gives the MSEs of the estimated source contributions under the two methods at 8 spatial locations. The estimates from the SDM are consistently more accurate across all 8 locations. For both methods, the errors are largest for source 1 and smallest for source 2. The MSE values are consistent across the 8 spatial locations for both methods. Table 1 gives MSPEs for the source contributions at the two sites. Clearly, the SDM gives much better prediction results, in particular for sources 1 and 3. Overall the errors at site 9 are bigger than those at site 10 and this is because, unlike site 10 that has a few nearby sites, site 9 does not have any nearby sites (see Figure 1). Both models have some trouble predicting source 1 at site 9. This may be due to the facts that site 9 is an isolated site and that the true source 1 has the smallest spatial range ($\beta_1 = 0.3$). Thus even under the SDM, we cannot borrow much information from the neighboring sites for source 1 at site 9.

4. APPLICATION TO REAL DATA

The method developed in the paper, the SDM, has been applied to the 24-hour Volatile Organic Compounds (VOC) data collected every 6 days from 8 monitoring sites in Harris County during January 1/1/2000 - 8/29/2009. Figure 5 shows the locations of the 8 monitoring sites used in this study.

The first important step in multivariate receptor modeling is to select an appropriate subset of species for an analysis; inclusion of noisy or unhelpful species could hinder source apportionment (Park et al. 2001). Ten VOC species (names listed in Table 2) that are major compounds at the sites considered (in Figure 5) were selected from 107 VOC species originally measured. There were a total of 669 days when VOC measurements were made for at least one of the 8 monitoring sites. The number of non-missing observations at each site ranges from 521 to 553, which implies that there were typically more than 100 missing observations (days with no VOC measurements) at each site during the study period. Figure 6 gives the location of missing observations for each site over time.

To build a reasonable multivariate receptor model in terms of the number of major pollution sources for the area and the identifiability conditions, the exploratory data analysis at each site preceded the analysis combining data from all 8 sites together. Based on the previ-

ous studies on the region (e.g., Buzcu and Fraser 2006), refineries, petrochemical production facilities, gasoline and natural gas/accumulation emissions were presumed to be the four most important sources affecting the region. This prior knowledge was utilized in pre-specification of zeros in the source composition profile matrix to achieve model identifiability as well as in selecting the appropriate subset of species that are contributed by those sources (see Table 2 for the pre-specification of zeros in the source composition matrix). Table 3 gives the major compounds for each of the four aforementioned sources.

We now fit the SDM model to the data to estimate the source composition matrix, mean and covariance parameters for source contributions along with the error variances. We then predict the source contribution at an unmonitored site. In fitting the model, we only use available observations and no imputation is performed.

Table 2 gives the estimated source composition matrix, P_{jk}'s, along with the means (ξ_i's) and standard deviations (σ_i's) of source contributions and the error standard deviations (η_i's). The table also provides the asymptotic standard errors of the estimates based on the inverse of the Hessian matrix of the loglikelihood function evaluated at the MLE estimates. For three parameters, P_{22} (source composition for Ethane from petrochemical production), η_5 (error standard deviation of Isopentane), and η_6 (error standard deviation of Propane), the estimates were too small and we were not able to obtain the asymptotic standard errors numerically. Overall, the estimated source composition profiles appear to be consistent with presumed four major sources for the region in terms of major compounds. No prior information from presumed sources, other than pre-assigned zeros (assuming that the species corresponding to pre-assigned zeros are not present in the emissions from that source), was used in fitting the source composition matrix of Table 2.

The estimated mean source contributions indicate that overall refineries and petrochemical production facilities play a major role in VOC emissions for the region and this agrees with the result in Buzcu and Fraser (2006). Estimated standard deviations for contributions from gasoline and natural gas are rather large compared to their estimated means.

For the spatial covariance parameters, we get $\hat{\beta}$=33.812 (6.061), $\hat{\nu}_1$=1.244 (0.326), $\hat{\nu}_2$=0.164 (0.026), $\hat{\nu}_3$=0.005 (0.001), and $\hat{\nu}_4$=0.126 (0.024). Here the order of the sources is the same as in Table 2. The numbers in parentheses are the asymptotic standard errors. The unit for $\hat{\beta}$ is Km. The estimated spatial range of roughly 33 Km is reasonable considering the size of the

spatial domain considered. Based on the estimated smoothness parameter values, gasoline gives the roughest spatial process ($\hat{\nu}_3$) and refinery gives the smoothest spatial process ($\hat{\nu}_1$). The estimated co-located correlation coefficients between source contributions is given by

$$(\hat{\rho}_{ik})_{i,j=1,\ldots,4} = \begin{pmatrix} 1.000 & 0.614 & 0.120 & 0.117 \\ 0.614 & 1.000 & 0.341 & 0.076 \\ 0.120 & 0.341 & 1.000 & 0.030 \\ 0.117 & 0.076 & 0.030 & 1.000 \end{pmatrix}.$$

The estimated cross-correlations are positive and mostly small except $\hat{\rho}_{12} = 0.614$, the co-located cross correlation between refinery and petrochemical production.

Now we estimate and predict the source contribution as described in Section 2.3. Figure 7 gives the time series plots of source contributions at site 2 (HRM-3 site), located to the south of a major interstate highway. Overall the estimated time series of the source contribution look reasonable. The contribution of evaporative gasoline emissions at this site is much higher than that can be anticipated from the overall mean contribution of gasoline for the entire region in Table 2. In fact, it is consistent with the observation of Buzcu and Fraser (2006) that the evaporative gasoline factor was a major contributor to VOC emissions together with the refinery factor at the HRM-3 site. It is worth to point out that even if we have missing observations on several days at the HRM-3 site, we can still estimate the source contribution for those days since we are borrowing information from the neighboring sites considering spatial dependence in estimating $\mathbf{G}$. This is a clear advantage of spatial modeling conducting the simultaneous analysis at all sites, rather than conducting a one-site-at-a-time analysis. We can obtain predictions of $\mathbf{G}$ at all of 669 time points at the HRM-3 site, although the HRM-3 site contains the observations only for 556 days.

Figure 8 gives the time series plots of the predicted source contributions at an unmonitored site given in Figure 5. This location belongs to super neighborhood in Houston, and while no monitoring site is available, air pollution epidemiologists or people in charge of developing air quality management plans may desire to know contributions of sources at such location. We can see that while the contributions of refineries and petrochemical production facilities are still in the same order of magnitude, the contribution from gasoline at this location is much smaller compared to that of the HRM-3 site, as expected.

5. CONCLUDING REMARKS

We have presented a new multivariate receptor modeling approach that can incorporate spatial dependence in the multiple pollutant data collected from multiple monitoring sites into estimation of source composition profiles and prediction of source contributions. The proposed method resulted in more precise estimates of source profiles by accounting for spatial dependence in the estimation. More importantly, it enabled predictions of source contributions when pollution measurements were not made at a specific monitoring site or even at an unmonitored site. These predicted source contributions can greatly enhance air pollution epidemiological studies and facilitate development of an effective air quality management plan by quantifying environmental impacts of pollution sources where no monitoring sites are available.

There are several possible directions for future work. First, we assumed isotropic covariance structure in our spatial model. When the spatial covariance structure of the multiple source contributions is nonisotropic or nonstationary, we may need to incorporate this into our covariance model for G_i's. Currently there are only a few such covariance models available. Jun (2009) gives a nonstationary cross-covariance models for multivariate spatial processes but the approach is geared towards global processes. The nonstationary version of the LMC model such as in Gelfand, Schmidt, Banerjee, and Sirmans (2004) may be applied to the situation, but the model may require quite a number of parameters and unless we have enough number of monitoring sites, the estimation of the parameters may be difficult. Currently the authors are pursuing the development of nonstationary covariance models for multivariate processes suitable for our problem.

Second, when pollutants are measured at hourly intervals, temporal dependence often exists in the data. In our spatial modeling, we assumed the independence of observations over time, which is typically satisfied when the data are measured at longer time intervals such as every 6 days. The spatial statistics extension of multivariate receptor modeling presented in this paper can be further generalized to account for spatio-temporal correlation in the data. In that case, we may use parametric spatio-temporal covariance functions for modeling the covariance structure of the source contributions. We could extend the multivariate version of Matérn covariance function used in this paper for spatio-temporal setting or we may consider

the covariance model developed in Apanasovich and Genton (2010).

Third, we assumed that the number of sources and model identifiability conditions are known or set a priori. When the number of sources and model identifiability conditions are unknown, accounting for such model uncertainty in multivariate receptor modeling is a challenging problem. Park, Oh, and Guttorp (2002) and Park et al. (2004) developed a Bayesian approach to account for model uncertainty in multivariate receptor models for the conventional multivariate receptor modeling data, i.e., for multiple pollutant data measured at a single monitoring site or a single pollutant data collected from multiple monitoring sites. Accounting for uncertainty in the number of sources and identifiability conditions in spatial multivariate receptor modeling such as the one developed in this paper is an important future research area.

Acknowledgments

Mikyoung Jun's research is supported by NSF grant DMS-0906532. Eun Sug Park's research is supported by a contract from Health Effects Institute (HEI), an organization jointly funded by the Environmental Protection Agency (EPA R824835) and automotive manufacturers. This publication is based in part on work supported by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). The contents of this article do not necessarily reflect the views and policies of HEI, nor do they necessarily reflect the views and Health Effects of Air Pollution 271 policies of EPA, or motor vehicles or engine manufacturers. The authors thank Ms. Hotchkiss, Dr. Jim Price, and Dr. Clifford Spiegelman for their help with the acquisition of the 24-hour canister VOC data.

REFERENCES

Anderson, T. (1984), *An introdiction to multivariate statistical analysis*: Wiley, New York.

Apanasovich, T. V. and Genton, M. G. (2010), "Cross-covariance functions for multivariate random fields based on latent dimensions," *Biometrika*, 97, 15–30.

Billheimer, D. (2001), "Compositional Receptor Modeling," *Environmetrics*, 12, 451–467.

Buzcu, B. and Fraser, M. (2006), "Source identification and apportionment of volatile organic compounds in Houston, TX," *Atmospheric Environment*, 40, 2385–2400.

Christensen, W. and Sain, S. (2002), "Accounting for dependence in a flexible multivariate receptor model," *Technometrics*, 44, 328–337.

Christensen, W., Schauer, J., and Lingwall, J. (2006), "Iterated confirmatory factor analysis for pollution source apportionment," *Environmetrics*, 17, 663–681.

Gajewski, B. and Spiegelman, C. (2004), "Correspondence Estimation of the Source Profiles in Receptor Modeling," *Environmetrics*, 15, 613–634.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004), "Nonstationary multivariate process modeling through spatially varying coregionalization," *Test*, 13, 263–312.

Gneiting, T., Kleiber, W., and Schlather, M. (2010), "Matérn cross-covariance functions for multivariate random fields," *Journal of the American Statistical Association*, 105, 1167–1177.

Goulard, M. and Voltz, M. (1992), "Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix," *Mathematical Geology*, 24, 269–282.

Heaton, M.J., R. C. and Christensen, W. (2010), "Incorporating time-dependent source profiles using the Dirichlet distribution in multivariate receptor models," *Technometrics*, 52, 67–79.

Henry, R. (1997a), "History and Fundamentals of Multivariate Air Quality Receptor Models," *Chemometrics and Intelligent Laboratory Systems*, 37, 37–42.

——— (1997b), "Receptor Model Applied to Patterns in Space (RMAPS) Part II - Apportionment of Airborne Particulate from Project MOHAVE 1997," *Journal of Air and Waste Management Association*, 47, 220–225.

Henry, R. and Kim, B. (1990), "Extension of self-modeling curve resolution to mixtures of more than three components. Part I: finding the basic feasible region," *Chemometrics and Intelligent Laboratory Systems*, 8, 205–216.

Hopke, P. (1985), *Receptor Modeling in Environmental Chemistry*, New York: Wiley.

——— (1991), "An Introduction to Receptor Modeling," *Chemometrics and Intelligent Laboratory Systems*, 10, 21–43.

——— (2003), "Recent developments in receptor modeling," *Chemometrics and Intelligent Laboratory Systems*, 17, 255–265.

Hopke, P., Ito, K., Mar, T., Christensen, W., Eatough, D., Henry, R., and et al. (2006), "PM source apportionment and health effects: 1. Intercomparison of source apportionment results," *J Expo Sci Environ Epidemiol*, 16, 275–286.

Jun, M. (2009), "Nonstationary cross-covariance models for multivariate processes on a globe," Technical Report 2009-110, IAMCS preprint series.

Kim, B. and Henry, R. (1999), "Extension of self-modeling curve resolution to mixtures of more than three components. Part II: finding the complete solution," *Chemometrics and Intelligent Laboratory Systems*, 49, 67–77.

——— (2000), "Extension of self-modeling curve resolution to mixtures of more than three components. Part III: atmospheric aerosol data simulation study," *Chemometrics and Intelligent Laboratory Systems*, 52, 145–154.

Lingwall, J., Christensen, W., and Reese, C. (2008), "Dirichlet based Bayesian multivariate receptor modeling," *Environmetrics*, 19, 618–629.

Nikolov, M., Coull, B., Catalano, P., and Godleski, J. (2010), "Multiplicative factor analysis with a latent mixed model structure for air pollution exposure assessment," *Environmetrics*, DOI: 10.1002/env.1039.

Paatero, P. (1997), "Least Squares Formulation of Robust, Non-Negative Factor Analysis," *Chemometrics and Intelligent Laboratory Systems*, 37, 23–35.

Paatero, P. and Tapper, U. (1994), "Positive Matrix Factorization: a nonnegative factor model with optimal utilization of error estimates of data values," *Environmetrics*, 5, 111–126.

Park, E., Guttorp, P., and Henry, R. (2001), "Multivariate receptor modeling for temporally correlated data by using MCMC," *Journal of the American Statistical Association*, 96, 1171–1183.

Park, E., Guttorp, P., and Kim, H. (2004), "Locating Major PM10 Source Areas In Seoul Using Multivariate Receptor Modelin," *Environmental and Ecological Statistics*, 11, 9–19.

Park, E., Oh, M., and Guttorp, P. (2002), "Multivariate receptor models and model uncertainty," *Chemometrics and Intelligent Laboratory Systems*, 60, 49–67.

Park, E., Spiegelman, C., and Henry, R. (2002), "Bilinear estimation of pollution source profiles and amounts by using multivariate receptor models," *Environmetrics*, 13, 775–798.

Pollice, A. (2009), "Recent statistical issues in multivariate receptor models," *Environmetrics*, DOI: 10.1002/env.1021.

Spiegelman, C. and Park, E. (2007), "A Computation Saving Jackknife Approach to Receptor Model Uncertainty Statements for Serially Correlated Data," *Chemometrics and Intelligent Laboratory Systems*, 88, 170–182.

Wackernagel, H. (2003), *Multivariate Geostatistics* (third ed.), Berlin: Springer-Verlag.

Wolbers, M. and Stahel, W. (2005), "Linear unmixing of multivariate observations: a structural model," *Journal of the American Statistical Association*, 100, 1328–1342.

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261.

Tables.

Table 1. MSPEs at the two unmonitored sites in the simulation study.

	site 9			site 10		
model	source 1	source 2	source 3	source 1	source 2	source 3
SIM	3.254	0.851	2.236	3.144	0.863	2.239
SDM	1.036	0.434	0.565	0.798	0.316	0.443

Table 2. Estimated source composition profile matrix (P_{ij}'s) along with the mean (ξ_i's), the standard deviation (σ_i's) estimates for the source contribution and the standard deviation (η_i's) estimates for the error terms. The asymptotic standard errors of the estimates are given in parentheses. Each entry of the source composition profile matrix displays the percentage and zeros in bold give the locations of pre-assigned zeros.

row	name	Refinery	Petrochem	Gasoline	Natural Gas	η_i
1	Benzene	2.431(0.148)	**0**	4.335 (0.173)	**0**	0.891 (0.131)
2	Ethane	44.605 (0.821)	0.000(-)	**0**	44.105 (0.703)	3.220 (0.506)
3	Ethylene	**0**	33.420(1.115)	**0**	**0**	2.404 (0.545)
4	Isobutane	5.057 (0.606)	13.154 (1.150)	16.433 (0.650)	7.572 (0.414)	3.702 (0.551)
5	Isopentane	7.077 (0.463)	**0**	35.750 (0.403)	1.116 (0.384)	0.007 (-)
6	Propane	26.401 (0.569)	8.370 (1.152)	1.002 (0.494)	36.096 (0.461)	0.0003 (-)
7	Propylene	**0**	36.578 (0.950)	**0**	**0**	7.358 (1.115)
8	n-Butane	11.734 (0.152)	4.949 (0.462)	26.436 (0.234)	9.090 (0.150)	2.676(0.393)
9	n-Hexane	**0**	3.630(1.384)	1.338 (0.108)	0.510 (0.063)	0.496 (0.078)
10	n-Pentane	2.694 (0.902)	**0**	14.704(0.703)	1.511 (0.810)	0.630 (0.093)
	ξ_i	22.755(1.533)	10.043(0.399)	2.588(0.372)	1.215(1.521)	
	σ_i	12.240 (0.882)	8.044 (0.309)	8.394 (0.148)	14.147 (0.480)	

Table 3. Major compounds for the four main pollution sources considered in the analysis.

	Refinery	Petrochemical Production	Gasoline Evaporation	Natural Gas
Major compounds	Propane	Ethylene	n-Butane	Ethane
	Ethane	Propylene	Isopentane	Propane
	n-Butane		Isobutane	
	Isobutane		n-Pentane	

Figures.

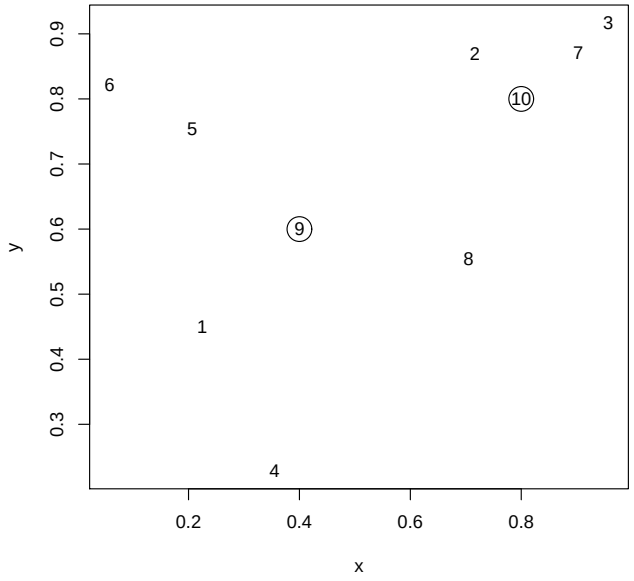

Figure 1. Location of 10 sites in the simulation study. Numbered circles give the locations of 2 prediction sites.

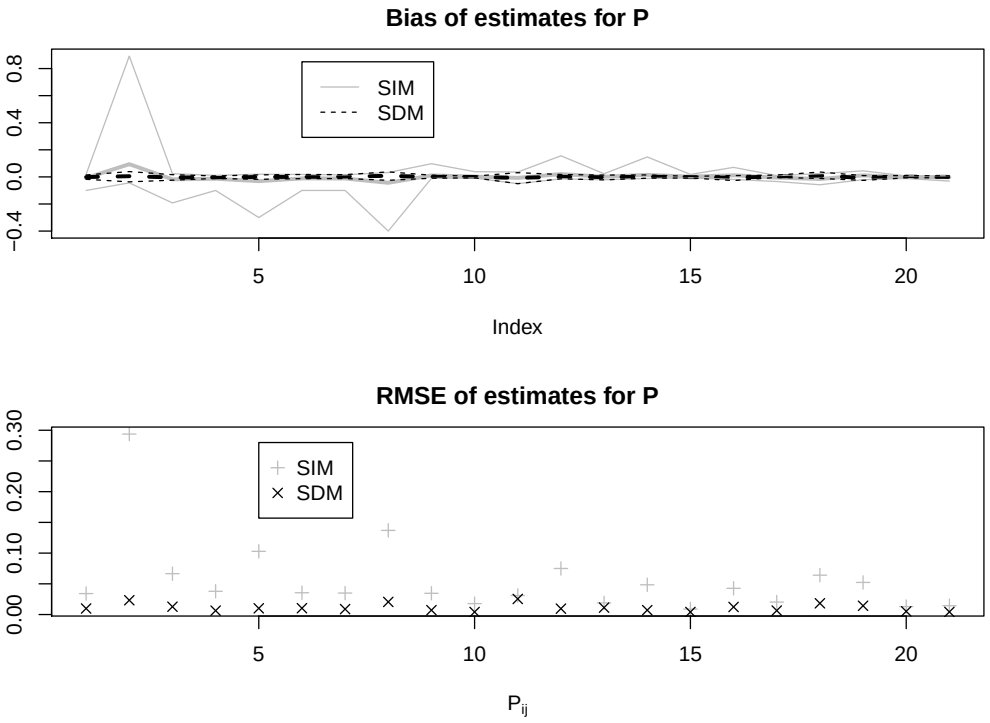

Figure 2. Biases and root-mean-squared-errors of the estimates of non-zero elements of the source composition matrix based on 100 simulations. We display non-zero elements of the matrix column-wise (there are a total of 21 non-zero elements). Top panel: thick lines give the mean and thin lines give 5th and 95th percentiles based on 100 simulations.

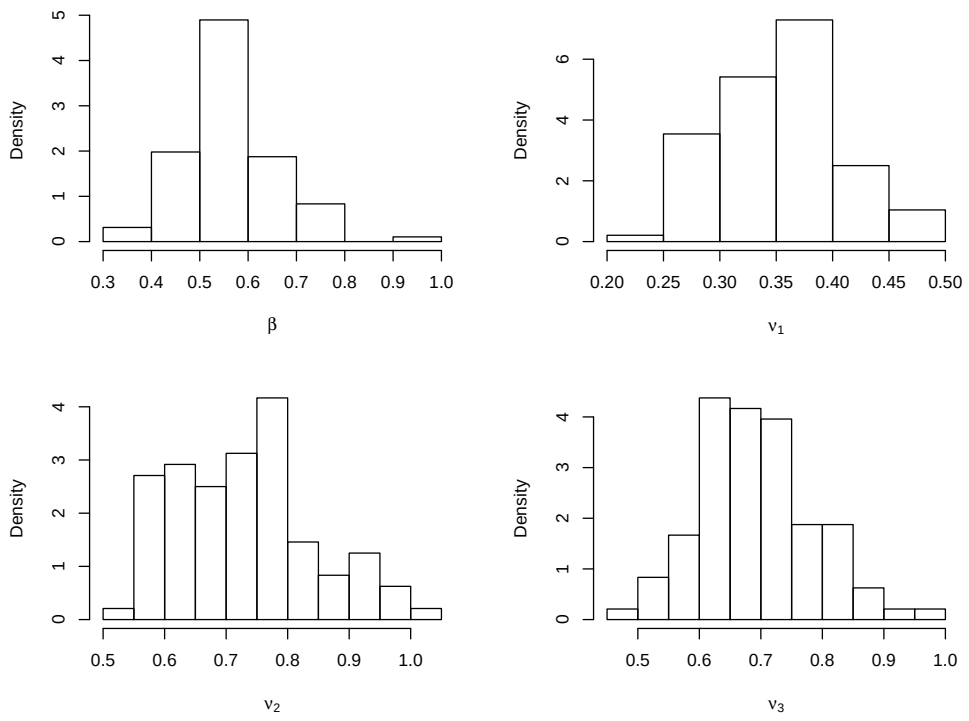

Figure 3. The estimates of the covariance parameters of the SDM model based on 100 simulations.

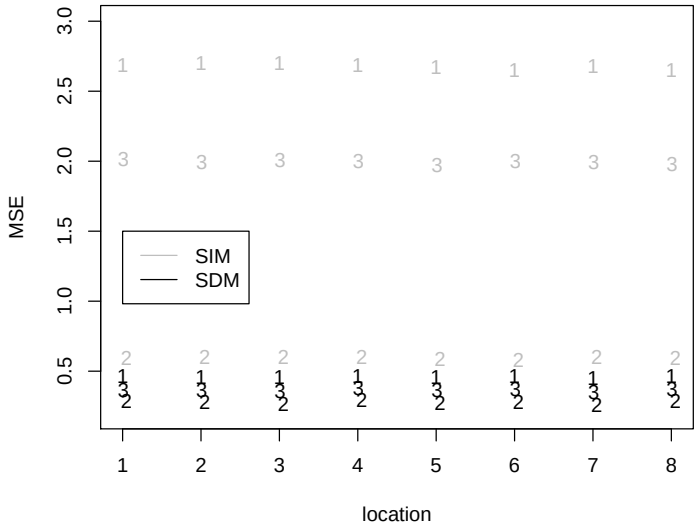

Figure 4. MSEs of the estimated source contributions at the 8 simulation locations given in Figure 1. The numbers in the figure represent the source number.

Figure 5. Map of 9 locations in Houston area. Observations over numbered 8 locations are used for model fitting and prediction of source contribution is made at an unmonitored location (cross).

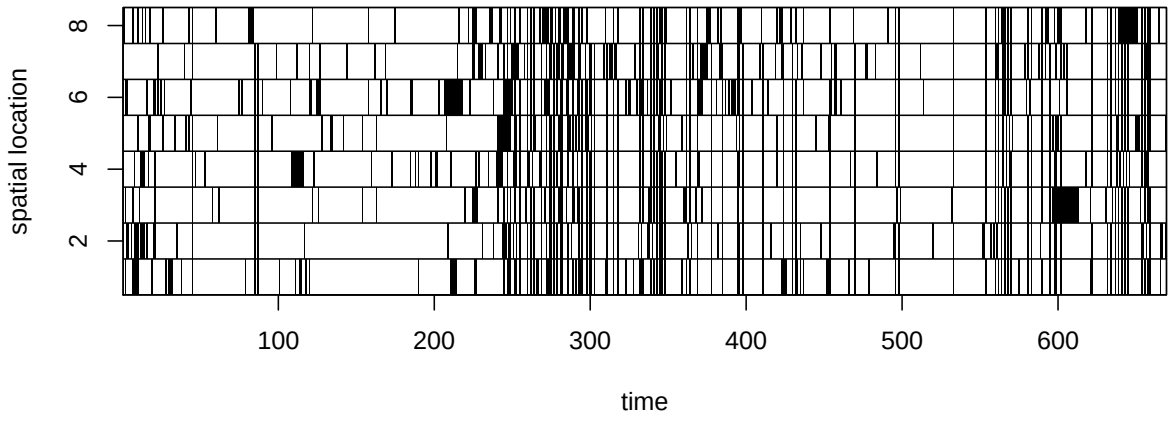

Figure 6. Locations of missing data in space and time are marked in black.

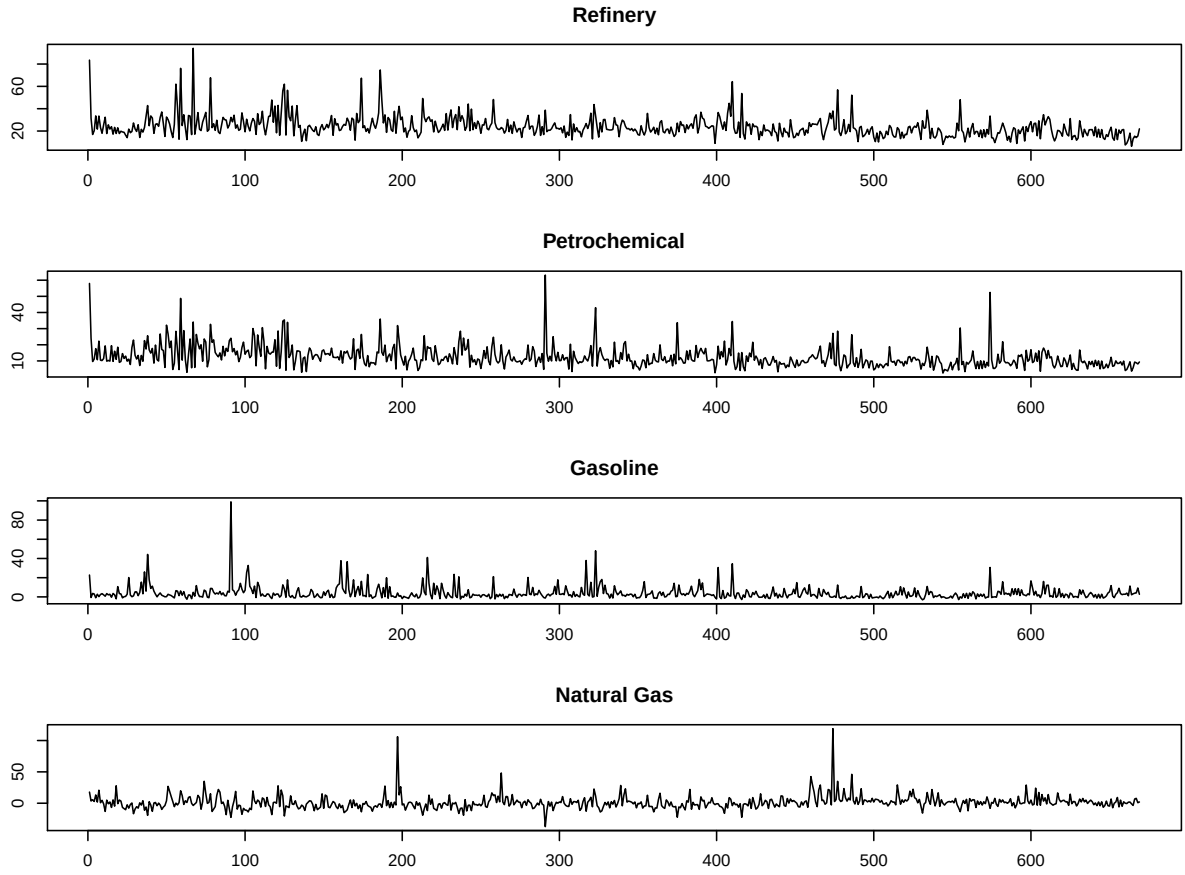

Figure 7. Estimated source contribution at the HRM-3 site (site 2 in Figure 5) over time.

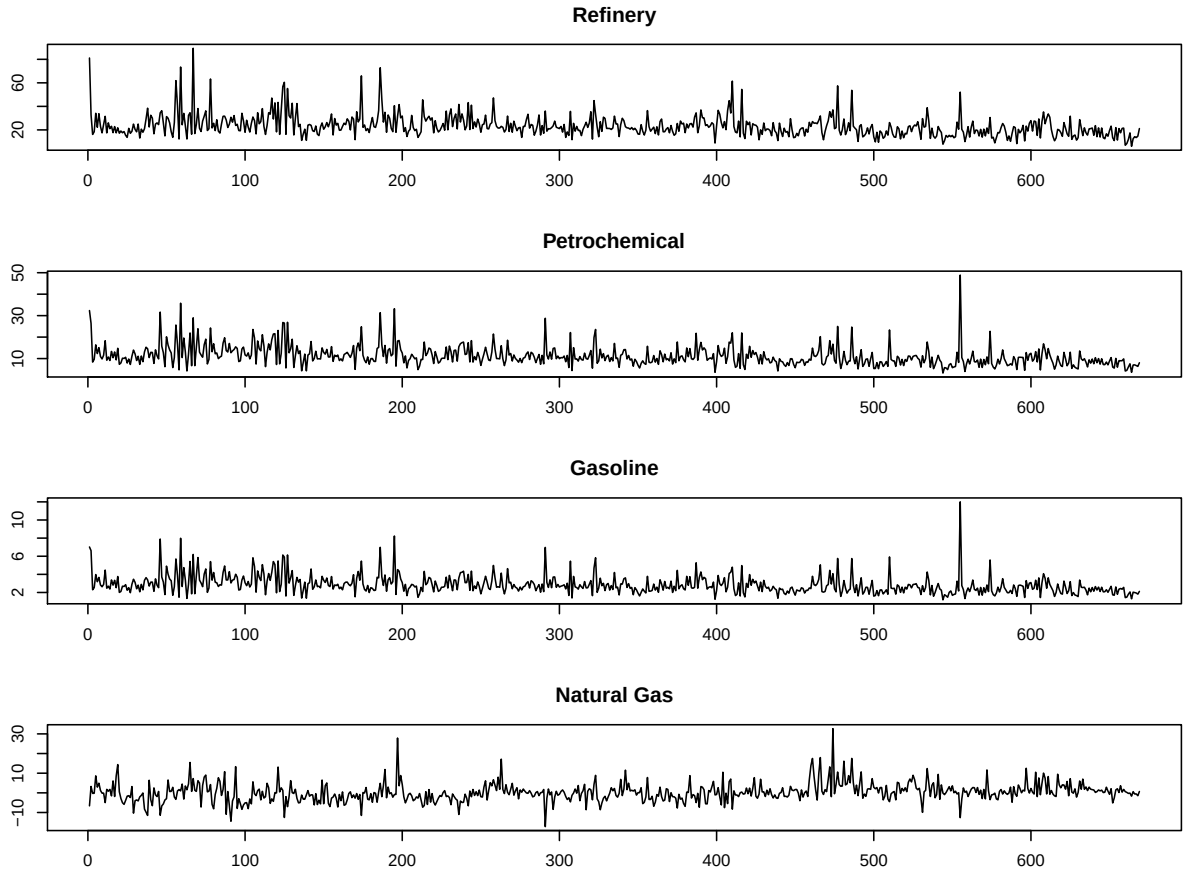

Figure 8. Predicted source contribution at the unmonitored site (cross in Figure 5) over time.