A Bayesian Multivariate Receptor Model for Estimating Source Contributions to Particulate Matter Pollution using National Databases

Amber J. Hackstadt and Roger D. Peng
Johns Hopkins University, Baltimore, USA

Abstract

Time series studies have suggested that air pollution can negatively impact health. These studies have typically focused on the total mass of fine particulate matter air pollution or the individual chemical constituents that contribute to it, and not source-specific contributions to air pollution. Source-specific contribution estimates are useful from a regulatory standpoint by allowing regulators to focus limited resources on reducing emissions from sources that are major contributors to air pollution and are also desired when estimating source-specific health effects. However, researchers often lack direct observations of the emissions at the source level. We propose a Bayesian multivariate receptor model to infer information about source contributions from ambient air pollution measurements. The proposed model incorporates information from national databases containing data on both the composition of source emissions and the amount of emissions from known sources of air pollution. The proposed model is used to perform source apportionment analyses for two distinct locations in the United States (Boston, Massachusetts and Phoenix, Arizona). Our results mirror previous source apportionment analyses that did not utilize the information from national databases and provide additional information about uncertainty that is relevant to the estimation of health effects.

1 Introduction

1.1 Background

Previous work has established an association between exposure to fine particulate matter ($PM_{2.5}$) air pollution and the risk for mortality and morbidity (Schwartz et al., 1996; Dominici et al., 2006; Pope and Dockery, 2006; Bell et al., 2009; Peng et al., 2009; Zanobetti et al., 2009; Brook et al., 2010; Rohr and Wyzga, 2012). Researchers have found an increase in the risk of mortality and morbidity with an increase in the total mass of particles as well as an increase in the risk of morbidity with an increase in the contributions of individual chemical constituents of particulate matter (PM) (Schwartz et al., 1996; Dominici et al., 2006; Bell et al., 2009; Peng et al., 2009; Zanobetti et al., 2009; Rohr and Wyzga, 2012). Thus, recent research suggests that sources of $PM_{2.5}$ will have varying effects on health due to differences in the amount each chemical constituent contributes to their emissions.

Estimates of source-specific contributions to $PM_{2.5}$ can be used to estimate source-specific health effects and identify sources with the most harmful contributions. They allow regulators to focus limited resources on reducing emissions of sources that contribute the most to morbidity and mortality. However, we often only have measurements of ambient $PM_{2.5}$ concentrations from national monitoring networks and not individual source emissions. Multivariate receptor models, or

source apportionment models, can be used to estimate source-specific contributions from ambient air pollution measurements.

1.2 Source Apportionment Model

Let $\boldsymbol{Y}_t$ be a $P \times 1$ column vector of PM concentrations (in $\mu g/m^3$) for P chemical constituents observed at time t. Let $\boldsymbol{\Lambda}$ be a $P \times K$ matrix of source profiles where the elements in each column sum to one. The element in the pth row and kth column of $\boldsymbol{\Lambda}$, λ_{pk}, is the contribution of chemical constituent p for source k. Let $\boldsymbol{f}_t$ be a $K \times 1$ column vector of source contributions from the K source categories where each source contribution, f_{kt}, is the amount of ambient PM (in $\mu g/m^3$) that is attributed to source category k for time point t. Assume that $\boldsymbol{Y}_t$ is observed and that $\boldsymbol{\Lambda}$ and $\boldsymbol{f}_t$ are unknown. Assume $\boldsymbol{\Lambda}$ does not depend on time.

There are two common types of source apportionment models. One could specify a model with additive errors, as in Christensen et al. (2006), Lingwall and Christensen (2007), Nikolov et al. (2007), and Heaton et al. (2010). The additive errors model is

$$\boldsymbol{Y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\epsilon}_t \tag{1}$$

where $\boldsymbol{\epsilon}_t$ is a $P \times 1$ vector of errors.

As in Nikolov et al. (2008) and Wolbers and Stahel (2005), one could instead specify a source apportionment model with multiplicative errors and assume

$$\boldsymbol{Y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t \circ \boldsymbol{\epsilon}_t \tag{2}$$

where $\circ$ denotes element-wise multiplication and $\boldsymbol{\epsilon}_t$ is a $P \times 1$ vector of errors.

1.3 Previous Work

Previous studies, such as studies using principal component analysis (PCA), have performed source apportionment analyses using an eigenvector analysis based on singular value decomposition (Thurston and Spengler, 1985; Koutrakis and Spengler, 1987; Gao et al., 1994). These approaches assume the additive errors model in (1) and that the number and/or sources are unknown. Other studies have used an approach called positive matrix factorization (PMF) which, unlike the eigenvector analysis, restricts the source contributions, f_{kt} to be positive (Paatero and Tapper, 1994; Paatero, 1997). PMF provides unique solutions under certain assumptions about the sources, and has been used in several studies (Song et al., 2001; Ramadan et al., 2003; Zhou et al., 2004; Hopke et al., 2006; Lingwall and Christensen, 2007; Kim and Hopke, 2008; Liming et al., 2009). Another approach used in several source apportionment analysis studies is UNMIX. UNMIX estimates the sources and source profiles by a data driven procedure that looks for hyperplanes in vector spaces to identify sources (Henry, 1997; Lewis et al., 2003; Henry, 2005; Hopke et al., 2006).

There also have been several source apportionment analyses that use a Bayesian approach (Billheimer, 2001; Park et al., 2001, 2002b; Nikolov et al., 2007, 2008; Lingwall et al., 2008; Heaton et al., 2010; Nikolov et al., 2011). Billheimer (2001) uses a Bayesian source apportionment model with multiplicative errors and models both the source contributions and the source profiles as unknown compositional quantities for K known sources. Nikolov et al. (2007) and Nikolov et al. (2008) propose source apportionment models that are part of Bayesian structural equations models (SEMs) to determine the source-specific health effects of air pollution. Lingwall et al. (2008), Park et al. (2001), and Park et al. (2002b) use a Bayesian approach assuming a model with additive errors as in (1). Heaton et al. (2010) consider a Bayesian source apportionment model with additive errors where the source profiles vary with time.

In this paper, we propose a Bayesian source apportionment model using the multiplicative errors model in (2) that incorporates information from three EPA databases. We use this model to estimate source contributions to ambient $PM_{2.5}$ in the Boston and Phoenix areas. Non-Bayesian approaches, such as PCA and PMF, do not lend themselves to the incorporation of *a priori* information as easily as Bayesian approaches. They also do not provide posterior distributions for the parameters, which can be used to obtain uncertainty estimates for parameters or functions of parameters. None of the previously mentioned Bayesian approaches have incorporated information from national databases that provide local information about source emissions as well as information about the chemical composition of the source emissions. The use of publicly available national databases prevents one from having to rely solely on previous source apportionment analyses in the area. It also allows our model to be extended to a national source apportionment analysis, which can be performed by first computing location-specific source contribution estimates for several locations across the United State, then comparing or combining these estimates.

2 Data

Our approach makes use of three key national databases provided by the United States Environmental Protection Agency (EPA): Chemical Speciation Network (CSN), National Emissions Inventory (NEI), and SPECIATE. The CSN is used to estimate ambient PM concentrations ($\boldsymbol{Y}_t$ in (2)), the SPECIATE database is used to create priors for the source profiles ($\boldsymbol{\Lambda}$ in (2)), and the NEI database is used to create priors for the unknown source contributions ($\boldsymbol{f}_t$ in (2)).

2.1 Chemical Speciation Network (CSN)

To obtain speciated measurements of ambient PM, $\boldsymbol{Y}_t$ in (2), we use the United States Environmental Protection Agency's CSN (United States Environmental Protection Agency, 2012g,f; Bell et al., 2007). From the CSN, we obtain daily concentrations of $PM_{2.5}$ in micrograms per cubic meters ($\mu g/m^3$) for 57 chemical constituents from monitors in the EPA's Speciation Trends Network (STN) (United States Environmental Protection Agency, 2012d; Bell et al., 2007). These monitors are placed at various locations throughout the United States (United States Environmental Protection Agency, 2012d; Monitoring and Quality Assurance Group, 1999; Bell et al., 2007). The frequency of the air pollution measurements and the years that the measurements span differ for each monitor. For our source apportionment analysis, we will focus on the monitor in Suffolk County, Massachusetts (MA) to get ambient $PM_{2.5}$ measurements for the city of Boston and a monitor in Maricopa County, Arizona (AZ) to get ambient $PM_{2.5}$ measurements for the city of Phoenix. The monitor in Suffolk county has measurements spanning the years from 2000 to 2009 with measurements approximately every 3 days starting in April 4, 2001. The monitor in Maricopa County has measurements spanning the years from 2000 to 2009 with measurements approximately every 3-6 days with the measurements becoming more frequent for the later dates. We select to focus on Boston and Phoenix because there exists previous source apportionment analyses at these locations to which we can compare our results and these are two distinct locations in the United States (US).

2.2 SPECIATE Database

We use the information in the SPECIATE database to create the priors for $\boldsymbol{\Lambda}$ in (2). The SPECIATE database is a data repository created by the EPA that provides speciated emission profiles for known sources of air pollutants, such as sources of PM, volatile organic compounds (VOC), and

other gases (United States Environmental Protection Agency, 2012j,i). The emission profiles are obtained from studies conducted in the 1970s and 1980s as well as recent air quality management studies and research studies (Hsu and Divita, 2011). The version of SPECIATE used in our analysis, version 4.2, contains a total of 3326 PM profiles, including profiles for "Coal-Fired Power Plant" and "Light Duty Vehicles - Unleaded" (United States Environmental Protection Agency, 2012i; Hsu and Divita, 2011). Using the descriptions of emission profiles in the SPECIATE database, the profiles can be grouped into various source categories. Figure 1 shows the median percent contribution for each chemical constituent for some of the source categories used in our analysis for Boston and Phoenix. For each chemical constituent in a given source category, the median percent contribution is found by taking the median of the percent that chemical constituent contributes to the source emissions across all profiles in that source category. Note that the chemical constituents have the following abbreviations: aluminum (Al), calcium (Ca), chlorine (Cl), elemental carbon (EC), iron (Fe), lead (Pb), nickel (Ni), organic carbon (OC), potassium (K), silicon (Si), sulfur (S), titanium (Ti), vanadium (V), and zinc (Zn). We scale the median percent contributions for a source category such that they sum to one. Not all profiles in SPECIATE have location information so we do not use any location information from this database. Regardless of the location information provided, we take the median of all profiles that fall into a given source category. However, since we focus on different chemical constituents at each location, the profile for a source category may be differ slightly between locations.

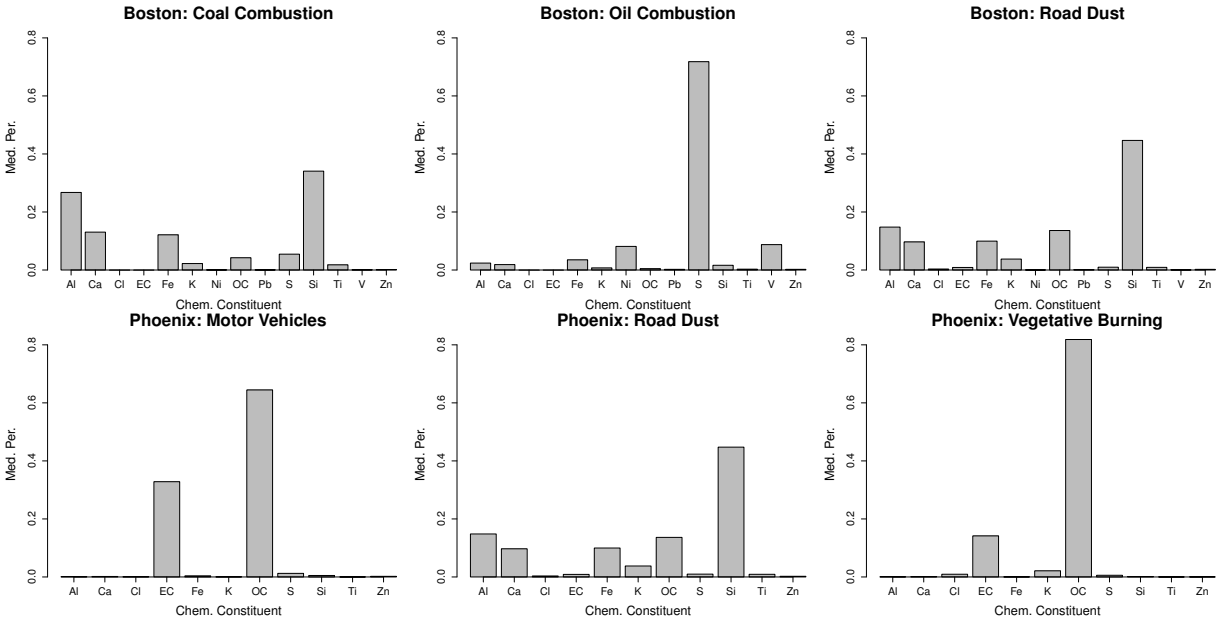

Figure 1: Plots providing the median percent contribution to source emissions for each chemical constituent, focusing on the source categories and chemical constituents used in our source apportionment analyses. The median percent contributions are obtained from the SPECIATE database.

We note here that the median percent estimates from SPECIATE do differ from the estimates for $\mathbf{\Lambda}$ from other source apportionment studies (Thurston and Spengler, 1985; Nikolov et al., 2007; Ramadan et al., 2000, 2003). For instance, the SPECIATE estimates suggest that Al and Si contribute relatively large proportions to the Coal Combustion category emissions. However, Thurston and Spengler (1985) estimate that Al and Si only contribute small proportions to the emissions for

Sulfate aerosols, which is similar to our Coal Combustion category. Likewise, Nikolov et al. (2007) assume that Si does not contribute to emissions from the Power Plant source category, which is similar to our Coal Combustion source category. They also estimate that Al contributes a small proportion to Power Plant emissions. These differences may be due to the fact that the $\boldsymbol{\Lambda}$ estimates from source apportionment analyses do not use the same information as in the SPECIATE database. Also, these analyses use different source categories and chemical constituents than in our Boston analysis.

We further note that the chemical constituents with the largest median percent contributions are similar for Coal Combustion and Road Dust source categories for Boston (Figure 1). This suggests that we may have some difficulty distinguishing the contributions from these sources. However, there are some noticeable differences between the median percent estimates for these two sources. The percent contribution estimates for EC and OC for Road Dust are higher than they are for Coal Combustion. Also, the percent contribution estimates for S and Al are higher for Coal Combustion than they are for Road Dust.

2.3 NEI Database

We use information in the NEI to create priors for $\boldsymbol{f}_t$ in (2). The NEI is a national database created and maintained by the EPA (United States Environmental Protection Agency, 2012h). It provides estimates for the primary emissions (in tons per year) of both criteria and hazardous pollutants at the county level. The NEI is created using estimates from state, local, and tribal air agencies and also from data developed by the EPA (United States Environmental Protection Agency, 2012h). Primary emissions are air pollution emissions that are emitted directly into the atmosphere from both human and natural sources, such as road dust and wood burning. Secondary $PM_{2.5}$, which also contributes to air pollution, is created in the atmosphere by chemical reactions with substances emitted into the atmosphere by various sources including power plants. Both primary emissions and secondary $PM_{2.5}$ contribute to ambient $PM_{2.5}$ levels (United States Environmental Protection Agency, 2012e). Note that the NEI database only provides estimates from primary emissions and does not offer information about secondary $PM_{2.5}$.

At the time of our analysis, there were inventories available for 2002, 2005, and 2008 (United States Environmental Protection Agency, 2012h). We selected to focus on the 2002 NEI because it was the most recent, reliable NEI at the time of our analysis (United States Environmental Protection Agency, 2012a; Emission Factor and Inventory Group, 2004; Emission Inventory and Analysis Group, 2006; E.H. Pechan & Associates, Inc., 2006; Assessment and Standards Division and E.H. Pechan & Associates, Inc., 2007). The 2005 NEI had some emission estimates copied from the 2002 NEI (United States Environmental Protection Agency, 2012b) and the 2008 NEI was still in the process of being updated at the time of the analysis (United States Environmental Protection Agency, 2012c). Figure 2 shows the annual primary emissions estimates for 4 source categories in Boston (Coal Combustion, Motor Vehicles, Oil Combustion, and Road Dust) and 3 source categories in Phoenix (Motor Vehicles, Road Dust, and Vegetative Burning). These estimates were found using the annual emission estimates from the National Emissions Inventory (NEI) for their corresponding counties, Suffolk and Maricopa, respectively.

3 Source Apportionment Model

Consider the source apportionment model with multiplicative errors given in (2) where we assume $\boldsymbol{Y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t \circ \boldsymbol{\epsilon}_t$. Let $\log(\boldsymbol{x})$, for any vector $\boldsymbol{x}$, denote a vector whose elements are the log of the corresponding elements of $\boldsymbol{x}$ and let $\log(\boldsymbol{A})$, for any matrix $\boldsymbol{A}$, denote a matrix whose elements are

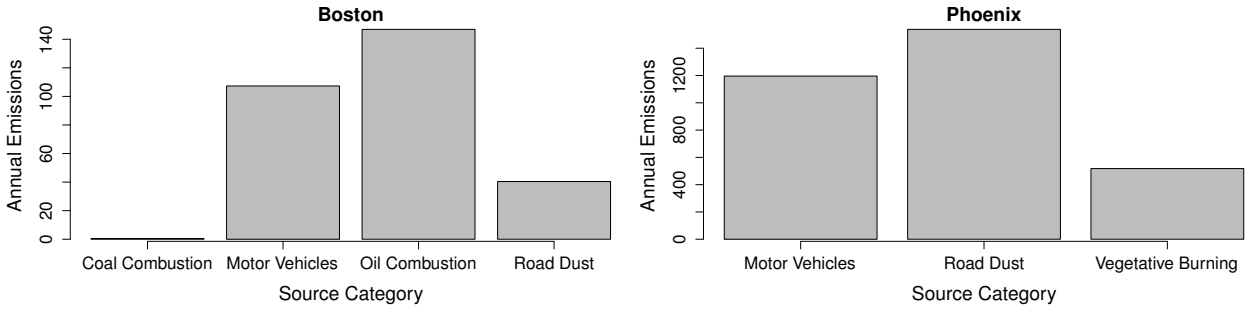

Figure 2: Plot of the annual primary emission estimates (in tons per year) for the source categories used in our analysis for Boston and Phoenix with emission estimates found using the annual emission estimates from the National Emissions Inventory (NEI) for their corresponding counties, Suffolk and Maricopa, respectively.

the log of the corresponding elements of $\boldsymbol{A}$. For the source apportionment model with multiplicative errors, $\log(\boldsymbol{Y}_t) = \log(\boldsymbol{\Lambda}\boldsymbol{f}_t) + \log(\boldsymbol{\epsilon}_t)$. Assume independent lognormal errors with $\log(\boldsymbol{\epsilon}_t)$ a having multivariate normal distribution with a mean equal to the $P \times 1$ zero vector and variance equal to diagonal matrix $\boldsymbol{\Psi}$. Denote this by $\mathrm{N}_P(\boldsymbol{0}, \boldsymbol{\Psi})$. Let $\left(\psi_1^2, \ldots, \psi_p^2\right)$ be the diagonal elements of $\boldsymbol{\Psi}$, $\boldsymbol{Y}$ be a $P \times T$ matrix with columns $\boldsymbol{Y}_t$ for $t = 1, \ldots, T$, and $\boldsymbol{F}$ be a $K \times T$ matrix with columns $\boldsymbol{f}_t$ for $t = 1, \ldots, T$. The likelihood for $\log(\boldsymbol{Y})$ given all other parameters, $f(\log(\boldsymbol{Y}); \boldsymbol{\Lambda}, \boldsymbol{F}, \boldsymbol{\Psi})$, is given by

$$\prod_{t=1}^{T}\prod_{p=1}^{P}\left[\left(2\pi\psi_p^2\right)^{-1/2}\exp\left\{\frac{-1}{2\psi_p^2}\left(\log(y_{pt}) - \log\left(\sum_{k=1}^{K}\lambda_{pk}f_{kt}\right)\right)^2\right\}\right]. \tag{3}$$

3.1 Informative Priors

3.1.1 Priors for Source Contributions

We use the NEI to obtain informative lognormal priors for the source contributions and assume $\log(f_{kt})$ is normally distributed with mean γ_k and standard deviation δ_k. We estimate the proportion of primary emissions for each source category, p_k, by $\hat{p}_k = q_k / \sum_{k=1}^{K} q_k$. For each county, q_k is the sum of all the NEI emission estimates that have corresponding EPA Source Classification Codes (SCC) indicating the estimates are from source category k. We use the CSN (Section 2.1) to estimate the daily total $\mathrm{PM}_{2.5}$ for time point t, denoted x_t. For day t, we find x_t by summing the measured ambient $\mathrm{PM}_{2.5}$ concentrations across the chemical constituents used in the analysis. We let $\tilde{f}_{kt} = \log(\hat{p}_k x_t)$ for $t = 1, \ldots, T$ and $k = 1, \ldots, K$. We find the sample mean and sample standard deviation of $\tilde{f}_{kt}$ for each source category k and denote them by $\bar{x}_{f_k}$ and s_{f_k}, respectively. For our lognormal priors for f_{kt}, we let $\gamma_k = \bar{x}_{f_k}$ and the standard deviation $\delta_k = 3s_{f_k}$ to inflate the variance. We use the inflated variance to allow some flexibility in our model for the source contribution estimates since the NEI only provides information on primary source emissions and our proportion estimates do not take into account secondary $\mathrm{PM}_{2.5}$.

3.1.2 Priors for Source Profiles

To improve estimation, we define variable λ_{pk}^* for $p = 1, \ldots, P$ and $k = 1, \ldots, K$ such that $\lambda_{pk} = \lambda_{pk}^* / \sum_{j=1}^{P} \lambda_{jk}^*$ and let $\boldsymbol{\Lambda}^*$ be the $P \times K$ matrix with elements λ_{pk}^*. Unlike the λ_{pk} values, which

only take on values between zero and one, the λ^*_{pk} range from zero to infinity. Note that the source apportionment model in (2) can be rewritten in terms of $\mathbf{\Lambda}^*$ by

$$\boldsymbol{Y}_t = \mathbf{\Lambda}\boldsymbol{f}_t \circ \boldsymbol{\epsilon}_t = \mathbf{\Lambda}^*\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{f}^*_t \circ \boldsymbol{\epsilon}_t = \mathbf{\Lambda}^*\boldsymbol{f}^*_t \circ \boldsymbol{\epsilon}_t \tag{4}$$

where $\boldsymbol{B}$ is a $K \times K$ diagonal matrix with the kth diagonal element $1/\sum_{j=1}^{P}\lambda^*_{jk}$ and $\boldsymbol{f}^*_t$ is such that $B^{-1}\boldsymbol{f}^*_t = \boldsymbol{f}_t$.

As indicated in (4), the source apportionment models with additive and multiplicative errors are not identifiable in a non-Bayesian setting. Yet, they become identifiable under certain constraints on the source profile matrix (Park et al., 2002a,b). We constrain the parameter space in our Bayesian approach to improve our estimates of the parameters using the constraints on $\mathbf{\Lambda}^*$ given in Park et al. (2002a) and Park et al. (2002b). The constraints are summarized as follows:

(C1) There are at least $K-1$ zeros in each column of $\mathbf{\Lambda}^*$.

(C2) Let $\mathbf{\Lambda}^{(k)}$ be a matrix created by selecting rows of $\mathbf{\Lambda}^*$ that have zero in the kth column and then deleting the kth column. The rank of $\mathbf{\Lambda}^{(k)}$ is $K-1$ for all k.

(C3) $\lambda^*_{pk} = 1$ for some $p \in \{1, \ldots, P\}$ for each $k = 1, \ldots, K$.

We utilize the information about the source profiles in the SPECIATE database when imposing constraints. For each source, we choose the $K-1$ λ^*_{pk}'s to fix at zero that correspond with the constituents that have the smallest median percent contributions (Figure 1) but also allow us to meet condition (C2). The (C3) constraint is required only to scale the values in $\mathbf{\Lambda}^*$ and does not help separate the sources (Park et al., 2002a). Thus, letting $\lambda^*_{pk} = 1$ for the same chemical constituent for several sources will not hinder the ability of the model to estimate contributions from these sources. However, setting $\lambda^*_{pk} = 1$ for a constituent that contributes a relatively large amount to $PM_{2.5}$ improves estimation (See Section 5 for more details). Therefore, we choose $\lambda^*_{pk} = 1$ such that the p corresponds to a chemical constituent that has a relatively large median percent contribution in the SPECIATE database and that contributes, on average, a relative large amount to $PM_{2.5}$ as indicated by CSN. The process used for selecting the constraints for the analysis of Boston and Phoenix is described in further detail in Section 1.1 in the Supplementary Material.

We call the values of λ^*_{pk} in $\mathbf{\Lambda}^*$ that are not set to either 0 or 1, the free values of $\mathbf{\Lambda}^*$ or free λ^*_{pk}'s. Note that the free λ^*_{pk}'s for source k are in terms of the contribution from the constituent for which the λ^*_{pk} is set to 1. For these free λ^*_{pk}'s, we assume prior independence and a truncated normal prior that is truncated on $(0, \infty)$ with mean μ_{pk} and variance σ^2_{pk} prior to truncation, denoted $\text{TruncN}\left(\mu_{pk}, \sigma^2_{pk}\right)$.

We use the SPECIATE database to estimate the hyperparameters for the truncated normal priors. First, note that the mean and variance of the truncated normal distribution after truncation, denoted by $\tilde{\mu}_{pk}$ and $\tilde{\sigma}^2_{pk}$, respectively, can be written in terms of the mean and variance prior to truncation, μ_{pk} and σ^2_{pk}, respectively. Hence

$$\tilde{\mu}_{pk} = \mu_{pk} + \sigma_{pk}\left\{\frac{\phi\left(-\mu_{pk}/\sigma_{pk}\right)}{1-\Phi\left(-\mu_{pk}/\sigma_{pk}\right)}\right\} \tag{5}$$

and

$$\tilde{\sigma}^2_{pk} = \sigma^2_{pk}\left\{1 - \left(\frac{\phi\left(-\mu_{pk}/\sigma_{pk}\right)}{1-\Phi\left(-\mu_{pk}/\sigma_{pk}\right)}\right)\left(\frac{\phi\left(-\mu_{pk}/\sigma_{pk}\right)}{1-\Phi\left(-\mu_{pk}/\sigma_{pk}\right)} + \frac{\mu_{pk}}{\sigma_{pk}}\right)\right\} \tag{6}$$

where $\phi\left(\right)$ and $\Phi\left(\right)$ denote the probability density function and cumulative distribution function for the standard normal distribution, respectively. After rescaling the SPECIATE profiles to be

in terms of the chemical constituents that are fixed at 1, we estimate $\tilde{\mu}_{pk}$ and $\tilde{\sigma}^2_{pk}$ by taking the sample mean and sample variance, respectively, of the profile values for constituent p. We denote these estimates by $\hat{\mu}_{pk}$ and $\hat{\sigma}^2_{pk}$ and find the hyperparameters μ_{pk} and σ^2_{pk} by minimizing

$$\left[\hat{\mu}_{pk} - \left\{\mu_{pk} + \sigma_{pk}\left(\frac{\phi\left(-\mu_{pk}/\sigma_{pk}\right)}{1-\Phi\left(-\mu_{pk}/\sigma_{pk}\right)}\right)\right\}\right]^2 + \left[\hat{\sigma}^2_{pk} - \sigma^2_{pk}\left\{1 - \left(\frac{\phi\left(-\mu_{pk}/\sigma_{pk}\right)}{1-\Phi\left(-\mu_{pk}/\sigma_{pk}\right)}\right)\left(\frac{\phi\left(-\mu_{pk}/\sigma_{pk}\right)}{1-\Phi\left(-\mu_{pk}/\sigma_{pk}\right)} + \frac{\mu_{pk}}{\sigma_{pk}}\right)\right\}\right]^2 \tag{7}$$

using an nonlinear constrained optimization algorithm.

3.2 Bayesian Model

For the Bayesian source apportionment model, assume the model in (2) and likelihood given in (3). Assume the priors for the source contributions, f_{kt}, given in Section 3.1.1 and the priors for the source profiles given in Section 3.1.2. For each ψ^2_p in the diagonal of $\mathbf{\Psi}$, assume an inverse-gamma prior with shape and scale α_p and β_p, respectively. We choose α_p and β_p such that the mean of the inverse gamma prior is equal to the sample variance of the observed $\log(y_{pt})$ values for constituent p and the standard deviation of the inverse gamma prior is twice the mean. The Bayesian model is summarized as follows

$$\begin{aligned}
\log(\boldsymbol{Y}_t)\,|\boldsymbol{\Lambda}, \boldsymbol{f}_t, \mathbf{\Psi} &\sim \mathrm{N}_P\left(\log\left(\boldsymbol{\Lambda}\boldsymbol{f}_t\right), \mathbf{\Psi}\right) \\
\lambda_{pk} &= \frac{\lambda^*_{pk}}{\sum_{j=1}^{P}\lambda^*_{jk}} \\
\lambda^*_{pk} &\sim \mathrm{TruncN}\left(\mu_{pk}, \sigma^2_{pk}\right) \\
\log(f_{kt}) &\sim \mathrm{N}\left(\gamma_k, \delta^2_k\right) \\
\psi_p &\sim \mathrm{InvGamma}\left(\alpha_p, \beta_p\right)
\end{aligned} \tag{8}$$

where λ^*_{pk} is as defined in Section 3.1.2, $\mathrm{InvGamma}(\alpha_p, \beta_p)$ denotes an inverse gamma distribution with probability density function of the form $f(\psi_p) = \beta_p^{\alpha_p}/\Gamma(\alpha_p)\,\psi_p^{-(\alpha_p+1)}\exp(-\beta_p/\psi_p)\,\mathrm{I}\{0 < \psi_p < \infty\}$ and $\Gamma()$ denotes the gamma function. $\mathrm{I}\{\}$ denotes the indicator function. Under the proposed model, the posterior distribution is proper but intractable so to obtain estimates of all the parameters of interest, Markov chain Monte Carlo (MCMC) techniques are used. Details of the estimation procedure can be found in Section 3 of the Supplemental Material.

4 Results

For our source apportionment analysis of Phoenix and Boston, unless otherwise indicated, the MCMC algorithms are run for 100,000 iterations, discarding the first 10,000 as burn-in. The source categories, chemical constituents, and constraints used in our analyses of Phoenix and Boston are summarized in Table 1. We choose to focus on these source categories and chemical constituents based on the information in the three EPA databases that suggest that these sources and chemical constituents are major contributors to $PM_{2.5}$ as well as previous research in the Boston area (Thurston and Spengler, 1985; Laden et al., 2000; Nikolov et al., 2007, 2008, 2011) and the Phoenix area (Mar et al., 2000, 2006; Ramadan et al., 2000, 2003; Lewis et al., 2003; Hopke et al., 2006; Thurston et al., 2005; Brown et al., 2007).

Table 1: Table summarizing the source categories and the chemical constituents used in the source apportionment analyses for Boston and Phoenix as well as the constraints on the source profile values. The $\lambda^*_{pk} = 1$ column gives the chemical constituent whose corresponding λ^*_{pk} value is fixed at one. The $\lambda^*_{pk} = 0$ gives the $K - 1$ chemical constituents whose corresponding λ^*_{pk} values are fixed at zero. P is total number of chemical constituents used in the analysis.

Location	Chem. Const.	P	Source Cat.	$\lambda^*_{pk} = 1$	$\lambda^*_{pk} = 0$
Boston	Al, Ca, Cl, EC, Fe,	14	Coal Comb.	S	Cl, Ni, Pb
	Pb, Ni, OC, K, Si,		Motor Veh.	OC	Ni, Ti, V
	S, Ti, V, Zn		Oil Comb.	S	Cl, EC, Pb
			Road Dust	OC	Ni, Pb, V
Phoenix	Al, Ca, Cl, EC,	11	Motor Veh.	OC	K, Ti
	Fe, OC, K, Si,		Road Dust	OC	Cl, Zn
	S, Ti, Zn		Veg. Burning	OC	Ti, Zn

To find the hyperparameters for λ^*_{pk} in our analyses of Phoenix and Boston, we first rescale the SPECIATE profiles to be in terms of the chemical constituents that are fixed at 1 (Table 1). We then obtain the hyperparameters using the procedure described in Section 3.1.2. We minimize (7) using the *optim* function in the R program (R Development Core Team, 2011) and the "L-BFGS-B" method (Byrd et al., 1995) with the lower bound for both parameters equal to $1e - 10$ and upper bound for both parameters equal to 5000. Figure 3 gives histograms of the rescaled SPECIATE source profiles values for EC for Motor Vehicles in the Boston analysis and for Cl for Vegetative Burning in the Phoenix analysis. The solid black line overlaid on top of each histogram is the density for the informative truncated normal prior for the corresponding λ^*_{pk}. To select the hyperparameters for the source contributions, except for the Coal Combustion category for the Boston analysis, we use the procedure described in Section 3.1.1 . For Coal Combustion, we let $\gamma_k = 0$ and $\delta_k = \log(30)$. Previous research indicates that large amounts of secondary $PM_{2.5}$ may be attributed to Coal Combustion and the NEI does not provide estimates for secondary $PM_{2.5}$ (Thurston and Spengler, 1985; Laden et al., 2000; Nikolov et al., 2007, 2008, 2011). These hyperparameters produce a relatively flat prior compared to the prior created using the NEI database, with a variance at least seven times as large as the variance for the prior created using the NEI database. This reflects that there is greater uncertainty in the source contributions from the Coal Combustion category.

4.1 Analyses of Phoenix and Boston

We focus on the estimates for the source contributions, f_{kt}, since these are likely to be used in subsequent time series analysis for source-specific health effects. The source contribution estimate for source k at time t, $\hat{f}_{kt}$, is found by taking the median of the posterior distribution for f_{kt}. We use the posterior median as opposed to the posterior mean because the posterior distributions from some of the sources and time points are skewed.

4.1.1 Phoenix

Figure 4 shows the source contribution estimates for Phoenix across time starting after January 1, 2000 with the dashed vertical lines marking the beginning of each year. It illustrates how the source contributions for Phoenix vary with time. The estimates for all source categories never exceed 20

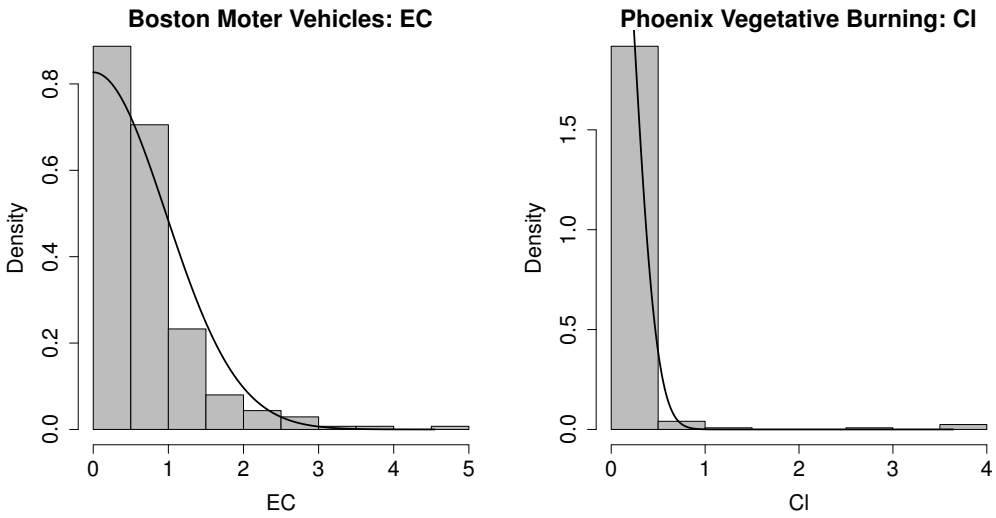

Figure 3: Histogram of the rescaled source profile values from the SPECIATE database for elemental carbon (EC) for the Motor Vehicles source category in the Boston analysis and for chlorine (Cl) for the Vegetative Burning source category in the Phoenix analysis. The solid black lines overlaid on top of the histograms are the densities for the informative truncated normal priors for the corresponding parameters.

μg/m^3. There is an apparent increase in the source contribution estimates for Motor Vehicles during the winter months, which agrees with Ramadan et al. (2000). This suggests that motor vehicle use increases during the colder months. There appears to be a spike in source contribution estimates for Vegetative Burning in the beginning and toward the end of the year for most years in the source apportionment analysis, which also agrees with Ramadan et al. (2000). This may be attributed to an increase in residential wood combustion during the winter months.

4.1.2 Boston

Figure 5 shows the source contribution estimates for Boston across time starting after January 1, 2001 since there were only 10 observations for 2000 for Boston and they were rather sparse. The dashed vertical lines mark the beginning of each year. Note that all the plots in Figure 5 suggest variation across time. The plot of the source contribution estimates for Coal Combustion shows occasional large spikes in estimated daily contributions but the estimates for the majority of the time points are below 20 μg/m^3. Motor Vehicle source contributions appear to increase towards the end of the year for many years in the Boston analysis. Again, this may be due to an increase in motor vehicle use during the colder months. The source contributions for Oil Combustion appear to increase towards the middle of the year. Since Boston is on a harbor and the Oil Combustion source category is likely to include residual fuel oil emissions from ships, this could reflect an increase in ship emissions during the summer months (Ault et al., 2007; Vutukuru and Dabdub, 2008; Mueller et al., 2011). Note that, for both the Boston and Phoenix analyses, the air monitors used are close to roadways. However, the analyses suggest that source categories other than those associated with roadways are contributing to $PM_{2.5}$, such as Oil Combustion in Boston and Vegetative Burning in Phoenix.

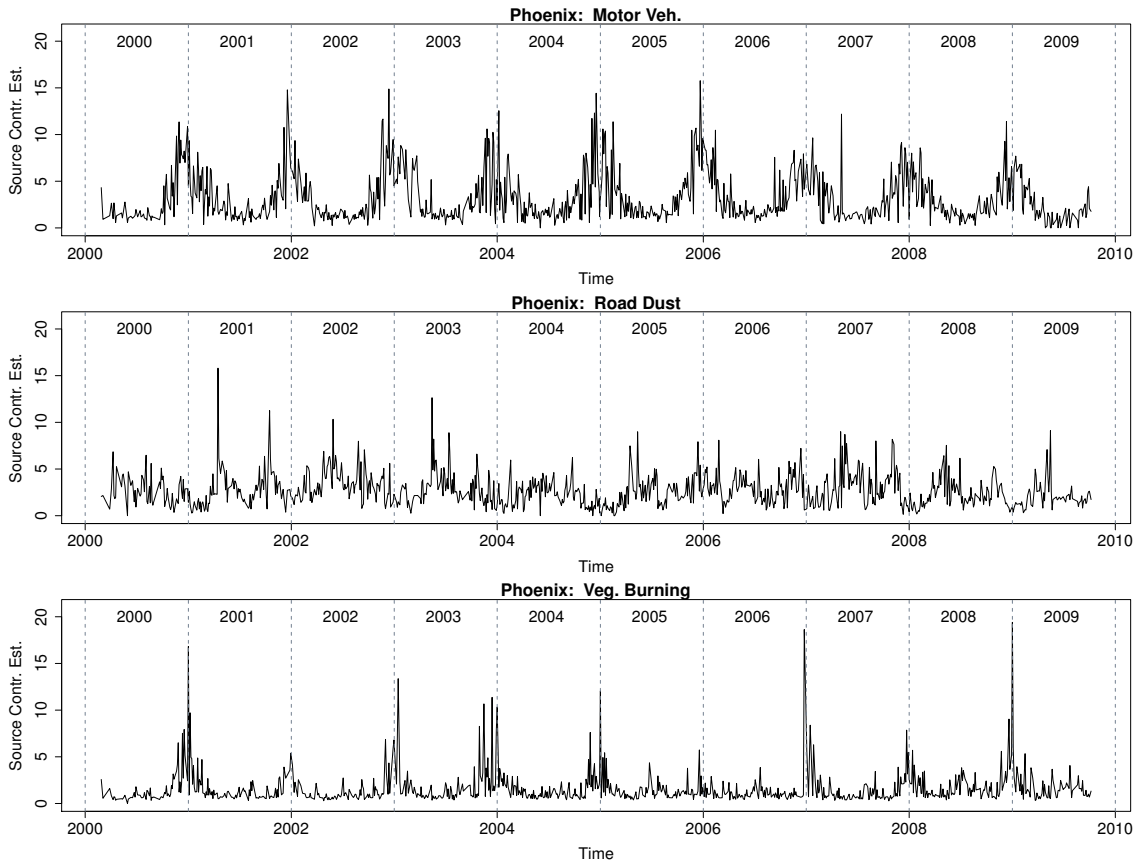

Figure 4: Plots of the source contribution estimates for Phoenix for all time points after January 1, 2000. The vertical dashed lines mark January 1st for the years 2000 through 2010.

4.2 Assessing Model Fit

We compare the estimated daily contributions for each chemical constituent to the observed values from the CSN. We find the estimates by taking the median of the posterior distribution for the daily contribution for each chemical constituent, $\sum_{k=1}^{K} \lambda_{pk} f_{kt}$. For Phoenix, there is strong agreement between observed and estimated values for Ca, Fe, K, OC, and Si. For example, as shown in the top left plot of Figure 6, the estimated contributions from OC for Phoenix are very similar to the corresponding observed values for the vast majority of the time points. There is not quite as strong agreement for S and EC but the observed and estimated values generally agree. The contributions from Zn tend to be underestimated but show fairly good agreement with the observed values. However, there is poor agreement between the observed and estimated contributions from Al, Cl, and Ti. These chemical constituents contribute relative small amounts to $PM_{2.5}$ and have observed contributions close to or below their corresponding minimum detection limit (MDL) for the majority of the time points. The top right plot of Figure 6 illustrates that the contributions for Ti are underestimated for most of the time points. Thus, we appear to estimate the source contributions well for chemical constituents that are measured well. However, we did not do as well for the chemical constituents that contribute small amounts to $PM_{2.5}$ and are not as easily measured.

We see similar results for the analysis of Boston in that we generally appear to estimate the

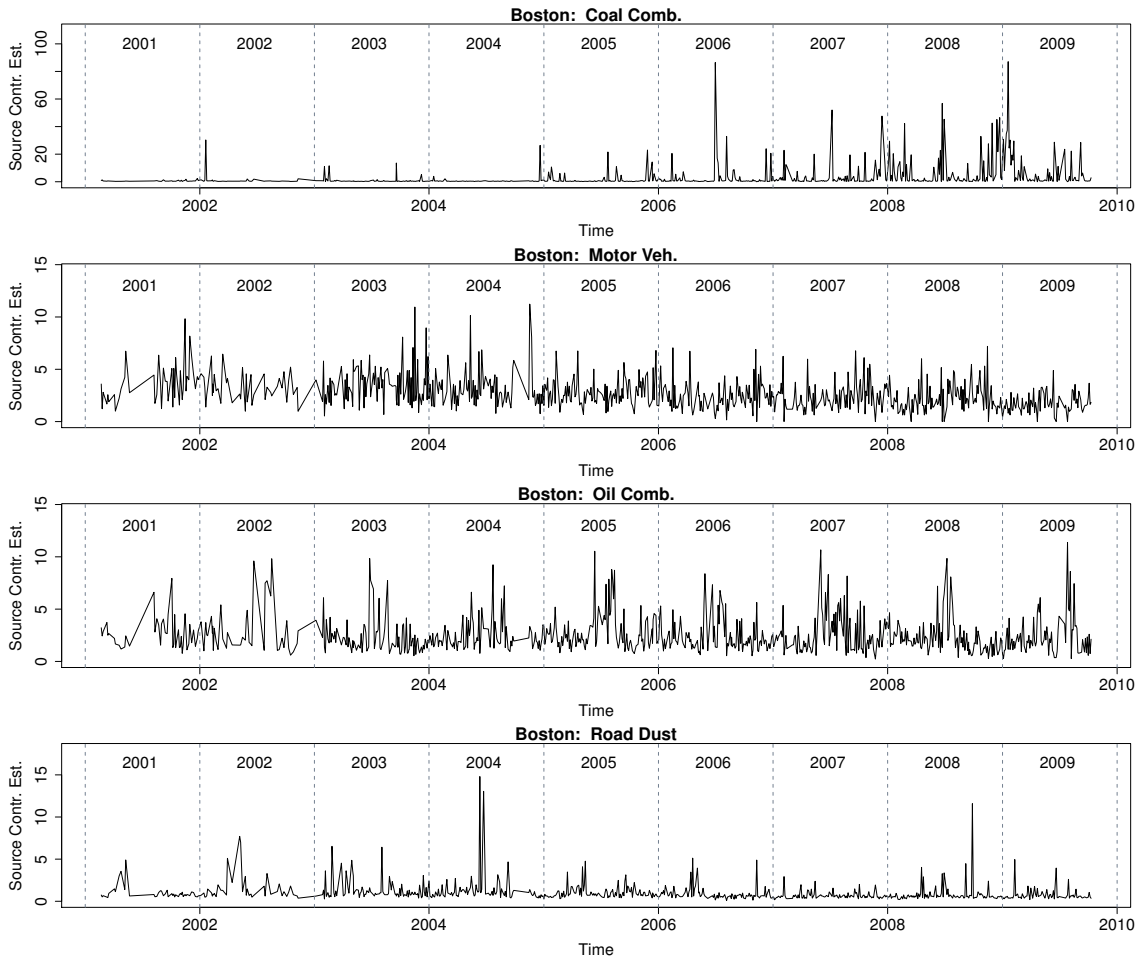

Figure 5: Plots of the source contribution estimates for Boston for all time points after January 1, 2001. The vertical dashed lines mark January 1st for the years 2001 through 2010.

source contributions well for chemical constituents that are measured well, but have poor agreement for the chemical constituents that contribute small amounts to $PM_{2.5}$. For example, as shown in the bottom left plot of Figure 6, the estimated contributions for OC for Boston are very similar to their corresponding observed values for the vast majority of the time points. We observe fairly good agreement between the observed and estimated contributions from Ca, EC, Fe, K, S,and Zn. However, we tend to underestimate the contributions from Ca, Fe, S,and Zn. There is generally poor agreement between the observed and estimated values for Cl, Ni, Pb, Si, Ti, and V. We tend to underestimate the contributions from these constituents. The bottom right plot of Figure 6 illustrates that the contributions for Ti are underestimated for most of the time points. We also note that the Boston analysis overestimates, sometimes drastically, the contributions from Al (Top left plot of Figure 6 in the Supplemental Material).

4.3 Sensitivity Analysis

A sensitivity analysis was performed to determine how sensitive the source contribution estimates are to the choice of prior for the source contributions as well as the sources used in the analysis.

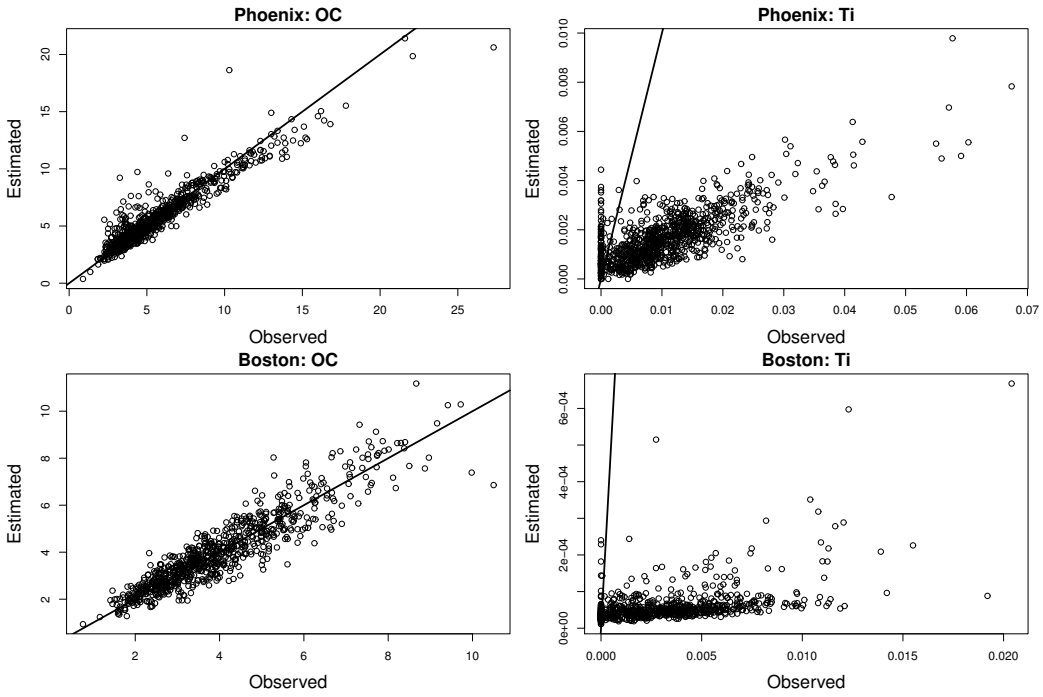

Figure 6: Top plots compare the estimated daily contributions to $PM_{2.5}$ for all time points for OC (top left) and Ti (top right) from the Phoenix analysis to the corresponding observed contributions found using the monitor in the Phoenix area from the Chemical Speciation Network (CSN). Bottom plots compare the estimated daily contributions to $PM_{2.5}$ for all time points for OC (bottom left) and Ti (bottom right) from the Boston analysis to the corresponding observed contributions found using a monitor in the Boston area from the CSN. The solid black lines are the identity functions.

Sections 4 and 5 in the Supplementary Material provide additional information on this sensitivity analysis.

4.3.1 Phoenix

Because there is less agreement among the previous studies in the Phoenix area as to what sources contribute to $PM_{2.5}$, we perform an analysis for Phoenix using different sources. The majority of studies in the Phoenix area included Motor Vehicles/Traffic, Wood/Vegetative Burning, and Dust/Soil/Crustal source categories but were less consistent with other source categories, such as source categories that have contributions from salt, coal, and smelting. Thus, we perform a source apportionment analysis with six sources that includes three additional source categories: Coal Combustion, Metals Processing, and Salt. Since there is greater uncertainty in the source contributions from the three additional sources, we use less informative priors for the source contributions (Table 2). The mean and IQR of the source contribution estimates for the Phoenix analyses are summarized in Table 2.

For the sources that overlap between the two analyses, the mean and the IQR source contribution estimates are similar for Motor Vehicles and Vegetative Burning but differ for Road Dust. Though the contribution estimates are not the same between the two Phoenix analyses, similar conclusions about the common sources between the two analyses can be drawn. For the Phoenix analysis with six sources, as with the analysis of Phoenix with three sources, the source contribution

Table 2: Summary of results from sensitivity analysis for Phoenix. For each source in each analysis, the table gives values for the mean, γ_k, and standard deviation, δ_k, of the normal prior for the log of the source contributions. It also gives the mean ($\bar{f}_k$) and interquartile range (IQR) of the source contribution estimates.

	Analysis	Motor	Road	Veg.	Coal	Metals	Salt
γ_k	Three	0.9	1.2	0.1	—	—	—
	Six	0.9	1.2	0.1	0	0	0
δ_k	Three	1.2	1.2	1.2	—	—	—
	Six	1.2	1.2	1.2	3.4	3.4	3.4
$\bar{f}_k$	Three	3.1	2.7	1.6	—	—	—
	Six	3.2	4.5	1.8	4.3	1.4	2.1
IQR	Three	(1.3, 4.3)	(1.5, 3.6)	(0.7, 1.8)	—	—	—
	Six	(1.4, 4.4)	(2.6, 6.1)	(0.7, 1.8)	(0.4, 3.8)	(0.3, 0.9)	(0.1, 0.7)

estimates for Motor Vehicles appear to increase during the winter months and the source contribution for Vegetative Burning appear to spike at the beginning and end of each year (Figure 2 in the Supplemental Material). We note that, for many the time points in the Phoenix analysis with six sources, the posterior and prior distributions for the log of the source contributions for the Salt and Metal Processing categories are very similar (See Section 4 in the Supplemental Material). This suggests that the data in our analysis provides little information about these additional sources for Phoenix. Therefore, we focus on the results from the Phoenix analysis with three sources.

4.3.2 Boston

The Boston analysis in Section 4.1.2 uses the NEI to create priors for $\log(f_{kt})$ for most sources but chooses a vague prior for the Coal Combustion source category since previous evidence suggests that much of its contributions to ambient PM can be attributed to secondary $PM_{2.5}$. We refer to this analysis as the original analysis. We compare these results to those from an analysis that does not consider the evidence about secondary $PM_{2.5}$ and creates priors based solely on the NEI. We call this analysis the NEI only analysis. Lastly, we perform an analysis that uses vague priors for all sources and does not use the information in the NEI. We refer to this analysis as the no NEI analysis. The different priors used for these three analyses are summarized in Table 3. Table 3 also gives the mean and IQR of the source contribution estimates for the three Boston analyses. The MCMC algorithm for the no NEI analysis was run from 200,000 iterations before convergence was obtained and required more tuning to obtain convergence than for the other two Boston analyses.

Note that the source contribution estimates differ depending on the priors used in the Boston analysis. For NEI only analysis, the source contribution estimates for Coal Combustion are all close to zero with very little temporal variability. The posterior and prior distributions for the $\log(f_{kt})$ for Coal Combustion are very similar for many of the time points suggesting that the prior is driving these estimates toward zero (See Section 5.1 in the Supplemental Material). For the no NEI analysis, the mean of the source contribution estimates for Coal Combustion, Motor Vehicles and Oil Combustion are smaller than for the original Boston analyses. The mean of the source contribution estimates for Road Dust is larger compared to the original and the NEI only analyses.

We also note that the analyses differed greatly in their estimated daily contributions from Al (top plots of Figure 6 in the Supplemental Material). For the original analysis, the contributions

Table 3: Summary of the results from the sensitivity analysis for Boston. For each source in each analysis, the table gives values for the mean, γ_k, and standard deviation, δ_k, of the normal prior for the log of the source contributions. It also gives the mean ($\bar{f}_k$) and interquartile range (IQR) of the source contribution estimates.

Parameter	Analysis	Coal	Motor	Oil	Road
γ_k	Original	0.0	0.7	1.0	-0.3
	NEI only	-4.7	0.7	1.0	-0.3
	No NEI	0.0	0.0	0.0	0.0
δ_k	Original	3.4	1.2	1.2	1.2
	NEI only	1.2	1.2	1.2	1.2
	No NEI	3.4	3.4	3.4	3.4
$\bar{f}_k$	Original	3.1	2.7	2.5	1.0
	NEI only	1.2e-2	2.2	2.3	1.0
	No NEI	1.1	1.2	0.8	1.6
IQR	Original	(0.3, 1.2)	(1.6, 3.5)	(1.4, 3.1)	(0.5, 1.1)
	NEI only	(7.7e-3, 1.2e-2)	(1.2, 2.9)	(1.3, 2.9)	(0.5, 1.1)
	No NEI	(0.2, 1.6)	(0.3, 1.7)	(0.3, 1.1)	(0.5, 2.3)

from Al are being drastically overestimated. This causes the estimate for daily total ambient $PM_{2.5}$ to be overestimated for the majority of time points (bottom left plot of Figure 6 in the Supplemental Material). For the NEI only and no NEI analyses, the contributions from Al are being underestimated (top middle and right plots of Figure 6 in the Supplemental Material). However, this underestimation of the contributions from Al does not cause the daily total ambient $PM_{2.5}$ to be systematically underestimated. There is generally good agreement between the estimated and observed daily total ambient $PM_{2.5}$ for the NEI only and no NEI analyses (bottom middle and left plots of Figure 6 in the Supplemental Material). If the daily total ambient $PM_{2.5}$ is of interest, one may prefer to use the results from the no NEI analysis since daily total ambient $PM_{2.5}$ is not overestimated for most of the time points and the estimate of the contribution from Coal Combustion is not trivially small for the majority of time points.

4.4 Sources of Variation in the Source Contribution Estimates

Because the daily source contributions are often used in subsequent health effects analyses, it is important to consider the sources of variation for these estimates. Our Bayesian analysis highlights two types of variation in the source contribution estimates: statistical uncertainty in the f_{kt} parameter values and natural variability across time for the source contributions. The points in Figure 7 represent the source contribution estimates, $\hat{f}_{kt}$, which vary across time. We use θ_k to denote the natural variability of f_{kt} across time for source k. Each vertical line in Figure 7 represents the statistical uncertainty in the f_{kt} parameter values. We use τ_{kt} to denote the statistical uncertainty in f_{kt}.

We estimate the statistical uncertainty in f_{kt} by taking the sample variance of the posterior draws for the f_{kt} parameter and denote the estimate by $\hat{\tau}^2_{kt}$. This provides an estimate for the variance of the posterior distribution for f_{kt}. Let $\bar{\tau}_k = \sqrt{1/T \sum_{t=1}^{T} \hat{\tau}^2_{kt}}$, so $\bar{\tau}_k$ is an estimate of the statistical uncertainty in the source contribution for an average day for source k. To estimate the natural temporal variability for source k, θ_k, we take the sample variance of the daily source

contribution estimates from the source apportionment analysis. Since we are only interested in providing a summary measure of the degree to which the f_{kt} parameter varies across time for each source, we do not model the temporal dependence between f_{kt} estimates. Thus, the estimate for the natural temporal variability is $\hat{\theta}_k = \sqrt{1/(T-1)\sum_{t=1}^{T}\left(\hat{f}_{kt} - 1/T\sum_{t=1}^{T}\hat{f}_{kt}\right)^2}$ where $\hat{f}_{kt}$ is the median of the posterior distribution for f_{kt} from the source apportionment analysis. The values for $\hat{\theta}_k$ and $\bar{\tau}_k$ for the no NEI Boston analysis and Phoenix analyses with 3 sources are given in Table 4.

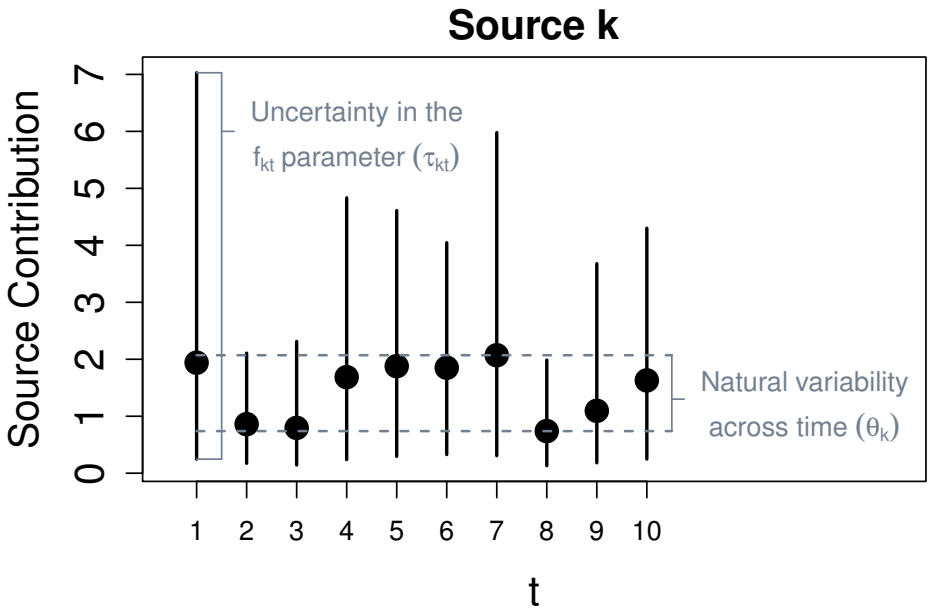

Figure 7: Figure to illustrate the types of variation for the source contributions, f_{kt}, where t is the time point. The points represent source contribution estimates, $\hat{f}_{kt}$, and the vertical lines represent the uncertainty in the source contribution estimate.

To account for the fact that some sources on average emit more than other sources, we divide both $\bar{\tau}_k$, and $\hat{\theta}_k$ by the mean of the source contribution estimates, $\bar{f}_k = 1/T\sum_{t=1}^{T}\hat{f}_{kt}$. We see that the $\bar{\tau}_k/\bar{f}_k$ values are not trivially small with $\bar{\tau}_k$ at least half as large as the mean contribution for all sources for Boston and all but one source for Phoenix. Note that $\hat{\theta}_k$ values are also not trivially small with the $\hat{\theta}_k$ values at least half the overall mean for all sources for both locations. This suggests that there is natural temporal variability in the source contribution estimates for both analyses. Also note that the $\hat{\theta}_k$ value is larger than the corresponding $\bar{\tau}_k$ values for all the sources for Boston, except Motor Vehicles, and all the sources for Phoenix. In the last column of Table 4, we compare our estimates for the two sources of variation and estimate the percentage of the variation that is attributed to temporal variability by $100 \times \hat{\theta}_k^2/\left(\hat{\theta}_k^2 + \bar{\tau}_k^2\right)$. For Motor Vehicles in the Boston analysis, the natural temporal variability is smaller than the statistical uncertainty. Our analysis suggests that using point estimates for source contributions in subsequent health effect analyses will result in confidence intervals for source-specific health effects with poor coverage.

Table 4: Table summarizing estimates for the types of variation in the source contribution estimates for the Bayesian analysis. Here, $\bar{\tau}_k$ denotes the estimate of the statistical variability in the source contributions for an average day for source k, $\hat{\theta}_k$ denotes the estimate of the natural temporal variability for source k, and $\bar{f}_k$ denotes the estimate of the mean contribution from source k. The estimate for the percent of the variation attributed to natural temporal variability (Per. Temp.) is found by $100 \times \hat{\theta}_k^2 / \left(\hat{\theta}_k^2 + \bar{\tau}_k^2\right)$.

Location	Source	$\bar{f}_k$	$\bar{\tau}_k$	$\hat{\theta}_k$	$\bar{\tau}_k/\bar{f}_k$	$\hat{\theta}_k/\bar{f}_k$	Per. Temp.
Boston	Coal. Comb.	1.1	1.3	1.5	1.2	1.3	54.9
	Motor Veh.	1.2	1.6	1.1	1.3	0.9	34.6
	Oil Comb.	0.8	0.8	1.0	1.0	1.2	56.9
	Road Dust	1.6	1.4	1.5	0.9	0.9	53.2
Phoenix	Motor Veh.	3.1	1.8	2.6	0.6	0.8	67.2
	Road Dust	2.7	0.6	1.8	0.2	0.6	90.1
	Veg. Burning	1.6	1.2	1.8	0.8	1.1	67.8

5 Discussion

In this paper, we propose a Bayesian source apportionment model using the multiplicative errors model in (2) that incorporates information from the CSN, NEI, and SPECIATE databases. The Bayesian approach allows us to incorporate *a priori* information about the source emissions. The use of national databases, as opposed to a local study, allows us to incorporate *a priori* information into source apportionment analyses at more than one U.S location. Furthermore, our Bayesian analysis provides us with posterior distributions for the parameters of interest. The posterior distributions are used to estimate the statistical uncertainty and natural variability across time for the source contributions, which provides additional information about uncertainty that is relevant to the estimation of health effects.

For the Boston analysis, we excluded the observations for 12 dates (April 22, 2001; July 2, 2003; July 5, 2003; July 2, 2004; July 5, 2004; July 3, 2005; July 4, 2006; May 27, 2007; July 2, 2007; July 5, 2007; July 2, 2008; and July 3, 2009). An analysis including these dates would have resulted in extremely large estimates for the source contributions from Coal Combustion (above 100 $\mu g/m^3$). For these dates, the CSN suggests there are relatively large contributions from K and S. Thus, a source apportionment analysis including them may be attributing the contributions from other sources, such as fireworks, to Coal Combustion. We believe that a fireworks source only contributes to ambient PM on a few select dates so it would be difficult to estimate the contributions from this source. Therefore, we believe these dates are outliers and perform the Boston analysis removing them. We, however, note that the mean and IQR of the source contribution estimates in the Boston analysis differs between the Boston analysis including these 12 dates and the analysis that excludes them.

Also note, for the Boston analysis, we have poor estimates for the contribution from Al, regardless of the priors used for the source contributions. It would be of interest to further examine what is causing these poor estimates. The SPECIATE data suggests that Al is a major contributor for both Coal Combustion and Road Dust. This difficulty in estimating the contribution from Al may be due to the difficulty in distinguishing the source contributions from Coal Combustion and Road Dust. Including more chemical constituents in the analysis may help distinguish between these

two source categories and improve the estimation of the contributions from Al. However, for our Boston analysis, the CSN estimates for the contributions from other chemical constituents, such as magnesium and selenium, are very low and below the MDL for the majority of the observed time points. An analysis including these constituents and imputing the missing values may better distinguish between contributions from Road Dust and Coal Combustion.

Our analysis for Phoenix demonstrates that the source contribution estimates can differ depending on the sources considered in the analysis. Thus, we suggest using evidence from previous studies to determine the number and type of sources for the Bayesian model. If there is no information available, the NEI database can be use to create several potential models. Model selection tools, such as BIC (Schwarz, 1978), DIC (Spiegelhalter et al., 2002; Celeux et al., 2006), or Bayes factor (Kass and Raftery, 1995), can be used to decide between the potential models. Model selection tools can also be used to determine which chemical constituents to include in the source apportionment model, especially when expert knowledge and previous source apportionment studies in the area are not available.

The Boston analysis shows that the source contribution estimates are sensitive to the choice of priors for the source contributions. The more informative priors improved convergence of the MCMC algorithm and incorporated location-specific prior information. However, the NEI only analysis that created priors based solely on the NEI resulted in source contribution estimates for Coal Combustion that are likely too small. Our original Boston analysis that created informative priors based on the NEI and previous studies drastically overestimated Al, and thus, overestimated the total daily source contributions. This suggests that using the NEI database to create informative priors may not improve estimation of source contributions for locations for which secondary $PM_{2.5}$ is a major contributor to ambient $PM_{2.5}$. In this case, we suggest using vague priors as opposed to ones created using the NEI database.

We also find sensitivity of the results to the parameter constraints on Λ^* (results not shown). We find that if $\lambda^*_{pk} = 1$ corresponds to a chemical constituent that generally contributes relatively small amounts to ambient $PM_{2.5}$, the estimated contributions to $PM_{2.5}$ for each chemical constituent do not agree as strongly with the observed values from the CSN as when $\lambda^*_{pk} = 1$ corresponds to a chemical constituent that generally contributes relatively large amounts to ambient $PM_{2.5}$. Having $\lambda^*_{pk} = 1$ for a chemical constituent that generally contributes relatively small amounts to $PM_{2.5}$ often results in rescaled SPECIATE profile values that are extremely large, which hinders the estimation of the hyperparameters for the informative priors for the free λ^*_{pk} values. Furthermore, it results in λ^*_{pk} parameter values that are on drastically different scales. This makes it difficult to get the MCMC algorithms to converge. Therefore, when setting $\lambda^*_{pk} = 1$ for each source, we recommend choosing a p that corresponds to a chemical constituent that has relatively large contributions for that source and its contributions are above the MDL for most of the observed time points.

Although our Bayesian analysis shows that the results are sensitive to the priors, number of sources, and parameter constraints, exploratory source apportionment analysis models also require user input which can influence the results. Interpretation by the users is required for PCA and PMF to attribute the source contributions and profiles to known sources of air pollution. PMF requires user input to make the model identifiable and the results of the analysis can depend on the constraints chosen to make the model identifiable (Lingwall and Christensen, 2007).

As in Nikolov et al. (2008) and Nikolov et al. (2011), we assume a multiplicative errors source apportionment model which models the ambient $PM_{2.5}$ concentrations using a lognormal distribution. We believe it models the observed data better than a source apportionment model with additive errors since the lognormal distributional assumption on the ambient $PM_{2.5}$ concentrations coincides with the observed skewness in the monitor data. Furthermore, Nikolov et al. (2011) found that a model with multiplicative errors performed better than a model with additive errors, in terms of

obtaining a lower Deviance Information Criteria (DIC) value (Spiegelhalter et al., 2002). However, the lognormal distributional assumption does not allow us to directly model $PM_{2.5}$ concentrations of zero.

Previous source apportionment analyses have been conducted in the Boston and Phoenix areas but none of these analyses use our Bayesian model and incorporate information available on air pollution from all three EPA databases. Yet, the results of our analyses generally agree with previous source apportionment analyses of $PM_{2.5}$ in the Boston and Phoenix areas. The mean of the source contribution estimates for the original and no NEI analyses (Table 3) are similar in scale to estimates found in other source apportionment analyses in the Boston area (Thurston and Spengler, 1985; Laden et al., 2000). Nikolov et al. (2007), Nikolov et al. (2008) and Nikolov et al. (2011), included four categories in their Boston source apportionment analysis: Road Dust, Power Plants, Oil Combustion, and Motor Vehicles, which are very similar to the four categories used in our analysis. Yet, due to differences in their and our analyses, no direct comparisons to our source contribution estimates can be made. However, like our analysis, Nikolov et al. (2011) find that the source categories of Motor Vehicles, Road Dust, Oil Combustion and Power Plants (similar to our Coal Combustion category) are contributing to $PM_{2.5}$ and that these contributions vary across time.

Although there is not strong agreement among the previous studies in the Phoenix area, our results do not contradict any of the key finding in previous studies. Like our Phoenix analyses using three sources (Section 4.1.1) and six sources (Section 4.3.1), many other analysis in the Phoenix area estimate an increase in traffic related contributions in the winter months (Ramadan et al., 2000, 2003; Lewis et al., 2003; Brown et al., 2007). The means of the source contribution estimates for the Phoenix analyses with three and six sources (Table 3) are also similar in scale to estimates found in other source apportionment analyses in the Phoenix area (Ramadan et al., 2000, 2003; Lewis et al., 2003; Brown et al., 2007; Hopke et al., 2006).

5.1 Conclusions

The proposed Bayesian source apportionment model incorporates information from national databases and gives similar results as other source apportionment analyses in the Boston and Phoenix areas using local information. The Bayesian approach allows us to quantify the uncertainty in the source contributions estimates. In our analyses of both Boston and Phoenix, we find that the uncertainties in the source contribution estimates, as well as the variabilities of the source contributions across time for many sources, are not trivially small and should not be ignored. So regardless of researchers' confidence in the *a priori* information, our results suggest that using point estimates for the source contributions in a subsequent analysis for source-specific health effects may lead researchers to underestimate the uncertainty in the health effects estimates. Further exploration is needed to determine how best to incorporate the uncertainty in source contributions into a health effect analysis. We believe that using a hierarchical Bayesian framework for a source-specific health effects analysis would help capture this uncertainty in the source contributions.

Additional information and supplementary material for this article is available online at the journal's website.

5.2 Acknowledgements

The project described was supported by Award Numbers R01ES019560, R21ES020152, and T32012871 from the National Institute of Environmental Health Sciences. The content is solely the responsi-

bility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health.

References

Ault AP, Moore MJ, Furutani H, Prather KA. 2009. Impact of Emissions from the Los Angeles Port Region on San Diego Air Quality during Regional Transport Events. *Environmental Science & Technology* **43**(10): 3500–3506.

Assessment and Standards Division and E.H. Pechan & Associates, Inc. 2007. Documentation for the final 2002 mobile National Emissions Inventory, version 3. Technical report, United States Environmental Protection Agency.

Bell M, Dominici F, Ebisu K, Zeger S, Samet J. 2007. Spatial and temporal variation in PM chemical composition in the United States for health effects studies. *Environmental Health Perspectives* **115**(7): 989–995.

Bell ML, Ebisu K, Peng RD, Samet JM, Dominici F. 2009. Hospital admissions and chemical composition of fine particle air pollution. *American Journal of Respiratory and Critical Care Medicine* **179**(12): 1115–1120.

Billheimer, D. 2001. Compositional receptor modeling. *Environmetrics* **12**(5): 451–467.

Brook RD, Rajagopalan S, Pope CA, Brook JR, ABhatnagar A, Diez-Roux AV, Holguin F, Hong Y, Luepker RV, Mittleman MA, Peters A, Siscovick D, Smith SC, Whitsel L, Kaufman JD. 2010. Particulate matter air pollution and cardiovascular disease. *Circulation* **121**(21): 2331–2378.

Brown SG, Frankel A, Raffuse SM, Roberts PT, Hafner HR, Anderson DJ. 2007. Source apportionment of fine particulate matter in Phoenix, AZ, using positive matrix factorization. *Journal of the Air & Waste Management Association* **57**(6): 741–752.

Byrd RH, Lu P, Zhu C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing* **16**: 1190–1208.

Celeux G, Forbes F, Robert CP, Titterington DM. 2006. Deviance information criteria for missing data models. *Bayesian Analysis* **1**(4): 651–674.

Christensen WF, Schauer JJ, Lingwall JW. 2006. Iterated confirmatory factor analysis for pollution source apportionment. *Environmetrics* **17**(6): 663–681.

Dominici F, Peng RD, Bell ML, Pham L, McDermot A, Zeger SL, Samet JM. 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* **295**(10): 1127–1134.

E.H. Pechan & Associates, Inc. 2006. Documentation for the final 2002 nonpoint sector (FEB 06 version) National Emission Inventory for criteria and hazardous air pollutants. Technical report, United States Environmental Protection Agency.

Emission Factor and Inventory Group 2004. 2002 National Emission Inventory (NEI) preparation plan: Final. Technical report, United States Environmental Protection Agency.

Emission Inventory and Analysis Group 2006. Documentation for the final 2002 point source National Emissions Inventory. Technical report, United States Environmental Protection Agency.

Gao N, Cheng MD, Hopke PK. 1994. Receptor modeling of airborne ionic species collected in SCAQS. *Atmospheric Environment* **28**(8): 1447–1470.

Heaton MJ, Reese CS, Christensen WF 2010. Incorporating time-dependent source profiles using the Dirichlet distribution in multivariate receptor models. *Technometrics* **52**(1): 67–79.

Henry RC. 1997. History and fundamentals of multivariate air quality receptor models. *Chemometrics and Intelligent Laboratory Systems* **37**(1): 37–42.

Henry RC. 2005. Duality in multivariate receptor models. *Chemometrics and Intelligent Laboratory Systems* **77**(1): 59–63.

Hopke PK, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV, Hao L, Neas L, Pinto J, Stolzel M, Suh H, Paatero P, Thurston GD. 2006. PM source apportionment and health effects: 1. Intercomparison of source apportionment results. *Journal of Exposure Science & Environmental Epidemiology* **16**(3): 275–286.

Hsu Y, Divita F 2011. SPECIATE 4.3: Addendum to SPECIATE 4.2 speciation database development documentation. Technical report, United States Environmental Protection Agency.

Kass RE, Raftery AE. 1995. Bayes factors. *Journal of the American Statistical Association* **90**(430): 773–795.

Kim E, Hopke PK 2008. Source characterization of ambient fine particles at multiple sites in the Seattle area. *Atmospheric Environment* **42**(24): 6047–6056.

Koutrakis P, Spengler JD. 1987. Source apportionment of ambient particles in Steubenville, OH using specific rotation factor analysis. *Atmospheric Environment* **21**(7): 1511–1519.

Laden F, Neas LM, Dockery DW, Schwartz J. 2000. Association of fine particulate matter from different sources with daily mortality in six U.S. cities. *Environmental Health Perspectives* **108**(10): 941–947.

Lewis CW, Norris GA, Conner TL, Henry RC. 2003. Source apportionment of Phoenix $PM_{2.5}$ aerosol with the Unmix receptor model. *Journal of the Air & Waste Management Association* **53**(3): 325–338.

Liming Z, Hopke PK, Weixiang Z. 2009. Source apportionment of airborne particulate matter for the Speciation Trends Network site in Cleveland, OH. *Journal of the Air & Waste Management Association* **59**(3): 321–331.

Lingwall JW, Christensen WF. 2007. Pollution source apportionment using *a priori* information and positive matrix factorization. *Chemical and Intelligent Laboratory Systems* **87**: 281–294.

Lingwall JW, Christensen WF, Reese CS. 2008. Dirichlet based Bayesian multivariate receptor modeling. *Environmetrics* **19**(6): 618–629.

Mar TF, Norris GA, Koenig JQ. 2000. Associations between air pollution and mortality in Phoenix, 1995-1997. *Environmental Health Perspectives* **108**(4): 347–353.

Mar TF, Ito K, Koenig JQ, Larson TV, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Neas L, Stolzel M, Paatero P, Hopke PK, Thurston GD. 2006. PM source apportionment and health effects. 3. Investigation of inter-method variations in associations between estimated source contributions of $PM_{2.5}$ and daily mortality in Phoenix, AZ. *Journal of Exposure Science and Environmental Epidemiology* **16**: 311–320.

Monitoring and Quality Assurance Group 1999. Particulate matter $PM_{2.5}$ speciation guidance document: Final draft. Technical report, United States Environmental Protection Agency.

Mueller D, Uibel S, Takemura M, Klingelhoefer D, Groneberg DA. 2011. Ships, ports and particulate air pollution - an analysis of recent studies. *Journal of Occupational Medicine and Toxicology* **6** (31).

Nikolov MC, Coull BA, Catalano PJ, Diaz E, Godleski JJ. 2008. Statistical methods to evaluate health effects associated with major sources of air pollution: a case-study of breathing patterns during exposure to concentrated Boston air particles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **57**(3): 357–378.

Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. 2007. An informative Bayesian structural equation model to assess source-specific health effects of air pollution. *Biostatistics* **8**(3): 609–624.

Nikolov MC, Coull BA, Catalano PJ, and Godleski JJ. 2011. Multiplicative factor analysis with a latent mixed model structure for air pollution exposure assessment. *Environmetrics* **22**(2): 165–178.

Paatero P. 1997. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems* **37**(1): 23–35.

Paatero P. Tapper U. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**(2): 111–126.

Park E, Guttorp P, Henry RC. 2001. Multivariate receptor modeling for temporally correlated by using MCMC. *Journal of the American Statistical Association* **96**(456): 1171–1183.

Park E, Spiegelman CH, Henry RC. 2002. Bilinear estimation of pollution source profiles and amounts by using multivariate receptor models. *Environmetrics* **13**: 775–798.

Park E, Oh M, Guttorp P. 2002. Multivariate receptor models and model uncertainty. *Chemometrics and Intelligent Laboratory Systems* **60**(12): 49–67.

Peng RD, Bell ML, Geyh AS, McDermott A, Zeger SL, Samet JM, Dominici F. 2009. Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental Health Perspectives* **117**(6): 957–963.

Pope CA, Dockery DW. 2006. Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association* **56**(6): 709–742.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ramadan Z, Eickhout B, Song XH, Buydens L, Hopke PK. 2003. Comparison of positive matrix factorization and multilinear engine for the source apportionment of particulate pollutants. *Chemometrics and Intelligent Laboratory Systems* **66**(1): 15–28.

Ramadan Z, Song XH, Hopke PK. 2000. Identification of sources of Phoenix aerosol by positive matrix factorization. *Journal of the Air & Waste Management Association* **50**(8): 1308–1320.

Rohr AC, Wyzga RE. 2012. Attributing health effects to individual particulate matter constituents. *Atmospheric Environment* **62**(0): 130–152.

Schwartz J, Dockery DW, Neas LM. 1996. Is daily mortality associated specifically with fine particles? *Journal of the Air & Waste Management Association* **46**(10): 927–939.

Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics* **6**(2): 461–464.

Song XH, Polissar AV, Hopke PK. 2001. Sources of fine particle composition in the Northeastern US. *Atmospheric Environment* **35**(31): 5277–5286.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**(4): 583–639.

Thurston GD, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV, Hao L, Neas L, Pinto J, Stölzel M, Suh H, Hopke PK. 2005. Workgroup report: Workshop on source apportionment of particulate matter health effects–intercomparison of results and implications. *Environmental Health Perspectives* **113**(12): 1768.

Thurston GD, Spengler JD. 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. *Atmospheric Environment* **19**(1): 9–25.

United States Environmental Protection Agency 2012a. 2002 National Emissions Inventory data & documentation. Website. `http://www.epa.gov/ttn/chief/net/2002inventory.html`. [accessed September 17, 2012]

United States Environmental Protection Agency 2012b. 2005 National Emissions Inventory data & documentation. Website. `http://www.epa.gov/ttn/chief/net/2005inventory.html`. [accessed September 17, 2012]

United States Environmental Protection Agency 2012c. 2008 National Emissions Inventory data. Website. `http://www.epa.gov/ttn/chief/net/2008inventory.html`. [accessed September 17, 2012]

United States Environmental Protection Agency 2012d. Air pollution monitoring. Website. `http://www.epa.gov/oaqps001/montring.html`. [accessed September 17, 2012]

United States Environmental Protection Agency 2012e. Basic information. Website. `http://www.epa.gov/pm/basic.html`. [accessed April 19, 2013]

United States Environmental Protection Agency 2012f. Chemical speciation. Website. `http:http://www.epa.gov/ttn/amtic/speciepg.html`. [accessed September 17, 2012]

United States Environmental Protection Agency 2012g. Download detailed AQS data. Website. `www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm`. [accessed September 17, 2012]

United States Environmental Protection Agency 2012h. Emissions inventories. Website. `http://www.epa.gov/ttn/chief/eiinformation.html`. [accessed September 17, 2012]

United States Environmental Protection Agency 2012i. SPECIATE. Website `http://www.epa.gov/ttnchie1/software/speciate`. [accessed September 17, 2012]

United States Environmental Protection Agency 2012j. SPECIATE data browser: Home. Website. `http://cfpub.epa.gov/si/speciate/`. [accessed September 17, 2012]

Vutukuru S, Dabdub D. 2008. Modeling the effects of ship emissions on coastal air quality: A case study of southern California. *Atmospheric Environment* **42**: 3751–3764.

Wolbers M, Stahel W. 2005. Linear unmixing of multivariate observations. *Journal of the American Statistical Association* **100**(472): 1328–1342.

Zanobetti A, Franklin M, Koutrakis P, Schwartz J. 2009. Fine particulate air pollution and its components in association with cause-specific emergency admissions. *Environmental Health* **8**(58).

Zhou L, Kim E, Hopke PK, Stanier CO, Pandis S. 2004. Advanced factor analysis on Pittsburgh particle size-distribution data special issue of aerosol science and technology on findings from the fine particulate matter supersites program. *Aerosol Science and Technology* **38**(sup1): 118–132.