A hierarchical modeling approach to estimate regional acute health effects of particulate matter sources

Jenna R. Krall, Emory University
A.J. Hackstadt, Vanderbilt University
R.D. Peng, Johns Hopkins University

Journal Title: Statistics in Medicine
Volume: Volume 36, Number 9
Publisher: Wiley: 12 months | 2017-04-30, Pages 1461-1475
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1002/sim.7210
Permanent URL: https://pid.emory.edu/ark:/25593/s9jbf

Final published version: http://dx.doi.org/10.1002/sim.7210

Copyright information:
© 2017 John Wiley & Sons, Ltd.

Accessed December 2, 2018 11:54 PM EST
A hierarchical modeling approach to estimate regional acute health effects of particulate matter sources

J. R. Krall\textsuperscript{a}, A. J. Hackstadt\textsuperscript{b}, and R. D. Peng\textsuperscript{c}

\textsuperscript{a}Department of Biostatistics & Bioinformatics, Emory University, 1518 Clifton Road, Mailstop 1518-002-3AA, Atlanta, GA 30322

\textsuperscript{b}Department of Biostatistics, Vanderbilt School of Medicine, 2525 West End Avenue, Suite 11000, Nashville, TN 37203

\textsuperscript{c}Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205

Abstract

Exposure to particulate matter (PM) air pollution has been associated with a range of adverse health outcomes, including cardiovascular disease (CVD) hospitalizations and other clinical parameters. Determining which sources of PM, such as traffic or industry, are most associated with adverse health outcomes could help guide future recommendations aimed at reducing harmful pollution exposure for susceptible individuals. Information obtained from multisite studies, which is generally more precise than information from a single location, is critical to understanding how PM impacts health and to informing local strategies for reducing individual-level PM exposure. However, few methods exist to perform multisite studies of PM sources, which are not generally directly observed, and adverse health outcomes. We developed SHARE, a hierarchical modeling approach that facilitates reproducible, multisite epidemiologic studies of PM sources. SHARE is a two-stage approach that first summarizes information about PM sources across multiple sites. Then, this information is used to determine how community-level (i.e. county- or city-level) health effects of PM sources should be pooled to estimate regional-level health effects. SHARE is a type of population value decomposition that aims to separate out regional-level features from site-level data. Unlike previous approaches for multisite epidemiologic studies of PM sources, the SHARE approach allows the specific PM sources identified to vary by site. Using data from 2000–2010 for 63 northeastern US counties, we estimated regional-level health effects associated with short-term exposure to major types of PM sources. We found PM from secondary sulfate, traffic, and metals sources was most associated with CVD hospitalizations.
**Keywords**
Cardiovascular health; Health effects; Particulate matter sources; Source apportionment; Statistical methods in Epidemiology

1. Introduction
Exposure to particulate matter air pollution less than 2.5 μm in aerodynamic diameter, referred to as PM$_{2.5}$, has been associated with acute cardiovascular health effects based on both epidemiologic and toxicological studies [1]. Cardiovascular health effects related to short-term PM$_{2.5}$ exposure include hospitalizations for cardiovascular diseases (CVD) [2, 3] as well as subclinical health measures such as autonomic dysfunction [4, 5] and increased blood pressure [6]. Current recommendations for those with heart disease and other susceptible individuals include referring to the Air Quality Index (AQI), which provides daily levels of ambient PM$_{2.5}$ concentrations and associated health risks, and limiting physical activity when PM$_{2.5}$ concentrations are high [7, 8]. However, PM$_{2.5}$ is a complex chemical mixture generated by sources such as traffic, industry, and vegetative burning [9, 10], and these sources emit combinations of chemical constituents that vary in their associations with adverse health outcomes [11, 12, 13]. Determining which types of PM$_{2.5}$ sources, or which combinations of chemical constituents, are most toxic could lead to development of more targeted recommendations to reduce health risks in susceptible subpopulations.

In the most recent US Environmental Protection Agency (US EPA) scientific review of the health effects of PM, emphasis was placed on results from multisite epidemiologic studies because such studies are critical for more precisely estimating health effects associated with PM exposure, identifying potential confounders and effect modifiers, and representing health effects across the US [1]. Multisite epidemiologic studies have identified positive associations between short-term PM exposure and CVD hospitalizations, including the National Morbidity and Mortality Study of PM$_{10}$ in 14 US cities [14], the Medicare Air Pollution Study (MCAPS), which analyzed PM$_{2.5}$ in 204 US counties [2], as well as multisite studies in Europe [15, 16]. These multisite studies of PM and CVD hospitalizations contributed to the US EPA conclusions that a “causal relationship exists between short-term PM$_{2.5}$ exposure and cardiovascular effects” [1]. Further analyses of the MCAPS data found associations between CVD hospitalizations and PM$_{2.5}$ elemental carbon (EC) and organic carbon (OC) matter in 119 US communities [17], and PM$_{2.5}$ vanadium, nickel, and EC were associated with CVD effect estimates for PM$_{2.5}$ in 106 US counties [18]. However, these PM$_{2.5}$ constituents can be emitted by multiple sources of PM$_{2.5}$ and therefore it is not currently known which PM$_{2.5}$ sources are most associated with adverse cardiovascular outcomes.

Estimating health effects associated with exposure to source-specific PM$_{2.5}$ is challenging because PM$_{2.5}$ sources are generally unobserved in ambient air and are frequently estimated using source apportionment models. Commonly, source apportionment models are applied to concentrations of PM$_{2.5}$ and its chemical constituents observed at single ambient
monitors. Performing multisite studies of PM$_{2.5}$ source types and health is not only challenging because sources are unobserved at each site but also because PM$_{2.5}$ sources vary spatially in chemical composition both across the US [9, 19, 20, 21] and within a single community [10].

In multisite epidemiologic studies of total PM$_{2.5}$, estimated community-level (i.e. county- or city-level) health effects are frequently pooled across sites to estimate regional-level effects. However, it is unclear how to pool community-level health effects for PM$_{2.5}$ source types because, unlike total PM$_{2.5}$, the presence of PM$_{2.5}$ source types varies across communities. Pooling estimated community-level health effects of PM$_{2.5}$ source types requires determining which, if any, source types are similar in chemical composition across monitors. Commonly, ad hoc approaches are used to match estimated sources between monitors. These methods include using the inferred chemical makeup of each source type, for example matching traffic-related sources based on amounts of EC and OC, as well as matching sources based on the temporal correlation of PM$_{2.5}$ by source type [22, 10, 23]. These methods have not been evaluated from a statistical perspective in multisite epidemiologic studies for their ability to estimate health effects corresponding to PM$_{2.5}$ source types. Some source apportionment models have been extended to handle multisite data, though they are not appropriate when PM$_{2.5}$ sources vary across the study site. Positive Matrix Factorization (PMF) [24] 5.0 can incorporate data from multiple sites to improve source estimation, but requires sources to be homogeneous across sites and therefore is not generally an appropriate approach to perform multisite epidemiologic studies. Two previous multisite studies of PM$_{2.5}$ sources have extended source apportionment models to multiple monitors in a region by assuming that each source type has the same chemical composition across monitors [25, 26], though these approaches are generally inappropriate because of the known spatial variability in PM$_{2.5}$ sources.

Therefore, existing methods have two major limitations when conducting multisite studies of acute health effects of PM$_{2.5}$ sources. Either they require ad hoc assessment of source similarity between monitors, which limits their utility in large, regional-level studies, or they require unreasonable assumptions about the homogeneity of pollution sources in a region. To address these limitations, we developed SHARE, a hierarchical modeling approach to estimate acute health effects of PM$_{2.5}$ sources in multisite epidemiologic studies. SHARE identifies which monitors measure PM$_{2.5}$ source types that are similar in chemical composition, and whose estimated health effects can be pooled across communities in multisite studies.

This paper is organized as follows. Section 2 provides details about the relevant methods including source apportionment approaches, epidemiologic models of the health effects of PM$_{2.5}$ sources, and our proposed SHARE method. We introduce the data in Section 3 and provide a simulation study of our SHARE approach in Section 4. In Section 5, we applied SHARE to the MCAPS dataset to estimate regional-level associations between daily cardiovascular (CVD) hospitalizations and short-term exposure to PM$_{2.5}$ sources for 63 counties in the northeastern US from 2000–2010. We have made software publicly available to apply SHARE (https://github.com/kralljr/share).
2. Methods

2.1. Source apportionment

Many source apportionment models have been proposed to estimate sources of PM$_{2.5}$ from PM$_{2.5}$ chemical constituent data. In this section, we briefly describe the standard source apportionment framework and common source apportionment models. For PM$_{2.5}$ chemical constituent concentrations from one ambient monitor, source apportionment generally assumes the matrix of observed concentrations for $T$ days and $P$ chemical constituents $X_{[T \times P]}$ is the product of two unobserved matrices $F$ and $\Lambda$ such that

$$X_{[T \times P]} = F_{[T \times L]} \Lambda_{[L \times P]} + \varepsilon_{[T \times P]}$$  \hspace{1cm} (1)$$

where $L$ is the number of sources. The source concentration matrix $F_{[T \times L]}$ represents the concentration of PM$_{2.5}$ from each unobserved source type $l$ ($l = 1 \ldots L$) on day $t$ ($t = 1 \ldots T$) and the profile matrix $\Lambda_{[L \times P]}$ describes the relative contribution of each chemical constituent $p$ ($p = 1 \ldots P$) to each source type $l$. The profile matrix characterizes the chemical composition of each source type and is used to link estimated sources to known sources of pollution at that monitor. The $L$ time series from the source concentration matrix $F_{[T \times L]}$ are frequently used in time series regression models to estimate community-level associations between sources and adverse health outcomes. The last matrix, $\varepsilon_{[T \times P]}$, represents measurement error or other variation not captured by the model. Source apportionment models differ from other latent variable models because they aim to estimate interpretable $F$ and $\Lambda$ such that $f_{tl} \geq 0$ and $\lambda_{lp} \geq 0$ for all $t$, $l$, $p$ and $\sum_{l=1}^{L} f_{tl}$ should be approximately equal to the total PM$_{2.5}$ mass observed on day $t$.

Examples of commonly applied source apportionment methods include Positive Matrix Factorization (PMF) [24], Absolute Principal Component Analysis (APCA) [10, 27], and Unmix [28]. These methods differ in how they implement the positivity constraints when estimating $F$ and $\Lambda$. For example, APCA estimates sources of PM$_{2.5}$ at one monitor using rescaled results from Principal Component Analysis (PCA). Briefly, to obtain mostly positive daily source concentrations, APCA estimates PCA scores using the uncentered data (whereas standard PCA estimates scores using centered data). Then to ensure the sum of daily source-specific PM$_{2.5}$ is approximately equal to daily total PM$_{2.5}$, APCA rescales the resulting scores by regressing daily total PM$_{2.5}$ on the estimated APCA scores. APCA can be easily implemented using standard statistical software. Technical details of APCA are summarized in the Supplementary Material, Appendix A.

One limitation of commonly applied source apportionment methods is that they are designed for data from individual monitors and cannot be easily extended to multiple monitors across a region. Thurston et al. [26] extended standard APCA to multiple monitors by assuming that PM$_{2.5}$ source types do not vary between monitors. We will refer to this method as multiple monitor APCA or mAPCA and the method is summarized in the Supplementary Material, Appendix A. The method for mAPCA is similar to APCA, except the concatenated data across monitors are used in place of the data from an individual monitor. The mAPCA
approach provides a framework for comparing and combining sources across monitors, but
assumes the source profiles are the same across monitors. This assumption means mAPCA
effectively estimates concentrations for PM$_{2.5}$ sources at a particular monitor that (1) may
not actually be present in the geographic area containing that monitor or (2) may have a
chemical composition that differs from the regionally-estimated source. While these
assumptions of mAPCA are problematic, mAPCA does not require ad hoc steps to estimate
PM$_{2.5}$ sources across a large region to estimate regional-level health effects. We
implemented APCA and mAPCA using R version 3.0 [29]

2.2. Estimating associations between PM$_{2.5}$ sources and CVD hospitalizations

In this section, we assume that PM$_{2.5}$ source concentrations $f_{c}$ have already been estimated
for one monitor in each county $c$ using source apportionment. Commonly, regional-level
health effects associated with short-term exposure to PM$_{2.5}$ are estimated by pooling
estimated community-level health effects; in our study a community corresponds to a US
county. To estimate county-level health effects of PM$_{2.5}$ sources, log-linear time series
models are fitted to daily counts of morbidity and each PM$_{2.5}$ source type $l$. Specifically in
our models, we assumed the number of CVD hospitalizations for day $t$ for a particular
county $c$, $y_{t,c} \sim \text{Poisson}(\mu_{t,c})$ and

$$\log(\mu_{t,c}) = \beta_{0,c} + \hat{f}_{t,c} \beta_{l,c} + \text{confounders} \quad (2)$$

where confounders may include control for meteorology, day of week, long-term trends in
CVD hospitalizations, and others. For each county $c$, we estimated associations between
PM$_{2.5}$ source type $l$ and CVD hospitalizations using the log relative risk $\hat{\beta}_{l,c}$ and its
corresponding standard error. We fitted separate regression models for each source type so
that the interpretation of the coefficient corresponding to source type $l$ will be the same
across counties with varying source types.

To estimate regional associations between PM$_{2.5}$ sources and CVD hospitalizations, we
fitted two-level Bayesian hierarchical models as in previous multisite studies of CVD
hospitalizations and PM$_{2.5}$ [2, 17]. These models allow the estimation of regional
associations by pooling county-specific associations from equation 2. The hierarchical
model assumes the estimated log relative risks for each source $l$ and county $c$, $\hat{\beta}_{1,c}$, are
normally distributed and centered around the true log relative risk $\beta_{l,c}$

$$\hat{\beta}_{1,c} \sim N(\beta_{1,c}, \hat{\sigma}^{2}_{l,c})$$

$$\beta_{l,c} \sim N(\theta_{l}, \phi^{2}_{l}) \quad (3)$$

where $\hat{\sigma}_{l,c}$ is the estimated standard error of $\hat{\beta}_{l,c}$ from the time series regression model
(equation 2) and is assumed to be known. In the second level of the model, the county-
specific log relative risks $\beta_{l,c}$ follow a normal distribution with mean $\theta_{l}$ the regional log
relative risk and the parameter of interest. This model allows the estimation of regional-level

Stat Med. Author manuscript; available in PMC 2018 April 30.
health effects, $\theta_k$ using a large number of counties, each with many days of data. We fitted the hierarchical models using the TLNise software [30, 17] implemented in R.

### 2.3. SHared Across a REgion (SHARE) method

Multisite studies of PM$_{2.5}$ source types and health are challenging because each source $i$ may vary in its presence and chemical composition across sites. Therefore, it is not known which sources are similar in chemical composition across sites to inform how estimated community-level health effects should be pooled across sites in equation 3. We propose SHARE, a hierarchical modeling approach that facilitates estimating regional-level associations between PM$_{2.5}$ sources and adverse health outcomes. SHARE is a two-stage approach. In the first stage, we compare estimated community-level sources to determine those *major sources*, for example traffic, that are present at many monitors in the study. In the second stage, we determine how similar these *major sources* are to sources present at each community. This information can then be used to guide pooling estimated community-level health effects to estimate regional-level health effects of *major* PM$_{2.5}$ sources as in equation 3.

Using standard source apportionment methods, we first estimate source profiles $\Lambda_i$ and source concentrations $F_i$ at each monitor $i$, as described in Section 2.1. For each monitor $i$, most source apportionment models roughly assume that the PM$_{2.5}$ chemical constituent concentrations, $X_i$, are approximately equal to the product of a source concentration matrix, $F_i$, and a source profile matrix, $A_k$, such that $X_i \approx F_i \Lambda_i$ (e.g. equation 1). The main aim of the SHARE approach is to determine how information can be pooled across monitors $i$.

#### 2.3.1. Estimating “major sources”—In the first stage of SHARE, we estimate a population-level matrix $\Lambda$ that represents those *major sources* whose chemical compositions are similar across monitors and therefore represent sources whose estimated health effects should be pooled across communities in a hierarchical model. It is important to note that the information contained within $\Lambda$ will not exactly correspond to source profiles in the standard source apportionment framework, but rather $\Lambda$ will only be used to guide pooling community-specific estimated health effects and to help interpret regional-level estimated health effects.

We estimate $\Lambda$ using ideas from Population Value Decomposition (PVD) [31], which is an approach that estimates population-level features from data across multiple individuals and uses these features to approximate individual-level data. As in PVD, we find the population-level profile matrix $\Lambda$ by applying PCA to the matrix of concatenated source profiles for all $M$ ambient monitors, $\Lambda = [\Lambda^T_1, \Lambda^T_2, \ldots, \Lambda^T_M]^T$. Then $\Lambda = \Lambda W W^T$, where $W$ is the matrix of principal component loadings of $\Lambda^T \tilde{\Lambda}$ and $W W^T$ is the identity matrix. Letting $A^T$ be the matrix of the first $L$ principal component loadings that explain most of the variability in $\tilde{\Lambda}$, $\Lambda$ will represent the profiles corresponding to major types of PM$_{2.5}$ sources.

#### 2.3.2. Pooling community-level data—In this second stage, we aim to determine which sources represented by the *major source* profile matrix $\Lambda$ are represented in each monitor-specific source concentration matrix, $F_i$. Using our *major source* profile matrix $\Lambda$, we can
express each monitor’s profile matrix as $A_i \approx A_i A^T \Lambda$ because $\tilde{\Lambda} \approx \tilde{\Lambda} \Lambda^T \Lambda$. Then, using $A_i$ and $\Lambda$, we can rewrite

$$X_i \approx F_i \Lambda_i \approx (F_i \Lambda_i \Lambda^T \Lambda) \Lambda = \tilde{F}_i \Lambda$$

This straightforward application of PVD estimates source concentrations at each monitor $i$ as $\tilde{F}_i = F_i (\Lambda_i \Lambda^T) = F_i \Psi_i$. Because $F_i$ represents the source concentrations present at monitor $i$, $\tilde{F}_i$ is a linear combination of the source concentrations estimated at monitor $j$ based on the major sources represented in $\Lambda$. Therefore, we cannot use $\tilde{F}_i$ to estimate community-level health effects of PM$_{2.5}$ sources because the linear combination may not represent exposures present at monitor $i$.

To address this limitation, we represent the relationship between sources at monitor $i$ and major sources as a bipartite graph. We note that entries in $\Psi_i = \Lambda_i \Lambda^T$ will be large for sources at monitor $i$ that are similar in chemical composition to the major sources represented by $\Lambda$. The bipartite graph representation will match sources present at monitor $i$ to major sources, where an edge indicates a source at monitor $i$ is similar in chemical composition to a major source. We then estimate $\Psi_i$ using the off-diagonal of the corresponding adjacency matrix of the bipartite graph, which consists of ones and zeroes with ones indicating the presence of an edge.

We used the Hungarian method [32] for optimal bipartite matching to estimate the adjacency matrix $\Psi_i$. The Hungarian method finds those sources at monitor $i$ that are similar in chemical composition to major sources by minimizing the sum of the corresponding edges in $\Psi_i$. Recall that $(\Psi_i)_{L_i \times \Lambda_i} = \Lambda_i \Lambda^T$. If we rescale $A_i$ and $\Lambda$ to contain only unit vectors, $\psi_i, l, j = \cos(\alpha_l, i, j)$, where $\alpha_l, i, j$ is the angle between the chemical composition of source $l$ at monitor $i$ and major source $j$. Since smaller angles correspond to sources at monitor $i$ that are similar in chemical composition to major sources, we applied the Hungarian method to the matrix of angles $\alpha_i, l, j$ to find $\Psi_i$ as in Figure 1A. We limited matches to angles less than 45 degrees, so that some sources at monitor $i$ may differ in chemical composition compared with major sources. This cutoff allows “local” sources, which are sources that are only found at one or a few monitors in the study (e.g. factories) or sources that have substantial variation in chemical composition across monitors. Local sources may be of interest in individual community studies, but are not the focus of this study of regional-level health effects of PM$_{2.5}$ sources. The cutoff of 45 degrees ensures a matched $A_i$ at monitor $i$ is closer to the major source represented in $\Lambda$ than to a vector orthogonal to the major source. We did not find that our results were sensitive to the cutoff angle selected.

Then using the estimated adjacency matrix $\Psi_i$, $\tilde{F}_i = F_i \Psi_i$, will be a reordering of $F_i$ based on the chemical composition of major sources. For a source at monitor $i$ that is not chemically similar to any major sources, the corresponding column $l$ in $\Psi_i$ will contain all zeros and $\tilde{F}_i, t, l, i = 0$ for all days $t$. Because we estimate $\Psi_i$ using an adjacency matrix, the concentrations in $\tilde{F}_i$ are estimated using only the chemical composition of sources at monitor $i$, $A_i$.
columns of $\mathbf{F}_i$ can then be used to estimate community-level health effects of short-term exposure to PM$_{2.5}$ sources. It is important to note that when a source type $l$ is not present at a monitor, $f_{t,i,l}=0$ for all $t$, therefore we do not estimate associations between PM$_{2.5}$ source type $l$ and health using data from that monitor.

Therefore, the SHARE approach gives two results: the major source profile matrix $\Lambda$, that can be used to summarize major sources within an area, and the monitor-specific source profile matrices $\mathbf{F}_i$ which can be used to estimate regional-level health effects. The $\mathbf{F}_i$ are critical because they represent a reordering of the source concentrations $F_i$ based on $\Lambda$.

Because the $\mathbf{A}$th columns of each $\mathbf{F}_i$ correspond to the same major source, each source $l$ can be pooled across monitors to estimate regional-level effects as in equation 3.

3. Data

The US Environmental Protection Agency’s Chemical Speciation Network (EPA CSN) is a national monitoring network of approximately 250 monitors that measure ambient air concentrations for total PM$_{2.5}$ mass and over 50 PM$_{2.5}$ chemical constituents roughly every third or sixth day. We restricted our analysis to 24 chemical constituents of PM$_{2.5}$ (Supplementary Material, Table S1) that contributed to previously identified PM$_{2.5}$ source types in the US [9, 10, 20]. These constituents include major ions (e.g. sulfate and nitrate), metals (e.g. zinc and vanadium), and carbon-containing constituents (EC and OC). For the eleven year period from 2000–2010, we created a dataset of 85 EPA CSN monitors in northeastern US counties (Figure 2) that each had more than 50 days measuring all 24 PM$_{2.5}$ chemical constituents and total PM$_{2.5}$ mass. These monitors fall within the northeast and the industrial midwest regions [1, 14] in coastal, industrial, and heavily populated counties. We also obtained daily temperature and dew point temperature for each county from the National Oceanic and Atmospheric Administration [33].

To estimate associations between PM$_{2.5}$ source types and CVD hospitalizations, we used daily emergency CVD hospitalizations for Medicare enrollees from the Centers for Medicare and Medicaid, aggregated by county. We restricted our dataset to 63 counties in the northeastern US containing at least one EPA CSN monitor such that all 85 monitors from our restricted EPA CSN dataset fall within one of these 63 counties. As in previous studies of PM and emergency CVD hospitalizations, we included primary diagnoses of heart failure, heart rhythm disturbances, cerebrovascular events, ischemic heart disease, and peripheral vascular disease in our daily counts for CVD hospitalizations [17, 34]. Because the Medicare data analyzed for this study did not include individual identifiers, we did not obtain consent from individuals. This study was reviewed and exempted by the Institutional Review Board at the Johns Hopkins Bloomberg School of Public Health.

4. Simulation study

We used a simulation study to test the performance of SHARE. We simulated PM$_{2.5}$ by source type for sources identified in the northeastern US including traffic, fireworks, soil dust, secondary sulfate, salt, metals, and a miscellaneous phosphorus/vanadium (P/V) source. For each monitor, we simulated which source types generated total PM$_{2.5}$ based on
one of 5 subregions, or areas with different source types (Table 1). These subregions represent a potential spatial distribution of identifiable sources across a region, where identifiable means the source is both present in the subregion and located close enough to an ambient monitor to be detected. Some sources such as traffic are spatially variable [10] and in order for an ambient monitor to identify a traffic source, there must both be traffic in the community and the monitor must be located reasonably close to a roadway. PM$_{2.5}$ from sources containing salt, such as sea salt or road salt, may only be present in coastal communities or communities with a lot of snow [35, 9]. All subregions included PM$_{2.5}$ from soil dust and secondary sulfate, since both soil (also frequently referred to as dust or crustal) and secondary sources have been identified across the US [9, 10, 36, 21]. We simulated multisite, regional datasets of PM$_{2.5}$ chemical constituent concentrations observed at multiple monitors, where some monitors were in the same subregion and identified the same source types and some monitors were in different subregions and identified different source types. Details about the simulated data can be found in the Supplementary Material, Appendix B.

4.1. Estimating PM$_{2.5}$ sources

We first tested whether SHARE could correctly determine sources similar in chemical composition across 25 monitors, and this information guides pooling community-level estimated health effects across monitors. We simulated data for 5 monitors in each of the 5 subregions (Table 1). We simulated a total of 120 sources across 25 monitors, where each monitor had 4, 5 or 7 source types depending on its corresponding subregion (Table 1). As a measure of whether SHARE correctly determines sources similar in chemical composition across monitors, we computed the percent of correct source identifications across sources and monitors. For example, if SHARE failed to identify a soil-related source at all 5 monitors in subregion I, but otherwise correctly identified sources, then SHARE was correct for 115/120 = 95.8% of sources. In this simulation study, SHARE correctly determined sources similar in chemical composition across monitors (100% source identification). Because mAPCA assumes the source profiles are the same across monitors, this approach frequently identified too many sources in subregions II–V (73% source identification). We also performed an array of additional simulations, whose details and corresponding results can be found in the Supplementary Material, Appendix B and Table S2. Across different simulation scenarios, SHARE was able to correctly determine sources similar in chemical compositions across monitors.

4.2. Estimating regional-level health effects

The primary aim of SHARE is to provide a hierarchical modeling approach that facilitates multisite time series studies of the short-term health effects of PM$_{2.5}$ sources. In the second part of the simulation study, we evaluated the ability of SHARE to estimate regional-level health effects of PM$_{2.5}$ sources. We also estimated regional-level health effects using mAPCA. We did not include the miscellaneous P/V source in this part of the simulation study because source apportionment studies frequently focus only on estimated sources that match reasonably well to known sources of PM$_{2.5}$ [10]. The details of the simulated hospitalizations data can be found in the Supplementary Material, Appendix B.
We considered two possible extreme cases of 25 monitors measuring PM$_{2.5}$ source types across a region. In case A, all 25 monitors measured the same source types (all monitors were in subregion I), while in case B, the 25 monitors were divided across subregions I–V, where subregions are defined as in Table 1. The assumption of mAPCA is met in case A since all monitors measured the same set of source types, but not in case B, where the source types varied by monitor. Using the simulated data, we applied both SHARE and mAPCA to estimate PM$_{2.5}$ concentrations by source type at each monitor. We fitted log-linear time series regression models (equation 2 with no covariates) to estimate associations with hospitalizations at each monitor. We estimated regional associations by pooling estimated associations from each monitor using a two-level Bayesian hierarchical model. For both SHARE and mAPCA, we pooled each source type in across all monitors. To compare differences in the estimated health effects across 100 simulated multisite datasets, we obtained the average regression coefficient and its corresponding standard error $\sqrt{W+(1+\frac{1}{100})B}$, where $W$ is the within-simulation variance and $B$ is the between-simulation variance. We used a 10% trimmed mean to compute the statistics across 100 simulated datasets. From these values, we found the average percent increase in hospitalizations for an interquartile range (IQR) increase in PM$_{2.5}$ concentration by source type and the corresponding estimated 95% confidence interval (CI). Also, we obtained the mean squared error (MSE) for the estimated health effects across simulated datasets.

Table 2 shows the average estimated regional-level health effects as the percent increase in hospitalizations associated with an IQR increase in PM$_{2.5}$ concentration by source type for SHARE and mAPCA. These estimated regional-level health effects were averaged across 100 simulated multisite datasets, with measurement error standard deviation $\sigma_e = 0.01$. The IQRs were computed as the median of monitor-specific IQRs using the simulated data and varied between simulated datasets. In case A where all monitors measured the same set of source types, estimated health effects were similar using SHARE and mAPCA for source estimation. SHARE also performed well in case B where the source types varied across monitors and the assumption of mAPCA was not met. The estimated health effects for mAPCA in case B were greatly overestimated for traffic and secondary sulfate and greatly underestimated for fireworks, salt, and metals. The results for $\sigma_e \in \{0.001, 0.1\}$ did not differ substantially from results using simulated data with $\sigma_e = 0.01$ (Supplementary Material, Tables S3–S4).

In this simulation study, we found that SHARE correctly identifies the set of PM$_{2.5}$ sources that are similar in chemical composition across monitors. Using SHARE, we can also estimate regional associations between PM$_{2.5}$ sources and adverse health outcomes.

### 5. Cardiovascular hospitalizations and PM$_{2.5}$ sources in the northeastern US

#### 5.1. PM$_{2.5}$ sources in the northeastern US

Across 85 EPA CSN monitors in our study, the number of days with complete data for PM$_{2.5}$ total mass and all 24 PM$_{2.5}$ constituents ranged from 51 days to 924 days with a median of 323 days. We first applied SHARE to determine the sources similar in chemical composition across monitors (major sources), for which we will pool community-specific...
health effects to estimate regional-level health effects. Table 3 shows the 9 major sources identified using SHARE along with those constituents most associated with each source type, which we defined as constituents with values in $A$ greater than 0.4 or less than −0.4. When possible, we named our major sources by matching them to PM$_{2.5}$ sources identified in the literature [10]. However, source names should be interpreted with caution since each identified source type may represent any PM$_{2.5}$ source that has similar contributing chemical constituents. In Table 3, we also included the number of monitors and counties where a source similar in chemical composition was identified as well as the median and IQR for the PM$_{2.5}$ concentration by source type in μg/m$^3$. Because both the IQR and average source concentrations vary by monitor, we displayed the median IQR and median average source concentration across monitors.

We applied both SHARE and mAPCA to data in the northeastern US. Using SHARE, Figure 3 shows the monitors with PM$_{2.5}$ sources similar in chemical composition, as represented by the major sources (open circles). For monitors where a source type similar in chemical composition was not found (plus signs), either this source type was not present at the monitor or we were unable to identify the source type at that monitor. Failure to identify the source type could occur when either the monitor’s profile corresponding to the source type did not explain much of the variability in the chemical constituent data at that monitor or the monitor’s profile was too noisy to match a major source. However, the aim of this study was to pool estimated health effects for sources with similar chemical composition across communities, as defined by the major sources. So estimated health effects were only pooled in counties where the estimated chemical composition of the source was similar to the major source.

We also estimated sources in the northeastern US using mAPCA. To match source types between SHARE and mAPCA, we used the Hungarian method as described in Section 2.3. Using mAPCA, we did not identify a P/V source or a traffic source, but otherwise found the other 7 major sources identified by SHARE (Table 3).

Sources of PM$_{2.5}$ may vary by season because of differences in heating use, meteorology, and other factors. To determine whether our estimated PM$_{2.5}$ sources varied by season, we applied both SHARE and mAPCA separately to our data divided into cold season days (October 1–March 31) and warm season days (April 1–September 30) (Supplementary Material, Figures S1–S2, Tables S5–S6). We found that SHARE did not identify PM$_{2.5}$ from fireworks in the cold season, which is reasonable because the US Independence Day holiday on July 4th is the day that drives most of the variability in PM$_{2.5}$ from fireworks. Additionally, SHARE did not identify a traffic source in the warm season. Previous results have found an increase in traffic PM$_{2.5}$ in the cold season [35]. While the salt source in the cold season consisted of sodium and chlorine, the closest warm season source contained primarily nitrate and sodium. By separating data by season, we were able to identify a traffic source using mAPCA in the cold season. In the cold season, mAPCA did not identify a metals source or a fireworks source. In the warm season, mAPCA did not identify a traffic source or an As/Br/Se source. There were fewer available monitors in the seasonal analysis compared with the main results because we only included monitors with more than 50 days of data in a season.

Stat Med. Author manuscript; available in PMC 2018 April 30.
5.2. Sensitivity analysis and validation substudy

To test whether the number of total monitors affected the performance of SHARE, we estimated PM$_{2.5}$ sources at 5 monitors in New York City, NY using data from (1) the 5 monitors in New York City (2) monitors in New York City, NY; Philadelphia, PA; Boston, MA; Providence, RI; Washington, DC; Baltimore, MD and (3) 41 monitors in major northeast counties. For the 3 datasets, we manually compared the source types identified using SHARE at the 5 New York City monitors and did not find substantial variation across datasets.

For 10 randomly selected monitors from our dataset, two researchers manually and independently determined which major sources were present at each monitor, an approach common in the literature [10]. The manual approach and SHARE had good agreement: of 65 sources identified across 10 monitors, they agreed for 50 sources (76.9%). Excluding poor matches where the angle between the monitor-specific source and the major source exceeded the threshold of 45 degrees, SHARE agreed with the manual approach in 50 of 55 sources (90.9%).

5.3. Associations between CVD hospitalizations and PM$_{2.5}$ sources

Our combined CVD hospitalizations and PM$_{2.5}$ constituent dataset had 85 EPA CSN monitors located within 63 counties. While most counties (n=46) had only one EPA CSN monitor, the other 17 counties had 2 monitors (n=13), 3 monitors (Hamilton, OH; Allegheny, PA; Philadelphia, PA), and 4 monitors (Cook, IL). A summary of the daily CVD hospitalizations by US county can be found in the Supplementary Material, Table S7. We applied SHARE and mAPCA to these PM$_{2.5}$ constituent concentrations to estimate PM$_{2.5}$ concentrations by source type at each monitor. For counties with more than one monitor, we averaged estimated concentrations from PM$_{2.5}$ sources across monitors for each day, as is commonly done in studies of PM and health [11, 17].

We estimated county-level associations between CVD hospitalizations and short-term exposure to PM$_{2.5}$ sources using overdispersed Poisson time series regression models (equation 2). Covariates in the model included indicators for day of week and age category (64, 65–74, ≥75). In addition, to control for confounding by weather, we included smooth functions (natural splines) of temperature and the 3-day running mean temperature, each with 6 degrees of freedom, and dew point temperature and 3-day running mean dewpoint temperature (3 degrees of freedom each). To account for long-term trends in hospitalizations, we also included a smooth function of time with 8 degrees of freedom per year. These covariates have been previously used in studies estimating health effects of PM$_{2.5}$ total mass and PM$_{2.5}$ chemical constituents [17, 34]. As in previous studies, we estimated associations between CVD hospitalizations and PM$_{2.5}$ sources for same-day exposure (lag 0), previous-day exposure (lag 1), and exposure 2 days before (lag 2) [17].

We estimated associations with CVD hospitalizations for the 6 major sources identified by SHARE that were similar in chemical composition to known sources in the northeastern US: traffic, soil, secondary sulfate, sea salt, metals, and residual oil [9, 10, 19, 26]. It is common in source apportionment analyses to focus on estimated source types that match known
sources of pollution in the area [10]. We did not estimate associations with short-term exposure to a fireworks source of PM$_{2.5}$, since this source type only has high concentrations within several days of July 4th and any estimated health effect for the source would be confounded by the US Independence Day holiday. For each of the major sources, we pooled relevant county-specific associations using a two-level Bayesian hierarchical model. We reported estimated associations as the percent increase in CVD hospitalizations associated with an IQR increase in each major source to allow comparisons across PM$_{2.5}$ source types (Table 3). The associations and 95% posterior intervals for lags 0, 1, and 2 exposure to PM$_{2.5}$ sources are shown in Figure 4. We also estimated associations separately for the warm season and cold season for same-day exposure to PM$_{2.5}$ sources (Supplementary Material, Figure S3).

Using SHARE, we found that an IQR increase in same-day exposure to PM$_{2.5}$ from traffic was associated with a 1.12% (95% posterior interval 0.22%, 2.02%) increase in CVD hospitalizations. Additionally, IQR increases in same-day PM$_{2.5}$ from metals and secondary sulfate were associated with increases in CVD hospitalizations of 0.82% (0.36%, 1.28%) and 0.74% (0.12%, 1.36%) respectively. Using mAPCA, we found evidence of associations of CVD hospitalizations with PM$_{2.5}$ secondary sulfate, salt, and residual oil at lag 0, though mAPCA did not identify a traffic source of PM$_{2.5}$. We did not find evidence that lag 1 or lag 2 exposure to PM$_{2.5}$ sources was associated with CVD hospitalizations using either SHARE or mAPCA. The seasonal results showed the largest differences in estimated health effects by season for secondary sulfate, salt, and soil (Supplementary Material, Figure S3). We did not estimate associations between PM$_{2.5}$ from metals and CVD hospitalizations in the cold season because the source was only identified in one county using SHARE.

6. Discussion

SHARE is a hierarchical modeling approach for estimating regional-level health effects of PM$_{2.5}$ sources in multisite time series studies. In our analysis of PM$_{2.5}$ source types and CVD hospitalizations in the northeastern US using SHARE, we identified positive associations between short-term exposure to PM$_{2.5}$ from traffic, secondary sulfate, and metals. Previous studies have identified combustion PM$_{2.5}$, such as PM$_{2.5}$ from traffic, to be most associated with adverse health outcomes [1]. Exposure to secondary sulfate, a regional pollutant that contributes substantially to PM$_{2.5}$ by mass [21, 10], may be well-represented by the daily Air Quality Index (AQI). However, PM$_{2.5}$ from traffic and metals sources may not be well-represented by the AQI because they are frequently spatially heterogeneous and can vary substantially within a city. Therefore, recommendations for reducing exposure based on the AQI alone may not sufficiently protect health. Other recommendations that could reduce exposure to PM$_{2.5}$ from traffic and metals sources may include not exercising near roadways or industrial sources of pollution and keeping the windows rolled up while commuting [8]. Our approach to identify the most harmful sources of pollution could also facilitate future policies aimed at reducing pollution by focusing on the most toxic sources.

In this study, we were interested in estimating regional-level health effects associated with major sources of PM$_{2.5}$ in a multisite US study. Previously, multisite studies have been used to examine confounding and effect modification across regions, including differences by air...
conditioning use [37], PM$_{2.5}$ composition [18], and oxidative potential of PM$_{2.5}$ [38]. We did not control for other pollutants that could potentially confound the association between PM$_{2.5}$ sources and CVD hospitalizations. The MCAPS study of 204 US counties did not find that CVD hospitalization effect estimates for PM$_{2.5}$ were modified by average ozone concentration [2]. A multisite study of PM$_{2.5}$ and CVD hospitalizations in Europe did not find that associations were confounded by ozone, though there was some evidence of confounding by NO$_2$ [3]. Other studies have not found much evidence of confounding of PM$_{2.5}$ by gaseous co-pollutants [1], though this has yet to be extensively examined for PM$_{2.5}$ constituents and sources. A challenge in conducting multisite studies of multiple pollutants is that ambient monitors are not always co-located and may measure pollution at different temporal scales [39, 40]. We also did not explore confounding by local sources of PM$_{2.5}$ present in each county. Exposure to local sources within a county, for example a specific factory, may be associated with adverse health outcomes; however, local sources can be better examined using county-level studies instead of multisite studies. In general, assessing confounding by other source types in multisite epidemiologic studies is an important area of future research, but this may require novel approaches to account for varying source types across sites. Because our SHARE approach facilitates multisite studies, it could be extended to investigate confounding by PM$_{2.5}$ sources that has not been previously explored.

When we applied SHARE to the data divided by season, we found that the chemical makeup of the sodium-containing source (labelled salt) differed substantially with a sodium-chlorine source present in the winter and a nitrate-sodium-bromine source in the summer (Supplementary Material, Tables S5–S6). The difference in estimated health effects of PM$_{2.5}$ from salt between SHARE and mAPCA may be driven by SHARE grouping together two different sodium sources: road salt crushed during winter months and an industrial nitrate source in the warm season. Because mAPCA estimates sources using constituent data concatenated across monitors, it may be more robust to this issue. While we matched factors identified using SHARE to known sources of PM$_{2.5}$ (e.g. traffic) using the chemical constituents in Table 3, source apportionment methods cannot definitively link latent factors to known PM$_{2.5}$ sources.

Using season-stratified data, neither SHARE nor mAPCA identified a traffic source of PM$_{2.5}$ across monitors in the warm season. Traffic PM$_{2.5}$ and its constituents are spatially heterogeneous [41], and it may be difficult to estimate the source using ambient monitoring data. In addition, we used APCA to estimate source concentrations at each monitor within the SHARE framework, which relies on PCA. Traffic PM$_{2.5}$ may explain little variation in the PM$_{2.5}$ constituent data, making it difficult to estimate with methods such as APCA. Future work could explore incorporating prior information about traffic PM$_{2.5}$ into SHARE to enable better estimation of this source.

In this study we did not account for the difference in spatial resolution between point measures of pollution and aggregated CVD hospitalizations over counties. Failing to account for this spatial misalignment may lead to estimated health effects that are biased towards the null [42, 41]. However, approaches to account for spatial misalignment have been generally focused on PM$_{2.5}$ and its constituents. More work is needed to determine how to account for
spatial misalignment in studies of PM$_{2.5}$ sources and health. Our proposed method, SHARE, also does not use the spatial correlation between monitors to determine whether two monitors measure similar sources of PM$_{2.5}$. Previous studies have demonstrated that sources of PM$_{2.5}$ vary across regions [9, 19, 20, 21] and even within a community [10], and therefore incorporating spatial correlations may not provide additional information about PM$_{2.5}$ sources.

We used APCA [27] and mAPCA [26] to estimate sources. While APCA and mAPCA are more simplistic source apportionment approaches than models such as PMF, they can be easily implemented using standard statistical software, which was necessary to perform extensive simulation studies to test the performance of SHARE. Additionally, mAPCA is an appropriate comparison to using APCA within SHARE because differences in estimated regional-level health effects between mAPCA and SHARE will likely be driven by the assumption of mAPCA that source profiles are the same across monitors. Previous studies have demonstrated that estimated health effects of PM$_{2.5}$ sources do not vary substantially between source apportionment approaches [43, 44, 21] and therefore we do not expect our estimated regional-level health effects were substantially impacted by the source apportionment method selected. Future work could examine the performance of SHARE using other source apportionment methods.

Many source apportionment models, including both mAPCA and APCA, do not yield uncertainties for estimated PM$_{2.5}$ concentrations by source type. To estimate associations between PM$_{2.5}$ sources and hospitalizations, we treated estimated concentrations from PM$_{2.5}$ sources as known in time series regression models and have likely underestimated the uncertainty of the resulting health effects. Future work could incorporate bootstrapped confidence intervals of the principal components used to estimate sources (e.g. [45]) or fully Bayesian models [19] to propagate this uncertainty.

In this work we developed SHARE, a hierarchical modeling approach for performing multisite studies of the associations between PM$_{2.5}$ sources and adverse health outcomes. Using SHARE, we found evidence that same-day exposure to PM$_{2.5}$ from traffic, secondary sulfate, and metals was associated with increased emergency CVD hospitalizations.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

The authors thank Dr. Ana Rule for her help interpreting the results and Dr. Alyssa Frazee for feedback on the manuscript.

Contract/grant sponsor: This work was supported by the National Institute on Aging [T32AG000247], the National Institute Of Environmental Health Sciences [T32ES012160, T32ES012871, R01ES019560, R21ES020152], and the U.S. Environmental Protection Agency [RD83587101]. The content is solely the responsibility 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. This publication was developed under Assistance Agreement No. RD83587101 awarded by the U.S. Environmental Protection Agency to Yale University. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors (JR Krall, AJ Hackstadt, RD Peng) and
do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication.

References


22. Bell ML, Ebisu K, Leaderer BP, Gent JF, Lee HJ, Koutrakis P, Wang Y, Dominici F, Peng RD. Associations of PM$_{2.5}$ constituents and sources with hospital admissions: Analysis of four counties in Connecticut and Massachusetts (USA) for persons ≥65 years of age Environmental Health Perspectives. 2013; 122(2):138–144. DOI: 10.1289/ehp.1306656


Figure 1.
Example of Hungarian method for estimating $\Psi_i$ corresponding to sources $i_1$–$i_4$ at monitor $i$ and major sources $M_1$–$M_5$. Figure 1A shows the matrix of angles, $\cos^{-1}(\Psi_i) \times \frac{180}{\pi}$, where $\Psi_i = A_iA_i^T$. Figure 1B shows the resulting matrix $\hat{\Psi}_i$ after applying the Hungarian method. Shaded boxes indicate sources at monitor $i$ that are similar in chemical composition to major sources. Note that source $i_2$ is not similar to any major source since all angles are greater than 45 degrees.
Figure 2.
Map of 85 PM$_{2.5}$ chemical constituent monitors from the US EPA chemical speciation network.
Figure 3.
Maps corresponding to the 9 major sources identified in the northeastern US. Each map shows the monitors where that source has similar chemical composition to the major source (open circles) and the monitors where there was not a PM$_{2.5}$ source with similar chemical composition (plus signs).
Figure 4.
Regional percent increase in CVD hospitalizations (95% posterior intervals) associated with an IQR increase in same-day (lag 0), previous-day (lag 1) and two days before (lag 2) PM$_{2.5}$ concentration for 6 major sources identified in the northeastern US. Results are shown for SHARE and mAPCA. Unlike SHARE, mAPCA does not identify a traffic source.
## Table 1

Simulated subregions with varying types of PM$_{2.5}$ sources.

<table>
<thead>
<tr>
<th>Subregion</th>
<th>Number</th>
<th>Source types</th>
<th>Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>7</td>
<td>traffic, fireworks, soil</td>
<td>secondary sulfate, salt, metals P/V</td>
</tr>
<tr>
<td>II</td>
<td>5</td>
<td>fireworks, soil</td>
<td>secondary sulfate, metals P/V</td>
</tr>
<tr>
<td>III</td>
<td>4</td>
<td>fireworks, soil</td>
<td>secondary sulfate P/V</td>
</tr>
<tr>
<td>IV</td>
<td>4</td>
<td>traffic, soil</td>
<td>secondary sulfate P/V</td>
</tr>
<tr>
<td>V</td>
<td>4</td>
<td>traffic, soil</td>
<td>secondary sulfate, metals</td>
</tr>
</tbody>
</table>
Table 2

Average regional percent increase in hospitalizations and 95% confidence interval (CI) associated with an IQR increase in PM$_{2.5}$ concentration by source type across 100 simulated multisite datasets for measurement error standard deviation $\sigma_e = 0.01$. For SHARE and mAPCA, results are shown for both case A, where all 25 monitors measure the same set of source types, and case B, where the 25 monitors measure different sets of source types as in subregions I–V. Each row also shows the true percent increase in hospitalizations ($\gamma$) and the mean squared error (MSE) corresponding to each case.

<table>
<thead>
<tr>
<th>Method</th>
<th>Source</th>
<th>$\gamma$</th>
<th>A. Estimate</th>
<th>95% CI</th>
<th>MSE</th>
<th>B. Estimate</th>
<th>95% CI</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHARE</td>
<td>Traffic</td>
<td>3.00</td>
<td>3.09</td>
<td>(2.89, 3.30)</td>
<td>0.01</td>
<td>3.04</td>
<td>(2.75, 3.33)</td>
<td>0.02</td>
</tr>
<tr>
<td>SHARE</td>
<td>Fireworks</td>
<td>1.00</td>
<td>1.10</td>
<td>(0.87, 1.33)</td>
<td>0.01</td>
<td>1.01</td>
<td>(0.69, 1.32)</td>
<td>0.01</td>
</tr>
<tr>
<td>SHARE</td>
<td>Soil</td>
<td>0.75</td>
<td>0.72</td>
<td>(0.50, 0.95)</td>
<td>0.00</td>
<td>0.75</td>
<td>(0.45, 1.04)</td>
<td>0.01</td>
</tr>
<tr>
<td>SHARE</td>
<td>Sec sulf</td>
<td>0.50</td>
<td>0.53</td>
<td>(0.47, 0.59)</td>
<td>0.00</td>
<td>0.54</td>
<td>(0.19, 0.90)</td>
<td>0.00</td>
</tr>
<tr>
<td>SHARE</td>
<td>Salt</td>
<td>1.00</td>
<td>1.12</td>
<td>(0.80, 1.44)</td>
<td>0.02</td>
<td>1.12</td>
<td>(0.22, 2.02)</td>
<td>0.03</td>
</tr>
<tr>
<td>SHARE</td>
<td>Metals</td>
<td>1.00</td>
<td>0.93</td>
<td>(0.74, 1.13)</td>
<td>0.01</td>
<td>1.00</td>
<td>(0.88, 1.13)</td>
<td>0.01</td>
</tr>
<tr>
<td>mAPCA</td>
<td>Traffic</td>
<td>3.00</td>
<td>3.12</td>
<td>(2.89, 3.34)</td>
<td>0.01</td>
<td>4.46</td>
<td>(1.59, 7.41)</td>
<td>2.50</td>
</tr>
<tr>
<td>mAPCA</td>
<td>Fireworks</td>
<td>1.00</td>
<td>1.10</td>
<td>(0.86, 1.35)</td>
<td>0.01</td>
<td>-2.76</td>
<td>(-4.47, -1.01)</td>
<td>14.33</td>
</tr>
<tr>
<td>mAPCA</td>
<td>Soil</td>
<td>0.75</td>
<td>0.71</td>
<td>(0.45, 0.97)</td>
<td>0.01</td>
<td>1.05</td>
<td>(0.60, 1.51)</td>
<td>0.11</td>
</tr>
<tr>
<td>mAPCA</td>
<td>Sec sulf</td>
<td>0.50</td>
<td>0.55</td>
<td>(0.48, 0.62)</td>
<td>0.02</td>
<td>4.56</td>
<td>(1.38, 7.85)</td>
<td>17.76</td>
</tr>
<tr>
<td>mAPCA</td>
<td>Salt</td>
<td>1.00</td>
<td>1.14</td>
<td>(0.76, 1.53)</td>
<td>0.05</td>
<td>-1.85</td>
<td>(-2.91, -0.77)</td>
<td>8.24</td>
</tr>
<tr>
<td>mAPCA</td>
<td>Metals</td>
<td>1.00</td>
<td>0.92</td>
<td>(0.72, 1.14)</td>
<td>0.00</td>
<td>-4.33</td>
<td>(-7.37, -1.19)</td>
<td>28.71</td>
</tr>
</tbody>
</table>
Table 3

Major PM$_{2.5}$ sources in the northeastern US identified using SHARE with the number of monitors (out of 85) where the chemical composition of the estimated PM$_{2.5}$ source was similar to the major source, the number of counties (out of 63) where those monitors were located, and the constituents most associated with each major source. Also shown is the median PM$_{2.5}$ concentration across monitors and the difference in PM$_{2.5}$ concentration for each source type between the third quartile and the first quartile, labeled as the interquartile range (IQR), in $\mu$g/m$^3$.

<table>
<thead>
<tr>
<th>Sources</th>
<th>Monitors</th>
<th>Counties</th>
<th>Contributing constituents</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metals</td>
<td>42</td>
<td>36</td>
<td>Lead, Zinc, Manganese</td>
<td>1.02</td>
<td>1.52</td>
</tr>
<tr>
<td>Soil</td>
<td>79</td>
<td>60</td>
<td>Silicon, Aluminum, Titanium, Calcium, Iron</td>
<td>1.02</td>
<td>1.31</td>
</tr>
<tr>
<td>Sec. Sulfate</td>
<td>74</td>
<td>58</td>
<td>Ammonium, Sulfate, OC, Selenium</td>
<td>6.74</td>
<td>7.59</td>
</tr>
<tr>
<td>Fireworks</td>
<td>70</td>
<td>53</td>
<td>Strontium, Potassium, Copper</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>Salt</td>
<td>49</td>
<td>41</td>
<td>Chlorine, Sodium ion, Nitrate, Bromine</td>
<td>0.32</td>
<td>0.96</td>
</tr>
<tr>
<td>P/V</td>
<td>37</td>
<td>31</td>
<td>Phosphorus, Vanadium</td>
<td>0.02</td>
<td>0.37</td>
</tr>
<tr>
<td>Residual oil</td>
<td>37</td>
<td>34</td>
<td>Nickel, Iron, Vanadium, Manganese</td>
<td>0.09</td>
<td>0.26</td>
</tr>
<tr>
<td>As/Se/Br</td>
<td>11</td>
<td>11</td>
<td>Arsenic, Selenium, Bromine</td>
<td>0.44</td>
<td>1.46</td>
</tr>
<tr>
<td>Traffic</td>
<td>29</td>
<td>24</td>
<td>EC, OC, Iron</td>
<td>3.58</td>
<td>2.87</td>
</tr>
</tbody>
</table>