Published in final edited form as:

Environ Ecol Stat. 2015 December 1; 22(4): 779–800. doi:10.1007/s10651-015-0319-6.

A model-based approach for imputing censored data in source apportionment studies

Jenna R. Krall,

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205, Tel.: 410-502-5870, Fax: 410-955-0958

Charles H. Simpson, and

Havoc Engineering, 24 N. Wolfe St., Baltimore, MD 21231, Tel.: 443-474-6549, Fax: 410-955-0958

Roger D. Peng

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205, Tel.: 410-955-2468, Fax: 410-955-0958

Jenna R. Krall: jenna.krall@gmail.com; Charles H. Simpson: csimpson@havocengineering.com; Roger D. Peng: rpeng1@jhu.edu

Abstract

Sources of particulate matter (PM) air pollution are generally inferred from PM chemical constituent concentrations using source apportionment models. Concentrations of PM constituents are often censored below minimum detection limits (MDL) and most source apportionment models cannot handle these censored data. Frequently, censored data are first substituted by a constant proportion of the MDL or are removed to create a truncated dataset before sources are estimated. When estimating the complete data distribution, these commonly applied methods to adjust censored data perform poorly compared with model-based imputation methods. Model-based imputation has not been used in source apportionment and may lead to better source estimation. However if the censored chemical constituents are not important for estimating sources, censoring adjustment methods may have little impact on source estimation. We focus on two source apportionment models applied in the literature and provide a comprehensive assessment of how censoring adjustment methods, including model-based imputation, impact source estimation. A review of censoring adjustment methods critically informs how censored data should be handled in these source apportionment models. In a simulation study, we demonstrated that model-based multiple imputation frequently leads to better source estimation compared with commonly used censoring adjustment methods. We estimated sources of PM in New York City and found estimated source distributions differed by censoring adjustment method. In this study, we provide guidance for adjusting censored PM constituent data in common source apportionment models, which is necessary for estimation of PM sources and their subsequent health effects.

Keywords

Chemical speciation; Factor analysis; Imputation; Censored data; Particulate matter

1 Introduction

Particulate matter (PM) air pollution is a complex chemical mixture (Bell *et al.*, 2007) that originates from many different sources (Maykut *et al.*, 2003; Ito *et al.*, 2004; Hopke *et al.*, 2006). Previous studies have found PM collected from ambient monitors with a cut point of $2.5\text{ }\mu\text{m}$ in aerodynamic diameter, $\text{PM}_{2.5}$, to be associated with increased mortality and morbidity (Zanobetti and Schwartz, 2009; Peng *et al.*, 2009; Environmental Protection Agency, 2009). $\text{PM}_{2.5}$ emitted from different sources likely varies in toxicity because $\text{PM}_{2.5}$ is a mixture of different chemical constituents that vary in toxicity (Ostro *et al.*, 2007; Peng *et al.*, 2009; Zhou *et al.*, 2011; Krall *et al.*, 2013). In epidemiologic studies, determining which sources of $\text{PM}_{2.5}$ are most harmful to human health generally requires daily concentrations of $\text{PM}_{2.5}$ emitted from different sources. Typically, $\text{PM}_{2.5}$ sources are not directly measured and must be inferred from daily $\text{PM}_{2.5}$ chemical constituent concentrations using source apportionment models. However, $\text{PM}_{2.5}$ constituent concentrations are frequently censored below the 99% confidence limit for not measuring a false positive (Environmental Protection Agency, 2015), referred to as the minimum detection limit (MDL) (Polissar *et al.*, 2001; Kim *et al.*, 2003; Maykut *et al.*, 2003). Since many common source apportionment models cannot handle censored data, censored $\text{PM}_{2.5}$ constituent data must be imputed or removed before sources can be estimated.

Previous studies have compared different methods to impute or remove censored environmental data, however the best method for adjusting censored data depends on what quantity is being estimated from the data (Helsel, 2010; Ganser and Hewett, 2010). Substituting censored concentrations with a constant between 0 and the MDL (e.g. $\frac{1}{2} \times \text{MDL}$) will lead to biased estimates of the mean and standard deviation (Helsel, 2006). In contrast, for dimension reduction or latent variable procedures, such as principal component analysis (PCA), substitution methods for censored data may perform satisfactorily. As an example, Farnham *et al.* (2002) applied PCA to groundwater chemicals and found that substituting censored values with $\frac{1}{2} \times \text{MDL}$ yielded principal component scores and loadings close to those obtained under no censoring. Aruga (1997) found that substituting censored values with a constant near zero led to acceptable results in factor analysis, where acceptable was defined by the number of factors obtained, the variables associated with each factor, and the variance explained by each factor. Model-based methods for imputing censored data are generally preferred to substitution methods (Helsel, 2010; Ganser and Hewett, 2010) and have been used to estimate the distributions of multivariate censored data (Hopke *et al.*, 2001; Chen *et al.*, 2013; Francis *et al.*, 2009). However, if the assumed distribution is incorrect, a model-based approach may produce biased estimates of the data distribution (Helsel, 2010).

Source apportionment methods are latent variable models designed specifically for environmental applications. Most source apportionment models assume $\text{PM}_{2.5}$ chemical

constituent concentrations are related to PM_{2.5} source concentrations using methods similar to factor analysis, which estimates latent variables (e.g. PM_{2.5} sources) from the observed data. Source apportionment models differ from traditional factor analysis because they estimate two non-negative matrices: the chemical contributions to each source, referred to as the source profiles, and the daily source concentrations. Source apportionment methods vary in their approaches to source estimation and common methods include Absolute Principal Component Analysis (APCA) (Thurston and Spengler, 1985), Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994; Norris *et al.*, 2008), and Unmix (Henry, 1997; Norris *et al.*, 2007). Some guidance exists on how to adjust censored data in source apportionment studies (Paatero and Hopke, 2003; Larson *et al.*, 2004), however no studies have comprehensively examined how different adjustment methods impact source estimation. Often in source apportionment, censored data are first imputed or removed to create a truncated dataset and then sources are estimated. Because source apportionment methods vary, the best method for adjusting censored data may vary by source apportionment model. It is not currently known how different censoring adjustment methods impact commonly applied source apportionment models, such as APCA and PMF, and this information is critical to guide future studies that use these models.

In source apportionment studies, censored data are often substituted with a constant between 0 and the MDL for each constituent (commonly $\frac{1}{2} \times \text{MDL}$) (Larson *et al.*, 2004; Marmur *et al.*, 2005; Lee *et al.*, 2008) or constituents with many censored observations are excluded or downweighted in the analysis (Kavouras *et al.*, 2001; Querol *et al.*, 2001; McDonald *et al.*, 2003; Paatero and Hopke, 2003; Song *et al.*, 2007). While model-based methods to impute censored data are preferred to substitution or exclusion methods for estimating the distribution of the data (Helsel, 2010), model-based imputation approaches have not been evaluated in source apportionment studies. Model-based imputation may not lead to better source estimation compared with other censoring adjustment methods because the assumed data distribution for the model-based method may be incorrect and censored constituents may not be important for estimating sources. If the censored constituent is necessary to distinguish similar sources, applying substitution or exclusion censoring adjustment methods may limit our ability to correctly resolve PM_{2.5} sources.

For example, Figure 1 shows a time series of concentrations in New York City for two PM_{2.5} constituents: aluminum, which has many concentrations that fall below the MDL, and calcium, which is completely observed. These constituents both contribute to a soil source of PM_{2.5} in New York City (Ito *et al.*, 2004) and estimation of this source depends on the censoring adjustment method applied. Substituting censored aluminum concentrations with a constant will lead to underestimation of the correlation between aluminum and calcium, which may limit our ability to estimate PM_{2.5} from soil. Excluding aluminum from the analysis will lead to less available information about the soil source. However, if aluminum does not contribute to sources in New York City, the method chosen to adjust censored data may not impact source estimation.

This work offers three contributions. We first conducted a simulation study to provide a comprehensive analysis of the impact of different censoring adjustment methods on source estimation using APCA and PMF. We examined two censoring adjustment methods that are

commonly applied in source apportionment: substituting censored concentrations using $\frac{1}{2} \times \text{MDL}$ and excluding or downweighting constituents with a large proportion of censored data. In addition, we developed a model-based approach for imputing censored constituent concentrations. We demonstrated through simulation how each commonly-applied censoring adjustment method (substituting, excluding/downweighting) and our model-based imputation method impact source estimation. In the second part of this study, we estimated PM_{2.5} sources in New York City and demonstrated how source estimation varies by censoring adjustment method. We also showed that the impact of censored data on source estimation depends on whether the censored constituent is critical for estimating PM_{2.5} sources. Last, we have made software publicly available for imputing censored PM_{2.5} constituent time series data using our model-based approach in the handles R package (<https://github.com/kralljr/handles>). This work fills a critical gap in the source apportionment literature by examining the impact of different censoring adjustment methods on APCA and PMF and demonstrating when model-based imputation methods improve source estimation compared with simpler approaches.

2 Methods

2.1 Source apportionment methods

We examined the impact of censoring adjustment methods on two common source apportionment models, APCA and PMF. Let $\mathbf{X}_{[n \times P]} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_P)$ be the observed PM_{2.5} chemical constituent data, where $\mathbf{x}_p$ is the vector of n daily concentrations for constituent p ($p = 1, \dots, P$). Both APCA and PMF estimate two quantities from $\mathbf{X}$: the source concentration matrix $\mathbf{F}$ and the source profile matrix $\mathbf{A}$. The source concentration matrix $\mathbf{F}_{[n \times L]} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_L)$, where $\mathbf{f}_l$ is n daily concentrations for source l . The source profile matrix $(\mathbf{A}_{[L \times P]})^T = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \dots, \boldsymbol{\lambda}_L)$, where $\boldsymbol{\lambda}_l$ is the profile for source l that gives the contributions of P chemical constituents to source l . We assume the number of sources L is known.

2.1.1 APCA—For each day t ($t = 1, \dots, n$) and constituent p , let $z_{tp} = \frac{x_{tp} - \bar{x}_p}{s_p}$ where $\bar{x}_p$ is the sample mean and s_p is the sample standard deviation for constituent p . Then, the vectors of length n that make up $\mathbf{Z}_{[n \times P]} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_P)$ are all mean zero with unit variance. APCA first finds absolute principal component scores, $\mathbf{A}$, by rotating and rescaling results from principal component analysis (PCA):

1. Using PCA, obtain the matrix $\tilde{\mathbf{V}}$ of the first L eigenvectors of $\mathbf{Z}^T \mathbf{Z}$ where the norm of each column of $\tilde{\mathbf{V}}$ is equal to the square root of the corresponding eigenvalue.
2. Find $\mathbf{V} = \tilde{\mathbf{V}}\mathbf{R}$, where $\mathbf{R}$ is the $L \times L$ varimax rotation matrix (Harris and Kaiser, 1964)

Define $\mathbf{A}_{[n \times L]} = (\mathbf{X}\mathbf{S}^{-1})[\text{Cor}(\mathbf{X})^{-1}\mathbf{V}]$, where $\mathbf{S} = \text{diag}(s_1, s_2, \dots, s_P)$ is a diagonal matrix of the sample standard deviations of the columns of $\mathbf{X}$. The absolute principal component scores, $\mathbf{A}$, can be roughly interpreted as the scaled but uncentered data rotated into the space spanned by $\mathbf{V}$. In the next step, total mass PM_{2.5} is regressed on $\mathbf{A}$. The estimated coefficients, $\hat{\boldsymbol{\eta}}$, are then used to estimate source concentrations $f_{il} = a_{il} \times \hat{\eta}_l$. To find the

source profiles, we regress daily concentrations for constituent p on the estimated source

concentrations $\mathbf{F}$, $E[x_{tp}|\mathbf{F}] = \nu_p + \sum_{l=1}^L f_{tl} \times \lambda_{lp}$ to obtain the source profile matrix $\mathbf{A}$. We implemented APCA using R version 3.0.2 (R Core Team, 2012).

2.1.2 Positive Matrix Factorization—Positive Matrix Factorization (PMF) finds $\mathbf{A}$ and $\mathbf{F}$ that minimize

$$\sum_{p=1}^P \sum_{t=1}^T \left(\frac{x_{tp} - \sum_{l=1}^L f_{tl} \times \lambda_{lp}}{u_{tp}} \right)^2 \quad (1)$$

subject to $\lambda_{lp} \geq 0$ and $f_{tl} \geq 0$ for all l, p, t . Uncertainties u_{tp} are selected to reflect the relative certainty about each x_{tp} . Since APCA assumes all uncertainties are equal, we set $u_{tp} = 1$ for all t, p in PMF to make results more comparable between source apportionment methods. The multilinear engine (ME) is a program that finds $\mathbf{A}$ and $\mathbf{F}$ by minimizing the objective function for PMF (equation 1) using a conjugate gradient algorithm (Paatero, 1999). We used the ME version 2 (ME-2) software released with the user interface program PMF version 3.0, which is distributed by the US EPA.

2.2 Adjusting censored data below the MDL

We compared three methods for adjusting censored PM_{2.5} constituent concentrations:

1. Substitute censored concentrations with $\frac{1}{2} \times \text{MDL}$
2. Exclude or downweight constituents with many censored concentrations
3. Impute censored concentrations using our proposed model-based approach

We chose the constant $\frac{1}{2} \times \text{MDL}$ for the substitution method because it has been shown to yield better PCA results than substituting 0 or the MDL (Farnham *et al.*, 2002). For APCA, method (2) is implemented by first excluding constituents with more than 25% daily concentrations below the MDL from the analysis and then substituting remaining censored concentrations with $\frac{1}{2} \times \text{MDL}$. For implementing method (2) for PMF, we substituted censored data with $\frac{1}{2} \times \text{MDL}$ and then estimated the signal-to-noise ratio (SNR) for each constituent, as defined by (Paatero and Hopke, 2003). If $0.2 < \text{SNR} < 2$, all uncertainties for the constituent are increased three-fold, and if $\text{SNR} \leq 0.2$, the constituent is excluded from the analysis. Because the computation of the SNR by (Paatero and Hopke, 2003) uses the number of censored concentrations, this approach effectively removes constituents with many censored concentrations and downweights constituents with a moderate number of censored concentrations. Both the $\frac{1}{2} \times \text{MDL}$ approach (Song *et al.*, 2001; Larson *et al.*, 2004) and the exclude and/or downweight method (Rizzo and Scheff, 2007; Song *et al.*, 2007) have been frequently applied in the source apportionment literature.

2.2.1 Model-based approach for imputing censored data—We developed a model-based multiple imputation approach that uses a parametric model to multiply impute our censored data (Hopke *et al.*, 2001; Little and Rubin, 2002). Model-based imputation is known to perform well for missing data (Mage *et al.*, 1999; Dempster *et al.*, 1977) and these

approaches also perform well in the case of censored data if the chosen distribution fits the data well (Helsel, 2010). Additionally, multiple imputation can be used to assess the uncertainty in source estimation driven by censoring. For each day t , we assumed $\log(\mathbf{x}_t) \sim \text{MVN}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where $\mathbf{x}_t$ is the vector of P constituent concentrations on day t . Similar multivariate normal distributions have been previously used to model environmental data (Hopke *et al.*, 2001; Chen *et al.*, 2013). We assumed the data are independent across time, which is an assumption of most source apportionment models and is likely reasonable in this application since $\text{PM}_{2.5}$ constituent monitors generally only sample $\text{PM}_{2.5}$ every sixth day.

Since censored data made using standard maximum likelihood estimators difficult, we used a Markov Chain Monte Carlo approach to sample from the posterior distributions of $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, and the censored data. We used conjugate priors $\boldsymbol{\theta} \sim \text{MVN}(\mathbf{0}, 10^5 \mathbf{I})$ and $\boldsymbol{\Sigma} \sim \text{inv-Wishart}(P + 1, \mathbf{I})$, where $\mathbf{I}$ is the $P \times P$ identity matrix. This approach is similar to the approach of Hopke *et al.* (2001), who used a flat prior for $\boldsymbol{\theta}$ instead of a normal prior. Our choice of conjugate priors allowed us to perform extensive simulations to determine the impact of this model on source estimation. We directly sampled from the full conditional distributions of $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, and the censored constituent concentrations using Gibbs sampling. Letting $\mathbf{Y} = \log(\mathbf{X})$, the full conditionals for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are

$$(\boldsymbol{\theta} | \boldsymbol{\Sigma}, \mathbf{Y}) \sim \text{MVN} \left((10^{-5} \mathbf{I} + n \boldsymbol{\Sigma}^{-1})^{-1} (n \boldsymbol{\Sigma}^{-1} \bar{\mathbf{y}}), (10^{-5} \mathbf{I} + n \boldsymbol{\Sigma}^{-1})^{-1} \right) \quad (2)$$

$$(\boldsymbol{\Sigma} | \boldsymbol{\theta}, \mathbf{Y}) \sim \text{inv-Wishart} \left(n + P + 1, \mathbf{I} + \sum_{t=1}^n (\mathbf{y}_t - \boldsymbol{\theta})(\mathbf{y}_t - \boldsymbol{\theta})^T \right) \quad (3)$$

where $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_P)^T$. For each day t , let y_{tp} be the logged concentration for a censored constituent p and $\mathbf{y}_{tq}$ be the logged concentrations for the remaining q constituents. The full conditional distribution of y_{tp} is truncated normal,

$$(y_{tp} | \mathbf{y}_{tq}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) \sim \text{trunc-N} \left(\theta_p + \sum_{pq} \sum_q^{-1} (\mathbf{y}_{tq} - \boldsymbol{\theta}_q), \boldsymbol{\Sigma}_p - \sum_{pq} \sum_q^{-1} \sum_{pq}^T \right) \quad (4)$$

where y_{tp} is truncated above by the log of its MDL, $\boldsymbol{\Sigma}_{pq}$ is the covariance between constituent p and the remaining constituents q , and θ_p , $\boldsymbol{\theta}_q$, $\boldsymbol{\Sigma}_p$, and $\boldsymbol{\Sigma}_q$ refer to the subsets of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ corresponding to constituents p and q .

We drew 50,000 samples from the joint distribution of $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, and the censored data by iteratively sampling from the three distributions (equations (2), (3), and (4) for each censored y_{tp}) and updating the values for $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$, and each censored y_{tp} . To create our imputed datasets, we drew 100 logged constituent datasets from the last 25,000 samples and then exponentiated each to obtain constituent concentrations on the original scale.

3 Simulation study

We performed a simulation study to compare how different censoring adjustment methods affect source estimation using APCA and PMF. This simulation study quantifies the impact of censored data on APCA and PMF and demonstrates the practical benefits and disadvantages of selecting specific censoring adjustment methods. We conducted a simulation study to compare three methods for adjusting censored data: (1) $\frac{1}{2} \times \text{MDL}$ (2) excluding and/or downweighting constituents and (3) our model-based approach.

3.1 SPECIATE database for source profiles

In our simulation study, we incorporated data from the US EPA SPECIATE database (version 4.2), which contains $\text{PM}_{2.5}$ source profiles collected throughout the US for 53 chemical constituents. We cleaned SPECIATE such that (1) there were $P = 23$ $\text{PM}_{2.5}$ chemical constituents (Table S1, Online Resource), (2) the source profiles were normalized to represent the percent contribution of each chemical constituent to a source and (3) the sources fell into one of 7 major source categories in the US: wood burning (wood), diesel motor vehicles (DMV), soil dust (dust), gasoline motor vehicles (GMV), coal combustion (coal), oil combustion, and metals production (details in the Online Resource, Section 1). While the 23 constituents are a fraction of the over 50 constituents that make up $\text{PM}_{2.5}$ (Bell *et al.*, 2007), generally source apportionment focuses on a subset of 20–30 constituents, which include constituents that contribute most to $\text{PM}_{2.5}$ by mass (e.g., sulfate, nitrate, organic carbon) (Bell *et al.*, 2007), smaller constituents previously identified as toxic (e.g., arsenic, nickel, vanadium, selenium) (Ito *et al.*, 2004; Franklin *et al.*, 2008; Bell *et al.*, 2009; Zanobetti *et al.*, 2009) and constituents associated with common sources (e.g. PM from soil frequently contains calcium, aluminum, and titanium) (Ito *et al.*, 2004; Nikolov *et al.*, 2007).

3.2 Simulating $\text{PM}_{2.5}$ constituent data

To simulate $\text{PM}_{2.5}$ chemical constituent data $\mathbf{X}$, we used the Schur (element-wise) product of lognormal errors $\mathbf{e}$ with the product of a source concentration matrix, $\mathbf{F}$, and a source profile matrix, $\mathbf{A}$,

$$\mathbf{X}_{[n \times P]} = (\mathbf{F}_{[n \times L]} \mathbf{A}_{[L \times P]}) \circ \mathbf{e}_{[n \times P]} \quad (5)$$

where $\log(e_{tp}) \stackrel{\text{IID}}{\sim} N(0, 0.01^2)$. We used lognormally distributed errors to ensure constituent concentrations were non-negative. For the rows of $\mathbf{A}_{[L \times P]}$, we selected observed source profiles from the cleaned SPECIATE database for wood, DMV, dust, GMV, and coal to represent two hypothetical communities with different numbers of sources: $L = 3$ sources (wood/ DMV/ dust) and $L = 5$ sources (wood/ DMV/ dust/ GMV/ coal) (Online Resource, Section 2, Table S2). To generate the concentration time series $\mathbf{f}_l$ for each source l , we generated $n = 1000$ independent lognormal concentrations with means and standard deviations chosen to approximately reflect the distribution of sources reported in the literature (Table S3, Online Resource) (Ito *et al.*, 2004; Lingwall *et al.*, 2008).

For both the 3-source and 5-source scenarios, we simulated 300 $\mathbf{F}$ datasets and created 300 datasets $\mathbf{X}$. We assumed that the sources' chemical compositions were fixed and so we used the same matrix of profiles $\mathbf{A}$ for each simulated dataset. To introduce different degrees of censoring, we first randomly selected 2 or 11 of the 23 constituents to be censored. Then we created five censored datasets for each $\mathbf{X}$ by censoring these randomly selected constituents at their 20%, 50%, or 80% quantiles from the observed constituent concentration distributions. We did not censor 11 constituents at 80% because nearly 40% of the total data would be censored and source apportionment is not practical in this setting. These censoring scenarios yielded between 1.7% and 23.9% censored data across all constituents. We adjusted each censored dataset using the three censoring adjustment methods. When constituents were censored at 20%, the exclude method used in APCA does not drop any constituents and therefore does not differ from the $\frac{1}{2} \times \text{MDL}$ method. The different simulation scenarios are shown in Table 1.

3.3 Comparing source apportionment results between censoring adjustment methods

To analyze the results from our simulation study, we compared sources estimated using the uncensored, simulated data $\mathbf{X}$ with sources estimated from censored and then adjusted data $\mathbf{X}$. While we could have focused on the differences between censored-then-adjusted source estimates and the true simulated source concentrations, this approach would conflate measurement error from the source apportionment methods, measurement error in the chemical constituent data, and censoring. By comparing results from source apportionment models, we focused our attention on the source estimation error driven by censored data. The measures used in this section can help determine best practices for handling censored data in APCA and PMF.

We computed (a) the number of incorrectly identified sources under censoring, (b) the source concentration bias due to censoring, and (c) the ratio of source concentration sample variances between data with and without censoring. Because the distributions of these measures were skewed, we computed the median value across 300 simulated datasets. These measures reflect information about sources that is commonly reported in the source apportionment literature (Ito *et al.*, 2004; Hopke *et al.*, 2006) and capture both the identification of the source (a) and distributional properties of the source concentrations (b and c).

Sources are frequently identified by linking constituents that have large values in estimated source profiles to expert knowledge about the chemical makeup of different sources (Ito *et al.*, 2004; Hopke *et al.*, 2006). However in a large simulation study, individually inspecting each profile would be impractical. To identify sources, we applied k-nearest neighbors classification using our cleaned SPECIATE dataset. Details about the classification method can be found in the Online Resource (Section 3). Source mis-classification under censoring acts as a measure of misattribution driven by censoring (e.g. naming PM from GMVs as PM from DMV), which is important since a major aim of source apportionment is to identify sources of PM_{2.5} in a community. Applying our classification method to each of the 300 uncensored datasets led to the correct identification of all 3 or all 5 sources. To compute (a),

we computed the median number of incorrectly identified sources across simulated, censored datasets.

We used the classification method to identify sources obtained from the censored data, which allowed us to compare source concentration means and variances between data with and without censoring. Although all 3 or all 5 sources were always correctly classified for the uncensored data, there were cases where the sources classified from the censored data did not match the uncensored sources. In these cases, we were unable to compare source means and variances between data with and without censoring. Therefore, the comparison of means and variances were only made for datasets whose sources were both correctly classified for both the uncensored and censored/adjusted data. As a measure of source concentration bias due to censoring, we compared the median differences in source concentration means for source l , d_l , across simulated datasets between data with and

without censoring by reporting $\left(\sqrt{\frac{1}{L} \sum_{l=1}^L d_l^2} \right)$ across sources. For an estimated source l and simulated dataset i , let $\tilde{s}_{il}^2$ be the sample temporal source concentration variance from censored data and s_{il}^2 be the corresponding quantity for the uncensored data. To compare sample source concentration variances for each source and each amount of censoring, we computed the median of $\frac{\tilde{s}_{il}^2}{s_{il}^2}$ across 300 simulated datasets. By computing the mean differences and variance ratios, we can both determine how the differences change with increasing censoring and how these differences vary by censoring adjustment method.

For our model-based method, we computed each quantity (a–c) for each of the 100 imputed datasets and took the 20% trimmed mean. The trimmed mean better reflected the average quantity (a–c) than the untrimmed mean because we sometimes observed quantities (a–c) from one imputed dataset that were far from the other imputed datasets.

3.4 Results for source estimation

Across all measures and censoring adjustment methods, the impact of censoring on source estimation increased with the amount of censored data. We obtained the average number of sources misclassified of 3 or 5 total sources across simulations using our k-nearest neighbors classification method. For APCA, while both the $\frac{1}{2} \times \text{MDL}$ method and the model-based method never misclassified sources, when constituents were excluded from the analysis, sources were frequently misclassified (Table 2). For example in the 3-source scenario using the exclude method, 1 of the 3 sources was incorrectly classified on average when more than 20% of the data were censored. DMV was often misclassified, though wood and dust were sometimes misclassified as well. Under wood/ DMV/ dust/ GMV/ coal, the most frequent misclassifications were wood, DMV, and GMV, while dust was hardly ever misclassified.

Table 3 shows the median source concentration bias due to censoring, d_l , collapsed across

sources as $\left(\sqrt{\frac{1}{L} \sum_{l=1}^L d_l^2} \right)$ for APCA. Compared to other censoring adjustment methods, the model-based method generally led to less source concentration bias due to censoring. When there were 3 sources, excluding constituents that had more than 25% censored daily

concentrations seemed to decrease concentration bias due to censoring compared with the $\frac{1}{2} \times \text{MDL}$ approach. However, excluding constituents did not perform uniformly better than $\frac{1}{2} \times \text{MDL}$ when there were 5 sources.

The estimated variance ratios between sources estimated with and without censored data for APCA are shown in Tables 4 and 5. Across censoring adjustment methods, some sources had overestimated variances under censoring and other sources had underestimated variances. For example, the variance of DMV was severely overestimated and the variance of GMV was underestimated when many data were censored in the 5-source scenario (Table 5). Substituting censored data with $\frac{1}{2} \times \text{MDL}$ led to good estimates of the source variances when few data were censored. The exclude method generally performed worse than imputing data with $\frac{1}{2} \times \text{MDL}$. Compared with other censoring adjustment methods, the model-based approach led to sample temporal source variances closer to those estimated from uncensored data for APCA.

Results from PMF are included in Tables S4–S7 (Online Resource). Because PMF is more time-consuming than APCA, we drew 10 imputed datasets for the model-based method and took the median across datasets. In general censored data made source classification difficult under PMF across censoring adjustment methods, but PMF frequently led to estimates of the source means that were less affected by censored data compared with APCA. When few data are censored under PMF, the model-based method provided less or equally biased estimates of source means and variances compared with other commonly applied censoring adjustment methods. When many data were censored, the relative performance of censoring adjustment methods for PMF was less consistent than when APCA was used for source apportionment.

3.5 Sensitivity analysis

We found the performances of censoring adjustment methods were consistent across an array of sensitivity analyses. For APCA, the model-based method still performed best under a different combination of sources (Tables S8–S11, Online Resource). Our results were robust to the choice of source concentration means and standard deviations and the choice of observed source profiles from SPECIATE. As a sensitivity analysis, we also generated the source concentration data with an arbitrary covariance between the sources, but found results were similar to assuming sources were independent.

The simulated constituent data were not lognormally distributed, since the data were generated using the Schur product of lognormal errors and a linear combination of the lognormally distributed sources and fixed source profiles (equation 5). If the simulated constituent data were distributed as lognormal by design, our model-based approach may have overperformed because it assumes a lognormal distribution. We tested whether the model-based method was overperforming because of the lognormally distributed errors by generating $\mathbf{X}$ as in equation 5 using Gamma distributed errors ϵ (shape=rate=25). We found results using Gamma distributed errors, in particular the results from our model-based approach, were unchanged.

4 PM_{2.5} sources in New York City

4.1 Data

To estimate PM_{2.5} sources in New York City, we used data from the New York State Department of Environmental Conservation (<http://bit.ly/198xZcm> accessed 26 Nov 2012) and the Environmental Protection Agency Chemical Speciation Network for the Queens College monitor. From these data, we created a dataset of 174 days from April 2001–December 2002 to match a previous source apportionment study (Ito *et al.*, 2004). In general, PM_{2.5} chemical constituent concentrations were measured every third day. Our dataset consisted of all constituents from the simulation study (Table S1, Online Resource) except sulfate and including barium, cadmium, chromium, cobalt, magnesium, molybdenum, sulfur, and ammonium ion, for a total of 30 constituents (Ito *et al.*, 2004). We matched our source apportionment results to 4 PM_{2.5} sources in New York City by using the several constituents that contribute most to each source, as reported by Ito *et al.* (2004). The four sources (and associated constituents) were: soil (aluminum, titanium, calcium, iron, silicon), secondary sulfate (sulfur, ammonium ion), traffic (organic carbon, elemental carbon, potassium), and residual oil (chlorine, lead, vanadium, nickel, nitrate, zinc).

In this application, we used a constant MDL over time for each constituent as in the simulation study. Specifically, we used the maximum MDL over time for each constituent to ensure that all values recorded below the MDL were treated as censored in our analysis. If we used other constants for the MDL (e.g. the mean MDL), we would be treating some concentrations below the MDL as observed. When the MDL varies over time, it is not appropriate to apply substitution-based censoring adjustment methods (e.g. 1/2 MDL) since the method can introduce false temporal variability in the data (Helsel, 2010). Future work could investigate how temporally varying MDLs impact source estimation, though both our simulation study and analysis use a constant MDL for each constituent to isolate the effects of censored constituents from the effects of time-varying MDLs.

For the model-based approach, we used 100 draws from the posterior distribution (equation 4) and used the 10% trimmed mean across imputations to create the concentration time series for each source. Sometimes the methods used to obtain PM_{2.5} chemical constituent concentrations from ambient monitors do not yield censored data, but instead report daily concentrations below the MDL. These reported data below the MDL are frequently treated as censored and discarded because they may have no relationship to the true, unobserved concentrations (Helsel, 2005a). In this data analysis, we compared source apportionment results between the three censoring adjustment methods used in the simulation study and results using the data “reported below the MDL.”

4.2 Results

In our New York City dataset, 15 of the constituents had less than 25% censored concentrations, 2 constituents had 25%–50% censored concentrations, and 13 constituents had more than 50% concentrations below the MDL. We found few censored concentrations for constituents contributing to secondary sulfate (1% censored concentrations for ammonium ion) and traffic (6% for elemental carbon). However, there were more censored

data for constituents associated with soil (60% for aluminum and 6% for titanium) and residual oil (65% for chlorine, 29% for lead, 5% for vanadium, and 1% for nickel). The remaining constituents most associated with these sources had less than 1% censored daily concentrations.

Using APCA for source apportionment, the concentration means and standard deviations of the sources were similar in magnitude across censoring adjustment methods, with some differences (Table 6). For example, the estimated mean concentration of secondary sulfate ranged from $3.29 \mu\text{g}/\text{m}^3$ using our model-based method to $7.23 \mu\text{g}/\text{m}^3$ using the data reported below the MDL. Figure 2 shows APCA-estimated source time series from April 2001-September 2001 for the reported data and three censoring adjustment methods. The pink band is the interquartile range of estimated time series using the model-based method and demonstrates uncertainty in estimating source concentrations driven by censored data. APCA does not explicitly constrain source concentrations to be positive and can estimate negative source concentrations, particularly for days with low pollution. The time series plots show similar trends across censoring adjustment methods but there exist some differences in the average source concentration and source concentration variability. For secondary sulfate, which was relatively unaffected by censored data, all censoring adjustment methods performed comparably. Though soil had some censored data, particularly from aluminum, there was little difference in censoring adjustment methods. For residual oil and traffic, there were more differences across censoring adjustment methods, though traffic had little censored data compared with residual oil or soil sources. Note that Figure 2 only displays data for 48 of 174 days of data in New York City so that the points can be seen clearly and therefore does not exactly reflect the means and standard deviations reported in Table 6.

We also applied PMF to data from New York City and found the source estimates were more similar across censoring adjustment methods (Table S12, Figure S1 Online Resource). As with the APCA results, we found the largest differences across censoring adjustment methods for residual oil. Differences between our results and those reported for New York City by Ito *et al.* (2004) may be due to a different dataset, different implementations of source apportionment models, and different methods of source identification.

5 Discussion

We have provided the first comprehensive examination of how different censoring adjustment methods impact estimation of $\text{PM}_{2.5}$ sources using APCA and PMF. Generally sources must be estimated from available $\text{PM}_{2.5}$ constituent concentrations, which frequently are censored below MDLs. Because common source apportionment methods cannot directly handle censored data, guidance on how to adjust censored constituent data is critical for $\text{PM}_{2.5}$ source estimation. While many previous studies have determined the best methods for adjusting censored data when estimating summary statistics or performing traditional factor analysis or PCA (Helsel, 2005b; Farnham *et al.*, 2002; Aruga, 1997), no studies have examined how censored data impacts common source apportionment models. Most source apportionment studies do not use model-based approaches to impute censored data. We demonstrated that while a model-based imputation approach frequently leads to

better source apportionment results, substitution methods are also appropriate when few data are censored.

If the assumed distribution is correct, model-based approaches estimate the complete data distribution well and are generally preferred to using standard substitution methods (Helsel, 2010; Ganser and Hewett, 2010). Since source apportionment models cannot handle censored data, sources are generally estimated by first imputing or removing censored data and then performing source apportionment. Because different source apportionment models make different assumptions about the data, the optimal method for estimating the data distribution may not be the optimal method for estimating sources. For example, APCA ignores temporal correlation in the data, which might explain why our model-based method, which assumes temporal independence and multivariate normality for the logged data, improves source estimation under APCA. For PMF, we found that censoring adjustment methods performed more similarly, which indicates that the choice of censoring adjustment method should depend on the source apportionment model. Future work could develop a novel model-based approach for PMF that inflates censored observation uncertainties in equation 1 using multiple imputation variability. Under both APCA and PMF, when few data are censored, substituting concentrations below the MDL with $\frac{1}{2} \times \text{MDL}$ leads to source apportionment results similar to those obtained using a model-based method. It is not known to what extent our results are applicable to other source apportionment models such as Unmix.

The exclude and/or downweight methods did not perform consistently better than other censoring adjustment methods in our simulation study. In our simulation study using PMF, we assumed all data had equal uncertainties to minimize differences from APCA. Paatero and Hopke (2003) suggested that if constituents with small concentrations have small signal-to-noise ratios, dropping or downweighting constituents may lead to less bias in PCA. If we generated data so that concentrations close to the MDL had larger uncertainties, downweighting constituents may have led to better source apportionment results. Previous studies have selected uncertainties as a function of the MDL and the analytical uncertainty (Song *et al.*, 2001; Larson *et al.*, 2004), but it is not clear how best to generate analytical uncertainties for a simulation study. Additionally, we randomly selected constituents to be censored, while in practice the constituents with many censored concentrations may not be necessary for source estimation and excluding them may, in fact, decrease bias. Randomly selecting constituents to be censored allowed us to examine the impact of censoring for constituents representing a range of importance for estimating sources.

We estimated sources of PM_{2.5} in New York City and found that the differences between censoring adjustment methods varied between PM_{2.5} sources. Both residual oil and soil PM_{2.5} had associated constituents with a large amount of censored data. While estimates of soil PM_{2.5} were more similar across censoring adjustment methods, residual oil PM_{2.5} differed substantially between censoring adjustment methods (Table 6). Additionally, elemental carbon had a small amount of censored data, but this heavily impacted estimation of traffic PM_{2.5}. Therefore both the importance of censored constituents for attributing sources and the proportion of censored data determine how censoring adjustment methods impact source estimation.

We used a classification algorithm (details in the Online Resource, Section 3) in our simulation study to match estimated sources to true sources. In practice, researchers match estimated sources to known sources of pollution manually by examining the source profiles or factor loadings, though this method is impossible to replicate in a large simulation study. While we calibrated our classification method to our simulated data, this method may not always provide the same classification as a researcher would assign. We estimated mean biases and variance ratios in the simulation study conditional on our ability to match the sources estimated from the censored data with the true sources, however we were not able to completely eliminate misclassification. In the 5-source scenario under APCA, the variance of DMV was overestimated under censoring and the variance of GMV was underestimated under censoring (Table 5). These extreme values likely occurred because the simulated GMV source was classified as DMV and the variance used to generate the DMV source was much smaller than the variance used for the GMV source (Table S3, Online Resource). This example reflects possible reporting errors in practice where the source mean and variance could be incorrectly estimated because the source was incorrectly misclassified. This misclassification causes some difficulty in comparing simulation study results between APCA and PMF. The results from the simulation study indicate that PMF misclassifies sources more frequently than APCA, but for PMF the mean biases were often smaller and the variance ratios were often closer to one. For $\frac{1}{2} \times \text{MDL}$, it is possible that PMF misclassifies sources under censoring, leading to better estimates of the mean and variance conditional on correct classification. The overall impact of censoring on the mean and variance were likely underestimated in the simulation study because we could only compute these quantities for sources correctly identified under censoring. For example if dust and DMV in the 3-source scenario were classified as coal and GMV under censoring, we were unable to determine which source was dust.

We did not directly compare APCA and PMF, though other studies have found that source apportionment results were similar across methods (Ito *et al.*, 2004; Hopke *et al.*, 2006; Rizzo and Scheff, 2007; Lingwall *et al.*, 2008). Our work adds to the body of research comparing APCA and PMF by demonstrating that censored data should be treated differently depending on the source apportionment model. Specifically, our model-based method led to less bias due to censoring in source estimation under APCA, but had a smaller impact on PMF results. Instead of using APCA or PMF, we could directly modify the source apportionment model to handle censored data. Tobit factor analysis can be applied to censored data, but does not yield non-negative results and is not frequently applied to environmental data (Muthén, 1989; Kamakura and Wedel, 2001). Additionally, a fully Bayesian factor analysis model (Lingwall *et al.*, 2008; Nikolov *et al.*, 2011) could be fitted to the observed and censored data that yields non-negative results. One limitation in source apportionment modeling is that the estimated “sources” may include latent factors that do not correspond to specific emissions sources, such as $\text{PM}_{2.5}$ generated by photochemical reactions. By incorporating prior information into Bayesian source apportionment models, we could ensure estimated sources more closely resemble known emission sources (Hackstadt and Peng, 2014).

In our model-based approach, we assume constituent concentrations follow a multivariate 2-parameter lognormal distribution. As with any parametric model, this distribution only serves as an approximation to the true, unknown data distribution. We did not find evidence that our data was not 2-parameter lognormally distributed using Kolmogorov-Smirnov goodness-of-fit tests. However, the multivariate 2-parameter lognormal distribution does not allow zero concentrations to occur and is not bounded above. The Johnson SB distribution better characterizes these properties of environmental data by allowing zero concentrations and bounding the concentrations from above (Mage, 1980; Johnson, 1949; Mage, 1996), but this distribution has not been previously applied to model multivariate PM_{2.5} chemical speciation data. In other applications of our model-based approach, the 2-parameter lognormal distribution may not fit the data well and other parametric models such as the Johnson SB distribution should be explored. Future work could extend our model-based approach to incorporate more flexible models for constituent concentrations that better match theoretical properties of pollution data.

In most time series studies of the short-term health effects of PM_{2.5} sources, source apportionment is first used to estimate PM_{2.5} sources (Laden *et al.*, 2000; Mar *et al.*, 2006). Then the associations between estimated source concentrations and daily adverse health counts, such as mortality, are estimated using log-linear time series regression. Measurement error in source concentration estimates can lead to biased health effect estimates in this two-stage approach (Zidek *et al.*, 1996; Fung and Krewski, 1999). Specifically, underestimated or overestimated variability in source concentrations can bias estimated health effects. Our simulation study results indicated that source concentration variances estimated from source apportionment differ between censored and observed data (Tables 4, 5, Tables S6–S7 (Online Resource)), which could lead to bias in subsequent health effects analyses. The model-based approach for imputing censored data allows us to examine the uncertainty in source estimation driven by censored data (e.g. Figure 2). By using our model-based approach to estimate sources prior to estimating source-specific health effects, we can propagate the uncertainty in source estimation driven by censoring to estimate the precision of our health effect estimate. The findings of this work should be incorporated in future studies examining sources of uncertainty and error in both source estimation and estimation of source-specific health effects. It is possible that the uncertainty driven by estimating latent sources from chemical constituent data is much larger than the error driven by censored data. We can better understand uncertainty in source estimation by exploring what uncertainty is driven by measurement error in the chemical constituent concentrations, error caused by censored constituent data, and uncertainty in source apportionment.

We have demonstrated how different censoring adjustment methods impact source estimation using APCA and PMF. When many data were censored, a model-based approach to multiply impute censored data improved source estimation when APCA was used for source apportionment. When few data were censored, substituting chemical constituent concentrations with $\frac{1}{2} \times \text{MDL}$ led to similar estimates of source means and variances as when uncensored data were used for both APCA and PMF. In general, excluding or downweighting constituents with many censored concentrations did not improve source estimation. We estimated PM_{2.5} sources in New York City and found estimated source

means and variances differed by censoring adjustment method for APCA, but were more similar across methods for PMF. In addition, the impact of censoring depended on which constituents were censored and whether they were important for estimating sources. Estimation of PM_{2.5} emitted from different sources can be impacted by the method chosen to impute or remove censored PM_{2.5} constituent concentrations. Therefore careful selection of censoring adjustment methods in source apportionment is necessary.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Aruga R. Treatment of responses below the detection limit: some current techniques compared by factor analysis on environmental data. *Analytica Chimica Acta*. 1997; 354(1–3):255–262.
- Bell ML, Dominici F, Ebisu K, Zeger SL, Samet JM. Spatial and temporal variation in PM_{2.5} chemical composition in the United States for health effects studies. *Environmental Health Perspectives*. 2007; 115(7):989–995. [PubMed: 17637911]
- Bell ML, Ebisu K, Peng RD, Samet JM, Dominici F. Hospital admissions and chemical composition of fine particle air pollution. *American Journal of Respiratory and Critical Care Medicine*. 2009; 179(12):1115–1120. [PubMed: 19299499]
- Chen H, Quandt SA, Grzywacz JG, Arcury TA. A Bayesian multiple imputation method for handling longitudinal pesticide data with values below the limit of detection. *Environmetrics*. 2013; 24(2): 132–142. [PubMed: 23504271]
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*. 1977:1–38.
- Environmental Protection Agency. Integrated Science Assessment for particulate matter. 2009
- Environmental Protection Agency. electronic code of federal regulations. 2015 title 40. 136.
- Farnham IM, Singh AK, Stetzenbach KJ, Johannesson KH. Treatment of nondetects in multivariate analysis of groundwater geochemistry data. *Chemometrics and Intelligent Laboratory Systems*. 2002; 60(1–2):265–281.
- Francis RA, Small MJ, VanBriesen JM. Multivariate distributions of disinfection by-products in chlorinated drinking water. *Water Research*. 2009; 43(14):3453–3468. [PubMed: 19539341]
- Franklin M, Koutrakis P, Schwartz J. The role of particle composition on the association between PM_{2.5} and mortality. *Epidemiology*. 2008; 19(5):680–9. [PubMed: 18714438]
- Fung KY, Krewski D. On measurement error adjustment methods in Poisson regression. *Environmetrics*. 1999; 10(2):213–224.
- Ganser GH, Hewett P. An accurate substitution method for analyzing censored data. *Journal of Occupational and Environmental Hygiene*. 2010; 7(4):233–244. [PubMed: 20169489]
- Hackstadt AJ, Peng RD. A bayesian multivariate receptor model for estimating source contributions to particulate matter pollution using national databases. *Environmetrics*. 2014; 25(7):513–527. [PubMed: 25309119]
- Harris CW, Kaiser HF. Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*. 1964; 29(4):347–362.
- Helsel D. Much ado about next to nothing: incorporating nondetects in science. *Annals of Occupational Hygiene*. 2010; 54(3):257–262. [PubMed: 20032004]
- Helsel DR. More than obvious: better methods for interpreting nondetect data. *Environmental Science & Technology*. 2005a; 39(20):419A–423A.
- Helsel, DR. Nondetects and data analysis. *Statistics for censored environmental data*. Wiley-Interscience; 2005b.
- Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*. 2006; 65(11):2434–2439. [PubMed: 16737727]

- Henry RC. History and fundamentals of multivariate air quality receptor models. *Chemometrics and Intelligent Laboratory Systems*. 1997; 37(1):37–42.
- Hopke PK, Liu C, Rubin DB. Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the arctic. *Biometrics*. 2001; 57(1):22–33. [PubMed: 11252602]
- Hopke PK, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV, et al. PM source apportionment and health effects: 1. Intercomparison of source apportionment results. *Journal of Exposure Science and Environmental Epidemiology*. 2006; 16(3):275–286. [PubMed: 16249798]
- Ito K, Xue N, Thurston G. Spatial variation of PM_{2.5} chemical species and source-apportioned mass concentrations in New York City. *Atmospheric Environment*. 2004; 38(31):5269–5282.
- Johnson NL. Systems of frequency curves generated by methods of translation. *Biometrika*. 1949:149–176. [PubMed: 18132090]
- Kamakura WA, Wedel M. Exploratory Tobit factor analysis for multivariate censored data. *Multivariate Behavioral Research*. 2001; 36(1):53–82.
- Kavouras IG, Koutrakis P, Cereceda-Balic F, Oyola P. Source apportionment of PM₁₀ and PM_{2.5} in five Chilean cities using factor analysis. *Journal of the Air & Waste Management Association*. 2001; 51(3):451–464. [PubMed: 11266108]
- Kim E, Hopke PK, Paatero P, Edgerton ES. Incorporation of parametric factors into multilinear receptor model studies of Atlanta aerosol. *Atmospheric Environment*. 2003; 37(36):5009–5021.
- Krall JR, Anderson GB, Dominici F, Bell ML, Peng RD. Short-term exposure to particulate matter constituents and mortality in a national study of US urban communities. *Environmental Health Perspectives*. 2013; 121(10):1148–1153. [PubMed: 23912641]
- Laden F, Neas LM, Dockery DW, Schwartz J. Association of fine particulate matter from different sources with daily mortality in six U.S. cities. *Environmental Health Perspectives*. 2000; 108(10):941–947. [PubMed: 11049813]
- Larson T, Gould T, Simpson C, Liu LJS, Claiborn C, Lewtas J. Source apportionment of indoor, outdoor, and personal PM_{2.5} in Seattle, Washington, using Positive Matrix Factorization. *Journal of the Air & Waste Management Association*. 2004; 54(9):1175–1187. [PubMed: 15468670]
- Lee S, Liu W, Wang Y, Russell AG, Edgerton ES. Source apportionment of PM_{2.5}: Comparing PMF and CMB results for four ambient monitoring sites in the southeastern United States. *Atmospheric Environment*. 2008; 42(18):4126–4137.
- Lingwall JW, Christensen WF, Reese CS. Dirichlet based Bayesian multivariate receptor modeling. *Environmetrics*. 2008; 19(6):618–629.
- Little, RJA.; Rubin, DB. *Statistical analysis with missing data*. Wiley; 2002.
- Mage D. A probability model for the age distribution of sids. *Journal of Sudden Infant Death Syndrome and Infant Mortality*. 1996; 1(1):13–31.
- Mage D, Wilson W, Hasselblad V, Grant L. Assessment of human exposure to ambient particulate matter. *Journal of the Air & Waste Management Association*. 1999; 49(11):1280–1291. [PubMed: 10589295]
- Mage DT. An explicit solution for sb parameters using four percentile points. *Technometrics*. 1980; 22(2):247–251.
- Mar TF, Ito K, Koenig JQ, Larson TV, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Neas L, et al. 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*. 2006; 16(4):311–320. [PubMed: 16288316]
- Marmur A, Unal A, Mulholland JA, Russell AG. Optimization-based source apportionment of PM_{2.5} incorporating gas-to-particle ratios. *Environmental Science & Technology*. 2005; 39(9):3245–3254. [PubMed: 15926575]
- Maykut NN, Lewtas J, Kim E, Larson TV. Source apportionment of PM_{2.5} at an urban IMPROVE site in Seattle, Washington. *Environmental Science & Technology*. 2003; 37(22):5135–5142. [PubMed: 14655699]

- McDonald JD, Zielinska B, Sagebiel JC, McDaniel MR, Mousset-Jones P. Source apportionment of airborne fine particulate matter in an underground mine. *Journal of the Air & Waste Management Association*. 2003; 53(4):386–395. [PubMed: 12708502]
- Muthén BO. Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology*. 1989; 42(2):241–250.
- Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. An informative Bayesian structural equation model to assess source-specific health effects of air pollution. *Biostatistics*. 2007; 8(3):609–624. [PubMed: 17032699]
- Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. Multiplicative factor analysis with a latent mixed model structure for air pollution exposure assessment. *Environmetrics*. 2011; 22(2):165–178.
- Norris, G.; Vedantham, R.; Duvall, R.; Henry, RC. EPA Unmix 6.0 fundamentals & user guide. US Environmental Protection Agency; Washington DC: 2007.
- Norris, G.; Vedantham, R.; Wade, K.; Brown, S.; Prouty, J.; Foley, C. EPA Positive Matrix Factorization 3.0 fundamentals & user guide. US Environmental Protection Agency; Washington DC: 2008.
- Ostro B, Feng WY, Broadwin R, Green S, Lipsett M. The effects of components of fine particulate air pollution on mortality in California: results from CALFINE. *Environmental Health Perspectives*. 2007; 115(1):13–19. [PubMed: 17366813]
- Paatero P. The multilinear engine: A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*. 1999; 8(4):854–888.
- Paatero P, Hopke PK. Discarding or downweighting high-noise variables in factor analytic models. *Analytica Chimica Acta*. 2003; 490(1–2):277–289.
- Paatero P, Tapper U. Positive Matrix Factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*. 1994; 5(2):111–126.
- Peng RD, Bell ML, Geyh AS, McDermott A, Zeger SL, Samet JM, Dominici F. Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental Health Perspectives*. 2009; 117(6):957–963. [PubMed: 19590690]
- Polissar AV, Hopke PK, Poirot RL. Atmospheric aerosol over Vermont: chemical composition and sources. *Environmental Science & Technology*. 2001; 35(23):4604–4621. [PubMed: 11770762]
- Querol X, Alastuey A, Rodriguez S, Plana F, Ruiz CR, Cots N, Massagué G, Puig O. PM₁₀ and PM_{2.5} source apportionment in the Barcelona metropolitan area, Catalonia, Spain. *Atmospheric Environment*. 2001; 35(36):6407–6419.
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2012.
- Rizzo MJ, Scheff PA. Fine particulate source apportionment using data from the USEPA speciation trends network in Chicago, Illinois: Comparison of two source apportionment models. *Atmospheric Environment*. 2007; 41(29):6276–6288.
- Song XH, Polissar AV, Hopke PK. Sources of fine particle composition in the northeastern US. *Atmospheric Environment*. 2001; 35(31):5277–5286.
- Song Y, Tang X, Xie S, Zhang Y, Wei Y, Zhang M, Zeng L, Lu S. Source apportionment of PM_{2.5} in Beijing in 2004. *Journal of Hazardous Materials*. 2007; 146(1–2):124–130. [PubMed: 17208371]
- Thurston GD, Spengler JD. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. *Atmospheric Environment*. 1985; 19(1):9–25.
- Zanobetti A, Schwartz J. The effect of fine and coarse particulate air pollution on mortality: a national analysis. *Environmental Health Perspectives*. 2009; 117(6):898–903. [PubMed: 19590680]
- Zanobetti A, Franklin M, Koutrakis P, Schwartz J. Fine particulate air pollution and its components in association with cause-specific emergency admissions. *Environmental Health*. 2009; 8(58):1–12. [PubMed: 19138417]
- Zhou J, Ito K, Lall R, Lippmann M, Thurston G. Time-series analysis of mortality effects of fine particulate matter components in Detroit and Seattle. *Environmental Health Perspectives*. 2011; 119(4):461–466. [PubMed: 21193387]

Zidek JV, Wong H, Le ND, Burnett R. Causality, measurement error and multi-collinearity in epidemiology. *Environmetrics*. 1996; 7(4):441–451.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1. New York City time series for observed constituent data from April–November 2001 for two chemical constituents of $\text{PM}_{2.5}$: aluminum and calcium. Data below the MDL are marked with an asterisk at the MDL.

Fig. 2.

New York City time series of sources estimated using APCA from April-September 2001 using 30 chemical constituents. Time series were estimated using different censoring adjustment methods: $\frac{1}{2} \times \text{MDL}$, Exclude, and data Reported below the MDL. Also shown is the interquartile range of estimated time series using our model-based approach.

Table 1

Simulation study comparing censoring adjustment methods.

Source apportionment method	APCA
	PMF
Censoring adjustment method	$\frac{1}{2} \times \text{MDL}$
	Exclude and/or downweight constituents
	Model-based
Source scenario	3 sources (wood/DMV/dust)
	5 sources (wood/DMV/dust/GMV/coal)
Number (out of 23 total constituents) and quantile of constituents censored	
	2 constituents at {20%, 50%, or 80% }
	11 constituents at {20%, or 50% }

Median number of sources misclassified under different amounts of censoring for two source scenarios: 3 sources (wood/ DMV/ dust) and 5 sources (wood/ DMV/ dust/ GMV/ coal). APCA was used for source apportionment of 23 chemical constituents.

Table 2

Method	Sources	2 constituents			11 constituents		
		20%	50%	80%	20%	50%	80%
Model	3 sources	0.00	0.00	0.00	0.00	0.00	0.00
1/2 MDL	3 sources	0.00	0.00	0.00	0.00	0.00	0.00
Exclude	3 sources	-	1.00	1.00	-	1.00	1.00
Model	5 sources	0.00	0.00	0.00	0.00	0.00	0.00
1/2 MDL	5 sources	0.00	0.00	0.00	0.00	0.00	0.00
Exclude	5 sources	-	0.00	0.00	-	2.00	2.00

Median source concentration bias due to censoring, d_l , collapsed across sources as $\left(\sqrt{\frac{1}{L} \sum_{l=1}^L d_l^2}\right)$ for two source scenarios: 3 sources (wood/ DMV/ dust) and 5 sources (wood/ DMV/ dust/ GMV/ coal). APCA was used for source apportionment of 23 chemical constituents.

Table 3

Method	Sources	2 constituents			11 constituents		
		20%	50%	80%	20%	50%	80%
Model	3 sources	0.00	0.00	0.01	0.03	0.29	0.29
1/2 MDL	3 sources	0.11	0.19	0.09	0.56	0.87	0.87
Exclude	3 sources	-	0.05	0.06	-	0.13	0.13
Model	5 sources	0.00	0.02	0.03	0.21	2.41	2.41
1/2 MDL	5 sources	0.10	0.08	0.08	1.05	2.84	2.84
Exclude	5 sources	-	0.07	0.09	-	2.61	2.61

Table 4

Median ratio of concentration variances between data with and without censoring across 300 simulated datasets for sources wood/DMV/dust. APCA was used for source apportionment of 23 chemical constituents.

Method	source	2 constituents			11 constituents		
		20%	50%	80%	20%	50%	80%
Model	wood	1.00	1.00	1.00	1.00	1.00	1.00
1/2 MDL	wood	1.00	1.00	1.00	1.00	1.00	1.00
Exclude	wood	-	1.00	1.01	-	-	1.02
Model	DMV	1.00	1.00	1.00	1.00	1.00	1.00
1/2 MDL	DMV	1.00	1.00	1.00	0.99	0.98	0.98
Exclude	DMV	-	0.99	0.99	-	-	0.98
Model	dust	1.00	1.00	1.00	1.00	1.00	1.00
1/2 MDL	dust	1.00	1.00	1.00	0.99	0.98	0.98
Exclude	dust	-	0.98	0.98	-	-	0.96

Table 5

Median ratio of concentration variances between data with and without censoring across 300 simulated datasets for sources wood/DMV/dust/GMV/coal. APCA was used for source apportionment of 23 chemical constituents.

Method	source	2 constituents			11 constituents		
		20%	50%	80%	20%	50%	80%
Model	wood	1.00	1.00	1.00	1.00	1.00	1.00
1/2 MDL	wood	1.00	1.00	1.00	1.00	1.00	1.00
Exclude	wood	-	1.02	1.02	-	-	1.17
Model	DMV	1.00	1.01	1.02	1.35	3.53	
1/2 MDL	DMV	1.01	1.05	1.09	8.61	10.75	
Exclude	DMV	-	1.33	1.32	-	79.82	
Model	dust	1.00	1.00	1.00	0.99	1.00	
1/2 MDL	dust	1.00	1.00	1.01	0.98	1.00	
Exclude	dust	-	1.05	1.06	-	1.67	
Model	GMV	1.00	1.00	1.00	1.00	0.50	
1/2 MDL	GMV	1.00	1.00	1.00	0.96	0.52	
Exclude	GMV	-	0.99	0.99	-	0.64	
Model	coal	1.00	1.00	1.00	1.00	0.99	
1/2 MDL	coal	1.00	1.00	1.00	1.00	0.98	
Exclude	coal	-	1.00	1.00	-	0.63	

Table 6

Mean concentrations (standard deviations) in $\mu\text{g}/\text{m}^3$ for four sources in New York City estimated using APCA including soil, secondary sulfate, traffic, and residual oil/incineration. Results from 30 chemical constituents using four different methods for adjusting censored data are shown: data Reported below the MDL, Model-based, $\frac{1}{2} \times \text{MDL}$, Exclude.

	Soil	Secondary sulfate	Traffic	Residual oil/incineration
Reported below the MDL	2.51 (2.59)	7.23 (6.86)	5.01 (5.35)	1.26 (1.82)
Model	2.03 (1.91)	3.29 (5.61)	7.89 (5.88)	4.05 (3.07)
1/2 MDL	1.78 (2.19)	5.88 (6.74)	4.73 (6.09)	0.4 (1.44)
Exclude	2.56 (2.14)	5.86 (6.22)	4.28 (6.63)	1.68 (2.38)