Please see the online version of this article for supplementary materials.

Incorporating Time-Dependent Source Profiles Using the Dirichlet Distribution in Multivariate Receptor Models

Matthew J. HEATON

Department of Statistical Science
Box 90251
Duke University
Durham, NC 27708-0251
(*matt@stat.duke.edu*)

C. Shane REESE

Department of Statistics
Brigham Young University
230 TMCB
Provo, UT 84602
(*reese@stat.byu.edu*)

William F. CHRISTENSEN

Department of Statistics
Brigham Young University
230 TMCB
Provo, UT 84602
(*william@stat.byu.edu*)

Multivariate receptor modeling is used to estimate profiles and contributions of pollution sources from concentrations of pollutants such as particulate matter in the air. The majority of previous approaches to multivariate receptor modeling assume pollution source profiles are constant through time. In an effort to relax this assumption, this article uses the Dirichlet distribution in a dynamic linear receptor model for pollution source profiles. The receptor model developed herein is evaluated using simulated datasets and then applied to a physical dataset of chemical species concentrations measured at the U.S. Environmental Protection Agency's St. Louis–Midwest supersite. Supplemental materials to this articles are available online.

KEY WORDS: Adaptive MCMC; Dynamic model; Source apportionment; Time-varying loading.

1. INTRODUCTION

Ambient air pollutants have been linked to detrimental environmental and public health effects. An important first step in mitigating harmful effects of pollutants is the identification of pollution sources and quantification of their environmental impacts. Source apportionment seeks to derive information about pollution sources from ambient measurements of chemical species concentrations obtained from one or more receptor sites. Specifically, source apportionment seeks to (1) identify the major sources of pollution and (2) estimate the contribution of each source to pollution measured at the receptor site. By estimating and tracking pollution source contributions over time, government agencies can regulate the amount of emitted pollution and develop strategies for minimizing health risks to persons in proximity to pollution sources.

The basic source apportionment model (see Miller, Friedlander, and Hidy 1972) is written as

$$y_{pt} = \sum_{k=1}^{K} \lambda_{pk} f_{kt} + e_{pt}, \qquad t = 1, \ldots, T, p = 1, \ldots, P, \quad (1)$$

where y_{pt} is the concentration of chemical species p measured at time t, λ_{pk} is the proportion of chemical p in emissions from source k, f_{kt} is the concentration of pollutants contributed by source k to the air at the receptor site at time t, e_{pt} is the error term associated with y_{pt}, K is the total number of pollution sources, P is the total number of measured chemical species, and T is the total number of time periods. To maintain the physical interpretation of model parameters mentioned above, f_{kt} and λ_{pk} are generally constrained to be nonnegative and $\sum_{p=1}^{P} \lambda_{pk} \leq 1$. Using source apportionment terminology, the vector $\boldsymbol{\lambda}_k = (\lambda_{1k}, \ldots, \lambda_{Pk})'$ is referred to as the kth pollution source profile and f_{kt} is called the kth source contribution at time t. Model (1) is written in matrix form as

$$\mathbf{Y}_{P\times T} = \boldsymbol{\Lambda}_{P\times K}\mathbf{F}_{K\times T} + \mathbf{E}_{P\times T}, \quad (2)$$

where $\mathbf{Y}$ is the matrix of chemical concentration measurements over T time periods, $\boldsymbol{\Lambda}$ is the matrix of pollution source profiles, $\mathbf{F}$ is the matrix of source contributions, and $\mathbf{E}$ is the matrix of model errors.

From a statistical modeling perspective, elements of $\boldsymbol{\Lambda}$ and $\mathbf{F}$ are unknown model parameters. The number of pollution sources, K, could also be treated as a model parameter. While methods for estimating K is a current area of research, the primary focus of this article is to propose a general class of statistical models for incorporating time-dependent source profiles. For this reason, K is assumed to be known throughout. This assumption is not completely without merit because the datasets

TECHNOMETRICS, FEBRUARY 2010, VOL. 52, NO. 1
DOI 10.1198/TECH.2009.08134

used in this article have been thoroughly studied using alternative statistical methods. For more on methods for estimating K in factor models, see Park, Oh, and Guttorp (2002) and Lopes and West (2004).

The assumptions made about (2) vary by modeling technique. For example, chemical mass balance (CMB) models assume $\mathbf{\Lambda}$ is constant through time and known up to some measurement error. This assumption implies that not only are all pollution sources known but the chemical composition of emissions from these sources are measured up to a degree of uncertainty. Some of the several techniques used to estimate (2) from the CMB modeling perspective include weighted least squares (Miller, Friedlander, and Hidy 1972), effective variance (Watson, Cooper, and Huntzicker 1984), method of moments (Fuller 1987, pp. 193–194), and Britt and Luecke's method (Britt and Luecke 1973). Each of these methods, however, require assumptions about (2), in addition to known $\mathbf{\Lambda}$, which are not reasonable. For example, weighted least squares assumes chemical concentrations, y_{pt}, are normally distributed and have support on the real line. This assumption is obviously false in that, by definition, $y_{pt} \geq 0$. Effective variance, method of moments, and Britt and Luecke's method make no attempt to alter the normality assumption proposed by weighted least squares, but rather develop iterative least squares algorithms to estimate $\mathbf{F}$ in the presence of measurement errors.

Approaches to the CMB problem have been developed which weaken, but do not remove, the assumption that $\mathbf{\Lambda}$ is known. For example, Marmur, Mulholland, and Russell (2007) consider a version of the chemical mass balance problem where profiles must be specified, but the specified profiles can be altered (subject to constraints) during the estimation process. Bandeen-Roche (1994) approaches the chemical mass balance problem by assuming some elements of $\mathbf{\Lambda}$ are known while others are treated as model parameters. Billheimer (2001) uses informative prior distributions for $\mathbf{\Lambda}$ and estimates $\mathbf{\Lambda}$ and $\mathbf{F}$ using a Markov chain Monte Carlo (MCMC) algorithm.

In contrast to CMB models, multivariate receptor (MR) models assume $\mathbf{\Lambda}$ is unknown. From the MR modeling perspective, (2) can be viewed as a factor analytic model where $\mathbf{F}$ is the factor scores matrix and $\mathbf{\Lambda}$ is the factor loadings matrix. While traditional factor analytic treatments have been used for source apportionment (see Thurston and Spengler 1985), these methods are not optimal due to the nonuniqueness of source contribution and source profile estimates. Bandeen-Roche (1994) points out that achieving model identifiability is nontrivial and outlines conditions for which identifiability is guaranteed. Park, Spiegelman, and Henry (2002) also investigates several approaches to achieve model identifiability which are appropriate for receptor models. In short, the constraints imposed by Park, Spiegelman, and Henry (2002) to ensure model identifiability equate to placing point mass priors on elements of $\mathbf{\Lambda}$. In contrast, the degree of model nonidentifiability can be decreased by placing informative priors on $\boldsymbol{\lambda}_k$ rather than using presumptuous point mass priors.

Among the methods that have been developed to either yield uniquely identifiable solutions or reduce the degree of model indeterminacy, include confirmatory factor analysis (Christensen and Sain 2002), iterated confirmatory factor analysis (Christensen, Schauer, and Lingwall 2006), Unmix (Henry 1997), and positive matrix factorization (Paatero and Tapper 1994). Both confirmatory and iterated confirmatory factor analysis provide unique estimates of $\mathbf{\Lambda}$ and $\mathbf{F}$ by fixing $q > K$ rows of $\mathbf{\Lambda}$ to prevent rotation. In iterated confirmatory factor analysis, however, the q fixed rows are chosen randomly at each iteration and the algorithm continues until a goodness-of-fit statistic is minimized.

Positive matrix factorization (PMF) differs from both confirmatory and iterated confirmatory factor analysis in that PMF obtains parameter estimates via minimization of error terms normalized by a measurement of uncertainty. Specifically, the PMF algorithm minimizes

$$\sum_{t=1}^{T}\sum_{p=1}^{P}\left[\frac{y_{pt}-\sum_{k=1}^{K}\lambda_{pk}f_{kt}}{s_{pt}}\right]^2, \tag{3}$$

where s_{pt} is the uncertainty estimate for y_{pt}. Positive matrix factorization does not guarantee uniquely identified estimates of $\mathbf{\Lambda}$ and $\mathbf{F}$, but merely reduces the degree of nonidentifiability. Lingwall and Christensen (2007) evaluate the use of PMF in pollution receptor models via simulation studies in the presence and absence of a priori information. As PMF is currently the most commonly used estimation technique, the methods developed in this article will be compared to the results based on a PMF analysis.

Lingwall, Christensen, and Reese (2008) estimate (2) from a Bayesian perspective by placing prior distributions over $\mathbf{\Lambda}$ and $\mathbf{F}$ and estimating the parameters via MCMC. Specifically, Lingwall, Christensen, and Reese (2008) use the Dirichlet distribution as a prior for the vector $\boldsymbol{\lambda}_k$ which maintains the multivariate structure of source profiles. The Bayesian approach to pollution receptor modeling, as illustrated by Lingwall, Christensen, and Reese (2008), is advantageous in that a priori information can be easily incorporated via informative prior distributions. Furthermore, estimation of complex quantities becomes almost trivial with draws from the joint posterior distribution.

A common assumption between CMB and MR models is that y_{pt} and $y_{pt'}$ are typically assumed to be independent for all $t \neq t'$. Because air pollution data are gathered as consecutive measurements of chemical concentrations across time, this assumption is neither physically nor empirically justifiable. Christensen and Sain (2002) propose to account for the temporal dependence in y_{pt} by using a nested-block bootstrap. The general idea of the nested-block bootstrap is to block y_{pt} into blocks of size l and resample these blocks when forming bootstrap replicates. For an optimal block size l, the dependence structure is preserved within each block and neighboring blocks are effectively independent. Park, Guttorp, and Henry (2001) propose a novel approach to account for the temporal dependence of y_{pt} by allowing f_{kt} and e_{pt} in (1) to evolve according to first-order autoregressive processes. Park, Guttorp, and Henry (2001) then uses MCMC techniques to sample from the full joint posterior distribution of model parameters.

In contrast to Park, Guttorp, and Henry (2001), this article investigates possible temporal dependence in y_{pt} arising from $\mathbf{\Lambda}$ being nonconstant and correlated through time. To date, all previous approaches to receptor modeling have assumed $\mathbf{\Lambda}$ is constant through time. The goal and contribution of this article is to relax this assumption by modeling temporal changes

in λ_k as a time-dependent process. To do so, the Dirichlet distribution is used as a prior distribution for source profiles at each time period. Let $\boldsymbol{\lambda}_{kt} = (\lambda_{1kt}, \ldots, \lambda_{Pkt})$ be the kth pollution source profile at time t. Specifically, this article models $\boldsymbol{\lambda}_{kt}$ as a time-dependent process where at each time t, $\boldsymbol{\lambda}_{kt}$ follows a Dirichlet distribution. Through this prior specification, not only is the multivariate structure of $\boldsymbol{\lambda}_{kt}$ maintained, but the degree of model indeterminancy can be reduced by informing this prior distribution. Because $\boldsymbol{\lambda}_{kt}$ follows a Dirichlet distribution at each time period, the model developed in this article is called the "Dirichlet process model." By modeling $\boldsymbol{\lambda}_{kt}$ as a time-dependent process, not only is the assumption of constant source profiles relaxed, but a degree of temporal dependence in y_{pt} is accounted for by the temporal structure of $\boldsymbol{\lambda}_{kt}$.

Section 2 investigates the temporal structure of pollution source profiles. Section 3 introduces the Dirichlet process (DP) model used in this article as well as computational issues and solutions for obtaining estimates of model parameters. Section 4 includes a discussion of the performance of the DP model as compared to PMF based on several simulated datasets. In Section 5, the DP model is used to analyze a physical dataset collected at the U.S. Environmental Protection Agency's St. Louis–Midwest supersite. Section 6 discusses conclusions and future areas of research.

2. THE TEMPORAL STRUCTURE OF POLLUTION SOURCE PROFILES

As previously mentioned, chemical concentrations (y_{pt}) are typically gathered as consecutive measurements over time and, hence, are temporally correlated. By way of illustration, consider a dataset containing daily measurements of 44 chemical species collected at a receptor site outside St. Louis for nearly two consecutive years (see Section 5 for a complete description of the St. Louis dataset). Figure 1 displays plots of autocorrelation functions (ACF) for y_{pt} as estimated using the St. Louis data. Figure 1 illustrates that different chemical species exhibit varying degrees of autocorrelation. Autocorrelation in chemical species can arise for various reasons. Atmospheric dispersion models suggest that the duration of a particles' ambient suspension depends on several factors including size and weight in addition to the meteorological conditions shared by all particles. Additionally, measurements of y_{pt} are collected as 24-hour aggregates. Thus, particles which can remain suspended for longer periods can have similar measurements on consecutive days.

Because of the autocorrelation present in air pollution data, proper statistical modeling needs to account for the dependence structure of y_{pt}. In addition to the Park, Guttorp, and

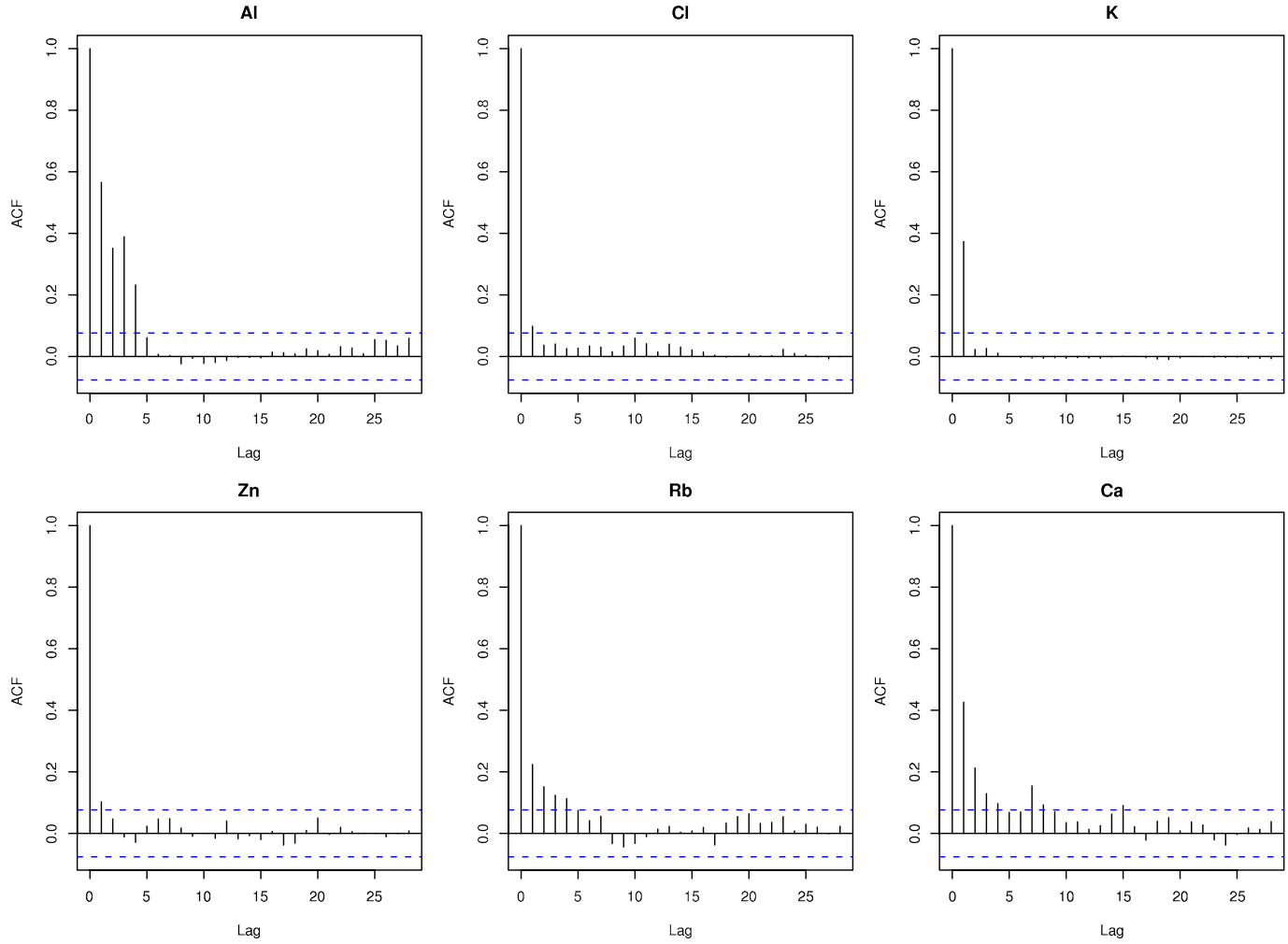

Figure 1. ACF plots of y_{pt} as estimated using data measured from a St. Louis receptor.

Henry (2001) assumption that the temporal structure of y_{pt} is accounted for by autocorrelation in f_{kt} and e_{pt}, the temporal structure of y_{pt} shown in Figure 1 may also be a byproduct of $\mathbf{\Lambda}$ being nonconstant and correlated over time. The possibility of nonconstant source profiles is supported by several physical facts. For example, the composition of vehicle emissions changes seasonally because fuel mixtures vary seasonally. Emissions also change depending on the proportion of diesel, hybrid, and unleaded gasoline vehicles on the road. Lee, Hopke, and Turner (2006) addressed this problem by separating diesel and auto emissions into two different sources. However, this separation can only be done with a priori information incorporated by the estimation method. Because such a priori information is often unavailable, most MR methods fall short in incorporating time-varying source profiles. Other pollution sources such as zinc and copper smelters are also subject to time-varying profiles because emission compositions from these sources change depending upon the choice of flux used to remove rock impurities prior to smelting.

To gather empirical evidence for potential autocorrelation in $\mathbf{\Lambda}$, a subset of the St. Louis dataset consisting of eight chemical species measured over 640 days was divided into 32 separate datasets of 20 days each. Let $\mathbf{\Lambda}_t$ be the estimate of $\mathbf{\Lambda}$ for the tth time period, for $t = 1, \ldots, 32$. Positive matrix factorization was then used to obtain estimates of $\mathbf{\Lambda}_t$ for $t = 1, \ldots, 32$ and these estimates were concatenated into a single time-varying estimate of $\mathbf{\Lambda}$. Figure 2 displays the time-varying estimate of $\boldsymbol{\lambda}_k$ for an identified winter-secondary pollution source along with an estimate for the constant source profile as estimated by PMF on all 640 days of data. Figure 2 shows that small fluctuations through time are exhibited by the most prominent elements of the source profile with the largest fluctuations occurring in Nitrate (NO_3). While this is a crude estimate of possible fluctuations in a source profile, the statistical evidence given in Figure 2 coupled with the physical justifications discussed above indicate a need to develop methodologies which account for time-varying source profiles.

As an extension of previous MR models, this article models the dependence structure of y_{pt} by allowing $\boldsymbol{\lambda}_k$ to vary through time. This article proposes using a Dirichlet prior distribution for $\boldsymbol{\lambda}_{kt}$ because the Dirichlet distribution correctly represents the multivariate structure of source profiles while maintaining proper constraints on $\boldsymbol{\lambda}_{kt}$. Furthermore, the temporal dependence in $\boldsymbol{\lambda}_{kt}$ is modeled using a generalized dynamic linear model (see West and Harrison 1997). The log-normal distribution is used as a convenient prior distribution for f_{kt} in order to maintain nonnegativity constraints. Additionally, y_{pt} is assumed to be log-normal in order to extend the Gaussian error assumption of Park, Guttorp, and Henry (2001) to a more physically realistic setting.

3. THE DIRICHLET PROCESS MODEL

In order to model time varying source profiles, (1) can be expressed in a dynamic linear model (DLM) context (see West and Harrison 1997) with observation equation,

$$y_{pt} | \mathbf{\Lambda}_t, \mathbf{f}_t, w_{pt} \sim \mathrm{LN}[\mathbf{\Lambda}_t(p) \times \mathbf{f}_t, w_{pt}], \tag{4}$$

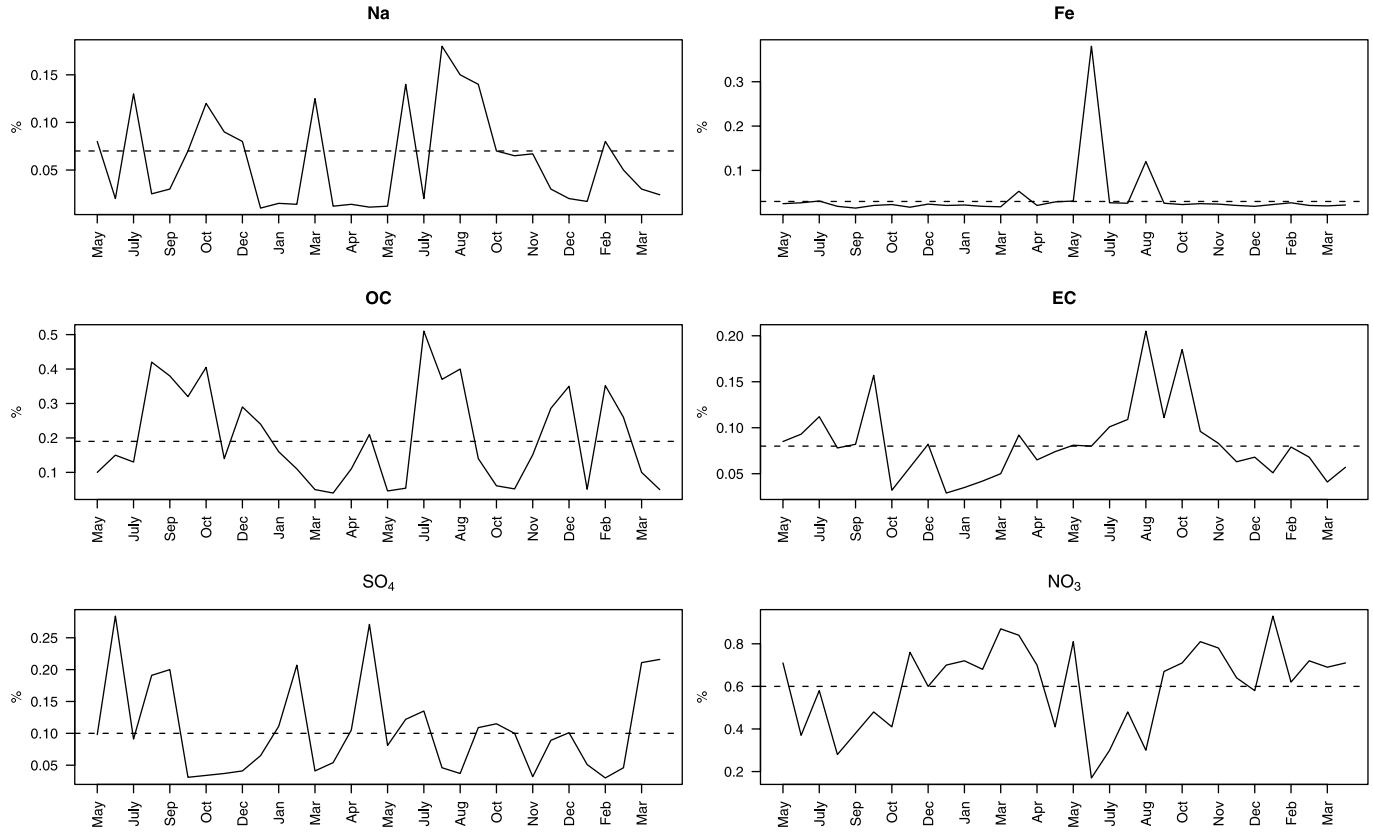

Figure 2. Time plot of winter secondary source profile. The dashed line indicates the PMF estimate of the source profile using all 640 days of data. Plots are on different scales to accent temporal variation in the profile.

where $\mathbf{\Lambda}_t(p)$ is the pth row of $\mathbf{\Lambda}_t = (\boldsymbol{\lambda}_{1t}, \ldots, \boldsymbol{\lambda}_{Kt})$, $\mathbf{f}_t = (f_{1t}, \ldots, f_{Kt})'$, $\mathbb{E}(y_{pt}) = \mathbf{\Lambda}_t(p) \times \mathbf{f}_t$, and w_{pt} is the coefficient of variation (CV) associated with y_{pt}. Thus, y_{pt} has density,

$$p(y_{pt}|\mathbf{\Lambda}_t, \mathbf{f}_t, w_{pt}) \propto \frac{1}{y_{pt}} \exp\{-(\ln(y_{pt}) - \ln(\mathbf{\Lambda}_t(p) \times \mathbf{f}_t) + 0.5\ln(w_{pt}^2+1)) / (2\ln(w_{pt}^2+1))\}. \quad (5)$$

Each $\boldsymbol{\lambda}_{kt}$ is then allowed to evolve through time according to the system equation,

$$\boldsymbol{\lambda}_{kt}|g_k, \boldsymbol{\lambda}_{k(t-1)} \sim \mathrm{DIR}[g_k \boldsymbol{\lambda}_{k(t-1)}], \quad (6)$$

where g_k is a precision parameter for the kth source. The density function for $\boldsymbol{\lambda}_{kt}$ is given by

$$p(\boldsymbol{\lambda}_{kt}|g_k, \boldsymbol{\lambda}_{k(t-1)}) \propto \prod_{p=1}^{P} \lambda_{pkt}^{g_k \lambda_{pk(t-1)}-1}. \quad (7)$$

The initial information of the DLM is

$$\boldsymbol{\lambda}_{k0}|\mathbf{m}_{k0} \sim \mathrm{DIR}[\mathbf{m}_{k0}], \quad (8)$$

$$f_{kt}|a_{kt}, b_{kt} \sim \mathrm{LN}[a_{kt}, b_{kt}]. \quad (9)$$

The hyperparameters $\mathbf{m}_{k0}$, a_{kt}, and b_{kt} are assumed known. For the remainder of this article (4)–(9) will be referred to as the "Dirichlet process" (DP) model. A special distinction should be made here in that the term "Dirichlet process" as it is used in this article is *not* the same as it is used in the Bayesian nonparametric literature. Rather, the use of the term "Dirichlet process" in this article denotes a time-dependent process where at each time period t, $\boldsymbol{\lambda}_{kt}$ follows a Dirichlet distribution.

First, in the DP model, chemical concentrations, y_{pt}, are assumed to follow a log-normal distribution with expectation $\mathbf{\Lambda}_t(p) \times \mathbf{f}_t$ and CV w_{pt}. The log-normal distribution is used here to extend the normality assumption of Park, Guttorp, and Henry (2001) to a physically realistic setting in which chemical concentrations are strictly nonnegative. The strong positive skewness imposed on y_{pt} by the log-normal distribution is physically justifiable in that large "outlying" concentrations can occur on days such as July 4th. Furthermore, high concentrations of chemicals can occur on days with severe weather.

Second, source profiles are allowed to vary according to the time-dependent process defined by (6). The precision parameter g_k represents a measure of autocorrelation in that it controls how similar $\boldsymbol{\lambda}_{kt}$ is to $\boldsymbol{\lambda}_{k(t-1)}$. Notice that, under the above parameterization, the expected value of λ_{pkt} is $\lambda_{pk(t-1)}$ with variance $\lambda_{pk(t-1)}(1-\lambda_{pk(t-1)})/(g_k+1)$; thus, g_k controls the variance of the temporal process while not affecting the expected value. For this reason, g_k is thought of as a measure of autocorrelation. Alternative methods and parameterizations for modeling temporal dependence in $\boldsymbol{\lambda}_{kt}$ have been proposed by Grunwald, Raftery, and Guttorp (1993) and Cargnoni, Muller, and West (1997) but the parameterization used in (6) is more intuitive for the reasons mentioned above. The parameter g_k is given a log-normal prior distribution with large variance to ensure $g_k > 0$ and allow the model to determine the amount of autocorrelation present in $\boldsymbol{\lambda}_{kt}$.

As the final component of the DP model, source contributions are assumed to follow a log-normal distribution with expectation a_{kt} and CV b_{kt}. Because f_{kt} must be restricted to the positive real line to maintain their physical interpretation, the use of the log-normal distribution as a prior distribution for all f_{kt} is justifiable. The log-normal distribution has the added benefit of having heavy tails which provides positive mass over possible large values of f_{kt}. The prior parameters a_{kt} and b_{kt} can be used at the discretion of the researcher to account for any a priori knowledge regarding f_{kt} but here each a_{kt} is set at 3 and each b_{kt} is set at 1. These values of a_{kt} and b_{kt} provide a sufficiently vague prior specification for each value of f_{kt}. An obvious extension of the DP model would be to model the temporal dependence as in Park, Guttorp, and Henry (2001) but, as the focus of this article is to introduce temporal dependence in $\mathbf{\Lambda}_t$, this extension is omitted here.

One aspect of temporal variation not explicity accounted for by the DP model is seasonal effects of pollution sources. Certainly some form of seasonal variation is to be expected. For example, the composition of mobile emissions could change seasonally as fuel mixtures vary by season. Explicitly building seasonality into the model, however, can be difficult because the pollution sources are unknown before hand and pollution sources can react different to seasonality. While not explicity built into the DP model, the model is sufficiently flexible to capture aspects of seasonality. Specifically, because f_{kt} and $\boldsymbol{\lambda}_{kt}$ are allowed to vary daily, these parameters can exhibit seasonal drifts over time. Indeed, seasonal drifts were identified when the DP model was applied to the St. Louis dataset (see Section 5).

The Bayesian formulation of the DP model provides an advantage over traditional approaches in that the results obtained therefrom are distributional results and probability statements regarding model parameters are perfectly legitimate. Additionally, Monte Carlo sampling and integration provide a simple approach to obtaining distributions over functions of model parameters. This added flexibility of the DP model over traditional MR models can be of great value to regulating bodies in obtaining probability distributions over complex quantities such as how many days in a year contributions from a certain pollution source exceed a prespecified threshold (see Section 5 for an example).

Due to the large number of model parameters that need to be estimated [a total of $(P \times K \times T) + (K \times T) + K$], MCMC sampling is employed for parameter estimation. A full description of the MCMC algorithm used in this article is included in the Appendix but a few points of interest are detailed here. First, at each iteration of the MCMC algorithm, g_k and f_{kt} are drawn using the random walk Metropolis algorithm. For example, when updating f_{kt}, proposal values are drawn from a $\mathrm{N}(f_{kt}^{(r-1)}, \sigma_{f_{kt}}^2)$ distribution where $f_{kt}^{(r)}$ is the value of f_{kt} at the rth iteration and $\sigma_{f_{kt}}^2$ is the variance of the proposal distribution. Each parameter vector $\boldsymbol{\lambda}_{kt}$ is drawn using the Metropolis–Hastings algorithm with proposal distribution $\mathrm{DIR}[l_{kt}\boldsymbol{\lambda}_{kt}^{(r-1)}]$. In this regard, l_{kt} becomes a scale parameter, similar to g_k in (6), which controls the variance of the proposal distribution.

In the MCMC algorithm, when updating $\boldsymbol{\lambda}_{k1}$, the Metropolis–Hastings acceptance probability requires the evaluation of

$$p(\boldsymbol{\lambda}_{k1}|\cdot) \propto L(\boldsymbol{\lambda}_{k1}) p(\boldsymbol{\lambda}_{kt}|g_k, \boldsymbol{\lambda}_{k0}) p(\boldsymbol{\lambda}_{k2}|g_k, \boldsymbol{\lambda}_{k1}), \quad (10)$$

where $L(\cdot)$ denotes the likelihood and "$|\cdot$" denotes conditioning on all other variables. A problem arises because, in this context, $\boldsymbol{\lambda}_{k0}$ is unknown and, hence, (10) cannot be evaluated. To circumvent this problem, $\boldsymbol{\lambda}_{k0}$ is treated as an additional parameter and sampled at each iteration of the algorithm. Specifically, at the rth iteration a proposal value $\boldsymbol{\lambda}_{k0}^{*}$ is drawn from $\mathrm{DIR}[l_{k0}\boldsymbol{\lambda}_{k0}^{(r-1)}]$ and accepted with probability,

$$\alpha = \min\left\{\frac{p(\boldsymbol{\lambda}_{k0}^{*}|\cdot)q(\boldsymbol{\lambda}_{k0}^{(r-1)}|l_{k0},\boldsymbol{\lambda}_{k0}^{*})}{p(\boldsymbol{\lambda}_{k0}|\cdot)q(\boldsymbol{\lambda}_{k0}^{*}|l_{k0},\boldsymbol{\lambda}_{k0}^{(r-1)})},1\right\}, \tag{11}$$

where $q(\cdot|\cdot)$ is the density of the proposal distribution given by (7). Note the acceptance probability given by (11) is the usual Metropolis–Hastings acceptance probability with limiting distribution $p(\boldsymbol{\lambda}_{k0}|\cdot)$. This sampled value is then used in evaluating the acceptance probability for $\boldsymbol{\lambda}_{k1}$.

To avoid excessive tuning of the scale of the distributions used to propose values of model parameters, the MCMC algorithm used in this article automatically adapts the scale of these distributions based on the acceptance rates. Specifically, if the algorithm "accepts" more than 60 of 100 consecutive proposals, the scale of the proposal distribution is increased by a factor of 10%. Alternatively, if the algorithm "rejects" more than 80 of 100 consecutive proposals, the proposal distribution is decreased by a factor of 10%. By using such an adaptation scheme, the algorithm adapts with the purpose of acheiving acceptance rates between 20% and 60%. Proposal distributions are adapted during the burn-in phase of the algorithm and left constant afterward. By tuning the scale of the proposal distribution during the burn-in phase only, the resulting Markov chain still converges to the correct stationary distribution because the proposal distribution is constant after the burn-in phase.

4. SIMULATION STUDY

To evaluate the performance of the DP model, datasets were simulated under a 2×3 full factorial design (see Cochran and Cox 1957, chapter 5) where the values of w_{pt} are taken to be $(0.2, 0.8)$ and the values of g_k are taken to be $(100, 250, \infty)$. The case where $g_k = \infty$ represents the assumption of constant source profiles. At each of the six combinations of w_{pt} and g_k, 50 datasets with $P = 44$, $K = 9$, and $T = 50$ were generated by first simulating values of $\boldsymbol{\Lambda}_t$ according to (6) where $\boldsymbol{\Lambda}_0$ was obtained from a previous analysis of the St. Louis dataset and then drawing y_{pt} from (4) where $\mathbf{f}_t$ was obtained from the same previous analysis. Using values of $\boldsymbol{\Lambda}_0$ and $\mathbf{F}$ from a previous analysis is preferable to individually specifying values because values based on previous analyses will be more realistic. The DP model was fit using 50,000 iterations of the MCMC algorithm with the first 25,000 constituting the burn-in phase. Positive matrix factorization was also fit to each of the simulated datasets with the uncertainty matrix calculated as w_{pt} times the data matrix.

Median absolute error (MAE) is used as a model performance metric for the DP model and PMF. Median absolute error for $\boldsymbol{\Lambda}_t$ (MAE_Λ) is calculated as $\sum_{p=1}^{P}|\lambda_{pkt} - \hat{\lambda}_{pkt}|$ where λ_{pkt} is the "true" value of λ_{pkt} used in simulating the datasets and $\hat{\lambda}_{pkt}$ is the median of the post-burn draws of the marginal posterior distribution for λ_{pkt} in the DP model and the point estimate when using PMF. Median absolute error for $\mathbf{F}$ is calculated in a similar manner as $\sum_{k=1}^{K}|f_{kt} - \hat{f}_{kt}|$ where f_{kt} is the "true" value and $\hat{f}_{kt}$ is the median of the posterior draws in the Bayesian setting and the point estimate under PMF. Smaller values for MAE indicate better model performance.

Plots (a) through (f) in Figure 3 display density plots for MAE_Λ under each combination of w_{pt} and g_k. Table 1 compares the median MAE_Λ for the DP model and for PMF. As compared to PMF, when profiles are time-dependent the DP model reduced MAE_Λ by an average of 68% and 67% when $w_{pt} = 0.2$ and $w_{pt} = 0.8$, respectively. Thus, the DP model is able to more correctly estimate $\boldsymbol{\Lambda}_t$ when $g_k \neq \infty$ than PMF. This performance is to be expected of the DP model under time-varying source profiles because the DP model is flexible enough to estimate time-varying profiles. A specific point of interest is the reduction in MAE_Λ when source profiles are constant over time; that is, $g_k = \infty$ (a key assumption for the use of PMF). In this case, the DP model reduces MAE_Λ by an average of 87%. Thus, even when the assumptions required for the use of PMF hold, the DP model still outperforms PMF in terms of model error.

The performance of the DP model in estimating $\mathbf{F}$ is dependent upon the value of w_{pt} as is displayed by plots (g) through (l) in Figure 3 as well as in Table 2. When $w_{pt} = 0.2$ and source profiles are time-variant, the DP model clearly achieves lower MAE_F than PMF with an average MAE_F of 4.166 under the DP model and an average MAE_F of 16.984 under PMF: a 75% reduction in error. When $w_{pt} = 0.8$ and source profiles are allowed to vary through time, the average MAE_F is 13.78 and 14.59 for the DP model and PMF respectively. This equates to a 6% difference in MAE_F across the two models. As was also seen in estimating $\boldsymbol{\lambda}_{kt}$, the DP model reduced MAE_F when the assumption of constant source profiles was true. This reduction of error is most prevalent by noting that MAE_F was reduced by 88% when $w_{pt} = 0.2$. When variation among y_{pt} was large ($w_{pt} = 0.8$) and source profiles are constant, model error under the DP model was 60% that of PMF.

While this article does not endeavor to comprehensively compare the DP model to PMF, early simulations indicate distinct advantages to using the DP model over PMF. For example, the DP model performs at least as well as PMF under the assumption of constant source profiles ($g_k = \infty$), a key assumption for the use of PMF. The DP model also has the added flexibility of incorporating time-varying profiles and outperforms PMF when time-varying profiles are present. Thus, the DP model is preferred to PMF in that it requires fewer assumptions for its use, it often has better performance in simulations, and it facilitates distributional analysis of model parameters rather than mere point estimates.

5. APPLICATION TO ST. LOUIS DATASET

The St. Louis Supersite is located on the Illinois side of St. Louis and monitors consolidated 24-hour measurements of fine particulate matter ($\mathrm{PM}_{2.5}$). Known pollution sources near the St. Louis Supersite include a steel mill to the norteast as well as zinc, copper, and lead smelters to the southwest. In total, the St. Louis dataset consists of 661 complete daily measurements taken over 749 days between May 2001 and May 2003 on 44

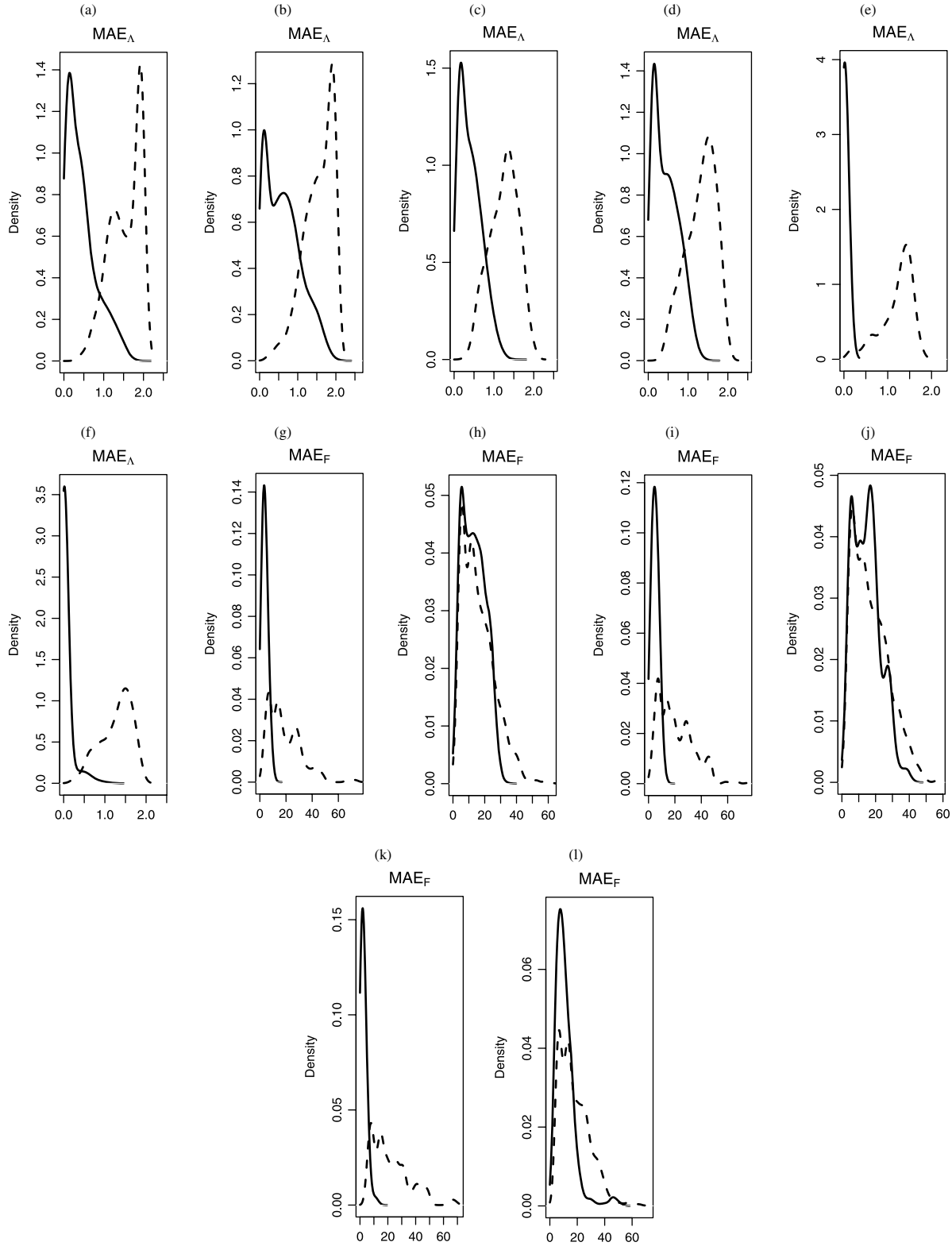

Figure 3. Density plots of MAE_Λ and MAE_F. The solid and dashed lines correspond to the MAE achieved by the DP model and PMF, respectively. The columns of plots correspond to $(w_{pt}, g_k) = (0.2, 100)$, $(0.8, 100)$, $(0.2, 250)$, $(0.8, 250)$, $(0.2, \infty)$, $(0.8, \infty)$, respectively. Across all specified values of (w_{pt}, g_k), MAE_Λ is lower for the DP model than PMF. When $w_{pt} = 0.2$ the DP reduces MAE_F compared to PMF, but when $w_{pt} = 0.8$ both models perform comparably.

Table 1. Median values of MAE_Λ for the DP and PMF models

	Factor levels					
	$g_k = 100$		$g_k = 250$		$g_k = \infty$	
Model	$w_{pt} = 0.2$	$w_{pt} = 0.8$	$w_{pt} = 0.2$	$w_{pt} = 0.8$	$w_{pt} = 0.2$	$w_{pt} = 0.8$
DP	0.349	0.566	0.351	0.401	0.160	0.182
PMF	1.599	1.605	1.292	1.387	1.321	1.365

chemical species including metals, organic, and elemental carbon (OC and EC), sulfate (SO_4), and nitrate (NO_3). Eighty-eight days had either complete or partial missing data. Other than one stretch of 5 consecutive missing measurements, none of the missing measurements occurred on consecutive days. The stretch of 5 consecutive days lasted between 11/21/2001 and 11/25/2001 (Wednesday–Sunday). Assuming source profiles do not change drastically over a 5 or 1 day period, the missing measurements were removed from the dataset and the DP model above was fit to the 661 days of complete data. A complete description of sampling methods for the St. Louis data set can be found in Lee, Hopke, and Turner (2006). Note that both S and SO_4 are included in the suite of chemical analyses because the two types of measurements can be useful in identifying subtle differences in source types. However, to keep from doubly impacting the source contribution estimates, the effect of S is eliminated and only SO_4 is considered when calculating the daily contribution amounts.

Prior to analysis, cross-validation techniques were used to assess the goodness of fit of the Dirichlet process model applied to the St. Louis dataset. Sixty-six days (approximately 10% of the complete dataset) were randomly selected to be cross-validated. For each of the 66 days, the Dirichlet process model was fit to the remaining 660 days in the dataset and the percentiles of the observed data in the posterior predictive distribution were recorded. For example, if day t was randomly selected to be cross-validated then the vector $\mathbf{y}_t$ was excluded and the Dirichlet process model was fit to the remaining data ($\mathbf{Y}_{-t}$). After fitting the model, the percentile (q_{pt}) of each $y_{pt} \in \mathbf{y}_t$ was calculated as

$$q_{pt} = 100 \times \int_0^{y_{pt}} p(\tilde{y}_{pt}|\mathbf{Y}_{-t})\, d\tilde{y}_{pt}, \tag{12}$$

where $p(\tilde{y}_{pt}|\mathbf{Y}_{-t})$ is the marginal posterior predictive distribution of y_{pt} given by

$$p(\tilde{y}_{pt}|\mathbf{Y}_{-t}) = \int_{\boldsymbol{\Theta}} p(\tilde{y}_{pt}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta}|\mathbf{Y}_{-t})\, d\boldsymbol{\Theta}, \tag{13}$$

$\tilde{y}_{pt}$ is the predicted value of y_{pt}, and $\pi(\boldsymbol{\Theta}|\mathbf{Y}_{-t})$ is the distribution of all model parameters ($\boldsymbol{\Theta}$) given all the data but the excluded time period. Due to the complexity of the integrals in (12) and (13), Monte Carlo integration techniques were used to calculate these quantities.

Using the above goodness-of-fit technique, the Dirichlet process model performed well in predicting concentrations for those chemicals with a high average concentration across time. For example, the distribution of q_{pt} for chemicals such as copper, zinc, lead, and elemental carbon were relatively uniform over the interval (0, 100). Additionally, the distribution of q_{pt} for other chemicals with high average concentrations such as sulfate and nitrate were unimodal, centered around 50 with high variation. Thus, the DP model was consistent with the observed data for chemicals with high average concentration.

In contrast, the Dirichlet process model exhibited positive or negative bias in predicting the concentration of those chemicals with low average concentration. For example, positive bias was exhibited by the DP model in predicting chemical concentrations such as those for magnesium and strontium. Additionally, the DP model exhibited negative bias in predicting chemicals such as titanium, vanadium, chromium, and gallium. In a few rare cases (e.g., cobalt and barium) the distribution of q_{pt} was bimodal with the observed values either falling near the 80th percentile or near the 20th percentile while rarely falling in the 40th to 60th percentile range. This model lack of fit is only of minor concern because research interest typically lies with those chemicals with high average concentration which are modeled correctly by the DP model.

For the analysis of the St. Louis dataset, the DP model was fit to all 44 observed chemical species using $K = 9$ sources due to the findings of Lingwall and Christensen (2007). Vague prior distributions for $\boldsymbol{\lambda}_{kt}$ and f_{kt} were used in fitting the DP model. Fifty thousand total iterations were used in the MCMC algorithm with the first 25,000 iterations constituting the burn-in phase.

After having fit the DP model, a common practice is to use the estimates of $\boldsymbol{\Lambda}_t$ and $\mathbf{F}$ to label the estimated sources. One useful quantity in labeling the pollution sources is the proportion of total predicted mass for chemical species p originating from source k. The total predicted mass for chemical species

Table 2. Median values of MAE_F for the DP and PMF models

	Factor levels					
	$g_k = 100$		$g_k = 250$		$g_k = \infty$	
Model	$w_{pt} = 0.2$	$w_{pt} = 0.8$	$w_{pt} = 0.2$	$w_{pt} = 0.8$	$w_{pt} = 0.2$	$w_{pt} = 0.8$
DP	3.440	12.797	4.891	14.590	2.180	9.453
PMF	15.884	13.901	18.083	15.279	18.676	15.373

Table 3. Distribution of explained chemical mass across sources

	Source								
Chemical	1	2	3	4	5	6	7	8	9
SO_4	0.857	0.069	0.033	0.015	0.019	0.000	0.002	0.002	0.002
OC	0.213	0.157	0.478	0.063	0.010	0.041	0.006	0.023	0.008
NO_3	0.090	0.897	0.002	0.002	0.005	0.001	0.002	0.000	0.002
EC	0.050	0.202	0.502	0.156	0.001	0.034	0.009	0.010	0.036
Na	0.406	0.158	0.356	0.005	0.010	0.036	0.006	0.022	0.001
K	0.048	0.074	0.088	0.003	0.081	0.019	0.667	0.010	0.009
Si	0.027	0.020	0.221	0.123	0.558	0.008	0.003	0.004	0.037
Fe	0.108	0.013	0.197	0.427	0.119	0.081	0.006	0.015	0.035
Ca	0.003	0.002	0.624	0.054	0.149	0.104	0.004	0.012	0.047
Cl	0.000	0.341	0.007	0.008	0.000	0.428	0.117	0.010	0.087
Mg	0.249	0.163	0.202	0.010	0.115	0.085	0.064	0.059	0.053
Al	0.097	0.110	0.021	0.027	0.573	0.014	0.124	0.011	0.022
Zn	0.110	0.007	0.004	0.004	0.007	0.610	0.013	0.130	0.115
Cu	0.001	0.001	0.001	0.003	0.001	0.002	0.054	0.913	0.024
Ba	0.155	0.089	0.056	0.374	0.004	0.004	0.306	0.006	0.005
Pb	0.116	0.028	0.014	0.007	0.014	0.006	0.031	0.045	0.739

p is given by $\sum_t \tilde{y}_{pt} = \sum_t \sum_k \hat{\lambda}_{pkt}\hat{f}_{kt}$ where $\hat{\lambda}_{pkt}$ and $\hat{f}_{kt}$ is the median of the posterior draws for λ_{pkt} and f_{kt}, respectively. The proportion of total predicted mass for chemical species p attributed to source k is then,

$$\frac{\sum_t \hat{\lambda}_{pkt}\hat{f}_{kt}}{\sum_t \sum_k \hat{\lambda}_{pkt}\hat{f}_{kt}}.$$

For brevity, Table 3 displays the proportion of total predicted mass attributed to the nine sources for the 16 of the total 44 chemical species with the largest average concentrations (the full table is available as a supplemental file). For example, the first number in the first row of Table 3 indicates that the fitted model estimates 85.7% of the total predicted mass of SO_4 is attributed to Source 1. Using the values listed in Table 3 and the plots of the source contributions over time in Figure 5 below, labels can be applied to the 9 estimated sources by matching the proportion of total predicted mass to knowledge of known pollution sources of the surrounding areas. For example, the steel mill to the northeast of the receptor location is a known large contributor to iron pollution and Source 4 in Table 3 is the largest contributor to the total predicted mass of iron. Thus, Source 4 is labeled as "steel mill." Sources 1 and 2 can likewise be labeled summer and winter secondary as these are high contributors of SO_4 and NO_3 and the corresponding source contribution plots exhibit strong summer and winter peaks. Source 3 is identified as the mobile source due to high contributions of OC. Silicon and aluminum are mostly explained by Source 5 which corresponds to soil. Interestingly, note that the soil source has a large spike in July of 2002 which corresponds to the Saharan dust storm which had global impacts (Lee and Hopke 2006). Zinc, copper, and lead are largely contributed by Sources 6, 8, and 9, respectively. These sources are then labeled as the zinc, copper, and lead smelters to the southwest of the receptor site. Lastly, Source 7 only shows large contributions on July 4th and July 5th and is high in potassium indicating a fireworks source.

The sources identified by the DP model coincide directly with those identified by Lingwall and Christensen (2007). This similarity between the DP model and PMF is encouraging because vague prior distributions were used for the DP model and the same nine sources were identified. In Lingwall and Christensen (2007), the source profiles were identified by fitting different source apportionment models for $K = 3, \ldots, 13$ and the final estimate of $\mathbf{\Lambda}$ was constructed by concatenating profile estimates from the different models. Thus, the DP model adequately identified pollution sources while PMF required additional guidance and analysis to identify the same pollution sources.

As noted previously, the DP model is sufficiently flexible to capture certain aspects of seasonal variation in pollution sources. For example, consider Figure 4 which display the time plots of the six largest elements of $\boldsymbol{\lambda}_{kt}$ for the zinc smelter source. Specifically, notice that in Figure 4 the percentage of chlorine (Cl) changes season to season. Chlorine seems to be more prevalent in the winter than in the summer. The average value for chlorine in $\boldsymbol{\lambda}_{kt}$ is 0.043 and 0.082 for summer and winter, respectively. Using the DP model, the average value of chlorine in the zinc smelter profile over time is 0.06 compared to 0.05 when using PMF. Thus, the PMF estimate of λ_{pkt} for chlorine in the zinc smelter appears to be a seasonal average while the DP model identifies seasonal trends. This seasonal phenomenon has posed a problem for traditional approaches to source apportionment. The analysis of the St. Louis dataset by Lee, Hopke, and Turner (2006) removed chlorine from the dataset to avoid this phenomenon while, as previously mentioned, Lingwall and Christensen (2007) found a yearly average when estimating a constant profile. In contrast, the DP model is flexible enough to capture such atmospheric phenomena as the seasonal variations in chlorine.

Figure 5 displays time plots of the posterior medians of f_{kt} for the nine identified sources. The median source contribution estimates over time are as follows: summer secondary (6.21 $\mu g/m^3$), winter secondary (3.86 $\mu g/m^3$), mobile (2.65 $\mu g/m^3$), steel mill (0.56 $\mu g/m^3$), soil (0.33 $\mu g/m^3$), zinc smelter (0.30 $\mu g/m^3$), fireworks (0.20 $\mu g/m^3$), copper smelter (0.17 $\mu g/m^3$), and lead smelter (0.13 $\mu g/m^3$). Each of these

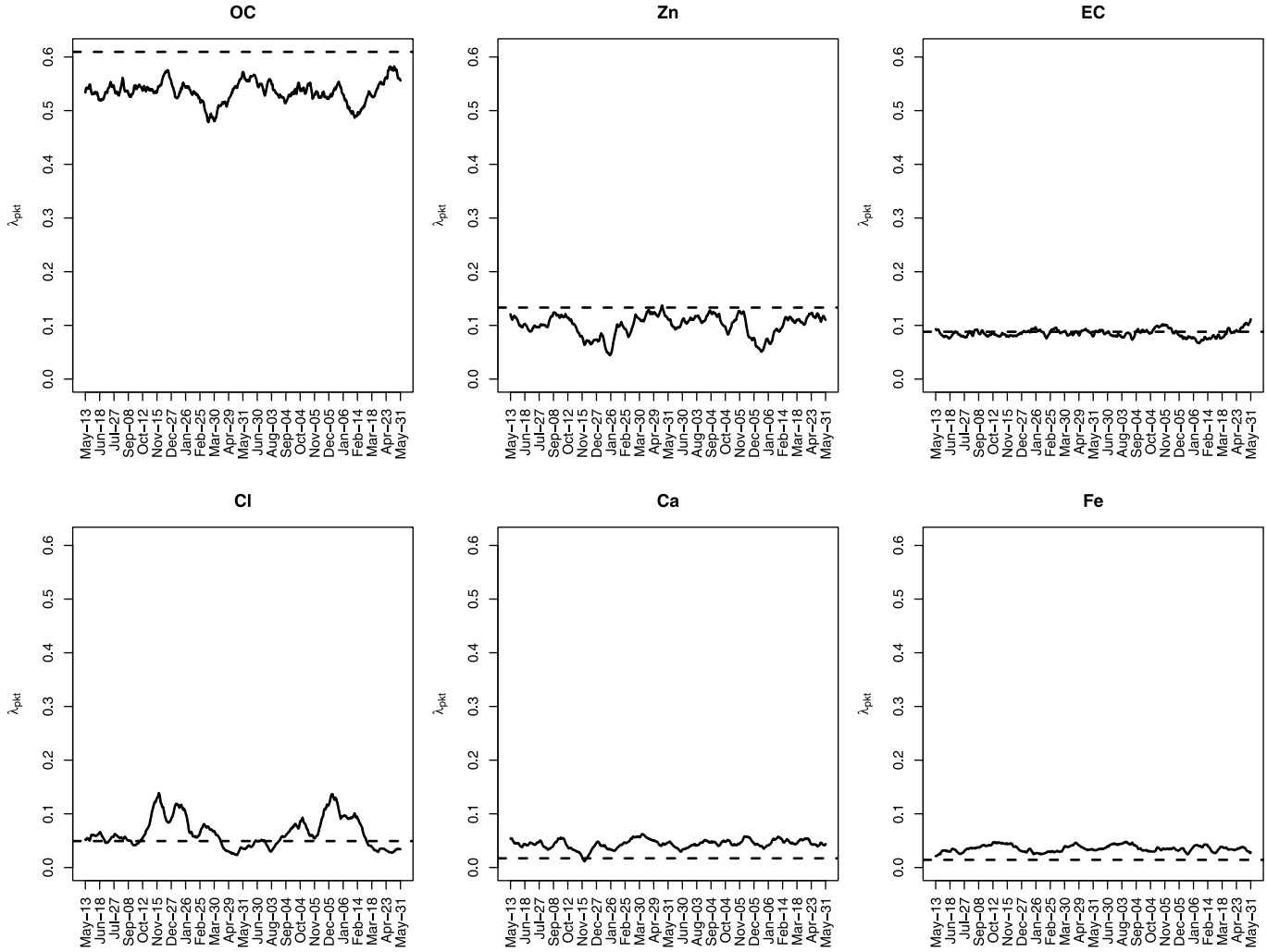

Figure 4. Time plot of the six largest elements of $\boldsymbol{\lambda}_{kt}$ for the zinc smelter profile as identified by the DP model. The dashed lines correspond to the time-constant PMF estimate.

medians are comparable to those obtained from Lingwall and Christensen (2007); however, the DP model has the added benefit of obtaining distributional results for all model parameters which is illustrated below.

Most of the identified sources (e.g., zinc smelter) correspond directly with a known pollution source. The summer and winter secondary sources, however, do not correspond with a specific pollution source but are sources used to explain seasonal variation in the chemical concentrations. The summer secondary source profile accounts for the fact that more of sulfate (SO_4) and sodium (Na) is seen during the summer months. Similarly, the winter secondary profile accounts for the higher concentrations of nitrate (NO_3) and elemental carbon (EC) during the winter months. By thus estimating the pollution sources, the DP model accounted for another known seasonal variation in chemical concentrations.

To demonstrate the advantage of using the DP model over PMF in obtaining distributional results for complex functions of model parameters, consider a regulatory body which has a goal to keep the daily auto emissions below 6.0 $\mu g/m^3$ for 90% of the measured days. Let Z represent the number of days that mobile emissions falls above 6.0 $\mu g/m^3$ in the two year period the St. Louis dataset was collected. The discrete probability distribution for Z, $p(Z|\mathbf{Y})$, is defined as

$$p(Z = z|\mathbf{Y}) = \int \mathrm{I}\left(\sum_{t=1}^{661} \mathrm{I}[f_{\text{auto},t} > 6.0] = z\right)\pi(\boldsymbol{\Theta}|\mathbf{Y})\, d\boldsymbol{\Theta}, \quad (14)$$

where $\mathrm{I}(\cdot)$ is an indicator function, $f_{\text{auto},t}$ is the mobile source contribution at time t, and $\pi(\boldsymbol{\Theta}|\mathbf{Y})$ is the posterior distribution of all model parameters ($\boldsymbol{\Theta}$). When using the DP model, the post burn-in draws from the MCMC are used to construct (14) via Monte Carlo integration. Figure 6 displays the model estimate of (14). Using Monte Carlo integration techniques, the complex quantity, $\Pr\{\text{Achieved Goal}\} = \Pr\{Z < 66.1 = 661 \times 0.1\} = \sum_{Z=0}^{\infty} \mathrm{I}(Z < 66.1) \times p(Z|\mathbf{Y})$ is estimated to be 0.86. If the regulatory body is satisfied with this probability of success then any policies implemented to achieve this goal should be continued; otherwise, the regulatory body would need to reevaluate its policies and reformulate a plan to reduce auto emissions.

6. CONCLUSIONS

In this article, an extension of the multivariate receptor model given by (1) was proposed where pollution source profiles are

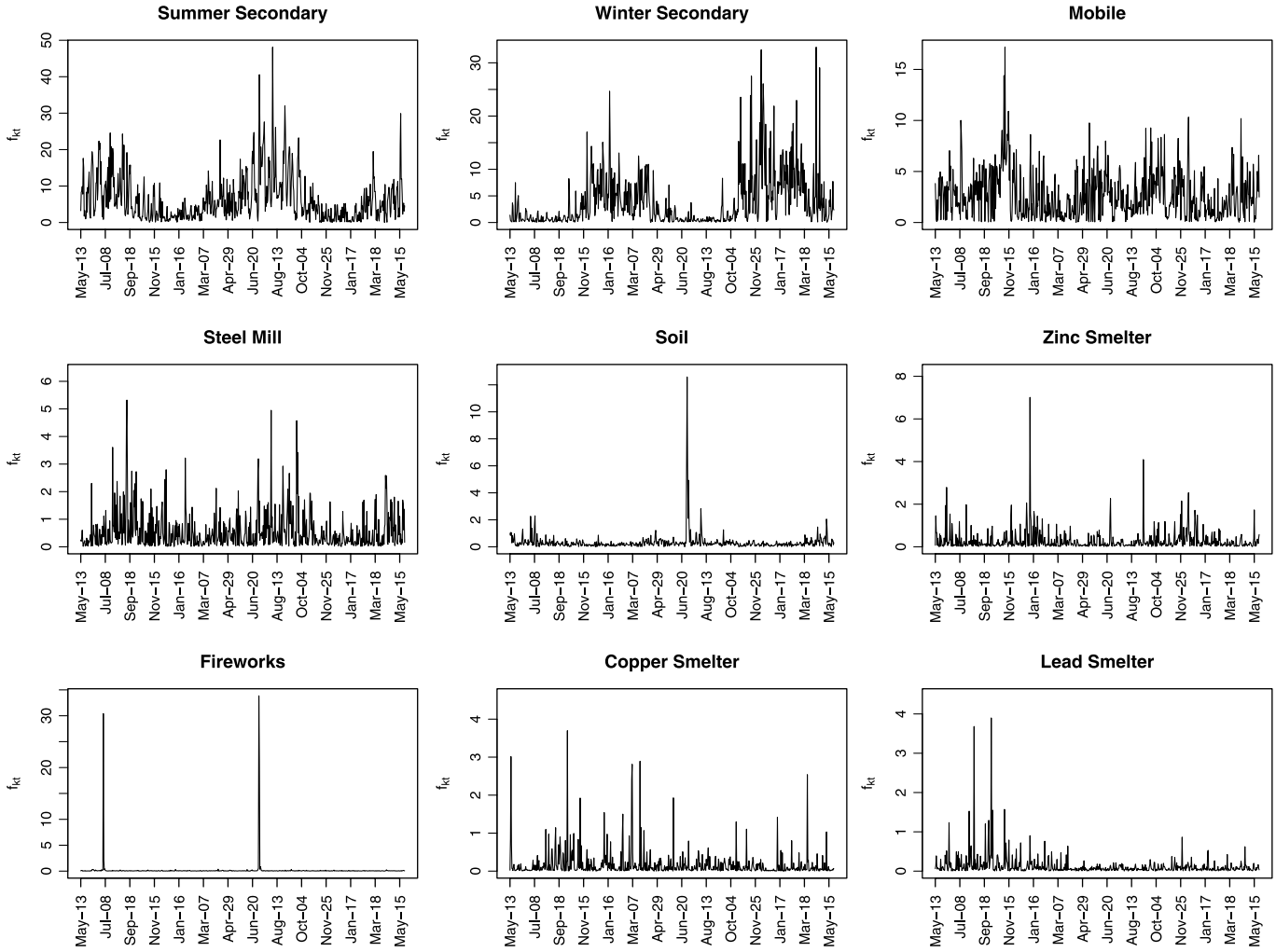

Figure 5. Time plot of f_{kt} for the nine identified sources as estimated by fitting the DP model to the St. Louis dataset.

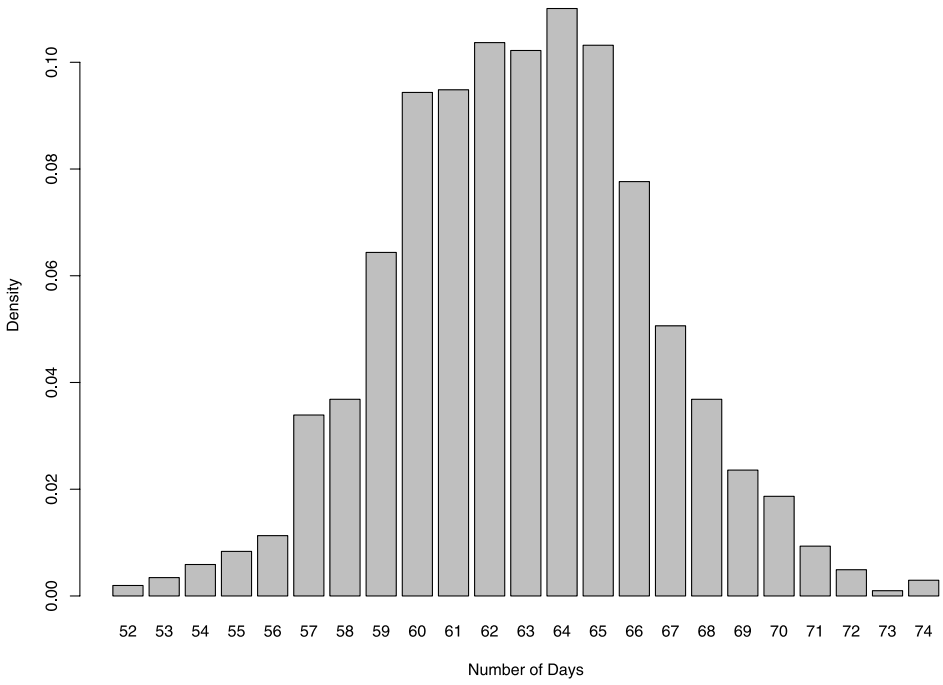

Figure 6. Distribution of the number of days mobile emissions exceeds 6.0 $\mu g/m^3$.

allowed to vary through time. Time-varying source profiles were shown to be empirically and physically justifiable. Furthermore, by allowing source profiles to vary through time, a degree of temporal dependence in pollution concentrations is accounted for. Model parameters were estimated via adaptive MCMC methods which provide draws from the joint posterior of all model parameters. These draws were shown to be useful in the estimation of complex functions of model parameters.

The methods proposed herein were found to reduce estimation error for $\boldsymbol{\lambda}_{kt}$ when source profiles were both time-varying and time-invariant as compared to PMF. When uncertainty associated with y_{pt} was small, the DP model reduced estimation error in f_{kt} as compared to PMF. However, when variation within y_{pt} was large, the DP model and PMF had comparable estimation error for f_{kt}. By applying the DP model to the St. Louis dataset, seasonal trends in chemical concentrations and $\boldsymbol{\lambda}_{kt}$ were discovered. Previous approaches to source apportionment either excluded chemicals with seasonal variation or, if included, estimates of λ_{pk} from these studies were shown to be a seasonal average. The flexibility of the DP model to estimate seasonal variations is a scientific asset when trying to understand the behavior of chemical species in ambient air.

The DP model presented here was developed to provide insight into the temporal evolution of source profiles where data series are long enough to warrant a suspicion of profile evolution. However, many datasets are collected over a 2–3 week period in a single season. The results in this article suggest that the source profiles change only slightly over such a short window. Hence, the assumption of constant source profiles for such datasets may be approximately correct rendering alternative methods such as Lingwall, Christensen, and Reese (2008) more appropriate for these cases.

As previously mentioned, Park, Guttorp, and Henry (2001) showed that source contributions f_{kt} and model errors e_{pt} also exhibit temporal correlation. As the main contribution of this article was to develop a model which allows for temporally correlated source profiles, the autocorrelation in f_{kt} and e_{pt} was ignored. Future approaches to multivariate receptor modeling should unite the methods proposed by Park, Guttorp, and Henry (2001) with the methods discussed in this article to completely model the temporal structure of y_{pt}.

A key assumption made throughout this article was that the degree of autocorrelation among all $\lambda_{pkt} \in \boldsymbol{\lambda}_{kt}$ was captured by the single smoothness parameter g_k. Figure 1 gives evidence against this assumption in that those chemicals with large concentrations display a greater degree of autocorrelation. To incorporate this, the use of a scalar precision parameter could be extended to vector valued as $\mathbf{g}_k = (g_{1k}, \ldots, g_{Pk})'$ where g_{pk} represents the degree of autocorrelation of chemical p in $\boldsymbol{\lambda}_{kt}$. Each g_{pk} would then need to be estimated via MCMC.

One interesting extension of the DP model, and source apportionment models in general, is the use of covariate information in identifying pollution sources and explaining the behavior of source contributions. For example, one would expect to see higher contributions from certain pollution sources on days where the wind carries pollutants from the source to the receptor site. Additionally, given wind direction, observing spikes in certain pollutants could provide some information regarding which direction pollution sources are relative to the receptor site. Methods which incorporate such covariate information are a current area of active research.

APPENDIX: MCMC DETAILS

Estimation of the DP model requires the calculation of the joint posterior distribution of all model parameters given the data. To acheive this, this article used a Metropolis within Gibbs adaptive MCMC algorithm. Specifically, a Gibbs sampler (Casella and George 1992) was used to sample each model parameter from its complete conditional distribution (the distribution of the parameter given all the other parameters and the data) in sequential order. The complete conditional distributions for the DP model are

$$\begin{aligned}
p\big(\boldsymbol{\lambda}_{kt}|\boldsymbol{\Theta}_{-\lambda_{kt}}, \mathbf{Y}\big) &\propto L(\boldsymbol{\lambda}_{kt}) p\big(\boldsymbol{\lambda}_{k(t+1)}|\boldsymbol{\lambda}_{kt}, g_k\big) p\big(\boldsymbol{\lambda}_{kt}|\boldsymbol{\lambda}_{k(t-1)}, g_k\big) \\
&\qquad \forall k, t = 1, \ldots, T-1, \\
p\big(\boldsymbol{\lambda}_{kT}|\boldsymbol{\Theta}_{-\lambda_{kT}}, \mathbf{Y}\big) &\propto L(\boldsymbol{\lambda}_{kT}) p\big(\boldsymbol{\lambda}_{kT}|\boldsymbol{\lambda}_{k(T-1)}, g_k\big) \qquad \forall k, \\
p\big(\boldsymbol{\lambda}_{k0}|\boldsymbol{\Theta}_{-\lambda_{k0}}, \mathbf{Y}\big) &\propto p(\boldsymbol{\lambda}_{k1}|\boldsymbol{\lambda}_{k0}, g_k) p(\boldsymbol{\lambda}_{k0}) \qquad \forall k, \\
p\big(f_{kt}|\boldsymbol{\Theta}_{-f_{kt}}, \mathbf{Y}\big) &\propto L(f_{kt}) p(f_{kt}) \qquad \forall k, t,
\end{aligned}$$

and

$$p\big(g_k|\boldsymbol{\Theta}_{-g_k}, \mathbf{Y}\big) \propto \left[\prod_{t=1}^{T} p\big(\boldsymbol{\lambda}_{kt}|\boldsymbol{\lambda}_{k(t-1)}, g_k\big)\right] p(g_k) \qquad \forall k,$$

where $\boldsymbol{\Theta}_{-\theta_i}$ represents the set of all model parameters *excluding* the parameter θ_i,

$$\begin{aligned}
L(\boldsymbol{\lambda}_{kt}) \propto \Bigg[&\prod_{p=1}^{P} \frac{1}{y_{pt}} \\
&\times \exp\Big\{-\big(\ln(y_{pt}) - \ln(\boldsymbol{\Lambda}_t(p) \times \mathbf{f}_t) + 0.5 \ln(w_{pt}^2 + 1)\big) \\
&/(2 \ln(w_{pt}^2 + 1))\Big\}\Bigg],
\end{aligned}$$

$L(f_{kt}) \propto L(\boldsymbol{\lambda}_{kt})$, are the likelihoods for $\boldsymbol{\lambda}_{kt}$ and f_{kt}, $p(\boldsymbol{\lambda}_{kt}|\boldsymbol{\lambda}_{k(t-1)}, g_k)$ is given by (7), $p(\boldsymbol{\lambda}_{k0})$ is given by (8), $p(f_{kt})$ is given by (9), and $p(g_k)$ is the log-normal prior distribution for g_k. Given starting values $\boldsymbol{\lambda}_{10}^{(0)}, \ldots, \boldsymbol{\lambda}_{KT}^{(0)}, f_{11}^{(0)}, \ldots, f_{KT}^{(0)}, g_1^{(0)}, \ldots, g_K^{(0)}$, the Gibbs sampling algorithm proceeds as follows:

1. Set $r = 1$.
2. For $k = 1, \ldots, K$ and $t = 1, \ldots, T$, sample $\boldsymbol{\lambda}_{kt}^{(r)}$ from $p(\boldsymbol{\lambda}_{kt}|\boldsymbol{\Theta}_{-\lambda_{kt}}^{(r)}, \boldsymbol{\Theta}_{-\lambda_{kt}}^{(r-1)}, \mathbf{Y})$ where $\boldsymbol{\Theta}_{-\theta_i}^{(r)}$ represents the set of parameters for which the rth value has already been drawn in the algorithm *excluding* the parameter θ_i.
3. For $k = 1, \ldots, K$ and $T = 1, \ldots, T$, sample $f_{kt}^{(r)}$ from $p(f_{kt}|\boldsymbol{\Theta}_{-f_{kt}}^{(r)}, \boldsymbol{\Theta}_{-f_{kt}}^{(r-1)}, \mathbf{Y})$.
4. For $k = 1, \ldots, K$, sample $g_k^{(r)}$ from $p(g_k|\boldsymbol{\Theta}_{-g_k}^{(r)}, \boldsymbol{\Theta}_{-g_k}^{(r-1)}, \mathbf{Y})$.
5. Repeat steps 2–4 for $r = 2, \ldots, R$.

As $R \to \infty$, the distribution of the generated parameters $\boldsymbol{\Theta} = \{\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_T, \mathbf{F}, g_1, \ldots, g_K\}$ tends to the joint posterior distribution and statistical inference can be performed on the resulting draws obtained from the algorithm.

Because each complete conditional distribution given above is known only up to a constant of proportionality, the Metropolis–Hastings algorithm (Chib and Greenberg 1995, see) was

used within the Gibbs sampler to sample from each complete conditional distribution. Specifically, values for $f_{kt}^{(r)}$ and $g_k^{(r)}$ were drawn using the random walk Metropolis algorithm where the proposal distribution is a normal distribution centered at $f_{kt}^{(r-1)}$ and $g_k^{(r-1)}$. In contrast, $\boldsymbol{\lambda}_{kt}^{(r)}$ was obtained using the Metropolis–Hastings algorithm where proposal values $\boldsymbol{\lambda}_{kt}^{*} \sim \text{DIR}[l_{kt}\boldsymbol{\lambda}_{kt}^{(r-1)}]$ where l_{kt} controls the scale of the proposal distribution.

To achieve good mixing properties, an adaptive MCMC algorithm was used to adapt the scale of each proposal distribution. Specifically, the acceptance rate of the chain was monitored in windows of size 100. If the algorithm accepted more than 60 of 100 proposals within a window, the scale of the corresponding proposal distribution was increased by a factor of 10%. Alternatively, if the algorithm accepted less than 20 of 100 proposals within a window, the scale of the corresponding proposal distribution was decreased by a factor of 10%. In this way, the adaptation is done to acheive an acceptance rate between 20% and 60%. The adaptation of the algorithm was done during the burn-in phase only. Thus, the limiting behavior of the resulting adaptive MCMC algorithm remains intact.

SUPPLEMENTAL MATERIALS

Explained Mass Table: Full table of explained chemical mass for St. Louis dataset. (ExplainedMassFullTable.txt, text file)

ACKNOWLEDGMENTS

This work was supported by the STAR Research Assistance Agreement RD-83216001-0 awarded by the U.S. Environmental Protection Agency. The article has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and the EPA does not endorse any products or commercial services mentioned in this publication. The authors would like to acknowledge Jay Turner for supplying the dataset used in this article as well as providing helpful comments. The authors also thank the editors and referees for many helpful comments that substantially improved the article.

[Received July 2008. Revised September 2009.]

REFERENCES

Bandeen-Roche, K. (1994), "Resolution of Additive Mixtures Into Source Components and Contributions: A Compositional Approach," *Journal of the American Statistical Association*, 89, 1450–1458. [68]

Billheimer, D. (2001), "Compositional Receptor Modeling," *Environmetrics*, 12, 451–467. [68]

Britt, H., and Luecke, R. (1973), "The Estimation of Parameters in Nonlinear, Implicit Models," *Technometrics*, 15, 233–247. [68]

Cargnoni, C., Muller, P., and West, M. (1997), "Bayesian Forecasting of Multinomial Time Series Through Conditionally Gaussian Dynamic Models," *Journal of the American Statistical Association*, 92, 640–647. [71]

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174. [78]

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis–Hasting Algorithm," *The American Statistician*, 49, 327–335. [78]

Christensen, W. F., and Sain, S. R. (2002), "Accounting for Dependence in a Flexible Multivariate Receptor Model," *Technometrics*, 44, 328–337. [68]

Christensen, W. F., Schauer, J. J., and Lingwall, J. W. (2006), "Iterated Confirmatory Factor Analysis for Pollution Source Apportionment," *Environmetrics*, 17, 663–681. [68]

Cochran, W. G., and Cox, G. M. (1957), *Experimental Designs*, New York: Wiley. [72]

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley. [68]

Grunwald, G. K., Raftery, A. E., and Guttorp, P. (1993), "Time Series of Continuous Proportions," *Journal of the Royal Statistical Society, Ser. B*, 55, 103–116. [71]

Henry, R. (1997), "History and Fundamentals of Multivariate Air Quality Receptor Models," *Chemometrics and Intelligent Laboratory Systems*, 37, 525–530. [68]

Lee, J. H., and Hopke, P. K. (2006), "Apportioning Sources of $PM_{2.5}$ in St. Louis, MO Using Speciation Trends Network Data," *Atmospheric Environment*, 40, S360–S377. [75]

Lee, J. H., Hopke, P. K., and Turner, J. R. (2006), "Source Identification of Airborne $PM_{2.5}$ at the St. Louis–Midwest Supersite," *Journal of Geophysical Research*, 111, d10S10. [70,74,75]

Lingwall, J. W., and Christensen, W. F. (2007), "Pollution Source Apportionment Using a priori Information and Positive Matrix Factorization," *Chemometrics and Intelligent Laboratory Systems*, 87, 281–294. [68,74-76]

Lingwall, J. W., Christensen, W. F., and Reese, C. S. (2008), "Dirichlet Based Bayesian Mulitvariate Receptor Modeling," *Environmetrics*, 19, 618–629. [68,78]

Lopes, H. F., and West, M. (2004), "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica*, 14, 41–67. [68]

Marmur, A., Mulholland, J. A., and Russell, A. G. (2007), "Optimized Variable Source-Profile Approach for Source Apportionment," *Atmospheric Environment*, 41, 493–505. [68]

Miller, M., Friedlander, S., and Hidy, G. (1972), "A Chemical Element Balance for the Pasadena Aerosol," *Journal of Colloid Interface Science*, 39, 65–176. [67,68]

Paatero, P., and Tapper, U. (1994), "Positive Matrix Factorization: A Nonnegative Factor Model With Optimal Utilization of Error Estimates of Data Values," *Environmetrics*, 5, 111–126. [68]

Park, E. S., Guttorp, P., and Henry, R. C. (2001), "Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC," *Journal of the American Statistical Association*, 96, 1171–1183. [68-71,78]

Park, E. S., Oh, M.-S., and Guttorp, P. (2002), "Multivariate Receptor Models and Model Uncertainty," *Chemometrics and Intelligent Laboratory Systems*, 60, 49–67. [68]

Park, E. S., Spiegelman, C. H., and Henry, R. C. (2002), "Bilinear Estimation of Pollution Source Profiles and Amounts by Using Multivariate Receptor Models," *Environmetrics*, 13, 775–798. [68]

Thurston, G. D., and Spengler, J. D. (1985), "A Quantitative Assessment of Source Contributions to Inhalable Particulate Matter Pollution in Metropolitan Boston," *Atmospheric Environment*, 19, 9–17. [68]

Watson, J. G., Cooper, J. A., and Huntzicker, J. J. (1984), "The Effective Variance Weighting for Least Squares Calculations Applied to the Mass Balance Receptor Model," *Atmospheric Environment*, 18, 1347–1355. [68]

West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Linear Models*, New York: Springer. [70]