Spatial variable selection methods for investigating acute health effects of fine particulate matter components

Laura F. Boehm Vock
St. Olaf College

Brian J. Reich, Montserrat Fuentes
North Carolina State University

and

Francesca Dominici
Harvard University

Summary: Multi site time series studies have reported evidence of an association between short term exposure to particulate matter (PM) and adverse health effects, but the effect size varies across the United States. Variability in the effect may partially be due to differing community level exposure and health characteristics, but also due to the chemical composition of PM which is known to vary greatly by location and time. The objective of this paper is to identify particularly harmful components of this chemical mixture. Because of the large number of highly-correlated components, we must incorporate some regularization into a statistical model. We assume that, at each spatial location, the regression coefficients come from a mixture model with the flavor of stochastic search variable selection, but utilize a copula to share information about variable inclusion and effect magnitude across locations. The model differs from current spatial variable selection techniques by accommodating both local and global variable selection. The model is used to study the association between fine PM (PM$_{<2.5\ \mu m}$) components, measured at 115 counties nationally over the period 2000-2008, and cardiovascular emergency room admissions among Medicare patients.

Key words: Bayesian modeling; Fine particulate matter; Stochastic search variable selection; Spatial data.
1. Introduction

Particulate matter (PM) is a complex mixture of airborne particles from both primary and secondary anthropogenic and natural sources. Major sources include combustion of fossil fuels and biomass, dust from industrial, construction, and mining operations, wildfires, and lightning strikes (Schlesinger et al., 2006). Particles are either emitted directly into the air (primary pollution) or formed by chemical interactions of gases and primary pollutant particles in the air (secondary pollution). Ambient levels of this complex mixture are currently regulated by the Environmental Protection Agency (EPA) based solely on particle size, not chemical composition or source. Yearly average and hourly maximum levels of the total amount of both coarse (PM10; particles $< 10\mu m$ in aerodynamic diameter) and fine (PM2.5; particles $< 2.5\mu m$) are restricted.

Numerous studies have quantified the relationship of fine particulate matter exposure and human health using multi-site time series studies (Schwartz et al., 1996; Dominici et al., 2000, 2006; Choi et al., 2009) comparing day to day changes in ambient measures of particulate matter to day to day changes in hospital admissions and mortality data at the county and city level. Both spatial and temporal heterogeneity in health effect estimates have been observed (Bell et al., 2007, 2008; Zhou et al., 2011; Ito et al., 2011). We hypothesize that spatial variation in the estimated health effects could be due to differing chemical composition of PM2.5 across space (Fuentes et al., 2006; Bell et al., 2007; Peng et al., 2009; Bell et al., 2009; Levy et al., 2012). While PM2.5 is currently regulated by total mass, deeper understanding of the toxicity of the PM2.5 chemical composition could lead to air pollution regulations that are more targeted to the most harmful emission sources.

A growing body of literature has investigated the potential health impact of specific PM2.5 components. While carbon fractions, including elemental carbon, black carbon, and organic carbon matter, are shown to have positive associations in a variety of health studies, no
components have been ruled out in all studies (Rohr and Wyzga, 2012). Even looking at individual chemical components does not eliminate between site variability in the health effect estimates. There is some variability in particle size within component depending on the source (Schlesinger, 2007), and also in particle chemistry and acidity, as these components may be part of larger molecules (Schlesinger et al., 2006). Site-specific effects of individual components may also vary due to population health characteristics which modify susceptibility or by lifestyle or housing characteristics which modify exposure to ambient air (Dominici et al., 2002, 2003; Zeka et al., 2005). For example, Dai et al. (2014) find PM2.5 health effects are greater in counties with higher rates of smoking and heavy alcohol use. However, Krall et al. (2013) did not find significant regional differences in mortality effects of PM2.5 constituents using single pollutant models.

Previous epidemiological and toxicological studies investigating components of fine particulate matter tend to focus on a few pollutants or pollutant groups. Peng et al. (2009) investigate the relationship of the seven most massive PM2.5 components and CVD hospitalizations. They use a hierarchical model to estimate national effects, but assume the effects at every site are independent. Some city-specific analyses investigate a larger number of pollutants (Ito et al., 2011; Zhou et al., 2011), but these results are hard to generalize to other locations.

To our knowledge, a model which accomplishes variable selection both locally (within a spatial location) and globally (on average for all the locations) across all sites within one model has not previously been attempted. Here we extend previous methods for multi-site variable selection to identify components of PM2.5 which may be important for health outcomes at all or specific sites. Our flexible spatial approach facilitates global and local selection simultaneously, unlike previous work. Reich et al. (2010) consider variable selection approaches for multiple predictors globally, that is, included or excluded from the model at all
Spatial variable selection methods for investigating acute health effects of fine particulate matter components

sites. The coefficients for the included variables are allowed to vary by site with spatial priors. Smith and Fahrmeir (2007) propose local selection for a single predictor. The predictor may be included in the regression at some sites and not others. This is accomplished by treating the coefficient of interest as the product of a spatially correlated binary random field and an independent Gaussian field. Similar work with multiple predictors in generalized linear models have been undertaken by Lum (2012) and Scheel et al. (2013). In all these works, the Gaussian field assumes the magnitudes of the coefficients are independent across sites, and there is no cross dependence between the magnitude and inclusion probability implied in the prior. We extend this approach to include multiple predictors and spatial correlation in the effect size, while simultaneously allowing global selection as a special case. Rather than treat each coefficient as the product of a binary field and a continuous field, we utilize a Gaussian copula to create a smooth predictive surface.

The key difference between the proposed method and the global methods of Reich et al. (2010) is that it addresses not only the question of which covariates are important, but also where the covariates are important. There are many cases when a practicing statistician would be interested in separating these two questions. For example, in an imaging study such as Smith and Fahrmeir (2007) one may attempt to find a particular spatial region of the brain which is affected by a covariate. Another example is meteorology, where certain factors may be important only in some regions because of different regional climate systems or regimes. In the air pollution application of this paper, the objective is to identify subregions where different pollutants are associated with adverse health outcomes. These difference may arise from different sources of pollution or different characteristics of the at-risk population. In summary, the proposed method is best suited to cases where covariates are thought to have an effect only in subregions of the overall spatial domain.

This spatial variable selection method facilitates a more comprehensive analysis of PM2.5
constituent data than previously attempted. We include 22 components of PM2.5 measured in 115 counties across the United States. For these data, we find that the spatial model without variable selection produces highly variable effect estimates across space and component, and fails to identify any particularly harmful components. By adding variable selection to the model, we are able to isolate elemental carbon as the component with largest effect. Compared to a non-spatial analysis, our spatial analysis provides more precise effect estimates, and reveals strong regional variation in the strength of this association, with the largest estimates in the Northeast, Midwest, and Pacific Coast.

2. Data

Of the approximately 50 speciated PM2.5 components measured by the EPA, we selected \( p = 22 \) components of interest. Each contributes at least 1% of total mass to PM2.5, or the literature has suggested a potential link with health outcomes, or both. The components and summary statistics are shown in Table 1. These speciated PM measurements are taken from the EPA’s Air Quality System (AQS) and AirExplorer databases (www.epa.gov/ttn/airs/airsaqs/, www.epa.gov/airexplorer/). The AQS data include raw monitor values and daily averages, while AirExplorer is a processed data product designed for use by health and epidemiology research. For twenty of the components we use the AQS data. Because of a high proportion of missingness for elemental carbon (EC) and organic carbon matter (OCM) in the AQS database, we use the AirExplorer data for these components. Any observations below the lower limit of detection are recorded as one half the detection limit. Following Peng et al. (2009), for counties that had more than one active monitor on a given day, an average was taken using 10% trimmed mean if more than 10 stations; for 3–10 stations, minimum and maximum values were excluded from the mean; and for 2 stations, we use the mean. All components are measured in \( \mu g/m^3 \) except EC, which is measured in inverse megameters, a measure of light extinction in haze.
We only used information from non-source-oriented monitors, and exclude values flagged by the EPA for data quality issues. Source-oriented monitors are placed with the intention of monitoring a known large pollutant source, and may not be in a populated area. To avoid biased pollution measurements, we exclude these and focus on non-source-oriented monitors, which are placed with the purpose of estimating the exposure in populated areas. Any days missing pollutant information were excluded. We include 117 counties in the US with at least 100,000 residents and PM2.5 components monitors active on at least 150 days in the time period 2000-2008. Of these we exclude two California counties with data quality issues. In these counties half the pollution measurements were 1000 times larger than expected based on nearby counties and measurements on preceding days. In total we include 115 counties.

As recommended by Gelfand et al. (2003) covariates are scaled, but not centered. We use the 90th percentile value to scale rather than standard deviation because of the skewness of the pollution data. We note that other scaling factors could be used, such as standard deviation or interquartile range (IQR), and that this is equivalent to putting different prior variances on the coefficients for each pollutant, and that scaling is important when pollutants are present in much different concentrations. As in Peng et al. (2009), we removed extreme pollution values from both the simulation study and analysis, where “extreme” is any value more than double the second highest value for that pollutant in that county. This results in a removal of approximately 0.4% of the observed data. We also conducted the data analysis with the extreme values as part of our sensitivity analysis (see Supplementary Material).

The health data includes Medicare beneficiary enrollment and Medicare Part-A inpatient records, aggregated to daily county totals. We count the number of patients hospitalized with a principal ICD-9 diagnosis code related to cardiovascular disease (CVD). These include heart failure, ischemic heart disease, and cerebrovascular events among others. The average CVD
hospitalization rates vary greatly by county, ranging from 8.5 to 32 per 100,000 Medicare enrollees per day.

[Table 1 about here.]

3. Statistical model

We observe $Y_t(s)$ CVD cases at county $s$ on day $t$, in addition to a vector of pollutants, $x_t(s)$, and confounding variables $z_t(s)$. The confounding variables include indicator functions for the day of the week and smooth functions of time, temperature, and dewpoint. To adjust for long term and seasonal trends, we include natural cubic spline smoothing function of time with 7 df per year as a predictor in the model. We also include indicators for the day of the week, and splines for daily average temperature (6 df) and the 3-day lag mean dewpoint (3 df) to control for confounding with weather. We chose these numbers of degrees of freedom for consistency with Dominici et al. (2002). Other work, including Peng et al. (2009) use different numbers of degrees of freedom. Details of sensitivity analysis are included in the supplementary material.

We utilize the spatially varying coefficients model described by Gelfand et al. (2003),

$$Y_t(s)|\beta(s), \eta(s) \sim \text{Poisson}\left[N_t(s)\exp\{\beta_0(s) + x_t(s)'\beta(s) + z_t(s)'\eta(s)\}\right]$$

(1)

where $N_t(s)$ is a known offset for the number of Medicare patients at county $s$ on day $t$ and $\beta_0(s)$ is a spatially varying intercept. We include spatially varying coefficients $\beta(s) = [\beta_1(s), \ldots, \beta_p(s)]^T$ and $\eta(s) = [\eta_1(s), \ldots, \eta_q(s)]^T$. Note that the dimension of $\eta(s)$, $q_s$, depends on the county $s$ due to the use of 7 knots per year for the spline function of time, and that each county may be observed over a different time interval. We control for unmeasured confounding independently within each location and then borrow information across locations for the health estimates. Therefore we select a low precision normal density as the prior for the confounding effects $\eta(s)$, which is independent across counties. We let
$\beta_k(s)$ denote the coefficient on pollutant $k$, for $k = 1, \ldots, p$, and county $s$, for $s = 1, \ldots, n$. The prior on $\beta_k(s)$ will imply both spatial correlation and variable selection.

### 3.1 Spatial variable selection model

To facilitate variable selection, the marginal prior distribution for each coefficient $\beta_k(s)$ is a mixture density,

$$f_k[\beta_k(s)] = \pi_k N(\alpha_k, \omega_k^2) + (1 - \pi_k) N(0, \omega_k^2/C),$$

where $N(m, v)$ is the Gaussian density function with mean $m$ and variance $v$. Global behavior is exhibited when $\pi_k = 1$ or 0. Coefficients will be smoothed either toward $\alpha_k$, the overall mean when $\beta_k$ is important, or toward 0. Thus $\pi_k \alpha_k$ is the overall mean effect across all counties. Variability is controlled by $\omega_k$, with $C$ representing the ratio of variance for coefficients “in” and “out” of the model. One interpretation of $C$ is that if $\beta_k(s)$ is within $\pm 3\omega_k/\sqrt{C}$ of 0 it may be safely replaced by zero (George and McCulloch, 1997). If this tuning parameter $C = \infty$, the second part of the mixture distribution is a point mass at zero. This is the formulation of SSVS used by Smith and Fahrmeir (2007) and Reich et al. (2010), in which the coefficients are modeled as the product of binary and continuous fields. In the simulation study, we let $C = 100$, while goodness of fit criteria are used to select $C = 400$ for the data analysis in Section 5.

In our model we apply a copula (Nelsen, 1999; Sklar, 1973, 1953) to the marginal distribution in (2). A copula is a general technique for modeling dependence between random variables while preserving desired marginal distributions. Many copulas have been proposed to capture various forms of dependence, including the class of Archimedean copulas and copulas constructed from multivariate densities such as Gaussian and $t$ (Nelsen, 1999). We select a Gaussian copula because the vast majority of spatial modeling deals with dependence through the covariance function of a Gaussian process, and the Gaussian copula allows us to leverage these models for our spatial variable selection model. To implement the copula
model, we introduce latent variables, $\theta_k(s)$, which follow a mean zero Gaussian process with correlation function $\rho(s, s')$. Here we use an exponential spatial correlation function,

$$\rho(s, s') = \text{cor} [\theta_k(s), \theta_k(s')] = \exp (-||s - s'||/r).$$

The range parameter $r$ has the interpretation that at distance $3r$ correlation is about 0.05. Although we choose an exponential correlation structure, any correlation function may be used, including a generalization to non-stationary areal correlation structure such as that implied by a conditional autoregressive model (CAR). We fix the variance of $\theta_k$ to one for identifiability purposes in the Gaussian copula. We write a Gaussian process with mean $m$, standard deviation $v$, and exponential spatial correlation with range $r$ as GPex($m$, $v$, $r$). Hence, $\theta_k(s) \sim \text{GPex}(0, 1, r)$. We force $\beta_k(s)$ to have the desired marginal distribution by the transformation

$$\beta_k(s) = F_k^{-1}\{\Phi[\theta_k(s)]\},$$

where $\Phi$ is the standard normal cdf and $F_k$ is the marginal cdf of $\beta_k(s)$ defined in (2) and dependent on $\pi_k$, $\alpha_k$, and $\omega_k$.

3.2 A comparison of copula and other model solutions

As noted previously, the usual SSVS formulation of the model in (2) is to set $C = \infty$ and write $\beta_k(s)$ as the product of a binary and a continuous field. That is, let $\beta_k(s) = \gamma_k(s)B_k(s)$ where $\gamma_k(s)$ is a binary spatial process and $B_k(s)$ is a continuous process. In the introduction, an alternative approach is to define each coefficient as the product of a binary field and a continuous field. In the local approach by Smith and Fahrmeir (2007) the binary indicator varies spatially according to an Ising (auto-logistic) prior, which is a binary analog to the Conditionally Autoregressive (CAR) model used for continuous data measured on a lattice. The continuous part, measuring the effect size, is assumed to vary across sites, but no spatial correlation is directly imposed in the prior. Using our notation, we can write $\gamma_k(s) \sim \text{Ising},$
and $B_k(s) \overset{iid}{\sim} N(0, \omega^2)$. The model described by Smith and Fahrmeir (2007) has $p = 1$, while Lum (2012) and Scheel et al. (2013) extend to multiple predictors. In Smith and Fahrmeir (2007) the approach is applied to functional MRI data in which spatial indication of importance through $\gamma_k(s)$ was of primary interest. The spatial variable selection approach we propose encourages sharing of information across sites by centering the distribution of $\beta_k(s)$ around $\alpha_k$ and inducing spatial correlation through the copula.

While possible to induce spatial correlation through $B_k(s)$, treating $\beta_k(s)$ as the product of two fields or variables requires choosing a correlation and cross-correlation structure for the two processes. The copula naturally incorporates the intuition that if a component has no effect within a given county, nearby counties are unlikely to have large coefficients and creates a continuous prior surface. The realization of the spatial process on a spatial grid in Supplemental Materials A shows large areas with $\beta(s) = 0$ (light blue) corresponding to regions of null effects. The effect is near zero for sites near null regions, and $\beta(s)$ varies smoothly across sites in non-null regions.

The global approach to spatial variable selection as implemented by Reich et al. (2010) fosters sharing of information across sites, but assumes the same set of predictors are included in the model at every site. Using our notation, their model can be summarized as $\gamma_k(s) = \gamma_k \sim \text{Bernoulli}(\pi)$ and $B_k(s)$ is a Gaussian Process with spatial correlation. The Reich et al. (2010) model also includes a second binary process which determines whether the variance in the Gaussian process is non-zero. That is, coefficients are equal to zero at all sites, equal to $\alpha_k$ at all sites, or are spatially varying with mean $\alpha$. Because this approach is global, it does not allow for the case that pollutants would have zero coefficients at some sites and non-zero effects at others.

The SpVS approach we implement here allows for both local and global inferences while incorporating spatial smoothing through a single continuous process. While the non-copula
approaches can be written with conjugate full-conditional Gibbs sampling updates for all parameters conditional on $\beta$, the copula approach is still computationally feasible through MCMC and code is provided in the supplementary material.

4. Simulation study

In the simulation study, we will investigate the performance of our model in terms of mean absolute deviation (MAD), power, Type I error, and ability to correctly identify null coefficients compared to a spatial model without variable selection and a variable selection model without spatial correlation. Because the primary focus of this study is to evaluate local variable selection, we compare performance for varying degrees of spatial correlation in the site-specific coefficients $\beta_k(s)$.

We generate coefficients $\beta_k(s)$ from multivariate normal latent variables $\theta_k(s)$ with exponential correlation using the transformation in (3) where $F$ is the cdf of (2) with $C = \infty$. Though we fit the continuous version of the model where $C < \infty$, we chose to generate the data for the $C = \infty$ case so that the true values of the coefficients will be exactly zero when they are out of the model, whether globally or locally.

Each model has $p = 9$ covariates. By varying $\pi_1, \ldots, \pi_9$ we determine the behavior for each of the nine pollutant coefficient vectors. We will define “global” behavior, as being generated with $\pi_k = 1$ and thus $\beta_k(s)$ are drawn from a multivariate normal distribution, $\beta_k(s) \sim N(\alpha_k, \omega_k^2 \Sigma)$, where $\Sigma$ is an exponential correlation matrix. Globally “null” covariates have $\pi_k = 0$ and hence $\beta_k(s) = 0$ for all $s$. Covariates exhibiting “local” behavior will have a non-zero mean $\alpha_k$, but $0 < \pi_k < 1$ so that the marginal distribution at each site is a mixture distribution.

For computational purposes, we restrict the simulation study to 48 counties in the Eastern United States. The maximum distance between these counties is 1360 km. We include the components which contribute the greatest mass to total PM2.5: sulfate, nitrate, elemental
carbon, organic carbon, silicon, and sodium, as well as arsenic, bromine, and calcium. By using the actual pollutant data to generate responses, we evaluate the models under the more realistic scenario in which covariates of interest are moderately to strongly correlated within site across time. Table 2(a) specifies the components and true parameters $\alpha_k$ and $\pi_k$ used to generate the data. The overall mean $\alpha$ is 0, 0.05, or 0.10, and the inclusion probability $\pi_k$ is 0, 0.3, 0.7, or 1, indicating null, local, or global inclusion. For all coefficients we set $\omega_k = 0.025$. We use the Poisson regression model (1) and the observed pollution data $x_t(s)$ to generate values of $Y_t(s)$ based on our draws of $\beta(s)$. That is, for each of $M$ datasets we generate $\beta_k(s)$ and then $Y_t(s)|\beta(s), x_t(s)$. For simplicity, we do not include any confounding variables in the simulation study.

Following the procedures outlined above, we generate $M = 50$ datasets using each of the following designs for the correlation in $\beta_k(s)$:

(1) **No Spatial Correlation**: The true coefficients for each county are independent.

(2) **Moderate Spatial Correlation**: The true coefficients for each county are exponentially spatially correlated with range $r = 60$ km (effective range 180 km).

(3) **Strong Spatial Correlation**: The true coefficients for each county are exponentially spatially correlated with range $r = 240$ km (effective range 720 km).

In all cases we also include a random intercept $\beta_0(s)$, with $E[\beta_0(s)] = -8$ yielding an approximate average risk similar to the median risk observed. For the model with no spatial correlation, the intercepts are also uncorrelated. The intercept in designs 2 and 3 are exponentially spatially correlated with $r = 120$ km, which implies an effective range of 360 km.

The standard deviation of $\beta_0(s)$, $\omega_0 = 1$, is similar to the between county standard deviation of average CVD rate.

We investigate the following three models:
(1) **Spatial Model** (Sp): \( \beta_k(s) \sim \text{GPex}(\alpha_k, \omega_k, r) \), without variable selection (i.e., \( \pi_k = 1, k = 1, \ldots, p \)).

(2) **Exchangeable VS Model** (EVS): \( \beta_k(s) \) are independent and follow the distribution in (2).

(3) **Spatial VS Model** (SpVS): The full model in Section 3. \( \beta_k(s) \) marginally follow a mixture distribution described in (2), and dependence is induced with a Gaussian copula.

In the simulation study, we use relatively uninformative priors, with \( \alpha_k \sim N(0, 1) \), \( \pi_k \sim \text{Beta}(1, 1) \), \( 1/\omega_k^2 \sim \text{Gamma}(0.001, 0.001) \), and \( r \sim \text{Uniform}(0, 2) \). For each dataset we run the model for 5500 iterations discarding the first 1000. We compare these three models using MAD, power and type-I error of \( \beta_k(s) \), and the ability to identify locally null coefficients.

4.1 *Simulation results*

Table 2(a) gives the MAD separately for global (\( \pi = 1 \)), local (\( 0 < \pi < 1 \)), and null (\( \pi = 0 \)) predictors. For global predictors, the spatial models (Sp and SpVS) have similar MAD and outperform the non-spatial model (EVS), as expected. The MAD for null covariates is considerably smaller for the variable selection models (EVS and SpVS). The spatial variable selection model outperforms both competitors for the local predictors, where both shrinking effects to zero and borrowing strength across space are beneficial.

Table 2(b) shows that by exploiting dependence between nearby locations, the SpVS model has substantially higher power for detecting important covariates than the EVS model. Type-I errors rates for local and global effects are near the nominal 0.05 level for all models, and much less than 0.05 for the globally zero coefficients for all three models (not shown). For the local covariates, there is not a significant difference in power when \( \pi_k = 0.3 \) and when \( \pi_k = 0.7 \), therefore power is only reported for each level of \( \alpha_k \). When the coefficients are uncorrelated, the spatial model has the highest power, with the EVS and SpVS models...
having slightly lower power very similar to each other. When the coefficients have weak or strong correlation, both spatial models significantly outperform the EVS model.

Finally, we check the ability of three goodness of fit measures to select the correct model: Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002); the predictive measure of Gelfand and Ghosh (1998) (GG); and log pseudo marginal likelihood (LPML). Smaller values indicate better fit for DIC and GG, while the preferred model is indicated by larger values of LPML. The generating model for Scenario 1 is EVS, while for Scenarios 2 and 3 the correct model is SpVS. In this study, DIC and LPML most often pick the correct “best” model, while GG is less accurate. Under Scenario 1, LPML and DIC pick SpVS and EVS equally often, while GG is split across all three models. LPML selects SpVS 38/50 and 48/50 times for Scenarios 2 and 3. DIC performs similarly, 37/50 and 48/50 times respectively. Neither measure selected the Sp model under any scenario. GG incorrectly selects the EVS model as best 32/50 and 43/50 times in Scenarios 2 and 3, and chooses the Sp model 14/50 and 6/50 times, selecting the correct SpVS model only 4/50 and 1/50 times.

[Table 2 about here.]

5. Analysis of CVD hospitalization data

5.1 Model comparisons

We fit the Spatial, EVS, and Spatial VS models to the CVD hospitalization and PM2.5 component data described in Section 2. Due to the large number of confounding variables per site, we chose to fit a two-stage approximation to the fully Bayesian model. The Poisson likelihood function in (1) is approximated by a normal likelihood, $\hat{\beta}(s) \sim N_p[\hat{\beta}(s), \hat{V}(s)]$, where $\hat{\beta}(s)$ is the maximum likelihood estimator and $\hat{V}(s)$ is the estimated covariance matrix. This two-stage model has the advantage of much faster computation time and model fit does
not depend on MCMC convergence of the confounding coefficients, as the approximation is fit marginally over the confounding coefficients $\eta(s)$.

In all of the models, we choose a flat prior between 0 and 1 for $\pi_k$, but note that one could use a beta prior for $\pi_k$ to encourage global selection by choosing both beta parameters to be less than one. Local selection ($0 < \pi_k < 1$) could also be weighted more heavily in the prior, and greater prior weight could be placed on inclusion or exclusion by choosing an asymmetric beta distribution for the prior. The prior for $\alpha_k$ is normal with standard deviation 0.025, which is approximately two to four times larger than the effect sizes we expect to see. Recalling that we model the log relative risk, 0.1 would correspond to a more than 10% increase in risk of CVD hospitalization for a 90th percentile increase in pollution, which previous work suggests is quite large. We have found that model fit is somewhat sensitive to the choice of the prior on $\omega_k$. When $\omega_k$ is large it becomes difficult to differentiate between the 0 and $\alpha_k$ modes of the distribution, and thus $\pi_k$ and $\alpha_k$ become difficult to estimate. Though $\alpha_k$ and $\pi_k$ are therefore somewhat sensitive to the choice of prior on $\omega_k$, we found that estimates of $\beta_k(s)$ are more stable over different choices of prior. We therefore choose a gamma(0.1, 0.001) prior on the inverse variance, $1/\omega_k^2$, to be relatively uninformative while encouraging $\omega_k$ to be small. We note in the simulation study that larger effect sizes are more robust to an uninformative prior on $\omega_k^2$, and that this prior is equivalent to fitting gamma(0.1, 0.1), a vague prior, on data with 10 times larger effect sizes, roughly equivalent to the effect sizes of the simulation study. The spatial range parameter is a uniform distribution from 0 to 1200 km. We fit the spatial VS model with different choices for the value of $C$; $C=100$, 225, 400, and 900. DIC was nearly identical for the first three choices, and highest for $C=900$. Here we present the model with $C = 400$. Though $C = 100$ minimizes DIC for the EVS model, the EVS model presented here also has $C = 400$ for consistency, and the model fits are very similar. We generate 50,000 MCMC samples, discarding the first 10,000 for burn in.
We compare models using DIC, LPML, and GG, the goodness of fit measures described in Section 4.1. We calculate LPML and GG statistics using the Poisson likelihood in (1) with \(\eta_k(s)\) replaced by its maximum likelihood estimate, \(\hat{\eta}_k(s)\). We calculate the DIC using the normal likelihood approximation. The goodness of fit measures for each model are shown in Table 3. The spatial VS model is selected as the best model by DIC and LPML, a substantial improvement over EVS and spatial. The spatial model without variable selection is selected as the best model by GG, but as shown in the simulation study, this result is questionable. We also note that the estimated spatial correlation is very strong, further evidence that the spatial models are a better choice than the exchangeable model.

In addition to these model comparisons, we conducted an extensive sensitivity analysis, which is summarized in the supplementary material. We investigated sensitivity to several modeling choices, including the effects of hyperpriors, confounders, and the variance ratio, \(C\), and found some sensitivity for hyperparameters, but that the effect estimates of interest were robust to these factors.

5.2 Results

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

The overall (spatial average) effects for all pollutants are reported in Table 3, where local covariates are defined as \(0.20 < \pi_k < 0.80\) and global covariates as \(\pi_k > 0.80\). In Table 4 we present the posterior estimates for the overall mean when included, \(\alpha_k\), the inclusion probability, \(\pi_k\), and the random effects standard deviation, \(\omega^2_k\), for our spatial VS model. Across all models, we consistently find positive overall effects for elemental carbon (EC). In the EVS model, the overall effect estimate of elemental carbon, \(\alpha_4\), is statistically significant

Spatial variable selection methods for investigating acute health effects of fine particulate matter components
(95% posterior credible interval excludes 0). For the spatial and spatial VS models, posterior probability that \( \alpha_4 > 0 \) is 0.88. No other overall effects are statistically significant, though many county specific relative risk estimates are significant. These results are in general agreement with Peng et al. (2009), who found 0.8% increase in CVD hospitalization risk for an IQR increase in elemental carbon, over the time period 2002-2006 in a six pollutant model.

The three models produce markedly different spatial patterns for the effect of EC, as shown in Figure 1. The spatial model without variable selection shows considerable variation in the effects across component and location. In the presence of the large number of potential predictors, EC does not stand out as particularly important in the spatial model without variable selection (Figure 1a). EC is clearly identified as the most relevant pollutant in both variable selection models. In the EVS model, the effect size varies considerably by city with no discernible spatial pattern. In contrast, the SpVS model reveals a strong spatial pattern, with the largest effects in the Northeast, Mid Atlantic, East North Central, West North Central, and Pacific regions. Therefore, the spatial variable selection model isolates both a specific PM component with the most impact, and further refines the spatial distribution of the effect size.

Figure 2 maps the county-specific EC effect estimates. The estimates vary smoothly across space, with the largest values in the Northeast, Midwest, and Pacific Coast. The estimates are small in other areas, with the notable exception of El Paso, Texas which has one of the largest effect estimates. The 95% posterior intervals for many counties exclude zero for the SpVS model, as shown in Figure 3c. Comparing the intervals for the EVS and SpVS models in Figure 3 illustrates the benefit of spatial smoothing; by borrowing strength across nearby counties, the intervals are narrower for the SpVS model and thus, unlike the EVS model, several individual counties have statistically significant EC effects.
6. Discussion

In this paper, we develop a statistical model which allows for local variable selection of spatially correlated coefficients. The copula framework creates a smooth prior surface reflecting the intuition that effect estimates near counties where the effect is zero will also be small. A simulation study confirms that including spatial correlation in the model, while adding computational complexity, results in better model fit and improved effect estimation than the spatial model or the exchangeable variable selection model. While the MAD for globally included covariates are similar between spatial and spatial VS models as expected, we found more precise estimates and more powerful tests of association when the data were generated with locally-relevant predictors. Under strong spatial correlation, our model had nearly half the MAD of the spatial model and 20% greater power than the EVS model.

In the data application, the SpVS model is clearly the most appealing. Figure 1 shows that the spatial models with and without variable selection give dramatically different results. With variable selection, a single pollutant (elemental carbon) is identified as the most relevant, whereas many pollutants have some effect in the model without variable selection. Compared to the EVS model, we find a substantial reduction in variance due to spatial smoothing.

This investigation can help us understand how observed spatial and temporal variations in the health effects of total PM2.5 concentration may be related to variation in chemical components. No previous study to our knowledge has utilized variable selection techniques on such a large list of components, and few have incorporated spatial correlation.

When interpreting these overall effect estimates, it is also important to note that these
numbers reflect only the average risk increase or decrease for the 115 counties included in the study, and are not a reflection of an average national risk. The counties selected for the study are generally densely populated urban counties which may differ significantly in pollution levels, demographic, socio-economic and health characteristics from other US counties. Also, while epidemiological studies such as this one can point to interesting relationships between pollutant levels and health effects, we must be careful in interpreting effects as causal links. As discussed by Thomas et al. (2007), the indicated pollutants may be correlated with other pollutants not included with the study, or in fact may be proxying for another pollutant which is included. Because we do not observe the relationship between ambient and personal exposure, the indicated pollutants may be signaling due to higher association with personal exposure, though another pollutant is more correlated with health. Measurement error, particularly that of using ambient pollution as a proxy for personal exposure, may also play a role, especially for pollutants which are extremely spatially heterogeneous. A spatially homogeneous pollutant may have stronger association between ambient and personal exposure, and thus may mask the more heterogeneous pollutant which has a relationship between personal exposure and health. Previous work studying measurement error and coarse PM10 (that greater than 2.5) (Chang et al., 2011) and total PM10 (Zeger et al., 2000) suggest relatively consistent estimates of that effect under different measurement error scenarios, but spatial heterogeneity of individual components of PM2.5 may be appreciably greater, and individual components vary in their level of heterogeneity (Bell et al., 2011). Further investigation of how spatial heterogeneity and measurement error may change health effect estimates is warranted, particularly in variable selection models.

Further research into the reason for heterogeneity in pollutant effects using both epidemiological and toxicological approaches is also needed. Heterogeneity may be due to differences in particle chemistry, interaction effects, differing population exposure characteristics, or the
measurement error issues described above. Though we have divided PM into its chemical components, the physical and toxicological properties are by no means homogeneous. For example, sulfate and nitrate particles are mostly secondary pollutants created by reactions of sulfur and nitrogen oxides with other gases and particles to form a diverse set of molecules, including sulfuric acid and ammonium bisulfate, which vary in acidity (Schlesinger et al., 2006). Toxicological studies suggest this acidity, not just the components, may be an important factor in health outcomes (Schlesinger et al., 2006). Individual chemical components may come from diverse primary sources as well, resulting in differing size and associated particles. For example, potassium becomes airborne from burning biomass, from blown crustal material (dust), and salts from sea spray which may have differing biological effects (Schlesinger, 2007). An extension of this SpVS model to a space-time setting would be useful for fully understanding possible seasonal effects. Though we adjust for seasonal confounding, we do not consider interactions of pollutant effect and season in the current model.

Supplementary Materials

Web Appendices and Figures referenced in Sections 2, 3, 3.2, and 5.1 and R code for implementing the SpVS model are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

We are grateful for the support provided by NIH/NIEHS grant R21ES022585-01, NIH grants R01ES019560, R21ES022795-01A1, and 5R01ES014843-02, EPA grant RD83479801, and funds from HEI. We are also grateful to the editor, associate editor, and referees for their helpful comments.


Received: November 24, 2013
Figure 1. Posterior median coefficients (in % RR increase per IQR increase) for each model. Each row includes coefficients for the labeled pollutant; each column indicates one county. The counties are grouped according to the 8 EPA subregions, defined as Northeast (NE), Mid Atlantic (Mid Atl), South Atlantic (S Atl), East South Central (ESC), West South Central (WSC), East North Central (ENC), West North Central (WNC), Mountain (Mtn), and Pacific (Pac). This figure appears in color in the electronic version of this article.
Figure 2. Posterior median county-specific % risk increase for an IQR increase in elemental carbon from the SpVS model. This figure appears in color in the electronic version of this article.
Figure 3. City-specific means (open dot), median (filled dot), and 95% posterior intervals (vertical lines) for the elemental carbon effect, arranged by EPA subregion, defined as Northeast (NE), Mid Atlantic (Mid Atl), South Atlantic (S Atl), East South Central (ESC), West South Central (WSC), East North Central (ENC), West North Central (WNC), Mountain (Mtn), and Pacific (Pac). The national average estimates are given for $\alpha_k$ (triangle), $\alpha_k \pi_k$ (square), and average of the 115 $\beta_k(s)$ (diamond). This figure appears in color in the electronic version of this article.
Table 1
Interquartile range (IQR), median, and maximum observed value in µg/m³ across all 115 sites in the period 2000-2008. Minimum values all approximately zero, and are not shown. The seven most massive components are listed first, with the rest in alphabetical order.

<table>
<thead>
<tr>
<th>Element</th>
<th>Median</th>
<th>IQR</th>
<th>Max</th>
<th>Element</th>
<th>Median</th>
<th>IQR</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfate</td>
<td>2.43</td>
<td>2.89</td>
<td>39.90</td>
<td>Chlorine</td>
<td>0.00</td>
<td>0.02</td>
<td>5.78</td>
</tr>
<tr>
<td>Nitrate</td>
<td>0.83</td>
<td>1.54</td>
<td>45.30</td>
<td>Chromium</td>
<td>0.00</td>
<td>0.00</td>
<td>1.17</td>
</tr>
<tr>
<td>Silicon</td>
<td>0.07</td>
<td>0.10</td>
<td>10.10</td>
<td>Copper</td>
<td>0.00</td>
<td>0.00</td>
<td>0.55</td>
</tr>
<tr>
<td>Elemental Carbon</td>
<td>0.54</td>
<td>0.52</td>
<td>11.00</td>
<td>Iron</td>
<td>0.07</td>
<td>0.08</td>
<td>24.10</td>
</tr>
<tr>
<td>Organic Carbon</td>
<td>3.30</td>
<td>3.42</td>
<td>101.18</td>
<td>Lead</td>
<td>0.00</td>
<td>0.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Sodium Ion</td>
<td>0.09</td>
<td>0.13</td>
<td>33.40</td>
<td>Magnesium</td>
<td>0.00</td>
<td>0.01</td>
<td>2.72</td>
</tr>
<tr>
<td>Ammonium Ion</td>
<td>1.12</td>
<td>1.43</td>
<td>21.70</td>
<td>Nickel</td>
<td>0.00</td>
<td>0.00</td>
<td>0.91</td>
</tr>
<tr>
<td>Aluminum</td>
<td>0.01</td>
<td>0.04</td>
<td>4.01</td>
<td>Potassium</td>
<td>0.05</td>
<td>0.05</td>
<td>12.80</td>
</tr>
<tr>
<td>Arsenic</td>
<td>0.00</td>
<td>0.00</td>
<td>0.16</td>
<td>Titanium</td>
<td>0.00</td>
<td>0.01</td>
<td>0.56</td>
</tr>
<tr>
<td>Bromine</td>
<td>0.00</td>
<td>0.00</td>
<td>0.19</td>
<td>Vanadium</td>
<td>0.00</td>
<td>0.00</td>
<td>0.17</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.04</td>
<td>0.05</td>
<td>4.90</td>
<td>Zinc</td>
<td>0.01</td>
<td>0.01</td>
<td>3.11</td>
</tr>
</tbody>
</table>
Table 2
(a) Median absolute deviation (MAD) × 100 for the simulation study. MAD is averaged across all 50 locations for each pollutant. The overall average over datasets is presented here. Simulation settings for α_k, π_k are shown in columns 2 and 3, and abbreviated names of the components used to generate the response are in column 1. A “*” indicates a statistically significant difference from the SpVS model. (b) Power and Type-I error for spatial model with no variable selection (Sp), exchangeable variable selection model (EVS), and spatial variable selection model (SpVS) for local and global covariates. We calculate power for each setting of the true value of α_k.

<table>
<thead>
<tr>
<th>Component</th>
<th>α × 100</th>
<th>π</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulf.</td>
<td>5</td>
<td>1</td>
<td>1.39*</td>
<td>1.66*</td>
<td>1.48</td>
<td>1.36</td>
<td>1.65*</td>
<td>1.38</td>
<td>1.10</td>
<td>1.55*</td>
<td>1.12</td>
</tr>
<tr>
<td>Nitr.</td>
<td>10</td>
<td>1</td>
<td>1.40*</td>
<td>1.46</td>
<td>1.45</td>
<td>1.37*</td>
<td>1.59*</td>
<td>1.41</td>
<td>1.16</td>
<td>1.43*</td>
<td>1.14</td>
</tr>
<tr>
<td>Sili.</td>
<td>5</td>
<td>0.3</td>
<td>1.34*</td>
<td>0.91</td>
<td>0.95</td>
<td>1.41*</td>
<td>0.99*</td>
<td>1.11</td>
<td>1.17*</td>
<td>0.95</td>
<td>0.91</td>
</tr>
<tr>
<td>El C.</td>
<td>5</td>
<td>0.7</td>
<td>1.91*</td>
<td>2.10</td>
<td>2.08</td>
<td>1.79</td>
<td>2.04*</td>
<td>1.78</td>
<td>1.49*</td>
<td>1.83*</td>
<td>1.44</td>
</tr>
<tr>
<td>Sodi.</td>
<td>10</td>
<td>0.3</td>
<td>1.53*</td>
<td>0.80</td>
<td>0.86</td>
<td>1.52*</td>
<td>0.79</td>
<td>0.84</td>
<td>1.21*</td>
<td>0.70</td>
<td>0.66</td>
</tr>
<tr>
<td>Or C.</td>
<td>10</td>
<td>0.7</td>
<td>2.30*</td>
<td>2.02</td>
<td>2.04</td>
<td>2.30*</td>
<td>2.01</td>
<td>1.95</td>
<td>1.91*</td>
<td>2.05*</td>
<td>1.60</td>
</tr>
<tr>
<td>Arse.</td>
<td>0</td>
<td>0</td>
<td>0.53*</td>
<td>0.12</td>
<td>0.12</td>
<td>0.50*</td>
<td>0.16</td>
<td>0.19</td>
<td>0.46*</td>
<td>0.13*</td>
<td>0.18</td>
</tr>
<tr>
<td>Brom.</td>
<td>0</td>
<td>0</td>
<td>0.60*</td>
<td>0.15</td>
<td>0.13</td>
<td>0.56*</td>
<td>0.14*</td>
<td>0.21</td>
<td>0.50*</td>
<td>0.12*</td>
<td>0.24</td>
</tr>
<tr>
<td>Calc.</td>
<td>0</td>
<td>0</td>
<td>0.63*</td>
<td>0.15</td>
<td>0.15</td>
<td>0.69*</td>
<td>0.15*</td>
<td>0.25</td>
<td>0.54*</td>
<td>0.12*</td>
<td>0.26</td>
</tr>
<tr>
<td>Avg. Local, β_k(s) = 0</td>
<td>1.61*</td>
<td>1.19*</td>
<td>1.23</td>
<td>1.64*</td>
<td>1.24</td>
<td>1.24</td>
<td>1.26*</td>
<td>0.72</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. Local, β_k(s) ≠ 0</td>
<td>1.93*</td>
<td>1.72</td>
<td>1.73</td>
<td>1.87*</td>
<td>1.67*</td>
<td>1.59</td>
<td>1.62</td>
<td>2.04</td>
<td>1.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Power and Type I error

<table>
<thead>
<tr>
<th>α_k</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Local</td>
<td>0.05</td>
<td>0.41</td>
<td>0.30</td>
<td>0.34</td>
<td>0.48</td>
<td>0.31</td>
<td>0.49</td>
<td>0.65</td>
<td>0.37</td>
</tr>
<tr>
<td>Local</td>
<td>0.10</td>
<td>0.80</td>
<td>0.76</td>
<td>0.75</td>
<td>0.84</td>
<td>0.74</td>
<td>0.81</td>
<td>0.90</td>
<td>0.77</td>
</tr>
<tr>
<td>Global</td>
<td>0.05</td>
<td>0.74</td>
<td>0.55</td>
<td>0.61</td>
<td>0.75</td>
<td>0.55</td>
<td>0.73</td>
<td>0.85</td>
<td>0.57</td>
</tr>
<tr>
<td>Global</td>
<td>0.10</td>
<td>1.00</td>
<td>0.92</td>
<td>0.92</td>
<td>0.99</td>
<td>0.88</td>
<td>0.95</td>
<td>1.00</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Type-I Error

<table>
<thead>
<tr>
<th>α_k</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
<th>Sp</th>
<th>EVS</th>
<th>SpVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>0.03</td>
<td>0.01</td>
<td>0.02</td>
<td>0.05</td>
<td>0.01</td>
<td>0.07</td>
<td>0.06</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>Global</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Table 3  
Estimated overall relative risk expressed as percent increase ($e^{\alpha_k} \times 100 - 100$) and goodness of fit statistics. For Spatial model (no VS), $e^{\alpha_k} \times 100 - 100$ is reported. For models with spatial correlation, range estimates in km are given. Statistically significant associations (0.05 level) marked by *. Local selection marked by $^L$ (posterior $0.20 < \pi_k < 0.80$). No estimates of $\pi$ were greater than 0.8. We also include goodness of fit criterion, DIC, GG, and LPML. The best model by each measure is highlighted in bold.

<table>
<thead>
<tr>
<th>Element</th>
<th>SpVS</th>
<th>Spatial</th>
<th>EVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfate</td>
<td>0.09$^L$</td>
<td>0.26</td>
<td>0.02</td>
</tr>
<tr>
<td>Nitrate</td>
<td>0.01$^L$</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>Silicon</td>
<td>-0.00</td>
<td>-0.1</td>
<td>-0.00</td>
</tr>
<tr>
<td>Elemental Carbon</td>
<td>0.42$^L$</td>
<td>0.46</td>
<td>0.64$^L$</td>
</tr>
<tr>
<td>Organic Carbon</td>
<td>0.03$^L$</td>
<td>-0.14</td>
<td>0.04$^L$</td>
</tr>
<tr>
<td>Sodium Ion</td>
<td>-0.01</td>
<td>-0.09</td>
<td>0.00</td>
</tr>
<tr>
<td>Ammonium Ion</td>
<td>0.05</td>
<td>0.17</td>
<td>0.16$^L$</td>
</tr>
<tr>
<td>Aluminum</td>
<td>-0.00</td>
<td>-0.06</td>
<td>-0.02</td>
</tr>
<tr>
<td>Arsenic</td>
<td>-0.05</td>
<td>-0.42</td>
<td>-0.06</td>
</tr>
<tr>
<td>Bromine</td>
<td>-0.02</td>
<td>0.09</td>
<td>0.00</td>
</tr>
<tr>
<td>Calcium</td>
<td>-0.00</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>Chlorine</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Chromium</td>
<td>0.02</td>
<td>-0.16</td>
<td>-0.01</td>
</tr>
<tr>
<td>Copper</td>
<td>0.01</td>
<td>-0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Iron</td>
<td>0.00</td>
<td>0.41</td>
<td>0.04</td>
</tr>
<tr>
<td>Lead</td>
<td>0.04</td>
<td>-0.09</td>
<td>-0.02</td>
</tr>
<tr>
<td>Magnesium</td>
<td>0.02</td>
<td>-0.06</td>
<td>-0.02</td>
</tr>
<tr>
<td>Nickel</td>
<td>0.00</td>
<td>0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>Potassium</td>
<td>0.02</td>
<td>-0.15</td>
<td>-0.03</td>
</tr>
<tr>
<td>Titanium</td>
<td>0.00</td>
<td>0.06</td>
<td>-0.01</td>
</tr>
<tr>
<td>Vanadium</td>
<td>0.04</td>
<td>-0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Zinc</td>
<td>-0.00</td>
<td>0.20</td>
<td>0.02</td>
</tr>
</tbody>
</table>

| Range (km)       | 1143   | 1175    |       |
| DIC              | **2647** | 2746    | 2684  |
| GG               | 27.335 | **27.292** | 27.350 |
| LPML             | **-155495** | -155551 | -155585 |
Table 4

Estimates of average county specific $\beta$, mean when included ($e^{\alpha_k} \times 100 - 100$), probability of inclusion $\pi_k$ from the Spatial VS model with 22 pollutants, and estimate of standard deviation, rescaled ($e^{\omega_k} \times 100 - 100$).

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Avg. $\beta_k^a$</th>
<th>$\alpha_k$</th>
<th>$\pi_k$</th>
<th>$\omega_k$</th>
<th>Avg. $\beta_k^a$</th>
<th>$\alpha_k$</th>
<th>$\pi_k$</th>
<th>$\omega_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfate</td>
<td>0.11</td>
<td>0.16</td>
<td>0.30</td>
<td>3.02</td>
<td>Chlorine</td>
<td>-0.01</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>Nitrate</td>
<td>0.15</td>
<td>-0.00</td>
<td>0.24</td>
<td>2.29</td>
<td>Chromium</td>
<td>-0.04</td>
<td>0.22</td>
<td>0.06</td>
</tr>
<tr>
<td>Silicon</td>
<td>0.03</td>
<td>-0.04</td>
<td>0.08</td>
<td>2.18</td>
<td>Copper</td>
<td>0.01</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Elemental Carbon</td>
<td>0.59</td>
<td>0.56</td>
<td>0.70</td>
<td>0.94</td>
<td>Iron</td>
<td>0.06</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>Organic Carbon</td>
<td>0.05</td>
<td>0.07</td>
<td>0.22</td>
<td>3.54</td>
<td>Lead</td>
<td>-0.09</td>
<td>0.36</td>
<td>0.08</td>
</tr>
<tr>
<td>Sodium</td>
<td>-0.01</td>
<td>-0.08</td>
<td>0.08</td>
<td>2.10</td>
<td>Magnesium</td>
<td>-0.06</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>Ammonium</td>
<td>0.14</td>
<td>0.15</td>
<td>0.17</td>
<td>4.36</td>
<td>Nickel</td>
<td>-0.02</td>
<td>-0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Aluminum</td>
<td>-0.05</td>
<td>-0.03</td>
<td>0.08</td>
<td>2.66</td>
<td>Potassium</td>
<td>-0.15</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>Arsenic</td>
<td>-0.07</td>
<td>-0.34</td>
<td>0.12</td>
<td>3.65</td>
<td>Titanium</td>
<td>-0.02</td>
<td>0.05</td>
<td>0.12</td>
</tr>
<tr>
<td>Bromine</td>
<td>0.06</td>
<td>-0.12</td>
<td>0.08</td>
<td>2.98</td>
<td>Vanadium</td>
<td>0.04</td>
<td>0.24</td>
<td>0.12</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.10</td>
<td>-0.03</td>
<td>0.08</td>
<td>2.72</td>
<td>Zinc</td>
<td>0.06</td>
<td>-0.04</td>
<td>0.09</td>
</tr>
</tbody>
</table>