A Functional Data Analysis of Spatiotemporal Trends and Variation in Fine Particulate Matter

Meredith C. King[a,*], Ana-Maria Staicu[a], Jerry M. Davis[b], Brian J. Reich[a], Brian Eder[c]

[a]*Department of Statistics, North Carolina State University , Raleigh, North Carolina, 27695*
[b]*Department of Marine, Earth & Atmospheric Sciences, North Carolina State University, Raleigh, NC 27695*
[c]*Computational Exposure Division, National Exposure Research Laboratory, U. S. Environmental Protection Agency, Research Triangle Park, NC 27711*

Abstract

In this paper we illustrate the application of modern functional data analysis methods to study the spatiotemporal variability of particulate matter components across the United States. The approach models the pollutant annual profiles in a way that describes the dynamic behavior over time and space. This new technique allows us to predict yearly profiles for locations and years at which data are not available and also offers dimension reduction for easier visualization of the data. Additionally it allows us to study changes of pollutant levels annually or for a particular season. We apply our method to daily concentrations of two particular components of $PM_{2.5}$ measured by two networks of monitoring sites across the United States from 2003 to 2015. Our analysis confirms existing findings and additionally reveals new trends in the change of the pollutants across seasons and years that may not be as easily determined from other common approaches such as Kriging.

Keywords: Particulate matter; Functional data; Air pollution; Kriging; Functional principal component analysis.

*Corresponding author
Email address: mcking4@ncsu.edu (Meredith C. King)

1. Introduction

Despite recent mitigation strategies promulgated by the U. S. Environmental Protection Agency (EPA), fine particulate matter ($PM_{2.5}$ - aerosols with diameters less than or equal to 2.5μ) continues to have a detrimental effect on human health and welfare in many areas of the nation. In a 2010 study, the EPA reported that over 20 million citizens lived in counties that exceeded the National Ambient Air Quality Standard (NAAQS) for $PM_{2.5}$. It is estimated that thousands of premature deaths occur in the U.S. annually due to elevated concentrations of $PM_{2.5}$ (Pope and Dockery, 2006). Additionally, $PM_{2.5}$ contributes greatly to visibility degradation through the scattering and absorption of visible light (Malm et al., 2004) and excessive nutrient and pollutant deposition.

In this paper we are concerned with characterizing how two main pollutant species of $PM_{2.5}$ - particulate nitrate (NO_3^-) and particulate sulfate (SO_4^{-2}) - vary during 2003-2015 across the U.S. The motivating data set contains concentrations of these pollutant species recorded by two monitoring networks, which operate independently and often have disparate sampling protocols and standard operating procedures; see Figure 1. Our objective is to study the variability of fine particulate nitrate and sulfate over time and across the U.S. using recent functional data analysis techniques.

Due to the health risks and visibility degradation associated with high concentrations of $PM_{2.5}$, its behavior has been extensively studied. Millar et al. (2010) provide a thorough review of approaches to modeling exposure to fine particulate matter. A popular class of methods are empirical (statistical) methods, common examples of which include Kriging (see Liao et al. (2006) and Leem et al. (2006) among others) and land use regression models. Hoek et al. (2008) provide a review of various land use regression models utilized to investigate spatial variation in concentrations.

Physical models, which apply mathematical equations from physical processes, are another common approach (Cyrys et al., 2005; Næss et al., 2007). Finally, many hybrid methods have been developed to incorporate different models and data sources (Hu et al., 2013; Liu et al., 2009). Berrocal et al. (2009) provide a way to incorporate data with different spatial supports; they use their method to analyze ozone levels using measurements taken from specific monitoring locations and the Community Multiscale Air Quality (CMAQ) model.

To study the spatiotemporal behavior of fine particulate matter, we consider an empirical modeling view in this work. In this case, the availability of complete data sets is necessary for many statistical approaches, and in the case of pollution data it is very common to have incomplete observations. For example, in the motivating application the planned schedule for measuring the pollution concentration is every third day. Various statistical approaches have been proposed to first impute the missing values and then model and predict fine particulate matter. Hierarchical Bayesian methods are a commonly used approach to model spatiotemporal behavior of particulate matter and predict missing observations (Sahu et al., 2006; Kibria et al., 2002; Zidek et al., 2002). Smith et al. (2003) use thin plate regression splines to model the temporal and spatial trends in the data and employ an expectation-maximization algorithm approach to predict missing observations. Sampson et al. (2011) consider data with a similar structure to ours and propose a spatiotemporal model that separates temporal trends and spatially varying coefficients and allows for non-stationary spatial correlation. Table 5 in the Supplementary Materials provides further comparison of these approaches.

The use of functional data analysis methods for environmental data has received attention recently (Gao and Niemeier, 2008; Park et al., 2013; Shaadan et al., 2012; Hörmann et al., 2015).

We propose incorporating the annual periodicity of the measurements into the model by viewing the annual concentration profiles at a location as a *functional time series* (Hörmann and Kokoszka, 2012) and modeling it as the sum of three components: 1) an annual mean level, 2) a linear combination of smooth annual trends with site-specific coefficients that vary over the years during the period of study, and 3) an annual specific residual effect. The annual trajectory may or may not vary over the space and the annual residual profile is assumed to be independent across sites and years, but allowed to exhibit dependence within a year. Although derived from a different perspective, our modeling technique shares several similarities with Sampson et al. (2011). Both approaches rely on a linear combination involving orthogonal smooth temporal trends. Sampson et al. (2011) works with the full time series and the temporal trends are functions defined over the entire period under study. In contrast we view the site level data as a *functional time series* where the functional argument is *day within year* and the series is indexed by *year*, and thus the trends are functions defined over a year-time. As a consequence of the different perspectives, the coefficients are space-dependent solely in Sampson et al. (2011), while they exhibit both spatial and yearly variation in the proposed approach. Furthermore the assumptions of the residual process are different: the residual component is allowed to have spatial dependence and is assumed independent over days/years in Sampson et al. (2011), while it is allowed to have dependence across the days within year, but is assumed independent over space and years in our method.

Our paper makes several contributions to the field. First, it proposes a dimension reduction approach of the complex dependent data over space and time. The methods are accompanied by an estimation approach that is distinct from other ideas considered in the literature and it leads to fast computations. This is in contrast to a full hierarchical Bayesian modeling approach, which is more computationally intensive. Second, our methodology relies on weaker assumptions than the ones commonly used in these settings. In particular, when Kriging is employed for prediction, stronger assumptions about the covariance structure - such as separable covariance structures which assume the dependence across space is independent of time and vice versa - are often needed to make computation feasible. By comparison, the proposed method considers a non-separable and non-stationary covariance structure. Third, the proposed method allows us to better visualize and gain insights from the data.

The remainder of the paper is structured as follows: Section 2 describes the data to be used in this paper. The modeling framework and estimation techniques are detailed in Section 3. The application of the proposed methods to our data and interpretation of the results are discussed in Section 4 and a description of the software implementation is found in Section 5. We conclude with a brief summary in Section 6.

2. Data description

Particulate nitrate and sulfate are recorded by two networks: the Interagency Monitoring of Protected Visual Environments (IMPROVE), and the Chemical Speciation Network (CSN).

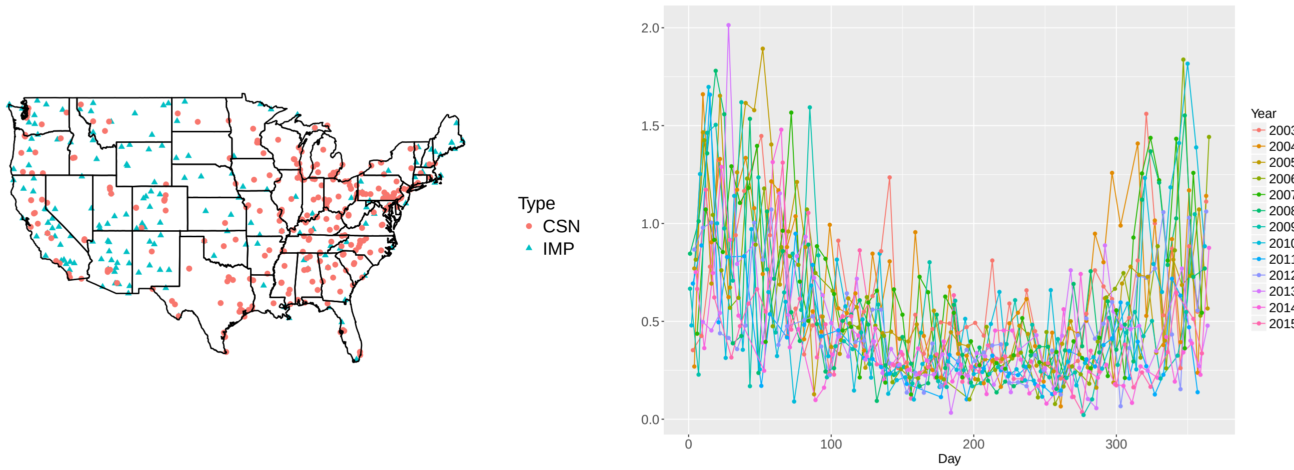

Figure 1: The left panel shows the location of the sites in the two networks: IMPROVE (blue triangles) and CSN (red circles). The right panel depicts the data for a CSN site in North Carolina observed from 2003-2015.

IMPROVE. The IMPROVE network, which began operations in 1985, represents a collaborative monitoring effort governed by a consortium of federal, regional, and state organizations. The majority of IMPROVE monitors are located in rural areas, often in national parks. There is a higher density of sites located in the western U.S. than in the eastern U.S.; the sites are depicted using filled triangles in Figure 1. They collect 24-hour integrated samples every third day (midnight to midnight LST). For a detailed description, see Malm et al. (2004).

CSN. The U.S. EPA's more recently established CSN also follows the one-in-three days collection protocol; see the CSN website for more information (EPA, 2016). Most of the sites monitored by this network are located in urban areas with a greater density in the eastern U.S.; the sites are depicted using filled circles in Figure 1.

Our study is limited to the period 2003-2015. In spite of the one-in-three day planned sampling schedule, several sites had no data for some years, or if they had observations during a year, they had only a few, in some cases covering less than a half a year period. A few CSN sites had multiple non-consistent recordings for sulfate during the same year and none for nitrate; those CSN sites during the respective years were omitted from the analysis. Our analysis is based on the remaining sites: 184 IMPROVE sites and 326 CSN sites. Most of these contain pollutant measurements for all 13 years; however there are sites observed only one year during this time frame, a few years but not necessarily consecutive years, or they had very sparse recordings during a year. Some analyses of similar data sets have utilized weekly or biweekly averages to handle missing daily measurements (Smith et al., 2003; Sahu et al., 2006; Sampson et al., 2011). However, daily $PM_{2.5}$ measurements are often used in studies as predictors for negative health outcomes (Bell et al., 2004; Dominici et al., 2006; Zanobetti et al., 2009). In these situations daily predictions for missing days would be beneficial. Further, by utilizing data on the daily scale we can avoid averaging over potentially unequal numbers of daily measurements.

The nitrate levels vary between 0 and 71 $\mu g/m^3$ and the sulfate levels vary between 0 and 41 $\mu g/m^3$ with higher values indicative of higher pollution levels. However, $\sim 0.5\%$ of all nitrate levels are larger than 10 $\mu g/m^3$ and $\sim 1\%$ of all sulfate levels are larger than 10 $\mu g/m^3$. To remove issues related to the skewness of the measurements, we take a log transformation. Because some measurements are close to zero we add a constant, 1, to each measurement before applying the log transformation. The right panel of Figure 1 depicts the CSN nitrate levels for a site in Winston-Salem, North Carolina.

3. Modeling framework

Environmental data often has a complex spatiotemporal dependence structure and modeling it poses many challenges. Let $Y_{ij}(d)$ be the response (log-transformed nitrate or sulfate concentration) at site i, year $13t_{ij}$ for $t_{ij} \in \{1/13, 2/13, \ldots, 1\}$, day $365d$, for $d \in \{1/365, 2/365, \ldots, 1\}$, and let s_i be the latitude/longitude of site i. We scale t_{ij} and d so they are both within the $[0, 1]$

interval. As d is a function of the day within year, we will often refer to d by day, with the understanding that the actual day within year is $365d$. Similarly, we will refer to t_{ij} by year, corresponding to the years of study from 2003 to 2015, even though the true year is $13t_{ij} + 2002$. We posit the following model

$$Y_{ij}(d) \quad = \mu(d, s_i, t_{ij}) + \textstyle\sum_{k \geq 1} \phi_k(d)\xi_{kij} + \epsilon_{ij}(d); \tag{1}$$

where $\mu(d, s_i, t_{ij})$ is the mean function which can depend on site, year or day specific covariates, the sum-term in the middle is the site-specific deviation from the overall mean and will be detailed next, and $\epsilon_{ij}(\cdot)$ is the site/time specific deviation. The term $\sum_{k \geq 1} \phi_k(d)\xi_{kij}$ is a linear combination of year-time functions $\phi_k(\cdot)$ that are assumed invariant across years and mutually orthogonal, in the sense that $\int_0^1 \phi_k(u)\phi_{k'}(u)\, du = 1$ if $k = k'$ and 0 otherwise. The basis coefficients ξ_{kij} quantify the dynamic variation over space (s_i) and time (t_{ij}) corresponding to the annual pattern represented by $\phi_k(\cdot)$. It is assumed that the spatiotemporal process $\xi_k(s_i, t_{ij}) = \xi_{kij}$ is independent from the noise measurement $\epsilon_{ij}(\cdot)$ for all k. Also for convenience we assume that the ξ_{kij}'s are independent across k. In practice we use a finite truncation K so that in model (1) the summation is for $k = 1, \ldots, K$. Finally, we assume that $\epsilon_{ij}(d)$ is the zero mean measurement error that is independent across i and j but possibly dependent over d.

The model in (1) is inspired from Park and Staicu (2015) in the way it models the dynamic behavior over time using time-invariant orthogonal basis functions. Nevertheless, (1) is different from Park and Staicu (2015), who assumes that the time varying curves $Y_{ij}(\cdot)$ are independent over i and thus are solely dependent over j. We assume that ξ_{kij} vary according to the following model

$$\xi_{kij} = a_{ki} + t_{ij}b_{ki} + e_{kij}, \tag{2}$$

where a_{ki} and b_{ki} are the random intercept and random slope, respectively, for year t_{ij} and site s_i, and e_{kij} is a nugget effect with variance denoted by σ_k^2. We assume that a_{ki} and b_{ki} are independent Gaussian processes over k and furthermore are mutually independent; to account for the dependence of the curves over sites, it is assumed that the two processes are each dependent across i. The Gaussian processes have mean zero and covariances $\text{cov}(a_{ki}, a_{ki'}) = \sigma_{ka}^2 \rho_{ka}(\|s_i - s_{i'}\|)$ and $\text{cov}(b_{ki}, b_{ki'}) = \sigma_{kb}^2 \rho_{kb}(\|s_i - s_{i'}\|)$; here σ_{ka}^2 and σ_{kb}^2 denote the variance of the intercepts and slopes corresponding to the kth component, while $\rho_{ka}(\cdot)$ and $\rho_{kb}(\cdot)$ are corresponding autocorrelation functions. This model assumption yields a somewhat simpler spatiotemporal covariance: $\text{cov}(\xi_{kij}, \xi_{ki'j'}) = \sigma_{ka}^2 \rho_{ka}(\|s_i - s_{i'}\|) + t_{ij} t_{i'j'} \sigma_{kb}^2 \rho_{kb}(\|s_i - s_{i'}\|)$ for $i \neq i'$ or $j \neq j'$. However, even in this case, the implied dependence structure of the data is described by a non-trivial spatiotemporal covariance:

$$\text{cov}\{Y_{ij}(d), Y_{i'j'}(d')\} = \sum_{k \geq 1} \phi_k(d)\phi_k(d')\{\sigma_{ka}^2 \rho_{ka}(\|s_i - s_{i'}\|) + t_{ij} t_{i'j'} \sigma_{kb}^2 \rho_{kb}(\|s_i - s_{i'}\|)\}. \quad (3)$$

This induced covariance model is non-separable in space and time (Schabenberger and Gotway, 2004; Cressie, 1993), in the sense that the dependence across space varies based on time and the reverse. The covariance model in (3) is isotropic in space (Schabenberger and Gotway, 2004), as the dependence across space depends solely on the distance between spatial locations. Isotropy, which represents a type of stationarity, may be an unreasonable assumption for $\text{PM}_{2.5}$ data. However, Stein (1999) demonstrated that in many cases predictions are insensitive to a misspecification of the covariance function when neighboring observations are highly correlated. Additionally, Parker et al. (2016) and Reich et al. (2011) both found in simulation studies that nonstationary covariance models do not dramatically improve prediction performance. Thus, we make the simplifying assumption of isotropy in our modeling approach.

The model in (1) relies on an orthogonal basis $\{\phi_k(\cdot)\}_k$ in $L^2[0,1]$. One option is to use pre-specified basis functions, such as Fourier basis functions or wavelets. However, such an approach would require a possibly large number of basis functions in order to capture the variability in the data. An appealing alternative is to use data-driven basis functions that would allow for a more parsimonious representation. Following the ideas of Park and Staicu (2015) we select $\phi_k(d)$'s as the eigenfunctions of the pooled covariance, obtained by ignoring the dependence across space and years. The resulting basis functions which will be the same across sites and years will then capture the key directions of variation within a year. More formally, denote the covariance function by $c\{(d, s_i, t_{ij}), (d', s_{i'}, t_{i'j'})\} = \text{cov}\{Y_{ij}(d), Y_{i'j'}(d')\}$ and let $\Sigma(d, d')$ be the weighted average across i and j; $\Sigma(d, d') = \sum_{j=1}^{13} P(T = t_j) \int_{\mathcal{S}} c\{(d, s, t_j), (d', s, t_j)\} g(s) ds$, where $P(T = t_j)$ is the relative frequency of the years $t_j \in \{t_1, t_2, \ldots, t_{13}\}$, $g(s)$ is the sampling density of the spatial locations and $\mathcal{S}$ is the spatial domain. For example, $g(s)$ could be the number of sites per km^2 in the U.S. for location s. Using similar arguments to Horváth and Kokoszka (2012), one can show that this function is a proper covariance function: see Section 5 in the Supplementary Materials for a full derivation. This covariance function is sometimes called the marginal covariance of an appropriate induced process and has been considered in the literature by other authors including Aston et al. (2016). Let $\Xi(d, d) = \Sigma(d, d') + \Gamma(d, d')$, where $\Gamma(d, d')$ is the smooth covariance function of the error term of (1). We take $\{\phi_k(\cdot)\}_k$ as the eigenbasis of the covariance function $\Xi(d, d)$. One simple approach to select the finite truncation K is using the percentage of explained variance of this covariance function.

Several important advantages of this modeling framework are that it is parsimonious, computationally efficient, and furthermore allows us to recover the trajectory for any spatial location and time in the domain under study. Specifically, once all the model components are estimated - the mean function $\mu(d, s_i, t_{ij})$, the orthogonal functions $\phi_k(d)$'s, the finite truncation K, and the covariance functions of processes $a_k(s)$ and $b_k(s)$ for all $k = 1, \ldots, K$, the proposed methodology allows us to reconstruct $Y(\cdot; s, t)$ for any location s in the U.S. and year t between 2003 and 2015 assuming the necessary covariate information is available. The following section describes the estimation of each of these terms in part as well as the prediction of full new trajectories. The methodology is illustrated on the nitrate data as recorded by the two networks, CSN and IMPROVE.

4. Estimation using U.S. nitrate concentrations from 2003-2015

The estimation approach encompasses three main steps: (i) estimate the overall mean function $\widehat{\mu}(\cdot)$; (ii) estimate the orthogonal functions $\{\widehat{\phi}_k(\cdot)\}$ and estimate the basis function coefficients $\widetilde{\xi}_{kij}$ for each k separately; (iii) estimate the spatiotemporal covariance function for each k and predict $\xi_k(s, t)$ for every s and t in the domain under study. In the following we discuss each step in turn; we use the nitrate data across the U.S. during 2003-2015 for illustration. Corresponding analysis for sulfate can be found in Section 1 of the Supplementary Materials.

There are species-specific differences in levels of accuracy, biases, and precision, which thereby complicate comparability across the networks. In particular, for nitrate, several challenges are measurement error associated with volatility, interference from gaseous organic species, and limitations of analytical methods; the calibration standards vary across networks. Because of these sampling differences we separately analyze the two networks and compare the results.

4.1. Mean nitrate profile in the U.S.

We consider the model framework in (1) to understand the variability of nitrate across the U.S., as monitored separately by each of the two networks. We choose to exploit the autocorrelation in the data and model the mean only as a function of day within year, $\mu(d)$. This choice also leads to a simpler interpretation of the basis coefficients.

To fix ideas, consider the nitrate data recorded by CSN monitors and assume that its variation is described by model (1); $\mu(d)$ denotes the overall nitrate level measured by CSN sites for day d. We assume that $\mu(\cdot)$ is a smooth cyclic function defined on $[0, 1]$. One popular approach to estimate an unknown smooth function is to use penalized spline smoothing (Wood, 2006; Eilers and Marx, 1996; Ramsay and Silverman, 2005). In particular let $\{B_\ell(\cdot)\}_{1\leq\ell\leq L}$ be a specified basis in $[0, 1]$; to account for the periodicity of the underlying function, we assume that this basis is cyclic, and use cyclic cubic splines (Wood, 2006). Let $\mu(d) = \sum_{\ell=1}^{L} B_\ell(d)\beta_\ell$ where L is the dimension of the basis and is specified by the number of knots. The choice of the basis dimension L, and thus the number of knots, is important in describing the smoothness of the mean function. A common way to bypass this is to select a relatively large value for L in order to capture the characteristics of the function and then penalize the basis coefficients. We consider the squared norm of the second derivative to describe the roughness of the function and use an additional parameter to control the size of the curvature relative to the model fit; see Wood (2006) among others.

In the case of independent observations, the nonparametric literature suggests selecting the smoothing parameter, λ, using restricted maximum likelihood (REML) or generalized cross-validation (GCV). There is limited research on smoothness parameter selection when the data exhibits dependence across space and time; we select λ using REML which has been shown to be more robust to data dependence (Krivobokova and Kauermann, 2007).

For the data applications we use a cyclic cubic basis with 11 interior knots placed at equal time points in $[0, 1]$; this leads to $L = 11$. For the CSN-nitrate data, the smoothing parameter was estimated to $\lambda = 114.63$; let $\widehat{\mu}(d) = \sum_{\ell=1}^{L} B_{\ell}(d)\widehat{\beta}_{\ell}$ denote the estimated, network specific, overall mean function. The estimated overall nitrate yearly profile in the U.S. on the log-scale is plotted in the leftmost panel of Figure 2. The result (shown in red) is compared with the estimated mean nitrate yearly profile for the IMPROVE network (blue color). The overall levels are higher for the CSN stations than for IMPROVE ones, and this is most likely because the majority of CSN sites are located in urban areas while the IMPROVE sites are primarily in rural locations, and pollution levels are typically higher in urban areas. Malm et al. (2004) noted this difference in nitrate levels for rural and urban locations as well.

However, irrespective of the monitoring network the nitrate levels exhibit similar behavior: they are higher in the cold seasons (fall and winter) than in the warmer seasons (spring and summer). Specifically, the nitrate levels start to decline roughly around the beginning of March until the middle of summer. The decline rate appears to be slower for the IMPROVE stations than for the CSN ones. Also the nitrate levels for IMPROVE sites seem to stay lower slightly longer than those of CSN sites, though by middle October they too increase steadily.

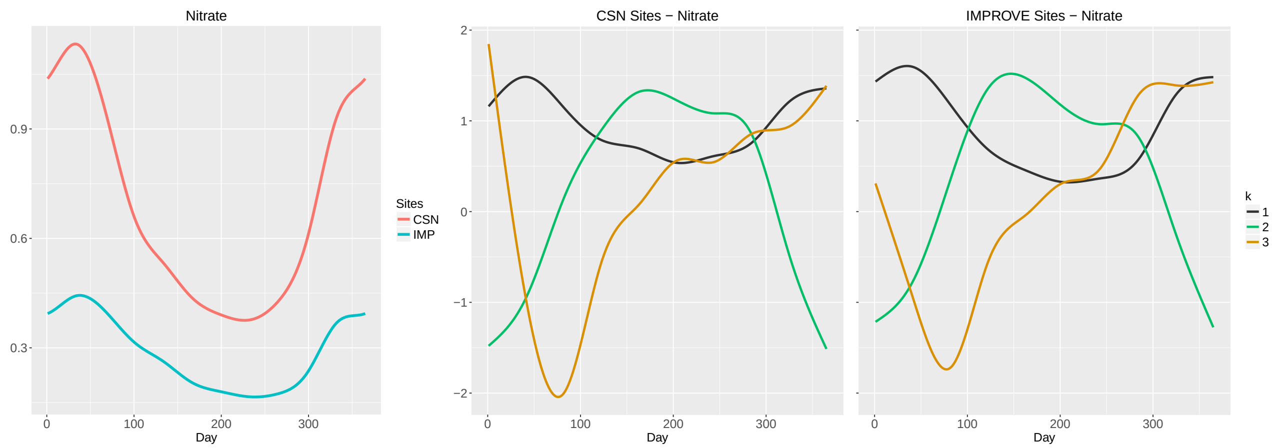

Figure 2: Left panel: estimated mean function $\widehat{\mu}(d)$ corresponding to CSN (red) and IMPROVE (blue). Results are shown on the log-scale. Middle and right panels: estimated eigenfunctions for nitrate variation, $\widehat{\phi}_k(d)$, for CSN sites (middle, K = 3) and IMPROVE sites (right, K = 3).

4.2. Main directions of annual variation across the U.S.

Let $\widetilde{Y}_{ij}(d) = Y_{ij}(d) - \widehat{\mu}(d)$ be the centered data; we use the centered data to estimate the data-driven orthogonal directions $\phi_k(d)$'s. Following the earlier intuition, the directions are estimated by the eigenfunctions of the pooled covariance function by ignoring the dependence over i and j. In order to ensure that the directions are smooth, a smooth estimator of the pooled covariance is obtained first. However the annual-profiles are not observed for every day of the year; to account for this we borrow ideas from sparse functional principal components (Yao et al., 2005).

Consider the pairwise product $G_{ij,ll'} = \widetilde{Y}_{ij}(d_{ijl})\widetilde{Y}_{ij}(d_{ijl'})$ for every observed pair $(d_{ijl}, d_{ijl'})$ and note that its expected value, $E[G_{ij,ll'}]$ - which is equal to the covariance between $Y_{ij}(d_{ijl})$ and $Y_{ij}(d_{ijl'})$ - is smooth over $(d_{ijl}, d_{ijl'})$ when $l \neq l'$. When $l = l'$ this expected value may be inflated by some positive constant σ_e^2; this could be viewed as some noise variance. It follows that we can obtain an estimator for the pooled covariance by using a bivariate smoother through the data $\{(d_{ijl}, d_{ijl'}), G_{ij,ll'}, i = 1, \ldots, n, j = 1, \ldots, m_i, l \neq l'\}$ and a working independence assumption. Let $\{D_\ell(d, d')\}_{\ell \geq 1}$ be a bivariate basis defined on $[0, 1] \times [0, 1]$ and assume that $E[G_{ij,ll'}] = \sum_{\ell=1}^{L} D_\ell(d_{ijl}, d_{ijl'})\gamma_\ell$, where γ_ℓ are basis coefficients. We estimate the basis coefficients by minimizing a penalized criterion that is similar to the one used for estimating the univariate smooth function $\mu(d)$, with the difference being that Y_{ijl} is replaced by $G_{ij,ll'}$ and the basis representation as well as the penalty are replaced by the ones corresponding to this setting (Wood, 2006; Eilers and Marx, 2003). A computationally faster alternative is to first obtain an estimate of the pooled covariance, called a raw pooled covariance estimator, by averaging across i and j for all observed pairs and then obtain the final pooled covariance estimator by passing a bivariate smoother through this pooled raw covariance estimator; see Di et al. (2009) and Goldsmith et al. (2013) who used this approach for a covariance estimator of a sample of independent functional observations. We used 100 bivariate basis functions obtained from a tensor product of two univariate bases, each with 10 functions. As in Yao et al. (2005) and Staniswalis and Lee (1998) the final estimator is adjusted to be symmetric and positive semidefinite by zeroing all the negative eigenvalues; this estimation allows us to estimate the noise variance σ_e^2. Let $\widehat{\Xi}(d, d')$ be the estimated pooled covariance and let $\{\widehat{\phi}_k(\cdot), \widehat{\lambda}_k\}_k$ be the pairs of eigenfunctions/eigenvalues corresponding to the spectral decomposition of this covariance. Denote by K the finite truncation determined by a percentage of explained variance equal to some fixed value. Common thresholds used in the literature are 90% or 95% (Di et al., 2009); we use a 95$\%$ threshold for the percentage of variation explained in our data application.

Figure 2 shows the leading annual directions in which nitrate varies for the CSN sites (middle panel) and the IMPROVE sites (rightmost panel). The number of directions selected to explain 95% of the variance is $K = 3$ for both CSN and for IMPROVE. Overall, for both the CSN and the IMPROVE network the top estimated directions seem to be related to the seasonality of the nitrate variation; this seasonality-related variation is in agreement to previous findings in the literature that studied nitrate among other components of $PM_{2.5}$ variation over time (Bell et al., 2007).

The first direction accounts for 77% of the total variance for CSN-nitrate and 83% of the variance for IMPROVE-nitrate. The first direction for CSN-nitrate is positive and roughly constant throughout the year with a decrease during the summer, and therefore generally represents a random effect for site within year. Sites with positive values for this direction tend to have an average annual nitrate level that is higher than that of the U.S. average. For the IMPROVE network, the first direction is positive and shows a more noticeable dip during the months from April to October, implying that the sites with a positive coefficient for this direction tend to have higher average annual nitrate levels than the U.S.

For both networks, the second direction looks similar and seems to indicate that the next most important direction of variation in nitrate is related to the contrast between the pollutant levels in the warm months and cold months. Specifically, it appears that sites with larger magnitude coefficients along this direction experience more seasonality - larger differences in the pollutant level between winter and summer - than the U.S. average corresponding to each network in part. The analysis of nitrate also depicts a third direction that is positive throughout the year except during the spring months implying a larger difference between the pollutant levels in the spring and those in the remaining months of the year.

4.3. Spatiotemporal variation

Once the mean function $\mu(d)$ and the orthogonal directions $\{\phi_1(d), \ldots, \phi_K(d)\}$ are estimated in (1), the basis coefficients ξ_{kij} can be predicted by $\int_0^1 \{Y_{ij}(u) - \widehat{\mu}(u)\}\widehat{\phi}_k(u)du$ which can be approximated numerically if the curves $Y_{ij}(\cdot)$ are observed at fine grids of points. Nevertheless as we specified in the beginning, the common protocol of sampling in our case is one every three days. However, there are many sites with considerably fewer, such as two or three, observations per year. To accommodate such designs, we predict ξ_{kij} in a mixed model framework-based approach.

Specifically, consider the following model, $\widetilde{Y}_{ij}(d) = \sum_{k=1}^{K} \phi_k(d)\xi_{kij} + e_{ij}(d)$ where we assume that the centered response $\widetilde{Y}_{ij}(d)$ is the outcome, $\phi_k(\cdot) = \widehat{\phi}_k(\cdot)$'s are known, fixed quantities, $\xi_{kij} \sim N(0, \lambda_k)$ are unknown random variables modeled to be independent over i, j and k with known variance $\lambda_k = \widehat{\lambda}_k$, and random noise $e_{ij}(d) \sim N(0, \sigma_e^2)$ independent over i, j, d with known variance $\sigma_e^2 = \widehat{\sigma_e^2}$. The assumption that the variance term $e_{ij}(d)$ is independent over d is made for convenience; $e_{ij}(d)$ should not be mistaken with the noise process $\epsilon_{ij}(\cdot)$ described by (1). The random components ξ_{kij} are predicted by conditional expectation $\widetilde{\xi}_{kij} = E[\xi_{kij}|\widetilde{Y}_{ij}]$; a simple closed form expression is available by using independence and normality assumptions.

Figure 3 shows the predicted basis coefficients for the first component, $\widetilde{\xi}_{1ij}$, for nitrate levels in 2003 and 2015 for the CSN sites (left panels, using filled circles) and IMPROVE sites (right panels, using filled triangles). The absolute magnitude of these loadings is reflected by the dimension of the circle or triangle and their sign is depicted by color (red for $+$ and blue for $-$). For CSN sites, there is a clear spatial trend in the sign of the coefficient values. We see generally positive coefficients in the Midwest and California, implying higher annual nitrate levels than the overall U.S mean at these CSN sites in 2003. By 2015, the region of positive coefficients for CSN has condensed indicating areas of possible decreasing trends in nitrate levels over this period. While the loadings of the CSN sites appear to be spatially correlated, the correlation is not nearly as strong for the IMPROVE sites. In fact this observation holds true for the loadings of the other directions that explain the nitrate variation across the U.S. We also note that the estimated variance of the random noise $\hat{\sigma}_e^2$ is 0.06 for the CSN and 0.02 for the IMPROVE network capturing the day-to-day specific variability in log-nitrate levels.

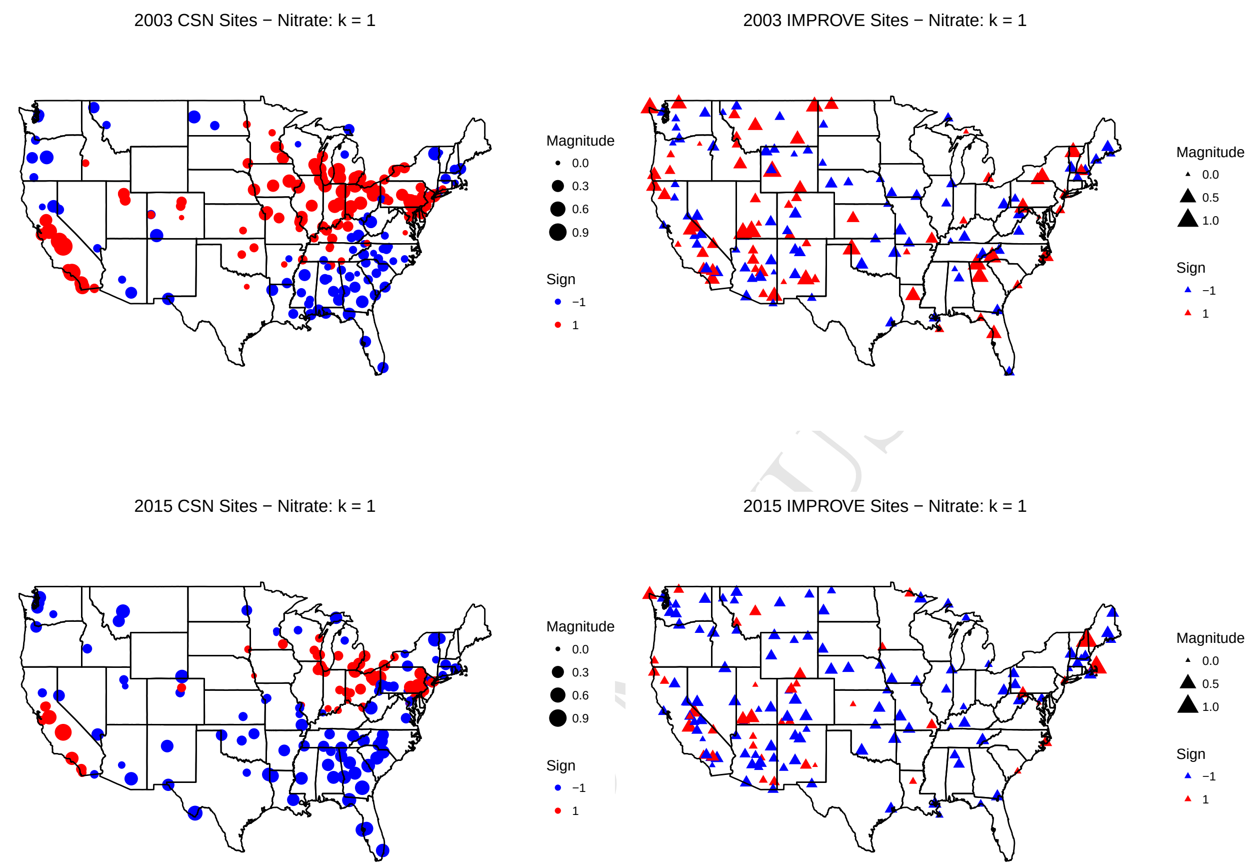

Figure 3: Preliminary predicted loadings for the first direction for nitrate variation in 2003 (top panels) and 2015 (bottom panels): CSN sites (left panels) and IMPROVE sites (right panels).

These preliminary estimates of $\widetilde{\xi}_{kij}$ allow us to understand the variation of nitrate solely for the sites and the years at which observed data are available. This is a limitation as many sites in our data do not have nitrate level measurements for all the 13 years from 2003 to 2015. Therefore it is preferable to use an approach that would allow us to predict these coefficients for years and sites within the respective network that have not been observed. For this purpose we use $\widetilde{\xi}_{kij}$ to gain insight into the space-time correlation in the data corresponding to each direction and in turn make predictions for the locations and times at which data are not available.

4.4. Prediction of annual nitrate

For each k, consider the "pseudo" data $\{\widetilde{\xi}_{kij}, t_{ij}, s_i : j = 1, \ldots, m_i, i = 1, \ldots, n\}$. In other words, we now treat the scores and corresponding sites and years as a new data set. Assume a working normal distribution with zero-mean and space-time parametric covariance model, as detailed in Section 3. Specifically, for each k assume a spatiotemporal behavior for $\widetilde{\xi}_{kij}$ as described by (2), using a random intercept and random slope, which comes down to assuming the following covariance model:

$$\mathrm{cov}(\widetilde{\xi}_{kij}, \widetilde{\xi}_{ki'j'}) = \sigma_{ka}^2 \rho_{ka}(\|s_i - s_{i'}\|) + t_{ij} t_{i'j'} \sigma_{kb}^2 \rho_{kb}(\|s_i - s_{i'}\|) + \sigma_k^2 I(i = i', j = j'). \quad (4)$$

Denote $I(\cdot)$ as the indicator variable that is equal to one if $i = i'$ and $j = j'$. Here σ_{ka}^2 and ρ_{ka} describe the variance and spatial dependence of the intercept and σ_{kb}^2 and ρ_{kb} describe the same characteristics of the slope. The dependence between coefficients as described by (4) may be unnecessarily complex for larger k. We propose the use of an information criterion to select among nested covariance models.

Before estimating the model parameters implied by (4), we conducted a preliminary investigation to check the assumptions made about the temporal and spatial dependence. For example to check the assumption of isotropy, we considered the sample semivariograms of the $\widetilde{\xi}_{kij}$'s for different angles between sites at fixed years. The results from 2003 for both networks are located in Figure 5 in the Supplementary Materials. The semivariograms for the IMPROVE network appear relatively consistent over different angles. However, there is some evidence of anisotropy for CSN. For this analysis we will continue to use the isotropic covariance model for CSN. The high spatial correlation between neighboring sites should lessen the potential effect of model misspecification.

In the future it could be helpful to consider using an anisotropic covariance function or incorporating covariates into the covariance function for the CSN. Examination of the omnidirectional sample semivariograms for fixed years showed that it is reasonable to assume $\rho_{ka}(\cdot)$ and $\rho_{kb}(\cdot)$ to be double exponential correlation functions (Rasmussen and Williams, 2006) with parameters δ_{ka} and δ_{kb} respectively. For example, $\rho_{ka}(\Delta) = \exp\left(-\Delta^2/2\delta_{ka}^2\right)$ where Δ is the distance between sites measured in kilometers and the parameter δ_{ka} is proportional to the spatial correlation range. Thus, larger values of δ_{ka} or δ_{kb} indicate higher spatial correlation. Residuals from initial fits of this model also indicated that the assumption of independence for the e_{kij} is reasonable.

Maximum likelihood estimation is used to estimate the model parameters for each network and k in part; the parametric modeling framework also allows us to calculate standard errors of the estimates. For both networks, it appears that the dependence of the $\widetilde{\xi}_{kij}$'s for $k = 3$ is somewhat less complex than for $k = 1, 2$. Thus we consider gradually simpler covariance models: the first covariance model is described by (4); the second model assumes (4) with $\delta_{kb} = 0$; finally a third model assumes (4) with $\delta_{kb} = 0$ and $\sigma_{kb}^2 = 0$. If $\delta_{kb} = 0$ it implies that the dependence of $\widetilde{\xi}_{kij}$, cov($\widetilde{\xi}_{kij}$, $\widetilde{\xi}_{ki'j'}$), for any two different locations, i and i', remains constant over time. If in addition $\sigma_{kb}^2 = 0$ then the dependence of $\widetilde{\xi}_{kij}$ is the same for any two different years.

We use Akaike information criterion (AIC) for covariance model selection. For CSN, we adopt (4) for $k = 1, 2$ and then assume $\delta_{kb} = 0$ for $k = 3$. For the IMPROVE network, we also use (4) for $k = 1, 2$ and assume $\delta_{kb} = 0$. Details about the reduced models and results of the AIC comparison can be found in Section 3.1 of the Supplementary Materials.

Table 1 shows the parameter estimates for the final covariance models and their associated standard errors in the case of nitrate for both networks and for each annual direction of variation.

The expression of the likelihood and its partial derivatives are not trivial; we use a numerical approximation of the Hessian to calculate the Fisher information matrix and thus estimate the standard deviations of the parameter estimates. We see that for the first direction ($k = 1$) the spatial correlation parameters are larger for the CSN than the estimates for the IMPROVE network implying more spatial dependence between CSN sites. In the Supplementary Materials, Section 3.2 discusses the interpretation of the spatial correlation parameter in the context of our problem. For example, CSN sites within around 172 km of each other will have spatially correlated random intercepts. On the other hand, IMPROVE sites need to be within around 1 km of one another to exhibit any spatial correlation between random intercepts. This aligns with what we saw in Figure 3 where there are large clusters of similar loadings for the CSN sites whereas the clusters are much smaller for the IMPROVE sites in 2003. It is also interesting to note that when $k = 1$ the variability of the intercepts and slopes is roughly equal for the IMPROVE and CSN sites. For CSN, the coefficients corresponding to the second and third principal directions of variation ($k = 2$ and $k = 3$) exhibit strong spatial correlation between random intercepts, though of course their variability decreases with k.

Network	k	$\hat{\sigma}^2_{ka}$	$\hat{\delta}_{ka}$	$\hat{\sigma}^2_{kb}$	$\hat{\delta}_{kb}$	$\hat{\sigma}^2$
CSN	1	0.07	43.39	0.03	427.51	0.01
		(0.007)	(2.940)	(0.009)	(56.824)	(0.000*)
	2	0.01	155.09	0.00*	288.55	0.00*
		(0.001)	(13.871)	(0.001)	(43.757)	(0.000*)
	3	0.00*	331.30	0.00*	N/A	0.00*
		(0.001)	(31.430)	(0.001)	N/A	(0.000*)
IMPROVE	1	0.09	0.00*	0.02	0.00*	0.01
		(0.011)	(0.000*)	(0.002)	(0.000*)	(0.000*)
	2	0.01	9.87	0.00*	27.72	0.00*
		(0.001)	(5.172)	(0.001)	(10.784)	(0.000*)
	3	0.00*	10.84	0.00*	N/A	0.00*
		(0.000*)	(5.482)	(0.000*)	N/A	(0.000*)

Table 1: Maximum likelihood estimates of the spatiotemporal covariance parameters separated by network and direction for nitrate. Standard errors for the estimates are found below in parentheses. Values denoted with an asterisk are rounded to zero, but their estimated values are not zero.

The estimated model covariance, obtained from the assumed covariance model with the estimated covariance parameters, and the normality assumption allow us to predict the basis coefficients $\xi_{kij} = \xi_k(s_i, t_{ij})$ for unobserved locations and years within the time frame studied. Specifically, for each k separately and each specific network, Kriging is used to predict the values of $\xi_k(s^*, t^*)$ corresponding to new network site s^* and year t^* (Cressie, 1993; Wackernagel, 2003). Thus, if we denote $\widehat{\xi}_k(s,t)$ to be the predicted temporal basis coefficients for the spatial location s and time t we can predict the full profiles by

$$\widehat{Y}(\cdot, s, t) = \widehat{\mu}(\cdot) + \sum_{k=1}^{K} \widehat{\phi}_k(\cdot)\widehat{\xi}_k(s,t). \tag{5}$$

where $\widehat{\mu}(\cdot)$ and the $\widehat{\phi}_k(\cdot)$ are as previously estimated in Sections 4.1 and 4.2.

Figure 4 displays the observed and estimated trajectories on the log-scale for the first CSN site in the data set which is located in southern Alabama from 2003 to 2006 from left to right. While the estimated profiles do not capture all of the day-to-day variability in nitrate levels, they successfully mimic the seasonal behavior of nitrate levels at this site.

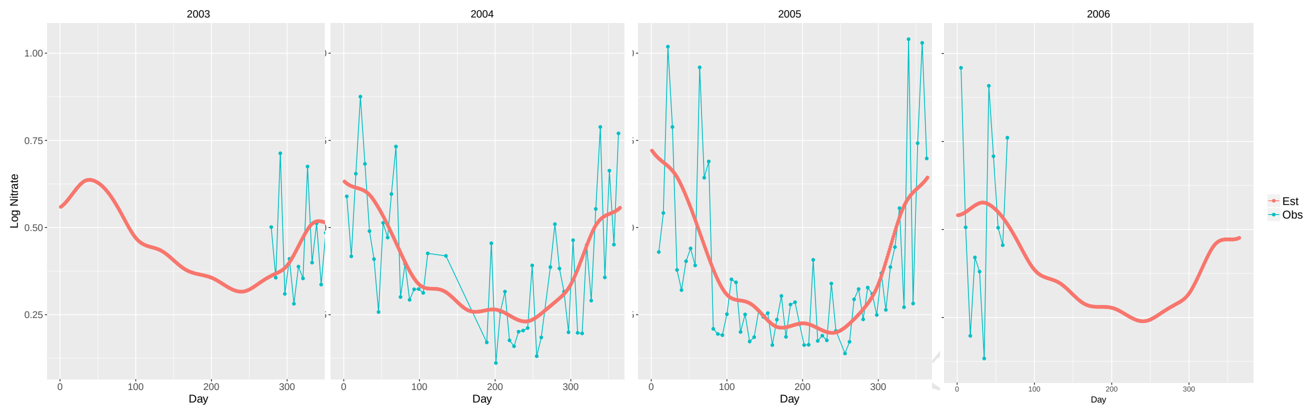

Figure 4: Observed and estimated nitrate levels on the log-scale for a CSN site in Alabama from 2003 to 2006 (from left to right).

Figure 4 highlights an important advantage of our approach. This site only has measurements from 2003 to 2006, and in that time period it was only observed for small portions of the year in 2003 and 2006. To compare the annual or seasonal nitrate levels across the U.S. and within the network, many existing methods use annual or seasonal averages (Bell et al., 2007; Pitchford et al., 2009). If monitoring sites are missing observations, the averages will be over differing numbers of days. For the site in Alabama, the annual average for 2003 would only include the 15 observations at the end of the year, when nitrate levels are at their highest and the 2006 average would utilize 11 observations at the beginning of the year. On the other hand, in 2004 and 2005 the annual average would include 49 and 59 observations, respectively, gathered over the entire year. Using our estimated trajectories we can avoid this issue and average over the entire annual profile. We investigated the prediction performance of our method in cases similar to 2006 in Figure 4, where the site was only observed for a portion of the year. Despite this lack of data for segments of the year, our trajectories still do a good job of predicting pollutant levels throughout the year. See Section 4.2 of the Supplementary Materials for description and results of this analysis.

Another feature of the data we investigated in the Supplementary Materials is potential nonstationarity and how this affects prediction. We tested for spatial stratified heterogeneity using a method proposed by Wang et al. (2016) based on the U.S. partition shown in Figure 7 of the Supplementary Materials. We found annual site averages over the period of study and conducted the test on the averages for each year. For all tests, we found evidence of spatial stratified heterogeneity at an $\alpha = 0.05$ significance level. Additionally, when we accounted for multiple testing and utilized a Bonferroni correction for these tests, we still rejected the null hypothesis for every year. The spatial stratified heterogeneity represents one type of nonstationarity that is present in the data. We also explored the potential differences in the correlation across these regions. For example, in the case of CSN nitrate, the regions differed in the number of estimated principal components. The Northwest region resulted in $K = 4$, while the Northeast and Southeast needed only $K = 2$. For additional comparison we constructed 95% confidence intervals for covariance parameters. Figure 7 of the Supplementary Materials also shows the intervals for the spatial range parameter of the random intercepts for $k = 1$, δ_{1a}, for the CSN nitrate data. While the uncertainty associated with these estimates varies across region due to the different sampling densities in each area, there are some noticeable regional differences in this range parameter. However, we investigated regional models in an analysis described in Section 4.4 of the Supplementary Materials, and in most cases the regional and overall models resulted in similar predictions.

Due to the multi-step estimation procedure, standard errors for predictions are difficult to estimate. However, we propose the use of a simplifying assumption to calculate standard errors for predictions. If we consider $\widehat{\mu}(d)$, $\widehat{\phi}_k(d)$ and the estimated covariance parameters to be fixed quantities, then the variance of a daily prediction is solely a function of the variance of the predicted scores and the errors. Specifically, the standard error for a prediction of $\widehat{Y}(d,s,t)$ for an unobserved site or year will be $\sqrt{\sum_{k=1}^{K}\widehat{\phi}_k^2(d)Var\{\widehat{\xi}_k(s,t)\}+\sigma_\epsilon^2}$ where $Var\{\widehat{\xi}_k(s,t)\}$ is the variance of the Kriging prediction for $\widehat{\xi}_k(s,t)$ (Cressie, 1993). We assume independence across k, so we do not have to account for covariance between $\widehat{\xi}_k(s,t)$ and $\widehat{\xi}_{k'}(s,t)$. Our pointwise prediction band will be calculated as

$$\widehat{Y}(\cdot,s,t)\pm z_{1-\alpha/2}\sqrt{\sum_{k=1}^{K}\widehat{\phi}_k^2(\cdot)Var\{\widehat{\xi}_k(s,t)\}+\sigma_\epsilon^2} \tag{6}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution. In the case of an observed site and year with missing daily measurements within the year, the prediction standard error for a given day would be the same except we would use $Var\{\widetilde{\xi}_k(s,t)\}$ as defined in (4) instead of $Var\{\widehat{\xi}_k(s,t)\}$. By ignoring some sources of estimation variability, we may underestimate the variance to a certain extent, but for large data sets the approximation should work fairly well.

4.5. Method performance

Using five-fold cross-validation, we can further assess the performance of our method for profile prediction. For each fold we use roughly 80% of the sites as training data and predict for the remaining 20% of sites. We are primarily interested in two settings: (1) prediction for a new, unobserved site and (2) prediction for a site that is observed one year but has no measurements for the remaining years of study. In the second case, we include the first year of measurements for the prediction sites in the training data and then predict for the remaining years at those sites. We calculate the mean-squared error (MSE) and mean absolute deviation (MAD) for daily predictions (Table 2) as well as seasonal average predictions (Table 3) on the log-scale. We include corresponding results for data on the original scale to aid interpretation, but in practice caution should be used when transforming predictions back to the original scale. For simplicity we divide each year into four "seasons" of 91 or 92 days yielding slightly different segments than the traditional seasons. For example, the winter average will be the average taken over days 1-92 while the fall average is over days 275-365. When comparing our predicted seasonal averages to the observed data, we only include sites that have 20 or more measurements in a given season so that we have an accurate seasonal average.

We compare our method (ST-FDA) to a k-nearest neighbors based approach (kNN) and a spatiotemporal Kriging method (STK). For a given prediction site s^* and year t^* the kNN approach takes the average annual profile of the k-closest sites to s^* observed during year t^*. Due to the every third day sampling procedure we then compute a k-day moving average to yield a complete predicted profile. This is easily the fastest method, requiring a few seconds for each fold. We report results for $k = 30$ which yielded the best kNN results, but we initially considered other choices of k. To apply the STK approach we utilize the `R` package, `gstat`, and consider the full time series for each site as in Sampson et al. (2011). Specifically, each site has a single time series of daily measurements from 2003-2015. Using the training data we first estimate a smooth mean function using penalized splines with 150 knots and also include site latitude and longitude as covariates. Then with the centered data we employ spatiotemporal Kriging separately for each month to make computation feasible. For each month, we calculate the sample variogram and then estimate model parameters by minimizing the squared difference between the sample and model variogram surfaces. After some initial investigation, we adopt a separable variogram model that is exponential in both space and time. Using the estimated variogram and training data for the current month, we then predict for the missing sites and days by Kriging. We primarily use the default settings of `gstat`, so it is possible the results for this method could be improved. Because calculating a sample variogram is computationally costly, the STK approach is by far the slowest method, taking around five hours per fold. Meanwhile, our ST-FDA approach requires a more reasonable 40-50 minutes per fold.

			New Site		Site observed 1 year	
Scale	Network	Method	MSE	MAD	MSE	MAD
Log	CSN	ST-FDA	0.18	0.31	0.15	0.29
		kNN, k = 30	0.21	0.34	0.21	0.34
		STK	0.24	0.36	0.24	0.36
	IMPROVE	ST-FDA	0.12	0.23	0.07	0.17
		kNN, k = 30	0.12	0.22	0.12	0.22
		STK	0.13	0.24	0.13	0.24
Original	CSN	ST-FDA	2.97	0.87	2.64	0.81
		kNN, k = 30	3.52	0.93	3.50	0.92
		STK	3.75	0.99	3.75	0.99
	IMPROVE	ST-FDA	1.19	0.41	0.74	0.33
		kNN, k = 30	1.19	0.40	1.16	0.39
		STK	1.19	0.42	1.19	0.42

Table 2: Average MSE and MAD for daily predictions for all folds on the log-scale and original scale for our method (ST-FDA), k-nearest neighbors approach (kNN) and spatiotemporal Kriging (STK) under two settings of missing observations

From Table 2 we see that our functional data analysis approach yields the smallest MSE and MAD of all the methods for CSN. This is especially true when we are predicting for a site at which we have measurements for one year. For prediction at a new site in the IMPROVE network, all methods perform about the same, but again we see our method benefits in the case when we observe data for one year and predict for the remaining years. However, as we saw in Figure 4, our daily predictions mimic the average behavior throughout the year and capture the day-to-day variability to a lesser extent.

			New Site				Site observed 1 year			
Scale	Network	Method	Win	Spr	Sum	Fall	Win	Spr	Sum	Fall
Log	CSN	ST-FDA	0.07	0.03	0.02	0.05	0.03	0.01	0.01	0.02
		kNN	0.12	0.05	0.03	0.07	0.11	0.05	0.03	0.07
		STK	0.11	0.05	0.04	0.07	0.10	0.05	0.04	0.07
	IMPROVE	ST-FDA	0.13	0.03	0.02	0.08	0.02	0.02	0.01	0.01
		kNN	0.12	0.03	0.01	0.08	0.12	0.03	0.01	0.08
		STK	0.13	0.03	0.02	0.08	0.12	0.03	0.02	0.08

Table 3: MSE of seasonal site averages on the log-scale for our method (ST-FDA), nearest neighbors approach (kNN) and spatiotemporal Kriging (STK) under two settings of missing observations.

Table 3 replicates the analysis in Table 2, but separately for each season; it shows the MSE for predicted seasonal averages for all the competing methods. For brevity we only include the results for the log-scaled data, but results on the original scale are included in Section 4.1 of the Supplementary Materials. Additionally, corresponding seasonal results for MAD are located in this section. On the log-scale, our method performs best for CSN and matches the kNN and STK approaches for IMPROVE when predicting for a new site. However when we return to the original data scale, the results do not indicate a uniform winner which is likely an effect of the transformation. Again ST-FDA noticeably improves in prediction accuracy when we observe one year of concentrations for a site of interest; see the block of columns under the label ‘Site observed 1 year.’

We also investigate the coverage of 90% prediction intervals for our method based on our proposed standard error estimation approach. The daily prediction intervals are slightly aggressive for CSN with coverages of 83% for new site prediction and 82% for prediction at a site observed one year. The corresponding average daily coverage for IMPROVE are 93% and 83%. The results for seasonal average prediction intervals are found in Table 4 and we generally maintain the desired coverage though there is some undercoverage as expected. These prediction intervals are calculated for the log-scaled data as standard errors are not easily transformed.

	New Site				Site observed 1 year			
Network	Win	Spr	Sum	Fall	Win	Spr	Sum	Fall
CSN	0.91	0.90	0.94	0.91	0.89	0.91	0.94	0.89
	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.03)
IMPROVE	0.93	0.94	0.94	0.94	0.88	0.83	0.86	0.91
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.02)	(0.02)

Table 4: 90% prediction interval coverage for seasonal average predictions on the log-scale. Corresponding standard errors are listed below in parentheses.

4.6. Spatiotemporal trend analysis

The model for ξ_{kij} in (2) includes a random intercept and random slope for each site. By fitting a random effects model to the scores with our estimated covariance model, we can gain a better understanding of the spatiotemporal trends in nitrate. Figure 5 contains the estimated random slopes for both networks for $k = 1$. In Figure 2 we noted that the first main direction of variation was positive and roughly constant throughout the year for both networks. All CSN sites have negative random slopes which indicates that nitrate levels at all sites in the network are decreasing to some extent from 2003-2015. For CSN, the smallest slopes are located at sites in California, coastal sites in the Northeast and sites in the Midwest, specifically those near Lake Michigan. Thus these sites experience the largest decreases in nitrate levels over the period of study. In their local analysis of the Bay Area of California, Fairley et al. (2011) also reported a decrease in nitrate levels from 2000 to 2009. For IMPROVE, most of the random slopes are also negative, but there are not strong regional trends.

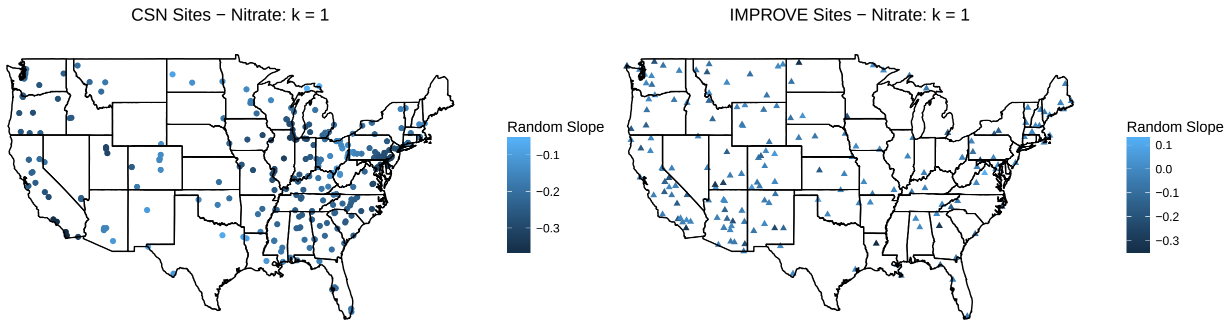

Figure 5: Estimated random slopes for the first direction of nitrate variation for CSN sites (left panel) and IMPROVE sites (right panel).

Predicting the annual level of additional pollutants across the U.S. at various times allows us to study the interplay between various pollutants over space and time in a more formalized manner. In particular by having the level of nitrate and sulfate at every day within the year, one can get a single number summary - such as annual average or average over a specific season - and calculate for each location and every year, the proportion of one pollutant relative to the combined pollutant level. Figure 6 depicts the average percentages of nitrate (blue color) at the state level for every year and every winter and summer for the CSN sites. The size of each pie reflects the combined pollutant level: larger pies correspond to states with higher average combined levels of nitrate and sulfate, while smaller pies correspond to states with lower average pollution levels of nitrate and sulfate. Generally, we see that the annual combined pollutant average of states decreases from 2003 to 2015 for CSN. However, we do see an increase in 2005 in the Midwest, especially in the winter that was also noted by Pitchford et al. (2009). Sulfate tends to comprise the majority of pollution totals in the summer, whereas nitrate is more dominant in the winter, especially for states in the Midwest and Northeast. This is additional evidence of the seasonal behavior of nitrate and sulfate found by Bell et al. (2007). The corresponding results for the IMPROVE sites are shown in Figure 7.

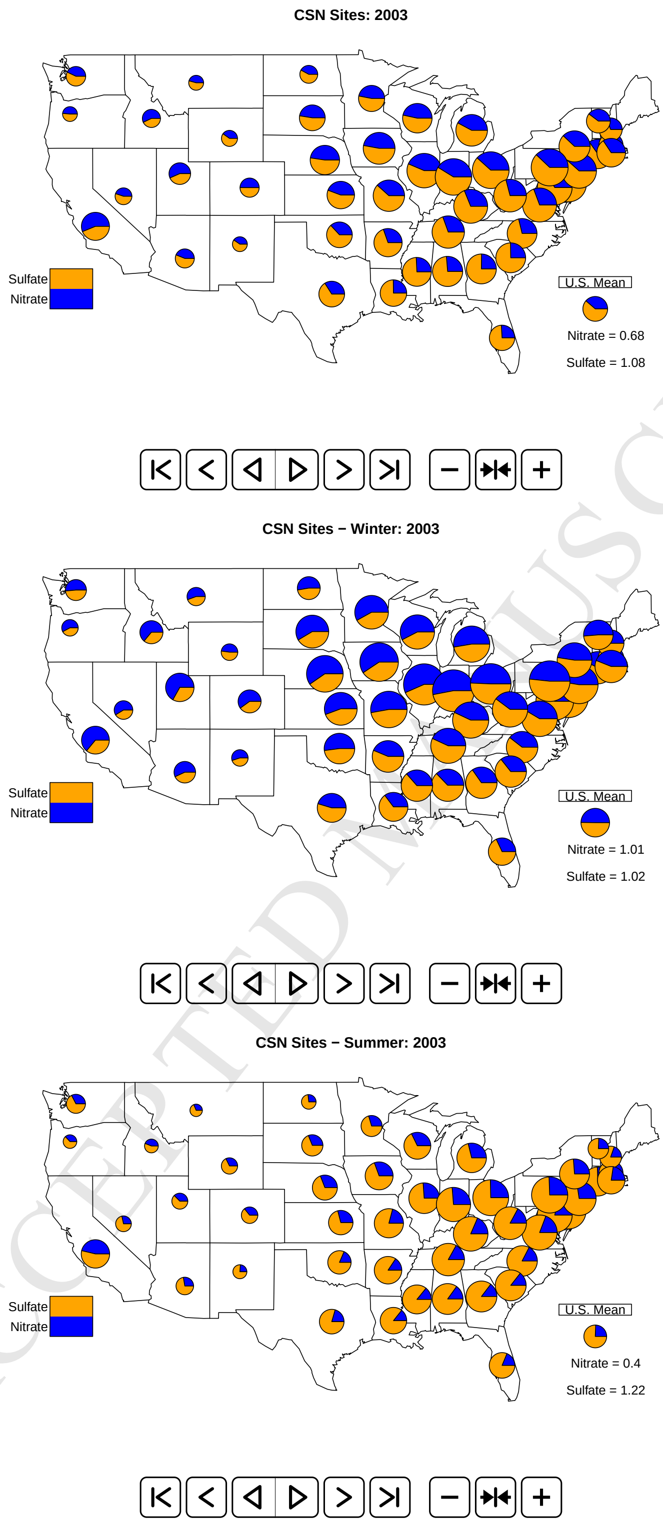

Figure 6: Annual (top panels) and seasonal (middle and bottom panels) average total (nitrate + sulfate) pollution levels for each state from 2003 until 2015 for CSN. Totals are on the log-scale. The sections of the pie represent the proportion of the total pollution accounted for by each pollutant. The radius of the circle represents the level of total pollution and is scaled appropriately so we can compare levels between the years and seasons.

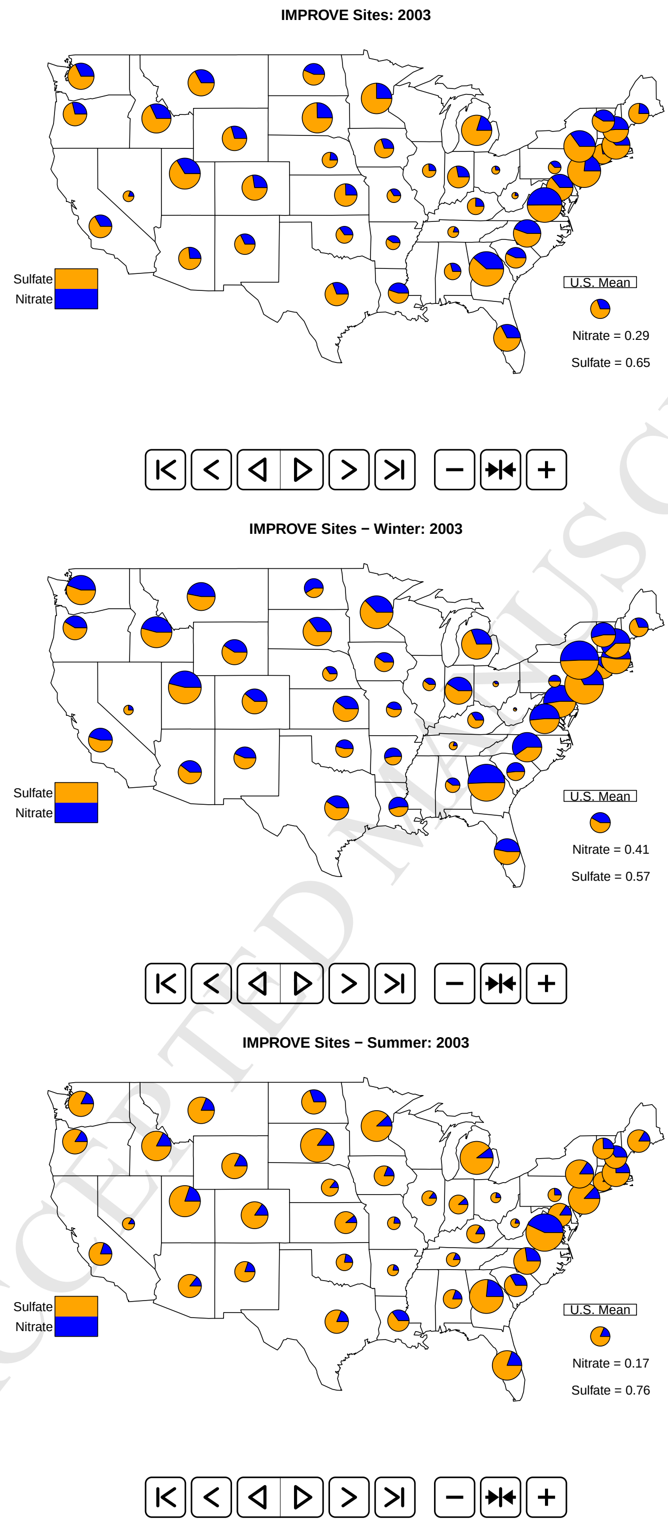

Figure 7: Annual (top panels) and seasonal (middle and bottom panels) average total (nitrate + sulfate) pollution levels for each state from 2003 until 2015 for IMPROVE. Totals are on the log-scale. The sections of the pie represent the proportion of the total pollution accounted for by each pollutant. The radius of the circle represents the level of total pollution and is scaled appropriately so we can compare levels between the years and seasons.

5. Software implementation

Our computational procedures can be divided into the three main estimation steps described at the beginning of Section 3. All of the analysis is carried out in `R` (R Core Team, 2013). To estimate the mean as discussed in Section 4.1 we utilize the `mgcv` package, specifically the flexible function, `gam()`. The `gam()` function allows us to fit the smooth mean function by using the function, `s()`, to define the smooth term for our model. Next we center the data and estimate the main directions of variation ϕ_k's and corresponding coefficients $\widetilde{\xi}_{kij}$'s by pooling all the data together, ignoring the dependence over space and time using functional principal component analysis tools implemented by the function `fpca.sc()` in the `R` package `refund` (Crainiceanu et al., 2014). Finally, while established software exists for the previous two steps, due to the complex nature of spatiotemporal data there are fewer resources to perform maximum likelihood estimation for covariance parameters and those that exist cannot accommodate our non-separable covariance function. Therefore, we had to develop our own code to carryout this procedure. Code for the complete analysis of nitrate can be found in the online supplementary materials.

6. Final remarks

This paper illustrates a functional data analysis methodology to gain insights into the variation of nitrate in the U.S. from 2003-2015, as measured by two main networks. The results for studying the variation of sulfate are included in the Supplementary Materials.

We make several modeling choices when applying our method to the $PM_{2.5}$ data. First, we opt to model the networks separately. Due to the similarity in the estimated eigenfunctions across the two networks, we also considered a joint model that assumes a network specific mean, a common eigenbasis across networks, and a common covariance model for the basis coefficients. It does allow for separate network monitoring errors. However, cross-validation analysis showed that the current modeling approach yielded more accurate predictions likely due to the differing levels of spatial correlation within the two networks. Details about this model and results of the analysis are included in Section 4.3 of the Supplementary Materials. Additionally, although there is some evidence of anisotropy or potential regional differences in $PM_{2.5}$ behavior, we choose to continue with our model for the entire continental U.S. with an isotropic covariance. In Section 4.4 of the Supplementary Materials we detail a cross-validation analysis in which we consider modeling regions separately. For the CSN, daily sulfate level predictions could be improved with a regional model in the Northeast and Southeast. However, in most cases the regional analysis resulted in similar prediction accuracy. Finally, we model the mean function only in terms of day within year, but including additional covariates could improve our approach.

Because of the health risks associated with $PM_{2.5}$, understanding the spatiotemporal behavior of nitrate and sulfate levels in the U.S. could help mitigate these key contributors to $PM_{2.5}$ concentrations. We presented a new approach to analyzing the $PM_{2.5}$ variability and change over space and time; the conclusions are consistent to other literature published in this area (Bell et al., 2007; Malm et al., 2004; Pitchford et al., 2009; Hand, 2011). However, our approach allows us to reconstruct the annual profile of the pollutants for every year under study and for any location in the continental U.S., allowing for a better understanding of the temporal trends in nitrate and sulfate levels. In addition, investigation of these complete estimated site profiles can potentially yield further insights about the various spatiotemporal trends in the behavior of pollutants in the U.S. While the $PM_{2.5}$ data represents one case where this functional data analysis approach could be beneficial, this process could also be applied to other similar data sets.

The proposed methodology allows us to analyze the variation of the pollutants across space and time. The precision of the estimates varies across the U.S. due to the differing densities of sites in each network. Regions where we have many sites will yield more precise estimates, whereas the standard error for a site average in an area with few neighboring sites would be larger. Much like Sampson et al. (2011), our multi-step estimation procedure complicates the estimation of standard errors. While we considered an approach to estimate standard errors, future work could focus on improving this effort by accounting for the lower bound of $PM_{2.5}$ concentrations and the uncertainty in each estimation step.

Acknowledgments

Dr. Staicu's research was funded by the NSF grant DMS 1454942. Dr. Reich's work was supported by the NIH grant R21ES022795-01A1. This research was also supported by the NIH grant T32 GM081057

C. A. Pope, D. W. Dockery, Health Effects of Fine Particulate Air Pollution: Lines that Connect, Journal of the Air & Waste Management Association 56 (6) (2006) 709–742.

W. Malm, B. Schichtel, M. Pitchford, L. Ashbaugh, R. Eldred, Spatial and monthly trends in speciated fine particle concentration in the United States, Journal of Geophysical Research: Atmospheres 109 (D3), ISSN 2156-2202.

G. Millar, T. Abel, J. Allen, P. Barn, M. Noullett, J. Spagnol, P. L. Jackson, Evaluating Human Exposure to Fine Particulate Matter Part II: Modeling, Geography Compass 4 (7) (2010) 731–749.

D. Liao, D. J. Peuquet, Y. Duan, E. A. Whisel, J. Dou, R. L. Smith, H.-M. Lin, J.-C. Checn, G. Heiss, GIS Approaches for the Estimation of Residential-Level Ambient PM Concentrations, Environmental Health Perspectives 114 (9) (2006) 1374–1380.

J.-H. Leem, B. M. Kaplan, Y. K. Shim, H. R. Pohl, C. A. Gotway, S. M. Bullard, J. F. Rogers, M. M. Smith, C. A. Tylenda, Exposures to Air Pollutants during Pregnancy and Preterm Delivery, Environmental Health Perspectives 114 (6) (2006) 905–910.

G. Hoek, R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, D. Briggs, A review of land-use regression models to assess spatial variation of outdoor air pollution, Atmospheric Environment 42 (33) (2008) 7561 – 7578.

J. Cyrys, M. Hochadel, U. Gehring, G. Hoek, V. Diegmann, B. Brunekreef, J. Heinrich, GIS-Based Estimation of Exposure to Particulate Matter and NO_2 in an Urban Area: Stochastic versus Dispersion Modeling, Environmental Health Perspectives 113 (8) (2005) 987–992.

O. Næss, P. Nafstad, G. Aamodt, B. Claussen, P. Rosland, Relation between Concentration of Air Pollution and Cause-Specific Mortality: Four-Year Exposures to Nitrogen Dioxide and Particulate Matter Pollutants in 470 Neighborhoods in Oslo, Norway, American Journal of Epidemiology 165 (4) (2007) 435–443.

X. Hu, L. A. Waller, M. Z. Al-Hamdan, W. L. Crosson, M. G. E. Jr., S. M. Estes, D. A. Quattrochi, J. A. Sarnat, Y. Liu, Estimating ground-level $PM_{2.5}$ concentrations in the Southeastern U.S. using geographically weighted regression, Environmental Research 121 (2013) 1 – 10.

Y. Liu, C. J. Paciorek, P. Koutrakis, Estimating Regional Spatial and Temporal Variability of $PM_{2.5}$ Concentrations Using Satellite Data, Meteorology, and Land Use Information, Environmental Health Perspectives 117 (6) (2009) 886–892.

V. J. Berrocal, A. E. Gelfand, D. M. Holland, A Spatio-Temporal Downscaler for Output from Numerical Models, Journal of Agricultural, Biological, and Environmental Statistics 15 (2) (2009) 176–197.

S. K. Sahu, A. E. Gelfand, D. M. Holland, Spatio-Temporal Modeling of Fine Particulate Matter, Journal of Agricultural, Biological, and Environmental Statistics 11 (1) (2006) 61–86.

B. G. Kibria, L. Sun, J. V. Zidek, N. D. Le, Bayesian Spatial Prediction of Random Space-Time Fields with Application to Mapping $PM_{2.5}$ Exposure, Journal of the American Statistical Association 97 (457) (2002) 112–124.

J. Zidek, L. Sun, N. Le, H. Özkaynak, Contending with Space-Time Interaction in the Spatial Prediction of Pollution: Vancouver's Hourly Ambient PM_{10} Field, Environmetrics 13 (5-6) (2002) 595–613.

R. L. Smith, S. Kolenikov, L. H. Cox, Spatiotemporal Modeling of $PM_{2.5}$ Data with Missing Values, Journal of Geophysical Research 108 (D24, 9004).

P. D. Sampson, A. A. Szpiro, L. Sheppard, J. Lindström, J. D. Kaufman, Pragmatic Estimation of a Spatio-Temporal Air Quality Model with Irregular Monitoring Data, Atmospheric Environment 45 (2011) 6593–6606.

H. O. Gao, D. A. Niemeier, Using functional data analysis of diurnal ozone and NOx cycles to inform transportation emissions control, Transportation Research Part D: Transport and Environment 13 (4) (2008) 221–238.

A. Park, S. Guillas, I. Petropavlovskikh, Trends in stratospheric ozone profiles using functional mixed models, Atmospheric Chemistry and Physics 13 (22) (2013) 11473–11501.

N. Shaadan, S. M. Deni, A. A. Jemain, Assessing and comparing PM10 pollutant behaviour using functional data approach, Sains Malaysiana 41 (11) (2012) 1335–1344.

S. Hörmann, L. Kidziński, M. Hallin, Dynamic functional principal components, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77 (2) (2015) 319–348.

S. Hörmann, P. Kokoszka, Functional time series, Handbook of statistics 30 (2012) 157–186.

EPA, Chemical Speciation Network (CSN), `https://www3.epa.gov/ttn/amtic/speciepg.html`, 2016.

M. L. Bell, J. M. Samet, F. Dominici, Time-Series Studies of Particulate Matter, Annual Review of Public Health 25 (1) (2004) 247–280.

F. Dominici, R. D. Peng, M. L. Bell, L. Pham, A. McDermott, S. L. Zeger, J. M. Samet, Fine particulate air pollution and hostpital admission for cardiovascular and respiratory diseases, JAMA: The Journal of the American Medical Association 295 (10) (2006) 1127–1134.

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

S. Y. Park, A.-M. Staicu, Longitudinal functional data analysis, Stat 4 (2015) 212–226.

O. Schabenberger, C. A. Gotway, Statistical Methods of Spatial Data Analysis, Chapman and Hall/CRC, 2004.

N. A. C. Cressie, Statistics for Spatial Data, Revised Edition, John Wiley & Sons, Inc., 1993.

M. L. Stein, Interpolation of Spatial Data: Some Theory for Kriging, Springer-Verlag, 1999.

R. J. Parker, B. J. Reich, J. Eidsvik, A Fused Lasso Approach to Nonstationary Spatial Covariance Estimation, Journal of Agricultural, Biological, and Environmental Statistics 21 (3) (2016) 569–587.

B. J. Reich, J. Eidsvik, M. Guindani, A. J. Nail, A. M. Schmidt, A class of covariate-dependent spatiotemporal covariance functions, The annals of applied statistics 5 (4) (2011) 2265–2687.

L. Horváth, P. Kokoszka, Inference for functional data with applications, vol. 200, Springer, 2012.

J. A. Aston, D. Pigoli, S. Tavakoli, Tests for separability in nonparametric covariance operators of random surfaces, The annals of statistics, to appear 1 (1) (2016) 40.

S. Wood, Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC, 2006.

P. H. Eilers, B. D. Marx, Flexible Smoothing with B-Splines and Penalties, Statistical Science 11 (2) (1996) 89–121.

J. Ramsay, B. Silverman, Functional Data Analysis, Springer, 2005.

T. Krivobokova, G. Kauermann, A note on penalized spline smoothing with correlated errors, Journal of the American Statistical Association 102 (480) (2007) 1328–1337.

F. Yao, H.-G. Müller, J.-L. Wang, Functional data analysis for sparse longitudinal data, Journal of the American Statistical Association 100 (470) (2005) 577–590.

P. H. Eilers, B. D. Marx, Multivariate Calibration with Temperature Interaction Using Two-Dimensional Penalized Signal Regression, Chemometrics and Intelligent Laboratory Systems 66 (2003) 159–174.

C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, N. M. Punjabi, Multilevel functional principal component analysis, The annals of applied statistics 3 (1) (2009) 458.

J. Goldsmith, S. Greven, C. Crainiceanu, Corrected Confidence Bands for Functional Data Using Principal Components, Biometrics 69 (1) (2013) 41–51.

J. G. Staniswalis, J. Lee, Nonparametric Regression Analysis of Longitudinal Data, Journal of the American Statistical Association 93 (1998) 1403–1418.

M. L. Bell, F. Dominici, K. Ebisu, S. L. Zeger, J. M. Samet, Spatial and temporal variation in $PM_{2.5}$ chemical composition in the United States for health effects studies, Environmental Health Perspectives 115 (7) (2007) 989–995.

C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.

H. Wackernagel, Multivariate Geostatistics, Springer Science & Business Media, 2003.

M. L. Pitchford, R. L. Poirot, B. A. Schichtel, W. C. Malm, Characterization of the Winter Midwestern Particulate Nitrate Bulge, Journal of the Air & Waste Management Association 55 (9) (2009) 1061–1069.

J.-F. Wang, T.-L. Zhang, B.-J. Fu, A measure of spatial stratified heterogeneity, Ecological Indicators 67 (2016) 250–256.

D. Fairley, S. Beaver, P. Martien, S. Tanrikulu, Trends in Bay Area Ambient Particulates, Tech. Rep. 201009-010-PM, Bay Area Air Quality Management District, Research and Modeling Section, 939 Ellis Street, San Francisco, CA 94109, 2011.

R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`, 2013.

C. Crainiceanu, P. Reiss, J. Goldsmith, L. Huang, L.and Huo, F. Scheipl, refund: Regression with functional data, r package version 3.0.2, 2014.

J. L. Hand, Spatial and Seasonal Patterns and Temporal Variability of Haze and its Constituents in the United States: Report V, 2011.

Highlights for “A Functional Data Analysis of Spatiotemporal Trends and Variation in Fine Particulate Matter

- A functional data analysis approach for spatiotemporal functional data is proposed
- The approach allows for complete profile prediction for sites or times without data
- The technique offers dimension reduction for easier data visualization
- The method confirms existing findings and yields new insights about $PM_{2.5}$ variation