Atmospheric Environment 45 (2011) 6593–6606

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data☆

Paul D. Sampson [a,*], Adam A. Szpiro [b], Lianne Sheppard [b], Johan Lindström [a,c], Joel D. Kaufman [d]

[a] Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA
[b] Department of Biostatistics, University of Washington, USA
[c] Centre for Mathematical Sciences, Lund University, Sweden
[d] Department of Environmental and Occupational Health Sciences, University of Washington, USA

ARTICLE INFO

Article history:
Received 15 November 2010
Received in revised form
1 April 2011
Accepted 8 April 2011

Keywords:
Exposure estimation
Hierarchical model
Nonstationary spatial covariance
Partial least squares
Particulate matter
Universal kriging

ABSTRACT

Statistical analyses of health effects of air pollution have increasingly used GIS-based covariates for prediction of ambient air quality in "land use" regression models. More recently these spatial regression models have accounted for spatial correlation structure in combining monitoring data with land use covariates. We present a flexible spatio-temporal modeling framework and pragmatic, multi-step estimation procedure that accommodates essentially arbitrary patterns of missing data with respect to an ideally complete space by time matrix of observations on a network of monitoring sites. The methodology incorporates a model for smooth temporal trends with coefficients varying in space according to Partial Least Squares regressions on a large set of geographic covariates and nonstationary modeling of spatio-temporal residuals from these regressions. This work was developed to provide spatial point predictions of $PM_{2.5}$ concentrations for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air) using irregular monitoring data derived from the AQS regulatory monitoring network and supplemental short-time scale monitoring campaigns conducted to better predict intra-urban variation in air quality. We demonstrate the interpretation and accuracy of this methodology in modeling data from 2000 through 2006 in six U.S. metropolitan areas and establish a basis for likelihood-based estimation.

1. Introduction

Statistical analyses of the health effects of air pollution have increasingly used GIS-based covariates for prediction of ambient air quality in "land-use" regression models. More recently these regression models have accounted for spatial correlation structure in combining monitoring data with land-use covariates. The current paper builds on these concepts to address spatio-temporal prediction of ambient concentrations of particulate matter with aerodynamic diameter less than 2.5 μm ($PM_{2.5}$) on the basis of a model representing spatially varying seasonal trends and nonstationary spatial correlation structures. Our hierarchical methodology provides a pragmatic approach that fully exploits regulatory and other supplemental monitoring data which jointly define a complex spatio-temporal monitoring design.

Our work has been developed for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). MESA Air is a cohort study funded by the U.S. Environmental Protection Agency (EPA) that emphasizes accurate prediction of intra-urban variation in individual exposures to ambient air pollutants in order to accomplish its primary aim of assessing the relationship between chronic exposure to air pollution and sub-clinical cardiovascular disease. The MESA Air cohort includes more than 6000 male and female subjects in six major U.S. metropolitan areas (Los Angeles, CA; New York, NY; Chicago, IL; Minneapolis-St. Paul, MN; Winston-Salem, NC; and Baltimore, MD). The primary MESA Air hypotheses relate to chronic exposure to $PM_{2.5}$. We note that final exposure estimates in MESA Air will integrate predictions of outdoor concentrations with additional subject-level data, including time-activity patterns, home infiltration characteristics, address history, and employment address (Cohen et al., 2009).

Our aim here is to provide point spatial predictions of long-term average concentrations at the residences of MESA Air subjects.

☆ Although the research described in this presentation has been funded in part by the United States Environmental Protection Agency through grant R831697 and assistance agreement CR-83407101 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. Additional support was provided by an award to the University of Washington under the National Particle Component Toxicity (NPACT) initiative of the Health Effects Institute (HEI).

* Corresponding author. Tel.: +1 206 685 2664; fax: +1 206 685 6419.
E-mail address: pds@stat.washington.edu (P.D. Sampson).

1352-2310/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2011.04.073

A primary source of monitoring data is the EPA's regulatory Air Quality System (AQS) repository (EPA, 2010). The AQS network includes a number of fixed site monitors in each region, each of which measures ambient air pollution levels on a regular basis, either daily, every third day or every sixth day in the case of $PM_{2.5}$. Although there are some missing data, most AQS sites provide nearly complete $PM_{2.5}$ concentration time series over several years at their spatial locations. The analysis here uses data for the time period 2000–2006.

The MESA Air study conducted supplemental air quality monitoring campaigns beginning in 2005 to provide additional concentration data in order to better model and predict intra-urban variation in air quality not well-represented by the AQS regulatory monitoring network. The objective of the MESA Air monitoring is to more completely sample a spatial design space that emphasizes traffic-related pollution at small spatial scales (100s of meters) and to capture data at actual subject home locations. For logistical reasons, the supplementary monitoring data are sampled as two-week averages based on an unbalanced design that results in significant amounts of missing data at many measurement locations (Cohen et al., 2009). In each region, the monitoring through 2006 includes: (a) from two to five fixed site monitors providing up to 1.5 years of 2-week observations, and (b) rotating sets of 2-week observations at (typically) 4 subject homes, with monitors moved every two weeks to cover a total of about 50 subject homes, each observed twice. MESA Air monitoring at fixed and home sites continued through July, 2009, providing a larger dataset for future modeling. The composite of the AQS and MESA Air monitoring data provide a rich but highly irregular spatio-temporal monitoring database for analysis.

Statistical approaches to the modeling of spatio-temporal air quality monitoring data for ambient exposure estimation typically derive from models that decompose observations into spatio-temporal *trend* and spatio-temporal *residuals,* or variation around the trend. A number of characteristics of the variation in air-quality data across the urban scales of interest in our current work motivate the modeling strategy and the approach to trend and residual that we present. We find that all air quality measurements demonstrate systematic time trends and seasonality, but these time trends vary in space, even over relatively small metropolitan area spatial scales. The details of the trend and seasonality vary somewhat from year to year, so that they are not modeled well by simple periodic functions like sinusoids or other trigonometric functions. Finally, characteristics of these systematic time trends, including the long-term mean and amplitude of seasonal variations, covary with spatial (or "land use") covariates. Even though the primary interest in MESA Air is predicting spatial variation of long-term average concentrations to estimate exposures, the complex spatio-temporal monitoring design necessitates a statistical modeling approach that accounts for spatio-temporal interactions in the data. Otherwise we cannot use the spatially rich but temporally sparse data at MESA Air homes to help in estimating the long-term averages.

For an overview of general techniques for modeling correlated spatio-temporal data, see Banerjee et al. (2004). See also the closely related strategy of Le and Zidek (2006). Smith et al. (2003) use an expectation-maximization (EM) algorithm to allow for arbitrary missing data patterns, but their model does not accommodate the complex spatio-temporal interactions addressed here. A recent paper by Fanshawe et al. (2008) demonstrates how carefully chosen covariates may eliminate the need to accommodate spatio-temporal correlation in the residuals, but the model in that paper assumes a uniform time trend across a relatively small spatial area. Paciorek et al. (2009) and Sahu et al. (2006) model particulate matter using techniques that allow for more complex spatio-temporal interactions, however their estimation and prediction procedures (based on a different approach to representing spatio-temporal trend and residual) are applicable only with a nearly complete (balanced) space-time monitoring data matrix.

The pragmatic modeling and prediction procedure described here includes sufficiently complex spatio-temporal interactions to account for variation in seasonal patterns at different locations, and it accommodates essentially arbitrary patterns of missing data with respect to an ideally complete space by time matrix of observations. We use an extensive database of GIS-based covariates in a spatio-temporal generalization of universal kriging in order to predict spatial variation in seasonal trends and in two-week ambient concentration levels. We compute predictions of concentrations at all MESA Air subject residences, and then compute estimates of long-term average concentrations as empirical averages of the predicted time series at these locations.

In Szpiro et al. (2010) we describe a likelihood-based version of the hierarchical model presented below and study its statistical properties by applying it to oxides of nitrogen (NO_x) in a simulation scenario based on a subset of the MESA Air geographic covariates, subject locations, and monitoring data. Lindström et al. (2011) extend this modeling to incorporate predictions from a source dispersion model as a spatio-temporal covariate, also for analyses of NO_x. The present paper focuses on a multi-step pragmatic approach to estimation and utilizes a more complete set of covariates and monitoring data to predict $PM_{2.5}$ concentrations at the homes of all MESA Air subjects. We use the Los Angeles study area as an example to illustrate the data structure and modeling approach. In order to make predictions at all MESA Air subject homes, we fit the model separately in four geographic regions covering the six MESA Air study areas: Southern California, a Midwest region spanning Minneapolis-St Paul and Chicago, a Northeast region spanning Baltimore and New York City, and North Carolina.

The outline of this paper is as follows. In Section 2 we describe the monitoring data as well as the geographic covariates that are available to inform predictions. Section 3 describes our hierarchical model, and Section 4 provides a detailed account of the pragmatic multi-step estimation and prediction procedure. In Section 5 we apply our model to predict concentrations at the home addresses of MESA Air subjects in all study areas. We illustrate the procedure using data from the Los Angeles region, and we also show predictions for subjects in all six MESA Air regions. Section 6 summarizes this work and discusses issues to be addressed in future research.

2. Data

2.1. AQS and MESA air monitoring data

We incorporate data from two types of AQS fixed site monitors, monitors recording $PM_{2.5}$ concentrations daily and monitors recording concentrations every third day. We use all 247 available "non-source oriented" monitors in counties nearby MESA Air subject locations that contain a minimum of at least 1 year of continuous data. We extended the spatial domain as necessary to include a minimum of 20 monitors around each MESA Air study city.

The MESA Air supplementary monitoring for $PM_{2.5}$ in each of the six study areas collects two-week average concentrations under two different spatio-temporal sampling plans: one for "fixed sites" and one for "home outdoor" sites. All of the locations at which data had been collected in the California region as of Dec 31, 2006 are shown on the map in Fig. 1.

There are a total of seven MESA Air fixed sites in the Los Angeles area, one of which is co-located with an AQS monitor to allow for

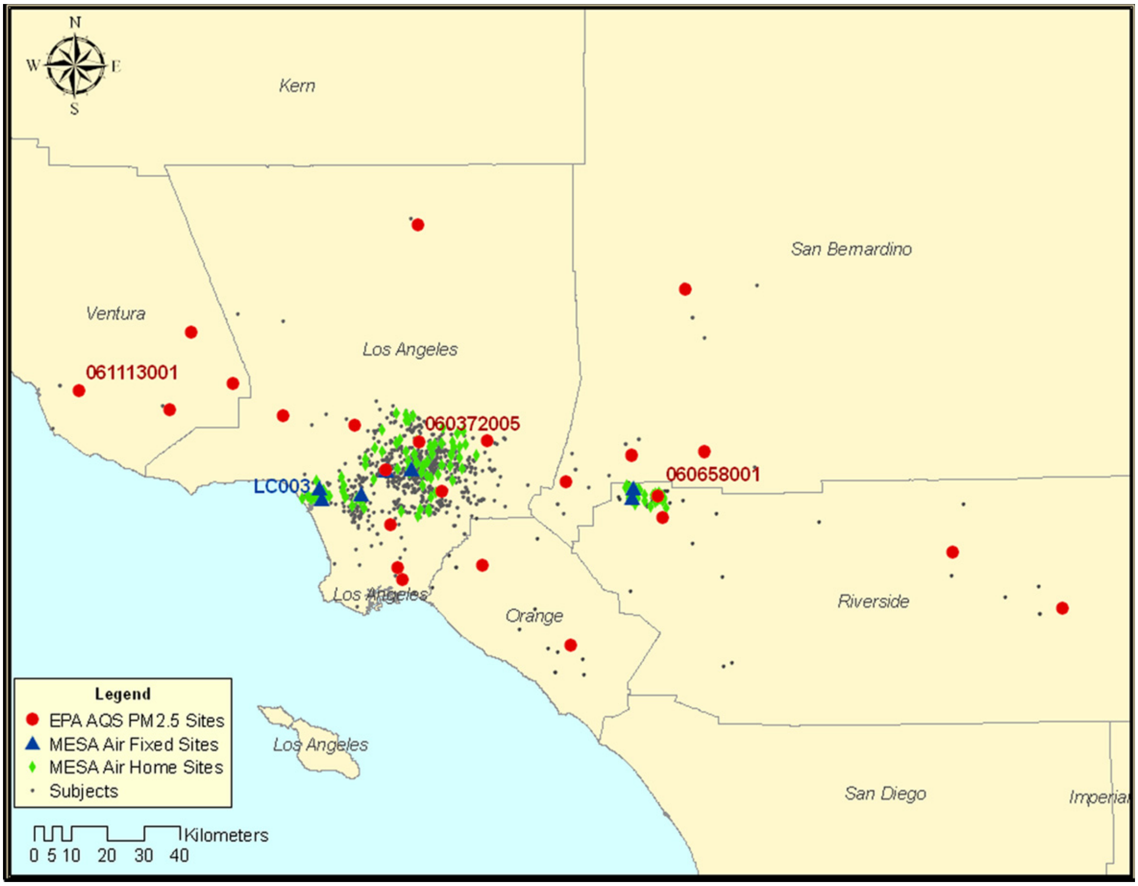

Fig. 1. Monitoring sites and subject home locations in the Los Angeles region.

instrument calibration. These fixed sites began measuring two-week average concentrations in November 2005. There were approximately 40 measurements per site and a total of 264 fixed site measurements during this timeframe. A total of 45 home outdoor monitoring locations in Los Angeles are also included; these were sampled during two-week periods starting in May 2006. The plan called for each home to be sampled two times, in different seasons. (Not all homes were sampled twice before the end of 2006, the closing date for the database for the analysis in this manuscript.) Fig. 2 presents a schematic illustration of the spatio-temporal sampling scheme combining the various AQS (EPA) and MESA Air monitoring sites.

For this analysis all AQS data are summarized at the 2-week time scale of the MESA Air monitoring campaigns. One practical feature of this data structure is that 2-week mean pollutant concentrations have far simpler temporal structure than daily data, which demonstrate high temporal autocorrelation, even after removing temporal trends. Fig. 3 shows four example time series on the 2-week time scale. We computed overlapping 2-week averages of $PM_{2.5}$ concentrations from AQS monitoring sites because the MESA Air monitoring periods in the Riverside area to the east were offset one week from the monitoring periods for central and coastal Los Angeles. These 2-week averages are centered on the Wednesday midpoints of the MESA Air 2-week sampling periods. We required at least 4 valid daily AQS observations for computation of a 2-week mean (actually a mean over 15 days). For this preliminary analysis we do not account for differences between monitor types or temporal sampling density (i.e. daily vs. every 3rd day). Monitoring sites sampling only every 6th day did not provide enough data to estimate valid 2-week averages.

The four $PM_{2.5}$ series in Fig. 3 present log transformations of two-week averages. The three AQS monitoring sites are located: in the Riverside area to the east (060658001), in the north central area of the Los Angeles concentration of MESA Air subjects (060372005), and to the northwest near the coast in Ventura County (061113001). We note similar, but slightly varying temporal trends as depicted by the smooth curves (explained in section 4) and variation in the long-term mean concentration which is highest in Riverside and lowest along the coast in Ventura County.

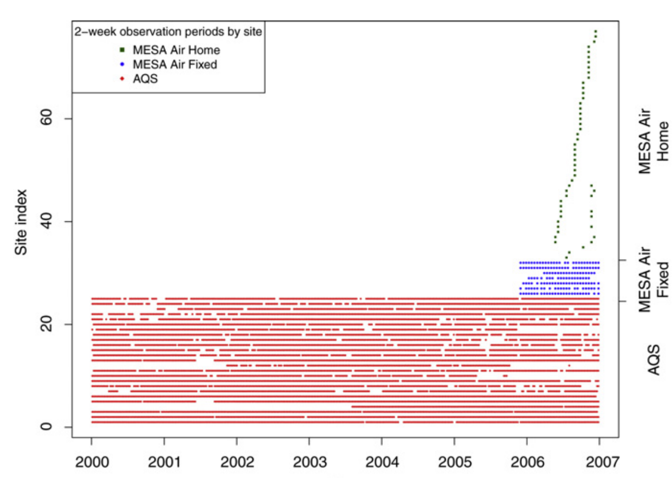

Fig. 2. Schematic of the temporal sampling pattern for AQS monitors and the MESA Air fixed and home sites. Each point in this figure represents a 2-week average measurement.

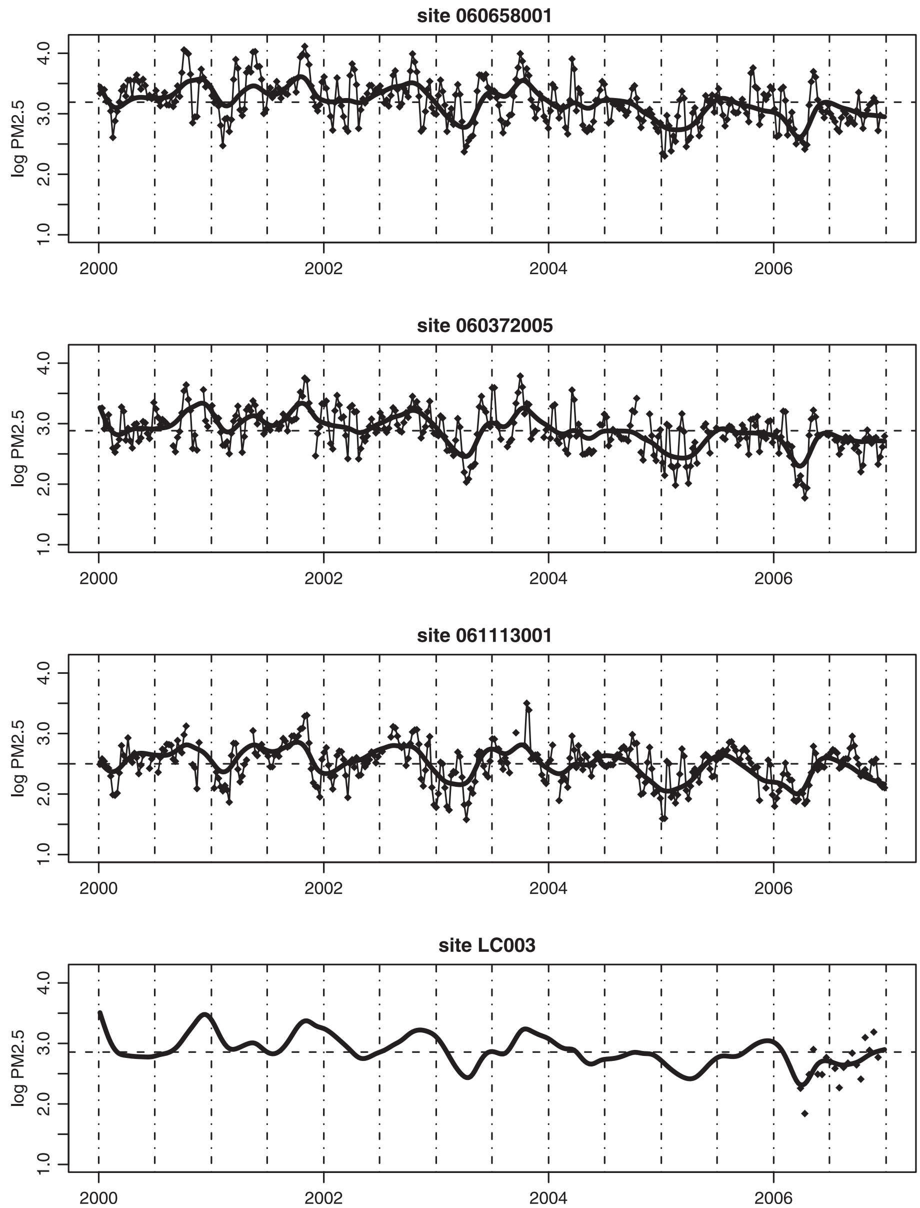

Fig. 3. Example log-transformed two-week average $PM_{2.5}$ data at AQS and MESA Air fixed sites in the Los Angeles region. The black points are measurements and the lines represent estimated temporal trends based on the SEOF model (see Section 4.2). The locations of each of these sites are shown on the map in Fig. 1.

The final short time series in Fig. 3 presents one of the MESA Air fixed sites in the coastal area of Los Angeles County. The black time trend drawn on this plot is largely determined by an average of the trends from the AQS sites nearest this MESA Air fixed site. The extrapolation of the time trend prior to the beginning of monitoring in 2006 has similar features to the trend curve for the central Los Angeles site, but is quite different from Ventura trend curve. Estimation of a long-term mean at MESA Air monitoring sites requires this spatially interpolated trend.

2.2. GIS-based geographical covariates

Our strategy for predicting concentrations at locations and times without measurements includes the use of regression models with geographic covariates. This is often termed "land use" regression (LUR) (Moore et al., 2007; Ross et al., 2007; Hoek et al., 2008). Our application of LUR is embedded in a hierarchical spatio-temporal model that incorporates flexible correlation structures. We consider a variety of geographic covariates, including: (i) indirect measures of traffic influences provided by distances to major roads (major roads identified by census feature class codes A1-A3), together with lengths of such roads in seven circular buffers from 50 to 750 meters around sites of interest, (ii) average population density (number of people per square km in the block group where the monitor or participant is located), and (iii) percentages of land in circular buffers described by various land use categories such as commercial property, cropland, industrial property, and residential property. These are all derived using the ArcGIS (ESRI, Redlands, CA) software package. The population density is calculated from publicly available U.S. Census Bureau data, and the roadway variables are derived from the proprietary TeleAtlas Dynamap 2000 roadway network. In total we considered approximately 200 possible covariates accounting for the road and land use variables measured in seven nested circular buffers. Covariates were screened prior to analysis and those with essentially no variability in a given study region (e.g. percent of forest in Los Angeles) were omitted. The number of covariates remaining for analysis ranged from 41 for Southern California to 66 for the Northeast region (including geographic coordinates derived from latitude and longitude).

3. Components of the space-time hierarchical model

Our statistical model is comprised of a spatio-temporal trend model and spatio-temporal residuals. This decomposition of space-time concentrations (for log-transformed 2-week averages) can be written

$$Y_{s,t} = \mu(s,t) + \varepsilon(s,t) \tag{1}$$

where s indexes space and t time, μ is the trend surface, and ε is the residual surface. Our approach accounts for spatial variability in temporal trends in order to make use of the complex spatio-temporal monitoring data from the combined AQS and MESA Air monitoring campaigns as well as a possibly nonstationary spatial covariance structure in the residuals. Our proposed model includes sufficiently complex spatio-temporal interactions to account for both variations in seasonal patterns at different locations and changes over time in the configuration of sites available for spatial predictions. It accommodates the fact that air quality measurements demonstrate systematic time trends and seasonality, that these time trends vary in space, even over relatively small metropolitan area spatial scales, and that the details of the trend and seasonality vary somewhat from year to year (and, hence, are not modeled well by simple sinusoids). Our notion of "systematic time trend" is not precisely defined, but our intention is to represent as "trend" both long-term (multi-year) changes in average concentrations and smooth intra-annual variation that is not easily represented with simple time series models. These trends might be driven by changes in emission patterns or possibly long scale climate patterns.

We write the temporal trend at location s as a linear combination of m smooth, orthogonal, temporal basis functions $f_j(t)$:

$$\mu(s,t) = \beta_{0s} + \sum_{j=1}^{m} \beta_{js} f_j(t) \tag{2}$$

We compute the $f_j(t)$ from data and refer to them as *smoothed empirical orthogonal functions* (SEOFs) (Fuentes et al., 2006). These basis functions are defined to have temporal averages equal to zero so that β_{0s} represents the long-term temporal mean. We will demonstrate that a linear combination of a small number of these SEOFs is sufficient to characterize the variation in temporal trends, resulting in residuals with negligible temporal correlation.

The coefficients of this model are the amplitudes of the temporal basis function patterns. They are modeled to vary systematically in space according to regressions on q_j spatial covariates assembled into an $N \times q_j$ design matrix:

$$\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{js_1} \\ \vdots \\ \beta_{js_N} \end{pmatrix} \sim N\left(\mathbf{X}\alpha_j, \Sigma_e\left(\phi_j\right)\right) = \sum_{k=1}^{q_j} X_k \alpha_{jk} + \mathbf{e}_j \tag{3}$$

where $s_1,\ldots s_N$ denote the spatial locations of the N monitoring sites, and X_k is the $N \times 1$ vector of values on the kth spatial covariate with regression coefficients α_{jk}. The spatial covariance model, $\Sigma_e(\phi_j)$, needs to be estimated for each of the $j = 1,\ldots,m$ spatial regressions (also called universal krigings). We fit standard exponential spatial correlation functions with nugget effects.

We assume the spatio-temporal residuals $\varepsilon(s,t)$ are temporally independent but spatially correlated with a common covariance for all time periods, written as

$$\boldsymbol{\varepsilon}_t = \begin{pmatrix} \varepsilon_{s_1,t} \\ \vdots \\ \varepsilon_{s_N,t} \end{pmatrix} \sim N(0, \Sigma_\varepsilon(\phi_\varepsilon)) \tag{4}$$

where the spatial covariance matrix $\Sigma_\varepsilon(\phi_\varepsilon)$, with spatial correlation parameters ϕ_ε, is computed using the Sampson–Guttorp spatial-deformation model for nonstationary spatial covariance (Damian et al., 2003). The applicability of this model is based on the assumption that the meteorological events that can drive high temporal autocorrelation structure on a daily time scale are largely averaged out on the 2-week time scale of the MESA Air supplemental monitoring. Our aim is to separate the spatially varying temporal trends from the residuals, thus leaving the residuals from the trend essentially uncorrelated in time.

4. Computational steps in estimation of the space-time model

The steps described below are: (1) the computation of Smoothed Empirical Orthogonal Basis functions (SEOFs), (2) the fitting of smooth temporal trends at each of the monitoring sites using these SEOFs, (3) analysis of spatial variation in the fitted smooth temporal trends by a Partial Least Squares approach to land use regression, (4) modeling of the spatial covariance structure of the residuals from the fitted spatio-temporal trends, and (5) cross-validated assessment of spatio-temporal predictions of $PM_{2.5}$ using a spatio-temporal universal kriging procedure (sometimes called "kriging with external drift").

4.1. Smoothed empirical orthogonal basis functions

The first step in fitting our hierarchical model to data is to derive the smoothed empirical orthogonal basis functions (SEOFs), $f_j(t)$, that we use to fit spatially varying temporal trend. Sampson introduced an approach to computing SEOFs in Fuentes et al. (2006). We review this approach here. Consider a spatio-temporal matrix $\mathbf{Y}$ of T observations (rows) at N locations (columns) in space and suppose that we want to approximate $\mathbf{Y}$ as a linear combination of m SEOFs. We write

$$\mathbf{Y} = \mathbf{M} + \mathbf{E} \tag{5}$$

where

$$\mathbf{M} = \mathbf{F}\boldsymbol{\beta}, \tag{6}$$

$\mathbf{F}$ being a $T \times m$ matrix with columns representing the values of m temporal basis functions, and $\boldsymbol{\beta}$ being an $m \times N$ matrix of coefficients.

$$\mathbf{F} = [f_0(t)\, f_1(t) \cdots f_m(t)]$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0N} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1N} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{m1} & \beta_{m2} & \cdots & \beta_{mN} \end{pmatrix}$$

The most parsimonious set of basis functions for a least squares minimization of $\mathbf{E}$ is obtained by taking $\mathbf{F}$ to be the matrix of the first m left singular vectors of the singular value decomposition (SVD):

$$\mathbf{Y} = \mathbf{UDV}' \tag{7}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $T \times N$ and $N \times N$ orthonormal matrices of left and right singular vectors, respectively, and $\mathbf{D}$ is an $N \times N$ diagonal matrix of singular values. The prime ($\mathbf{V}'$) indicates the matrix transpose. That is, using the superscript (m) to denote sub-matrices with m columns, write

$$\mathbf{Y} = \mathbf{U}^{(m)}\left(\mathbf{D}^{(m)}(\mathbf{V}^{(m)})'\right) + \mathbf{E} \tag{8}$$

This suggests that we take the matrix of empirical orthogonal functions, $\mathbf{F}$, to be the matrix of left singular vectors $\mathbf{U}^{(m)}$. However we wish our temporal basis functions to be defined as *smooth* functions of time, and the usual left singular vectors will not vary smoothly over the row (time) index. In addition, in practice every data matrix $\mathbf{Y}$ that we consider will have some missing data so that we cannot simply compute the usual singular value decomposition. We therefore embed the following "EM-like" procedure (*algorithm SVD.em*) for computation of an "SVD" of a matrix with missing data in a cross-validated smoothing loop to derive SEOFs (*algorithm SEOF*).

Algorithm *SVD.em*:

1. Specify a dimension (rank), m, for the model.
2. Scale the observations at each monitoring site (columns of $\mathbf{Y}$) to norm (variance) one; call this matrix $\tilde{\mathbf{Y}}$.
3. Compute an initial average temporal vector u_1 as the set of row averages of nonmissing values in $\tilde{\mathbf{Y}}$. Initialize missing observations in the data matrix $\tilde{\mathbf{Y}}$ using elements of a rank-one approximation provided by a regression (without intercept) of the data in each column (site) on u_1.
4. Compute the rank-m SVD-approximation of the now complete data matrix.
5. Impute the missing values in $\tilde{\mathbf{Y}}$ by the elements of the rank-m SVD approximation just computed.
6. Return to step 4 and iterate to convergence.

We specify the scaling in step 2 so that all sites contribute equally to the characterization of variability in patterns of temporal trend regardless of the amplitude of trends at the sites. This algorithm, coded in the R system (R Development Core Team, 2009), converges adequately fast in all of the applications we have faced. The computation of EOFs in the presence of missing data has also been addressed in the oceanographic literature (Beckers and Rixen, 2003).

As a practical approach to smoothing the EOFs, we compute smoothing spline regressions on the time index using the smooth.spline function in the R language (with generalized cross-validation specified by the argument cv = F). To choose the dimension of the SEOF model we use the Bayesian Information Criterion (BIC) computed for predictions of trend in the following cross-validation loop:

Algorithm *SEOF*:

In a cross-validation loop leaving out a random subset of sites in each iteration:

1. Compute the SVD using the *SVD.em* algorithm above.
2. Smooth vectors consisting of every other 2-week value of m left-singular vectors using smooth.spline (with generalized cross-validation specified by argument cv = F). Evaluate the fitted spline on every weekly observation using the R function predict.smooth.spline.
3. Compute the trend prediction of every site in the left-out cross validation group by least squares regression on the smoothed trend components and evaluate a BIC criterion for fit.

The smoothing in step (2) uses every other value so that the smoothing spline would not be applied to time series with temporal correlations artificially inflated by the overlapping of 2-week averages. The smoothed singular vectors are no longer exactly orthogonal, but exact orthogonality is not a concern. We choose as optimal the number of components m with best (or near-best) distribution of BIC values in the cross-validated fits.

Although the *SVD.em* algorithm can be run on matrices with arbitrary amounts of missing data, MESA Air home sites have too little data, only one or two observations, to permit meaningful multiple regressions on the smoothed temporal basis functions. Furthermore, as MESA Air fixed site monitors began operation in 2005 or 2006, these sites also have few observations compared to the AQS sites. For this reason the specification of the temporal trend functions just described was computed using only the AQS monitoring sites.

4.2. Trend fits at MESA air monitoring sites

Once the SEOFs are determined, the next step in our pragmatic procedure is to estimate values for the corresponding trend coefficients $\beta_{0s}, \beta_{1s}, \ldots, \beta_{ms}$ at each spatial location with monitoring data. Since there are nearly complete temporal data at each AQS monitor, we estimate the coefficients at these locations by linear regression of the data on the corresponding SEOFs.

As the data at MESA Air sites are more limited, we fit trends at these locations using information provided by AQS sites in a spatial neighborhood. In principal we might use a model like that of Banerjee and Johanson (2006), which we would call a "spatial random effects" model. However, the fitting of that model as currently implemented is computationally impractical for the size of our spatio-temporal datasets. We choose instead to use local random effects modeling strategy so that the fitted trend at a MESA AIR site is a simple empirical Bayes fit with shrinkage of trend model coefficients to the average trend of those AQS sites in a local neighborhood. We chose neighborhood sizes of 25–40 km in different regions in order to assure that at least three neighboring AQS sites were used in this fitting at each of the MESA Air sites. The random effects regression models were computed by REML using the lmer function of the lme4 package for the R system (http://cran.r-project.org/web/packages/lme4/).

The shape of the temporal trends determined by the coefficients $\hat{\beta}_{0s}, \hat{\beta}_{1s}, \ldots, \hat{\beta}_{ms}$ computed this way is necessarily determined almost entirely by the neighboring AQS sites as the MESA Air time series are too short to carry much weight. To assure that the MESA Air observations contribute as much as possible to estimation of the long-term average coefficients β_{0s}, rather than use the REML estimate of overall trend we used the estimated coefficients $\hat{\beta}_{1s}, \hat{\beta}_{2s}, \ldots, \hat{\beta}_{ms}$, to detrend the MESA Air observations, computing

$$Y_{st}^{d} = Y_{st} - \sum\nolimits_{j=1}^{m} \hat{\beta}_{js} f_j(t) \tag{9}$$

and then take the average of the detrended observations as an estimate of the long-term average

$$\hat{\beta}_{0s} = \frac{1}{T}\sum_{t} Y_{st}^{d} \tag{10}$$

Fitted trends derived from trend coefficients $\hat{\beta}_{0s}, \hat{\beta}_{1s}, \ldots, \hat{\beta}_{ms}$ computed by this procedure appear appropriate by visual inspection. In principle, estimated trend coefficients should be influenced by the spatial covariates according to the regression model of equation (3). That is the case in our likelihood-based approach to the hierarchical model (Szpiro et al., 2010; Lindström et al., 2011), but covariates are not incorporated at this stage of our sequential fitting procedure. Fitted trend coefficients are used as outcomes in the computation of spatial regression models as explained in the following sub-section.

4.3. Spatial regression by partial least squares (PLS)

Our hierarchical model proposes that the spatial variation in the parameters $\hat{\beta}_{js}$ of the temporal trend models can be modeled via regression on spatial covariates. The GIS-based dataset of spatial covariates for $PM_{2.5}$ concentrations provides substantial numbers of highly correlated spatial covariates. For example, the composite lengths of A1 roads in buffers of, say, 400 meters around specified locations are highly correlated with lengths of roads in 500 meter buffers and negatively correlated with the distance to the nearest A1 road. Percentages of property in various land use categories are similarly correlated across buffer sizes; for example, percentages of land classified residential in a buffer are substantially negatively correlated with percentages of land classified as commercial. Model specification with large sets of multicollinear predictors typically involves either (a) variable selection, (b) shrinkage or regularization, perhaps including variable selection as in a "lasso" approach (Tibshirani, 1996), or (c) regression on a smaller number of composite covariate scores. While our fundamental concern is the quality of predictions, we prefer not to choose a method that would select one particular buffer size for inclusion in a model ignoring neighboring buffer sizes, or one particular land use categorization at the expense of a correlated land use categorization. We choose, instead, to regress on a small number of composite covariate scores using the method of PLS regression to define the composite scores (see, for example, Garthwaite, 1994 or Abdi, 2010). The description of the composite scores in terms of individual variable loadings can be useful as it facilitates comparison of regression models across the four geographic modeling regions. PLS regressions were computed using the pls package for the R system (http://cran.r-project.org/web/packages/pls).

Almost all the numeric land use covariates considered here have skewed distributions and are log-transformed for analysis. Most of the spatial covariates fall into one of three groups: (1) shortest distances to roads and commercial properties, (2) lengths of roads of different classes (A1, A2, A3) within buffers, and (3) percentage of property in difference land use categories in buffers. We have chosen to log-transform all of the numerical scores except for the "angle" to A1/A2/A3 road variables and the "residential" land use variables, which span the entire 0-100 percent range. Log-transformations were computed after selection of an empirically determined constant to add in order to deal with zero scores. To be considered for analysis we require a covariate to have more than 5 non-zero observations.

The code in the pls package for the R system computes conventional leave-one-out cross-validatory assessments of predictions of these regression models. However, these cross-validations assume a modeling framework with spatially independent errors. Since our hierarchical model includes spatial correlation, we choose the dimension of the PLS regression by cross-validation with a universal kriging prediction involving component scores defined by the PLS algorithm and a model for the spatial covariance structure of the residuals from those regressions as indicated in equation (3). We computed spatial covariance matrices $\sum_e(\phi_j)$ in terms of exponential variogram models fitted to the residuals of the PLS regressions using the likfit function in the geoR package for the R system (http://cran.r-project.org/web/packages/geoR).

4.4. Residual covariance modeling

The hierarchical model of equations (1)–(4) involves a spatial covariance matrix $\sum_\varepsilon(\phi_\varepsilon)$ for the spatio-temporal residuals in equation (3). The structure of $\sum_\varepsilon(\phi_\varepsilon)$ is estimated using the Sampson–Guttorp deformation model for nonstationary spatial covariance fitted to an empirical spatial covariance matrix

Table 1
Site descriptive statistics for selected spatial covariates and average $PM_{2.5}$ concentration.

City – sites	#sites	Min #obs/site	Max #obs/site	#Sites < 150 m to A1 (%)		#Sites < 150 m to A3 (%)		Med dist to A1, A2, A3	Med dist to Commerce	Med Pop dens	Ave PM2.5 (sd)	
Los Angeles												
AQS	24	177	363	0	(0)	0	(0)	671	28	1621	17	(5)
Fixed	7	19	28	3	(43)	0	(0)	851	417	5103	17	(3)
Home	45	1	2	2	(4)	1	(2)	1049	274	3718	16	(3)
Chicago												
AQS	44	133	364	1	(2)	5	(11)	464	100	867	14	(2)
Fixed	7	7	34	2	(29)	0	(0)	338	0	1581	13	(2)
Home	65	1	2	3	(5)	1	(2)	727	130	3544	12	(3)
St. Paul												
AQS	41	88	365	1	(2)	2	(5)	902	188	1256	10	(2)
Fixed	3	27	29	1	(33)	0	(0)	764	130	2351	9	(1)
Home	37	1	2	2	(5)	4	(11)	461	697	2906	9	(3)
New York												
AQS	45	111	356	6	(13)	1	(2)	446	117	3207	13	(1)
Fixed	3	31	32	1	(33)	1	(33)	99	133	4403	13	(3)
Home	78	1	2	9	(12)	0	(0)	402	249	43819	14	(4)
Baltimore												
AQS	39	205	365	1	(3)	6	(15)	360	102	703	14	(1)
Fixed	5	14	33	1	(20)	0	(0)	210	156	844	14	(1)
Home	39	1	1	2	(5)	2	(5)	418	372	1809	15	(4)
Winston-Salem												
AQS	29	116	365	0	(0)	3	(10)	547	366	461	14	(1)
Fixed	4	18	35	1	(25)	0	(0)	208	478	481	14	(1)
Home	52	1	2	0	(0)	0	(0)	698	988	516	15	(4)

Notes: "Fixed" and "Home" refer to MESA Air supplemental fixed and home outdoor monitoring sites. Note that the simple average and standard deviation of reported $PM_{2.5}$ concentrations reported here for MESA Air sites are influenced by seasonality and unbalanced temporal sampling (see Fig. 2) that is not accounted for and the values for the MESA Air Home sites, in particular, are based on very small numbers of observations. "Med dist to A1, A2, A3" is the median over sites of the minimum distance to a major road of class A1, A2, or A3. "Med dist to Commerce" is the median over sites of the distance to the nearest commercial land use property. "Med Pop dens" is the median over sites of the block group population density.

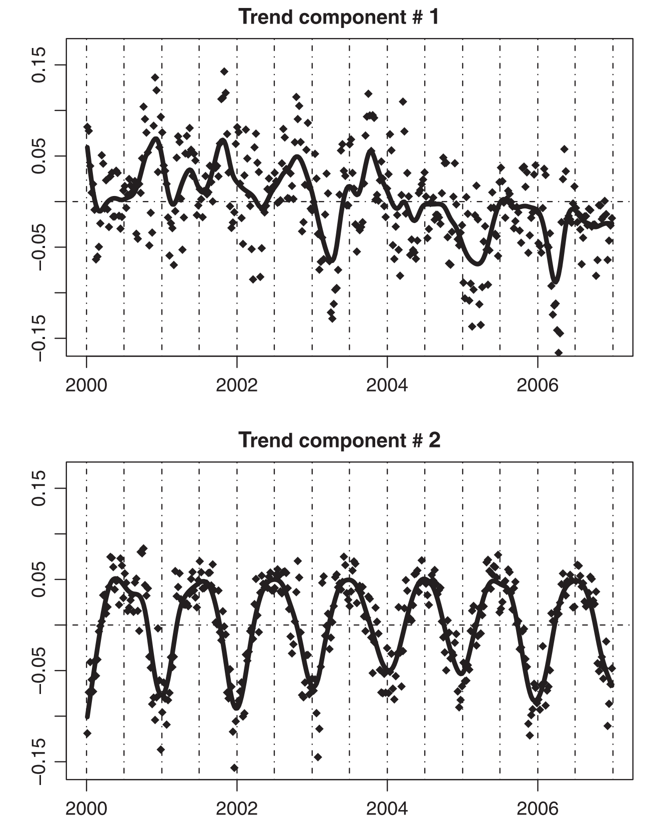

Fig. 4. Smoothed empirical orthogonal basis functions for log-transformed two-week average concentrations of PM2.5 in the Los Angeles area. The first smoothed component explains 27.5% of the variation in the matrix of log-transformed 2-week average PM2.5 concentrations while the second component explains only 9.4% of the variation.

computed from the spatio-temporal matrix of residuals from the site-specific trend models for the AQS monitoring sites. The spatial deformation model expresses spatial correlations as

$$\mathrm{cor}(\varepsilon(s_i,t),\varepsilon(s_j,t)) = \gamma_\theta(|d(s_i)-d(s_j)|) \tag{11}$$

where $d(s)$ is a smooth deformation of the coordinate system and $\gamma_\theta(h)$ is the exponential spatial correlation model with parameter θ. We fit this model, estimating θ and the smooth deformation $d(s)$ as a pair of thin-plate splines, using the Bayesian computations explained and illustrated in Damian et al. (2001, 2003).

4.5. Cross-validated spatio-temporal predictions

We predict concentrations at subject homes using the hierarchical model described above. Our methodology incorporates covariates and spatial interpolation of long-term averages and seasonal trends based a on spatial correlation model, a procedure that may be regarded as a generalization of "universal kriging" or "kriging with external drift" (Wackernagel, 2010). Specifically, given estimates of the parameters of the components of the hierarchical model as explained in sections 4.1–4.4, spatio-temporal predictions at target locations s_0 are computed as

$$\widehat{Y}_{s_0,t} = \widehat{\mu}(s_0,t) + \widehat{\varepsilon}(s_0,t) \tag{12}$$

where

$$\widehat{\mu}(s_0,t) = \widehat{\beta}_{0s_0} + \sum_{j=1}^{m} \widehat{\beta}_{js_0} f_j(t) \tag{13}$$

Each of the $\widehat{\beta}_{js_0}$ is computed by universal kriging using the PLS component scores as spatial covariates and the spatial covariance models underlying the estimates of the matrices $\sum_e(\phi_j)$. The spatio-temporal residual field $\widehat{\varepsilon}(s_0,t)$, being mean zero, is computed by a simple kriging calculation with the spatial covariances defined by the spatial deformation model underlying the

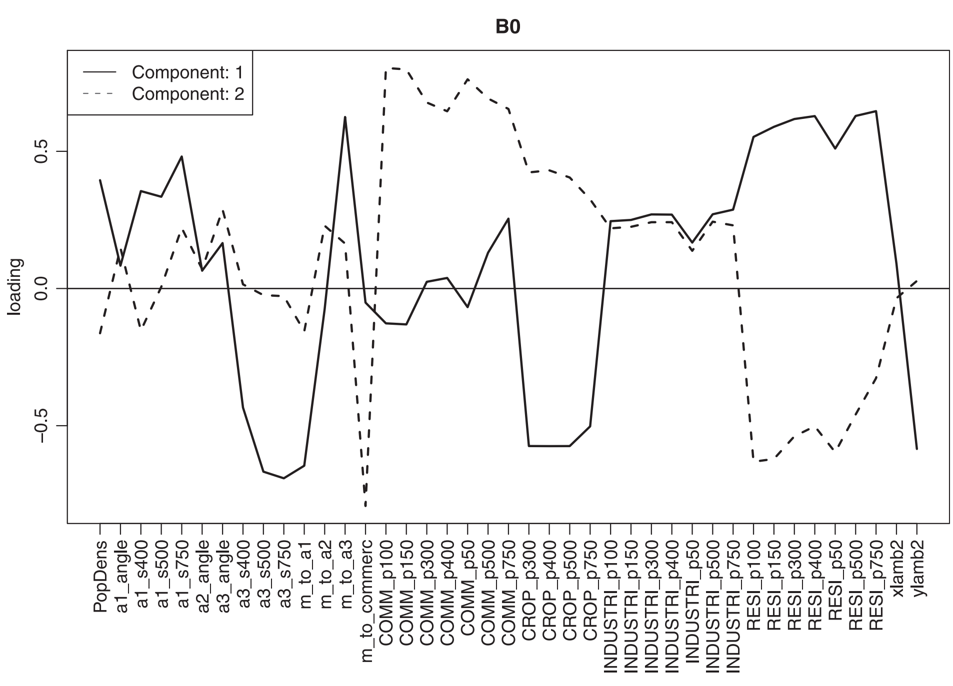

Fig. 5. Loadings (correlations) of the spatial covariates on the two Partial Least Squares (PLS) components for the spatial regression model for the intercept (long-term mean), β_{0s}, for the Los Angeles study region. The a1, a2, and a3 variables refer to major road census feature class codes; they include the angle to the nearest such roadway, the lengths of roadway segments in circular buffers of varying radii, and the distances in meters to the nearest roadway. The variables beginning COMM, CROP, INDUSTRI, and RESI refer to fractions of the property in circular buffers designated as Commercial, Cropland, Industrial, and Residential using the ArcGIS software system (ESRI).

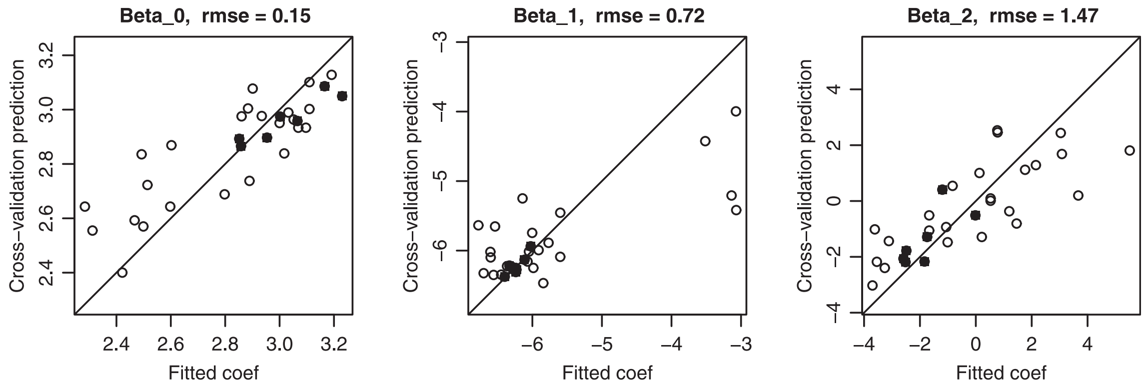

Fig. 6. Cross-validated predictions of the β_{0s}, β_{1s} and β_{2s} spatial fields of coefficients for the long-term average and two SEOF temporal trends in the Los Angeles region. The open dots represent the AQS locations while the shaded dots are the MESA Air fixed sites. The predictions are based on PLS regression models with 2, 1, and 1 components, respectively.

spatial covariance matrix $\sum_\varepsilon(\phi_\varepsilon)$. Estimates are computed for each 2-week period indexed by t using all monitoring data available at that time point from the AQS, MESA Air fixed, and MESA Air home monitors.

5. Application

Table 1 presents descriptive statistics for $PM_{2.5}$ concentrations and key spatial covariates including proximities to highways, distance to commercial properties, and median population densities. $PM_{2.5}$ concentrations are clearly highest, on average, around Los Angeles and lowest in St. Paul. Home sites in New York are closest to A1 highways while very few homes in Los Angeles and Winston–Salem are near highways. Homes in Chicago are closest to commercial properties. Population density is (obviously) highest in New York City and lowest in Winston–Salem.

The detailed results that follow, illustrated in Figs. 4–8, concern only the Los Angeles modeling region. Fig. 4 presents the results of the calculation of the SEOFs by the algorithm of section 4.1. The unsmoothed EOFs include a lot of short temporal scale variation and there is a considerable loss of explanatory power as a result of the smoothing. This short temporal scale variation is accounted for separately in the spatio-temporal residuals, allowing us to focus the modeling of the effects of spatial covariates (below) on smooth temporal trend patterns. The unsmoothed first singular vector explains over 70% of the variation in the 24 log-transformed AQS time series while the SEOF explains only 27.5%. This first component reflects a long-term decrease in $PM_{2.5}$ levels across most of the sites over these six years and seasonal structure that is quite variable from year-to-year. The second SEOF, explaining only 9.4% of the variation in the original time series, turns out to describe a relatively simple seasonal structure. The nature of these SEOFs varies across the four modeling regions. In North Carolina, for example, the dominant SEOF is a simple seasonal pattern like that of the second SEOF here (data not shown). In all cases, fitted trends based on these SEOFs, such as those illustrated in Fig. 3, leave residuals with very little temporal autocorrelation (results not shown here).

The cross-validatory assessment of the number of PLS components for our land use regressions based on universal kriging predictions suggests at most two predictive components for all of the trend coefficients in all regions. This result is strikingly different from the result of conventional PLS regression cross-validation without a spatial correlation model. As many as 10 PLS components are suggested for the intercept parameter β_0 by conventional

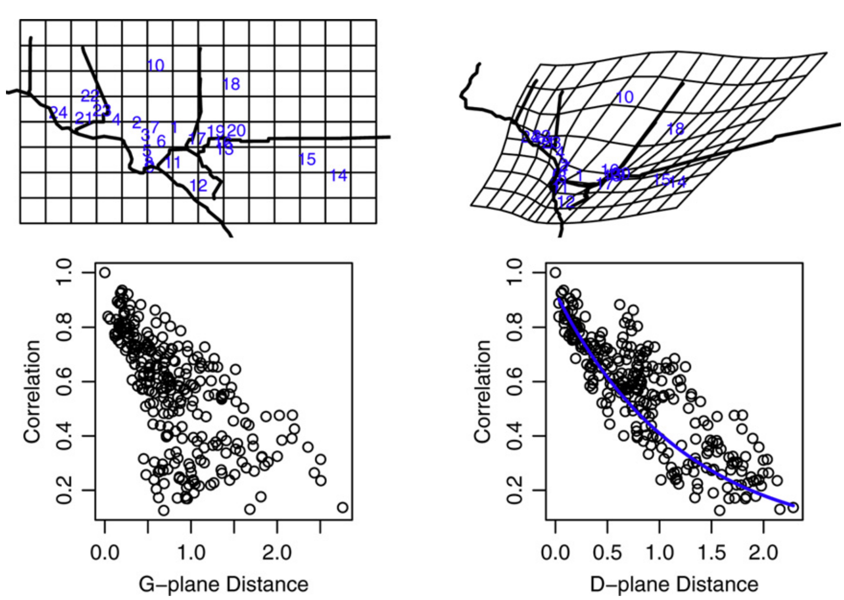

Fig. 7. Spatial structure of the spatio-temporal residuals before (left) and after (right) transformation using the Sampson–Guttorp method to account for nonstationarity.

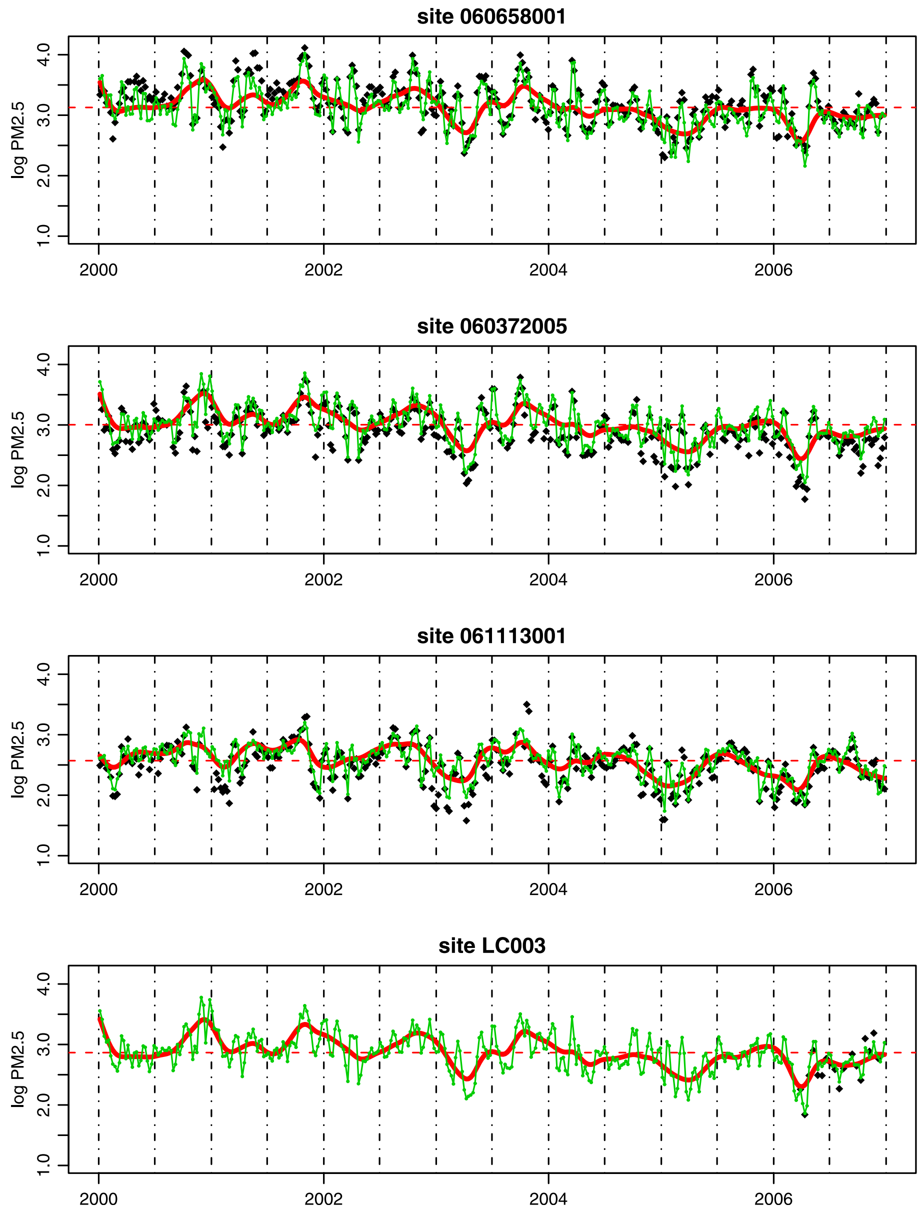

Fig. 8. Example cross-validated predictions of log-transformed two-week concentrations at AQS and MESA Air fixed sites in the Los Angeles region. The black dots are the measured data, the solid curves show predicted trends based on the SEOF part of the spatio-temporal model, and the jagged lines show two-week average predictions that incorporate the spatio-temporal residuals. The predicted trend curves drawn here can be compared with the fitted trend curves depicted in Fig. 3.

cross-validation, but substantially better cross-validatory predictions are found using only two components together with a spatial correlation model.

Loadings (correlations) of the spatial covariates on the two PLS components for the spatial regression model for the intercept (long-term mean), β_{0s}, are illustrated in Fig. 5. We see that the first component is highly negatively correlated with distance to the nearest A1 highway and positively correlated with the distances to the nearest A2 and A3 highways. Lengths of A1 vs A2 and A3 roads in buffers show the expected opposite correlation pattern. That is, this first component, which is predictive of long-term mean concentrations, defines a score that contrasts sites relatively near A1 highways but not necessarily near A2 and/or A3 roads with sites that are near A2 and/or A3 roads but far from A1 highways. The former sites have higher $PM_{2.5}$ concentrations on average while the latter sites have lower concentrations. The sites with positive scores (near A1 highways) are largely residential as indicated by the positive loadings on the residential land use covariates and contrasting negative correlations with the cropland covariates. The second PLS component defines a commercial vs. residential property contrast.

Scatterplots of cross-validated universal kriging predictions using two component PLS regressions for each of the three trend coefficients are presented in Fig. 6. The shaded dots are the MESA Air fixed sites. We see reasonably good regressions for the long-term mean $\widehat{\beta}_{0s}$ and the 3rd coefficient $\widehat{\beta}_{2s}$ for the amplitude of the simple cyclic seasonal structure of Fig. 4, but little ability to predict variation in the 2nd trend coefficient $\widehat{\beta}_{1s}$ for the component carrying the long-term decrease in concentrations.

The strength of the spatial correlation structure in the deviations from the fitted trends is illustrated in Fig. 7. The lower left panel shows inter-site correlations vs inter-site distance for the geographic configuration of sites given in the upper left panel. The horizontal axis has units of 100s of kms. After a Sampson–Guttorp deformation of the geographic coordinate system, illustrated in the upper right panel, we obtain the lower right scatterplot, which shows a much clearer spatial correlation structure. Note that correlations do not die out to zero even over the greatest spatial separation.

Fig. 8 presents cross-validated predictions of the 2-week observations for the four sites seen in Fig. 3. The dashed horizontal lines are the long-term means of the cross-validation predictions while the smooth solid lines are the predicted trends. These may be compared with the fitted trends drawn in Fig. 3. The jagged lines connect the predicted 2-week values. The predictions generally track the observations quite well with modest levels of over- or under-estimation of the long-term means. Site 060372005 displays the most noticeable over-estimation. We note that the magnitude of the error in these cross-validated predictions is probably greater than the error of prediction expected at most MESA subject homes as, for example, prediction of concentrations at locations near site 060372005 will benefit from the monitoring data at that site, data which were excluded in its own prediction.

Model fitting and model predictions as illustrated above for the MESA Air Southern California study area were carried out similarly for the other three geographic modeling domains, the Midwest region spanning Minneapolis-St Paul and Chicago, the Northeast region spanning Baltimore and New York City, and North Carolina. These regions are depicted in Figs. 9–11. Prediction of long-term averages at the locations of the MESA Air subjects in all six study areas are summarized in the boxplots of Fig. 12.

Results demonstrate the substantially higher and spatially variable $PM_{2.5}$ concentration levels for the MESA Air subject locations in Southern California in contrast to all the other MESA Air study areas. St Paul, MN, Baltimore, MD, and New York City all demonstrate similar ranges of concentrations while there is relatively little variation in concentration levels for the MESA Air subjects in and around Winston-Salem, NC.

Conventional "plug-in" estimates of the standard errors of predictions of long-term averages could be computed for the final step of our prediction strategy (section 4.5). However, we refrain from computing these estimates as the current model has some recognized deficiencies and because such standard errors do not account for prediction uncertainty that derives from the multi-step pragmatic model-fitting procedure we have employed. We report instead the accuracy of these pragmatic ambient concentration predictions in terms of descriptive statistics on the cross-validated (leave-one-out) errors of prediction of long-term mean concentrations for 6 MESA Air study areas.

Table 2 reports first (in columns two and three) means and standard deviations for the fitted long-term mean concentrations across all the AQS and MESA Air monitoring sites in each of the study regions. We then report in column four the root-mean-square error of the cross-validated predictions at these sites. The maps in Figs. 1 and 9–11 show that these descriptive statistics pertain to monitoring sites over geographic areas substantially exceeding the spatial domain of the MESA Air subjects. We therefore computed these same summary statistics also on just the MESA Air fixed sites, which were located within the domains of the MESA Air subjects. These sites provide more relevant characterizations of the accuracy of our model predictions, albeit for a relatively small number of sites in each region (N = 3 to 7).

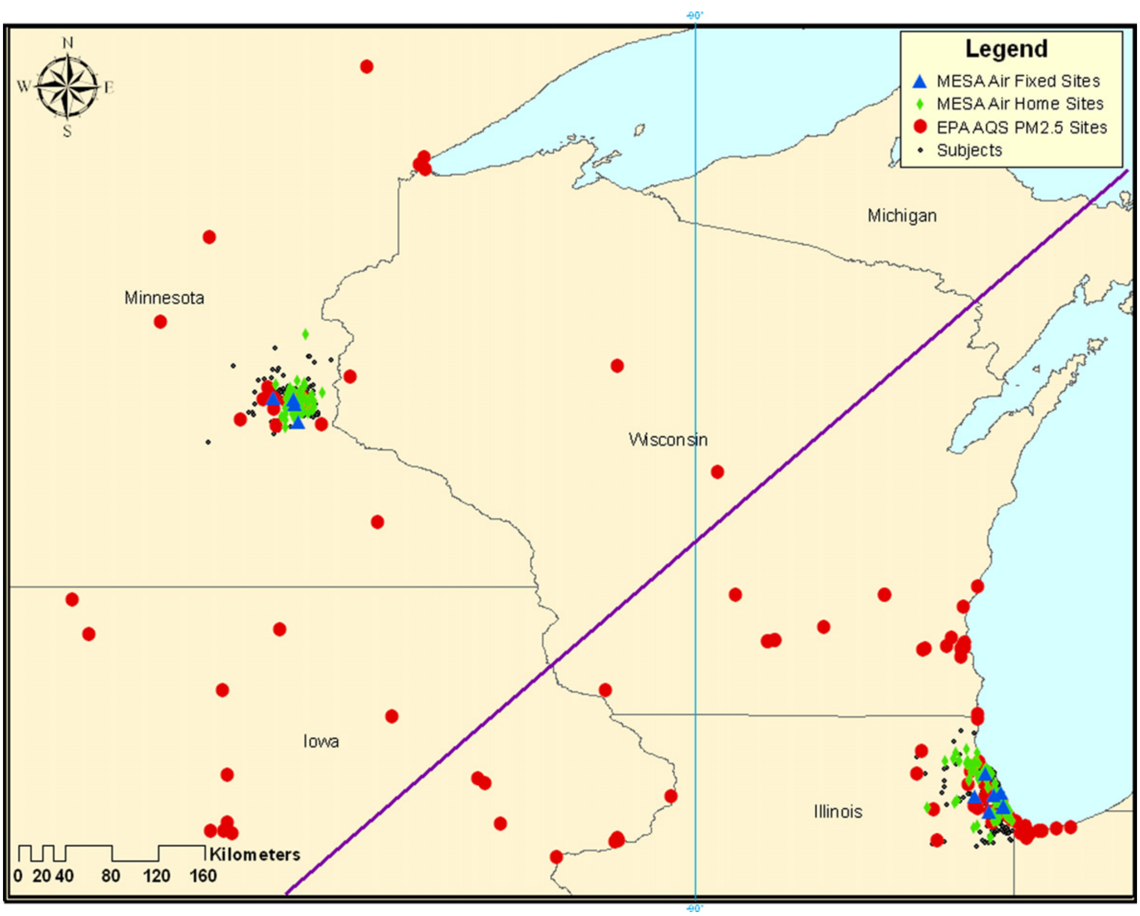

Fig. 9. Monitoring sites and subject home locations in the midwest region spanning the Minneapolis-St Paul and Chicago MESA Air study areas.

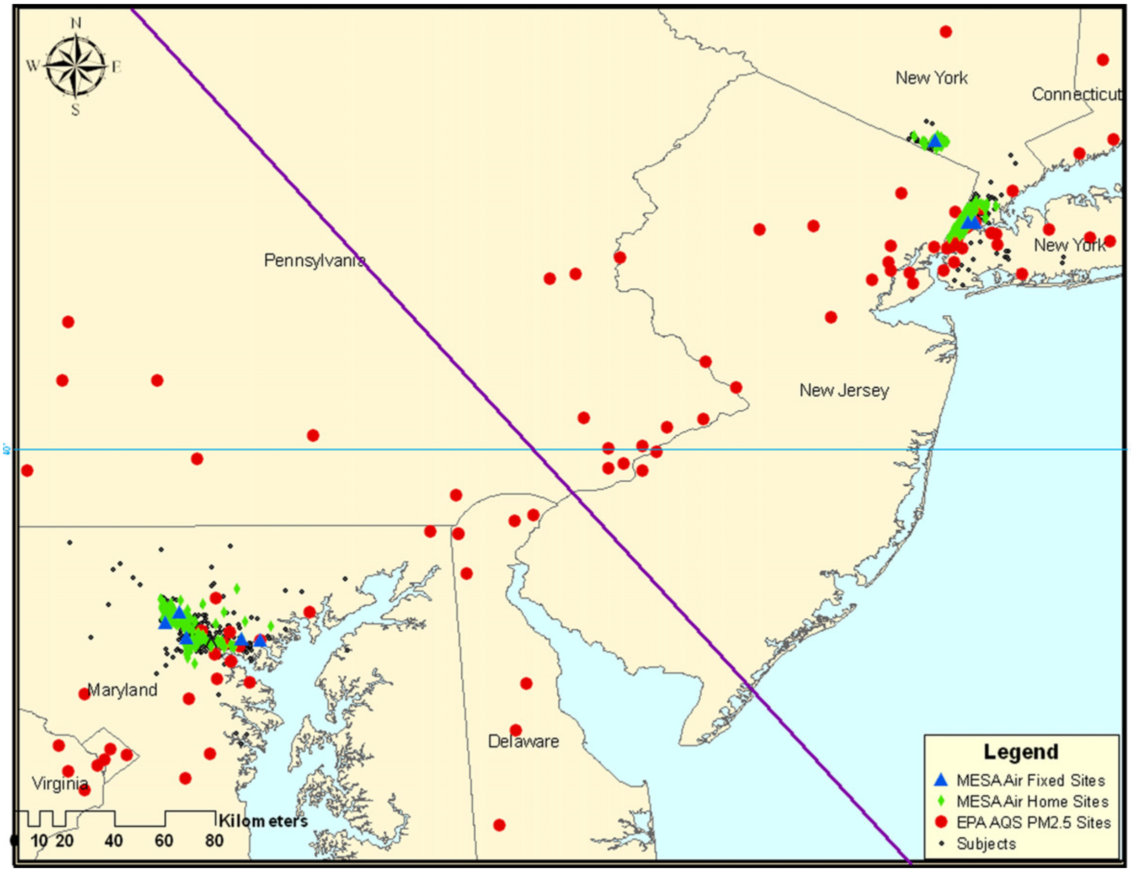

Fig. 10. Monitoring sites and subject home locations in the northeast region spanning the New York and Baltimore MESA Air study areas.

We see that the error of long-term estimates is similar across regions, generally near 10% of the mean for all sites and for the subsets of MESA Air Fixed sites. Comparison of standard deviations of long-term means to cross-validated root-mean-squared errors indicates that predictions are capturing meaningful intra-urban spatial variability in some areas, including California, but these root-mean-square errors are nearly as large as the standard deviations in some of the other study areas, notably North Carolina. In future work we expect improved city-specific spatio-temporal predictions and lower uncertainty using new GIS computations and additional land use and traffic covariates. Furthermore, uncertainty will be assessed by model-based standard errors using the likelihood modeling computations of Szpiro et al. (2010) and Lindström et al. (2011) as well as cross-validation.

6. Discussion

The current pragmatic $PM_{2.5}$ modeling and long-term ambient concentration predictions described here are based on the first phase of data collection by the MESA Air study. Revised analyses will be based on monitoring data complete through 2009 and improved measures of our GIS-based covariates including measures of traffic volumes not available for the current analyses.

There are a number of important contributions of the data analyses and results presented here. We have obtained insight into model selection, including the importance of accounting for the spatial correlation model in using cross-validation to select the number of PLS components for the mean model. We benefitted by using the MESA Air supplemental monitoring data in addition to AQS data to determine estimates of model parameters $\widehat{\beta}_{0s}, \widehat{\beta}_{1s}, \widehat{\beta}_{2s}$ at both AQS and MESA Air fixed sites for this regression modeling. Estimates of long-term average ambient $PM_{2.5}$ exposure described here are being used in preliminary health effect analyses with the MESA Air cohort (Adar et al., 2010; Krishnan et al., 2009).

We will ultimately use the likelihood method reported in Szpiro et al. (2010) and Lindström et al. (2011) because it provides a unified framework and gives standard errors (as opposed to cross-validation in this paper). However, much of the model selection work will still be done outside of the likelihood framework using the methods presented in this paper. This includes the specification of the SEOFs and the PLS or selection of covariates for a similar regression model with a universal kriging cross-validation approach. The work here, carried out in parallel to the development of the likelihood method, provided the most pragmatic approach to obtaining initial ambient concentration estimates for use in our epidemiology studies.

The current modeling and analysis leaves some problematic issues to be addressed in future work. The most important is dealing appropriately with the fact that 2-week average concentrations derive from different monitoring networks (AQS and MESA Air) with different monitoring instruments and temporal sampling protocols. This will require a nested specification of spatio-temporal correlation at a daily time scale. This daily time scale structure is also fundamental to a "downscaling" extension of the current model predictions in order to obtain estimates at a daily time scale for analysis of epidemiologic outcomes such as cardiovascular events that are expected to be sensitive to acute as well as chronic exposure.

The current model includes only temporal factors (the temporal basis functions) and spatially varying covariates. Recent extensions of the likelihood model fitting can incorporate spatio-temporal

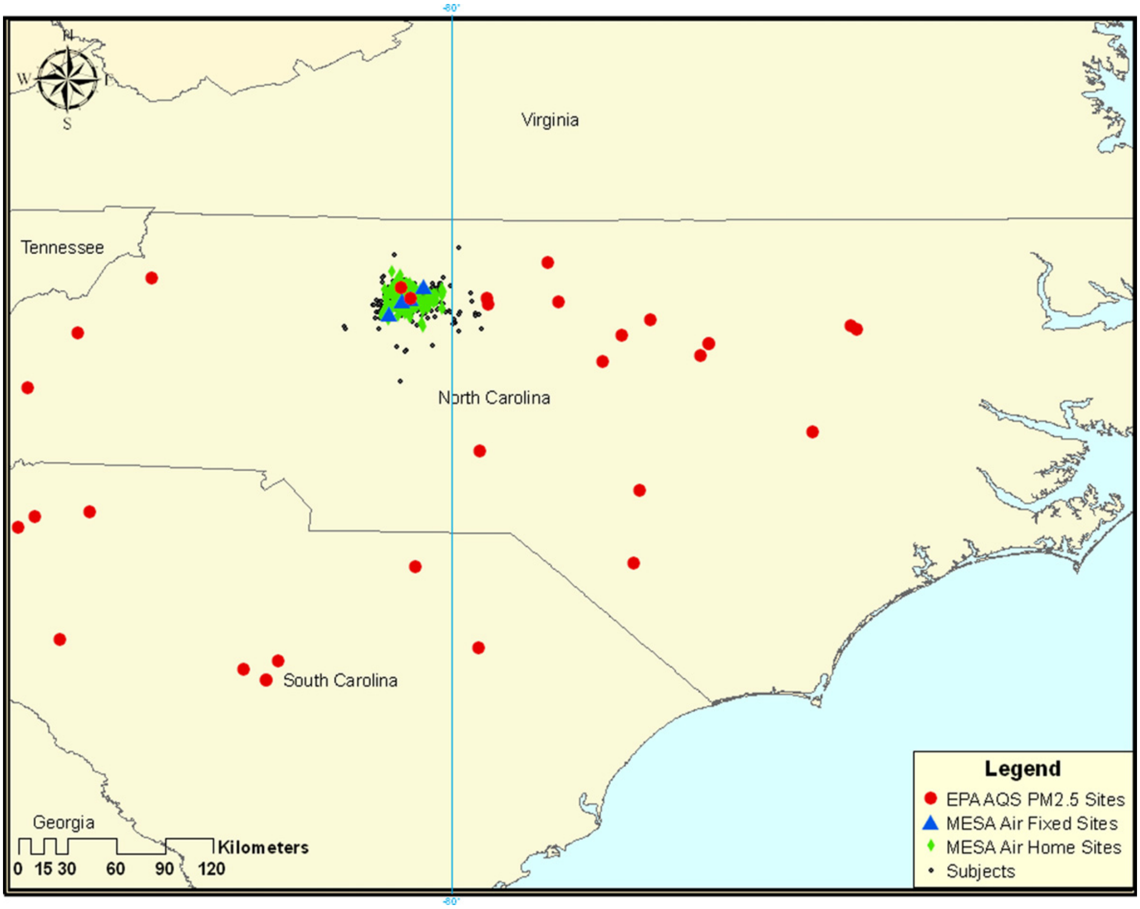

Fig. 11. Monitoring sites and subject home locations in the North Carolina region.

covariates, the most important of which are spatio-temporally varying characterizations of the effects of traffic on ambient exposure. Lindström et al. (2011) models NO_X (rather than $PM_{2.5}$) using as a covariate theoretical predictions of pollutant concentrations provided by a physics-based plume dispersion model, EPA's CALINE model (Wilton et al., 2010).

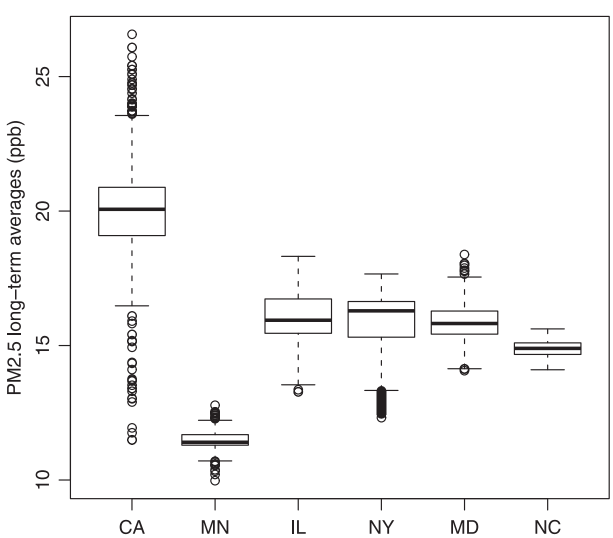

Fig. 12. Predicted long-term average concentrations of $PM_{2.5}$ (ppb) at all subject home locations in each of the six MESA Air study areas.

The ultimate objective of the modeling described here is to provide predicted exposures for estimating health effects in epidemiology studies. Up to now, we have used a "plug-in" approach that does not account for the additional variability resulting from uncertainty in the spatio-temporal prediction procedure. We have recently developed an efficient bootstrap-based approach to incorporating this uncertainty in health effect estimation (Szpiro et al., in press). In future work, we will apply this methodology based on fitting of the current spatio-temporal model using likelihood methods in order to obtain corrected standard errors for the disease model parameters of interest.

Table 2
Descriptive statistics on the cross-validated (leave-one-out) errors of prediction of long-term mean concentrations at monitoring sites in the six cities of the four major modeling regions.

Data Scale	Site	All Sites				MESA Air Fixed Sites			
		N	Mean	SD	RMSE	N	Mean	SD	RMSE
Log	CA	31	2.85	0.27	0.15	7	3.02	0.15	0.09
	IL	51	2.67	0.13	0.09	7	2.67	0.12	0.12
	MN	44	2.36	0.15	0.09	3	2.36	0.07	0.09
	MD	44	2.70	0.08	0.07	5	2.76	0.06	0.08
	NY	48	2.61	0.11	0.09	3	2.68	0.18	0.12
	NC	33	2.66	0.06	0.07	4	2.70	0.02	0.07
Original	CA	31	17.84	4.47	2.42	7	20.64	3.04	1.98
	IL	51	14.52	1.83	1.32	7	14.59	1.93	1.86
	MN	44	10.68	1.49	0.88	3	10.61	0.72	1.05
	MD	44	14.90	1.23	1.05	5	15.77	0.90	1.24
	NY	48	13.66	1.58	1.21	3	14.70	2.52	1.75
	NC	33	14.38	0.90	0.95	4	14.84	0.33	1.02

Acknowledgements

Although the research described in this presentation has been funded in part by the United States Environmental Protection Agency through grant R831697 and assistance agreement CR-83407101 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. Additional support was provided by an award to the University of Washington under the National Particle Component Toxicity (NPACT) initiative of the Health Effects Institute (HEI).

References

Abdi, H., 2010. Partial least squares regression and projection on latent structure regression (PLS regression). WIREs Computational Statistics 2, 97–106.

Adar, S.D., Klein, R., Klein, B.E.K., Szpiro, A.A., Cotch, M.F., Wong, T.Y., O'Neill, M.S., Shrager, S., Graham Barr, R., Siscovick, D., Daviglus, M.L., Sampson, P.D., Kaufman, J., 2010. Air pollution and the microvasculature: A cross-sectional assessment of in vivo retinal images in the population-based Multi-Ethnic Study of Atherosclerosis (MESA). PLoS Medicine 7 (11), e1000372.

Banerjee, S., Gelfand, A.E., Carlin, P., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall, 472 pp.

Banerjee, S., Johanson, G.A., 2006. Coregionalized single- and multiresolution spatially varying growth curve modeling with application to weed growth. Biometrics 62, 864–876.

Beckers, J., Rixen, M., 2003. EOF calculations and data filling from incomplete oceanographic datasets. Journal of Atmospheric and Oceanic Technology 20, 1839–1856.

Cohen, M.A., Adar, A.D., Allen, R.W., Avol, E., Curl, C.L., Gould, T., Hardie, D., Ho, A., Kinney, P., Larson, T.V., Sampson, P.D., Sheppard, L., Stukovsky, K.D., Swan, S.S., Liu, L.-J.S., Kaufman, J.D., 2009. Approach to estimating participant pollutant exposures in the multi-ethnic study of atherosclerosis and air pollution (MESA Air). Environmental Science & Technology 43, 4687–4693.

Damian, D., Sampson, P.D., Guttorp, P., 2001. Bayesian estimation of semi-parametric non-stationary spatial covariance structures. Environmetrics 12, 161–178.

Damian, D., Sampson, P.D., Guttorp, P., 2003. Variance modeling for nonstationary processes with temporal replications. Journal of Geophysical Research – Atmosphere 108 (D24).

EPA, 2010. The U.S. Environmental Protection Agency, Technology Transfer Network (TTN) Air Quality System (AQS). http://www.epa.gov/ttn/airs/airsaqs/ (accessed Sept 2010).

ESRI. ArcGIS (Version 9.1). Redlands, CA.

Fanshawe, T.R., Diggle, P.J., Rushton, S., Sanderson, R., Lurz, P.W.W., Glinianaia, S.V., Pearce, M.S., Parker, L., Charlong, M., Pless-Mulloli, T., 2008. Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. Environmetrics 19, 549–566.

Fuentes, M., Guttorp, P., Sampson, P.D., 2006. Using transforms to analyze space-time processes. In: Finkenstadt, B., Held, L., Isham, V. (Eds.), Statistical Methods for Spatio-Temporal Systems. Chapman and Hall/CRC, Boca Raton, FL, USA, pp. 77–150.

Garthwaite, P., 1994. An interpretation of partial least squares. Journal of the American Statistical Association 89, 122–127.

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

Krishnan, R.M., Adar, S.D., O'Neil, M.S., Barr, G.R., Polak, J.F., Herrington, D., Jorgensen, N.S., Van Hee, V.C., Kaufman, J.D., 2009. Vascular responses to short and long-term ambient air pollution: The multi-ethnic study of atherosclerosis and air pollution. Circulation 120 (Suppl 2), S462.

Le, N.D., Zidek, J.V., 2006. Statistical Analysis of Environmental Space-Time Processes. Springer, New York, NY, USA.

Lindström, J., Szpiro, A.A., Sampson, P.D., Sheppard, L., Oron, A., Richards, M., Larson, T., 2011. A flexible spatio-temporal model for air pollution: allowing for spatio-temporal covariates (January 19, 2011). UW Biostatistics Working Paper Series. Working Paper 370. http://www.bepress.com/uwbiostat/paper370.

Moore, D.K., Jerrett, M., Mack, W.J., Kunzli, N., 2007. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. Journal of Environmental Monitoring 9, 246–252.

Paciorek, C.J., Yanosky, J.D., Puett, R.C., Laden, F., Suh, H.H., 2009. Practical large-scale spatio-temporal modeling of particulate matter concentrations. Annals of Applied Statistics 3, 370–397.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Ross, Z., Jerrett, M., Ito, K., Tempalski, B., Thurston, G.D., 2007. A land use regression for predicting fine particulate matter concentrations in the New York City region. Atmospheric Environment 41, 2255–2269.

Sahu, S.K., Gelfand, A.E., Holland, D.M., 2006. Spatio-temporal modeling of fine particulate matter. Journal of Agricultural, Biological, and Environmental Statistics 11, 61–86.

Smith, R.L., Kolenikov, S., Cox, L.H., 2003. Spatio-temporal modeling of $PM_{2.5}$ data with missing values. Journal of Geophysical Research 108 (D24), 9004.

Szpiro, A.A., Sampson, P.D., Sheppard, L., Lumley, T., Adar, S.D., Kaufman, J.D., 2010. Predicting intraurban variation in air pollution concentrations with complex spatio-temporal interactions. Environmetrics 21, 606–631.

Szpiro, A.A., Sheppard, L., Lumley, T. Efficient measurement error correction with spatially misaligned data. Biostatistics, in press (published online January 20, 2011).

Wackernagel, H., 2010. Multivariate Geostatistics: An Introduction with Applications, third ed. Springer-Verlag, Berlin, Germany.

Wilton, D., Szpiro, A., Gould, T., Larson, T., 2010. Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles, CA and Seattle, WA. Science of the Total Environment 408, 1120–1130.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.