12-5-2012

A National Model Built with Partial Least Squares and Universal Kriging and Bootstrap-based Measurement Error Correction Techniques: An Application to the Multi-Ethnic Study of Atherosclerosis

Silas Bergen
University of Washington - Seattle Campus, srbergen@uw.edu

Lianne Sheppard
University of Washington - Seattle Campus, sheppard@u.washington.edu

Paul D. Sampson
University of Washington - Seattle Campus, pds@u.washington.edu

Sun-Young Kim
University of Washington - Seattle Campus, puha0@u.washington.edu

Mark Richards
University of Washington - Seattle Campus, markr9@u.washington.edu

See next page for additional authors

Authors
Silas Bergen, Lianne Sheppard, Paul D. Sampson, Sun-Young Kim, Mark Richards, Sverre Vedal, Joel Kaufman, and Adam A. Szpiro

A national model built with partial least squares and universal kriging and bootstrap-based measurement error correction techniques: an application to the Multi-Ethnic Study of Atherosclerosis (MESA)

Silas Bergen [*] Lianne Sheppard [†] Paul D. Sampson [‡] Sun Young Kim [§]
Mark Richards [¶] Sverre Vedal [‖] Joel D. Kaufman [**] Adam A. Szpiro [††]

Last revised: 4th December 2012

Abstract

Studies estimating health effects of long-term air pollution exposure often use a two-stage approach, building exposure models to assign individual-level exposures which are then used in regression analyses. This requires accurate exposure modeling and careful treatment of exposure measurement error. To illustrate the importance of carefully accounting for exposure model characteristics in two-stage air pollution studies, we consider a case study based on data from the Multi-Ethnic Study of Atherosclerosis (MESA). We present national spatial exposure models that use partial least squares and universal kriging to estimate annual average concentrations of four $PM_{2.5}$ components: elemental carbon (EC), organic carbon (OC), sulfur (S), and silicon (Si). Our models perform well, with cross-validated R^2s ranging from 0.62 to 0.95. We predict $PM_{2.5}$ component exposures for the MESA cohort and estimate cross-sectional associations with carotid intima-media thickness (CIMT), adjusting for subject-specific covariates. In naïve analyses that do not account for measurement error, we find statistically significant associations between CIMT and increased exposure to OC, S, and Si. We correct for measurement error using recently developed methods that account for the spatial structure of predicted exposures. OC exhibits little spatial correlation, and the corrected inference is unchanged from the naïve analysis. The S and Si exposure surfaces display notable spatial correlation, resulting in corrected confidence intervals (CIs) that are 50% wider than the naïve CIs, but that are still statistically significant. The impact on health effect inference is concordant with the degree of spatial correlation in the exposure surfaces.

[*]Department of Biostatistics, University of Washington
[†]Departments of Environmental and Occupational Health Sciences and Biostatistics, University of Washington
[‡]Department of Statistics, University of Washington
[§]Department of Environmental and Occupational Health Sciences, University of Washington
[¶]Department of Environmental and Occupational Health Sciences, University of Washington
[‖]Department of Environmental and Occupational Health Sciences, University of Washington
[**]Department of Environmental and Occupational Health Sciences, University of Washington
[††]To whom correspondence should be directed: Department of Biostatistics, University of Washington, (email) aszpiro@u.washington.edu, (p) 206-616-6846

1 Introduction

The relationship between air pollution and adverse health outcomes has been well-documented (Samet et al., 2000, Pope et al., 2002). Many studies focus on particulate matter, specifically particulate matter less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$) (Miller et al., 2007, Kim et al., 2009). Health effects of $PM_{2.5}$ could depend on characteristics of the particles, including shape, solubility, pH, or chemical composition (Vedal et al., 2012), and a deeper understanding of these differential effects could help inform policy. One of the challenges in assessing the impact of different chemical components of $PM_{2.5}$ in an epidemiology study is the need to assign exposures to study participants based on monitoring data at different locations (i.e., spatially misaligned data). When doing this for many components, the assignment or prediction procedure needs to be streamlined in order to be practical. Whatever the prediction algorithm, using the estimated rather than true exposures induces measurement error in the subsequent epidemiologic analysis. This paper describes a flexible and efficient prediction model that can be applied on a national scale to assign estimates of long-term exposure levels for components of $PM_{2.5}$. It then illustrates application of the predictions in a health analysis of the Multi-Ethnic Study of Atherosclerosis (MESA) cohort and incorporates appropriate techniques to correct for the impact of measurement error. We find that the importance of accounting for measurement error varies between chemical components, depending on the level of spatial structure inherent in the respective pollutant surfaces.

Current methods for assigning exposures include land-use regression (LUR) with Geographic Information System (GIS) covariates (Brauer et al., 2003, Hoek et al., 2008) and universal kriging (UK) (Jerrett et al., 2005, Künzli et al., 2005) that also exploits residual spatial structure (Kim et al., 2009, Mercer et al., 2011). There are often many candidate GIS covariates, including some that are correlated with each other, necessitating a dimension reduction procedure. Variable selection methods that have been considered in the literature include exhaustive search, stepwise selection, and shrinkage by the "lasso"(Tibshirani, 1996, Mercer et al., 2011). However, variable selection methods tend to be computationally intensive and require significant analyst time, feasible perhaps when considering a single pollutant but quickly becoming impractical when attempting to develop predictions for multiple pollutants. A more streamlined alternative is partial least squares (PLS) (Sampson et al., 2009, Abdi, 2003). This method finds a small number of linear combinations of the GIS covariates that most efficiently account for variability in the measured concentrations. These linear combinations, known

as "scores", effectively reduce the covariate space to a much smaller dimension. The resulting scores can then be used as the mean structure in a LUR or UK model in place of individual GIS covariates. This provides the advantages of using all available GIS covariates and eliminating potentially time-consuming variable selection processes. We employ a combination of PLS and universal kriging to develop a flexible and efficient national prediction model to estimate long-term average concentrations of four chemical species of $PM_{2.5}$: elemental carbon (EC), organic carbon (OC), silicon (Si) and sulfur (S).

Using exposures predicted from spatially misaligned data rather than true exposures in health models introduces measurement error that may have implications for $\hat{\beta}_X$, the estimated health model coefficient of interest (Madsen et al., 2008, Gryparis et al., 2009, Szpiro et al., 2011b). Zeger et al. (2000) point out that this exposure measurement error may have substantial implications for interpreting epidemiologic air pollution studies and emphasize that great care must be taken with interpretation of such epidemiologic studies when measurement error is present. Molitor et al. (2007) discuss the impact on the health model coefficient of interest and confidence interval width when spatial autocorrelation is exploited, using hierarchical Bayesian models. Those models do not explicitly estimate subject-level exposure but rather treat it as a latent variable. Our treatment of measurement error focuses instead on a 2-step approach in which exposures are explicitly calculated and used in health modeling. In this context, Berkson-like error that arises from smoothing the true exposure surface may inflate the standard error of $\hat{\beta}_X$. Classical-like error results from estimating the prediction model parameters, and may bias $\hat{\beta}_X$ in addition to inflating its standard error. Bootstrap methods to adjust for the effects of measurement error have been discussed in Szpiro et al. (2011b).

We derive predictions of component concentrations using our national exposure models, and use them as the covariates of interest in health analyses assessing associations between carotid intima-media thickness (CIMT), a subclinical measure of atherosclerosis, and air pollution exposure. These results have also been described elsewhere (Vedal et al., 2012). We then apply measurement error correction methods to account for the fact that predicted rather than true exposures are being used in these health models.

This article is organized as follows: Section 2.1 describes the MESA cohort used for the health modeling, and the monitoring data used to develop the exposure models. Section 2.2 gives notation for and describes the cross-validation procedure used to assess the exposure models. Section 2.3 describes

the health models and measurement error correction techniques. Section 3 gives results of the exposure modeling and cross-validation and the results of health analyses with and without measurement error correction. Section 4 discusses characteristics of the individual pollutant exposure models and how features of the pollutant surfaces are reflected in the measurement error correction results.

2 Methods

2.1 Data

2.1.1 MESA Cohort

The Multi-Ethnic Study of Atherosclerosis (MESA) study is a population-based study that began in 2000, with a cohort consisting of 6,814 participants from six U.S. cities: Los Angeles, CA; St. Paul, MN; Chicago, IL; Winston-Salem, NC; New York, NY; and Baltimore, MD. Four ethnic/racial groups were targeted: white, African American, Hispanic, and Chinese American. All participants were free of clinical cardiovascular disease at time of entrance.

We illustrate our approach using the common carotid intima-media thickness (CIMT) endpoint in MESA. CIMT, a subclinical measure of atherosclerosis, was measured by B-mode ultrasound using a GE Logiq scanner, and the endpoint was quantified as the right far wall CIMT measures conducted during MESA exam 1, which took place for 2000-2002 (Vedal et al., 2012). We considered the 5,501 MESA participants who had CIMT measures during exam 1; our analysis was based on the 5,298 MESA participants who had IMT measures during exam 1 and complete values of confounding variables.

2.1.2 Monitoring data

Data on EC, OC, Si and S were collected to build the national model. These data consisted of annual averages from 2009-2010 as measured by the EPA's Interagency Monitoring for Protected Visual Environments (IMPROVE) (Eldred et al., 1988) and Chemical Speciation Network (CSN) (EPA 2009). The IMPROVE monitors are a nation-wide network located mostly in national parks and other remote areas. The CSN monitors are in more urban areas. These two networks provide data that are evenly dispersed throughout the lower 48 states (Figure 1).

All CSN and IMPROVE monitors that had at least 10 data points per quarter and a maximum of 45 days between measurements were included in our analyses. For Si and S, averages were over

01/01/2009-12/31-2009. The EC/OC data set consisted of 204 IMPROVE and CSN monitors averaged over 01/01/2009-12/31-2009, and 51 CSN monitors averaged over 05/01/2009-04/30/2010. The latter period was used since prior to 05/01/2009 these monitors used a protocol that was incompatible with the IMPROVE network. Comparing averages over 05/01/2009-04/30/2010 to those which used comparable protocol over 01/01/2009-12/31-2009 indicated little difference between the time periods. The annual averages were square-root transformed prior to modeling.

2.1.3 Geographic covariates

For all monitor and subject locations, approximately 600 LUR covariates were available. These included distances to A1, A2, and A3 roads (Census Feature Class Codes (CFCC)); land use within a given buffer; population density within a given buffer; and normalized difference vegetation index (NDVI) which measures the level of vegetation in a monitor's vicinity. CFCC A1 roads are limited access highways; A2 and A3 roads are other major roads such as county and state highways without limited access (Mercer et al., 2011). For NDVI a series of 23 monitor-specific, 16-day composite satellite images were obtained, and the pixels within a given buffer averaged for each image. PLS incorporated the 25th, 50th and 75th percentile of these 23 averages. The median of "high-vegetation season" image averages (defined as April 1-September 30) and "low-vegetation season" averages (October 1-March 31) were also included. For more detailed information about the land use variables see Anderson (1976).

Before building our exposure models, we conducted variable pre-processing to eliminate LUR covariates that were too homogeneous or outlier-prone to be of use. We eliminated variables with $> 85\%$ identical values, and those with the most extreme standardized outlier > 7. We log-transformed and truncated all distance variables at 10 km and computed additional "compiled" distance variables such as minimum distance to major roads, distance to any port, etc. These compiled variables were then subject to the same inclusion criteria. All selected covariates were mean-centered and scaled by their respective standard deviations.

2.2 Spatial prediction models

2.2.1 Notation

To describe our exposure models, we introduce some notation. Let $\mathbf{X}^*$ denote the $N^* \times 1$ vector of observed square-root transformed concentrations at monitor locations; $\mathbf{R}^*$ the $N^* \times p$ matrix of

geographic covariates at monitor locations; $\mathbf{X}$ the $N \times 1$ vector of unknown square-root transformed concentrations at the unobserved subject locations; and $\mathbf{R}$ the $N \times p$ matrix of geographic covariates at subject locations. PLS was used to decompose $\mathbf{R}^*$ into a set of linear combinations of much smaller dimension than the space of the covariates $\mathbf{R}^*$. Specifically,

$$\mathbf{R}^*\mathbf{H} = \mathbf{T}^*.$$

Here, $\mathbf{H}$ is a $p \times k$ matrix of weights for the geographic covariates, and $\mathbf{T}^*$ is an $N^* \times k$ matrix of PLS components or scores. Let $\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_k\}$ and $\mathbf{T}^* = \{\boldsymbol{t}_1^*, ..., \boldsymbol{t}_k^*\}$. The weights $\mathbf{h}_i$ for $i = 1, ..., k$ are found in such a way that $\boldsymbol{t}_1^* = \mathbf{R}^*\mathbf{h}_1$ is the score of maximum covariance with $\mathbf{X}^*$, subject to $\mathbf{h}_1^T\mathbf{h}_1 = 1$; $\boldsymbol{t}_2^* = \mathbf{R}^*\mathbf{h}_2$ is orthogonal to $\boldsymbol{t}_1^*$ and is the score that explains as much remaining covariance as possible with $\mathbf{X}^*$, etc. In practice, $k \ll p$, leaving us with a set of orthogonal scores that are designed to explain the covariance of $\mathbf{R}^*$ and $\mathbf{X}^*$ as efficiently as possible. More details can be found in Abdi (2003). PLS scores at unobserved locations are then found by simply computing $\mathbf{T} = \mathbf{R}\mathbf{H}$.

Once the PLS components $\mathbf{T}$ and $\mathbf{T}^*$ were obtained for the unobserved and monitoring locations, the following joint model was assumed to motivate predictions,

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{X}^* \end{pmatrix} = \begin{pmatrix} \mathbf{T} \\ \mathbf{T}^* \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\eta}^* \end{pmatrix}. \tag{1}$$

It is now implicitly assumed that $\mathbf{T}$ and $\mathbf{T}^*$ each contain a $\mathbf{1}$ vector of appropriate length. Here, $\boldsymbol{\alpha}$ is a $(k+1) \times 1$ vector containing an intercept and the coefficients for the k PLS scores, and $\boldsymbol{\eta}$, $\boldsymbol{\eta}^*$ are $N \times 1$ and $N^* \times 1$ vectors of errors. For the prediction models which used only PLS, $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$ were assumed to be independent. Under this assumption, we estimated $\boldsymbol{\alpha}$ using a least-squares fit to regression of $\mathbf{X}^*$ on $\mathbf{T}^*$, let this estimate be denoted by $\hat{\boldsymbol{\alpha}}_{pls}$. PLS-only predictions at the unobserved locations were then obtained by computing $\mathbf{T}\hat{\boldsymbol{\alpha}}_{pls}$.

The second prediction modeling scenario did not assume independence of $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$ but instead assumed a joint distribution, specified as

$$\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\eta}^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11}(\boldsymbol{\theta}) & \Sigma_{12}(\boldsymbol{\theta}) \\ \Sigma_{21}(\boldsymbol{\theta}) & \Sigma_{22}(\boldsymbol{\theta}) \end{pmatrix} \right). \tag{2}$$

Here each corner of the covariance matrix is a kriging covariance matrix parameterized by a common

vector of parameters $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ (Cressie, 1992). τ^2 is the nugget, interpretable as the amount of variability in the pollution exposures that is unexplainable by spatial structure; σ^2 is the partial sill, interpretable as the amount of variability that is explainable by spatial structure; and ϕ is the range, interpretable as the maximum distance between two locations beyond which they may no longer be considered spatially correlated. Several variogram models are available to define the covariance between two locations. The exponential variogram was used for EC, OC and S, but provided a poor fit for Si. We therefore examined cubic and spherical variograms and found the spherical variogram provided a much better fit and used it to model Si in our exposure models. Given an assumed variogram, the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}_{uk}$ were estimated by profile maximum likelihood.

Once the estimates $\hat{\boldsymbol{\alpha}}_{uk}$ and $\hat{\boldsymbol{\theta}}$ were obtained, the universal kriging predictions were computed by the conditional expectation formula,

$$\mathbf{W}_{uk} = \mathbf{T}\hat{\boldsymbol{\alpha}}_{uk} + \Sigma_{12}(\hat{\boldsymbol{\theta}})\Sigma_{22}(\hat{\boldsymbol{\theta}})^{-1}(\mathbf{X}^* - \mathbf{T}^*\hat{\boldsymbol{\alpha}}_{uk}). \tag{3}$$

2.2.2 Cross-validation and Model Selection

10-fold cross-validation (Hastie et al., 2001) was used to assess the models' prediction accuracy, to compare predictions generated using PLS only to those generated using PLS combined with UK, and to select the number of PLS components to use in the final prediction models. Data were randomly assigned to one of ten groups. One group (a "test set") was omitted, and the remaining groups (a "training set") were used to fit the model and generate test set predictions. Each group played the role of test set until predictions were obtained for the entire data set. At each iteration, the following steps were taken:

1. PLS was fit using the training set, and K scores were computed for the test set, for $K = 1, ..., 10$. $\hat{\boldsymbol{\alpha}}_{pls}$ was calculated using the training set for each set of scores.

2. UK parameters $\boldsymbol{\theta}$ and coefficients $\boldsymbol{\alpha}_{uk}$ were estimated using the training set. The first K PLS scores played the role of $\mathbf{T}^*$ in Equation 1, for $K = 1, ..., 10$.

3. PLS-only predictions were generated using the first K components and corresponding $\hat{\boldsymbol{\alpha}}_{pls}$. An analogous set of predictions was derived using the first K PLS components and the corresponding UK, using $(\hat{\boldsymbol{\alpha}}_{uk}, \hat{\boldsymbol{\theta}})$ estimated from the training set.

The `R` package `pls` (Wehrens et al., 2006, R Development Core Team, 2010) was used to fit the PLS. UK was done using the `R` package `geoR` (Ribeiro Jr and Diggle, 2001). The best-performing models were selected based on their cross-validated root mean squared prediction error (RMSEP) and corresponding R^2. For a data set with N^* observations and corresponding predictions, the formulae for these performance metrics are given by

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N^*}(Obs_i - Pred_i)^2}{N^*}},$$
$$R^2 = \max\left(0, 1 - \frac{\text{RMSEP}^2}{\text{Var}(Obs)}\right).$$

These metrics are sensitive to scale; accordingly they are useful for evaluating model performance for a given pollutant, but not for comparing models across pollutants.

2.3 Health modeling

2.3.1 Disease model

Multivariable linear regression models were used to estimate the effects of $PM_{2.5}$ component exposure on CIMT. Each model included a single $PM_{2.5}$ component along with a vector of subject-specific covariates. Let $\mathbf{Y}$ be the 5501×1 vector of health outcomes, $\mathbf{W}^2$ the 5501×1 vector of predictions, and $\mathbf{Z}$ a matrix of potential confounders. Note that $\mathbf{W}^2$ and $\mathbf{X}^2$ respectively refer to the predicted and true exposures on the native scale, as all the exposure modeling was done on the square root scale. We assumed the following linear model relating $\mathbf{Y}$ to $\mathbf{X}$

$$\mathbf{Y} = \beta_0 + \mathbf{X}^2\beta_X + \mathbf{Z}\boldsymbol{\beta}_Z + \boldsymbol{\epsilon}, \tag{4}$$

where $(\epsilon_1, ..., \epsilon_N)$ were assumed i.i.d. $N(0, \sigma_\epsilon^2)$ random variables, and derived $\hat{\beta}_X$ by ordinary least squares (OLS).

2.3.2 Measurement Error Correction

The model in Equation 4 was fit using the predicted exposures $\mathbf{W}^2$ instead of the true exposures $\mathbf{X}^2$ as the covariate of interest. Using predictions rather than true exposures in health modeling introduces

two sources of measurement error that potentially influence the behavior of $\hat{\beta}_X$. Berkson-like error arises from smoothing the true exposure surface and could inflate the standard error of $\hat{\beta}_X$. Classical-like error arises from estimating the exposure model parameters $\boldsymbol{\alpha}_{uk}$ and $\boldsymbol{\theta}$. The classical-like error potentially inflates the standard error of $\hat{\beta}_X$ and could also bias the estimate. The parametric and parameter bootstraps were used to correct for the effects of measurement error. See Szpiro et al. (2011b) for additional background and details.

We describe the parametric bootstrap in the context of predictions that use both PLS and UK; the approach would be very similar if PLS alone was used (though we did not implement that correction here). The parametric bootstrap is defined given a set of health outcomes $\mathbf{Y}$ and subject-specific covariates $\mathbf{Z}$ for participants, true square-root transformed exposures $\mathbf{X}^*$ at monitoring sites, and geographic covariates $\mathbf{R}$ and $\mathbf{R}^*$ at participant and monitoring locations, respectively;

1. Derive geographic covariate weights $\mathbf{H}$ from $\mathbf{R}^*$ and $\mathbf{X}^*$, and hence PLS scores $\mathbf{T}$ and $\mathbf{T}^*$ at unobserved and observed locations, respectively.

2. Estimate exposure model parameters $\boldsymbol{\alpha}_{uk}$ and $\log(\boldsymbol{\theta})$ by nonlinear optimization using $\mathbf{X}^*$ and $\mathbf{T}^*$, exploiting the likelihood defined by Equations 1 and 2.

3. Derive $\mathbf{W}_{uk}$ from Equation 3, use $\mathbf{W}_{uk}^2$ in Equation 4 place of $\mathbf{X}^2$ along with $\mathbf{Z}$ to estimate health model parameters $\hat{\beta}_0$, $\hat{\beta}_X$, $\hat{\boldsymbol{\beta}}_Z$ and $\hat{\sigma}_\epsilon^2$.

4. For $j = 1, ..., B$ bootstrap samples:

   (a) Simulate $\mathbf{X}_j^*$ (and $\mathbf{X}_j$) from Equation 1 and $\mathbf{Y}_j$ from Equation 4, using $\hat{\boldsymbol{\alpha}}_{uk}$, $\hat{\boldsymbol{\theta}}$, $\hat{\beta}_0$, $\hat{\beta}_X$, $\hat{\boldsymbol{\beta}}_Z$ and $\hat{\sigma}_\epsilon^2$ in place of unknown parameters

   (b) Estimate new exposure parameters $\hat{\boldsymbol{\alpha}}_{uk,j}$ and $\log(\hat{\boldsymbol{\theta}}_j)$ by nonlinear optimization using $\mathbf{X}_j^*$

   (c) Plug $\hat{\boldsymbol{\alpha}}_{uk,j}$, $\hat{\boldsymbol{\theta}}_j$ and $\mathbf{X}_j^*$ into Equation 3 to derive $\mathbf{W}_{uk,j}$

   (d) Calculate $\hat{\beta}_{X,j}$ using $\mathbf{W}_{uk,j}^2$ by OLS in Equation 4.

5. Calculate a bootstrap bias estimate, $\widehat{Bias}_p(\hat{\beta}_X)$, and the corresponding corrected effect estimate $\hat{\beta}_{X,p}^{corrected} = \hat{\beta}_X - \widehat{Bias}_p(\hat{\beta}_X)$. We give details on calculating $\widehat{Bias}_p(\hat{\beta}_X)$ below.

6. Estimate the bootstrap standard error as

$$\widehat{SE}(\hat{\beta}_X) = \sqrt{\frac{\sum_{j=1}^{B}\left(\hat{\beta}_{X,j} - \frac{1}{B}\sum_{j=1}^{B}(\hat{\beta}_{X,j})\right)^2}{B}}.$$

For the parametric bootstrap we set $B = 15,000$. Note that using Step 6 to estimate the standard error of $\hat{\beta}_{X,p}^{corrected}$ will give an underestimate, as it does not account for the additional variability introduced by $\widehat{Bias}_p(\hat{\beta}_X)$. To fully account for this variability we would need to perform a nested bootstrap within each original bootstrap sample; however as discussed in Szpiro et al. (2011b) (and as exemplified in our results) the estimated bias is so small that this underestimation is ignorable in practice.

An undesirable trait of the parametric bootstrap is the computational time required to implement it, as it requires B non-linear optimizations. The parameter bootstrap is similar to the parametric bootstrap, but is much more time-efficient. It involves estimating a sampling distribution for $\hat{\boldsymbol{\alpha}}_{uk}$ and $\log(\hat{\boldsymbol{\theta}})$, and sampling $\hat{\boldsymbol{\alpha}}_{uk,j}$ and $\log(\hat{\boldsymbol{\theta}}_j)$ from this distribution rather than estimating them via nonlinear optimization as in Step 4(b) of the parametric bootstrap (Szpiro et al., 2011b). For our implementation we estimated the sampling distribution of $(\hat{\boldsymbol{\alpha}}_{uk,j}, \log(\hat{\boldsymbol{\theta}}_j))$ with a multivariate Gaussian distribution centered at $(\hat{\boldsymbol{\alpha}}_{uk}, \log(\hat{\boldsymbol{\theta}}))$ with a covariance matrix specified by the estimated inverse Hessian of the likelihood.

Another attractive feature of the parameter bootstrap is the ability to control the amount of variability in the sampling distribution of $\hat{\boldsymbol{\alpha}}_{uk,j}$ and $\hat{\boldsymbol{\theta}}_j$, by simply multiplying the estimated covariance matrix by a factor $\lambda \geq 0$. This generalization of the parameter bootstrap allows for investigation of how variability in the sampling distribution of $(\hat{\boldsymbol{\alpha}}_{uk}, \log(\hat{\boldsymbol{\theta}}))$ affects the bias of $\hat{\beta}_X$, which can be useful in refining our bootstrap bias estimates by simulation extrapolation (SIMEX) (Stefanski and Cook, 1995). In the Appendix we describe our approach to SIMEX in greater detail and give the results of applying it to the MESA data.

The partial parametric bootstrap simulates bootstrap exposures from Equation 1 using the original estimates $(\hat{\boldsymbol{\alpha}}_{uk}, \log(\hat{\boldsymbol{\theta}}))$ throughout. This approach corrects for the Berkson-like error only, since the parameter estimates remain fixed. The partial parametric bootstrap is equivalent to the generalization of the parameter bootstrap with $\lambda = 0$.

Let $E_\lambda(\hat{\beta}_X^B)$ denote the empirical mean of the parameter bootstrapped $\hat{\beta}_X$ implemented with a given value λ, and $E_p(\hat{\beta}_X^B)$ the empirical mean of the parametric bootstrapped $\hat{\beta}_X^B$. The bias of $\hat{\beta}_X$

estimated by the parameter bootstrap with multiplier λ is then defined as

$$\widehat{Bias}_{\lambda}(\hat{\beta}_X) = \left(E_{\lambda}(\hat{\beta}_X^B) - E_0(\hat{\beta}_X^B)\right).$$

For our main analysis we implemented the parameter bootstrap for $\lambda = 1$, the value which makes the parameter bootstrap theoretically equivalent to the parametric bootstrap. The corresponding bias-corrected effect estimate is

$$\hat{\beta}_{X,1}^{corrected} = \hat{\beta}_X - \widehat{Bias}_1(\hat{\beta}_X).$$

The bias estimated by the parametric bootstrap is similarly derived as

$$\widehat{Bias}_p(\hat{\beta}_X) = \left(E_p(\hat{\beta}_X^B) - E_0(\hat{\beta}_X^B)\right),$$

with corresponding correction

$$\hat{\beta}_{X,p}^{corrected} = \hat{\beta}_X - \widehat{Bias}_p(\hat{\beta}_X).$$

3 Results

3.1 Data

3.1.1 MESA cohort

Summary statistics for the MESA cohort are in Table 1. Mean CIMT was 0.68 mm. The other variables summarized are the ones that were included as covariates in the health model. The mean age was 62 years, and the cohort was 52% female. 39% were white, 27% African-American, 22% Hispanic, and 12% Chinese. 44% had hypertension and 15% used a statin drug. The highest percentage of participants resided in Los Angeles (19.7%), but the distribution across the 6 cities was quite homogeneous. Only the 5,298 participants that had complete values of all the variables listed in Tables 1 were included in the analysis.

3.1.2 Monitoring data

Concentrations of the four pollutants by monitoring network are shown in Table 2. Table 2 indicates that the EC and OC concentrations measured by CSN monitors tended to be higher than those

measured by IMPROVE monitors. Average Si and S concentrations measured by CSN monitors were also higher than the IMPROVE averages, but relative to their standard deviations the differences between CSN and IMPROVE monitors in Si and S concentrations were not as great as the EC and OC concentrations.

3.1.3 Geographic Covariates

The geographic variables that were selected as a result of the pre-processing procedure discussed in Section 2.1.3 are shown in Table 3. Table 2 shows the distributions of select geographic covariates, by monitoring network and at MESA locations. The summaries in Table 2 reflect the difference in placement between the IMPROVE and CSN monitors. Although relatively few monitors belonging to either IMPROVE or CSN were within 150 m of an A1 road there was a larger proportion of CSN monitors within 150 m of an A3 road (44%) than IMPROVE (19%). The median distance to commercial and service centers was much smaller for CSN monitors (127 m) than it was for IMPROVE monitors (4696 m). The median population density was much larger for CSN monitors (805 people/mi^2) than for IMPROVE monitors (only 3 people/mi^2). The median summer NDVI values within 250 m were slightly smaller for CSN monitors than for IMPROVE monitors, indicating IMPROVE monitors were located in greener areas. Table 2 also shows that MESA participant locations had covariate distributions that more closely mirrored the CSN monitors, as is especially evident for the number of sites less than 150 m from an A3 road and median population density.

3.2 Spatial prediction models

3.2.1 Model evaluation

The selected models corresponding to lowest cross-validated R^2 all used PLS and UK. Table 4 shows the number of PLS components used and the R^2 and RMSEP for the selected prediction models. The CV statistics for the PLS only models are shown to illustrate the extent to which UK improved prediction accuracy. EC and OC were minimally improved; there was more improvement evident for Si and substantial improvement for the S predictions. The ratio of the nugget to the sill given in Table 4 also indicates the importance of spatial smoothing. For a fixed range, smaller values of this ratio indicate that there is more information about concentration variability at nearby locations and thus UK predictions draw heavily from nearby monitors.

3.2.2 Interpretation of partial least squares

Figure 2 can be used to examine which of the geographic covariates were most important for explaining pollutant variability. Specifically, Figure 2 summarizes the $p \times 1$ vector $\mathbf{M} = \mathbf{H}\hat{\boldsymbol{\alpha}}_{pls}$, the vector such that $\mathbf{RM}$ equals the PLS-only predicted exposures. Positive coefficient values indicate that increases in that geographic covariate were associated with higher predicted exposure; the larger the value, the more marked the association. For EC, OC and S, population density was heavily relied on to explain exposure. In addition, the NDVI, intense land use, emissions, and line-length variables were all positively associated with exposure, while the distance to source variables were negatively associated. The NDVI variables were more heavily exploited for OC and S than they were for EC. For Si, the NDVI and intense land use variables appeared to be the most informative and were mostly negatively associated with Si exposure. Proximity to features appeared to be informative for all four pollutants.

3.2.3 Exposure predictions

Figure 1 shows the predicted national concentrations, with finer detail illustrated for St. Paul, MN. The EC and OC predictions were much higher in the middle of urban areas, and quickly dissipated further from urban centers. S predictions were high across the midwestern and eastern states and in the Los Angeles area, and lower in the plains and mountains. Si predictions were low in most urban areas, and high in desert states.

Table 2 summarizes the predicted exposures for the MESA participants. Mean predicted EC and OC exposure concentration were 0.74 and 2.17 μg/m^3, respectively. Mean predicted Si and S exposure concentration was 0.09 ng/m^3 and 0.78 μg/m^3, respectively.

3.3 Health models

The results from the naïve health model that did not include any measurement error correction, as well as the results from the health modeling that included bootstrap-corrected point estimates and standard errors of $\hat{\beta}_X$, are displayed in Table 5. The naïve analysis found significant associations between OC, Si, and S and elevated CIMT. There was also evidence of association with EC, but this was not statistically significant. The point estimates and standard errors for the EC and OC health effects were virtually unchanged when measurement error correction was implemented, while the bootstrap-corrected standard errors for Si and S were about 50% larger than their respective naïve

estimates. The estimated biases resulting from the classical-like measurement error were so small as to be uninteresting from an epidemiologic perspective; however we found notable differences between the bias estimates from the parameter and parametric bootstraps which we discuss in the Appendix.

4 Discussion

We have presented a comprehensive 2-step approach to analyzing long-term effects of air pollution exposure. We estimated exposure to $PM_{2.5}$ components and assessed their impact on a sub-clinical measure of atherosclerosis (CIMT) in the MESA cohort. This approach includes a national prediction model for individual components and correction for measurement error in the epidemiologic analysis using a methodology that accounts for differing amounts of spatial structure in the exposure surfaces. We find that a national approach to exposure modeling is reasonable and performs well in terms of prediction accuracy, with R^2 no lower than 0.62 for any of the $PM_{2.5}$ components and ranging to as large as 0.95. Our exposure models are also useful in terms of understanding the spatial nature of our exposure surfaces, which can be ascertained by comparing cross-validation results from models based purely on PLS to those from models that also incorporate kriging.

To interpret the measurement error results, it is helpful to take note of the relative importance of the PLS and kriging aspects of the four prediction models. For EC and OC, using PLS alone was sufficient to make accurate predictions, whereas the spatial smoothing from UK was much more important in improving prediction accuracy for Si and S. It is accordingly no coincidence that the bootstrap-corrected standard error estimates for EC and OC were unchanged from the naïve estimates, while the corrected SE estimates for Si and S were about 50% larger than their respective naïve estimates. The fact that the EC and OC exposure predictions were derived mostly from the PLS components only with independent residuals implies that the Berkson-like error was almost pure Berkson error (i.e., independent across location), which is correctly accounted for by naïve standard error estimates. On the other hand, the importance of kriging for Si and S indicates that much more spatial smoothing took place for these pollutants which induced spatial correlation in the residual difference between true and predicted exposure. Accordingly, standard errors that correctly account for the Berkson-like error in these two pollutants are inflated because the correlated errors in the predictions translate into correlated residuals in the disease model that are not accounted for by naïve standard error estimates (Szpiro et al., 2011b) . The fact that the standard error estimates from the parameter and

parametric bootstraps (which account for both Berkson-like and classical-like error) and the partial parametric bootstrap (which accounts only for Berkson-like error) were so similar further indicates that the larger corrected SE estimates were indeed most likely a result of the Berkson-like error. None of our measurement error analyses indicated that any important bias was induced by the classical-like error.

We can interpret our measurement error findings from the perspective of how the exposure variability is partitioned into between-city and within-city variability. For Si and S most of the variability is between-cities, so naïve standard error estimates fail to account for uncertainty due to the choice of the specific six MESA Air cities. On the other hand, for EC and OC the variability in exposure is primarily within-city so the choice of specific cities is not as important. For all components, we cannot rule out confounding by city in the present analysis without adjusting for city with a fixed effect. Doing this results in a significant loss of study power (results not shown).

We also note that our measurement error correction methods rely on a linear health model. Since the exposure modeling was done on the square root scale and the health modeling on the native scale, the Berkson-like error could potentially induce bias in $\hat{\beta}_X$. However, if this were a significant source of bias the bootstrap methods would detect it. In our application we did not see any evidence of bias from the Berkson-like error.

Our results show that careful investigation of the exposure model characteristics can help to clarify the implications for the subsequence epidemiologic analyses that use the predicted exposures. As is pointed out in Szpiro et al. (2011a), such an overarching framework that considers the end goal of health modeling is a more reasonable and scientifically valid approach than treating exposure models as if they exist for their own sake; doing the latter can even potentially lead to selecting exposure models that generate accurate predictions but lead to anti-conservative or more biased health effect inference. This analysis serves as an example that will inform ongoing efforts by our group and others to construct and utilize exposure prediction models that are most suitable for epidemiologic studies.

Acknowledgments

Funding for this research was provided by grants T32 ES015459 and P50 ES015915 from the National Institute of Environmental Health Sciences. 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). MESA is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159 through N01-HC-95169 and UL1-RR-024156. MESA Air is funded by the US EPA's Science to Achieve Results (STAR) Program Grant #RD831697.

References

U.S. EPA 2009. Integrated Science Assessment for Particulate Matter (Report No. EPA/600/R-08/139F) U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC. 3-31.

H. Abdi. Partial Least Squares (PLS) regression. *Encyclopedia of Social Sciences Research Methods (ed. M. Lewis–Beck, A. Bryman and T. Futing)*, pages 1–7, 2003.

J.R. Anderson. *A land use and land cover classification system for use with remote sensor data.* Number 964. US Govt. Print. Off., 1976.

M. Brauer, G. Hoek, P. van Vliet, K. Meliefste, P. Fischer, U. Gehring, J. Heinrich, J. Cyrys, T. Bellander, M. Lewne, et al. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. *Epidemiology*, 14(2):228–239, 2003.

N. Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

R.A. Eldred, T.A. Cahill, and M. Pitchford. IMPROVE- A New Remote Area Particulate Monitoring System for Visibility Studies. *Proceedings of the 81st Annual Meeting of the APCA, Dallas, TX*, 1988.

A. Gryparis, C.J. Paciorek, A. Zeka, J. Schwartz, and B.A. Coull. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10(2):258–274, 2009. ISSN 1465-4644.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

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

M. Jerrett, R.T. Burnett, R. Ma, C.A. Pope III, D. Krewski, K.B. Newbold, G. Thurston, Y. Shi, N. Finkelstein, E.E. Calle, et al. Spatial analysis of air pollution and mortality in los angeles. *Epidemiology*, 16(6):727–736, 2005.

S.Y. Kim, L. Sheppard, and H. Kim. Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology*, 20(3):442–450, 2009.

N. Künzli, M. Jerrett, W.J. Mack, B. Beckerman, L. LaBree, F. Gilliland, D. Thomas, J. Peters, and H.N. Hodis. Ambient air pollution and atherosclerosis in los angeles. *Environmental Health Perspectives*, 113(2):201–206, 2005.

L. Madsen, D. Ruppert, and NS Altman. Regression with spatially misaligned data. *Environmetrics*, 19(5):453–467, 2008.

L.D. Mercer, A.A. Szpiro, Lianne Sheppard, Johan Lindström, S.D. Adar, R.W. Allen, E.L. Avol, A.P. Oron, T. Larson, S.L.-J. Liu, and J.D. Kaufman. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NO_X) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric Environment*, 45(26):4412–4420, 2011.

K.A. Miller, D.S. Siscovick, L. Sheppard, K. Shepherd, J.H. Sullivan, G.L. Anderson, and J.D. Kaufman. Long-term exposure to air pollution and incidence of cardiovascular events in women. *New England Journal of Medicine*, 356(5):447–458, 2007.

J. Molitor, M. Jerrett, C.C. Chang, N.T. Molitor, J. Gauderman, K. Berhane, R. McConnell, F. Lurmann, J. Wu, A. Winer, et al. Assessing uncertainty in spatial exposure models for air pollution health effects assessment. *Environmental Health Perspectives*, 115(8):1147–1153, 2007.

C.A. Pope, R.T. Burnett, M.J. Thun, E.E. Calle, D. Krewski, K. Ito, and G.D. Thurston. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA: The Journal of the American Medical Association*, 287(9):1132–1141, 2002.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

P.J. Ribeiro Jr and Peter J. Diggle. geoR: a package for geostatistical analysis. *R-NEWS*, 1(2):14–18, June 2001. URL `http://CRAN.R-project.org/doc/Rnews/`.

J.M. Samet, F. Dominici, F.C. Curriero, I. Coursac, and S.L. Zeger. Fine particulate air pollution and mortality in 20 US cities, 1987–1994. *New England Journal of Medicine*, 343(24):1742–1749, 2000.

P.D. Sampson, A.A. Szpiro, L. Sheppard, J. Lindström, and J.D. Kaufman. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*, 45 (36):6593–6606, 2009.

LA Stefanski and JR Cook. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, pages 1247–1256, 1995.

A.A. Szpiro, C.J. Paciorek, and L. Sheppard. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology*, 22(5):680–685, 2011a.

A.A. Szpiro, L. Sheppard, and T. Lumley. Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12(4):610–623, 2011b.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

S. Vedal, J.D. Kaufman, T.V. Larson, P.D. Sampson, L Sheppard, C.D. Simpson, A.A. Szpiro, J.D. McDonald, A.K. Lund, and M.J. Campen. University of Washington/Lovelace Respiratory Research Institute National Particle Component Toxicity (NPACT) Initiative: Integrated Epidemiological and Toxicological Cardiovascular Studies to Identify Toxic Components and Sources of Fine Particulate Matter (DRAFT). *Heath Effects Institute, Boston, MA*, 2012.

R. Wehrens, B.H. Mevik, and M.B.H. Mevik. The PLS Package. *Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*, 2006.

S.L. Zeger, D. Thomas, F. Dominici, J.M. Samet, J. Schwartz, D. Dockery, and A. Cohen. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental Health Perspectives*, 108(5):419–426, 2000.

Variable	N	Mean (SD) or %
CIMT	5501	0.68 (0.19)
Age	5501	61.9 (10.1)
Weight (lb)	5501	173.0 (37.5)
Height (cm)	5501	166.6 (10.0)
Waist (cm)	5500	97.8 (14.1)
Body surface area	5501	1.9 (0.2)
BMI (kg/m^2)	5501	28.2 (5.3)
DBP	5499	71.8 (10.3)
Gender		
Female	2872	52.2
Male	2629	47.8
Race		
White, caucasian	2168	39.4
Chinese American	675	12.3
Black, African-American	1459	26.5
Hispanic	1199	21.8
Site		
New York	867	15.8
Baltimore	776	14.1
St. Paul & Minneapolis	899	16.3
Chicago	998	18.1
Los Angeles	1083	19.7
Education		
Complete high school	991	18.0
Some college	1571	28.6
Complete college	2010	36.5
Missing	13	0.2
Income		
$< \$12,000$	566	10.3
\$12,000-24,999	1022	18.6
\$25000-49999	1543	28.0
\$50000-74999	901	16.4
$> \$75000$	1271	23.1
Missing	198	3.6
Hypertension		
No	3106	56.5
Yes	2395	43.5
Statin use		
No	4681	85.1
Yes	817	14.9
Missing	3	0.1

Table 1: Summary of characteristics of the MESA cohort. Only variables that were used in the health modeling are summarized here.

Location	# Sites	EC ($\mu g/m^3$)	OC ($\mu g/m^3$)	Si (ng/m^3)	S ($\mu g/m^3$)	#Sites <150m to A1 (%)	# Sites <150m to A3 (%)	Med dist to Comm[a]	Med Pop dens[b]	NDVI[c]
IMPROVE	190	0.19 (0.18)	0.93 (0.55)	0.16 (0.12)	0.41 (0.27)	4 (2)	36 (19)	4696	3	150
CSN	98	0.66 (0.24)	2.23 (0.71)	0.10 (0.09)	0.69 (0.25)	3 (3)	43 (44)	127	805	140
All monitors	288	0.37 (0.30)	1.43 (0.88)	0.14 (0.11)	0.51 (0.29)	7 (2)	79 (27)	1235	20	146
MESA Air	5501	0.74 (0.18)	2.17 (0.36)	0.09 (0.03)	0.78 (0.15)	349 (6)	2763 (50)	302	3496	137

[a] Median distance to commercial or service centers, in meters
[b] People/mi^2 for census block/block group monitor/subject belongs to
[c] Median value of summer NDVI medians within 250m buffer

Table 2: Mean (SD) concentration at IMPROVE and CSN monitoring networks and over both networks taken together; and predicted concentrations for the MESA Air cohort. Also shown are summary statistics of selected land-use regression covariates.

Figure 2 abbreviation	Variable description	Buffer sizes
	Distance to features, in km[a]	
distance to features	A1 road	NA
	Nearest road	NA
	Airport	NA
	Large airport	NA
	Port	NA
	Coastline‡	NA
	Commercial or service center	NA
	Railroad	NA
	Railyard	NA
	Emissions[b]	
so2	SO_2	30km
pm25	$PM^{\dagger}_{2.5}$	30km
pm10	$PM^{\dagger}_{10}$	30km
nox	NO_X	30km
	Population	
population	$\log_{10}$ population density	500m, 1km, 1.5km, 2km, 2.5km, 3km, 5km, 10km, 15km
	NDVI	
ndvi.winter	Median winter	250m, 500m, 1km, 2.5km, 5km, 7.5km, 10km
ndvi.summer	Median summer	250m, 500m, 1km, 2.5km, 5km, 7.5km, 10km
ndvi.q75	75^{th} %ile	250m, 500m, 1km, 2.5km, 5km, 7.5km, 10km
ndvi.q50	50^{th} %ile	250m, 500m, 1km, 2.5km, 5km, 7.5km, 10km
ndvi.q25	25^{th} %ile	250m, 500m, 1km, 2.5km, 5km, 7.5km, 10km
	Land use	
transport	Transportation, communities and utilities	750m, 3km, 5km, 10km, 15km
transition	Transitional areas	15km
stream	Streams and canals	3km†, 5km, 10km, 15km
shrub	Shrub and brush rangeland	1.5km, 3km, 5km, 10km, 15km
resi	Residential	400m, 500m, 750m, 1km, 1.5km, 3km, 5km, 10km, 15km
oth.urban	Other urban or built-up	400m†, 500m, 1.5km, 3km, 5km, 10km, 15km
mix.range	Mixed rangeland	3km, 5km, 10km, 15km
mix.forest	Mixed forest land	750m, 1km, 1.5km, 3km, 5km, 10km, 15km
lakes	Lakes †	10 km
industrial	Industrial	1km*, 1.5km*, 3km, 5km, 10km, 15km
industcomm	Industrial and commercial complexes †	15km
herb.range	Herbaceous rangeland	3km†, 5km, 10km
green	Evergreen forest land	400m, 500m, 750m, 1km, 1.5km, 3km, 5km, 10km, 15km
forest	Deciduous forest land	750m, 1km, 1.5km, 3km, 5km, 10km, 15km
crop	Cropland and pasture	400m, 500m, 750m, 1km, 1.5km, 3km, 5km, 10km, 15km
comm	Commercial and services	500m, 750m, 1km, 1.5km, 3km, 5km, 10km, 15km
	Line lengths	
a23	Total dist of A2 and A3 roads within buffer	100m, 150m, 300m, 400m, 500m, 750m, 1km, 1.5km, 3km, 5km
a1	Total dist of A1 roads within buffer	1km, 1.5km, 3km, 5km

[a] Truncated at 25km and $\log_{10}$ transformed
[b] Tons per year of emissions from tall stacks
† Variable used for modelling Si, S only
* Variable used for modelling EC, OC only
‡ $\log_{10}$ and untransformed values both included

Table 3: Land-use regression covariates and (where applicable) covariate buffer sizes that made it through pre-processing and were considered by PLS. Most variables were used in each of the four $PM_{2.5}$ component models; however the pre-processing procedure selected some variables for EC and OC that were not selected for Si and S, and vice versa. This is due to the fact that the monitors used to measure EC and OC were not all identical to the ones used to measure Si and S.

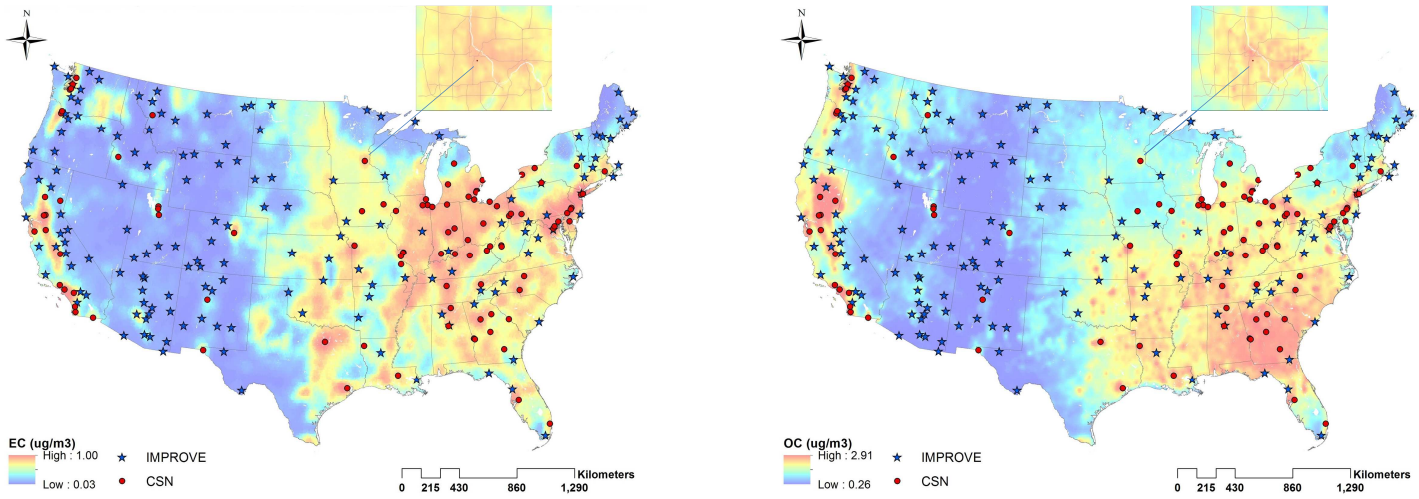
		R^2		RMSEP		Est. UK pars			
Pollutant	# Scores	PLS only	PLS+UK	PLS only	PLS+UK	$(\tau^2)^a$	$(\sigma^2)^b$	$(\phi)^c$	τ^2/σ^2
EC	3	0.79	0.82	0.11	0.10	0.0074	0.0025	413	2.96
OC	2	0.60	0.69	0.22	0.20	0.0251	0.0199	304	1.26
Si	2	0.36	0.62	0.10	0.08	0.0043	0.0086	2789	0.50
S	2	0.63	0.95	0.13	0.05	0.0007	0.0251	2145	0.03

[a] Nugget used in kriging
[b] Partial sill used in kriging
[c] Range used in kriging

Table 4: Cross-validated R^2 and RMSEP for each component of $PM_{2.5}$, both for when PLS only was used and when PLS was used in conjunction with universal kriging. The estimated kriging parameters from the likelihood fit on the entire data set for each pollutant is also shown. The R^2 and RMSEP for the PLS + UK models reflect the effectiveness of the final prediction models.

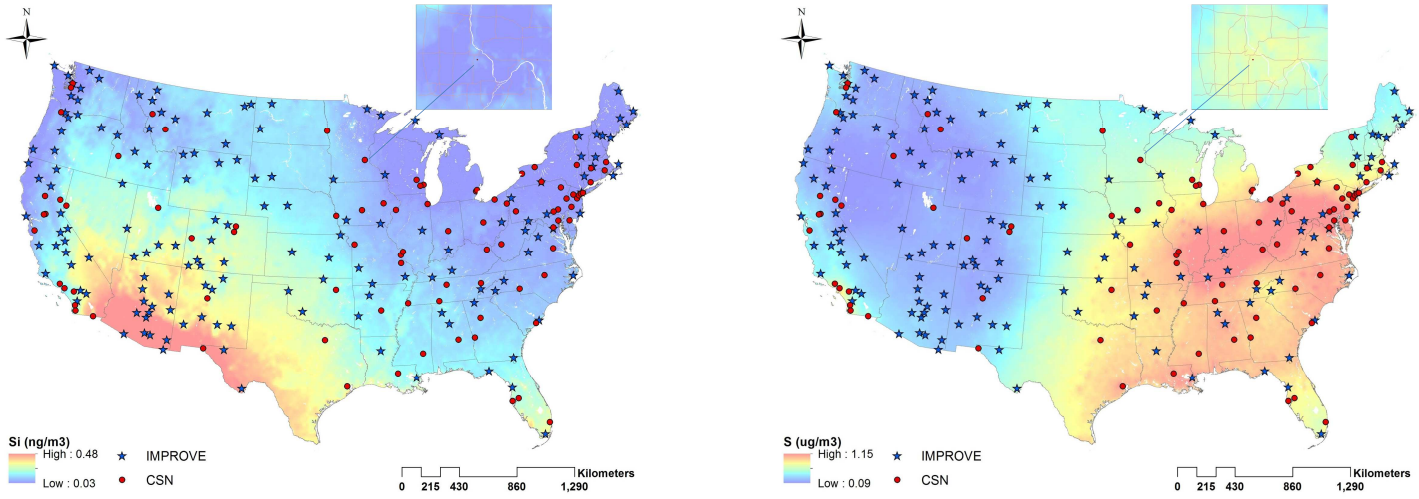
	EC		OC		Si		S	
	$\hat{\beta}_X$	$\hat{SE}(\hat{\beta}_X)$	$\hat{\beta}_X$	$\hat{SE}(\hat{\beta}_X)$	$\hat{\beta}_X$	$\hat{SE}(\hat{\beta}_X)$	$\hat{\beta}_X$	$\hat{SE}(\hat{\beta}_X)$
Naïve	0.001	0.014	0.025	0.008	0.408	0.081	0.055	0.017
Partial parametric	0.001	0.015	0.025	0.008	0.408	0.126	0.055	0.025
Parameter	0.001	0.015	0.025	0.008	0.408	0.127	0.055	0.025
Parametric	0.001	0.015	0.025	0.008	0.405	0.131	0.055	0.025

Table 5: Point estimates and standard errors for the different pollutants, using naïve analysis and with bootstrap correction for measurement error in covariate of interest. The two parameter bootstrap results are from 30,000 bootstrap samples, while the parametric results are from 15,000 samples.



(a) EC

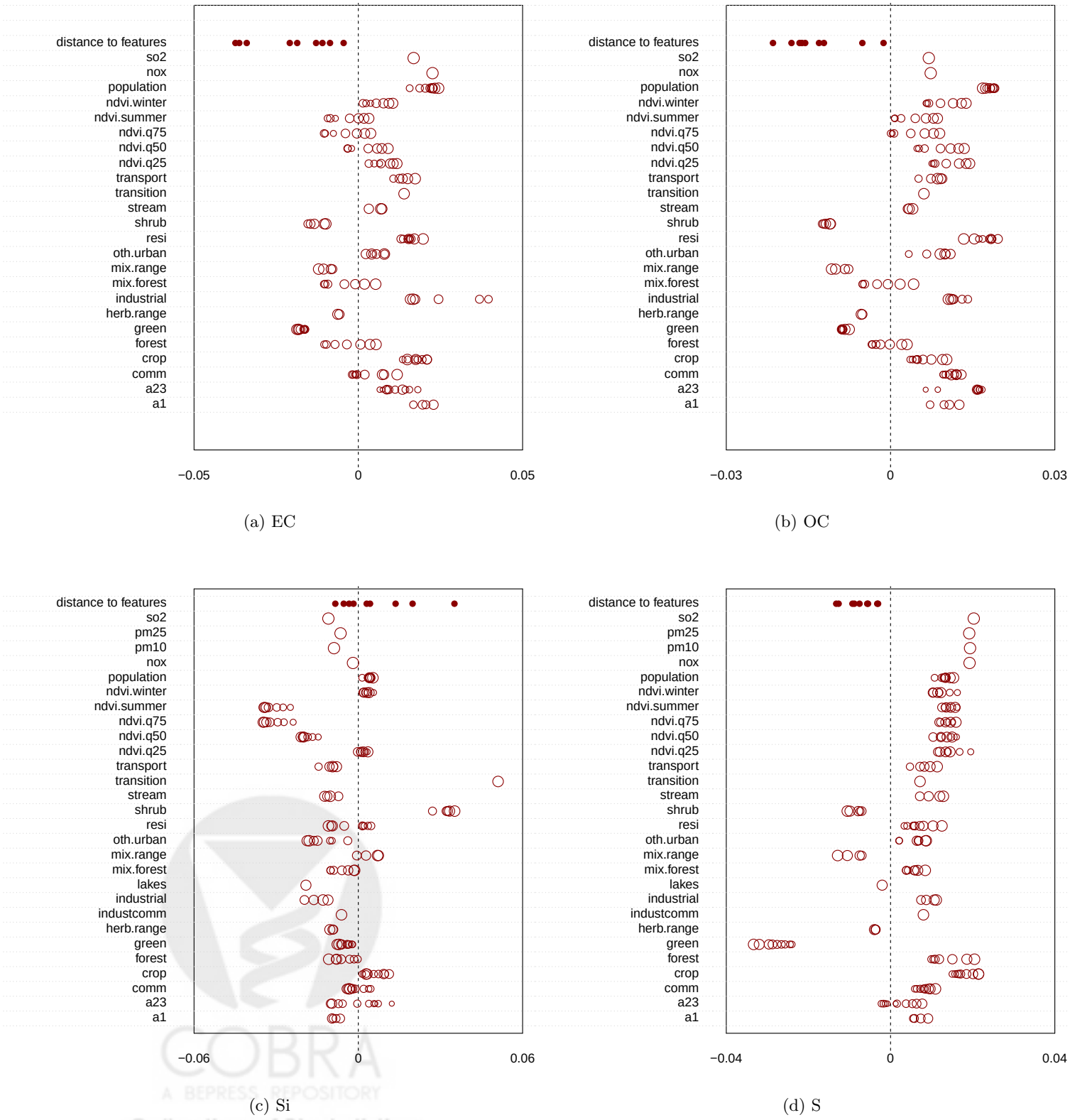
(b) OC



(c) Si

(d) S

Figure 1: Locations of IMPROVE and CSN monitors and predicted national average $PM_{2.5}$ component concentrations from final predictions models. Predictions are also shown for St. Paul, MN.



(a) EC

(b) OC

(c) Si

(d) S

Figure 2: Coefficients of the PLS fit, by geographic covariate type. The size of each circle represents the buffer size, with larger circles indicating larger buffers. Explanation of variable abbreviations are given in Table 3.

Appendix

The extension of the parameter bootstrap discussed in Section 2.3.2 wherein we sample $(\hat{\boldsymbol{\alpha}}_{uk,j}, \log(\hat{\boldsymbol{\theta}}_j))$ from a covariance multiplied by a non-negative λ provides additional insight into how to estimate the bias of $\hat{\beta}_X$ resulting from the classical-like error, and can be thought of in the framework of simulation extrapolation (SIMEX) (Stefanski and Cook, 1995). Consider first the partial parametric bootstrap. Though we use the originally estimated parameters $(\hat{\boldsymbol{\alpha}}_{uk}, \log(\hat{\boldsymbol{\theta}}))$ throughout (corresponding to $\lambda = 0$), these original estimates are a realization from a sampling distribution with some amount of variance, which can be thought of heuristically as "one unit of variance." The bias estimate from the parameter bootstrap with $\lambda = 1$ (corresponding to "two units of variance") assumes that the bias from the classical-like error obtained by going from $\lambda = 0$ to $\lambda = 1$ is the same as the bias induced by using the originally estimated parameters instead of the true, unknown parameters; in other words, the bias is treated as linear in λ. However if we perform the parameter bootstrap using different values of λ and estimate the bias for each one, we can get a more flexible representation of how the bias varies as a function of λ. Plotting realized $\widehat{Bias}_\lambda(\hat{\beta}_X)$ versus λ for several values of λ and extrapolating to $\widehat{Bias}_{-1}(\hat{\beta}_X)$ gives an alternative estimate of the bias. This extension is equivalent to performing a SIMEX analysis to extrapolate to the hypothetical setting where the variance of the measurement error is zero (Stefanski and Cook, 1995). We performed the parameter bootstrap using sample sizes of 30,000, sampling $(\hat{\boldsymbol{\alpha}}_{uk,j}, \log(\hat{\boldsymbol{\theta}}_j))$ from the inverse Hessian inflated by factors of $\lambda \in \{0, 0.5, 1, 1.5, 2\}$ and plotted the corresponding $\widehat{Bias}_\lambda(\hat{\beta}_X)$ against these values of λ. We then performed both linear and quadratic extrapolation to $\widehat{Bias}_{-1}(\hat{\beta}_X)$. The SIMEX-corrected estimate of $\hat{\beta}_X$ is defined as:

$$\hat{\beta}_{X,S}^{corrected} = \hat{\beta}_X + \widehat{Bias}_{-1}(\hat{\beta}_X)$$

This estimate was compared to the other corrected estimates defined in Section 2.3.2. Note that although the generalizations of the parameter bootstrap used 30,000 samples, the 15,000 parametric bootstrap samples were only compared to the first 15,000 partial parametric samples.

Figure A1 shows the results of the SIMEX implementation of the parameter bootstrap using linear and quadratic extrapolation. The green line indicates the linear correction from the parametric bootstrap. The choice of extrapolating function did not affect $\hat{\beta}_X$ for Si or OC, while there were slight differences for the other two components. Overall, while the SIMEX bias corrections did not suggest

any meaningful bias for any of the pollutants, all of these plots suggest that the bias from classical-like measurement error is away from the null, similar to previously published simulation results (Szpiro et al., 2011b). This is different from the usual bias toward the null from classical measurement error, confirming that additional caution is needed in the air pollution setting since we cannot always assume that ignoring measurement error results in conservative inference.

It is of interest that the bias estimated using SIMEX and the parameter bootstrap is noticeably different from the corresponding parametric bootstrap bias estimates; in the cases of OC and S, the parametric bootstrap estimates a bias toward the null. There are two key assumptions that justify regarding the parameter bootstrap as a more efficient version of the parametric bootstrap. The first assumption is that the estimated sampling distribution of $\log(\hat{\boldsymbol{\theta}}_j)$ from which the parameter bootstrap estimates are sampled is a good approximation to the true sampling distribution. The second, heuristically stated, is that the process of sampling exposures and estimating exposure model parameters can be considered to be independent of each other. In other words, the exposure model parameters in the parametric bootstrap are "decoupled" from the data from which they were derived, which is clearly the case for the parameter bootstrap since $\log(\hat{\boldsymbol{\theta}}_j)$ is sampled independently of $\mathbf{X}_j^*$. We examined both of these assumptions for OC, Si and S, as the differences between the parameter and parametric bias estimates were most pronounced for these pollutants. To examine the first assumption we compared density plots of $\log(\hat{\boldsymbol{\theta}}_j)$ from the parametric bootstrap to the ones used for the parameter bootstrap. This showed some mismatch between the distributions, indicating that this first assumption might be violated. To examine the second assumption, we sampled exposure model parameters at random from the empirical distribution of the parametric bootstrap $\log(\hat{\boldsymbol{\theta}}_j)$, and compared the resulting $\hat{\beta}_X^B$ to those derived using matching $\log(\hat{\boldsymbol{\theta}}_j)$ and $\mathbf{X}_j^*$ (see Szpiro et al. (2011b) for more details). We found that for OC and Si, the mean of the bootstrapped coefficients from the two approaches were significantly different from each other, indicating mild violation of the decoupling assumption. For S, however, the decoupling did not affect the bias, indicating that the discrepancy between the parametric and parameter bootstrap bias estimates must be due to violation of the first assumption.

In the end, however, it is important to point out that the estimated biases are so small that the bias-corrected effect estimates are no different from the naïve estimates, to a reasonable number of significant digits. Indeed we are only able to detect these biases with a very large number of bootstrap samples. The important impact of measurement error appears in the standard error estimates, and

the parameter and parametric bootstrap methods agree closely there.

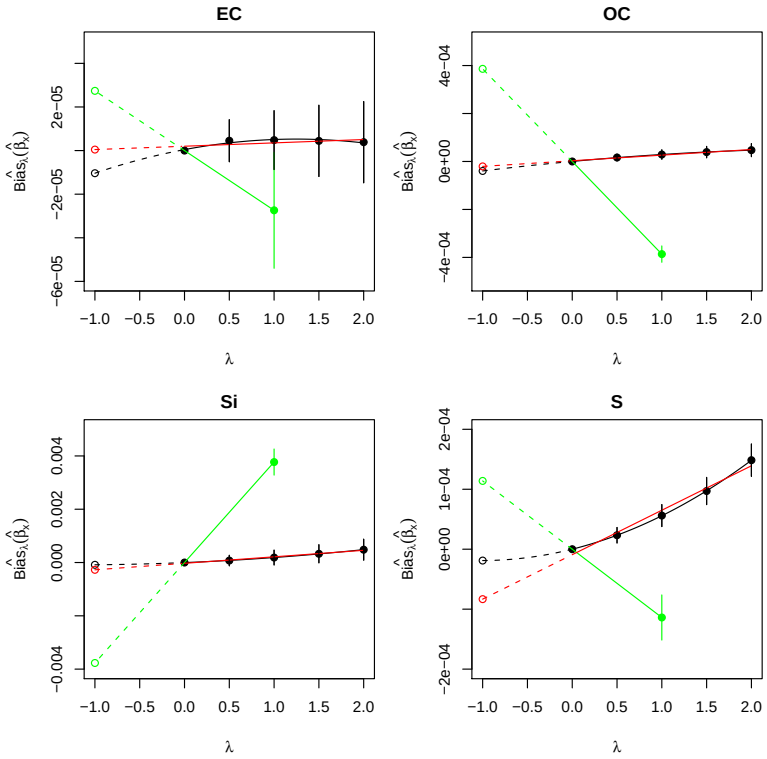


Figure A1: SIMEX bias estimates, when using either a linear or a quadratic extrapolation. The green line represents the bias extrapolation as estimated by the parametric bootstrap. Confidence intervals from a t-test testing zero bias are also shown.