12-11-2012

A Regionalized National Universal Kriging Model Using Partial Least Squares Regression for Estimating Annual PM2.5 Concentrations in Epidemiology

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

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

Adam A. Szpiro
University of Washington - Seattle Campus, aszpiro@u.washington.edu

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

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

See next page for additional authors

Authors
Paul D. Sampson, Mark Richards, Adam A. Szpiro, Silas Bergen, Lianne Sheppard, Timothy V. Larson, and Joel Kaufman

A Regionalized National Universal Kriging Model using Partial Least Squares Regression for Estimating Annual $PM_{2.5}$ Concentrations in Epidemiology

Paul D. Sampson[a*]

Mark Richards[b]

Adam A. Szpiro[c]

Silas Bergen[c]

Lianne Sheppard[d]

Timothy V. Larson[e]

Joel D. Kaufman[d]

[q] *Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA*

[b]*Department of Applied Mathematics, University of Washington, Box 352420, Seattle, WA 98195-2420, USA*

[c] *Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232, USA*

[d] *Department of Environmental and Occupational Health Sciences, University of Washington, Box 357234, Seattle, WA 98195-7234, USA*

[e]*Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700, USA*

Correspondence to: Paul D. Sampson, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, U.S.A. E-mail: pds@u.washington.edu; phone: 206-685-2664; fax: 206-685-6419

Abstract

Many cohort studies in environmental epidemiology require accurate modeling and prediction of fine scale spatial variation in ambient air quality across the U.S. This modeling requires the use of small spatial scale geographic or "land use" regression covariates and some degree of spatial smoothing. Furthermore, the details of the prediction of air quality by land use regression and the spatial variation in ambient air quality not explained by this regression should be allowed to vary across the continent due to the large scale heterogeneity in topography, climate, and sources of air pollution. This paper introduces a regionalized national universal kriging model for annual average fine particulate matter ($PM_{2.5}$) monitoring data across the U.S. To take full advantage of an extensive database of land use covariates we chose to use the method of Partial Least Squares, rather than variable selection, for the regression component of the model (the "universal" in "universal kriging") with regression coefficients and residual variogram models allowed to vary across three regions defined as West Coast, Mountain West, and East. We demonstrate a very high level of cross-validated accuracy of prediction with an overall R^2 of 0.88 and well-calibrated predictive intervals. In accord with the spatially varying characteristics of $PM_{2.5}$ on a national scale and differing kriging smoothness parameters, the accuracy of the prediction varies by region with predictive intervals being notably wider in the West Coast and Mountain West in contrast to the East.

Keywords: Ambient air quality, Land use regression, National air quality model, Partial Least Squares, Particulate matter, Universal kriging

1. Introduction

Residential predictions of ambient air quality concentrations are important for epidemiological cohort studies, particularly those conducted on a national scale. We focus on long-term averages of concentrations of fine particular matter, $PM_{2.5}$. Predictions of average air quality levels may be derived from spatial (Hart et al., 2009; Hystad et al., 2011; Mercer et al., 2011; Novotny et al., 2011) or spatio-temporal models (Yanosky et al., 2009; Paciorek et al., 2009; Szpiro et al., 2010; Sampson et al., 2011; Lindström et al., 2011). In this paper we develop a continental U.S national scale spatial model for year 2000 annual average $PM_{2.5}$ concentrations based on data from monitors in regulatory monitoring networks.

As monitoring sites in national regulatory networks are relatively sparse across broad regions of the country and as air quality levels are influenced by many small- and large-scale spatial features, accurate prediction requires a combination of geographic covariates such as distances from roads and other pollutant sources to capture small-scale variation and spatial smoothing for large-scale patterns. Regression models based on geographic covariates are traditionally termed "land use" regression (LUR, e.g. Moore et al. 2007; Ross et al. 2007; Hoek et al. 2008). We use a database of 265 GIS-based geographic covariates with multiple indirect measures of traffic, population density, land use, satellite-based vegetative index (NDVI), nearby pollutant emissions derived from emissions inventories, and distances to major sources of pollution. We incorporate spatial smoothing with the LUR by means of a geostatistical correlation model in order to exploit spatial information available in the monitoring dataset. Our model provides the basis for predictions at arbitrary spatial locations (assuming covariate values can be computed at all spatial locations) by universal kriging or "kriging with external drift" (see Wackernagel, 2010).

Almost all current applications of LUR in the literature, whether combined with a spatial correlation model or not, use some kind of variable selection procedure to choose a subset of variables that provide good, if not the optimal, predictions. As an alternative to variable selection, we use a Partial Least Squares (PLS) approach. We develop our LUR regression model from a small number of composite PLS scores, each defined as a linear combination of all available covariates. This approach is conceptually related to more the widely known method of regression on principal components (PCA); the distinction is that PLS components are based on the maximum covariance between the covariates and the monitoring data whereas PCA components are based on the covariance of the covariates alone. Details are presented in the methods section below.

A geostatistical spatial correlation model is a statistical characterization of spatial variation in pollutant levels not explained by the covariates in the regression model. This spatial variation in (residual) pollutant levels is influenced by variation in topography and climatological meteorological patterns that is not captured in our collection of geographic covariates and can be difficult to model explicitly.

We consider models developed on national and regional scales (with the United States partitioned into three large regions as shown below) for both the regression and spatial smoothing (kriging) parts of the models. We demonstrate that the best results in terms of the accuracy of cross-validated predictions and the coverage of cross-validated prediction intervals are obtained using regional regressions with regional residual variogram models.

The following sections detail the monitoring data and our extensive database of geographic covariates. We review the methods of Partial Least Squares regression and maximum likelihood estimation of a universal kriging model. We then explain our strategy for defining regional

analyses and assessing model fits using cross-validation. The last two sections present results and evaluative discussion of the methodology.

2. Methods

2.1 Monitoring data

Daily $PM_{2.5}$ concentration data from both the AQS and IMPROVE networks (http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm/, http://views.cira.colostate.edu/web/) were utilized to calculate an annual average at each monitoring location with data that met minimum inclusion criteria. We required a minimum of

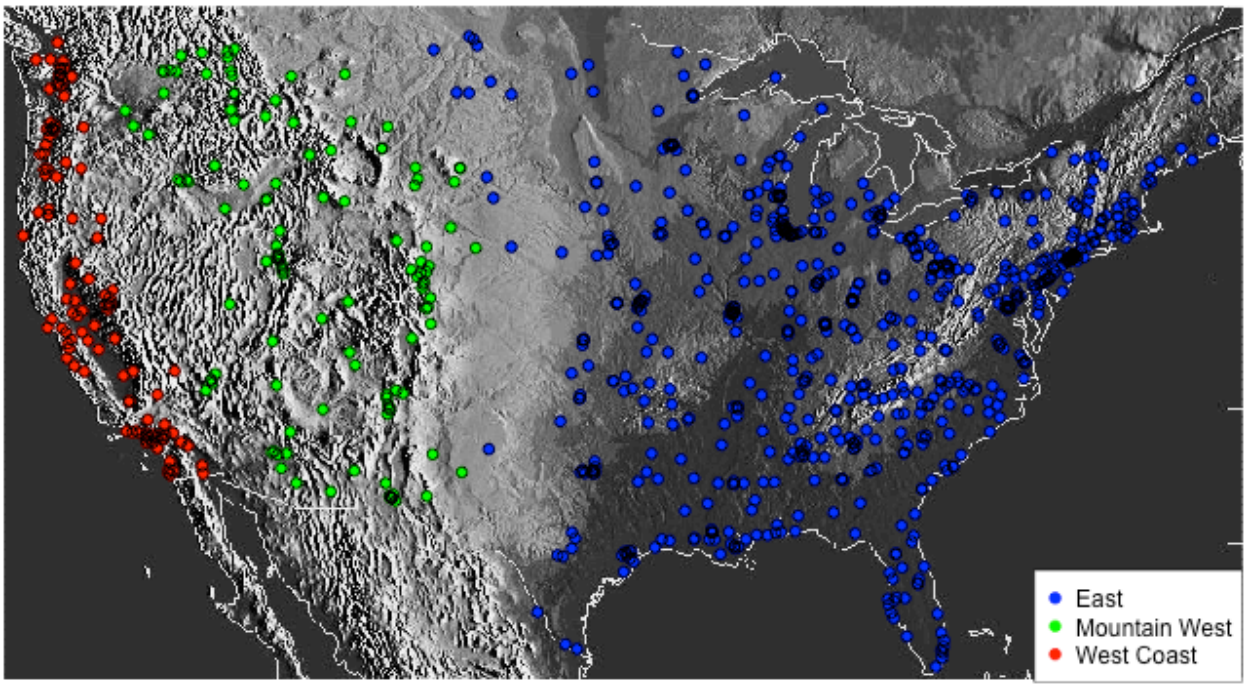

Fig. 1. U.S. topographic map with monitoring sites color coded according to the three modeling regions defined as explained in Section 2.5: East, Mountain West, and West Coast.

14 measurements per quarter for all four quarters. Of 1211 monitors providing $PM_{2.5}$ data in 2000, we found 903 met these criteria. Two sites were dropped for erroneous scoring of geographic covariates, leaving a total of 901. We subdivided these monitors into sets covering the eastern two thirds of the country, called the "East" (n=673), the "Mountain West" (n=120) and the "West Coast" (n=108) as explained in Section 2.5 and illustrated in Figure 1.

2.2 Geographic covariates

Our analysis considered an initial set of 265 distinct GIS-based geographic covariates, which was reduced to 171 prior to analysis. We combined some variables such as the within buffer road lengths in census feature class codes A2 and A3, and dropped others due to sparse discrete values (i.e. more than 85% of the values were identical). As summarized in Table 1, the final set of geographic covariates includes: (i) population in buffers from 5 to 15 km around target locations, (ii) total emissions of CO, NOx, PM_{10}, $PM_{2.5}$, and SO_2 (tons per year) in 15 and 30 km buffers, (iii) percentages of land according to 12 land use categories in circular buffers from 50 meters to 5 km, (iv) summaries of the distribution of the satellite-based MODIS Normalized Difference Vegetation Index, NDVI, in buffers from 250 meters to 10 km, (v) measures of impervious surfaces within buffers from 50 to 5000 meters, (vi) 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 circular buffers from 50 to 5000 meters around sites of interest, and (vii) distances to commercial zones, airports, small shipping ports, railroads, and railway yards. Data sources are provided in Appendix 1.

Table 1. Geographic covariates considered in the land use regression component of the universal kriging model. (Data sources listed in Appendix 1.)

Predictor Variable Category (n)	Units	Buffer Radii (meters)
Population (3)	Sum of people	5000, 10000, 15000
Emissions (8): NO_x, PM_{10}, $PM_{2.5}$, SO_2	Tons per year	15000, 30000
Land use (86)[a]: mixed forest, deciduous forest, evergreen, crop, pasture, grass, shrub, water, high development, medium development, low development, open development	Percent	50, 100, 150, 300, 400, 500, 750, 1000, 3000, 5000
Vegetative index (NDVI) (35): Winter average, Summer average; $75^{th,}$, 50^{th}, and 25^{th} quantiles	n/a	250, 500, 1000, 2500, 5000, 7500, 10000
Impervious Surfaces (10): Average within buffer	n/a	50, 100, 150, 300, 400, 500, 750, 1000, 3000, 5000
Roadway (18): Sum of A1 road lengths; Sum of A2 + A3 road lengths	Meters	A1: 400, 500, 750, 1000, 1500, 3000, 5000 A2+A3: 50, 100, 150, 300, 400, 500, 750, 1000, 1500, 3000, 5000
Distance to features (11): Commercial zone; A1, A2, A3 roadways; Large airport; Any airport; Large, Medium, Small shipping port; Railroad; Railyard	Meters, log10	n/a

[a] The number of buffers with data vary by land use category. Only the high, medium, low, and open development categories have all 10 buffer sizes.

2.3 Partial Least Squares regression

The GIS-based dataset of spatial covariates for $PM_{2.5}$ concentrations provides groups of highly correlated spatial covariates. For example, the composite lengths of A1 roads in buffers of varying sizes from 50 to 5000 meters around specified locations are necessarily highly correlated. Furthermore these variables are highly negatively correlated with the distance to the nearest A1 road. Percentages of property in various land use categories are similarly correlated across buffer sizes and 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 (e.g. Su et al., 2009; Mercer et al., 2011), (b) shrinkage or regularization, perhaps including variable selection as in a "lasso" approach (Tibshirani, 1996; Mercer et al., 2011), or (c) dimension reduction via regression on a smaller number of composite covariate scores. Our fundamental objective is high quality predictions and we prefer not to choose a method that would select one particular buffer size for inclusion in our model while ignoring neighboring buffer sizes, or one particular land use categorization at the expense of correlated land use categorizations.

The method of regressing on a small number of composite covariate scores using PLS regression to define the composite scores is well-established, especially in chemometric fields of application (see, for example, Garthwaite, 1994, Wold et al., 2001, or Abdi, 2010). The description of the composite scores in terms of individual variable loadings facilitates comparison of regression models across the three geographic modeling regions we consider here. PLS regressions were computed using the `pls` package for the R system (`http://cran.r-project.org/web/packages/pls`).

We summarize in brief the essentials of PLS regression following the classical notation of Wold et al (2001) or Abdi (2010). Let Y denote an $n \times 1$ vector of annual mean concentrations at n monitoring sites and let $\mathbf{X}$ denote the $n \times p$ matrix of geographic covariates. To simplify the exposition in this section, we will assume that Y has been centered to have mean zero and that the columns of $\mathbf{X}$ have been standardized to have covariate variances equal to 1 (or column sums of squares equal to 1). As in a principal components analysis (PCA), the covariate matrix is decomposed into a product of an $n \times p$ matrix of orthogonal scores $\mathbf{T}$ (often considered as a set of "latent vectors") and a $p \times p$ matrix of loadings $\mathbf{P}$ so that $\mathbf{X} = \mathbf{TP}'$. The columns of the PLS scores matrix $\mathbf{T}$ are computed with a sequential algorithm (most commonly the NIPALS algorithm; see Abdi 2010) to reflect the covariances between the Y and the columns of $\mathbf{X}$, rather than to explain the variances and covariances among the columns of $\mathbf{X}$ as is done in PCA.

The first column of scores, t_1 is a linear combination of the geographic covariates $\mathbf{X}w_1$, with normalized weight vector $w_1 \propto \mathbf{X}'Y$. That is, the weights are proportional to the simple covariances between the geographic covariates and the vector of annual mean concentrations. Stated differently, the weights are proportional to the *simple* linear regression coefficients (as opposed to *multiple* linear regression coefficients) of Y on each of the columns of $\mathbf{X}$. It follows that this vector of scores, $t_1 = \mathbf{X}w_1$ is the score of *maximum covariance* with the vector Y subject to $w_1'w_1 = 1$.

Subsequent score vectors are computed in the same simple way after replacing Y and $\mathbf{X}$ by the vector and matrix of residuals from the regressions on t_1: $Y - \hat{Y}_1 = Y - t_1 c_1$ and

$\mathbf{X} - \hat{\mathbf{X}}_1 = \mathbf{X} - t_1 p_1'$, with p_1 being the first vector of the loadings matrix **P**, proportional to the simple correlations of the geographic covariates with the first PLS score. In summary, PLS provides a decomposition of the large geographic covariate matrix $\mathbf{X}$ into a sequence of orthogonal PLS scores computed to maximize the covariance between Y and its prediction by these score vectors. If we complete the iterations to obtain all p vectors t_i, we obtain a re-expression of the covariate matrix $\mathbf{X}$ in terms of a set of orthogonal scores, but we typically stop with a small set of k PLS scores, $k \ll p$. Typically k is chosen by cross-validation to give the best predictions. We chose to compute PLS scores on the entire national covariate database in order to define them using the largest possible sample size. These definitions were held fixed across regions in models with regionally varying regression parameters. As discussed in Sections 2.4 and 2.5, we choose *k* based on 10-fold cross-validation of models that include kriging of the residuals.

2.4 Universal kriging with PLS regression and maximum likelihood estimation

The PLS computation described in Section 2.3 does not consider the fact that residuals from this regression will almost certainly be correlated in space, to an extent that will depend on the number of PLS components in the model. The complete spatial regression or universal kriging model using PLS scores can be written

$$Y = \mathbf{T}\beta + \varepsilon$$

where we change definitions slightly to let *Y* represent the uncentered vector of annual mean concentrations and correspondingly add a constant vector to the matrix of PLS scores **T**, now restricted to $k < p$ scores. The vector of regression coefficients β is therefore $(k+1)\times 1$, and we complete the model specification by assuming the errors ε are mean zero with spatial

covariance (or variogram) function depending on a parameter vector θ. Conditional on β, we can write

$$Y \sim N\left(\mathbf{T}\beta, \Sigma(\theta)\right)$$

where the parameter vector θ specifies the nugget, range, and sill of an exponential variogram model. If we identify regions by the subscript *j*, the fully regional model, with both PLS regression parameters β and covariance parameters θ varying by region, can be written

$$Y_j \sim N\left(\mathbf{T}_j\beta_j, \Sigma(\theta_j)\right),\ j = 1,2,3,$$

where $\mathbf{T}_j$ denotes the rows of $\mathbf{T}$ corresponding to the j^{th} region.

We estimate all the parameters, $(\beta_j, \theta_j),\ j = 1,2,3,$ jointly by maximizing the profile log-likelihood. In the normal log-likelihood function we replace the regression parameters by formulae for their generalized least squares estimates in terms of the covariance parameters. This resulting "profile" log-likelihood, a function only of the covariance parameters, is then maximized (using the R function `optim`).

We require expressions for the universal kriging (or generalized least squares) predictions of concentrations at unmonitored sites given observations at monitoring sites. Let Y_{mj} denote the vector of observations at monitored sites ("*m*") in region *j* and let $\mathbf{T}_{mj}$ denote the matrix of PLS component scores computed at these monitored sites. Similarly, let Y_{uj} and $\mathbf{T}_{uj}$ denote a corresponding vector and matrix for a set of unmonitored ("*u*") locations at which predictions are desired. Furthermore, let the modeled covariance matrix among the monitored sites be denoted by $\Sigma_{mm}(\theta_j)$ and the matrix of covariances between the monitored and unmonitored locations

$\Sigma_{um}(\theta_j)$. Then predictions of concentrations at unmonitored sites using estimates of the regression and covariance parameters can be written

$$\hat{Y}_{uj} = \mathbf{T}_{uj}\hat{\beta}_j + \Sigma_{um}(\hat{\theta}_j)\Sigma^{-1}_{mm}(\hat{\theta}_j)\left(Y_{mj} - \mathbf{T}_{mj}\hat{\beta}_j\right) .$$

2.5 Regional analysis strategy and cross-validation

For the modeling framework described in Section 2.4, we need to determine how best to divide the country into regions, how many (nationally defined) PLS components to use in the spatial regressions (universal kriging), and whether either or both of the two model components, the PLS regression model and/or the variogram model, are best defined on a regional or national basis.

We divided the country into regions in order to account for possible differences across the continental U.S. in the mean regression structure of $PM_{2.5}$ as well as the residual smoothness (hence the variogram) of spatial variation. Consideration of the PLS and variogram characteristics of preliminary models led to a final choice of three regions based on an assessment of topology and elevation. These three regions, called East, Mountain West, and West Coast, are illustrated in Figure 1. Although one could argue that the large eastern region should be further subdivided, diagnostic statistics suggest that a single universal kriging model suffices for this region.

The decisions about the numbers of PLS components to retain, and whether to estimate PLS regression coefficients and covariance coefficients regionally or nationally were based on 10-fold cross-validations. Monitoring sites in each of the three regions were randomly assigned to one of ten groups. In turn, each group was set aside as a "test set" and the remaining groups combined for a "training set" to fit the model and generate test set predictions using the universal

kriging prediction equation given above. Each group played the role of test set until predictions were obtained for the entire data set. We assessed the performance of each fitted model based on their cross-validated root mean squared prediction error, corresponding R^2, and the width and accuracy (percent coverage) of 95% predictive intervals.

3. Results

Table 2 gives descriptive statistics for year 2000 annual average $PM_{2.5}$ concentrations at the 901 monitoring sites shown in Figure 1. The eastern 2/3 of the country clearly demonstrates the highest average concentration across sites while the West Coast, with the second highest average concentration, shows much greater spatial variability.

Table 2. Descriptive statistics for annual mean $PM_{2.5}$ concentrations in 2000 on the native ($\mu g/m^3$) and square root scales for the regions illustrated in Figure 1.

Region (n)	($\mu g/m^3$)		Square root	
	Mean	SD	Mean	SD
National (901)	12.73	4.06	3.51	0.62
East (673)	13.72	3.10	3.68	0.45
Mt. West (120)	7.85	3.17	2.74	0.59
West Coast (108)	11.93	5.79	3.35	0.83

Models were fit to annual mean $PM_{2.5}$ concentrations on a square root scale since diagnostic analyses of residuals from fitted models suggested that the assumptions for the normal likelihood model were reasonably satisfied on this scale. Table 3 and Figures 2-4 provide summaries of the predictive quality of the models based on 10-fold cross-validation. Table 3 presents cross-validated R^2 values for various choices of the number of PLS components. The cross-validated

R^2 is computed as 1-RMSEP^2/Var(Obs) where RMSEP represents the root mean-squared error of the predictions and Var(Obs) is the variance of the observations, both on the transformed scale. While the details of the predictions vary depending on whether the PLS regression coefficients and variogram models are specified nationally or regionally, the cross-validated R^2 values change little. All the models show quite good performance nationally with R^2 values mostly exceeding 0.86 and surprisingly little sensitivity to the number of PLS components. Figure 2 provides the scatterplots underlying the R^2 values for the 2-component model in units of square root $\mu g/m^3$. The regional coefficient/regional variogram model gives not only the highest national R^2 of 0.88, but also the highest within-region R^2 values, which are 0.82 West Coast (red dots), 0.64 Mountain West (green dots), and 0.90 East (blue dots).

Table 3. 10-fold cross-validation R^2 statistics for the models considered. The values in the first row of this table for 2 PLS component models are represented in the scatterplots in Figure 2.

Coefficients:		National	National	Regional	Regional
Variogram:		National	Regional	National	Regional
Number of PLS	2	0.85	0.84	0.87	0.88
Components	3	0.86	0.86	0.88	0.88
	4	0.86	0.86	0.88	0.88
	5	0.86	0.86	0.88	0.87

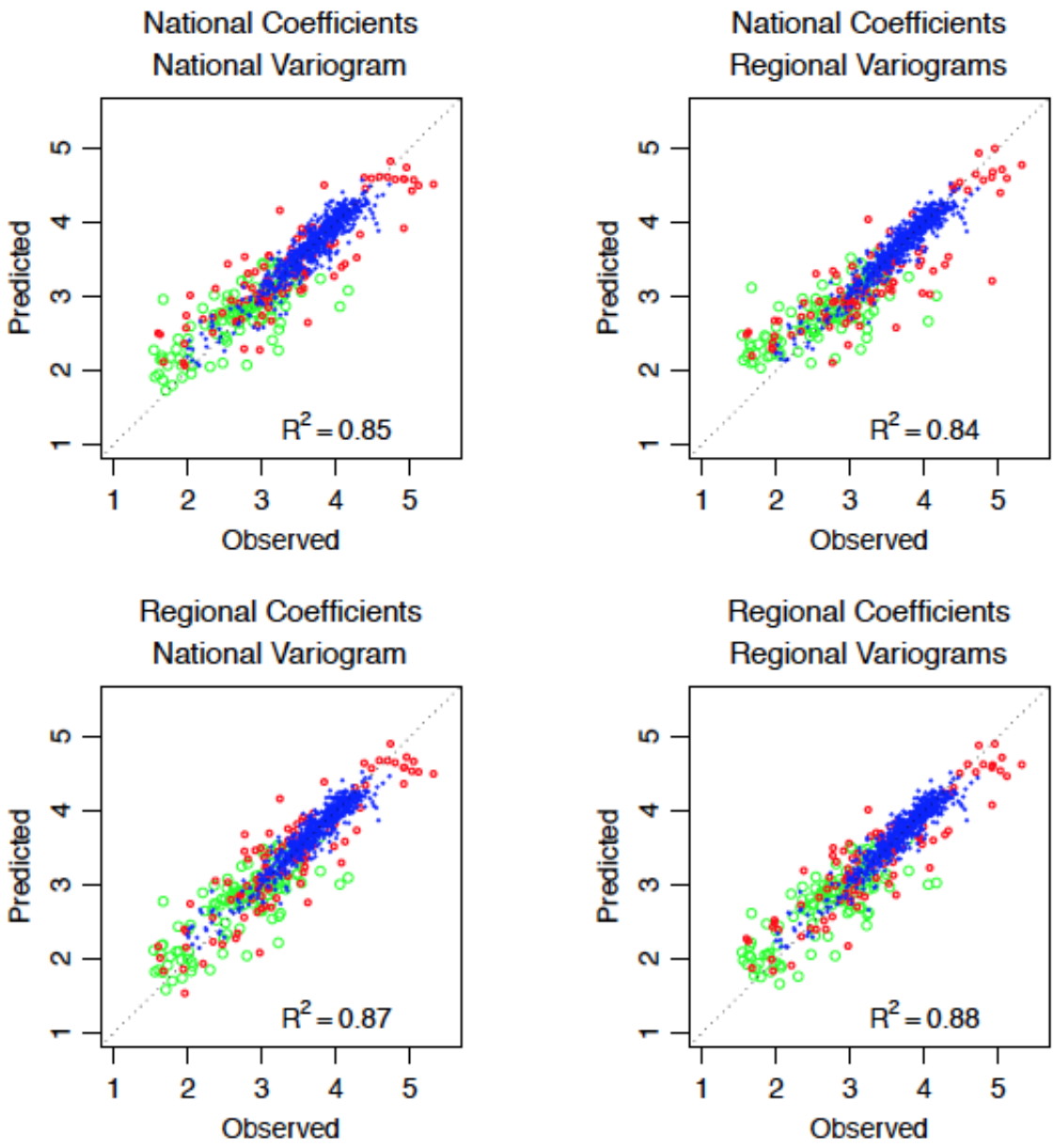

Fig. 2. Scatterplots of observed PM2.5 concentrations and cross-validation predictions (square root scale) for 2 PLS component universal kriging models by regional vs. national model specification. Points are colored coded by region (as in Figure 1): East: blue, Mountain West: green, West Coast: red.

Figure 3 presents boxplots showing distributions of widths of cross-validation-based 95% prediction intervals at monitoring sites. We also report the coverage of these conventional 95% predictive intervals from the cross-validation. We see that coverage is close to the nominal 95%

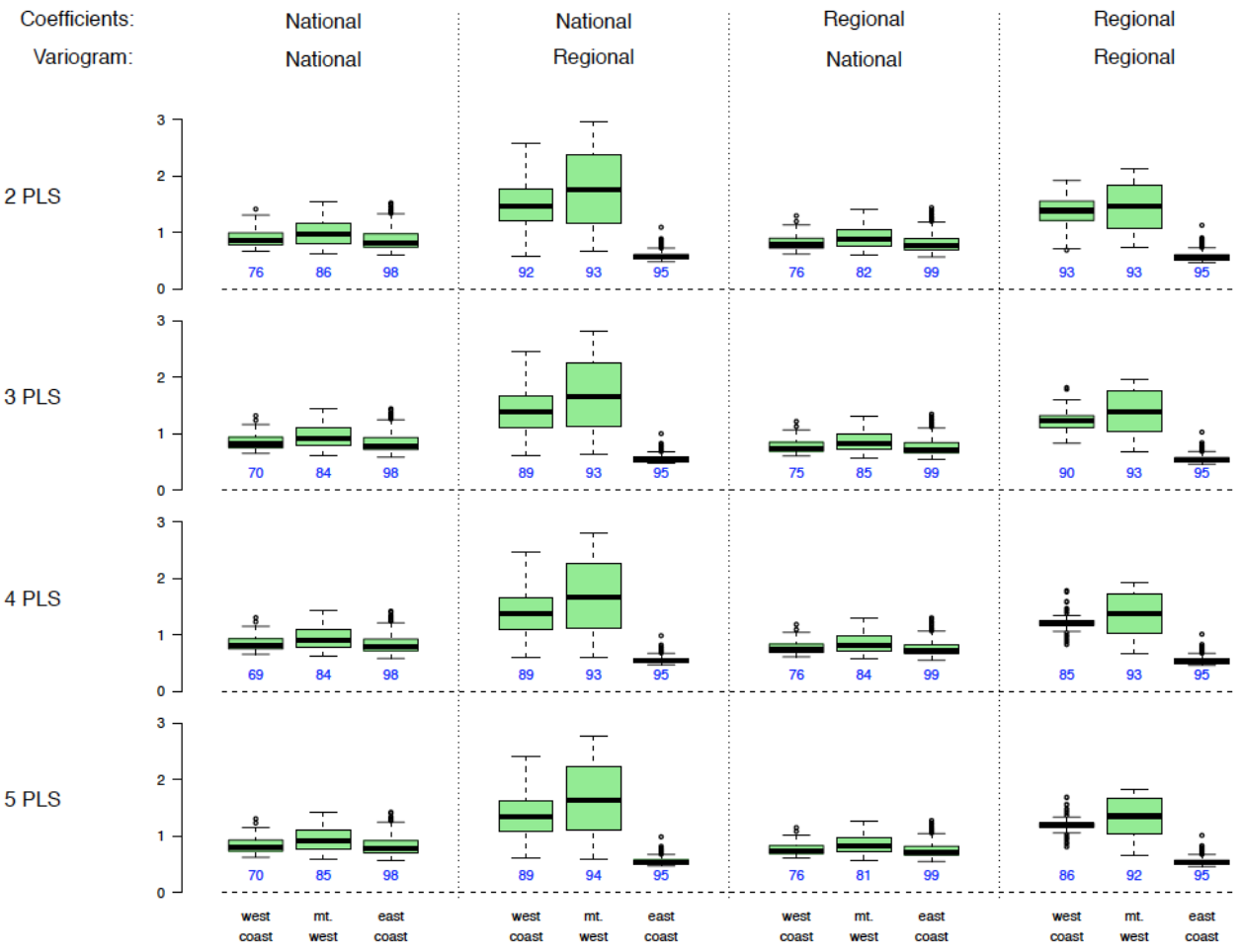

Fig. 3. Boxplots showing distributions of widths of cross-validation 95% prediction intervals at monitoring sites. Coverage percentages of the confidence intervals are printed below the boxplots.

level across all three regions only in the case of models with variograms varying by region. The fully regional model with two PLS components provides the highest R^2 values (to two decimal places) and the narrowest prediction intervals while achieving coverage that is closest to the 95% target for all three regions. We select this regional coefficient/regional variogram model with two PLS components as the primary model for further discussion of results.

Figure 4 depicts the magnitudes of cross-validation prediction errors. There is a greater range of prediction errors, both positive and negative, in the West Coast and Mountain West regions, but the predictions are approximately unbiased, on average, within each region.

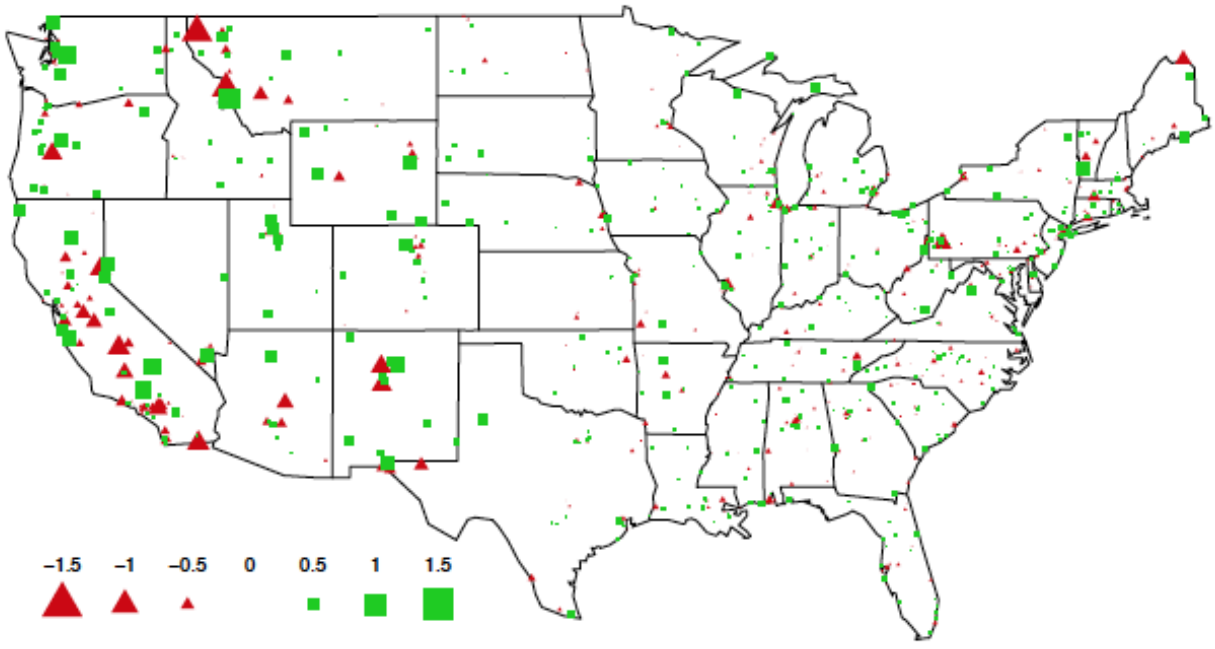

Fig. 4. Graphical depiction of positive (green) and negative (red) cross-validation prediction errors (in sqrt(ug/m3)) for the model with regionally estimated PLS regression coefficients and regional variogram models.

(Predictions in individual regions were not unbiased when either the PLS regression or the variogram model was specified on a national scale.)

Figure 5 presents a graphical depiction of the elements of the 171×2 matrix of loadings (**P)** which provides insight into the two PLS component scores. The first and most important component is dominated by high positive loadings on the "development" land use scores, lengths of roads, and amounts of impervious surface in contrast to high negative loadings on "natural" land use features like shrubs, grass and evergreen, along with negative loadings on distances to roads and pollutant sources. This score is an interpretable composite measure of development (or urbanization) positively correlated with $PM_{2.5}$ concentrations. The second PLS component, constrained to be orthogonal to the first component, clearly contrasts the NDVI vegetative index scores with some of the developmental measures, notably impervious surfaces and "high development" land use. While this score is easy to summarize conceptually, it has less straightforward interpretation since its contribution to the prediction of $PM_{2.5}$ concentrations varies in sign across regions.

Table 4 presents a summary of the parameter estimates from the likelihood fitting of the model. The coefficients of the PLS component scores show much larger coefficients for the dominant first PLS component in the Mountain West and West Coast in contrast to the East. The coefficient of the second PLS component, the contrast between NDVI and developmental measures, is clearly significant only in the East where it has a negative coefficient.

Figure 6 shows region-specific regional variogram estimates and Table 4 gives their associated parameter estimates. We used the residuals from the PLS regression part of the fitted

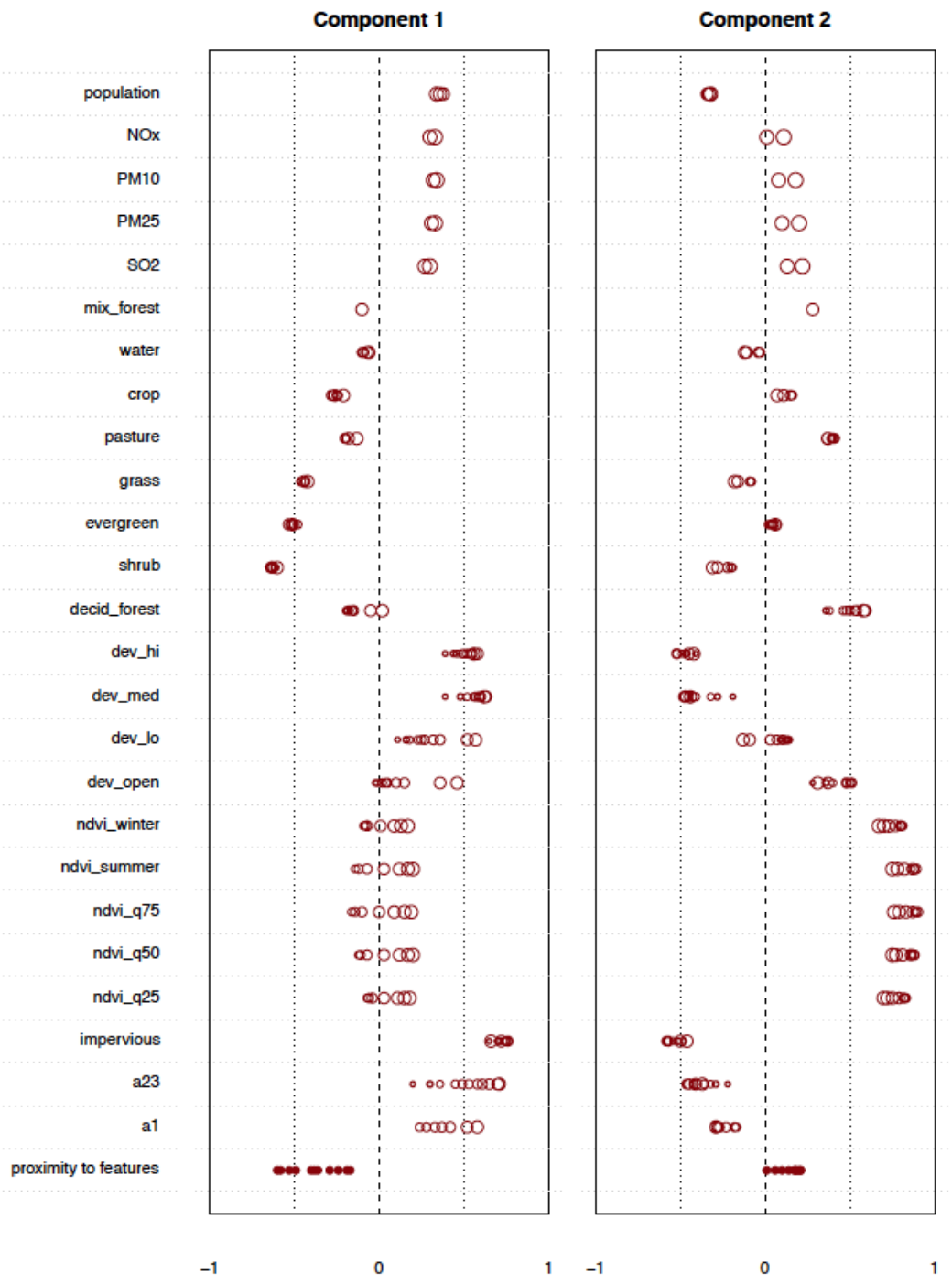

Fig. 5. Two PLS components characterized by the loadings of the 171 covariates of Table 1 on the component scores. These loadings have been scaled as correlations with the component score. Sets of circles with increasing radii denote a particular measure (such as sum of A1 highway road lengths, "a1") evaluated in buffers of increasing size, as specified in Table 1.

model to compute these empirical variograms. These results indicate there is much stronger residual spatial correlation structure in the East in contrast to the Mountain West and West Coast where the ranges are short and the nuggets higher. Thus geostatistical smoothing is contributing much more strongly to predictions in the East.

Table 4. Maximum likelihood parameter estimates and standard errors for the universal kriging model. (Standard errors, computed from the hessian of the full likelihood, are provided in parentheses only for the regression coefficients.) Regression coefficients B1 and B2 multiply the PLS component scores depicted by their loadings in Figure 5. Variograms corresponding to the parameters here are illustrated in Figure 6.

Region	Coefficients			Variogram Parameters		
	Intercept	B1	B2	Range	Sill	Nugget
East	2.604 (0.520)	0.024 (0.002)	-0.013 (0.002)	2944	0.433	0.013
Mountain West	3.069 (0.102)	0.062 (0.006)	0.014 (0.011)	35	0.156	0.014
West Coast	3.263 (0.128)	0.086 (0.009)	-0.016 (0.009)	86	0.254	0.018

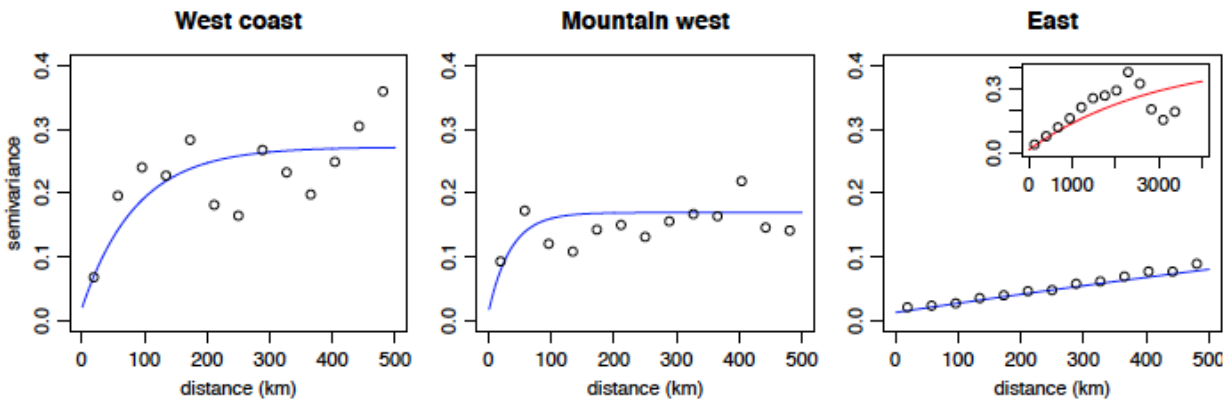

Fig. 6. Regional exponential variogram model fits estimated by maximum likelihood for 2 PLS component models. The three main panels are all drawn to the same distance of 500 km. Because of the much greater range of the variogram for the east, the third panel includes an inset illustrating the variogram drawn to a distance of 4000 km.

Finally, Figure 7 presents an image map of predicted $PM_{2.5}$ concentration evaluated on a regular 25 km grid across the United States and smoothed for display purposes. The inset illustrates the smaller spatial scale structure of the predictions using the example of the southern California region around Los Angeles. There are some clear features to note at this 25 km scale. The West and Mountain West regions are generally lower in concentration, on average, but with great variability and pockets of the highest concentrations located in urban areas, especially notable around Los Angeles. The eastern part of the country includes a broad region of higher concentrations generally extending east of the Mississippi with the exception of southern Florida, northern Wisconsin and Michigan, and northern New England.

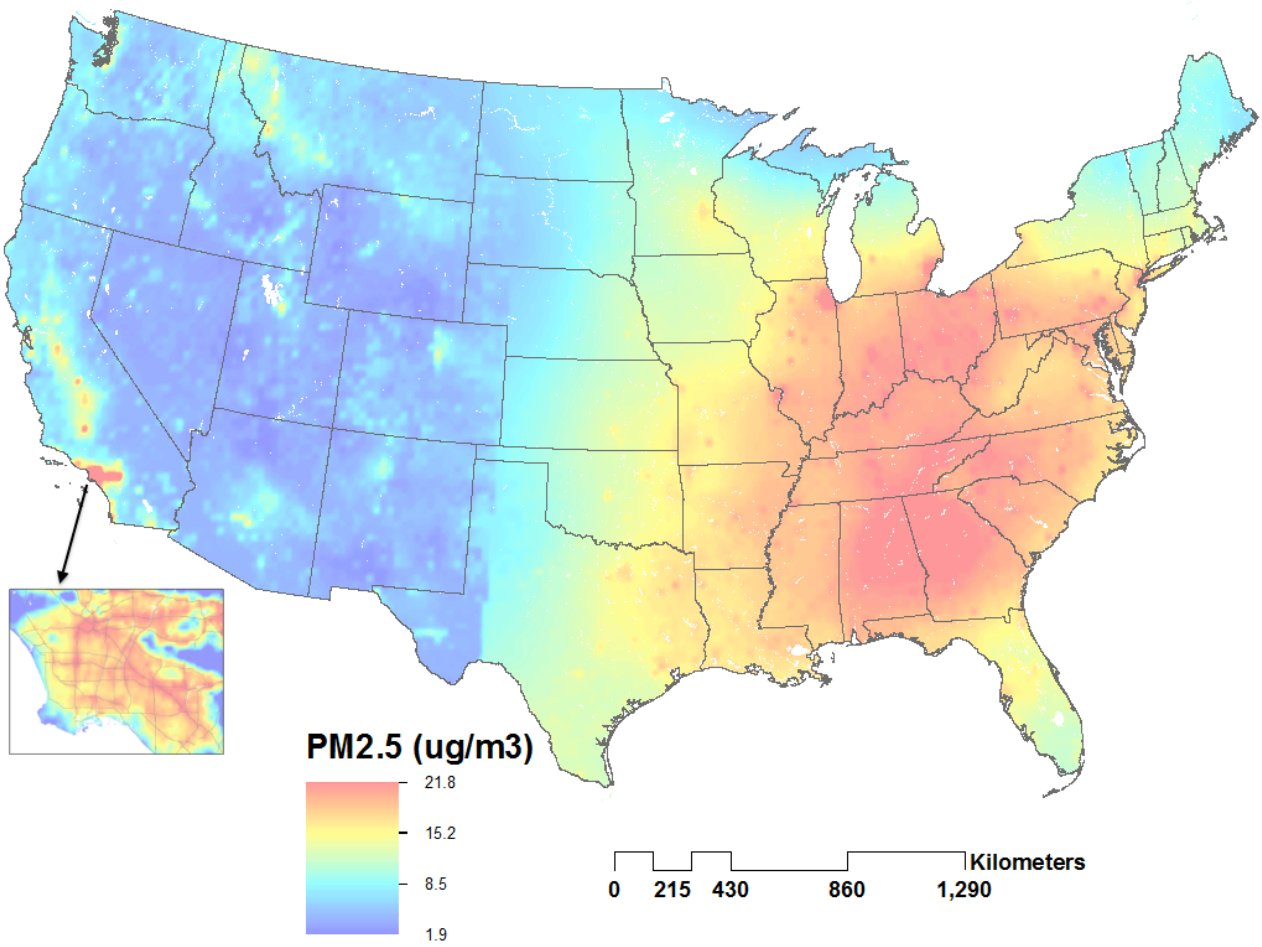

Fig. 7. Image plot of $PM_{2.5}$ concentration predictions over the entire US with an inset image of the Los Angeles area to demonstrate some of the fine scale spatial structure of the predictions. The national map was derived from predictions computed at points on a regular 25 km grid and the high resolution Los Angeles inset from predictions on nested 1 km and 0.5 km grids. The raster images were computed in ArcGIS with inverse distance weighting using the five nearest grid points.

4. Discussion

The regionalized national universal kriging model with Partial Least Squares regression presented here provides a number of important practical and methodological results. First it yields predictions of annual average $PM_{2.5}$ concentrations with good predictive accuracy (cross-validated R^2 value of 0.88 for our selected fully regional model; see Figure 2) and well-calibrated predictive intervals (Figure 3). Second, it has demonstrated the usefulness of PLS to simplify dimension reduction in comparison to other approaches to variables selection (e.g., Su et al., 2009; Mercer et al., 2011). In addition, our regionalized strategy to universal kriging has proved to be valuable in addressing aspects of large-scale nonstationarity.

Impressive as the national R^2 statistic may be, it can hide regionally varying biases and inaccuracies. We show that absolute prediction errors are substantially larger in the West and Mountain West regions, with corresponding larger 95% prediction intervals (Figures 3, 4). Fortunately, our modeling accurately represents this heterogeneity in predictive intervals that are well-calibrated in the sense of achieving near nominal 95% coverage. The greater errors are to be expected due to the greater variation in $PM_{2.5}$ concentrations in the western part of the country and the greater spatial variability (lower spatial correlation) reflected in the variogram models for the West Coast and Mountain West (Figure 6). The smoother spatial structure in the eastern two-thirds of the country leads to more accurate predictions mostly driven by the kriging component of the universal kriging estimates.

Our regional modeling for the spatially varying structure described above results in discontinuity in the predictions at the boundaries between regions. These discontinuities are relatively minor except in the transition from East to Mountain West due to a near absence of monitoring sites in the region from the southern border of the U.S. in the midwestern region of

Texas, north through the Texas panhandle, eastern Colorado, western Kansas, and southern Nebraska. There are little data to validate the relatively sharp transition in predicted $PM_{2.5}$ concentrations illustrated in Figure 7. (Fortunately, this area is relatively sparsely populated and there are few target subjects in any of the epidemiologic studies of interest.)

The PLS regression component of the modeling strategy obviates the more common approaches to variable selection in land use regression modeling. While there is no guarantee that the PLS approach will yield better predictions than a variable selection approach, we find it to be a convenient and scientifically attractive way to synthesize the predictive value of a very large number of highly correlated GIS-based covariates. Cross-validatory choice of the number of PLS components in the universal kriging model is very important since without the kriging component many more PLS components would be selected by cross-validation. We were able to achieve the best calibrated model (defined in terms of the coverage of prediction intervals) with only two PLS components.

Realistic spatial (and spatio-temporal) models on a national scale must be nonstationary in the sense that both of the components of the universal kriging approach, the spatial regression model and the residual spatial correlation structure, almost certainly vary regionally. In the case of the regression model, one might argue that features like roads and traffic should correlate with or predict pollutant levels similarly across the country, but even here, different vehicle mixes, vehicle speeds, road surfaces and meteorology can influence the details of these spatial predictions. Furthermore, certain types of covariates are (more) relevant in some parts of the country than others, as is the scale and extent of secondary pollutant formation. For example, in the east, there is the well-known phenomena of secondary particle formation from oxidation of sulfur and nitrogen oxides emitted from tall stacks by coal fired-power plants. These elevated

emissions result in a regional-scale particulate 'haze' rich in sulfate and nitrates that comprise major fraction of the fine particle mass (Malm et.al., 2004; Tesche et.al., 2006; Hand et.al., 2012). In contrast, the organic carbon fraction of the $PM_{2.5}$ is higher in the Western U.S. and is more variable over the region (Malm et.al. 2004; Hand et.al., 2012).

Our regional universal kriging models used simple stationary and isotropic spatial covariance or variogram models. It would certainly be attractive to consider nonstationary models for individual regions (Sampson, 2012), and it is possible that nonstationary covariance models would help, especially in the complex western regions of the U.S. However, results with the naïve stationary models are reasonably well-calibrated. We have obtained similarly accurate results in applications to other annual averages and other pollutants of interest, including NO_2.

The ultimate objective of the prediction model described here is to provide exposure predictions for epidemiologic analysis of health effects of long-term exposure to air pollutants, estimated by average annual exposure. Our spatial model provides very accurate predictions, but there will still be differences between the true and predicted exposures for study subjects, resulting in covariate measurement error that can bias health effect estimates and standard errors (Kim et al. 2009). In fact, more accurate exposure predictions do not necessarily result in the best health effect estimates, depending on exposure assessment study design and components of the exposure estimation errors which can lead to Berkson-like and classical-like errors in health effect estimates (Szpiro et al., 2011b). Since we employ likelihood-based methods to fit our universal kriging model, recently published computationally efficient bootstrap methods are available to correct for the measurement error and give valid health effect confidence intervals (Szpiro et al. 2011a, Bergen et al. 2012).

Acknowledgements

Although the research described in this presentation has been funded in part by the United States Environmental Protection Agency through grant R831697, 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), the NIEHS DISCOVER Center (P50 ES015915) and the Biostatistics, Epidemiologic & Bioinformatic Training in Environmental Health Training Grant (ES015459).

References

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

Bergen, S., Sheppard, L., Sampson, P.D., Kim S.-Y., Richards, M., Vedal, S., Kaufman, J.D., and Szpiro, A.A. 2012. A national prediction model for components of $PM_{2.5}$ and measurement error corrected health effect inference. Submitted.

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

Hand, J.L., Schichtel, B.A., Pitchford, M., Malm, W.C., and Frank, N.H. 2012. Seasonal composition of remote and urban fine particulate matter in the United States. Journal of Geophysical Research, 117, D05209, doi:10.1029/2011JD017122

Hart, J.E., Yanosky, J.D., Puett, R.C., Ryan, L., Dockery, D.W., Smith, T.J., Garshick, E., Laden, F. 2009. Spatial modeling of PM_{10} and NO_2 in the continental United States, 1985-2000. Environmental Health Perspectives 117, 1690-1696.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed.

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.

Hystad, P., Setton, E., Cervantes, A., Poplawski, K., Deschenes, S., Brauer, M., van Donkelaar, A., Lamsal, L., Martin, R., Jerrett, M., Demers, P. 2011. Creating national air pollution models for population exposure assessment in Canada. Environmental Health Perspectives 119, 1123-1129.

Kim, S.-Y., Sheppard, L., and Kim, H. 2009. Health effects of long-term air pollution: influrnce of exposure prediction models. Epidemiology 20, 442-450.

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.

Malm, W.C., Schichtel, B.A., Pitchford, M.L., Asbaugh, L.L., and Eldred, R.A. 2004. Spatial and monthly trends in speciated fine particle concentration in the United States. Journal of Geophysical Research, 109, D03306, doi:10.1029/2003JD003739

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

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.

Novotny, E.V., Bechle, M.J., Millet, D.B., Marshall, J.D. 2011. National satellite-based land-use regression: NO_2 in the United States. Environmental Science and Technology 45, 4407-4414.

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.

Sampson, P.D., Szpiro, A.A., Sheppard, L., Lindstrom, J. Kaufman, J.D. 2011. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. Atmospheric Environment 45, 6593-6606.

Sampson, P.D. 2012. Spatial covariance, in Encyclopedia of Environmetrics, 2nd ed, pp. ??-??.

Su, J.G., Jerrett, M., Beckerman, B., Wilhelm, M., Ghosh, J.K., Ritz, B., 2009. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. Environmental Research 109, 657-670.

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., and Lumley, T., 2011a. Efficient measurement error correction for spatially misaligned data. Biostatistics, 12:610-623.

Szpiro, A.A., Paciorek, C.J., and Sheppard, L., 2011b. Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology, 22:680-685.

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

Tesche, T.W., Morris, R., Tonnesen, G., McNally, D., Boylan, J., and Brewer, P., 2006. CMAQ/CAMx annual 2002 performance evaluation over the eastern U.S.. Atmospheric Environment 40 (2006) 4906-4919.

Wackernagel, H., 2010. Multivariate Geostatistics: An Introduction with Applications, 3rd ed. Springer-Verlag, Berlin, Germany.

Wold, S., Sjostrom, M., Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58, 109-130.

Yanosky, J.D., Paciorek, C.J., Suh, H.H., 2009. Predicting chronic fine and coarse particulate exposures using spatio-temporal models for the Northeastern and Midwestern United States. Environmental Health Perspectives 117, 522-529.