



UW Biostatistics Working Paper Series

11-30-2009

Pragmatic Estimation of a Spatio-Temporal Air Quality Model With Irregular Monitoring Data

Paul D. Sampson

University of Washington - Seattle Campus, pds@stat.washington.edu

Adam A. Szpiro

University of Washington, aszpiro@u.washington.edu

Lianne Sheppard

University of Washington, sheppard@u.washington.edu

Johan Lindström

Lund University, johanl@maths.lth.se

Joel D. Kaufman

University of Washington, joelk@u.washington.edu

Suggested Citation

Sampson, Paul D.; Szpiro, Adam A.; Sheppard, Lianne; Lindström, Johan; and Kaufman, Joel D., "Pragmatic Estimation of a Spatio-Temporal Air Quality Model With Irregular Monitoring Data" (November 2009). *UW Biostatistics Working Paper Series*. Working Paper 353.

<http://biostats.bepress.com/uwbiostat/paper353>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

1 **PRAGMATIC ESTIMATION OF A SPATIO-TEMPORAL AIR QUALITY MODEL**
2 **WITH IRREGULAR MONITORING DATA**

3
4 Paul D. Sampson^{1*}

5 Adam A. Szpiro²

6 Lianne Sheppard²

7 Johan Lindström^{1,3}

8 Joel D. Kaufman⁴

9
10 ¹ *Department of Statistics, University of Washington, U.S.A.*

11 ² *Department of Biostatistics, University of Washington, U.S.A.*

12 ³ *Centre for Mathematical Sciences, Lund University, Sweden*

13 ⁴ *Department of Environmental and Occupational Health Sciences, University of Washington,*
14 *U.S.A.*

15
16
17
18
19
20 Correspondence to: Paul D. Sampson, Department of Statistics, University of Washington, Box 354322, Seattle,
21 WA 98195-4322, U.S.A. E-mail: pds@stat.washington.edu

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

1 **1. Introduction**

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

11 The specification of a modeling approach depends on a number of factors that vary with the
12 details and scientific purpose of the study in which the predicted exposures will be computed.

13 Among these are:

- 14 • The spatial resolution and scale of the monitoring data, including the number and spacing of
15 monitors and the spatial extent of the modeling domain. Different strategies may be
16 appropriate depending on whether one is addressing an urban area, “mesoscale” regions, or
17 larger scale regions such as the United States east of the Mississippi River.
- 18 • The temporal scales of monitoring data, with hourly, daily, 2-week, and monthly scales
19 appropriate in different situations.
- 20 • The characteristics of the monitoring devices, which may vary for data from multiple
21 monitoring networks.
- 22 • The characteristics of relevant spatio-temporal covariates, including their spatial and
23 temporal scales.

24 Our work has been developed for the Multi-Ethnic Study of Atherosclerosis and Air
25 Pollution (MESA Air). MESA Air is a cohort study funded by the U.S. Environmental Protection
26 Agency (EPA) that emphasizes accurate prediction of intra-urban variation in individual

1 exposures to ambient air pollutants in order to accomplish its primary aim of assessing the
2 relationship between chronic exposure to air pollution and sub-clinical cardiovascular disease.
3 The MESA Air cohort includes more than 6000 male and female subjects in six major U.S.
4 metropolitan areas (Los Angeles, CA; New York, NY; Chicago, IL; Minneapolis-St. Paul, MN;
5 Winston-Salem, NC; and Baltimore, MD). The primary MESA Air hypotheses relate to chronic
6 exposure to $PM_{2.5}$. We note that final exposure estimates in MESA Air will integrate predictions
7 of outdoor concentrations with additional subject-level data, including time-activity patterns,
8 home infiltration characteristics, address history, and employment address (Cohen et al., 2009).

9 Our aim here is to provide point spatial predictions of long-term average concentrations at
10 the residences of MESA Air subjects. A primary source of monitoring data is the EPA's
11 regulatory Air Quality System (AQS) repository (EPA 2010). The AQS network includes a
12 number of fixed site monitors in each region, each of which measures ambient air pollution
13 levels on a regular basis, either daily, every third day or every sixth day in the case of $PM_{2.5}$.
14 Although there are some missing data, most AQS sites provide nearly complete $PM_{2.5}$
15 concentration time series over several years at their spatial locations. The analysis here uses data
16 for the time period 2000-2006.

17 The MESA Air study conducted supplemental air quality monitoring campaigns beginning in
18 2005 to provide additional concentration data in order to better model and predict intra-urban
19 variation in air quality not well-represented by the AQS regulatory monitoring network. The
20 objective of the MESA Air monitoring is to more completely sample a spatial design space that
21 emphasizes traffic-related pollution at small spatial scales (100s of meters) and to capture data at
22 actual subject home locations. For logistical reasons, the supplementary monitoring data are
23 sampled as two-week averages based on an unbalanced design that results in significant amounts
24 of missing data at many measurement locations (Cohen et al. 2009). In each region, the
25 monitoring through 2006 includes: (a) from two to five fixed site monitors providing up to 1.5
26 years of 2-week observations, and (b) rotating sets of 2-week observations at (typically) 4 subject
27 homes, with monitors moved every two weeks to cover a total of about 50 subject homes, each

1 observed twice. MESA Air monitoring at fixed and home sites continued through July, 2009,
2 providing a larger for future modeling. The composite of the AQS and MESA Air monitoring
3 data provide a rich but highly irregular spatio-temporal monitoring database for analysis.

4 Statistical approaches to the modeling of spatio-temporal air quality monitoring data for
5 ambient exposure estimation typically derive from models that decompose observations into
6 spatio-temporal *trend* and spatio-temporal *residuals*, or variation around the trend. Even after
7 specification of the scientific questions and the spatio-temporal scales of interest, different
8 statisticians will hold different perspectives on how-much spatial and temporal structure is to be
9 represented as trend and how much as residual. For example, considering just a single time
10 series of air quality data one may explicitly model a trend using iterated moving averages as in
11 the KZ filter (Rao and Zurbenko, 1994; Wise and Comrie, 2005), various state-space modeling
12 approaches (e.g., Taylor et al. 2006), spline smoothing, or simple parametric models such as
13 polynomial or trigonometric series. Alternatively, one can incorporate much of the variation in
14 the residuals by using a sufficiently rich autocorrelation model. Somewhat similar choices arise
15 in modeling spatial trends and residuals, with different perspectives provided by splines that
16 model the spatially correlated component semi-parametrically as part of the mean model and
17 geostatistical approaches that account for spatial structure not predicted by covariates using a
18 variogram model for the autocorrelation. However, there are mathematical frameworks that
19 allow one to show equivalence between certain spline and kriging predictions (e.g., Kent and
20 Mardia, 1994; Furrer and Nychka, 2007).

21 A number of characteristics of the variation in air-quality data across the urban scales of
22 interest in our current work motivate the modeling strategy and the approach to trend and
23 residual that we present. We find that all air quality measurements demonstrate systematic time
24 trends and seasonality, but these time trends vary in space, even over relatively small
25 metropolitan area spatial scales. The details of the trend and seasonality vary somewhat from
26 year to year, so that they are not modeled well by simple periodic functions like sinusoids or
27 other trigonometric functions. Finally, characteristics of these systematic time trends, including

1 the long-term mean and amplitude of seasonal variations, covary with spatial (or “land use”)
2 covariates. Even though the primary interest in MESA Air is predicting spatial variation of long-
3 term average concentrations to estimate exposures, the complex spatio-temporal monitoring
4 design necessitates a statistical modeling approach that accounts for spatio-temporal interactions
5 in the data. Otherwise we cannot use the spatially rich but temporally sparse data at MESA Air
6 homes to help in estimating the long-term averages.

7 For an overview of general techniques for modeling correlated spatio-temporal data, see
8 Banerjee et al. (2004). See also the closely related strategy of Le and Zidek (2006). Smith et al.
9 (2003) use an expectation-maximization (EM) algorithm to allow for arbitrary missing data
10 patterns, but their model does not accommodate the complex spatio-temporal interactions
11 addressed here. A recent paper by Fanshawe et al. (2008) demonstrates how carefully chosen
12 covariates may eliminate the need to accommodate spatio-temporal correlation in the residuals,
13 but the model in that paper assumes a uniform time trend across a relatively small spatial area.
14 Paciorek et al. (2009) and Sahu et al. (2006) model particulate matter using techniques that allow
15 for more complex spatio-temporal interactions, however their estimation and prediction
16 procedures (based on a different approach to representing spatio-temporal trend and residual) are
17 applicable only with a nearly complete (balanced) space-time monitoring data matrix.

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

1 In Szpiro et al. (2009) we describe a likelihood-based version of the hierarchical model
2 presented below and study its statistical properties by applying it to oxides of nitrogen (NO_x) in a
3 simulation scenario based on a subset of the MESA Air geographic covariates, subject locations,
4 and monitoring data. Lindström et al. (2010) extend this modeling to incorporate predictions
5 from a source dispersion model as a spatio-temporal covariate, also for analyses of NO_x . The
6 present paper focuses on a multi-step pragmatic approach to estimation and utilizes a more
7 complete set of covariates and monitoring data to predict $\text{PM}_{2.5}$ concentrations at the homes of
8 all MESA Air subjects. We use the Los Angeles study area as an example to illustrate the data
9 structure and modeling approach. In order to make predictions at all MESA Air subject homes,
10 we fit the model separately in four geographic regions covering the six MESA Air study areas:
11 Southern California, a Midwest region spanning Minneapolis-St Paul and Chicago, a Northeast
12 region spanning Baltimore and New York City, and North Carolina.

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

20 **2. Data**

21 *2.1 AQS and MESA Air Monitoring data*

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

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

5 --- Figure 1 ---

6 There are a total of seven MESA Air fixed sites in the Los Angeles area, one of which is co-
7 located with an AQS monitor to allow for instrument calibration. These fixed sites began
8 measuring two-week average concentrations in November 2005. There were approximately 40
9 measurements per site and a total of 264 fixed site measurements during this timeframe. A total
10 of 45 home outdoor monitoring locations in Los Angeles are also included; these were sampled
11 during two-week periods starting in May 2006. The plan called for each home to be sampled two
12 times, in different seasons. (Not all homes were sampled twice before the end of 2006, the
13 closing date for the database for the analysis in this manuscript.) Figure 2 presents a schematic
14 illustration of the spatio-temporal sampling scheme combining the various AQS (EPA) and
15 MESA Air monitoring sites.

16 --- Figure 2. ---

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

1 every 3rd day). Monitoring sites sampling only every 6th day did not provide enough data to
2 estimate valid 2-week averages.

3 --- Figure 3. ---

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

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

18 2.2 *GIS-based geographical covariates*

19 Our strategy for predicting concentrations at locations and times without measurements
20 includes the use of regression models with geographic covariates. This is often termed “land use”
21 regression (LUR) (Moore et al. 2007; Ross et al. 2007; Hoek et al. 2008). Our application of
22 LUR is embedded in a hierarchical spatio-temporal model that incorporates flexible correlation
23 structures. We consider a variety of geographic covariates, including: (i) indirect measures of
24 traffic influences provided by distances to major roads (major roads identified by census feature
25 class codes A1-A3), together with lengths of such roads in seven circular buffers from 50 to 750
26 meters around sites of interest, (ii) average population density (number of people per square km

1 in the block group where the monitor or participant is located), and (iii) percentages of land in
2 circular buffers described by various land use categories such as commercial property, cropland,
3 industrial property, and residential property. These are all derived using the ArcGIS (ESRI,
4 Redlands, CA) software package. The population density is calculated from publicly available
5 U.S. Census Bureau data, and the roadway variables are derived from the proprietary TeleAtlas
6 Dynamap 2000 roadway network. In total we considered approximately 200 possible covariates
7 accounting for the road and land use variables measured in seven nested circular buffers.
8 Covariates were screened prior to analysis and those with essentially no variability in a given
9 study region (e.g. percent of forest in Los Angeles) were omitted. The number of covariates
10 remaining for analysis ranged from 41 for Southern California to 66 for the Northeast region
11 (including geographic coordinates derived from latitude and longitude).

12 **3. Components of the space-time hierarchical model**

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

$$16 \quad Y_{s,t} = \mu(s,t) + \varepsilon(s,t)$$

17 where s indexes space and t time, μ is the trend surface, and ε is the residual surface. Our
18 approach accounts for spatial variability in temporal trends in order to make use of the complex
19 spatio-temporal monitoring data from the combined AQS and MESA Air monitoring campaigns
20 as well as a possibly nonstationary spatial covariance structure in the residuals. Our proposed
21 model includes sufficiently complex spatio-temporal interactions to account for both variations
22 in seasonal patterns at different locations and changes over time in the configuration of sites
23 available for spatial predictions. It accommodates the fact that air quality measurements
24 demonstrate systematic time trends and seasonality, that these time trends vary in space, even
25 over relatively small metropolitan area spatial scales, and that the details of the trend and

1 seasonality vary somewhat from year to year (and, hence, are not modeled well by simple
 2 sinusoids).

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

$$5 \quad \mu(s, t) = \beta_{0s} + \sum_{j=1}^m \beta_{js} f_j(t) \quad (1)$$

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

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

$$14 \quad \boldsymbol{\beta}_j = \begin{pmatrix} \beta_{js_1} \\ \vdots \\ \beta_{js_N} \end{pmatrix} \sim N(\mathbf{X}\boldsymbol{\alpha}_j, \Sigma_e(\phi_j)) = \sum_{k=1}^q X_k \alpha_{jk} + \mathbf{e}_j \quad (2)$$

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

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

$$21 \quad \boldsymbol{\varepsilon}_t = \begin{pmatrix} \varepsilon_{s_1, t} \\ \vdots \\ \varepsilon_{s_N, t} \end{pmatrix} \sim N(0, \Sigma_\varepsilon(\phi_\varepsilon)) \quad (3)$$

22 where the spatial covariance matrix $\Sigma_\varepsilon(\phi_\varepsilon)$, with spatial correlation parameters ϕ_ε , is computed
 23 using the Sampson-Guttorp spatial-deformation model for nonstationary spatial covariance
 24 (Damian et al., 2003). The applicability of this model is based on the assumption that the

1 meteorological events that can drive high temporal autocorrelation structure on a daily time scale
 2 are largely averaged out on the 2-week time scale of the MESA Air supplemental monitoring.
 3 Our aim is to separate the spatially varying temporal trends from the residuals, thus leaving the
 4 residuals from the trend essentially uncorrelated in time.

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

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

13 *4.1 Smoothed Empirical Orthogonal basis Functions*

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

$$20 \quad \mathbf{Y} = \mathbf{M} + \mathbf{E} \quad (4)$$

21 where

$$22 \quad \mathbf{M} = \mathbf{F}\boldsymbol{\beta}, \quad (5)$$

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

$$25 \quad \mathbf{F} = [f_0(t) \ f_1(t) \ \cdots \ f_m(t)]$$

1

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

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

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}'. \quad (6)$$

6 That is, using the superscript (m) to denote sub-matrices with m columns, write

$$\mathbf{Y} = \mathbf{U}^{(m)} \left(\mathbf{D}^{(m)} (\mathbf{V}^{(m)})' \right) + \mathbf{E} \quad (7)$$

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

16 *Algorithm SVD.em:*

- 17 1. Specify a dimension (rank), m , for the model.
- 18 2. Scale the observations at each monitoring site (columns of \mathbf{Y}) to norm (variance)
 19 one; call this matrix $\tilde{\mathbf{Y}}$.
- 20 3. Compute an initial average temporal vector u_1 as the set of row averages of
 21 nonmissing values in $\tilde{\mathbf{Y}}$. Initialize missing observations in the data matrix $\tilde{\mathbf{Y}}$ using
 22 elements of a rank-one approximation provided by a regression (without intercept) of
 23 the data in each column (site) on u_1 .
- 24 4. Compute the rank- m SVD-approximation of the now complete data matrix.

- 1 5. Impute the missing values in \tilde{Y} by the elements of the rank- m SVD approximation
- 2 just computed.
- 3 6. Return to step 4 and iterate to convergence.

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

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

13 *Algorithm SEOF:*

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

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

23 The smoothing in step (b) uses every other value so that the smoothing spline would not be
24 applied to time series with temporal correlations artificially inflated by the overlapping of 2-
25 week averages. The smoothed singular vectors are no longer exactly orthogonal, but exact

1 orthogonality is not a concern. We choose as optimal the number of components m with best (or
2 near-best) distribution of BIC values in the cross-validated fits.

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

9 *4.2 Trend fits at MESA Air monitoring sites*

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

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

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

$$7 \quad Y_{st}^d = Y_{st} - \sum_{j=1}^m \hat{\beta}_{js} f_j(t) \quad (8)$$

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

$$9 \quad \hat{\beta}_{0s} = \frac{1}{T} \sum_t Y_{st}^d \quad (9)$$

10 Fitted trends derived from trend coefficients $\hat{\beta}_{0s}, \hat{\beta}_{1s}, \dots, \hat{\beta}_{ms}$ computed by this procedure
 11 appear appropriate by visual inspection. In principal, estimated trend coefficients should be
 12 influenced by the spatial covariates according to the regression model of equation (2). That is
 13 the case in our likelihood-based approach to the hierarchical model (Szpiro et al., 2009), but
 14 covariates are not incorporated at this stage of our sequential fitting procedure. Fitted trend
 15 coefficients are used as outcomes in the computation of spatial regression models as explained in
 16 the following sub-section.

17 4.3 Spatial regression by Partial Least Squares (PLS)

18 Our hierarchical model proposes that the spatial variation in the parameters $\hat{\beta}_{js}$ of the
 19 temporal trend models can be modeled via regression on spatial covariates. The GIS-based
 20 dataset of spatial covariates for PM_{2.5} concentrations provides substantial numbers of highly
 21 correlated spatial covariates. For example, the composite lengths of A1 roads in buffers of, say,
 22 400 meters around specified locations are highly correlated with lengths of roads in 500 meter
 23 buffers and negatively correlated with the distance to the nearest A1 road. Percentages of
 24 property in various land use categories are similarly correlated across buffer sizes; for example,
 25 percentages of land classified residential in a buffer are substantially negatively correlated with
 26 percentages of land classified as commercial. Model specification with large sets of

1 multicollinear predictors typically involves either (a) variable selection, (b) shrinkage or
2 regularization, perhaps including variable selection as in a “lasso” approach (Tibshirani, 1996),
3 or (c) regression on a smaller number of composite covariate scores. While our fundamental
4 concern is the quality of predictions, we prefer not to choose a method that would select one
5 particular buffer size for inclusion in a model ignoring neighboring buffer sizes, or one particular
6 land use categorization at the expense of a correlated land use categorization. We choose,
7 instead, to regress on a small number of composite covariate scores using the method of PLS
8 regression to define the composite scores (see, for example, Garthwaite, 1994 or Abdi, 2010).
9 The description of the composite scores in terms of individual variable loadings can be useful as
10 it facilitates comparison of regression models across the four geographic modeling regions. PLS
11 regressions were computed using the `pls` package for the R system ([http://cran.r-](http://cran.r-project.org/web/packages/pls)
12 [project.org/web/packages/pls](http://cran.r-project.org/web/packages/pls)).

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

22 The code in the `pls` package for the R system computes conventional leave-one-out cross-
23 validatory assessments of predictions of these regression models. However, these cross-
24 validations assume a modeling framework with spatially independent errors. Since our
25 hierarchical model includes spatial correlation, we choose the dimension of the PLS regression
26 by cross-validation with a universal kriging prediction involving component scores defined by
27 the PLS algorithm and a model for the spatial covariance structure of the residuals from those

1 regressions as indicated in equation (2). We computed spatial covariance matrices $\Sigma_\varepsilon(\phi_j)$ in
 2 terms of exponential variogram models fitted to the residuals of the PLS regressions using the
 3 `likfit` function in the `geoR` package for the R system ([http://cran.r-](http://cran.r-project.org/web/packages/geoR)
 4 [project.org/web/packages/geoR](http://cran.r-project.org/web/packages/geoR)).

5 4.4 Residual covariance modeling

6 The hierarchical model of equations (1)-(3) involves a spatial covariance matrix $\Sigma_\varepsilon(\phi_\varepsilon)$ for
 7 the spatio-temporal residuals in equation (2). The structure of, $\Sigma_\varepsilon(\phi_\varepsilon)$, is estimated using the
 8 Sampson-Guttorp deformation model for nonstationary spatial covariance fitted to an empirical
 9 spatial covariance matrix computed from the spatio-temporal matrix of residuals from the site-
 10 specific trend models for the AQS monitoring sites. That is, we compute residuals

$$11 \hat{\varepsilon}_{it} = \hat{\varepsilon}(s_i, t) = Y_{s_i, t} - \hat{\mu}(s_i, t), \quad i = 1, \dots, N \quad (10)$$

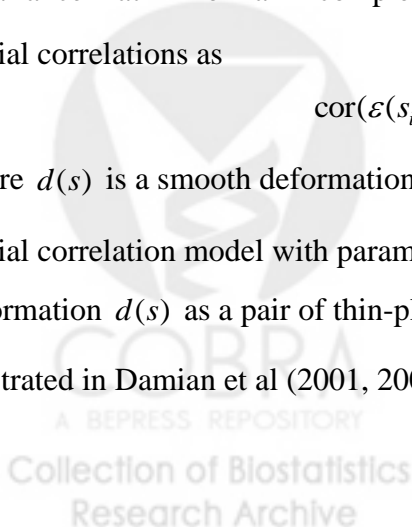
12 where

$$13 \hat{\mu}(s_i, t) = \hat{\beta}_{0s_i} + \sum_{j=1}^m \hat{\beta}_{js_i} f_j(t) \quad (11)$$

14 with parameter estimates $\hat{\beta}_{0s_i}, \hat{\beta}_{1s_i}, \dots, \hat{\beta}_{ms_i}$ computed at these AQS sites as explained in section
 15 4.2. We compute a covariance matrix with elements $s_{ij} = \text{cov}(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{jt})$ based on empirical
 16 covariances over time between time series of residuals at locations s_i and s_j . Because of
 17 possibly substantial amounts of missing data, we use an EM algorithm to estimate this spatial
 18 covariance matrix from an incomplete data matrix. The spatial deformation model expresses
 19 spatial correlations as

$$20 \text{cor}(\varepsilon(s_i, t), \varepsilon(s_j, t)) = \gamma_\theta(|d(s_i) - d(s_j)|) \quad (12)$$

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



1 4.5 Cross-validated spatio-temporal predictions

2 We predict concentrations at subject homes using the hierarchical model described above.

3 Our methodology incorporates covariates and spatial interpolation of long-term averages and
4 seasonal trends based a on spatial correlation model, a procedure that may be regarded as a
5 generalization of “universal kriging” or “kriging with external drift” (Wackernagel 2010).

6 Specifically, given estimates of the parameters of the components of the hierarchical model as
7 explained in sections 4.1-4.4, spatio-temporal predictions at target locations s_0 are computed as

$$8 \hat{Y}_{s_0,t} = \hat{\mu}(s_0,t) + \hat{\varepsilon}(s_0,t) \quad (13)$$

9 where

$$10 \hat{\mu}(s_0,t) = \hat{\beta}_{0s_0} + \sum_{j=1}^m \hat{\beta}_{js_0} f_j(t) \quad (14)$$

11 Each of the $\hat{\beta}_{js_0}$ is computed by universal kriging using the PLS component scores as spatial
12 covariates and the spatial covariance models underlying the estimates of the matrices $\Sigma_e(\phi_j)$.

13 The spatio-temporal residual field $\hat{\varepsilon}(s_0,t)$, being mean zero, is computed by a simple kriging
14 calculation with the spatial covariances defined by the spatial deformation model underlying the
15 spatial covariance matrix $\Sigma_\varepsilon(\phi_\varepsilon)$. Estimates are computed for each 2-week period indexed by t
16 using all monitoring data available at that time point from the AQS, MESA Air fixed, and MESA
17 Air home monitors.

18 5. Application

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

Collection of Biostatistics
Research Archive

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

16 --- Figure 4 ---

17 The cross-validatory assessment of the number of PLS components for our land use
18 regressions based on universal kriging predictions suggests at most two predictive components
19 for all of the trend coefficients in all regions. This result is strikingly different from the result of
20 conventional PLS regression cross-validation without a spatial correlation model. As many as 10
21 PLS components are suggested for the intercept parameter β_0 by conventional cross-validation,
22 but substantially better cross-validatory predictions are found using only two components
23 together with a spatial correlation model.

24 Loadings (correlations) of the spatial covariates on the two PLS components for the spatial
25 regression model for the intercept (long-term mean), β_{0s} , are illustrated in Figure 5. We see that
26 the first component is highly negatively correlated with distance to the nearest A1 highway and
27 positively correlated with the distances to the nearest A2 and A3 highways. Lengths of A1 vs

1 A2 and A3 roads in buffers show the expected opposite correlation pattern. That is, this first
2 component, which is predictive of long-term mean concentrations, defines a score that contrasts
3 sites relatively near A1 highways but not necessarily near A2 and/or A3 roads with sites that are
4 near A2 and /or A3 roads but far from A1 highways. The former sites have higher $PM_{2.5}$
5 concentrations on average while the latter sites have lower concentrations. The sites with
6 positive scores (near A1 highways) are largely residential as indicated by the positive loadings
7 on the residential land use covariates and contrasting negative correlations with the cropland
8 covariates. The second PLS component defines a commercial vs. residential property contrast.

9 --- Figure 5 ---

10 Scatterplots of cross-validated universal kriging predictions using two component PLS
11 regressions for each of the three trend coefficients are presented in Figure 6. The red dots are the
12 MESA Air fixed sites. We see reasonably good regressions for the long-term mean $\hat{\beta}_{0,s}$ and the
13 3rd coefficient $\hat{\beta}_{2,s}$ for the amplitude of the simple cyclic seasonal structure of Figure 4, but little
14 ability to predict variation in the 2nd trend coefficient $\hat{\beta}_{1,s}$ for the component carrying the long-
15 term decrease in concentrations.

16 --- Figure 6 ---

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

24 --- Figure 7 ---

25 Figure 8 presents cross-validated predictions of the 2-week observations for the four sites
26 seen in Figure 2. The dashed black lines represent the long-term means of the trend models
27 fitted to the data (black dots) while the dashed red lines are the long-term means of the cross-

1 validation predictions (green lines varying about predicted trends drawn in red). The predictions
2 generally track the observations quite well with modest levels of over- or under-estimation of the
3 long-term means. We note that the magnitude of the error in these cross-validated predictions is
4 probably greater than the error of prediction expected at most MESA subject homes as, for
5 example, prediction of concentrations at locations near site 060372005 will benefit from the
6 monitoring data at that site, data which were excluded in its own prediction.

7 --- Figure 8 ---

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

14 --- Figure 9 ---

15 --- Figure 10 ---

16 --- Figure 11 ---

17 --- Figure 12 ---

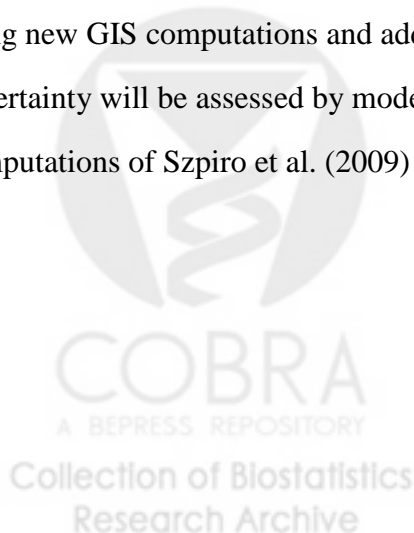
18 Results demonstrate the substantially higher and spatially variable $PM_{2.5}$ concentration
19 levels for the MESA Air subject locations in Southern California in contrast to all the other
20 MESA Air study areas. St Paul, MN, Baltimore, MD, and New York City all demonstrate
21 similar ranges of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

23 Conventional “plug-in” estimates of the standard errors of predictions of long-term averages
24 could be computed for the final step of our prediction strategy (section 4.5). However, we
25 refrain from computing these estimates as the current model has some recognized deficiencies
26 and because such standard errors do not account for prediction uncertainty that derives from the
27 multi-step pragmatic model-fitting procedure we have employed. We report instead the

1 accuracy of these pragmatic ambient concentration predictions in terms of descriptive statistics
2 on the cross-validated (leave-one-out) errors of prediction of long-term mean concentrations for
3 6 MESA Air study areas.

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

13 We see that the uncertainty of long-term estimates is similar across regions, about 3% to 5%
14 of the mean. Comparison of standard deviations of long-term means to cross-validated root-
15 mean-squared errors indicates that predictions are capturing meaningful intra-urban spatial
16 variability in some areas including California, but these root-mean-square errors are nearly as
17 large as the standard deviations in some of the other study areas such as North Carolina. In
18 future work we expect improved city-specific spatio-temporal predictions and lower uncertainty
19 using new GIS computations and additional land use and traffic covariates. Furthermore,
20 uncertainty will be assessed by model-based standard errors using the likelihood modeling
21 computations of Szpiro et al. (2009) and Lindström et al. (2010) as well as cross-validation.



1 **6. Discussion**

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

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

15 We will ultimately use the likelihood method reported in Szpiro et al. (2009) and Lindström
16 et al. (2010) because it provides a unified framework and gives standard errors (as opposed to
17 cross-validation in this paper). However, much of the model selection work will still be done
18 outside of the likelihood framework using the methods presented in this paper. This includes the
19 specification of the SEOFs and the PLS or selection of covariates for a similar regression model
20 with a universal kriging cross-validation approach. The work here, carried out in parallel to the
21 development of the likelihood method, provided the most pragmatic approach to obtaining initial
22 ambient concentration estimates for use in our epidemiology studies.

23 The current modeling and analysis leaves some problematic issues to be addressed in future
24 work. The most important is dealing appropriately with the fact that 2-week average
25 concentrations derive from different monitoring networks (AQS and MESA Air) with different
26 monitoring instruments and temporal sampling protocols. This will require a nested

1 specification of spatio-temporal correlation at a daily time scale. This daily time scale structure
2 is also fundamental to a “downscaling” extension of the current model predictions in order to
3 obtain estimates at a daily time scale for acute exposure estimation in epidemiologic outcomes
4 and, especially cardiovascular events that are expected to be sensitive to acute as well as chronic
5 exposure.

6 The current model includes only temporal factors (the temporal basis functions) and
7 spatially varying covariates. Recent extensions of the likelihood model fitting can incorporate
8 spatio-temporal covariates, the most important of which are spatio-temporally varying
9 characterizations of the effects of traffic on ambient exposure. Lindström et al. (2010) models
10 NO_x (rather than $\text{PM}_{2.5}$) using as a covariate theoretical predictions of pollutant concentrations
11 provided by a physics-based plume dispersion model, EPA’s CALINE model (Wilton et al.
12 2009).

13 The ultimate objective of the modeling described here is to provide predicted exposures for
14 estimating health effects in epidemiology studies. Up to now, we have used a “plug-in”
15 approach that does not account for the additional variability resulting from uncertainty in the
16 spatio-temporal prediction procedure. We have recently developed an efficient bootstrap-based
17 approach to incorporating this uncertainty in health effect estimation (Szpiro et al. 2009a). In
18 future work, we will apply this methodology based on fitting of the current spatio-temporal
19 model using likelihood methods in order to obtain corrected standard errors for the disease model
20 parameters of interest.

22 **Acknowledgements**

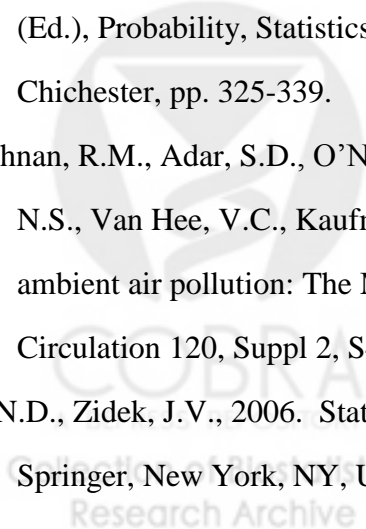
23 Although the research described in this presentation has been funded in part by the United
24 States Environmental Protection Agency through grant R831697 and assistance agreement CR-
25 83407101 to the University of Washington, it has not been subjected to the Agency’s required
26 peer and policy review and therefore does not necessarily reflect the views of the Agency and no

1 official endorsement should be inferred. Additional support was provided by an award to the
2 University of Washington under the National Particle Component Toxicity (NPACT) initiative
3 of the Health Effects Institute (HEI).

4 **References**

- 5 Abdi, H. 2010. Partial least squares regression and projection on latent structure regression (PLS
6 regression). *WIREs Computational Statistics* 2, 97-106.
- 7 Adar, S.D., Klein, R., Klein, B.E.K., Szpiro, A.A., Cotch, M.F., Wong, T.Y., O'Neill, M.S.,
8 Shrager, S., Graham Barr, R., Siscovick, D., Daviglius, M.L., Sampson, P.D., Kaufman, J.,
9 2010. Air pollution and the human microvasculature in vivo assessed via retinal imaging:
10 The Multi-Ethnic Study of Atherosclerosis and Air pollution (MESA Air), accepted for
11 publication in *PLoS Medicine*.
- 12 Banerjee, S., Gelfand, A.E., Carlin P., 2004, *Hierarchical Modeling and Analysis for Spatial*
13 *Data*. Chapman and Hall, 472 pp.
- 14 Banerjee, S., Johanson, G.A., 2006, Coregionalized single- and multiresolution spatially varying
15 growth curve modeling with application to weed growth. *Biometrics* 62, 864-876.
- 16 Beckers J., Rixen, M., 2003. EOF calculations and data filling from incomplete oceanographic
17 datasets. *Journal of Atmospheric and Oceanic Technology* 20, 1839-1856.
- 18 Cohen, M.A., Adar, A.D., Allen, R.W., Avol, E., Curl, C.L., Gould, T., Hardie, D., Ho, A.,
19 Kinney, P., Larson, T.V., Sampson, P.D., Sheppard, L., Stukovsky, K.D., Swan, S.S., Liu,
20 L-J. S., Kaufman, J.D., 2009. Approach to estimating participant pollutant exposures in the
21 Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Environmental*
22 *Science & Technology* 43, 4687-4693.
- 23 Damian, D., Sampson, P.D., Guttorp, P., 2001. Bayesian estimation of semi-parametric non-
24 stationary spatial covariance structures, *Environmetrics* 12, 161-178.
- 25 Damian, D., Sampson, P.D., Guttorp, P., 2003. Variance modeling for nonstationary processes
26 with temporal replications. *Journal of Geophysical Research – Atmosphere*, 108 (D24).

- 1 EPA, 2010. The U.S. Environmental Protection Agency, Technology Transfer Network (TTN)
2 Air Quality System (AQS), <http://www.epa.gov/ttn/airs/airsaqs/> (last accessed, Sept 2010).
- 3 ESRI. ArcGIS (Version 9.1). Redlands, CA.
- 4 Fanshawe, T.R., Diggle, P.J., Rushton, S., Sanderson, R., Lurz, P.W.W., Glinianaia, S.V.,
5 Pearce, M.S., Parker, L., Charlong, M., Pless-Mullooli, T., 2008. Modelling spatio-temporal
6 variation in exposure to particulate matter: a two-stage approach. *Environmetrics* 19, 549-
7 566.
- 8 Fuentes, M., Guttorp, P., Sampson, P.D., 2006. Using transforms to analyze space-time
9 processes, in: Finkenstadt, B., Held, L., Isham, V. (Eds.), *Statistical Methods for Spatio-
10 Temporal Systems*, Chapman and Hall/CRC, Boca Raton, FL, USA, pp. 77-150.
- 11 Furrer, E.M., Nychka, D.W., 2007. A framework to understand the asymptotic properties of
12 Kriging and splines. *Journal of the Korean Statistical Society* 36, 57-76.
- 13 Garthwaite, P. 1994. An interpretation of partial least squares. *Journal of the American
14 Statistical Association* 89, 122-127.
- 15 Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D. 2008. A
16 review of land-use regression models to assess spatial variation of outdoor air pollution.
17 *Atmospheric Environment* 42, 7561-7578.
- 18 Kent, J.T., Mardia, K.V. 1994. The link between kriging and thin-plate splines, in: Kelly, F.P.
19 (Ed.), *Probability, Statistics, and Optimization: a Tribute to Peter Whittle*, Wiley,
20 Chichester, pp. 325-339.
- 21 Krishnan, R.M., Adar, S.D., O'Neil, M.S., Barr, G.R., Polak, J.F., Herrington, D., Jorgensen,
22 N.S., Van Hee, V.C., Kaufman, J.D., 2009. Vascular responses to short and long-term
23 ambient air pollution: The Multi-Ethnic Study of Atherosclerosis and Air Pollution,
24 *Circulation* 120, Suppl 2, S462.
- 25 Le, N.D., Zidek, J.V., 2006. *Statistical Analysis of Environmental Space-Time Processes*,
26 Springer, New York, NY, USA.



- 1 Lindström, J., Szpiro, A.A., Sheppard, L., Sampson, P.D., Oron, A., Richards, M., Larson, T.,
2 2010. A flexible spatio-temporal model for air pollution: incorporating the output from a
3 source dispersion model.
- 4 Moore, D.K., Jerrett, M., Mack, W.J., Kunzli, N., 2007. A land use regression model for
5 predicting ambient fine particulate matter across Los Angeles, CA. *Journal of*
6 *Environmental Monitoring* 9, 246-252.
- 7 Paciorek, C.J., Yanosky, J.D., Puett, R.C., Laden, F., Suh, H.H., 2009. Practical large-scale
8 spatio-temporal modeling of particulate matter concentrations. *Annals of Applied Statistics*
9 3, 370-397.
- 10 R Development Core Team, 2009. R: A Language and Environment for Statistical
11 Computing. R Foundation for Statistical Computing, Vienna, Austria ([http://www.R-](http://www.R-project.org)
12 [project.org](http://www.R-project.org)).
- 13 Rao, S.T. Zurbenko, I.G., 1994. Detecting and tracking changes in ozone air quality, *Journal of*
14 *the Air & Waste Management Association* 44, 1089-1092.
- 15 Ross, Z., Jerrett, M., Ito, K., Tempalski, B., Thurston, G.D., 2007. A land use regression for
16 predicting fine particulate matter concentrations in the New York City region. *Atmospheric*
17 *Environment* 41, 2255-2269.
- 18 Sahu, S.K., Gelfand, A.E., Holland, D.M., 2006. Spatio-temporal modeling of fine particulate
19 matter. *Journal of Agricultural, Biological, and Environmental Statistics* 11, 61–86.
- 20 Su, J.G., Jerrett, M., Beckerman, B., Wilhelm, M., Ghosh, J.K., Ritz, B., 2009. Predicting traffic-
21 related air pollution in Los Angeles using a distance decay regression selection strategy.
22 *Environmental Research* 109, 657-670.
- 23 Smith, R.L., Kolenikov, S., Cox, L.H., 2003. Spatio-temporal modeling of PM_{2.5} data with
24 missing values. *Journal of Geophysical Research* 108(D24), 9004.
- 25 Szpiro, A.A., Sampson, P.D., Sheppard, L., Lumley, T., Adar, S.D., Kaufman, J.D., 2009.
26 Predicting intraurban variation in air pollution concentrations with complex spatio-
27 temporal interactions. *Environmetrics*, in press.

- 1 Szpiro, A.A., Sheppard, L., Lumley, T., 2009a. Efficient measurement error correction with
2 spatially misaligned data. UW Biostatistics Working Paper Series, Working Paper 350.
3 <http://www.bepress.com/uwbiostat/paper350>
- 4 Wackernagel, H., 2010. Multivariate Geostatistics: An Introduction with Applications, 3rd ed.
5 Springer-Verlag, Berlin, Germany.
- 6 Wilton, D., Szpiro, A., Gould, T., Larson, T., 2010. Improving spatial concentration estimates
7 for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in
8 Los Angeles, CA and Seattle, WA. *Science of the Total Environment*, 408 1120-1130.
- 9 Wise, E.K., Comrie, A.C., 2005. Extending the Kolmogorov–Zurbenko filter: Application to
10 ozone, particulate matter, and meteorological trends. *Journal of the Air & Waste*
11 *Management Association* 55, 1208–1216.
- 12 Taylor, C.J., Pedregal, D.J., Young, P.C., Tych, W., 2006. Environmental time series analysis
13 and forecasting with the Captain toolbox. *Environmental Modelling & Software* 22, 797-
14 814.
- 15 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal*
16 *Statistical Society, Series B* 58, 267-288.
- 17 Yanosky, J.D., Paciorek, C.J., Suh, H.H., 2009. Predicting chronic fine and coarse particulate
18 exposures using spatio-temporal models for the Northeastern and Midwestern United
19 States. *Environmental Health Perspectives* 117, 522-529.

20



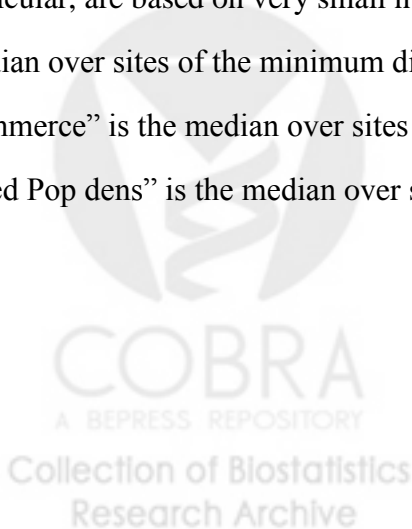
1 **Table 1.** Site descriptive statistics for selected spatial covariates and average PM_{2.5}
 2 concentration.

3

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

4

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



1 **Table 2.** Descriptive statistics on the cross-validated (leave-one-out) errors of prediction of
 2 long-term mean concentrations for 4 major modeling regions

3

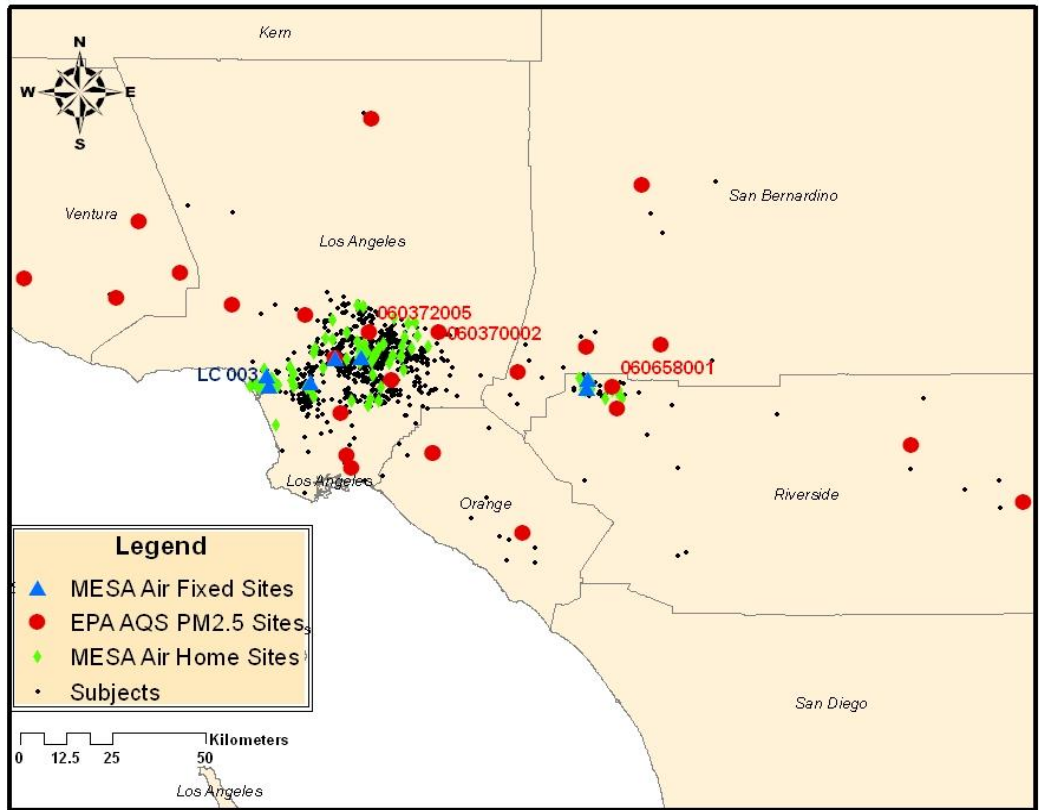
Data Scale	Site	All Sites			MESA Air Fixed Sites		
		Mean	SD	RMSE	Mean	SD	RMSE
Log	CA	2.85	0.27	0.15	3.02	0.15	0.04
	IL	2.66	0.13	0.09	2.67	0.12	0.04
	MN	2.30	0.14	0.10	2.36	0.07	0.03
	MD	2.70	0.08	0.07	2.76	0.06	0.03
	NY	2.61	0.11	0.09	2.68	0.18	0.03
	NC	2.66	0.06	0.07	2.69	0.03	0.03
Original	CA	17.84	4.47	2.42	20.64	3.04	0.94
	IL	14.45	1.89	1.31	14.59	1.93	0.68
	MN	10.07	1.32	0.88	10.61	0.72	0.34
	MD	14.90	1.23	1.05	15.77	0.90	0.42
	NY	13.66	1.58	1.21	14.70	2.52	0.44
	NC	14.34	0.92	0.95	14.72	0.39	0.40

4



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

2

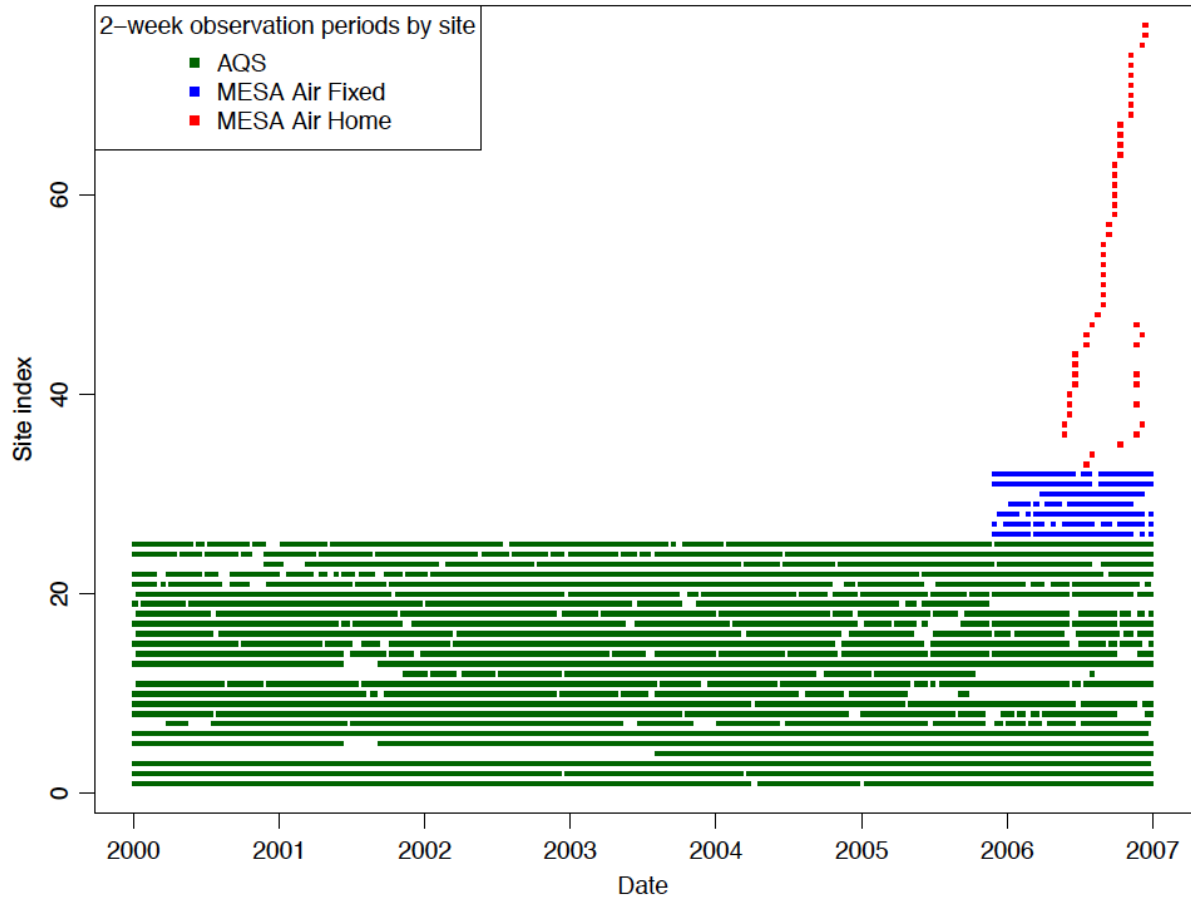


3

4



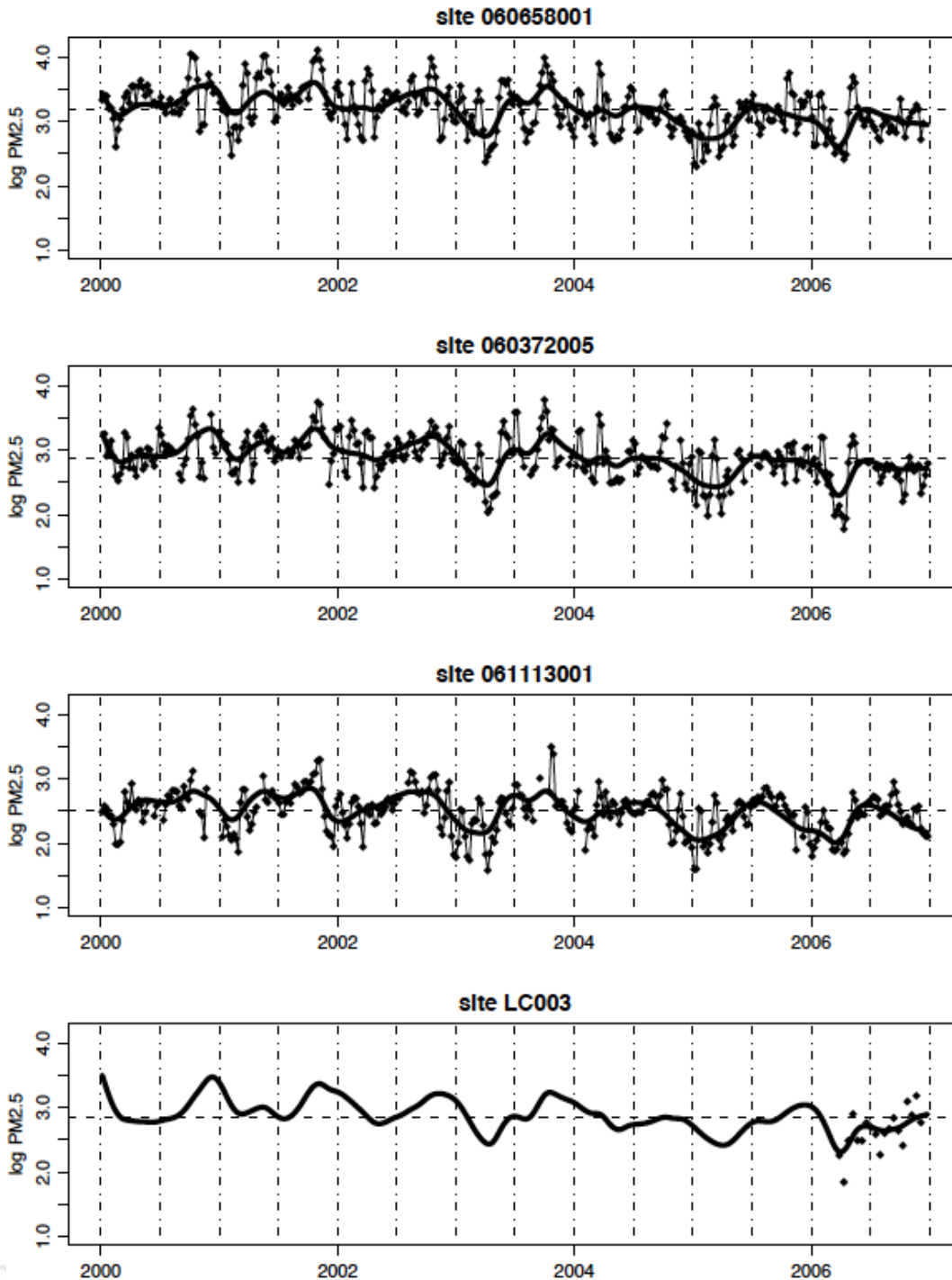
1 **Figure 2.** Schematic of the temporal sampling pattern for AQS monitors and the MESA Air
2 fixed and home sites. Each point in this figure represents a 2-week average measurement.
3



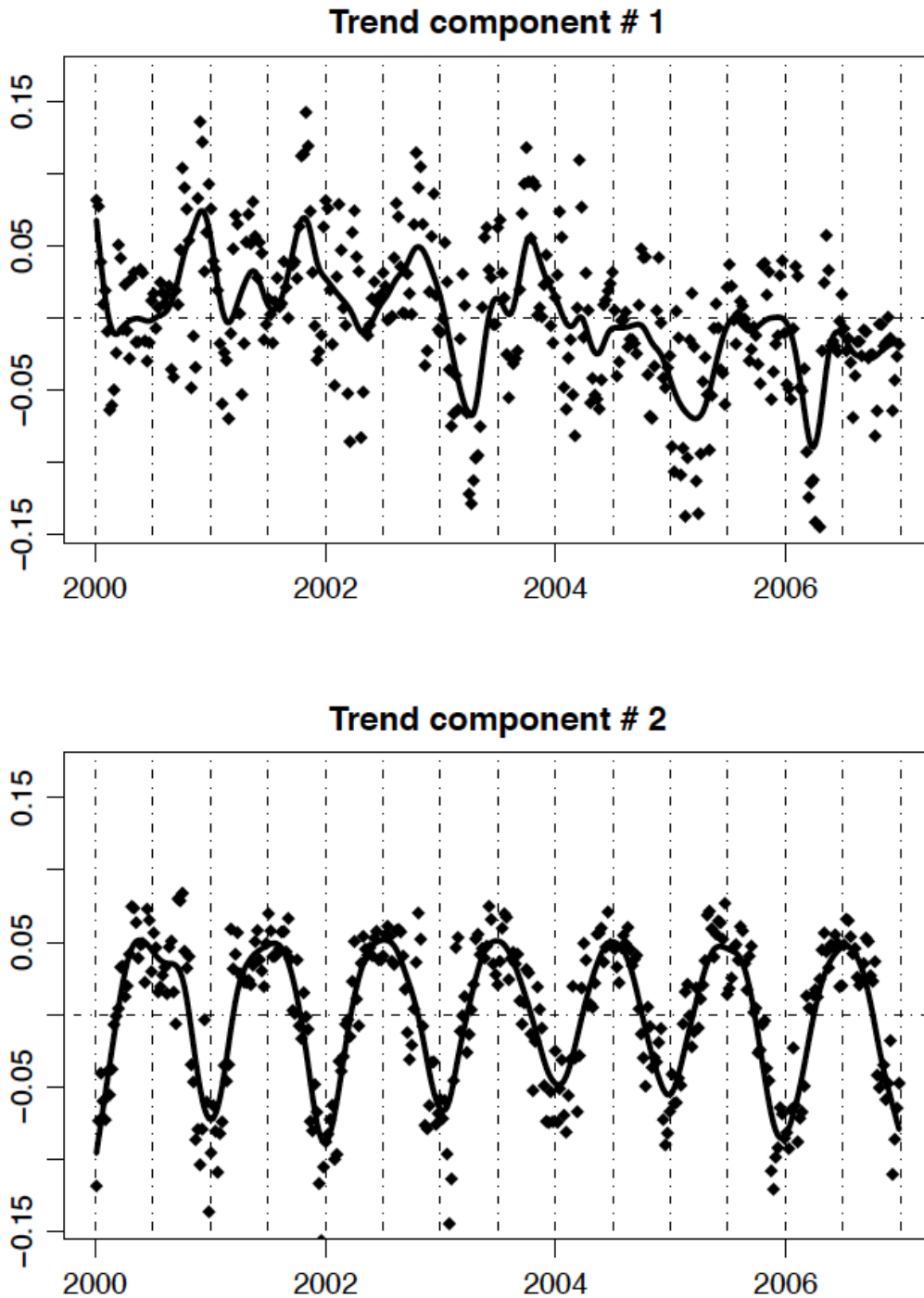
4



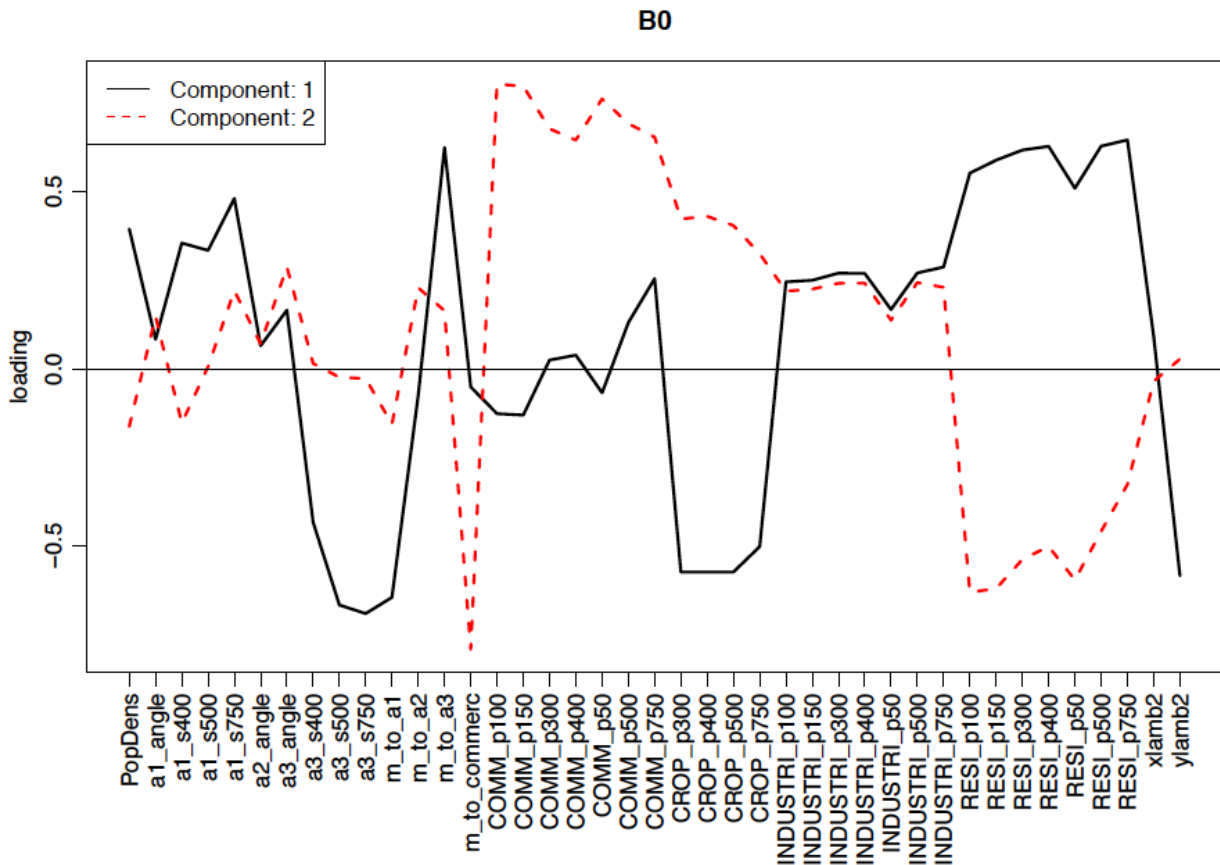
1 **Figure 3.** Example log-transformed two-week average $PM_{2.5}$ data at AQS and MESA Air
2 fixed sites in the Los Angeles region. The black points are measurements and the lines
3 represent estimated temporal trends based on the SEOF model (see Section 4.2). The
4 locations of each of these sites are shown on the map in Figure 1.
5



1 **Figure 4.** Smoothed empirical orthogonal basis functions for log-transformed two-week
2 average concentrations of PM_{2.5} in the Los Angeles area. The first smoothed component
3 explains 27.5% of the variation in the matrix of log-transformed 2-week average PM_{2.5}
4 concentrations while the second component explains only 9.4% of the variation.
5
6



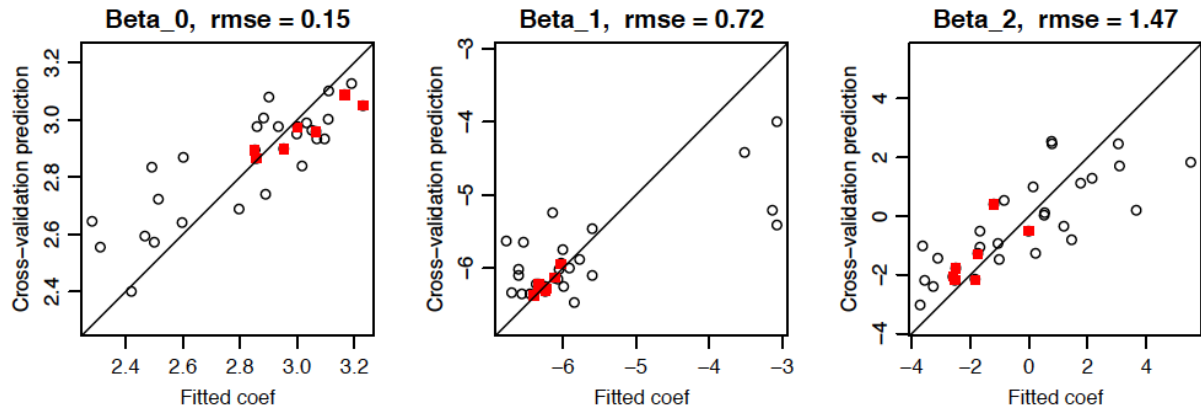
1 **Figure 5.** Loadings (correlations) of the spatial covariates on the two Partial Least Squares
 2 (PLS) components for the spatial regression model for the intercept (long-term mean), β_{0s} ,
 3 for the Los Angeles study region. The a1, a2, and a3 variables refer to major road census
 4 feature class codes; they include the angle to the nearest such roadway, the lengths of
 5 roadway segments in circular buffers of varying radii, and the distances in meters to the
 6 nearest roadway. The variables beginning COMM, CROP, INDUSTRI, and RESI refer to
 7 fractions of the property in circular buffers designated as Commercial, Cropland, Industrial,
 8 and Residential using the ArcGIS software system (ESRI).
 9



10



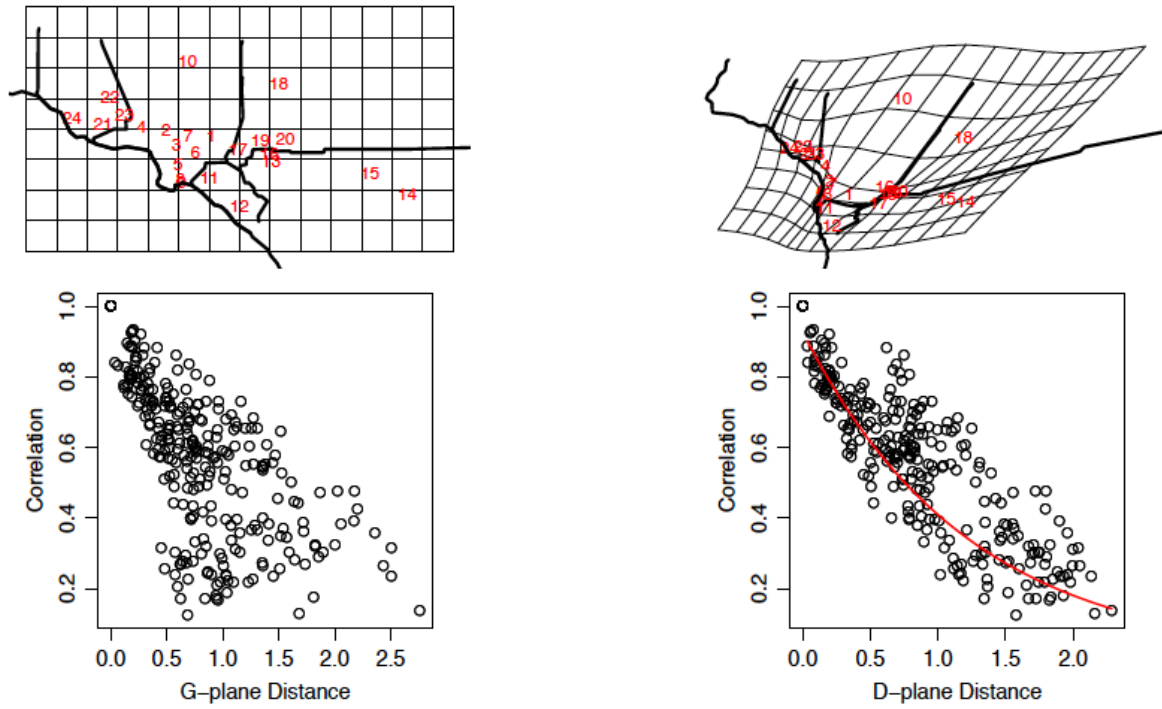
1 **Figure 6.** Cross-validated predictions of the β_{0s} , β_{1s} , and β_{2s} spatial fields of coefficients
2 for the long-term average and two SEOF temporal trends in the Los Angeles region. The
3 black dots represent the AQS locations while the red dots are the MESA Air fixed sites. The
4 predictions are based on PLS regression models with 2, 1, and 1 components, respectively.
5



6
7



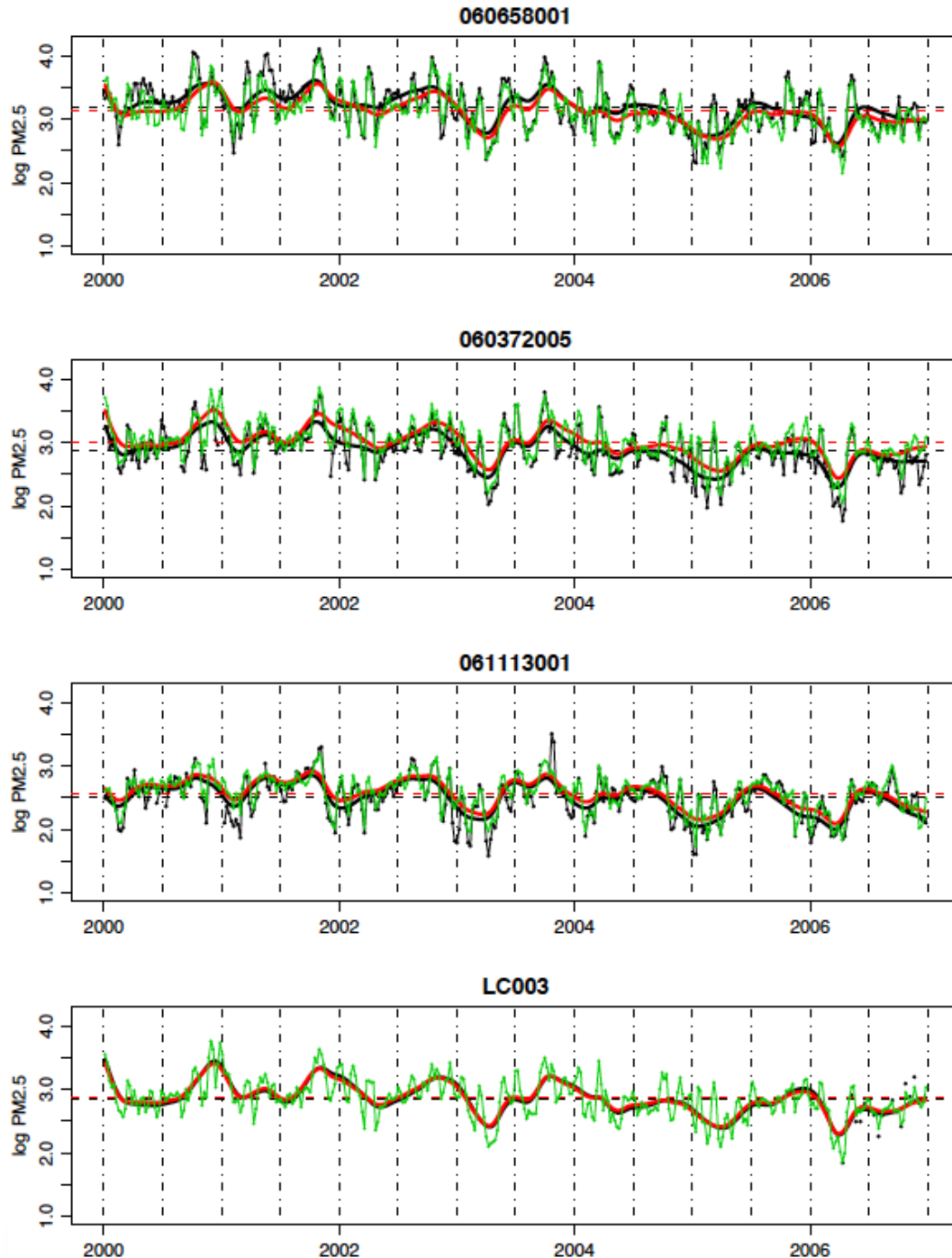
1 **Figure 7.** Spatial structure of the spatio-temporal residuals before (left) and after (right)
2 transformation using the Sampson-Guttorp method to account for nonstationarity.
3



4

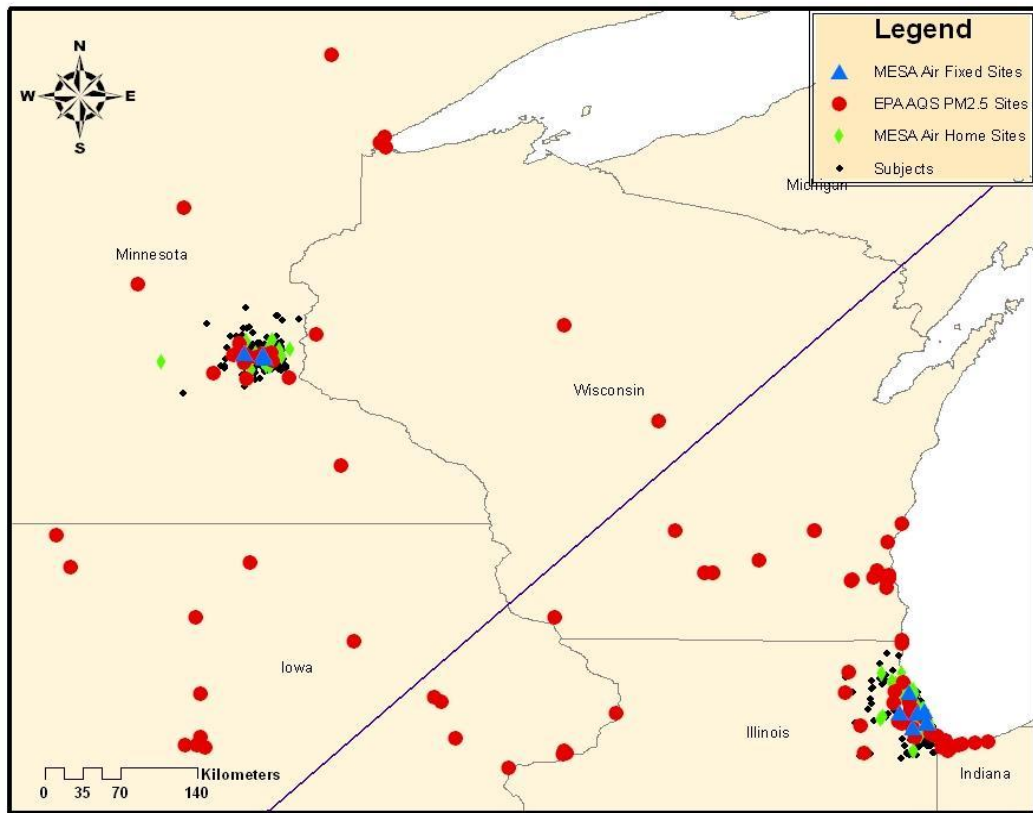


1 **Figure 8.** Example cross-validated predictions of log-transformed two-week concentrations
2 at AQS and MESA Air fixed sites in the Los Angeles region. The black dots are the
3 measured data, the red curves show predicted trends based on the SEOF part of the spatio-
4 temporal model, and the green lines show two-week average predictions that incorporate the
5 spatio-temporal residuals.
6



Research Archive

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

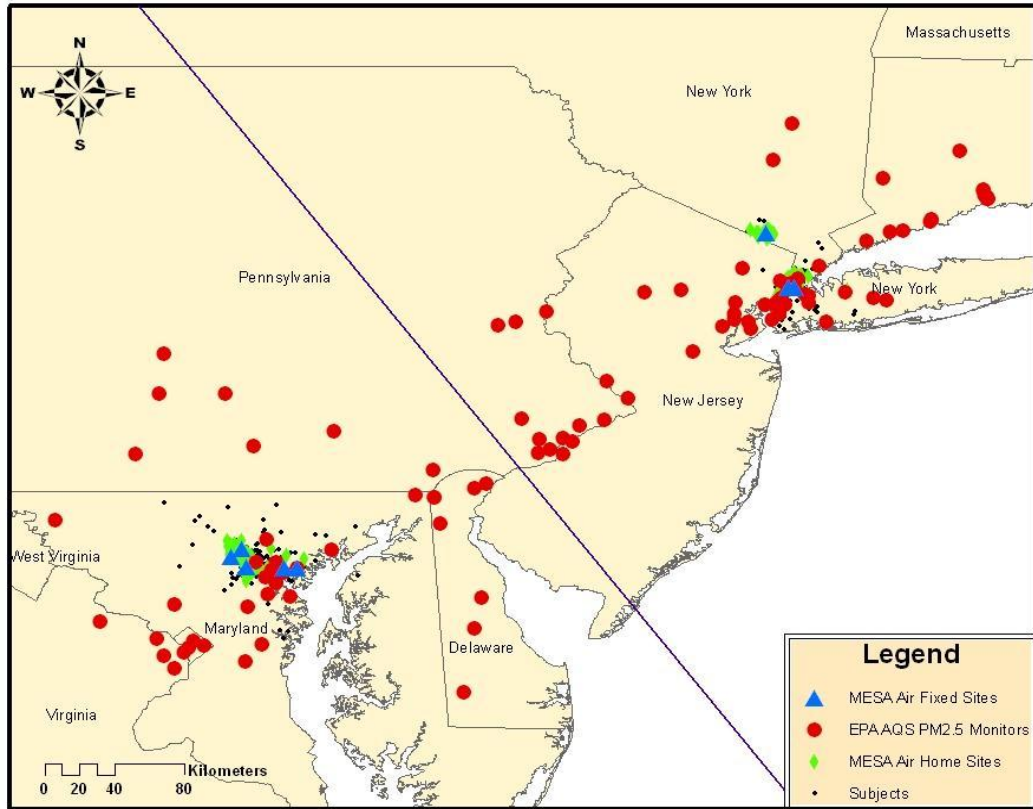


3
4



1 **Figure 10.** Monitoring sites and subject home locations in the northeast region spanning the New
2 York and MESA Air study areas.

3

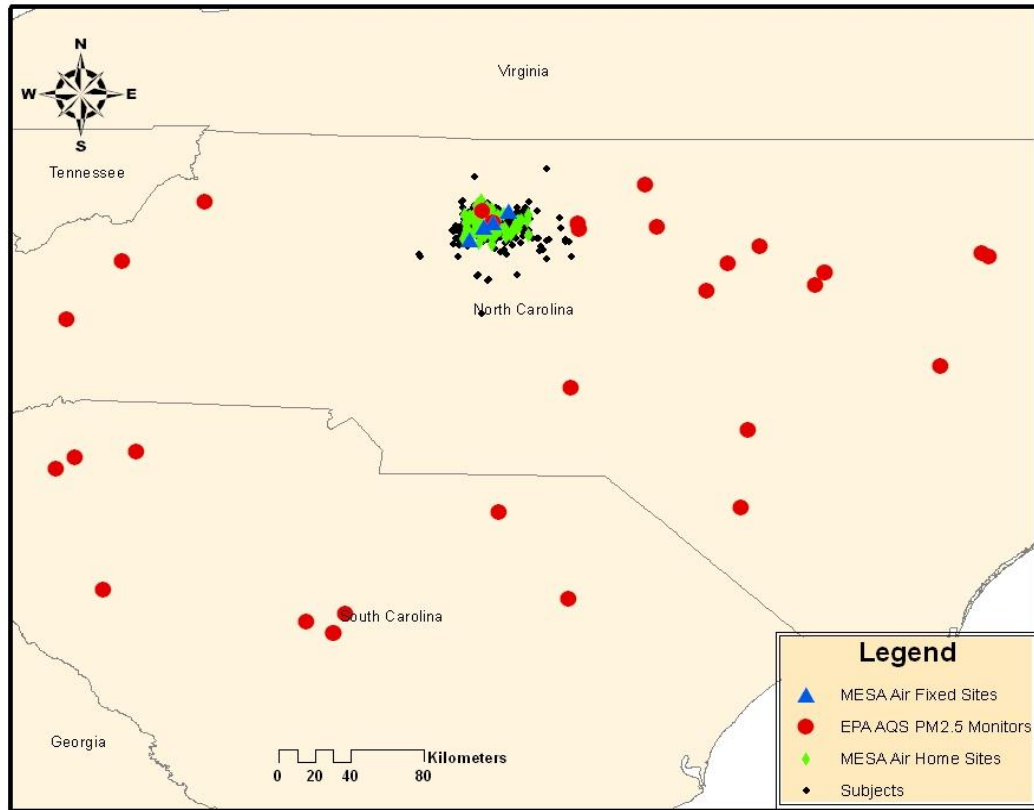


4

5



1 **Figure 11.** Monitoring sites and subject home locations in the North Carolina region.

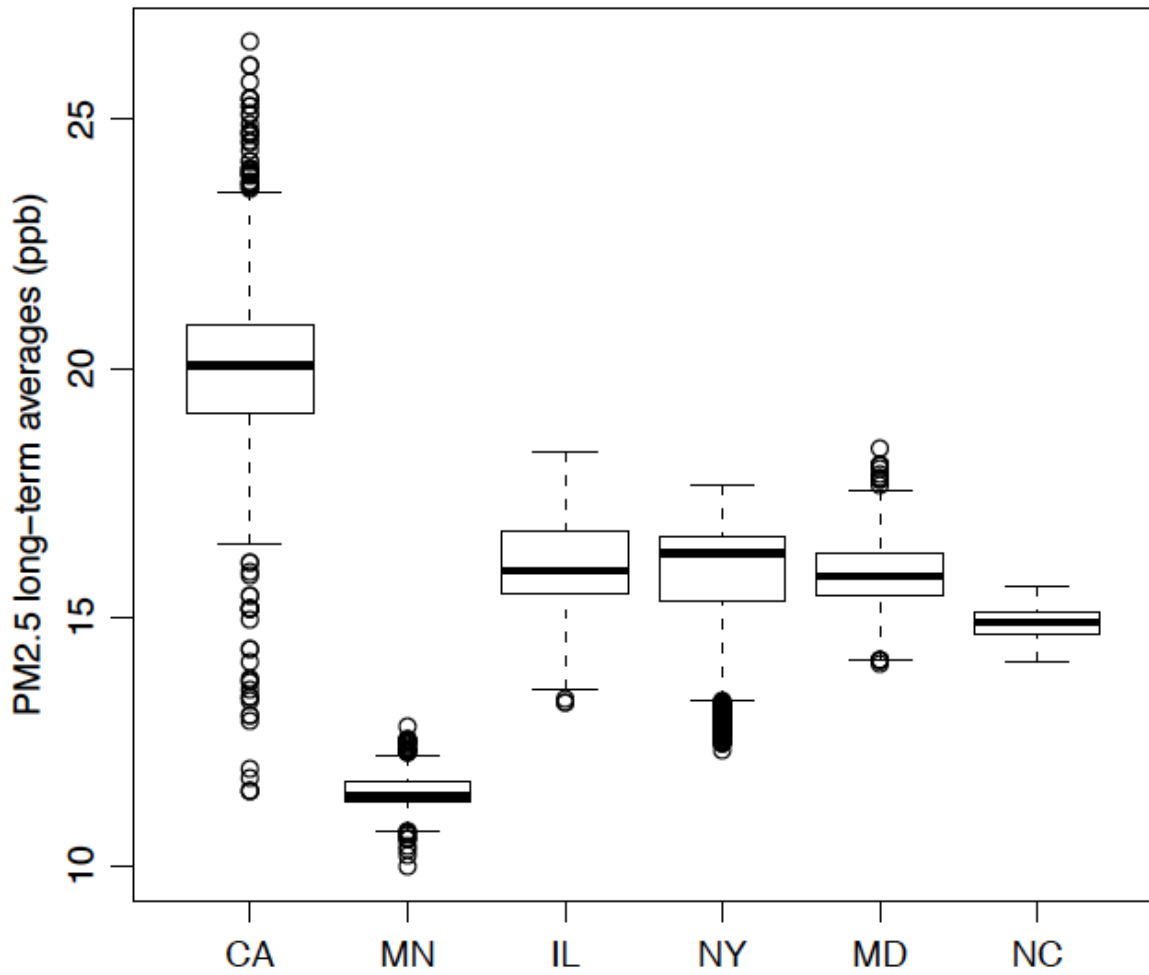


2

3



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



3

