UW Biostatistics Working Paper Series

12-16-2013

Prediction of fine particulate matter chemical components for the Multi-Ethnic Study of Atherosclerosis cohort: A comparison of two modeling approaches

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

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

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

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

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

See next page for additional authors

Suggested Citation

Kim, Sun-Young; Sheppard, Lianne; Bergen, Silas; Szpiro, Adam A.; Sampson, Paul D.; Kaufman, Joel; and Vedal, Sverre, "Prediction of fine particulate matter chemical components for the Multi-Ethnic Study of Atherosclerosis cohort: A comparison of two modeling approaches" (December 2013). *UW Biostatistics Working Paper Series.* Working Paper 398.
http://biostats.bepress.com/uwbiostat/paper398

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

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

This article is available at Collection of Biostatistics Research Archive: http://biostats.bepress.com/uwbiostat/paper398

1. Introduction

Modern epidemiological studies focusing on the association between long-term exposure to fine particulate matter ($PM_{2.5}$) and health rely on predictions of $PM_{2.5}$, because long-term and representative $PM_{2.5}$ measurements at each individual's actual location are infeasible. Common prediction approaches include assigning average $PM_{2.5}$ concentrations to an administrative unit such as county or city based on agency monitor(s) within that area, assigning measurements based on the agency monitor nearest that participant's home, or applying weighted averages of monitors to participants' locations with weights based on distances (Dockery et al., 1993; Lipsett et al., 2011; Miller et al., 2007). These approaches, however, do not represent well all the spatial variability in the underlying exposure surface; this in turn results in exposure measurement error in the health effect analysis. Recent advances in prediction models have led to better representation of variability in $PM_{2.5}$ concentrations across cohort locations than relatively simple and commonly used approaches. For instance, land use regression models represent variability using geographic variables that affect long-term average $PM_{2.5}$ concentrations (Eeftens et al., 2012; Hoek et al., 2008). More sophisticated spatio-temporal models, based on shorter-term average concentrations over two weeks or a month, characterize spatial and temporal variability using regression and smoothing techniques (Paciorek et al., 2009; Sampson et al., 2011; Szpiro et al., 2011; Yanosky et al., 2009).

Unlike other criteria air pollutants, $PM_{2.5}$ is not a single compound or chemical, but instead is a complex mixture of numerous components, including acids, organic chemicals, metals and soil or dust particles. Study of the health effects of long-term concentrations of $PM_{2.5}$ chemical components has been limited and few prediction models for these components have been developed. While most studies of $PM_{2.5}$ components have investigated associations of

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

short-term concentrations (Bell et al., 2007; Ostro et al., 2007; Peng et al., 2007), a few cohort studies have focused on the effects of long-term $PM_{2.5}$ component exposures. Ostro et al (2010) investigated long-term associations of eight $PM_{2.5}$ components and mortality in the California Teachers Study based on the nearest-monitor prediction approach. Sun et al (2013) adopted area-averaging, nearest-monitor, and inverse-distance-weighting methods to predict four $PM_{2.5}$ components to examine the associations with subclinical atherosclerosis outcomes in the Multi-Ethnic Study of Atherosclerosis (MESA). Because some $PM_{2.5}$ components such as EC and OC are affected largely by local sources such as traffic, it is likely that these simple prediction approaches provide poor predictions at residential locations, particularly when distant monitors were used. A recent study of eight trace elements from $PM_{2.5}$ in twenty European cities for the European Study of Cohorts for Air Pollution Effects (ESCAPE) demonstrated good capacity to represent local-scale spatial variability based on land use regression (De Hoogh et al., 2013).

The National Particle Component and Toxicity (NPACT) study at the University of Washington focused on $PM_{2.5}$ components and investigated the association with cardiovascular outcomes in the MESA cohort (Vedal et al., 2013). This study developed two distinct exposure models to predict $PM_{2.5}$ component concentrations at MESA participant homes. The spatio-temporal model used 2-week average samples of $PM_{2.5}$ component concentrations collected by a cohort-dedicated monitoring campaign. The national spatial model used annual average $PM_{2.5}$ component concentrations from the nation-wide agency monitoring networks sponsored by the U.S. Environmental Protection Agency (EPA) and other agencies. The spatio-temporal model was fit separately in each city, whereas the national spatial model was constructed as a single model over the continental U.S. The national spatial model has been described within the context of measurement error correction in health effect analysis (Bergen et al., 2013).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

This paper describes the spatio-temporal modeling approach to predict long-term concentrations of $PM_{2.5}$ components for NPACT and compares its characteristics to results from the national spatial model for the same cohort. We focused on four $PM_{2.5}$ components: elemental and organic carbon (EC and OC), sulfur, and silicon. These are considered to be markers for combustion-related traffic, secondary process of inorganic aerosol, and airborne crustal matter, respectively.

2. Material and methods

2.1. Data

2.1.1. NPACT/MESA Air monitoring data

The NPACT study obtained $PM_{2.5}$ chemical component measurements from the MESA and Air Pollution (MESA Air) study monitoring campaign (Cohen et al., 2009; Kaufman et al., 2012). This campaign lasted for four years, concentrated on the geographic areas covered by the MESA subject residences, and consisted of both fixed and home outdoor monitoring sites in each of the six MESA city regions: Los Angeles, Chicago, Minneapolis-St. Paul, Baltimore, New York, and Winston-Salem (Figure 1). Three to seven fixed sites operated for the entire study period, whereas approximately 50 rotating home-outdoor sites were sampled in each of two seasons. While the NPACT/MESA Air monitoring sites are located where most MESA participants live, there were very few regulatory monitoring sites near these subjects (Figure 1). Although NPACT sampled for trace elements, including sulfur and silicon, between August 2005 and August 2009, sampling for EC and OC was limited to March 2007 through August 2008. Figure 2 shows the sampling design of fixed and home-outdoor sites in Los Angeles; similar patterns hold for all MESA cities.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

We summarize the sampling and analysis methods here; details can be found in Vedal et al. (2013). The NPACT/MESA Air monitoring campaign collected 2 week samples of $PM_{2.5}$ components using the Harvard Personal Environmental Monitors with a 2.5 um cut size when operated with pump flow rate of 1.8 L/min. Sulfur and silicon were quantified by the X-Ray Fluorescence analysis of Teflon filters. EC and OC were determined by the IMPROVE_A thermal optical reflectance method from quartz filters. All data used in this analysis passed strict data cleaning and quality assurance criteria. In addition, we excluded a few measurements flagged in the quality assurance review to have equipment problems and two unreasonably high silicon measurements possibly contaminated during filter handling. To fit the spatio-temporal model, we additionally excluded a limited number of outlying measurements because those measurements dramatically affected model fitting and evaluation in our preliminary analysis. These exceeded a 2.5 times inter-quartile range distance from temporally and spatially defined quartiles in each city (Vedal et al., 2013). For statistical modeling we added 1 and log-transformed the 2-week average measurements. Silicon was modeled in nanograms per cubic meter whereas other components were in micrograms per cubic meter.

2.1.2. Regulatory monitoring data

There are two nation-wide regulatory monitoring programs for $PM_{2.5}$ components: the U.S. EPA Chemical Speciation Network (CSN) and the Interagency Monitoring of Protected Visual Environments (IMPROVE). The objective of the CSN program is to monitor temporal and spatial distribution of $PM_{2.5}$ components to identify and control potential sources (U.S. EPA, 2004). CSN monitoring sites are located mostly in urban areas and have collected $PM_{2.5}$ components on an every 3rd or 6th day schedule since 1999. The IMPROVE program was established in 1987 to assess and regulate visibility; most monitoring sites, sampling every 3rd

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

day, are deployed in national parks and rural areas (Hand et al., 2011). The sampling and analysis protocols of these two networks were described elsewhere (U.S. EPA, 2004; Hand et al., 2011). We initially planned to combine the CSN and IMPROVE data with the NPACT/MESA Air monitoring data to develop our spatio-temporal models. However, we found that there were important differences between the two networks in their sampling and analysis protocols. These led to inconsistencies in the data that we judged to be too severe to permit combining the data into one unified model (Kim et al., 2013). Instead, we used the NPACT/MESA Air monitoring data for the spatio-temporal model and the CSN and IMPROVE monitoring data for the national-spatial model. For the national spatial model, data was downloaded from the EPA Air Quality System data base for both CSN and IMPROVE for 2009 and 2010. We computed annual averages and square-root transformed these to reduce skewness.

2.1.3. Geocoding and geographic variables

Residential addresses of 7014 MESA and MESA Air participants who consented to use of their addresses were geocoded using TeleAtlas 2000 according to standardized procedures. Geocoded locations of NPACT/MESA Air and agency monitoring sites were obtained from geocoding and hand-held GPS devices, and EPA sources, respectively.

We created more than 800 candidate geographic variables at monitoring and cohort locations (Supplemental Table 1). These variables included population density from the U.S. census, normalized difference vegetative index (NDVI) and impervious surface measurement based on satellite imagery, land cover and elevation from the U.S. Geological Survey, emissions of primary pollutants from the National Emission Inventory, and road variables based on the TeleAtlas road network. We preprocessed these covariates, eliminating those that did not vary across locations, log transforming distance variables, and recoding variables by truncating to

COBRA
A SERVICE OF THE BEPRESS
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

avoid implausible extreme values. After this area-specific data processing, the number of candidate geographic variables in each area ranged between 52 and 116.

2.2. Exposure prediction model

2.2.1. Spatio-temporal model framework

We developed separate models for 2-week average log concentration measurements in each region and for each component. Our spatio-temporal modeling approach was based on the MESA Air study framework, previously described for $PM_{2.5}$ and NOx (Lindstrom et al., 2013; Sampson et al., 2011; Szpiro et al., 2010). The component models relied on much less monitoring data as they were based only on the NPACT/MESA Air monitoring campaign and not supplemented with additional data from the regulatory monitoring network. Thus we used a simplified version of the spatio-temporal model with a single temporal trend characterized by a single geographic covariate and no spatial correlation structure.

The spatio-temporal model represents the log 2-week average component concentration ($\mathrm{C}(\mathrm{s},\mathrm{t})$) in terms of of a long-term mean ($\beta_0(\mathrm{s})$), a temporal trend ($\beta_1(\mathrm{s})f(\mathrm{t})$), and spatio-temporal residuals ($\varepsilon(\mathrm{s},\mathrm{t})$), shown in the equation below.

$$\mathrm{C}(\mathrm{s},\mathrm{t}) = \beta_0(\mathrm{s}) + \beta_1(\mathrm{s})f(\mathrm{t}) + \varepsilon(\mathrm{s},\mathrm{t})$$

$$\beta_0(\mathrm{s}) \sim \left(\alpha_{00} + \sum_{j=1}^{j=m} \alpha_{j0} X_{j0}(s) \, , \Sigma(\phi_0, \sigma_0^2, \tau_0^2) \right)$$

$$\beta_1(\mathrm{s}) \sim (\alpha_{01} + \alpha_{11} X_{11}(s), \ \tau_1^2)$$

$$\varepsilon(\mathrm{s},\mathrm{t}) \sim \left(0, \Sigma(\phi_\varepsilon, \sigma_\varepsilon^2, \tau_\varepsilon^2)\right)$$

The long-term mean and temporal trend vary spatially with a trend coefficient ($\beta_1(\mathrm{s})$) scaling the spatially-constant temporal basis function ($f(\mathrm{t})$). The temporal basis function was estimated by smoothing the first temporal component of a singular value decomposition (SVD) of the space-

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics

http://biostats.bepress.com/uwbiostat/paper398

time monitoring data matrix. The long-term mean was characterized by a universal kriging model with a land use regression mean model and spatial correlation modeled with an exponential covariance function (Banerjee et al. 2004). The covariance function had parameters for the range (ϕ), partial sill (σ^2), and nugget (τ^2) which represent the spatial correlation distance, spatial variability, and non-spatial variability, respectively. Geographic covariates (X) were selected from a subset identified by the least absolute shrinkage and selection operator (lasso) (Tibshirani 1996) followed by an exhaustive search. The spatially varying trend coefficient was modeled by the one geographic variable most associated with the trend coefficient; its variance model had no spatial structure (i.e. zero range and partial sill). The spatio-temporal residual field was assumed to be temporally independent with mean zero and spatially correlated with an exponential covariance model.

2.2.2. Spatio-temporal model fitting and prediction procedure

Estimation of the temporal basis function was restricted to the $PM_{2.5}$ component data at fixed sites. To determine the set of geographic variables to be included in the long-term mean, we performed the variable selection using data from home outdoor sites for provisionally-computed long-term averages after removing a temporal trend. The geographic variables were rescaled to have common mean and unit variance. We selected twelve candidate variables from the lasso and then chose the final set of up to five (for sulfur and silicon) or four (for EC and OC) based on 5-fold cross-validated R^2 in an exhaustive search. Given the estimated temporal basis function, selected geographic variables and monitoring data, we estimated regression and covariance parameters. For the model evaluation, we performed 10-fold cross-validation for 2 week average measurements across home-outdoor sites and computed summary statistics such as mean square error (MSE) and R^2. To focus on the spatial prediction ability of our spatio-

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

temporal models, we computed temporally-adjusted R^2 statistic adjusting for temporal variability in addition to the usual (unadjusted) R^2. The temporally-adjusted R^2 accounted for temporal variability using either an estimated trend based on fixed sites or spatial averages of fixed sites at each time (Vedal et al., 2013).

We predicted log 2-week average concentrations at participant addresses conditional on the estimated spatio-temporal model parameters and geographic covariates. These were exponentiated and 1 was subtracted to obtain 2-week predictions on the native scale. In addition we computed the unit of silicon back to the original microgram per cubic meter units. We restricted the prediction area to participants living within 10 kilometers of any NPACT/MESA Air monitors to avoid extrapolation. In addition, we excluded a few extremely high or low predictions at addresses where covariate values for a particular geographic variable were far outside the range of covariate values across monitoring locations. Finally, we averaged the 2-week average predicted concentrations for one year from May 2007 to April 2008 when all four component data are available. Our spatio-temporal models were implemented in the R package `SpatioTemporal` on the Comprehensive R Archive Network (CRAN) (Lindstrom et al., 2013).

2.2.2. National Spatial model

We briefly summarize the national spatial modeling approach based on annual averages of $PM_{2.5}$ component concentrations from the CSN and IMPROVE monitoring network; for more detail see Bergen et al. (2013). Instead of variable selection, this model adopted partial least squares (PLS) to handle the large amount of collinear geographic variables. The PLS method finds the linear combinations of geographic variables, called PLS scores, that are most correlated with the long-term concentrations of $PM_{2.5}$ components (Abdi, 2003; Sampson et al., 2011;

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

Sampson et al., 2013). The first few PLS scores were used to characterize the mean structure in a universal kriging model. A 10-fold cross-validation procedure was implemented to determine the number of PLS scores. Two PLS scores were selected for all components except for EC with three. Given selected PLS scores, we estimated regression coefficients and covariance parameters. While the national spatial model was developed over the entire U.S. in NPACT, we evaluated it in the MESA areas by restricting the comparison between observations and cross-validated predictions to CSN and IMPROVE monitoring sites within 200 kilometers of the centers of the six MESA cities. Finally, we predicted annual average concentrations for the $PM_{2.5}$ components at MESA participant addresses in the same prediction area (i.e. within 10 kilometers of any NPACT/MESA Air monitors) and back transformed these to the original microgram per cubic meter units.

3. Results

3.1. NPACT/MESA Air monitoring data

Table 1 shows the summary statistics of 2 week concentrations for four $PM_{2.5}$ components in each of the six MESA regions from the NPACT/MESA Air monitoring network. Sulfur concentrations were high in the cities on the East Coast, while those of EC were high in highly-populated cities such as Los Angeles and New York. Silicon concentrations were high in Los Angeles as expected given the dry climate contributing to kicking up dust.

3.2. Spatio-temporal model fitting

3.2.1. Trend estimation

Figure 3 shows the computed SVD and trend function for log-transformed $PM_{2.5}$ components in Los Angeles. The results for the other five cities are shown in Supplemental Figure 1. Sulfur generally showed a clear seasonal pattern in all six cities. In Los Angeles, 2-

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

week average sulfur concentrations were higher in summer and lower concentrations in winter. There were reverse patterns in Baltimore, New York, and Winston-Salem on the East Coast. (Supplemental Figure 1). EC was higher in summer whereas OC was higher in winter in Los Angeles.

3.2.2. Variable selection

Table 2 and Supplemental Table 2 gives the classes of geographic variables included in the final selected models for each component and area from the potential variables described in Supplemental Table 1. For most pollutants and areas, the final models included traffic variables and urban and rural land use characteristics; inclusion of geographic coordinates, distances to sources, emission variables, vegetation, imperviousness, and elevation varied across $PM_{2.5}$ components and areas. Vegetation index was selected only in St. Paul, and Baltimore, and impervious surface measurement was chosen only in St. Paul. The variable selection R^2s from our cross-validation approach using selected variables for the regression of "long-term average" $PM_{2.5}$ component concentrations are also shown in Table 2. They were generally higher in all areas for EC and OC than for sulfur and silicon. Sulfur and silicon in St. Paul as well as New York and sulfur in Baltimore showed cross-validated variable selection R^2 lower than 0.2, possibly due to our conservative approach computing R^2 statistics, less spatial variability of sulfur and silicon, or absence of important geographic variables.

3.2.3. Parameter estimation

The parameter estimates for the regression coefficients and variance model parameters in Los Angeles and other five cities are shown in Figure 4 and Supplemental Figure 3, respectively. Los Angles and Chicago tended to show larger range and partial sill representing stronger spatial

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

correlation structure than other areas. In general, the estimated regression parameters for EC and OC were significantly different from zero, whereas those for silicon and sulfur were not.

3.3. Features of spatio-temporal and national spatial models

3.3.1. Model evaluation

Table 3, Figure 5, and Supplemental Figure 4 show statistics and scatter plots for cross-validated predictions of 2-week concentrations from the city-specific spatio-temporal model across MESA home-outdoor sites and cross-validated predictions of annual averages from the national spatial model across the CSN/IMPROVE sites in the MESA areas. Not surprisingly, in the spatio-temporal predictions many of the temporally-adjusted R^2s, with adjustment based on either the estimated trend or spatial means of fixed sites, were much lower than the unadjusted R^2s. Across all areas, the temporally-adjusted R^2s, particularly when spatial averages were used, were generally higher for EC and OC—for which variation is primarily spatial rather than temporal—than for sulfur and silicon. Los Angeles and Baltimore gave higher temporally-adjusted R^2 than other cities. Temporally-adjusted R^2s for sulfur, silicon, EC and OC across all six cities combined were 0.84, 0.38, 0.79, and 0.59, respectively. These MESA-wide statistics were generally higher than the city-specific temporally-adjusted R^2s, because the contribution of between-city variability to this estimate is larger than that of within-city variability. R^2s for sulfur, silicon, EC and OC in the national spatial model were 0.94, 0.45, 0.70, and 0.79, respectively.

3.3.2. Predicted long-term $PM_{2.5}$ component concentrations

The city-specific summaries and spatial distributions of predicted long-term $PM_{2.5}$ component concentrations at MESA Air participant homes varied by component and city (Table 4, Figure 6). Figure 7 and Supplemental Figure 5 display predicted long-term average

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

concentrations of sulfur, silicon, EC, and OC in each city. Predicted concentrations were generally higher from the spatio-temporal model than from the national spatial model. In addition, predictions generally varied more between cities than within each city (Figure 6). Whereas the degree of correlation between the two kinds of predictions differed by city, predictions from one model were positively correlated with those from the other across cities for all components with much lower correlation for OC than other components (correlation coefficient=0.91, 0.55, 0.82, and 0.19 for sulfur, silicon, EC, and OC, respectively). Correlations of $PM_{2.5}$ component predictions from the two models across cities were higher than those within cities (Figure 6).

4. Discussion

This study developed two exposure prediction modeling approaches to obtain long-term average residential concentrations of four $PM_{2.5}$ chemical components at participant addresses, specifically for epidemiological study application. Spatio-temporal and national spatial models were developed based on different monitoring data and modeling approaches. We, however, found generally consistent model performance across the six MESA cities driven by the large between-city variability of $PM_{2.5}$ components; predicted long-term concentrations of $PM_{2.5}$ components from the two models were fairly or highly correlated across cities. In contrast, the predictions are less highly correlated within each city.

We developed rich exposure prediction models in order to reduce measurement error in predicted individual-level concentrations in order to provide more valid and precise health effect estimates. To our knowledge, this study is the one of a few studies focusing on the development of exposure prediction approaches for $PM_{2.5}$ components. Previous cohort studies assessed health effects of long-term $PM_{2.5}$ component concentrations using relatively simple prediction

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

approaches in representing spatial distribution. Early cohort studies based on the Harvard Six City cohort and the American Cancer Society (ACS) cohort predicted sulfate, as a component of $PM_{2.5}$, using an area-averaging approach (Dockery et al., 1993; Pope et al., 1995). Ostro et al (2010) used the nearest-monitor method for eight $PM_{2.5}$ components in the California Teaches study cohort. These approaches, however, could have high exposure measurement error over space given spatially-limited regulatory monitoring networks which do not represent fine-scale spatial heterogeneity of $PM_{2.5}$ components, particularly for those strongly affected by local sources. This measurement error could then affect inference in the health effect analysis. Jerrett et al. (2005) found that estimated relative risk of mortality was about three times higher in the ACS cohort in California when long-term $PM_{2.5}$ concentrations were predicted using kriging in comparison to the area-averaging approach. We have shown by simulation that nearest-monitor predictions give more biased health effect estimates than kriging when the underlying pollution field has spatial structure (Kim et al., 2009). Sun et al. (2013) investigated the association of $PM_{2.5}$ components and subclinical atherosclerotic outcomes using area-averaging, nearest-monitor, and inverse-distance-weighting approaches based on the same NPACT/MESA Air monitoring data used in our spatio-temporal model. Supplemental Figure 6 showed that these predictions were highly correlated with predictions from the sptaio-temporal model across cities but present little or no within-city variability. The correlations with the predictions from the spatio-temporal model were higher than those from the national spatial model, most likely due to their reliance on the same monitoring data. Recently, De Hoogh et al. (2013) adopted land use regression on long-term concentrations of eight trace elements of $PM_{2.5}$ in the ESCAPE study. This study deployed 20 monitoring sites in each of 20 European cities and collected three 2-week samples over one year period to represent within-city spatial variability. They additionally

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

located one reference site for continuous 2-week sampling for the year to represent a temporal trend in each city. Using long-term averages, estimated by adjusting for a temporal trend, they fitted the land use regression using covariates chosen by a supervised stepwise selection across 20 monitoring sites in each city. Their approach is similar to our provisional approach for variable selection in the spatio-temporal modeling procedure. However, their city-specific cross-validated R^2s for sulfur and silicon were generally higher than our cross-validated R^2s. This difference may be their approach to compute R^2 statistic. Their R^2 was computed based on the leave-one-out cross-validation which can overestimate model performance particularly given a small number of training sites (Wang et al., 2012).

We chose highly conservative approaches in evaluating our two exposure prediction models to avoid overestimating model performance. One of the common evaluation approaches in land use regression studies is the leave-one-out cross-validation which fits a model for the data leaving out one site and predicts at the left-out site in sequence, and then compares predictions to observations (Hoek et al., 2008). However, this approach was overly optimistic for model performance when the number of sites was limited (Wang et al., 2012; Wang et al., 2013). Given about a hundred home-outdoor sites, our cross-validation was based on the 5 or 10 group cross-validation. In addition, we computed MSE-based R^2 by subtracting mean square prediction error relative to data variability from 1 as opposed to model-based R^2 calculated by the squared correlation coefficient. The model-based R^2 tended to overestimate prediction ability because observations are compared to predictions based on the regression line instead of the identity line as in MSE-based R^2 (Wang et al., 2012). Our evaluation approach using MSE-based R^2 in the 5- or 10-fold cross-validation was likely to provide more reasonable but lower R^2s than those reported in other studies.

COBRA
BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

Within-city predicted concentrations of $PM_{2.5}$ components were generally higher from the spatio-temporal model than those from the national spatial model. Features contributing to this within-city difference between the two exposure prediction models include the data sources, modeling approaches, and evaluation methods. Whereas the spatio-temporal model was developed based on the NAPCT monitoring data, the national spatial model relied on the CSN/IMPROVE monitoring data. These two data sources had distinct sampling schedules, used different sampling equipment, and were mostly sampled at non-overlapping locations (Kim et al., 2013). Higher predictions for sulfur, silicon, and EC in the spatio-temporal model compared to those in the national spatial model correspond to higher measurements of sulfur, silicon, and EC at NPACT/MESA Air monitoring sites relative to those at CSN/IMPROVE sites (Supplemental Figure 6). In addition, operationally the land use information was incorporated differently in the two models. The spatio-temporal model relied on variable selection techniques to choose a subset of geographic variables. In the national spatial model we reduced the dimension of the covariate data with PLS and included only a few of the resulting scores. We note that the spatio-temporal model variable selection was based on detrended provisional "long-term averages"; these were quite uncertain and thus limited our confidence in applying the PLS approach. Lastly, we devised a temporally-adjusted R^2 to evaluate spatial prediction ability for the spatio-temporal model. In the national spatial model our use of long-term averages removed all temporal variability so the traditional R^2 only represents spatial performance. All these fundamental differences between the two prediction models make it difficult to directly compare their performance statistics and conclude that one model is preferable to the other.

Despite different data sources and modeling characteristics, the two exposure models also showed consistent features across pollutants and similar ordering of predicted concentrations in

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

all cities. The two prediction models presented relatively strong mean structures for EC and OC and prominent spatial dependence structure for sulfur and silicon. The proportion of the variability represented by the long-term mean model was larger for EC and OC than for sulfur and silicon in the spatio-temporal model (Supplemental Table 2). Similarly, the regression part of the national spatial model explained most of the variability for EC and OC while the covariance structure characterized in the kriging part of the model was important only for sulfur and silicon (Bergen et al., 2013). Predicted concentrations from the spatio-temporal model were higher in some cities than others; similar patterns applied for predictions from the national spatial models for all components except for OC.

This study includes some limitations and implications for future studies. In the NPACT study, while we originally intended to characterize within-city distribution of $PM_{2.5}$ components based on the dedicated monitoring campaign for the target cohort combined with additional regulatory data, our preliminary exploratory analysis led us to limit our analysis to only the NPACT data given its incompatibility with CSN/IMPROVE data (Kim et al., 2013). The simplified spatio-temporal model and limited monitoring data did not allow us to represent all the spatial variation within each city and may affect the ensuing health effect analyses. In addition, we focused on the four $PM_{2.5}$ components which are considered as least ambiguous markers for pollution sources of our interest. We plan to expand our modeling approaches to other components which are also treated as being strongly related to specific pollution sources.

5. Conclusions

We described two modeling approaches for predicting long-term concentrations of $PM_{2.5}$ components. Both performed reasonably well across cities. Predictions were generally consistent across the six study areas except for organic carbon; consistency was relatively weak within each

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

city. These predictions of $PM_{2.5}$ components allow us to assess associations of long-term exposure to $PM_{2.5}$ components and health.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

References

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

Banerjee S., Carlin B.P., Gelfand A.E., 2004. Hierarchical Modeling and Analysis for Spatial Data. Boca Raton, FL: Chapman & Hall/CRC Press. 21-68.

Bell M.L., Dominici F., Ebisu K., Zeger S.L., Samet J.M., 2007. Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies. Environ Health Perspect. 115, 989-995.

Bergen S., Sheppard L., Sampson P.D., Kim S.Y., Richards M., Vedal S., Kaufman J.D., Szpiro A.A., 2013. A national prediction model for components of PM2:5 and measurement error corrected health effect inference. [published online ahead of print June 11, 2013]. Environ Health Perspect (doi:10.1289/ehp.1206010).

Bild D.E., Bluemke D.A., Burke G.L., Detrano R., Diez Roux A.V., Folsom A.R., Greenland P., Jacob D.R. Jr, Kronmal R., Liu K., Nelson J.C., O'Leary D., Saad M.F., Shea S., Szklo M., Tracy R.P., 2002. Multi-ethnic study of atherosclerosis: objectives and design. Am J Epidemiol. 156, 871-881.

Cohen M.A., Adar S.D., Allen R.W., Avol E., Curl C.L., Gould T, Hardie D., Ho A., Kinney P., Larson T.V., Sampson P., Sheppard L., Stukovsky K.D., Swan S.S., Liu L.J., Kaufman J.D., 2009. Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Environ Sci Technol 43, 4687-4693.

de Hoogh K., Wang M., Adam M., Badaloni C., Beelen R., Birk M., Cesaroni G., Cirach M., Declercq C., Dėdelė A., Dons E., de Nazelle A., Eeftens M., Eriksen K., Eriksson C., Fischer P., Gražulevičienė R., Gryparis A., Hoffmann B., Jerrett M., Katsouyanni K., Iakovides M., Lanki T., Lindley S., Madsen C., Mölter A., Mosler G., Nádor G., Nieuwenhuijsen M., Pershagen G., Peters A., Phuleria H., Probst-Hensch N., Raaschou-Nielsen O., Quass U., Ranzi A., Stephanou E., Sugiri D., Schwarze P., Tsai M.Y., Yli-Tuomi T., Varró M.J., Vienneau D., Weinmayr G., Brunekreef B., Hoek G., 2013. Development of land use regression models for particle composition in twenty study areas in Europe. Environ Sci Technol. 47, 5778-5786.

Dockery D.W., Pope 3rd C.A., Xu X., Spengler J.D., Ware J.H., Fay M.E., Ferris B.G.,Speizer F.E., 1993. An association between air pollution and mortality in six U.S. cities. N Engl J Med. 329, 1753-1759.

Hand JL, Copeland SA, Day DE, Dillner AM, Indresand H, Malm WC, McDade CE, Moore CT, Pitchford ML, Schichte BA, Watson JG. Spatial and Seasonal Patterns and Temporal Variability of Haze and its Constituents in the United States: Report V. 2011. June (http://vista.cira.colostate.edu/improve/publications/Reports/2011/PDF/Cover_TOC.pdf).

Research Archive

http://biostats.bepress.com/uwbiostat/paper398

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. Atmos Environ. 42, 7561-7578.

Jerrett M., Burnett R.T., Ma R., Pope 3rd C.A., Krewski D., Newbold K.B., Thurston G., Shi Y., Finkelstein N., Calle E.E., Thun M.J., 2005. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology. 16, 727-36.

Kaufman J.D., Adar S.D., Allen R.W., Barr R.G., Budoff M.J., Burke G.L., Casillas A.M., Cohen M.A., Curl C.L., Daviglus M.L., Diez Roux A.V., Jacobs D.R. Jr, Kronmal R.A., Larson T.V., Liu S.L., Lumley T., Navas-Acien A., O'Leary D.H., Rotter J.I., Sampson P.D., Sheppard L., Siscovick D.S., Stein J.H., Szpiro A.A., Tracy R.P., 2012. Prospective Study of Particulate Air Pollution Exposures, Subclinical Atherosclerosis, and Clinical Cardiovascular Disease: The Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Am J Epidemiol. 176, 825-837.

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

Kim S.Y., Sheppard L., Larson T., Vedal S., 2013. Issues related to combining multiple speciated $PM_{2.5}$ data sources in spatio-temporal exposure models for epidemiology: The NPACT case study. UW Biostatistics Working Paper Series. XXX.

Lindstrom J., Szpiro A.A., Sampson P.D., Bergen S., Oron A., 2013a. SpatioTemporal: An R package for spatio-temporal modelling of air-pollution. J stat softw (in press). (http://cran.r-project.org/web/packages/SpatioTemporal/index.html).

Lindstrom J., Szpiro A.A., Sampson P.D., Oron A.P., Richards M., Larson T.V., Sheppard L., 2013b. A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. Environ Ecol Stat (in press).

Lipsett M.J., Ostro B.D., Reynolds P., Goldberg D., Hertz A., Jerrett M., Smith D.F., Garcia C., Chang E.T., Bernstein L., 2011. Long-term exposure to air pollution and cardiorespiratory disease in the California teachers study cohort. Am J Respir Crit Care Med. 184, 828-835.

Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, Kaufman JD. Long-term exposure to air pollution and incidence of cardiovascular events in women. *N Engl J Med*. 2007. 356(5):447-58.

Ostro B., Lipsett M., Reynolds P., Goldberg D., Hertz A., Garcia C., Henderson K.D., Bernstein L., 2010. Assessing Long-Term Exposure in the California Teachers Study. Environ Health Perspect. 118, 363-369.

Ostro B., Burnett R.T., Shin H., Hughes E., Garcia C., Henderson K.D., Bernstein L., Lipsett M., 2011. Assessing Long-Term Exposure in the California Teachers Study. Environ Health Perspect. 119, A242–A243 Erratum.

Hosted by The Berkeley Electronic Press

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. Ann Appl Stat. 3, 370–397.

Peng R.D., Bell M.L., Geyh A.S., McDermott A., Zeger S.L., Samet J.M., Dominici F., 2009. Emergency admissions of cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. Environ Health Perspect 117:957-963.

Pope 3rd C.A., Thun M.J., Namboodiri M.M., Dockery D.W., Evans J.S., Speizer F.E., Heath Jr C.W., 1995. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Car Med 151, 669-674.

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

Sampson P.D., Richards M., Szpiro, A.A., Bergen S., Sheppard L., Larson T.V., Kaufman J.D., 2013. A regionalized national universal kriging model using partial least squares regression for estimating annual PM2.5 concentrations in epidemiology. [published online ahead of print April 25, 2013]. Atmos environ. (doi: 10.1016/j.atmosenv.2013.04.015).

Sun M., Kaufman J.D., Kim S.Y., Larson T.V., Gould T.R., Polak J.F., Budoff M.J., Diez Roux A.V., Vedal S., 2013. Particulate matter components and subclinical atherosclerosis: common approaches to estimating exposure in a Multi-Ethnic Study of Atherosclerosis cross-sectional study. [published online ahead of print May 3, 2013]. Environ res. (doi: 10.1186/1476-069X-12-39).

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.

Tibshirani R., 1996. Regression shrinkage and selection via the Lasso. J Royal Stat Soc Series B. 58, 267-288.

U.S. EPA. 2004. Air quality criteria for particulate matter (Report No. EPA 600/P-99/002aF-bF): Volume 1. U.S. Environmental Protection Agency. Washington, DC. 1-9, 2-93-2-94.

Vedal S., Kim S-Y., Miller K.A., Fox J.R., Bergen S., Gould T., Kaufman J.D., Larson T.V., Sampson P.D., Sheppard E.A., Simpson C.D., Szpiro A.A., 2013. NPACT epidemiologic study of components of fine particulate matter and cardiovascular disease in the MESA and WHI-OS cohorts. Research Report 178. Health Effects Institute, Boston, MA (in press).

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

COBRA
Research Archive

Wang M., Beelen R., Eeftens M., Meliefste K., Hoek G., Brunekreef B., 2012. Systematic evaluation of land use regression models for NO_2. Environ Sci and Technol. 46, 4481-4489.

Wang M, Beelen R, Basagana X, Becker T, Cesaroni G, de Hoogh K, Dedele A, Declercq C, Dimakopoulou K, Eeftens M, Forastiere F, Galassi C, Gražulevičienė R, Hoffmann B, Heinrich J, Iakovides M, Künzli N, Korek M, Lindley S, Mölter A, Mosler G, Madsen C, Nieuwenhuijsen M, Phuleria H, Pedeli X, Raaschou-Nielsen O, Ranzi A, Stephanou E, Sugiri D, Stempfelet M, Tsai MY, Lanki T, Udvardy O, Varró MJ, Wolf K, Weinmayr G, Yli-Tuomi T, Hoek G, Brunekreef B. 2013. Evaluation of land use regression models for NO2 and particulate matter in 20 European study areas: the ESCAPE project. Environ Sci Technol. 47:4357-4364.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 1. Summary statistics of 2 week concentrations of four $PM_{2.5}$ components in the NPACT/MESA Air monitoring network

City	Type	Sulfur				Silicon				EC				OC			
		N of sites	N of samples	Mean	(SD)	N of sites	N of samples	Mean	(SD)	N of sites	N of samples	Mean	(SD)	N of sites	N of samples	Mean	(SD)
Los Angeles	Fixed	7	535	1.15	(0.59)	7	536	0.16	(0.08)	7	200	1.81	(0.79)	7	200	2.15	(1.06)
	Home	89	153	1.08	(0.62)	108	172	0.15	(0.08)	70	88	1.79	(0.87)	70	87	2.24	(1.03)
Chicago	Fixed	5	375	1.12	(0.44)	5	374	0.11	(0.04)	5	138	1.38	(0.39)	5	138	1.81	(0.63)
	Home	104	187	1.09	(0.36)	89	152	0.10	(0.06)	50	80	1.27	(0.32)	50	82	1.88	(0.62)
St. Paul	Fixed	3	257	0.73	(0.23)	3	256	0.11	(0.05)	3	93	0.87	(0.23)	3	95	1.71	(0.37)
	Home	104	187	0.70	(0.23)	104	187	0.11	(0.04)	54	89	0.79	(0.21)	54	90	1.70	(0.40)
Baltimore	Fixed	4	331	1.53	(0.62)	4	329	0.09	(0.04)	4	133	1.45	(0.52)	4	133	2.18	(0.71)
	Home	85	156	1.73	(0.67)	85	156	0.09	(0.05)	61	99	1.23	(0.35)	61	99	2.19	(0.89)
New York	Fixed	3	191	1.34	(0.56)	3	191	0.11	(0.05)	3	80	2.22	(0.93)	3	81	1.84	(0.74)
	Home	107	190	1.38	(0.57)	105	186	0.10	(0.05)	49	78	1.83	(0.77)	49	81	2.09	(0.71)
Winston-Salem	Fixed	4	352	1.51	(0.75)	4	352	0.09	(0.05)	4	105	1.07	(0.24)	4	105	2.55	(0.69)
	Home	92	177	1.71	(0.72)	92	177	0.11	(0.05)	47	84	1.05	(0.27)	48	86	2.75	(0.79)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

Table 2. Provisional cross-validation statistics and selected variables from trend-adjusted "long-term average" concentrations at home-outdoor sites

		Cross-validation statistics [a]		Geographic variables [b]									
City	Pollutant	MSE	R^2	Traffic	Land use (urban)	Land use (rural)	Position	Source	Emission	Vegetation	Imperviousness	Elevation	Residual oil [c]
LA	Sulfur	0.030	0.21	■	■			■				■	
	Silicon	0.001	0.47	■		■							
	EC	0.060	0.80	■	■	■						■	
	OC	0.456	0.18		■	■						■	
Chicago	Sulfur	0.025	0.35	■		■		■					
	Silicon	0.001	0.19	■	■	■							
	EC	0.031	0.48	■		■		■					
	OC	0.123	0.47	■	■			■					
St. Paul	Sulfur	0.007	0.17	■				■		■	■	■	
	Silicon	0.001	0.10		■	■				■			
	EC	0.014	0.39	■	■					■			
	OC	0.016	0.60	■	■	■							
Baltimore	Sulfur	0.023	0.12		■	■							
	Silicon	0.000	0.59				■	■		■		■	
	EC	0.030	0.58	■		■						■	
	OC	0.041	0.70	■		■			■	■			
NY	Sulfur	0.070	0.13		■	■						■	
	Silicon	0.002	0.04									■	■
	EC	0.356	0.55		■	■							
	OC	0.163	0.44	■	■	■							
Winston -Salem	Sulfur	0.033	0.25		■	■							
	Silicon	0.000	0.21	■			■					■	
	EC	0.018	0.43		■	■						■	

Hosted by The Berkeley Electronic Press

OC	0.128	0.19			

a. Provisional cross-validation approach based on lasso variable selection followed by all subset universal kriging
b. List of geographic variables for each category is shown in Supplemental Table 1
c. Considered only for New York

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

Table 3. Cross-validation statistics of predicted concentrations of four $PM_{2.5}$ components between spatio-temporal and national spatial models in six MESA Air areas

City	Pollutant	Spatio-temporal model [a]				National spatial model [b]	
		MSE	R^2	Temporally-adjusted R^2 [c]		MSE	R^2
				Estimated trend	Average		
LA	Sulfur	0.013	0.97	0.77	0.35		
	Silicon	0.001	0.68	0.66	0.49		
	EC	0.246	0.73	0.54	0.51		
	OC	0.354	0.67	0.49	0.37		
Chicago	Sulfur	0.035	0.74	0.54	0.15		
	Silicon	0.002	0.35	0.07	0.00		
	EC	0.031	0.69	0.51	0.49		
	OC	0.198	0.48	0.20	0.20		
Minneapolis-St. Paul	Sulfur	0.003	0.94	0.78	0.59		
	Silicon	0.001	0.65	0.39	0.19		
	EC	0.019	0.57	0.33	0.32		
	OC	0.023	0.85	0.47	0.46		
Baltimore	Sulfur	0.016	0.96	0.77	0.48		
	Silicon	0.000	0.82	0.58	0.35		
	EC	0.047	0.62	0.56	0.59		
	OC	0.111	0.86	0.36	0.35		
NY#	Sulfur	0.093	0.71	0.12	0.00		
	Silicon	0.001	0.33	0.27	0.36		
	EC	0.422	0.15	0.58	0.52		
	OC	0.229	0.46	0.63	0.57		
Winston-Salem	Sulfur	0.056	0.89	0.41	0.09		
	Silicon	0.001	0.77	0.14	0.04		
	EC	0.038	0.48	0.19	0.18		
	OC	0.179	0.72	0.13	0.12		
Overall [d]	Sulfur	0.035	0.92	0.84	0.82	0.003	0.94
	Silicon	0.001	0.61	0.38	0.28	0.002	0.45

Hosted by The Berkeley Electronic Press

EC	0.099	0.75	0.79	0.79	0.044	0.70
OC	0.176	0.75	0.59	0.55	0.078	0.79

a. City-specific model using 2 week concentrations on log scale
b. Nation-wide model using annual average concentrations on square root scale
c. Adjusted temporal trend was defined by two approaches which are unsmoothed temporal trend estimated using measurements across NAPCT fixed sites and average of measurements across fixed sites at each time
d. Evaluation of the national spatial model was restricted to EPA monitors within 200 kilometers from the centers of six MESA cities; city-specific evaluation was not carried out given limited numbers of EPA monitors in each city area

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

Table 4. Area-specific summary statistics for the spatio-temporal and national spatial models of predicted long-term average concentrations of four $PM_{2.5}$ components for 5,493 MESA Air participants residing within 10 kilometers of any MESA Air monitor based on addresses for 2000-2002

			Spatio-temporal model		National spatial model	
Pollutant	City	N	Mean	SD	Mean	SD
Sulfur	LA	1,073	1.13	0.04	0.53	0.02
	Chicago	999	1.24	0.06	0.75	0.03
	Minneapolis-St. Paul	898	0.81	0.01	0.53	0.01
	Baltimore	775	1.67	0.06	1.00	0.03
	NY	856	1.46	0.13	0.78	0.02
	Winston-Salem	892	1.69	0.08	0.94	0.02
Silicon	LA	1,073	0.14	0.02	0.14	0.00
	Chicago	999	0.11	0.01	0.06	0.00
	Minneapolis-St. Paul	898	0.10	0.01	0.08	0.00
	Baltimore	775	0.09	0.01	0.08	0.00
	NY	856	0.12	0.01	0.07	0.00
	Winston-Salem	892	0.10	0.01	0.11	0.00
EC	LA	1,073	1.98	0.34	0.83	0.14
	Chicago	999	1.40	0.20	0.74	0.13
	Minneapolis-St. Paul	898	0.85	0.08	0.58	0.07
	Baltimore	775	1.35	0.21	0.68	0.15
	NY	856	2.38	0.44	1.13	0.13
	Winston-Salem	892	1.10	0.08	0.57	0.07
OC	LA	1,073	2.33	0.29	2.43	0.20
	Chicago	999	1.92	0.28	1.71	0.23
	Minneapolis-St. Paul	898	1.73	0.13	2.09	0.19
	Baltimore	775	2.19	0.38	2.12	0.24
	NY	856	2.18	0.39	1.64	0.20
	Winston-Salem	892	2.63	0.16	1.95	0.29

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Figure Captions

Figure 1. Administrative and MESA Air/ NPACT monitoring sites for $PM_{2.5}$ components located within 200 km from city centers in 6 city areas

Figure 2. Temporal and spatial sampling design for silicon and EC for administrative and NPACT/MESA Air monitors in Los Angeles

Figure 3. Estimated smooth temporal trend for four log-transformed $PM_{2.5}$ components in Los Angeles

Figure 4. Estimated parameters for the selected geographical variables (scaled) and covariance structure in the spatio-temporal model for the four log-transformed $PM_{2.5}$ components in Los Angeles

Figure 5. Component-specific scatter plots of observations and cross-validated predictions from the spatio-temporal model for 2-week average concentrations (top) and for 2-week average concentrations after accounting for temporal variability (bottom) across home-outdoor sites in Los Angeles

Figure 6. Component-specific scatter plots and box plots for spatio-temporal and national spatial model predictions of long-term average concentrations across six MESA city areas

Figure 7. Predicted long-term average concentrations from the spatio-temporal and national spatial models of four $PM_{2.5}$ components at participant locations in Los Angeles (different colors represent quintiles of the range of concentrations)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

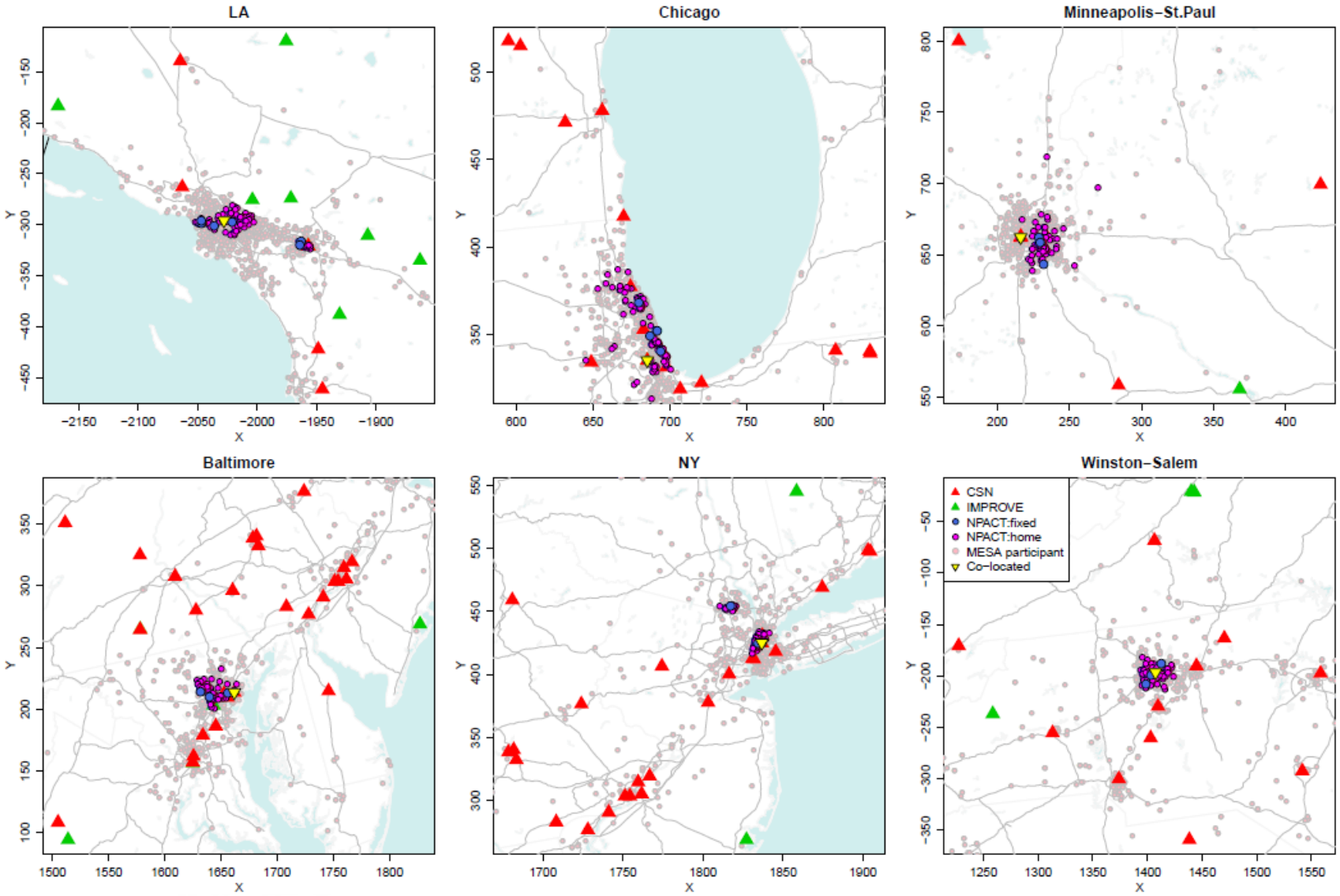


Figure 1. Administrative and MESA Air/ NPACT monitoring sites for $PM_{2.5}$ components located within 200 km from city centers in 6 city areas

Research Archive

Hosted by The Berkeley Electronic Press

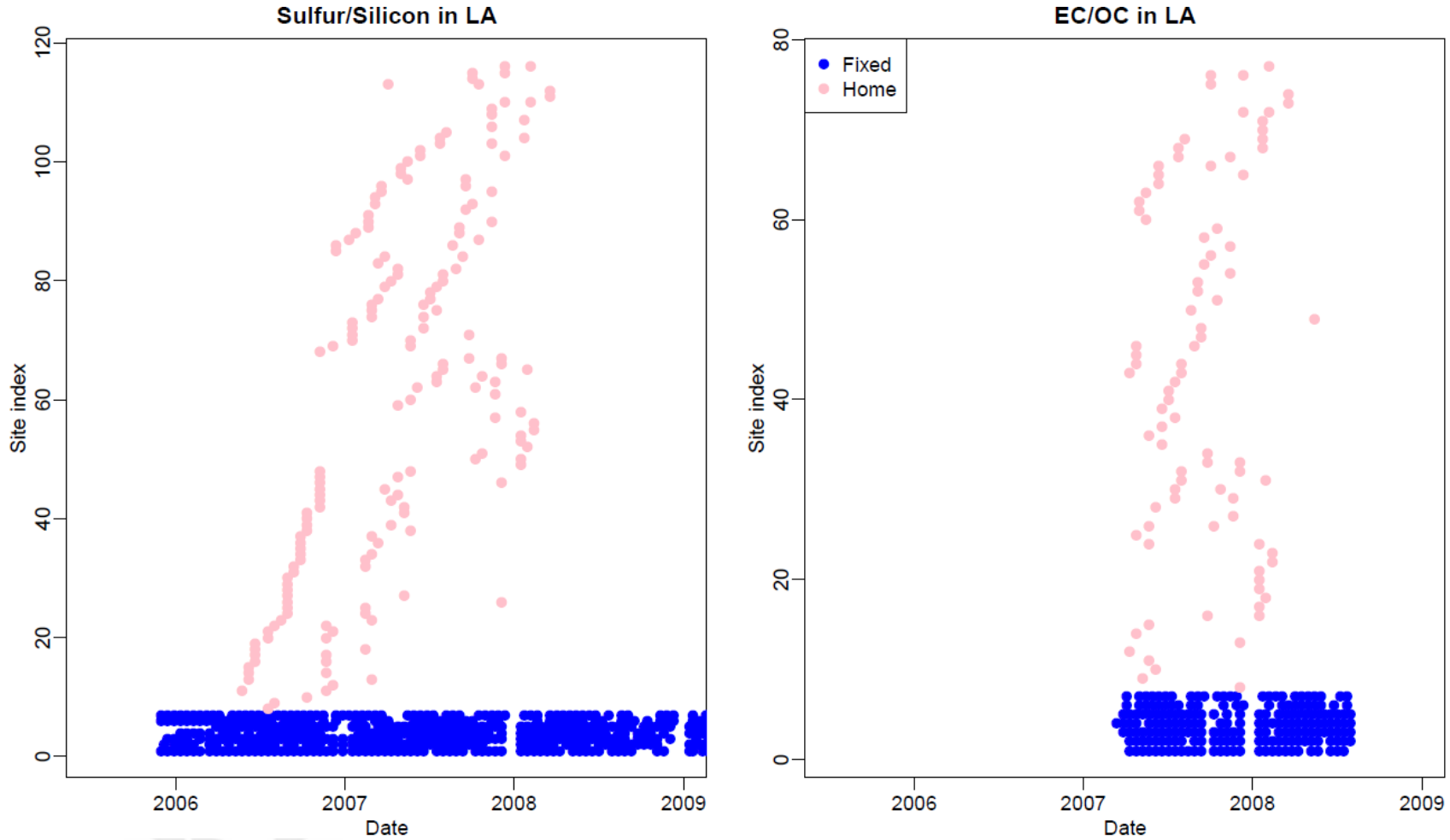


Figure 2. Temporal and spatial sampling design for silicon and EC for administrative and NPACT/MESA Air monitors in Los Angeles

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

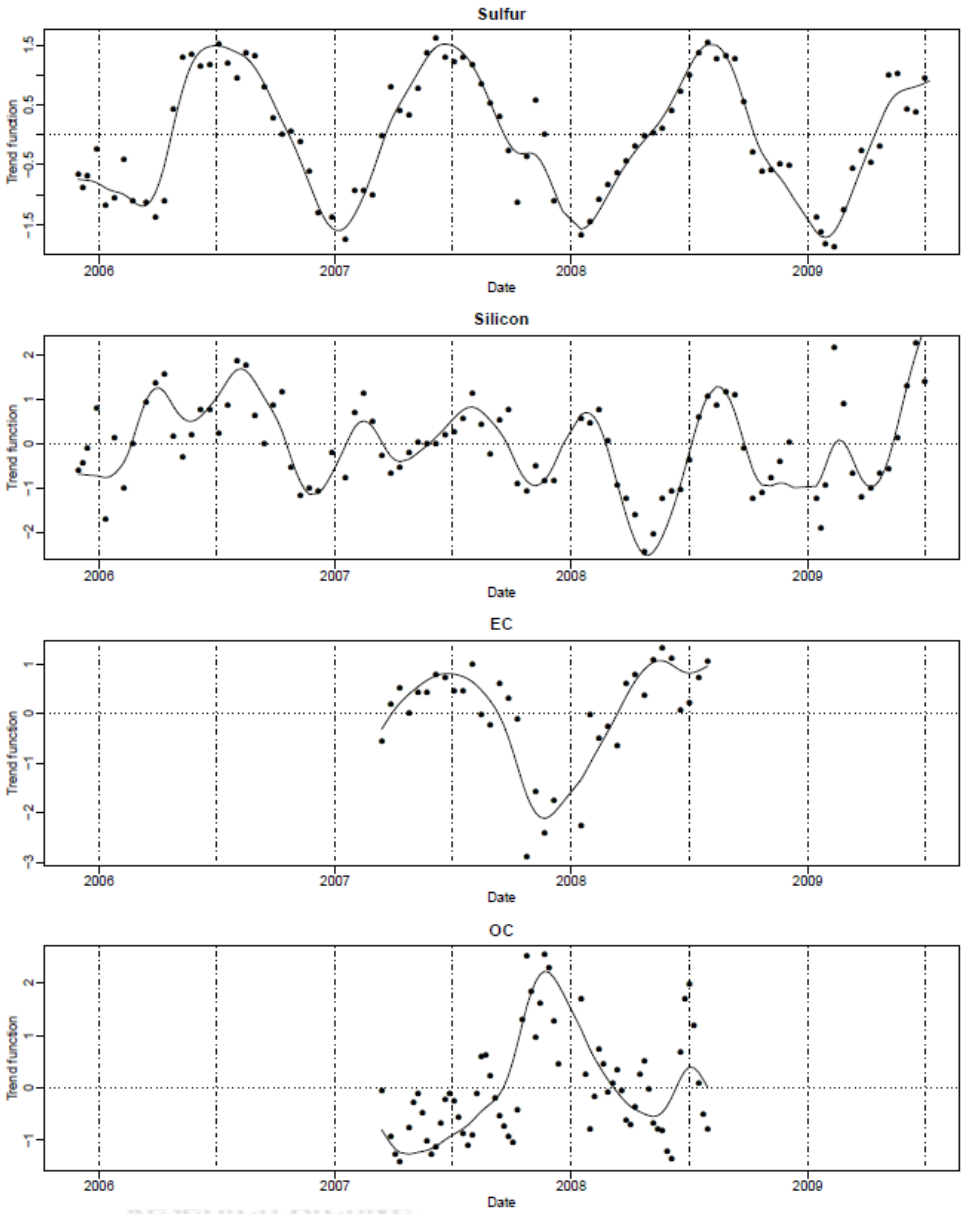


Figure 3. Estimated smooth temporal trend for four log-transformed $PM_{2.5}$ components in Los Angeles

Hosted by The Berkeley Electronic Press

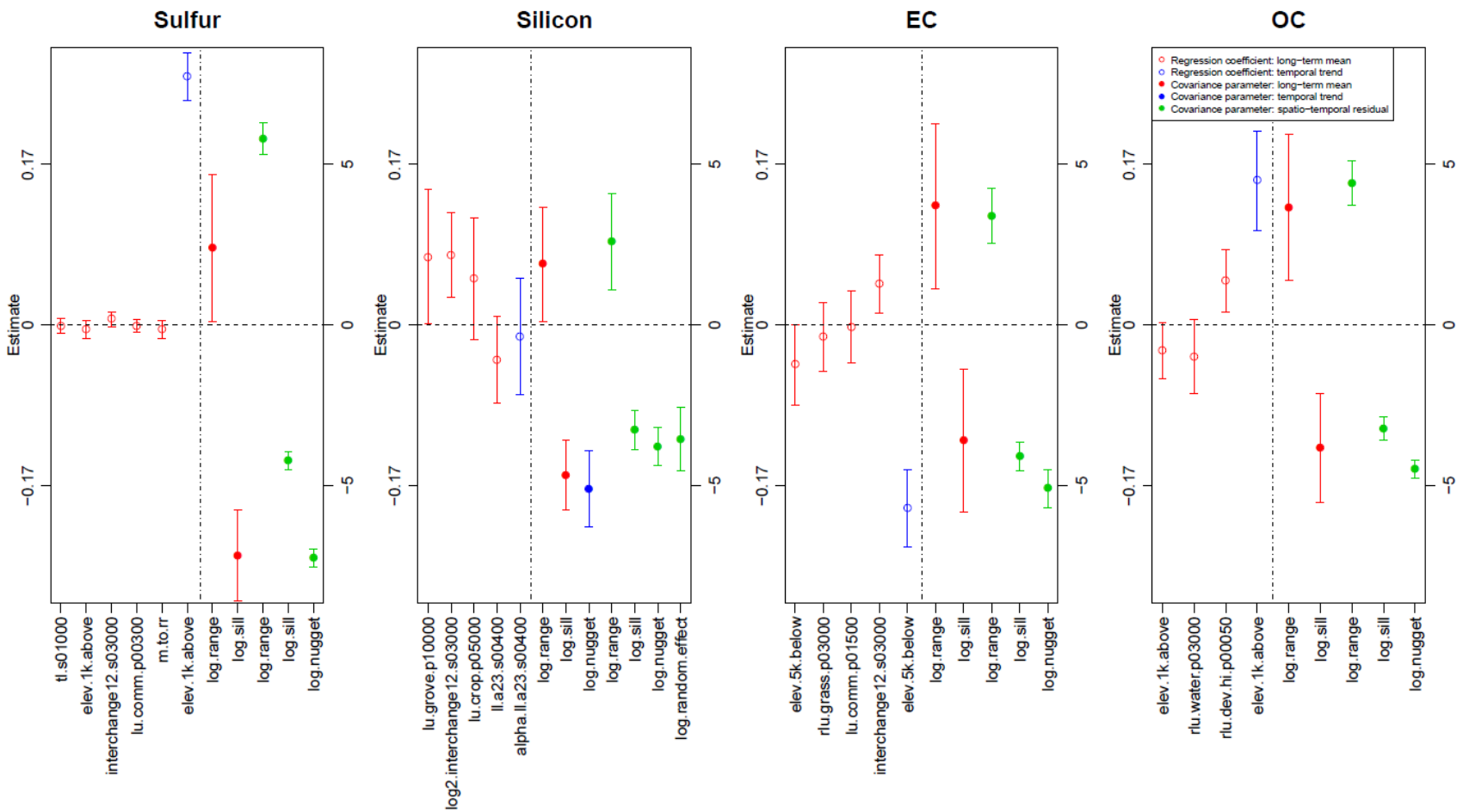


Figure 4. Estimated parameters for the selected geographical variables (scaled) and covariance structure in the spatio-temporal model for the four log-transformed $PM_{2.5}$ components in Los Angeles. Descriptions for the names of the geographical variables are shown in Supplemental Table 2.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

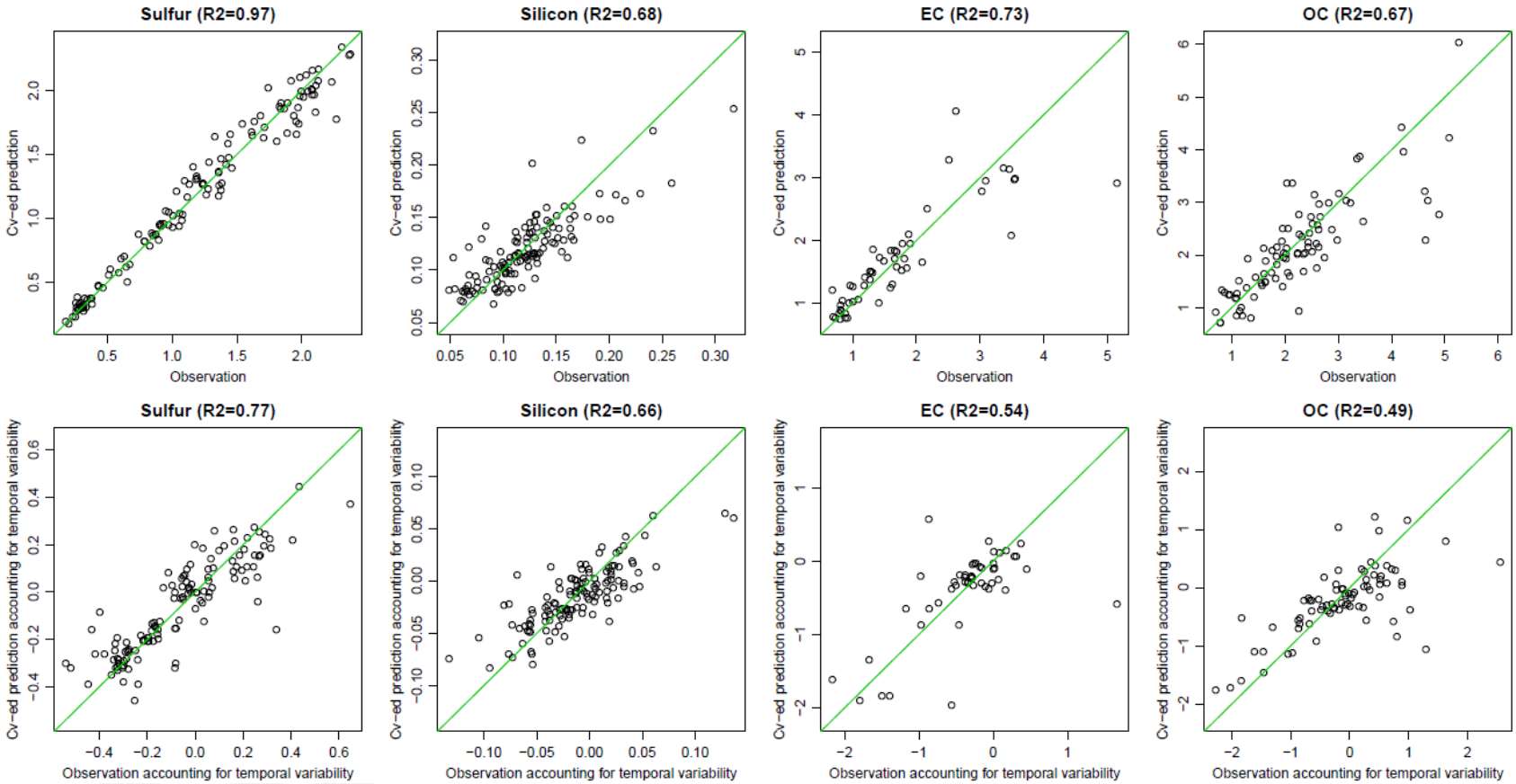


Figure 5. Component-specific scatter plots of observations and cross-validated predictions from the spatio-temporal model for 2-week average concentrations (top) and for 2-week average concentrations after accounting for temporal variability (bottom) across home-outdoor sites in Los Angeles

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

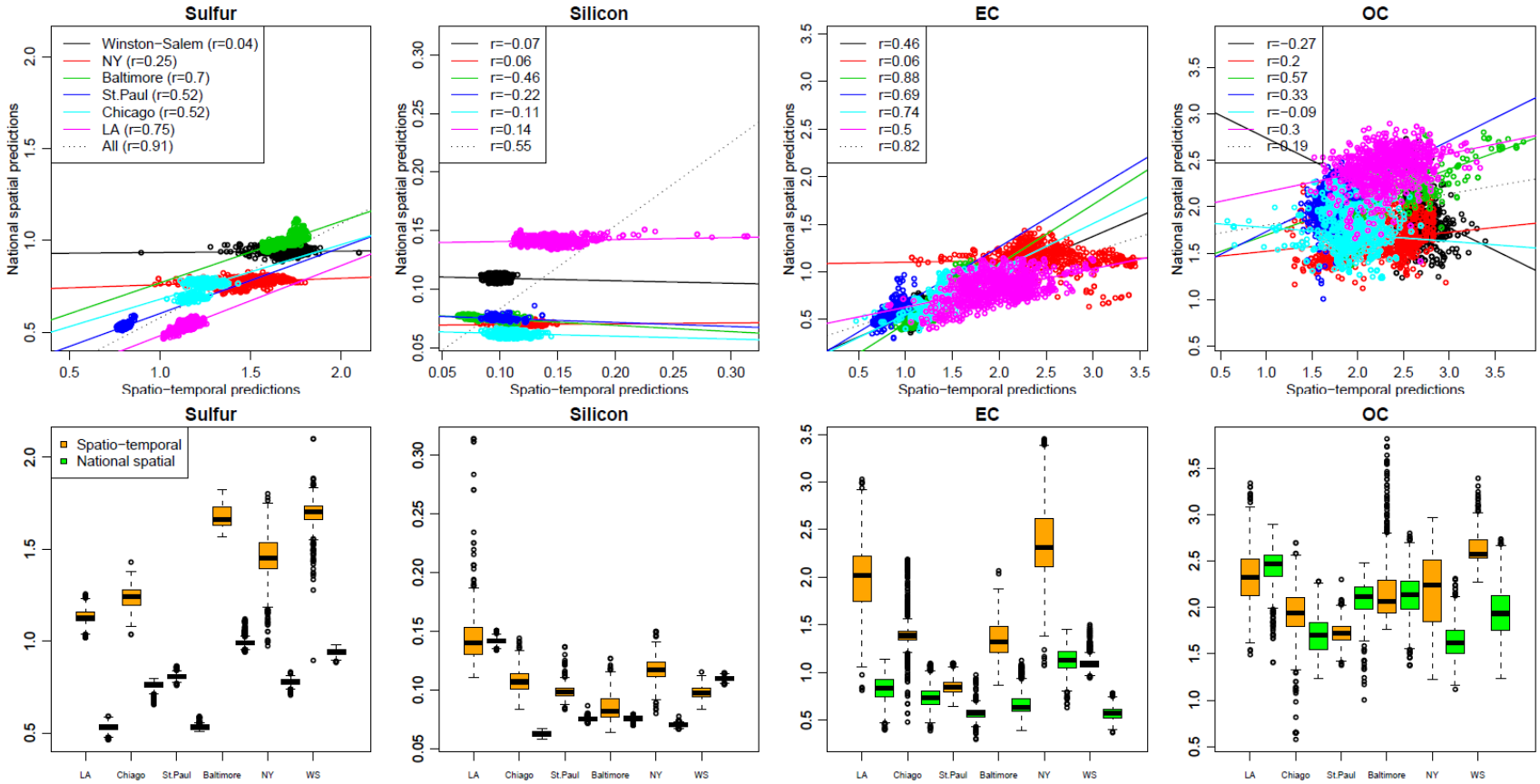


Figure 6. Component-specific scatter plots and box plots for spatio-temporal and national spatial model predictions of long-term average concentrations across six MESA city areas

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

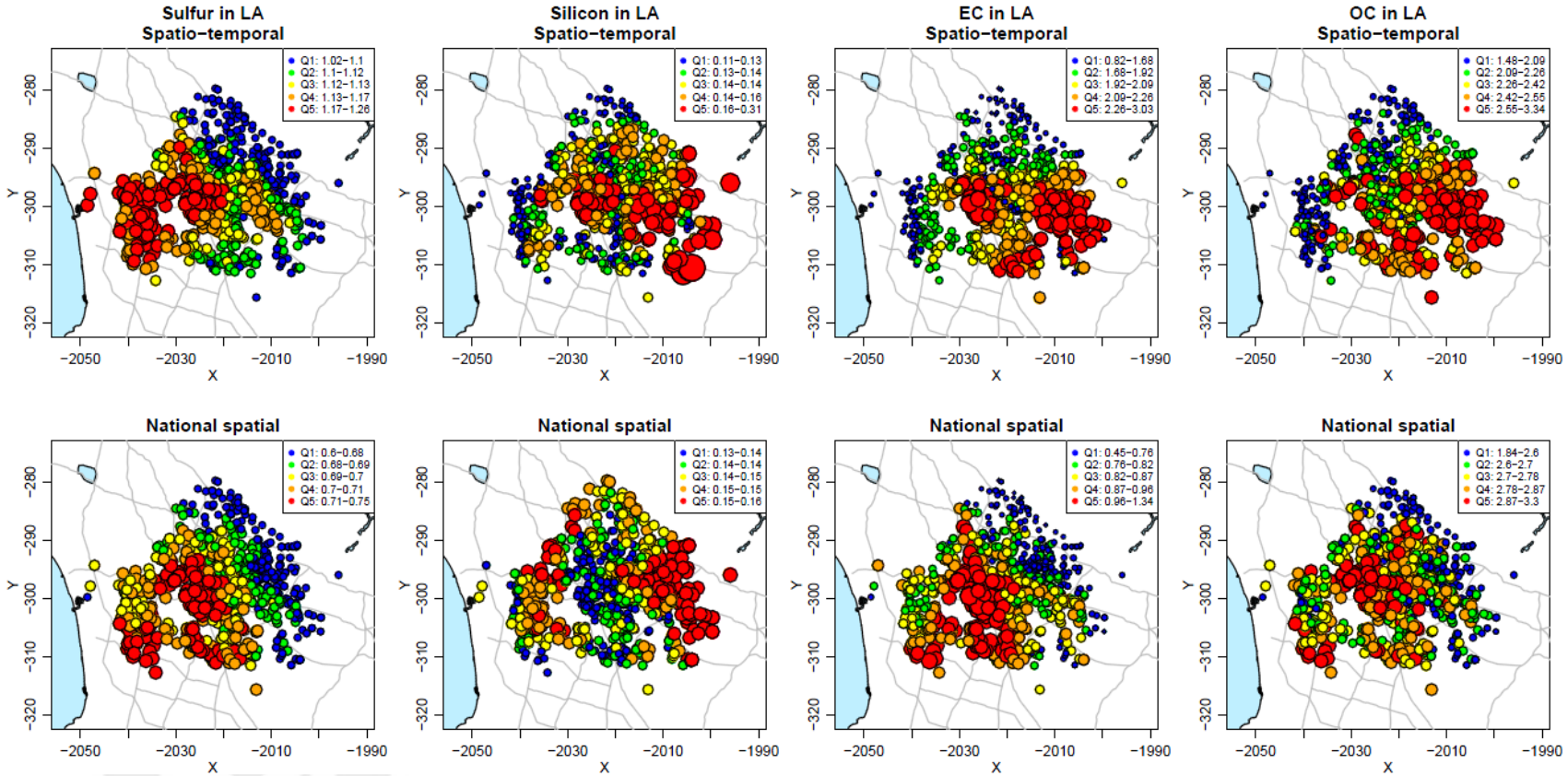


Figure 7. Predicted long-term average concentrations from the spatio-temporal and national spatial models of four $PM_{2.5}$ components at participant locations in Los Angeles (different colors represent quintiles of the range of concentrations)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

SUPPLEMENTAL MATERIALS

Supplemental Table 1. List of geographic variables

Category	Measure	Variable description
Traffic	Distance to the nearest road	Any road, A1, intersection
	Sum within buffers of 0.05-15 km	A1, A2+A3, truck route, intersections
Population	Sum within buffers of 0.5-3 km	Population in block groups
Land use (Urban)	Percent within buffers of 0.05-15 km	Urban or Built-Up land (residential, commercial, industrial, transportation, urban)
		Developed low, medium, and high density
		Developed open space
Land use (Rural)	Percent within buffers of 0.05-15 km	Agricultural land (cropland, groves, feeding)
		Rangeland (herbaceous, shrub)
		Forest land (deciduous, evergreen, mixed)
		Water (streams, lakes, reservoirs, bays)
		Wetland
		Barren land (beaches, dry salt flats, sand, mines, rock)
		Tundra
		Perennial snow or Ice
Position	Coordinates	Longitude, latitude
Source	Distance to the nearest source	Coastline, Coastline (rough)
		Commercial area
		Railroad
		Railyard
		Airport
		Major airport
		Large port
		City hall

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398

Emission	Sum within buffers of 3-30 km	$PM_{2.5}$ PM_{10} CO SO2 NOx
Vegetation	Quantiles within buffers of 0.5-10 km	Normalized Difference Vegetation Index (NDVI)
Imperviousness	Percent within buffers of 0.05-5 km	Impervious surface value
Elevation	Elevation above sea levels Counts of points above or below a threshold within buffers of 1-5 km	Elevation value
Residual oil	Distance to the nearest boiler Sum within buffers of 0.1-3 km	Residual oil grade 4 or 6 Total residual oil active heating capacity

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Supplemental Table 2. List of selected geographical variables for four $PM_{2.5}$ components

Category	Variable name	Variable description
Traffic	ll.a1.s<radius>	Sum of CFCC (Census Feature Class Code) A1 roads within a <radius> meter buffer in meters
	ll.a23.s<radius>	Sum of CFCC A2 and A3 roads within a <radius> meter buffer in meters
	tl.s<radius>	Sum of truck rout lengths within a <radius> meter buffer in meters
	interchange12.s<radius>	Sum of intersections of A1 and A2 roads within a <radius> meter buffer
	interchange3.s<radius>	Sum of intersections of A3 and any roads within a <radius> meter buffer
	intersect.s<radius>	Sum of intersections of any roads within a <radius> meter buffer
	log10.m.to.a1	Log-transformed meters to nearest CFCC A1 road
	log10.m.to.interchange3	Log-transformed meters to nearest intersection of A3 roads
	log10.m.to.truck	Log-transformed meters to nearest truck route
	log2.interchange12.s<radius>	Log-transformed sum of intersections of A1 and A2 roads within a <radius> meter buffer
	log2.interchange3.s<radius>	Log-transformed sum of intersections of A3 and any roads within a <radius> meter buffer
	log2.intersect.s<radius>	Log-transformed sum of intersections of any roads within a <radius> meter buffer
Land use	lu.bays.p<radius>	Percentage of 1980 land use type of bays and estuaries within a <radius> meter buffer
	lu.comm.p<radius>	Percentage of 1980 land use type of commercial and service areas within a <radius> meter buffer
	lu.crop.p<radius>	Percentage of 1980 land use type of cropland and pasture within a <radius> meter buffer
	lu.forest.p<radius>	Percentage of 1980 land use type of deciduous forest land within a <radius> meter buffer
	lu.green.p<radius>	Percentage of 1980 land use type of evergreen forest land within a <radius> meter buffer
	lu.grove.p<radius>	Percentage of 1980 land use type of orchards, groves, vineyards, nurseries within a <radius> meter buffer
	lu.industrial.p<radius>	Percentage of 1980 land use type of industrial areas within a <radius> meter buffer
	lu.oth.urban.p<radius>	Percentage of 1980 land use type of other urban or built-up land within a <radius> meter buffer
	lu.transition.p<radius>	Percentage of 1980 land use type of transitional areas within a <radius> meter buffer
	lu.transport.p<radius>	Percentage of 1980 land use type of transportation, communications and utilities within a <radius> meter buffer
	lc.anyforest.p<radius>	Percentage of 1980 land use type of woody wetland or forest area within a <radius> meter buffer
	lc.openbasic.p<radius>	Percentage of 1980 land use type of any forest, water, crop, shrub, pasture, herb, grass, and barren areas within a <radius> meter buffer
	lc.openplus.p<radius>	Percentage of 1980 land use type of any forest, water, crop, shrub, pasture, herb, grass, barren, and developed open spance within a <radius> meter buffer
	lc.water.p<radius>	Percentage of 1980 land use type of open water within a <radius> meter buffer

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

	rlu.grass.p<radius>	Percentage of 2006 land use type of grasslands and herbaceous vegatation within a <radius> meter buffer
	rlu.water.p<radius>	Percentage of 2006 land use type of open water within a <radius> meter buffer
	rlu.dev.hi.p<radius>	Percentage of 2006 land use type of developed high intensity within a <radius> meter buffer
	rlu.dev.med.p<radius>	Percentage of 2006 land use type of developed medium intensity within a <radius> meter buffer
	rlu.dev.open.p<radius>	Percentage of 2006 land use type of developed open space within a <radius> meter buffer
	rlc.dev.medhi.p<radius>	Percentage of 2006 land use type of developed high and medium intensity within a <radius> meter buffer
	rlc.openbasic.p<radius>	Percentage of 2006 land use type of any forest, water, crop, shrub, pasture, herb, grass, and barren areas within a <radius> meter buffer
	rlc.anyforest.p<radius>	Percentage of 2006 land use type of woody wetland or forest area within a <radius> meter buffer
Position	long	GPS longitude coordinate in decimal degrees
Source	m.to.rr	Meters to the nearest railroad
	log10.m.to.comm	Log-transformed meters to nearest commercial zone
	log10.m.to.l.airp	Log-transformed meters to nearest large airport
	log10.m.to.rr	Log-transformed meters to nearest railroad
Emission	em.CO.s30000	Sum of CO emissions in tons per year from tall stacks within 30 kilometer, minus the emissions from tall stacks within 3 kilometers
Vegetation	ndvi.q25.a<radius>	Average 2006 NDVI value at the first quantile within a <radius> meter buffer
	ndvi.q75.a<radius>	Average 2006 NDVI value at the third quantile within a <radius> meter buffer
Imperviousness	imp.a<radius>	Average impervious surface value (percent imperviousness) within a <radius> meter buffer
Elevation	elev.1k.above	Count of points (out of 24) more than 20 meter uphill of the location within 1 kilometer buffer
	elev.1k.below	Count of points (out of 24) more than 20 meter downhill of the location within 1 kilometer buffer
	elev.1k.rabove	Squre root-transformed count of points (out of 24) more than 20 meter uphill of the location within 1 kilometer buffer
	elev.1k.rbelow	Squre root-transformed count of points (out of 24) more than 20 meter downhill of the location within 1 kilometer buffer
	elev.1k.below	Count of points (out of 24) more than 50 meter downhill of the location within 5 kilometer buffer
Residual oil	oil.edf4.s<radius>	Sum of potential mega BTU (British Thermal Unit) from oil number 4 per hour within a <radius> meter buffer
	log10.m.to.6oil	Log-transformed meters to nearest Number 6 oil boiler

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Supplemental Table 3. Proportion of total variance of the predictions captured by the long-term mean, temporal trend, and spatio-temporal residuals across MESA Air monitoring home-outdoor sites

City	Pollutant	Long-term mean [a]		Temporal trend [a]	Spatio-temporal residual [a]
		Regression [b]	Regression+kriging+Error		
LA	Sulfur	0.00	0.01	0.82	0.17
	Silicon	0.10	0.20	0.32	0.48
	EC	0.04	0.35	0.50	0.15
	OC	0.04	0.28	0.36	0.36
Chicago	Sulfur	0.02	0.04	0.52	0.43
	Silicon	0.07	0.08	0.76	0.15
	EC	0.26	0.31	0.54	0.16
	OC	0.19	0.19	0.60	0.22
St. Paul	Sulfur	0.00	0.00	0.56	0.44
	Silicon	0.02	0.02	0.60	0.38
	EC	0.39	0.41	0.41	0.18
	OC	0.12	0.13	0.71	0.15
Baltimore	Sulfur	0.01	0.01	0.79	0.20
	Silicon	0.11	0.11	0.69	0.19
	EC	0.48	0.48	0.35	0.17
	OC	0.10	0.10	0.72	0.18
NY	Sulfur	0.09	0.09	0.54	0.37
	Silicon	0.11	0.11	0.39	0.51
	EC	0.66	0.64	0.36	0.00
	OC	0.46	0.46	0.43	0.11
Winston-Salem	Sulfur	0.01	0.01	0.84	0.15
	Silicon	0.02	0.02	0.74	0.24
	EC	0.17	0.21	0.57	0.22
	OC	0.06	0.06	0.43	0.51

a. Sum of ratios of long-term mean including regression, kriging, and error, temporal trend, and spatio-temporal residual is equal to 1; Total variance used as denominator for calculating ratios was sum of variances of long-term mean, temporal trend, and spatio-temporal residual instead of variance of predictions given correlation structure between three parts

http://biostats.bepress.com/uwbiostat/paper398

b. Ratio of regression part for long-term mean is separately presented to show its contribution to total variability

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Supplemental Table 4. Cross-validation statistics of predicted concentrations of $PM_{2.5}$ between the NPACT sptaio-temporal model and the MESA Air likelihood model in six MESA Air areas

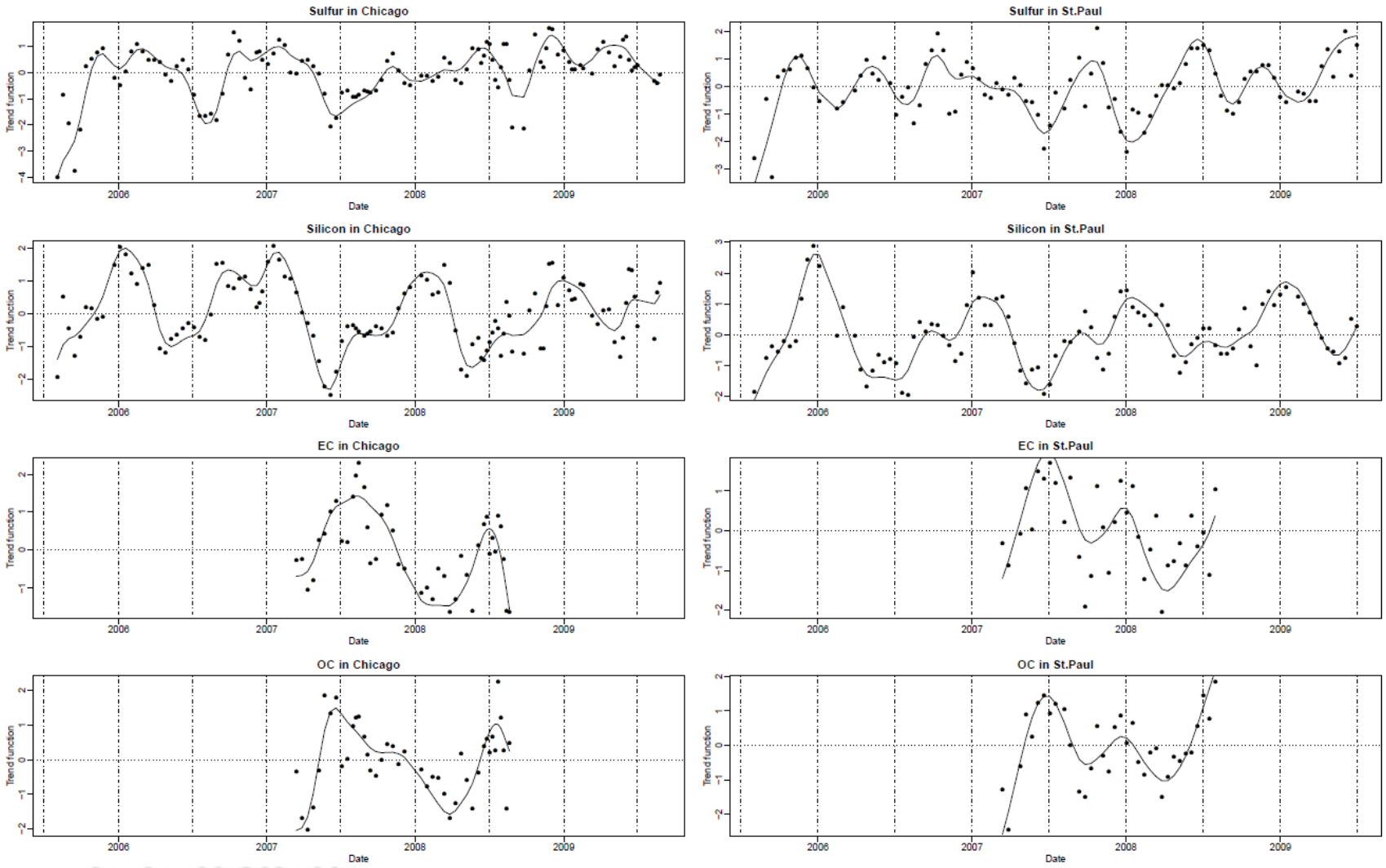
City	Spatio-temporal model					
	NPACT			MESA Air [b]		
	MSE	R^2	Temporally-adjusted R^2 [a]	MSE	R^2	Temporally-adjusted R^2 [a]
Los Angeles	4.51	0.84	0.37	8.58	0.77	0.41
Chicago	3.57	0.67	0.13	2.04	0.80	0.23
St. Paul	0.77	0.94	0.34	3.17	0.78	0.23
Baltimore	1.15	0.93	0.60	1.54	0.84	0.46
New York	15.32	0.36	0.00	15.76	0.35	0.39
Winston-Salem	2.18	0.89	0.22	1.00	0.85	0.29

a. Temporal variability was adjusted by spatial averages of fixed site measurements in NPACT and fixed sites as well as EPA sites in MESA Air

b. Keller J, Kim SY, Sheppard L, Sampson PD, Szpiro AA, Vedal S, Kaufman JD (2013). A unified spatiotemporal modeling approach for prediction of multiple air pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (in preparation).

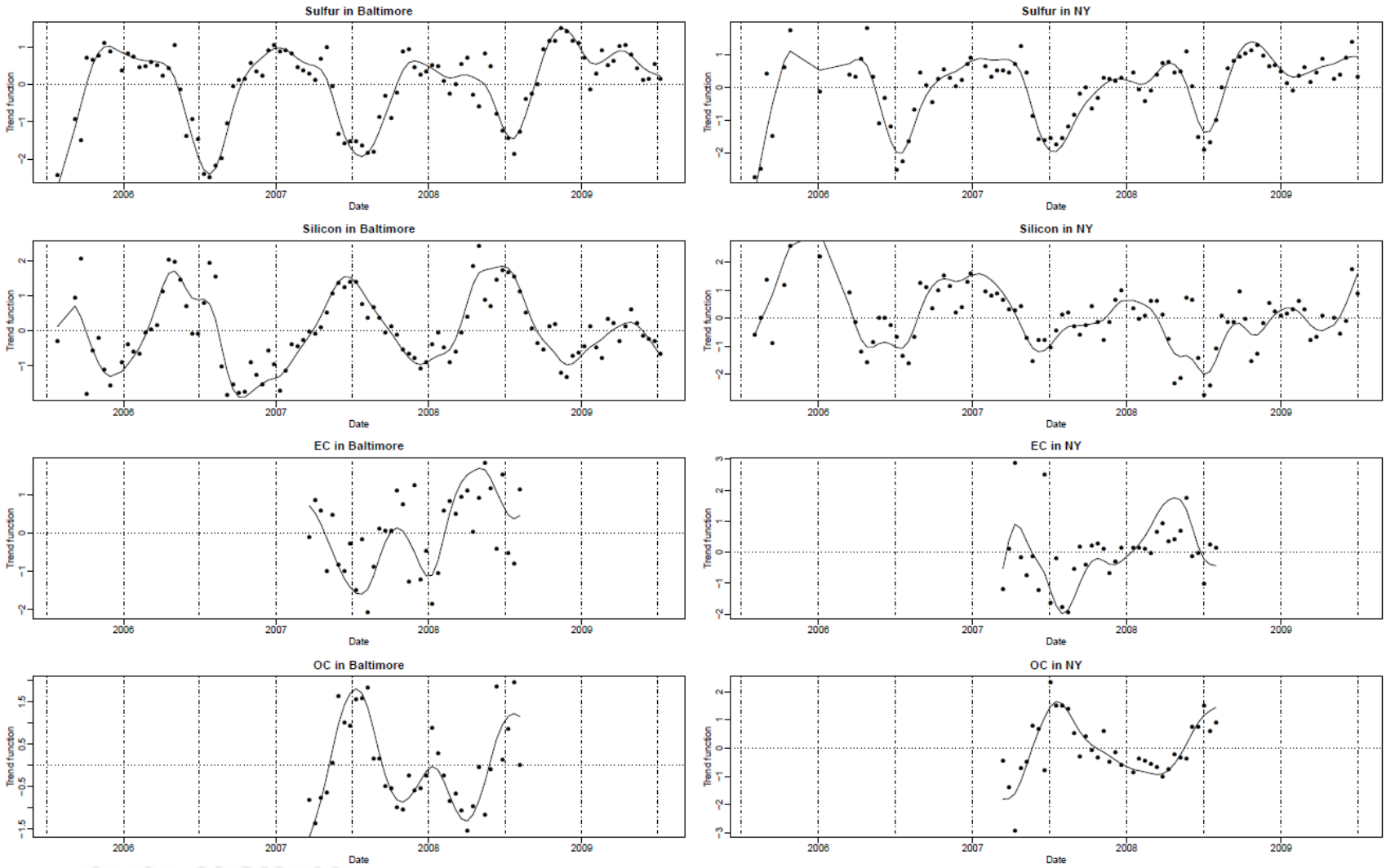
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper398



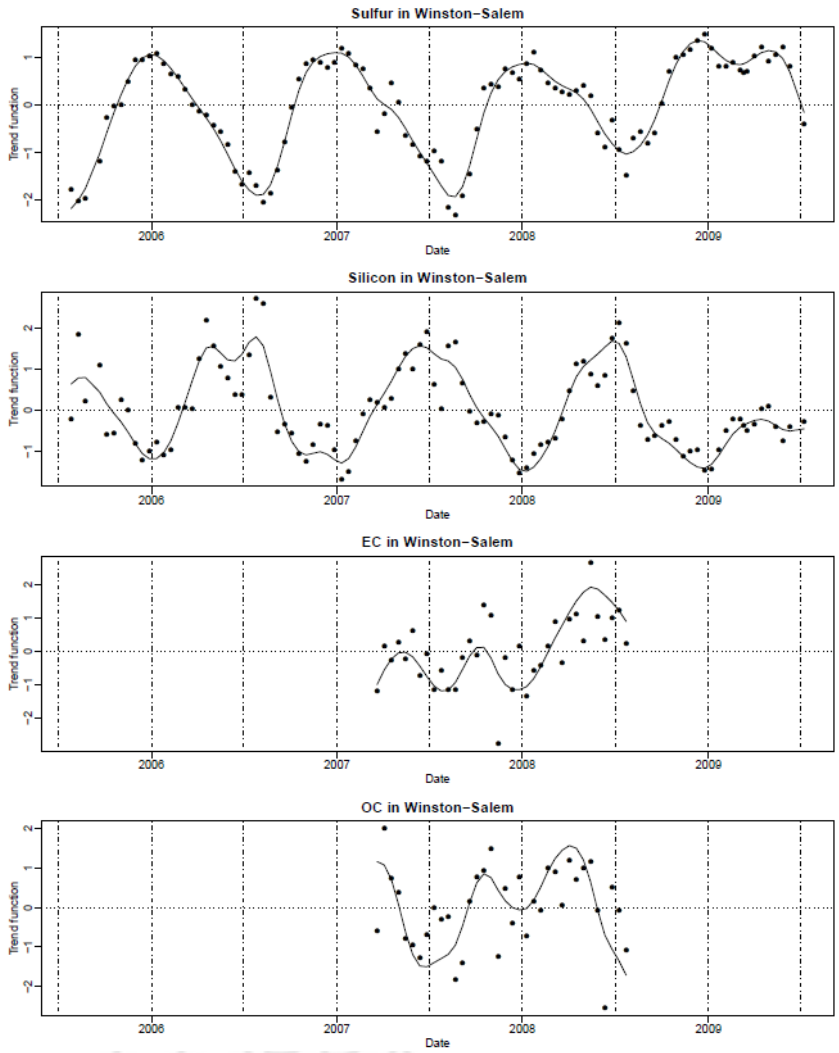
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

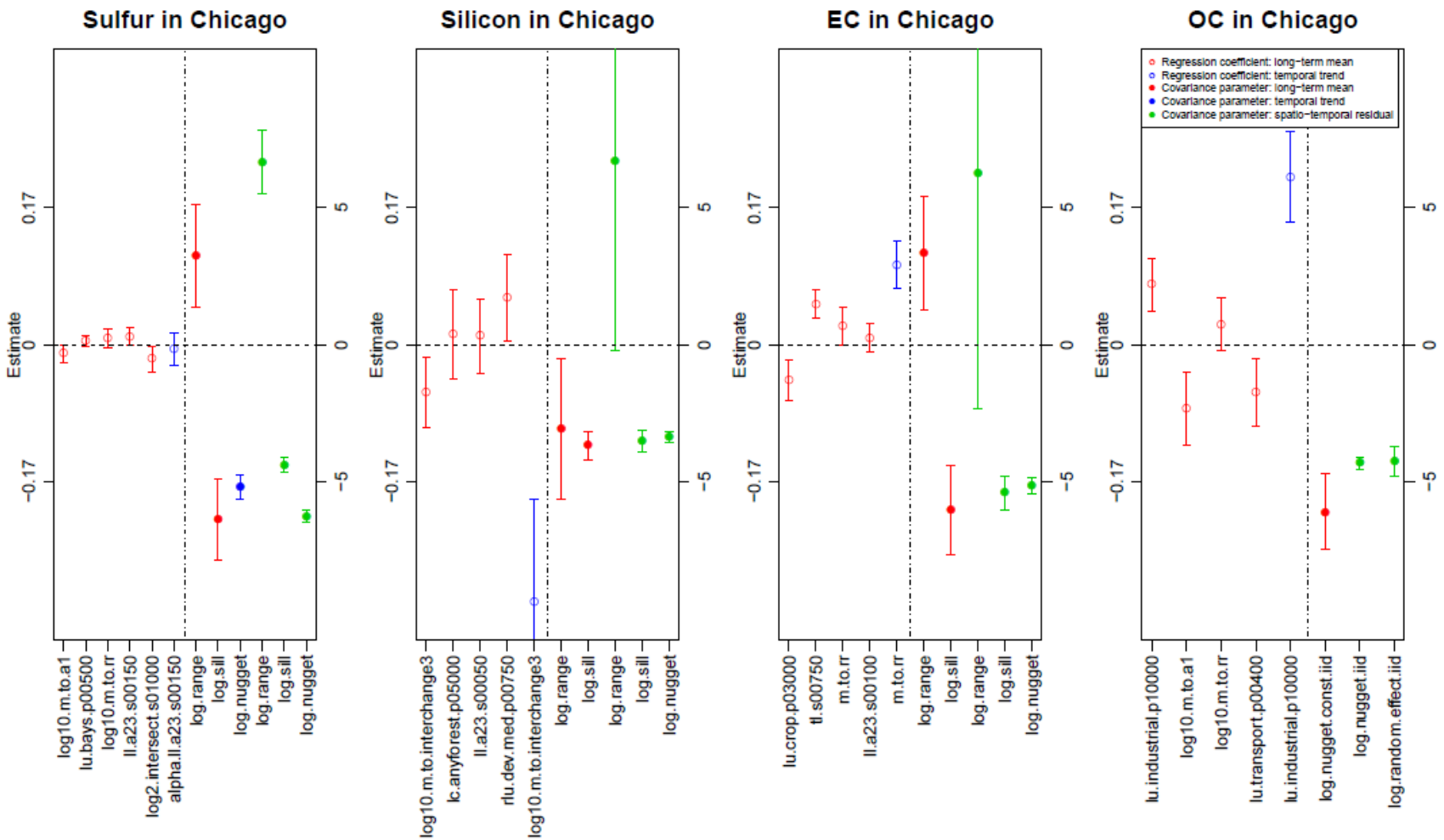
http://biostats.bepress.com/uwbiostat/paper398



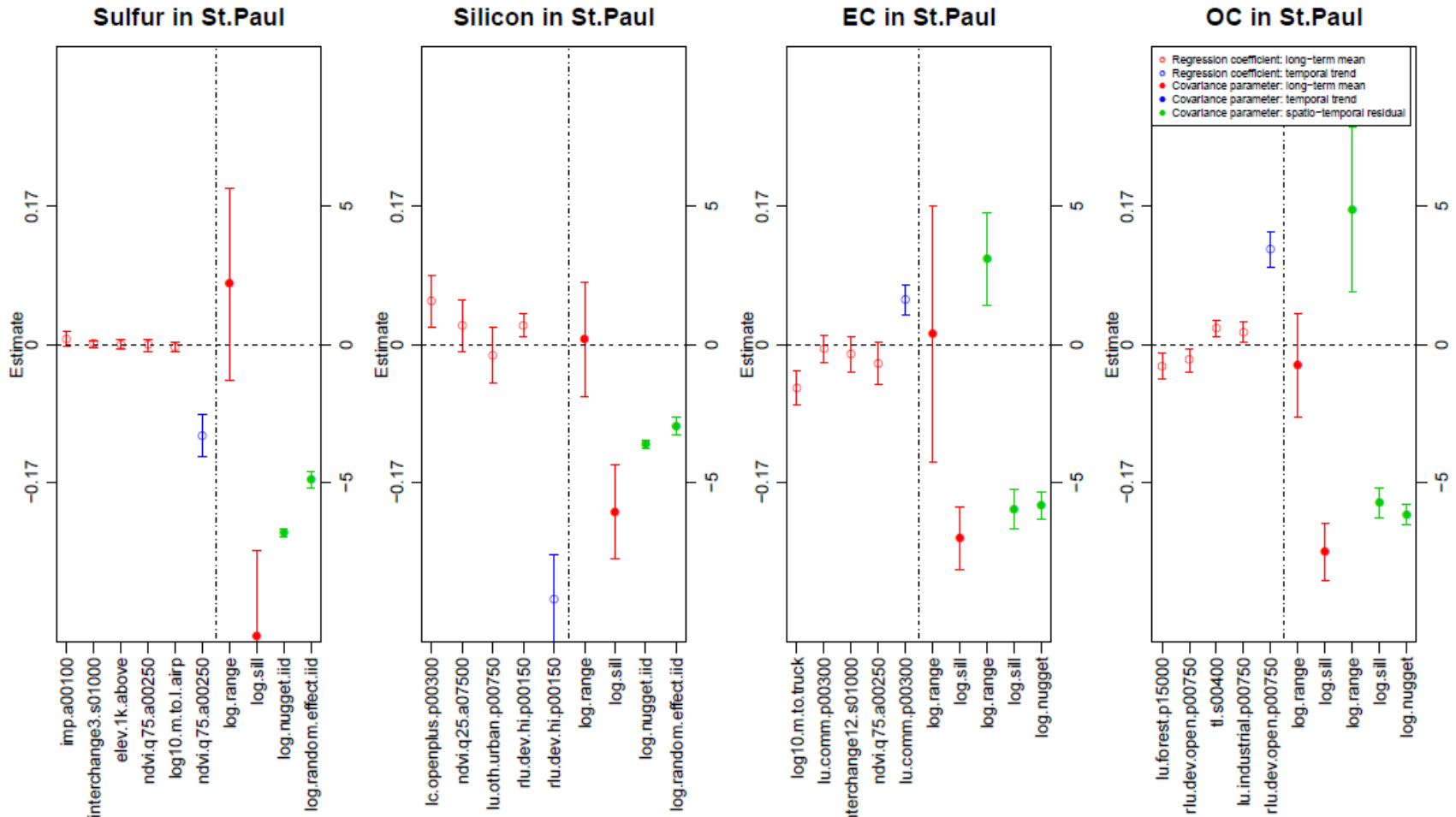
Supplemental Figure 1. Estimated temporal trend for four log-transformed $PM_{2.5}$ components in the five cities

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

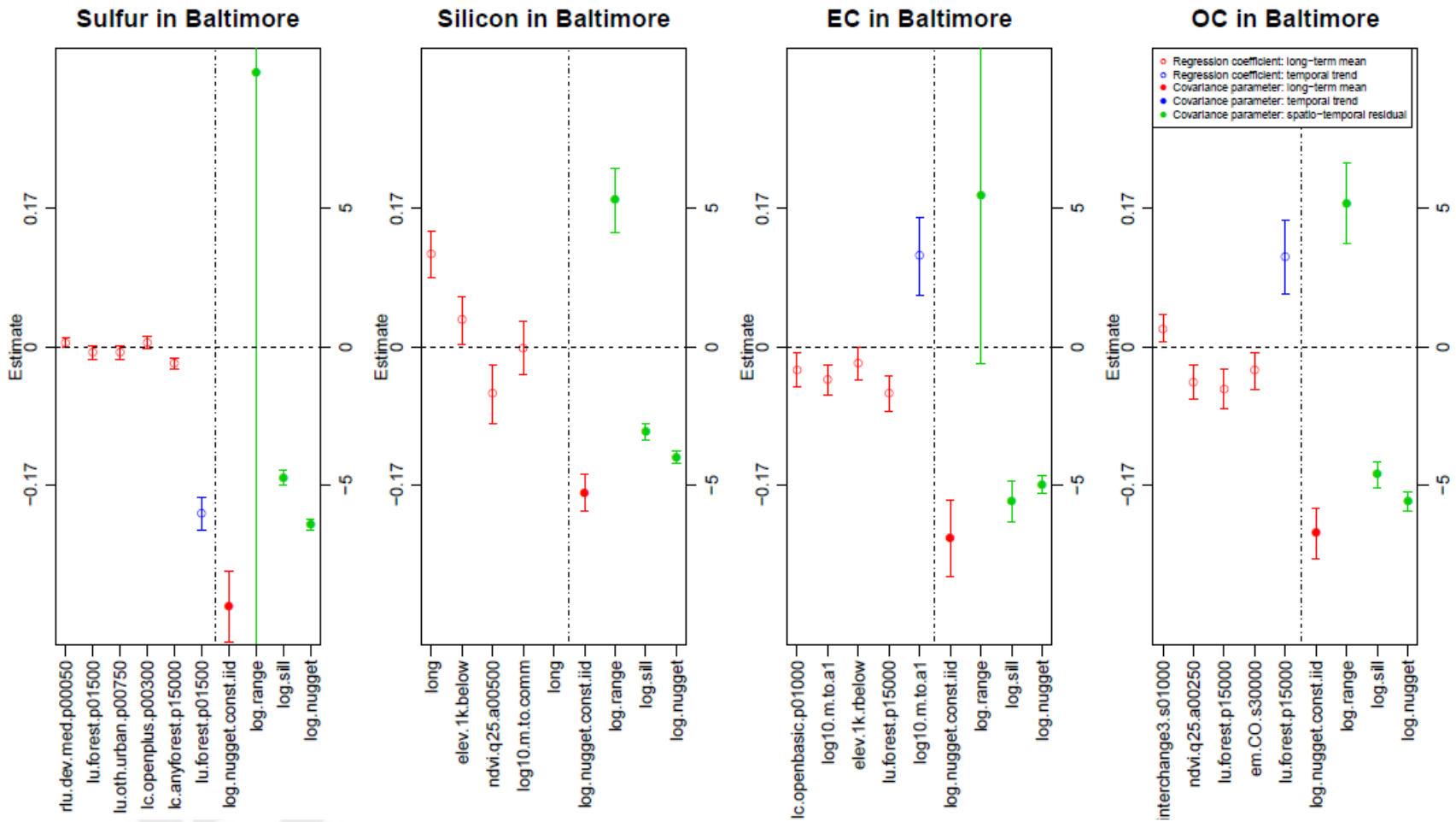


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

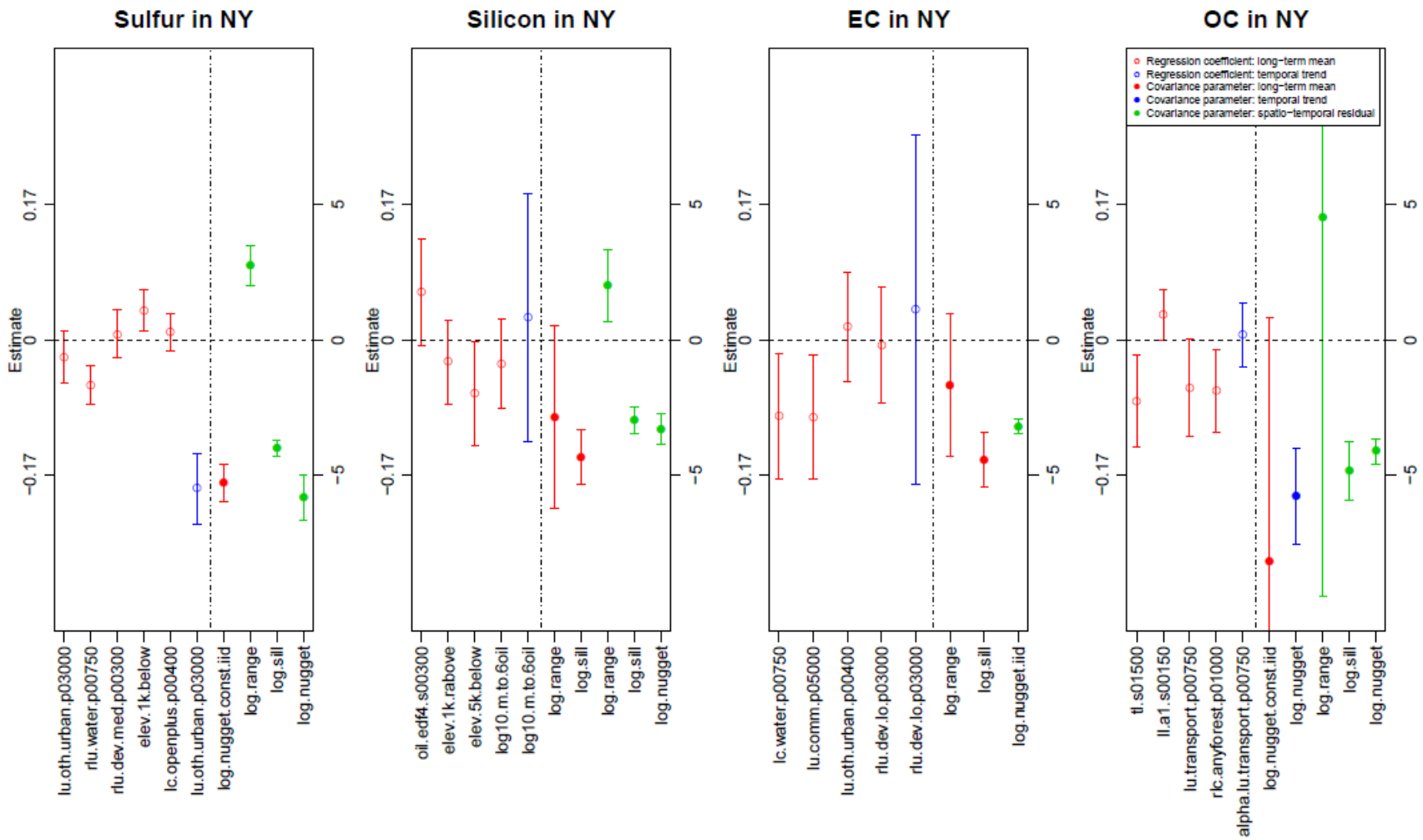


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

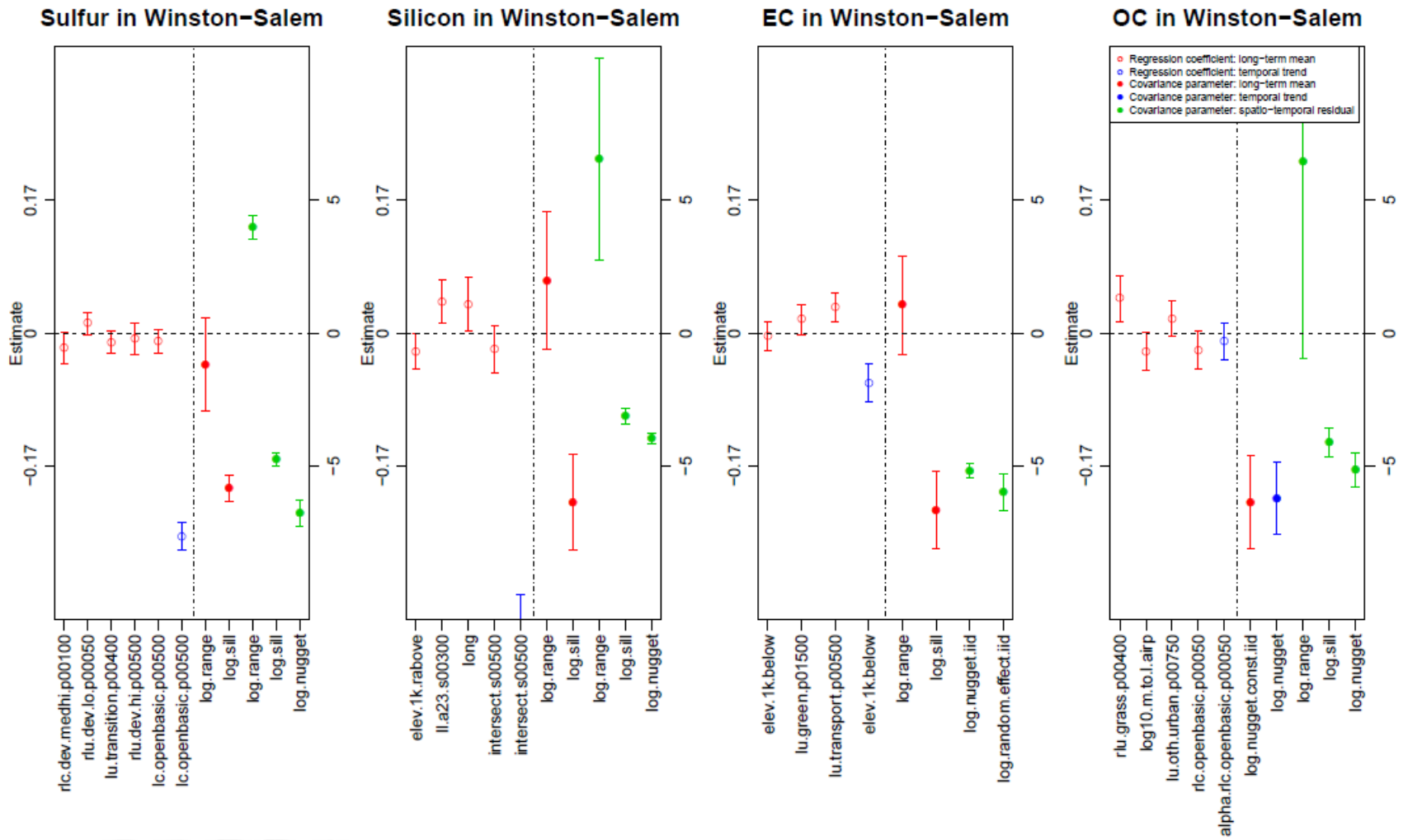


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive



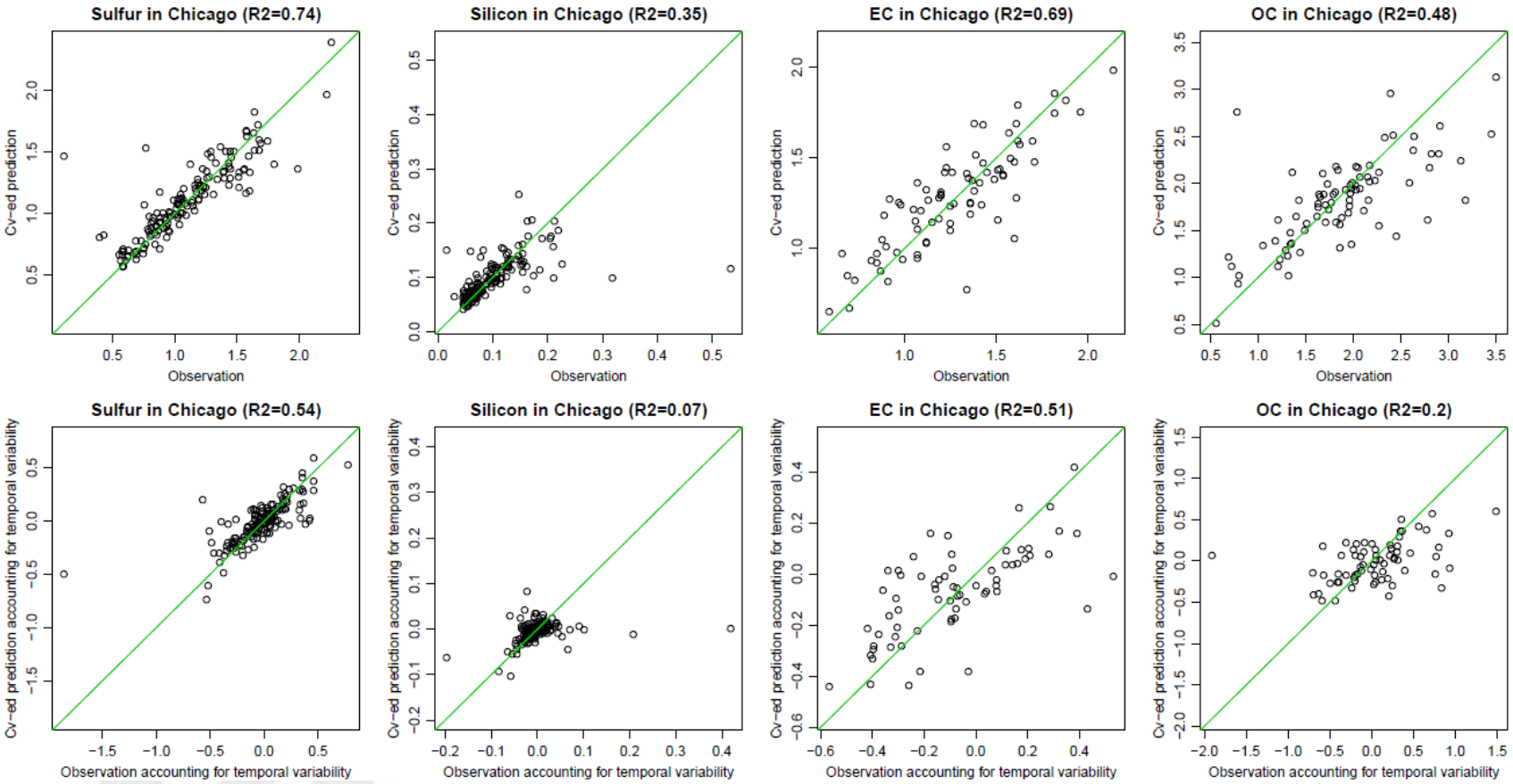
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press



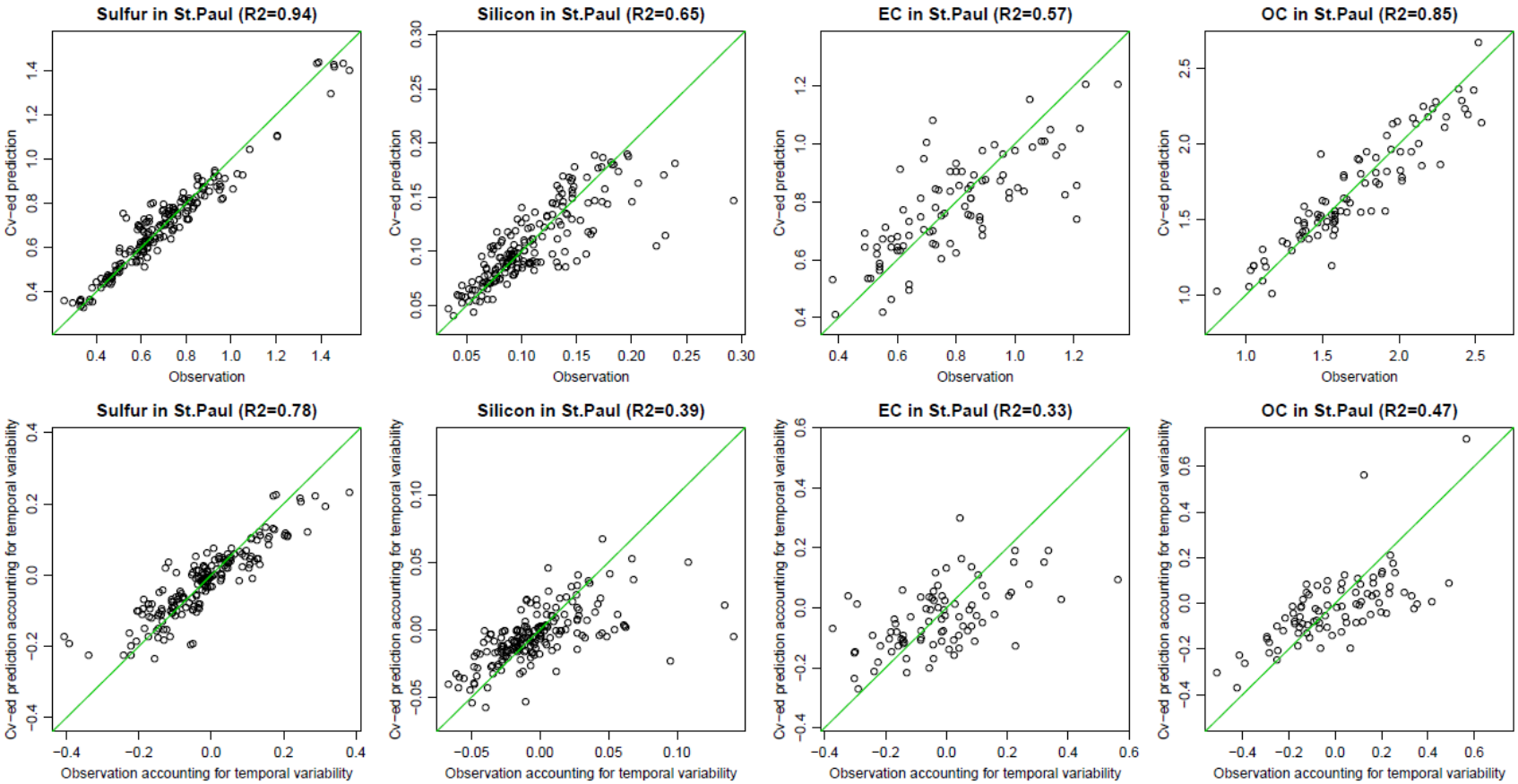
Supplemental Figure 2. Estimated parameters for the selected geographical variables (scaled) and covariance structure in the spatio-temporal model for the four log-transformed $PM_{2.5}$ components in the five cities

Collection of Biostatistics Research Archive

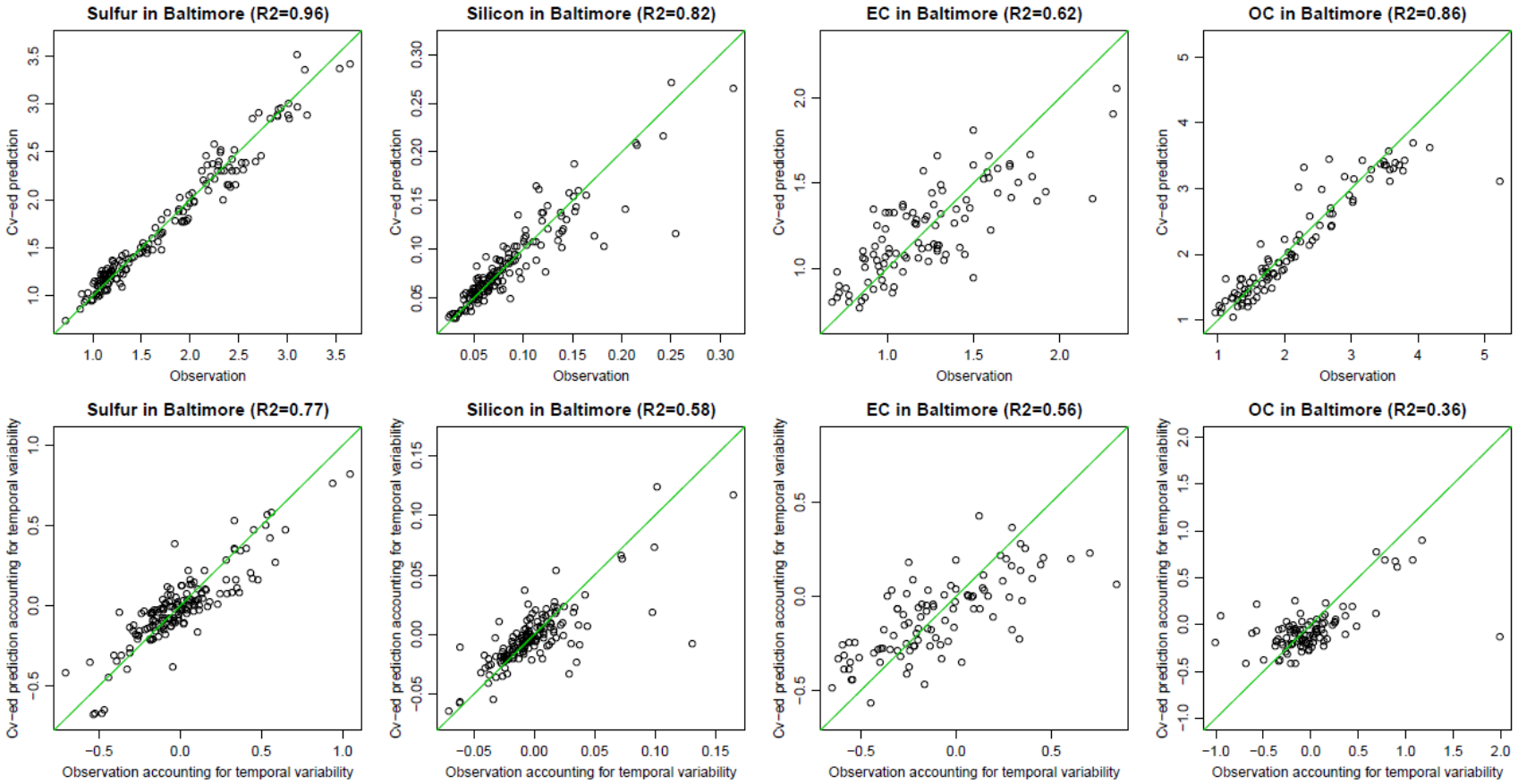


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

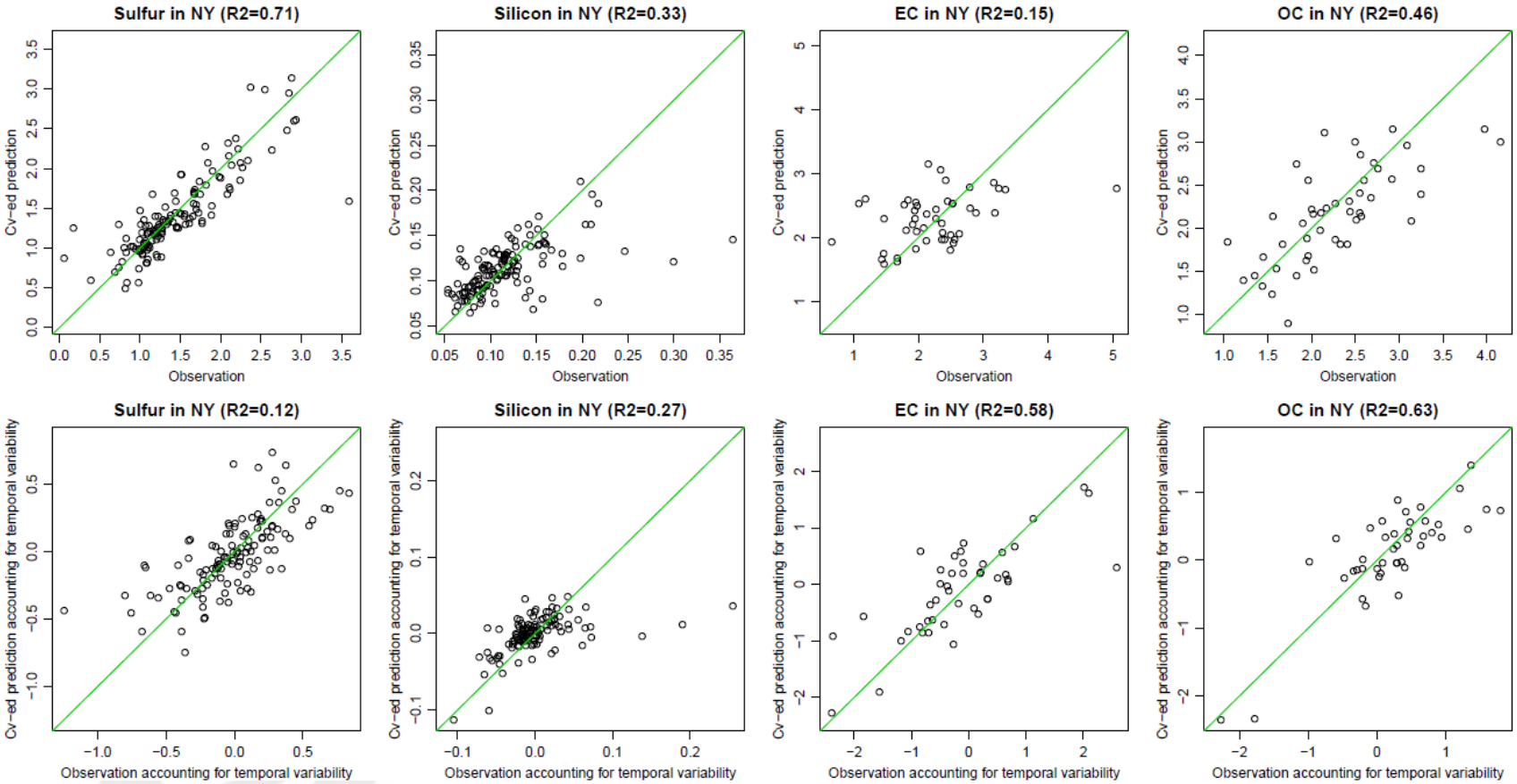


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

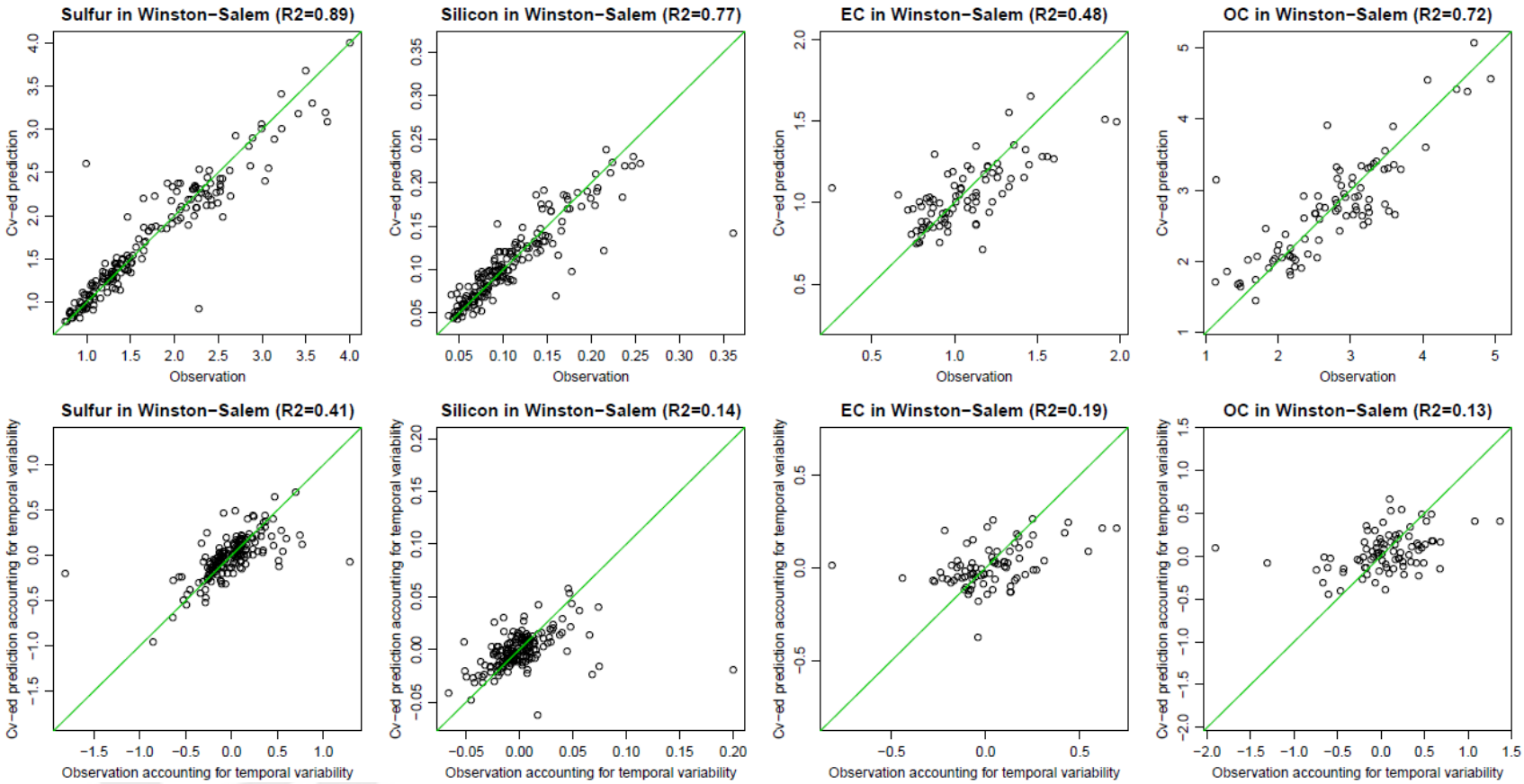


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press



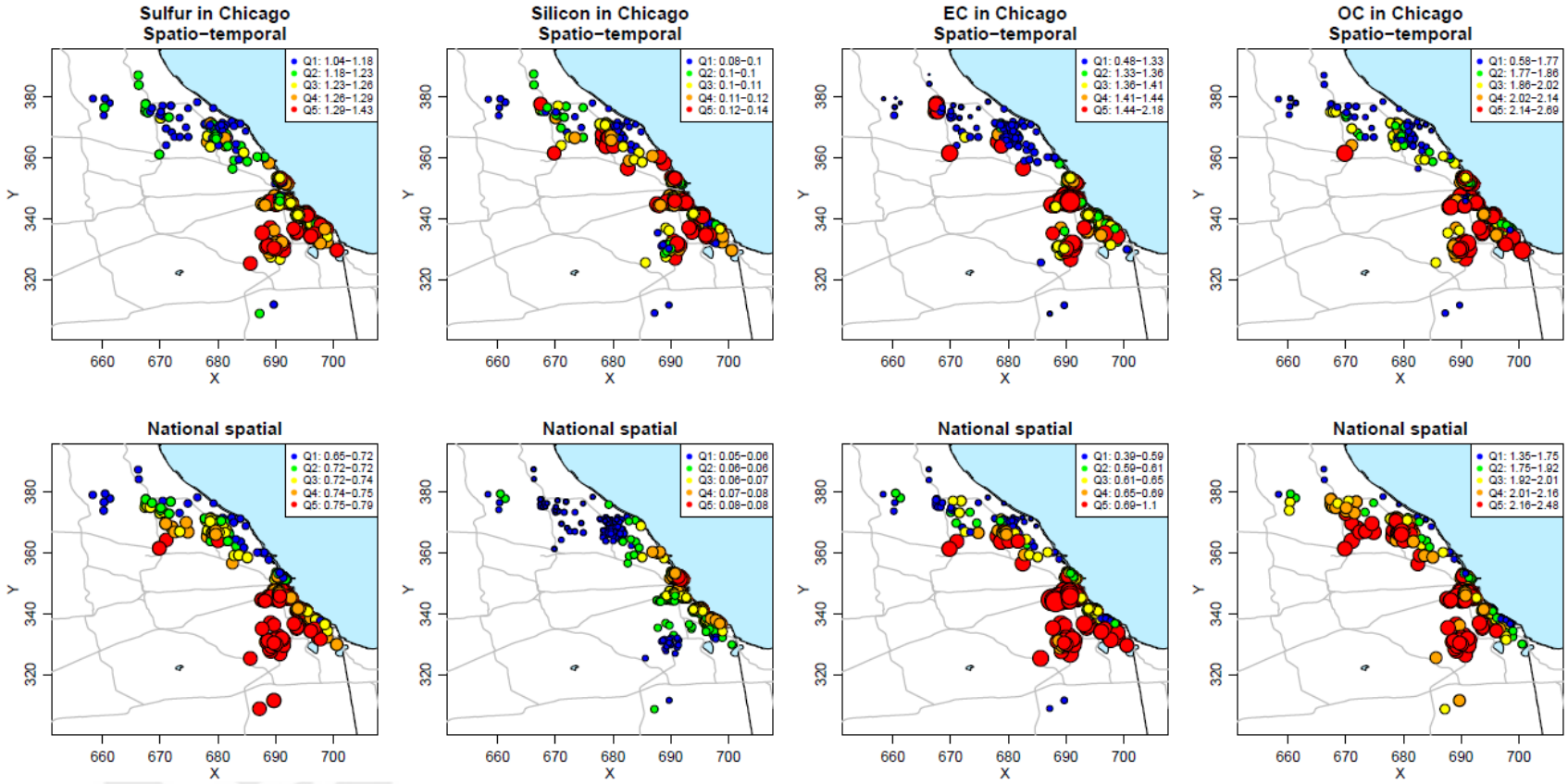
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive



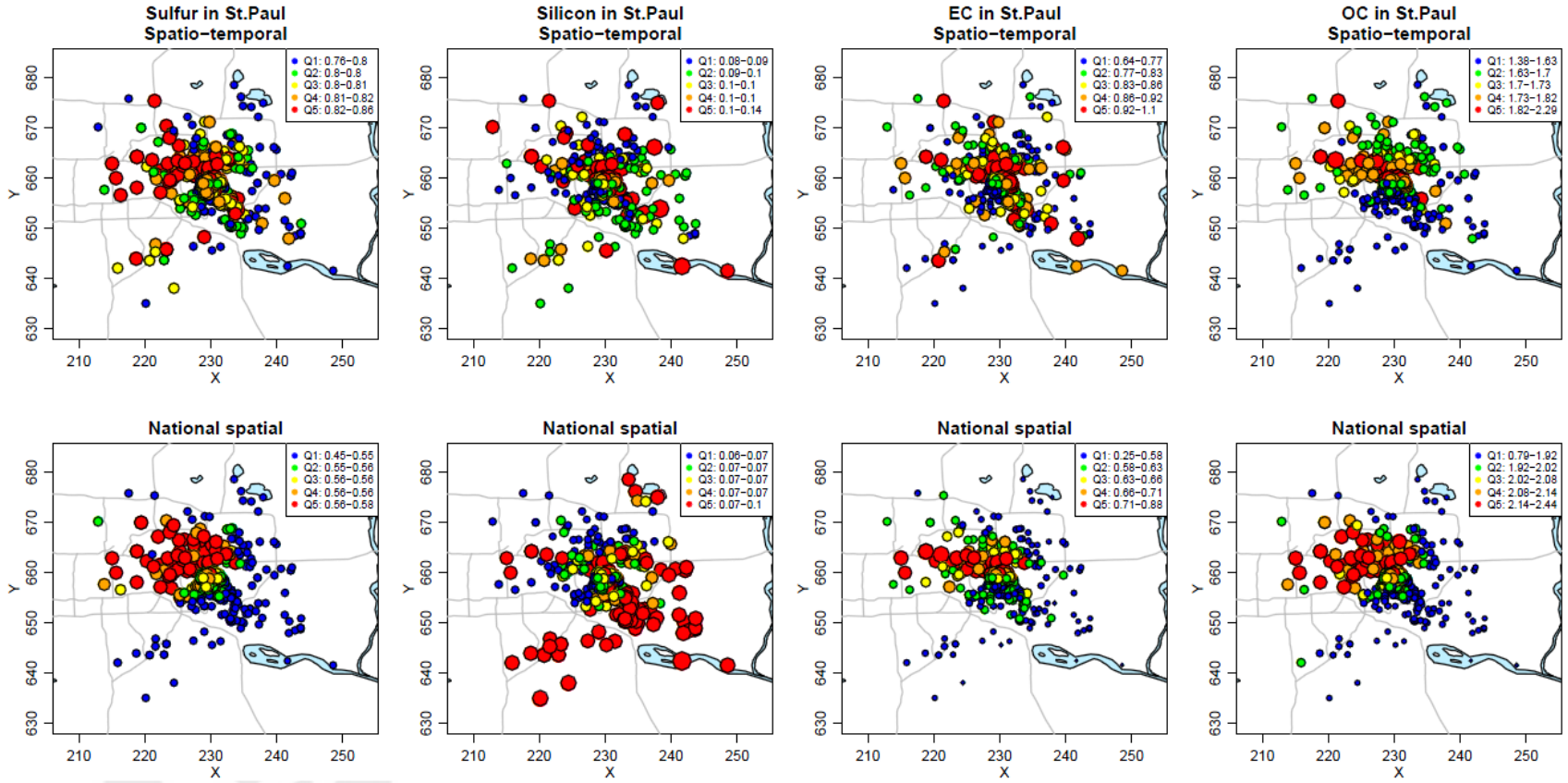
Supplemental Figure 3. Scatter plots of observations and cross-validated predictions from the spatio-temporal model for 2-week concentrations (left) and for 2-week concentrations accounting for temporal variability (right) for each component across home-outdoor sites in the five cities

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

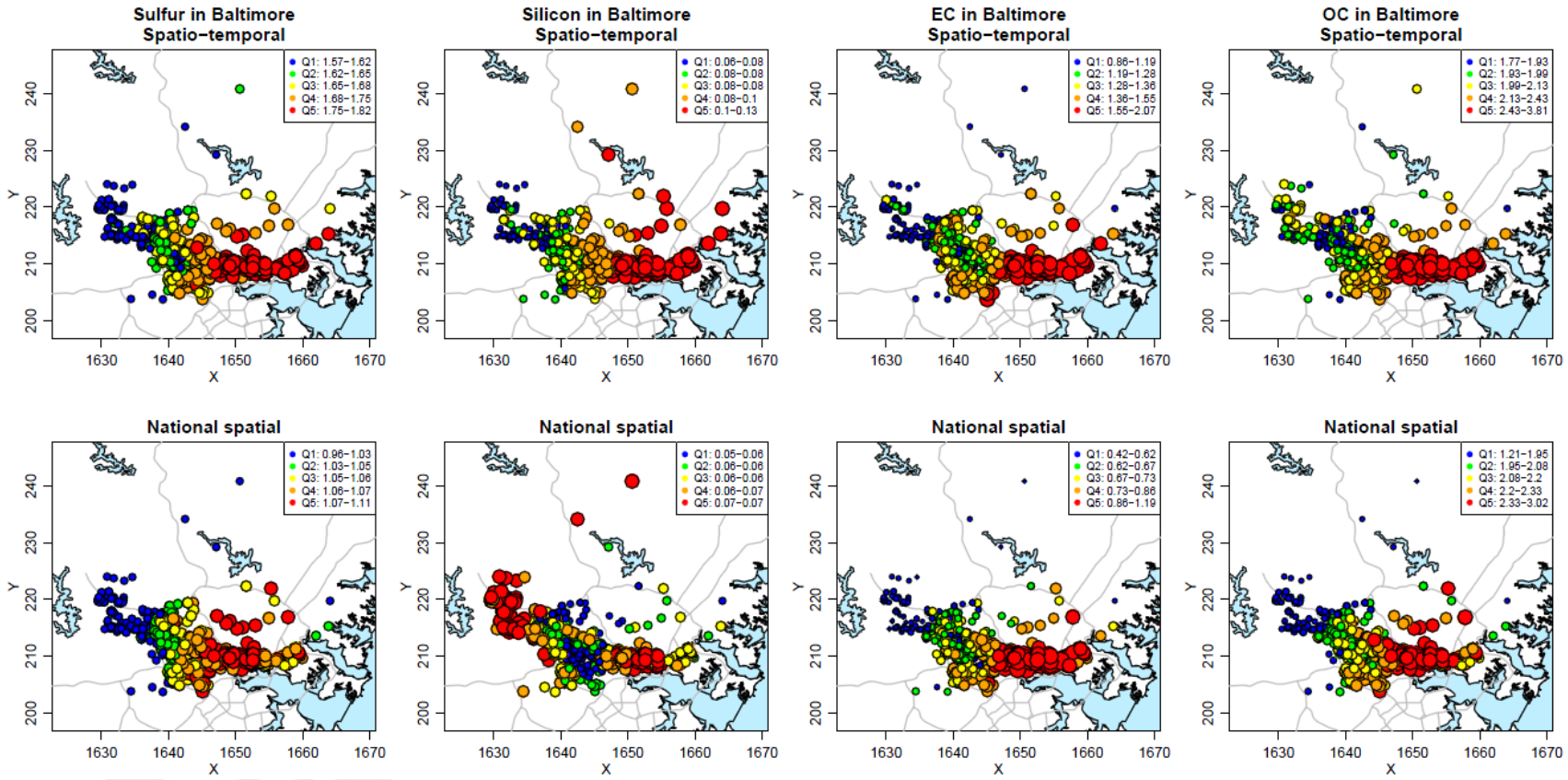


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

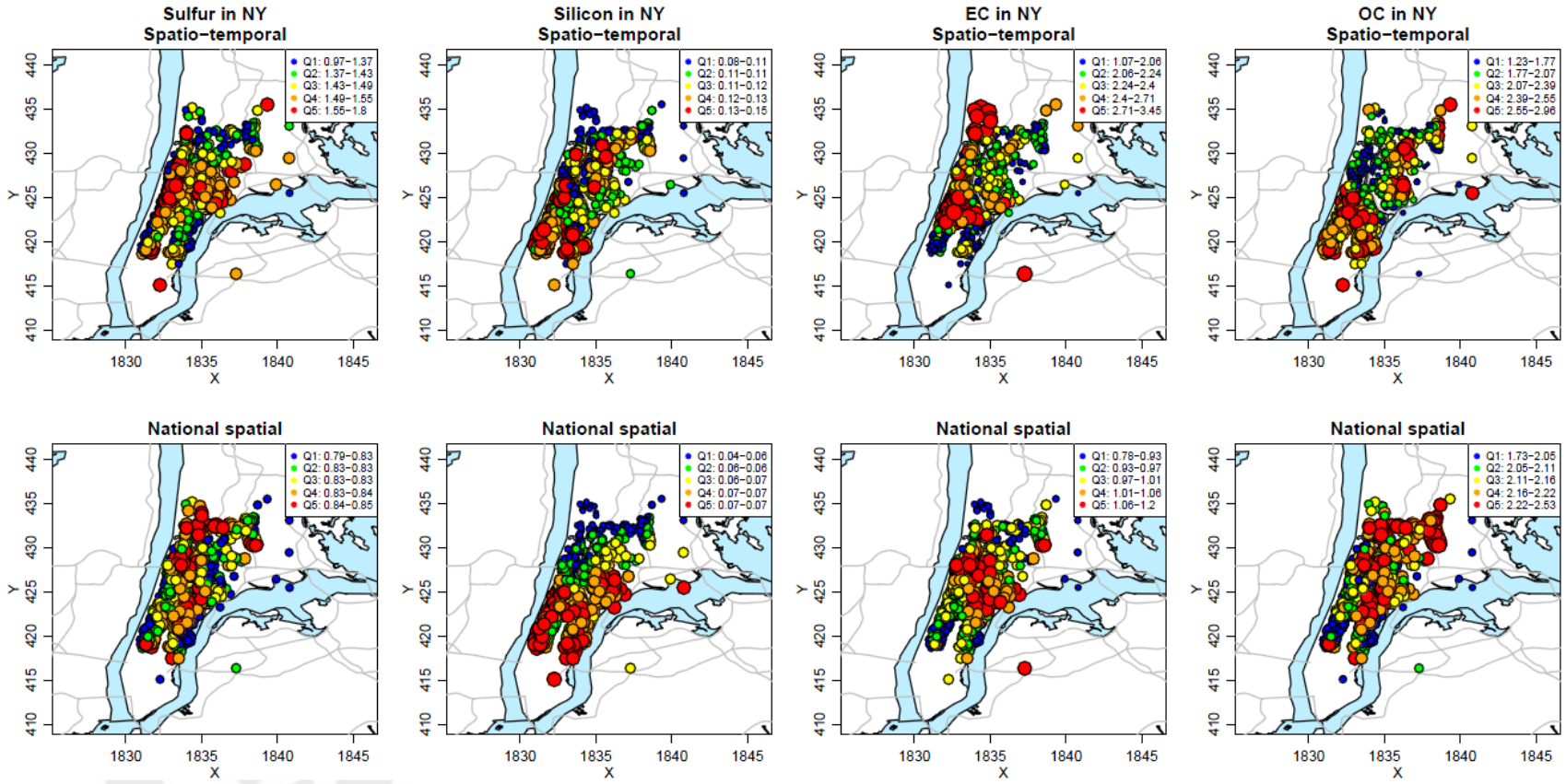


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

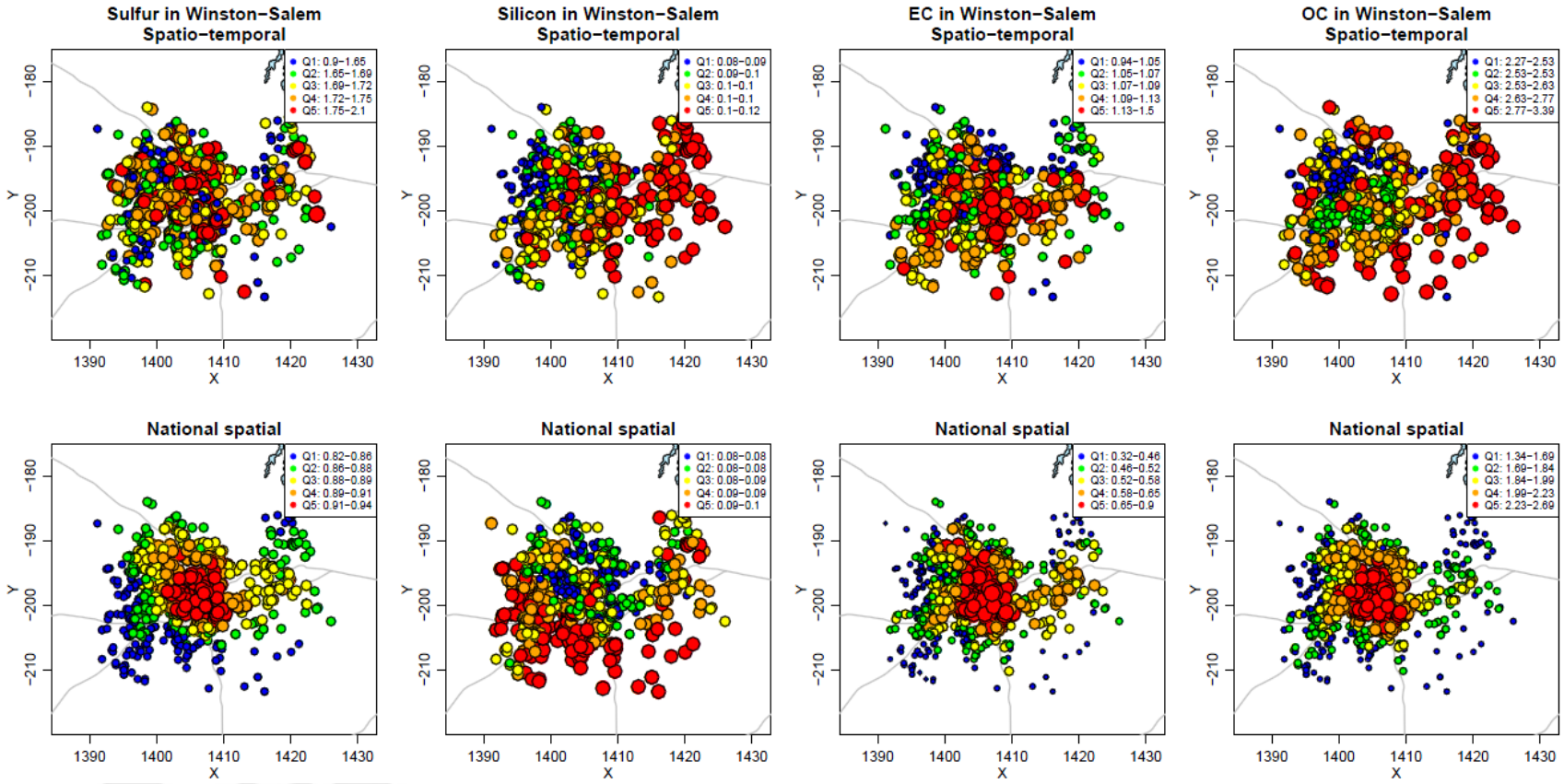


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive



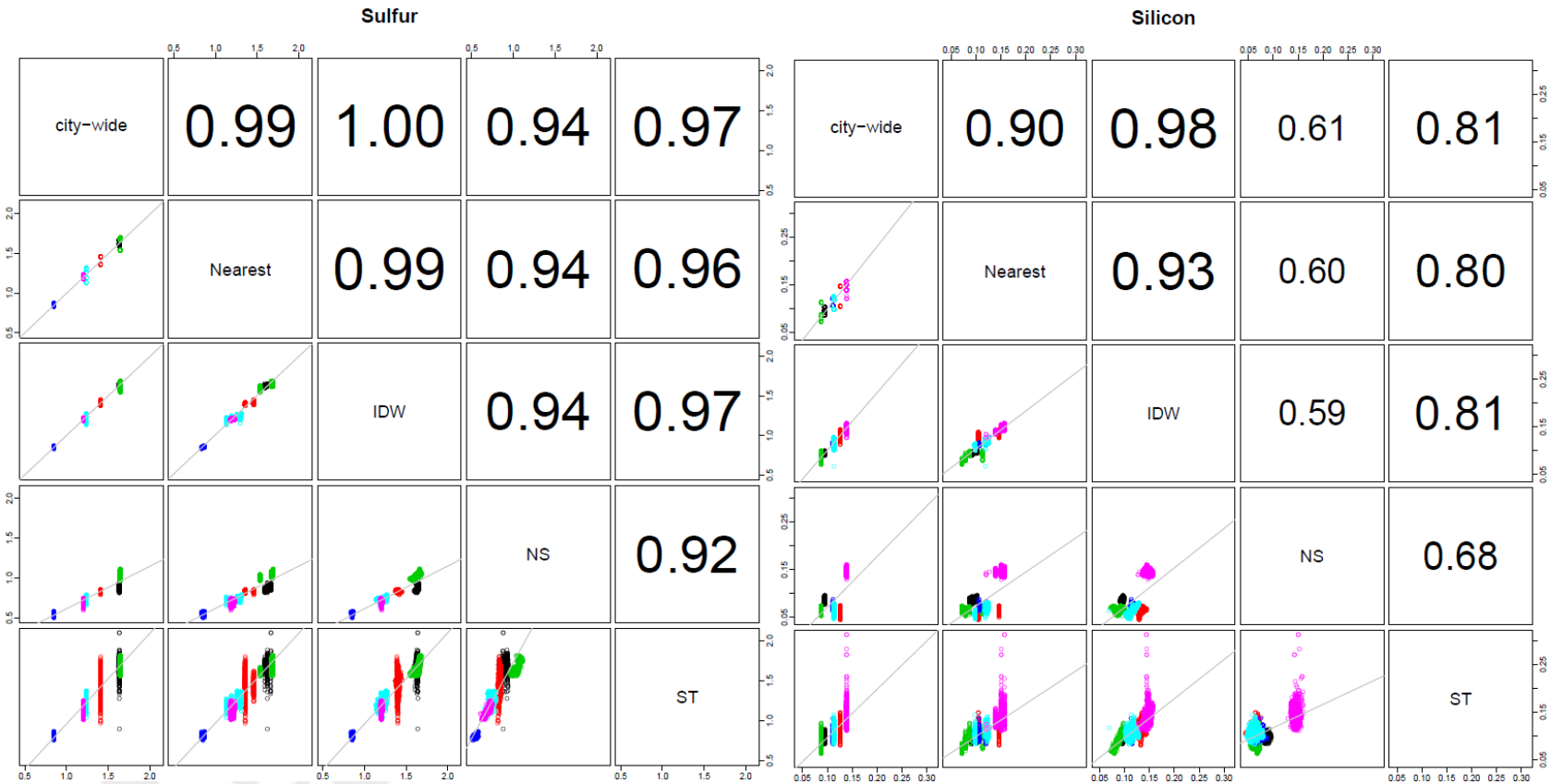
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press



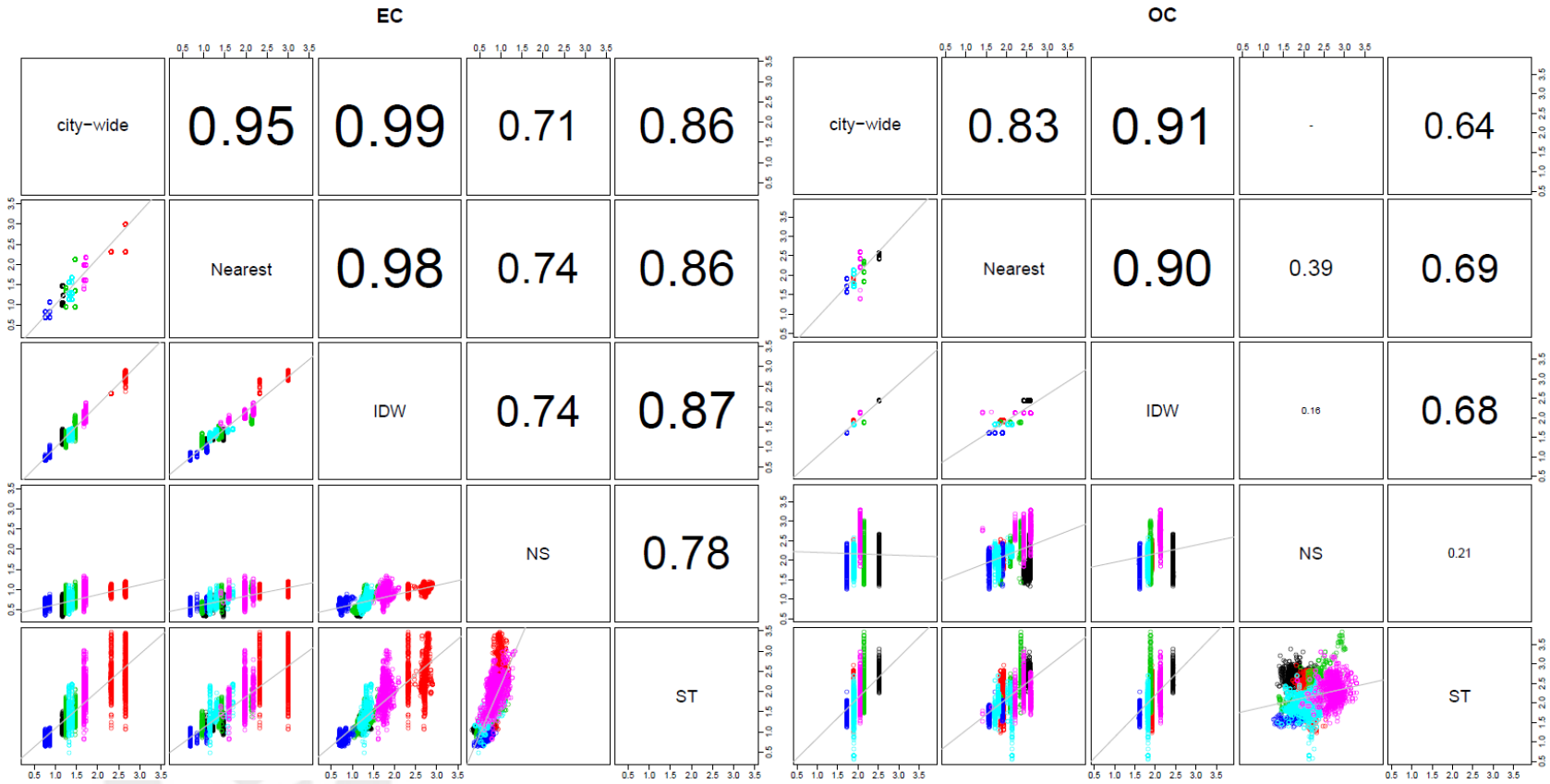
Supplemental Figure 4. Predicted long-term concentrations of four $PM_{2.5}$ components from the spatio-temporal and the national spatial models at participant locations in five cities (different colors represent quintiles of the range of concentrations)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive



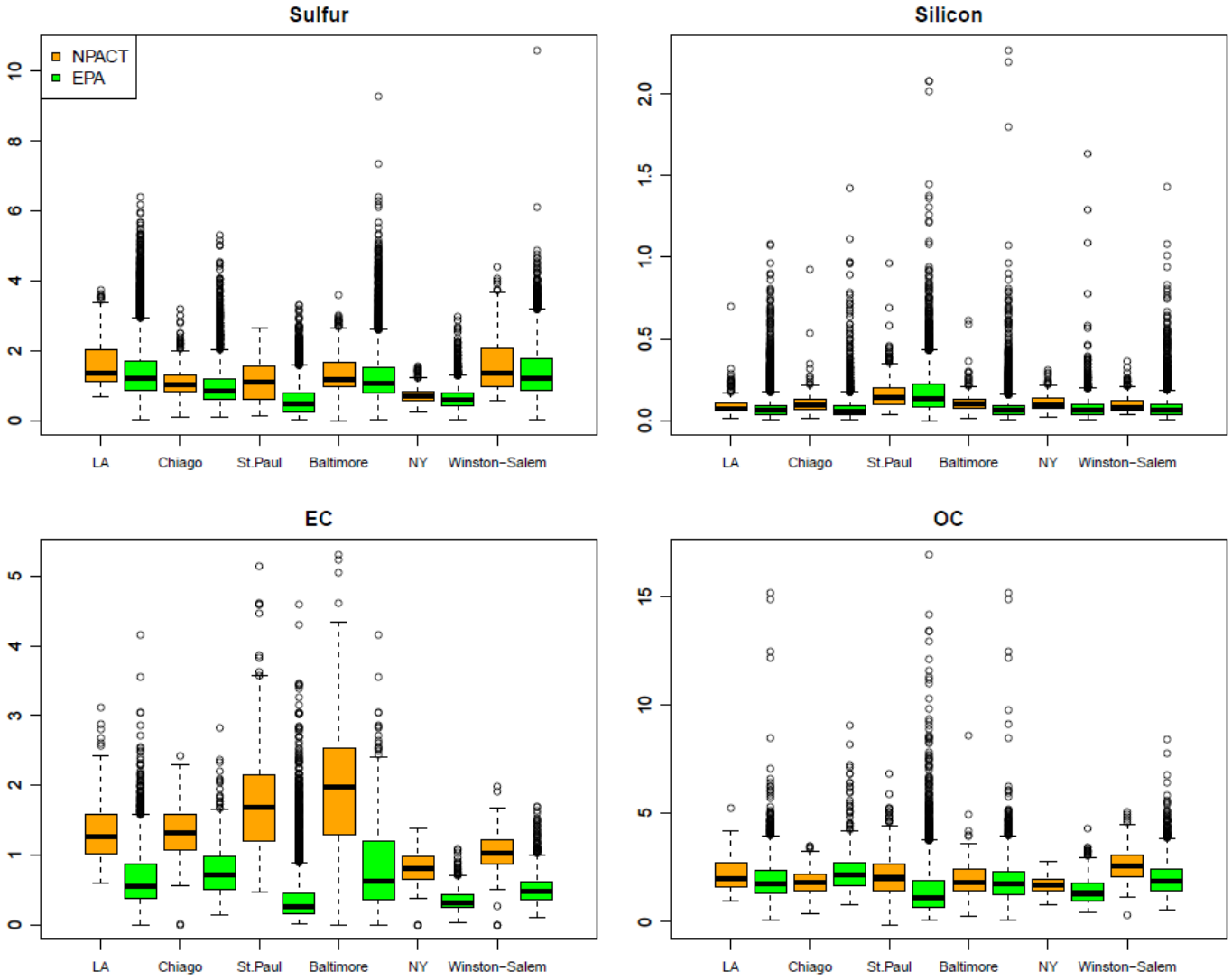
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press



Supplemental Figure 5. Scatter plots and correlation coefficients of predicted long-term concentrations of four $PM_{2.5}$ components between the five different prediction models (city-wide, nearest monitor, IDW (inverse distance weighting), NS (national spatial model), and ST (spatio-temporal model)) for MESA participant addresses within 10 kilometers of any NPACT/MESA Air monitors in six MESA Air cities (color code: black = Winston-Salem, red = NY, green = Baltimore, blue = St. Paul, light blue = Chicago, and pink = LA).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive



Supplemental Figure 6. Box plots of measurements for four $PM_{2.5}$ components by NPACT/MESA Air and CSN/IMPROVE monitoring campaign by six MESA city areas defined by 200 kilometers within the centers of six MESA cities (2 week samples for NPACT/MESA Air and daily samples on every 3rd and 6 day schedule for EPA)

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press