ENVIRONMENTAL HEALTH PERSPECTIVES

http://www.ehponline.org

A Unified Spatiotemporal Modeling Approach for Predicting Concentrations of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution

Joshua P. Keller, Casey Olives, Sun-Young Kim, Lianne Sheppard, Paul D. Sampson, Adam A. Szpiro, Assaf P. Oron, Johan Lindström, Sverre Vedal, and Joel D. Kaufman

http://dx.doi.org/10.1289/ehp.1408145

Received: 17 January 2014
Accepted: 11 November 2014
Advance Publication: 14 November 2014

This article will be available in its final, 508-conformant form 2–4 months after Advance Publication. If you require assistance accessing this article before then, please contact Dorothy L. Ritter, *EHP* Web Editor. *EHP* will provide an accessible version within 3 working days of request.

NIH National Institute of Environmental Health Sciences

A Unified Spatiotemporal Modeling Approach for Predicting Concentrations of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution

Joshua P. Keller,[1] Casey Olives,[2] Sun-Young Kim,[2,3] Lianne Sheppard,[1,2] Paul D. Sampson,[4] Adam A. Szpiro,[1] Assaf P. Oron,[5] Johan Lindström,[6] Sverre Vedal,[2] and Joel D. Kaufman[2]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA; [2]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, Washington, USA; [3]Institute of Health and Environment, Seoul National University, Seoul, Korea; [4]Department of Statistics, University of Washington, Seattle, Washington, USA; [5]Core for Biomedical Statistics, Seattle Children's Hospital, Seattle, Washington, USA; [6]Centre for Mathematical Science, Lund University, Lund, Sweden

Address correspondence to Joshua P. Keller, Department of Biostatistics, University of Washington, Health Sciences Building, Box 357232, 1705 NE Pacific Street, Seattle, WA 98195 USA. Telephone: 206-543-1044. E-mail: kellerjp@u.washington.edu

Running title: Spatiotemporal model for multiple air pollutants

Acknowledgments: This publication was developed under a STAR research assistance agreement, No. RD831697 (MESA Air), awarded by the U.S Environmental Protection Agency. It has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and the EPA does not endorse any products or commercial services mentioned in this publication. Additional support was provided by EPA grants RD-83479601-0

and CR-834077101-0, NIEHS grant K24ES013195, and the Biostatistics, Epidemiologic, and Bioinformatic Training in Environmental Health Training Grant from the NIEHS (T32 ES015459).

Competing financial interests: None of the authors has any actual or potential competing financial interests.

Abstract

Background: Cohort studies of the relationship between air pollution exposure and chronic health effects require predictions of exposure over long periods of time.

Objectives: We developed a unified modeling approach for predicting fine particulate matter, nitrogen dioxide, oxides of nitrogen, and black carbon (as measured by light absorption coefficient) in six U.S. metropolitan regions from 1999 through early 2012 as part of the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air).

Methods: We obtained monitoring data from regulatory networks and supplemented those data with study-specific measurements collected from MESA Air community locations and participants' homes. In each region, we applied a spatiotemporal model that included a long-term spatial mean, time trends with spatially varying coefficients, and a spatiotemporal residual. The mean structure was derived from a large set of geographic covariates that was reduced using partial least squares regression. We estimated time trends from observed time series and used spatial smoothing methods to borrow strength between observations.

Results: Prediction accuracy was high for most models, with cross-validation R^2 (R^2_{CV}) greater than 0.80 at regulatory and fixed sites for most regions and pollutants. At home sites, overall R^2_{CV} ranged from 0.45 to 0.92 and temporally adjusted R^2_{CV} ranged from 0.23 to 0.92.

Conclusions: This novel spatiotemporal modeling approach provides accurate fine-scale predictions in multiple regions for four pollutants. We have generated participant-specific predictions for MESA Air in order to investigate health effects of long-term air pollution exposures. These successes highlight modeling advances that can be adopted more widely in modern cohort studies.

Introduction

Cohort studies have shown an association between long-term exposure to air pollution and cardiovascular morbidity and mortality (Brook et al. 2010; Miller et al. 2007; Pope and Dockery 2006). To estimate these associations, epidemiologic studies develop exposure prediction models to predict pollutant concentrations over long periods of time at cohort home addresses based on monitoring data from regulatory networks or study-specific monitoring campaigns. While early models were based on region-wide averages or nearest-monitor approaches, more current methods include land-use regression (LUR) (Hoek et al. 2008; Jerrett et al. 2007), the use of satellite and remote sensing data (Kloog et al. 2011), geostatistical methods such as kriging (Beelen et al. 2009), generalized additive models (Hart et al. 2009), or a combination of these techniques (Beckerman et al. 2013; Bergen et al. 2013; Mercer et al. 2011).

The Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air) is investigating the association between long-term air pollution exposure and measures of cardiovascular health (Kaufman et al. 2012). MESA Air is following a cohort of more than 6000 individuals in six metropolitan regions: Baltimore, Maryland; Chicago, Illinois; Los Angeles, California; New York, New York; St. Paul, Minnesota; and Winston-Salem, North Carolina. The primary exposures of interest in MESA Air are fine particulate matter ($PM_{2.5}$), nitrogen dioxide (NO_2), oxides of nitrogen (NO_X), and black carbon, as measured by Light Absorption Coefficient (LAC). One goal of MESA Air is the development of advanced statistical methods that incorporate extensive supplemental monitoring to improve the prediction of intra-urban pollutant variability. MESA Air health effect analyses require spatiotemporal predictions of ambient outdoor concentrations for all four pollutants in all six metropolitan regions for times ranging from 1999 through 2012.

In general, exposure prediction models developed in one city do not transfer well to another city (Allen et al. 2011), and so prediction models are often study- and city-specific (e.g. Franklin et al. 2012). A challenge for multi-city studies like MESA Air that combine data from sub-cohorts and include several pollutant measures is generating predictions that are of comparable quality across pollutants and cities. Here we present a unified and flexible spatiotemporal modeling framework for the four MESA Air pollutants. We apply a standardized approach to model selection for all pollutants and regions, allowing the intrinsic flexibility of our modeling framework to account for differences in the way pollution processes behave in different regions.

Methods

To predict outdoor concentration of pollutants at MESA Air participant residences, we fit a separate spatiotemporal exposure prediction model for each pollutant ($PM_{2.5}$, NO_2, NO_X, and LAC) in each metropolitan region. Briefly, our model decomposes the space-time field of concentrations into spatially varying long-term (i.e. duration of study period) averages, spatially varying seasonal and long-term trends, and spatially correlated but temporally independent residuals, and accommodates data from the complex monitoring design described below. We modeled on a two-week time scale due to the two-week sampling of MESA Air supplementary monitoring instruments (see below); this allows for flexible aggregation of predictions to time scales of interest for health effects analyses (e.g. twelve months prior to clinic visit).

Monitoring data

We used three sources of outdoor air monitoring data. We obtained hourly NO_2 and NO_X and daily $PM_{2.5}$ concentration measurements in each metropolitan region from January 1, 1999 to March 31, 2012 from the Environmental Protection Agency (EPA) Air Quality System (AQS),

including data from the Interagency Monitoring of Protected Visual Environments (IMPROVE) network (U.S. EPA 2013). No AQS data were used for black carbon due to differences in collection methods from the MESA Air LAC measurement methods described below. We aggregated hourly data into daily averages, and subsequently averaged daily values to the two-week scale. AQS monitors that had less than two years of data or had irregular temporal coverage (e.g., operated only in the summer) were not used.

In each metropolitan region, we defined the modeling area to be locations within approximately 75 km of each metropolitan center (see Figure 1). AQS monitors within the modeling regions were considered for inclusion in the model, and predictions at participant residences were restricted to locations within these modeling regions. In New York, MESA Air participants were recruited from both New York City and Rockland County, and so the modeling region included locations near both areas. In Winston-Salem, only one AQS monitoring location for NO_2 and NO_X met inclusion criteria. In order to have a complete time series for the 14-year modeling period, an AQS monitor in Charlotte, NC was included for estimating time trends. In Chicago, the modeling region was further restricted to locations west of -87.5°W longitude due to some covariates being unavailable east of that meridian. In Los Angeles, only locations south and/or west of the San Gabriel Mountains were included.

To better capture the within-city variability of pollutant concentrations, MESA Air conducted a umentary monitoring campaign targeting the study cohort from July 2005 through August 2009. The MESA Air measurements were two-week cumulative measurements that began and ended on Wednesdays. Measurements of NO_2 and NO_X were made using Ogawa passive samplers and $PM_{2.5}$ mass was measured on Harvard Personal Environmental Monitor impactors using Teflon

filters. LAC was computed from the Teflon filters via reflectance. A detailed description of the data collection and site selection procedures has been previously published (Cohen et al. 2009).

The MESA Air monitoring campaign included three types of monitoring sites: fixed, home, and snapshot. Fixed sites were operated for the duration of the four-year MESA Air sampling period to provide long time-series of measurements, with one fixed site collocated with an AQS monitor in each region. Samples of participant residences in each metropolitan region were selected for monitoring as home outdoor sites on a rotating basis, with a majority of locations monitored 1 to 3 times in different seasons. Snapshot sites, which measured only NO_2 and NO_X, were located in clusters to capture gradients near sources (e.g., primary roadways) and monitored for three two-week periods, one each in winter, summer, and either spring or fall.

In New York City, data from the New York City Community Air Survey (NYCCAS) were used to supplement the AQS and MESA Air data (Matte et al. 2013; NYC Department of Health 2014). The NYCCAS data consist of two-week measurements of $PM_{2.5}$, NO_2, NO_X, and LAC collected during December 2008–December 2010 in a manner consistent with MESA Air sampling protocols. Five NYCCAS reference sites (one in each borough) collected measurements throughout the sampling period, and 150 NYCCAS distributed sites were monitored once per season during this time. Because of the similarity in monitoring scheme, we treated NYCCAS reference sites in the same manner as MESA Air fixed sites, and NYCCAS distributed sites in the same manner as MESA Air home sites, in our models. The NYCCAS data and a small subset of the MESA Air two-week data were centered on different weeks than the majority of the MESA Air measurements. To align these measurements with the rest of the MESA Air data, we treated these measurements as if they were made one week earlier or later, as appropriate.

Between 0.4% (LAC) and 1.6% (NO_2) of the pollutant measurements were below the limit of detection (LOD) and were replaced with the value LOD/2. The number of each type of monitoring site by region and pollutant is provided in Table 1. The range of the number of $PM_{2.5}$ observations at each monitoring site during the study period is provided in Table 2, along with summary statistics for the site means. Corresponding statistics for NO_2, NO_X, and LAC observations are provided in Supplemental Material Tables S1 through S3.

Geographic covariates

We compiled more than 300 geographic covariates for use in the model (Supplemental Material, Table S4). These covariates included proximity measures (distance to nearest major road, intersection, truck route, railway, railyard, coastline, airport, and port) and buffer measures (major road length, truck route length, land-use category, long-term vegetation index, population density, and emission sources). We included a long-term average of the dispersion model output from a modified implementation of the Caline3QHCR line-source model (Eckhoff and Braverman 1995). The Caline3QHCR model incorporates distance, traffic volume, meteorology, and diurnal traffic patterns in each region.

Geographic covariates with minimal variation or potentially highly influential values were excluded from the modeling process. Specifically, variables were removed if (a) more than 80% of monitoring sites had the same value, (b) more than 2% of observations were more than five standard deviations away from mean, (c) the standard deviation of the distribution of values at participant residences was more than five times the standard deviation of the distribution of values at monitoring locations, or (d) the maximum value was 10% among all monitoring sites (for land-use variables only). These filters were applied separately for each pollutant and region.

Spatiotemporal model

The monitoring data were highly unbalanced, with a small number of locations providing long time series of several years' duration and a larger number of locations providing broader spatial coverage, but at a relatively small number of time points. A hierarchical spatiotemporal model had been previously developed to accommodate the unbalanced nature of the MESA Air data (Lindström et al. 2013; Sampson et al. 2011; Szpiro et al. 2010). This model can be written as

$$C(s,t) = \mu(s,t) + v(s,t), \qquad [1]$$

where $C(s,t)$ represents the log-transformed two-week average pollutant concentration at location s and time t. The $\mu(s,t)$ term is the spatiotemporal mean surface and the $v(s,t)$ term represents spatiotemporal residual variation.

We break down the spatiotemporal mean into components

$$\mu(s,t) = \beta_0(s) + \Sigma_{i=1}^{m} \beta_i(s) f_i(t), \qquad [2]$$

where $\beta_0(s)$ is the long-term mean at location s, $f_i(t)$ are smooth time trends, and $\beta_i(s)$ are spatially-varying coefficients for the time trends.

The time trends are estimated from AQS and MESA Air fixed sites (and NYCCAS reference sites in New York) using a procedure developed by Fuentes et al. (2007) and previously described in detail by Sampson et al. (2011). In brief, we applied an Expectation-Maximization procedure to fill in missing values in the time series and derived the trends from a singular value decomposition. We smoothed the trends using splines, controlling the smoothness with the degrees of freedom (df) parameter. The model assumes the time trends account for enough of the temporal structure that the residuals are uncorrelated in time.

The long-term averages $\beta_0(s)$ and time trend coefficients $\beta_i(s)$ are modeled as spatial random fields with a spatial mean, distributed as

$$\beta_i \sim N(\boldsymbol{X}_i(s)\boldsymbol{\alpha}_i, \Sigma_i(\phi_i, \sigma_i, \tau_i)), i = 0,1,\ldots, m. \quad [3]$$

Here, $\boldsymbol{X}_i(s)$ are reduced-dimension summaries of the geographic covariates (described in detail below) at location s and $\boldsymbol{\alpha}_i$ are vectors of coefficients to be estimated. The covariance structure for β_i, denoted by Σ_i, is either an independence model with variance τ_i or a spatial smoothing model with exponential covariance function parameterized by range ϕ_i, partial sill σ_i, and nugget τ_i (Cressie 1993).

The zero-mean spatiotemporal residual term $v(s,t)$ in [1] has a spatial correlation structure and is assumed independent at each time point. It includes a random effect for each time point to model short-term variations that affect an entire region, such as large-scale meteorological events.

PLS scores

Rather than include each of the hundreds of geographic covariates directly in the model or use variable selection methods, we reduced the dimensionality of the covariates using partial least squares (PLS). In a manner similar to principal components analysis (PCA), PLS computes linear combinations, called scores, of the columns of a data matrix. Unlike PCA, the PLS procedure constructs scores that maximize the covariance between the scores and an outcome rather than the variance between the scores. A technical explanation of the PLS algorithm is provided by Abdi (2010). Sampson *et al.* (2013) described the application of PLS for spatial models, and here we describe how we applied the method to spatiotemporal data.

PLS regression requires a single outcome value for each location. Because the MESA Air data are unbalanced time series, we first derived values that could be used as outcomes in PLS regression. For each AQS, fixed, and NYCCAS reference site s, we regressed the time series of observations $C(s,t)$ on the smoothed time trends using ordinary least squares regression with mean function $E[C(s,t)] = \gamma_0^s f_0(t) + \ldots + \gamma_m^s f_m(t)$ to get estimates $(\hat{\gamma}_0^s, \ldots, \hat{\gamma}_m^s)$ for each location. For each time trend, PLS regression was performed separately with the $\hat{\gamma}_i^s$ as the outcomes and the matrix of geographic covariates as the predictors. This gave a set of PLS scores for each location that was different for each time trend. PLS scores at home and snapshot monitoring sites were predicted using the geographic variables at those locations and the score definitions defined from the regression at fixed sites. PLS regression was performed using the pls package (Mevik et al. 2011) in R 2.15.1 (R Core Team). Scores were included in the model as the $\boldsymbol{X}_i(s)$ in [3].

Parameter estimation and model selection

Once the PLS scores $\boldsymbol{X}_i(s)$ and time trends $f_i(t)$ were computed, the remaining parameters were calculated via maximum likelihood using the SpatioTemporal package (Lindström et al. 2012) in R. We varied several model parameters and used cross-validation to find the best-fitting model in each metropolitan region, as described below. We considered different values for the number of time trends (either 1 or 2), the df for smoothing time trends (either 4 or 8 per year), the number of PLS scores per time trend (2 or 3), and the covariance structure of the β_i fields (spatial smoothing or no spatial smoothing).

Cross-validation procedure

The primary interest of MESA Air is in long-term average exposures, and so we assessed model performance using cross-validation of long-term averages (LTAs). Because the highly

unbalanced structure of the monitoring data means that LTAs at home sites are computed from a handful of observations of a few weeks' duration, while LTAs at fixed sites are computed from long time series, we performed cross-validation separately for each site type. For home sites and NYCCAS distributed sites, we used ten-fold cross-validation, which leaves out one-tenth of the data in turn. For AQS, fixed, and NYCCAS references sites we used leave-one-out cross-validation since the total number of monitors was relatively small. For snapshot sites, we used ten-fold cross-validation, with monitors in the same cluster left out together. For all three schemes, the covariance parameters (but not the time trends or PLS scores) were re-estimated using all but the left-out sites. Pollutant concentrations at the left-out sites were predicted using the parameters estimated from the remaining data.

We assessed cross-validation performance using two measures: root mean-squared error (RMSE) and cross-validation R^2 (denoted by R^2_{CV}). Letting y_j denote the mean observed value and $\hat{y}_j$ the mean of the predicted values for the observed time points at monitoring site j, RMSE and R^2_{CV} were computed on the original scale of the data according to the formulas

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \qquad [4]$$

$$R^2_{CV} = max\,(0, 1 - RMSE^2/MSE_{obs}), \qquad [5]$$

where $MSE_{obs} = 1/n\ \Sigma_{j=1}^{n}\ (y_j - \overline{y})^2$ is the mean-squared error of the observed values. R^2_{CV} provides a measure of fit to the 1-1 line, in contrast to the typical regression-based R^2 (R^2_{CVreg}), which measures fit to the regression line and is computed as the square of the correlation coefficient between the cross-validation predictions and the observed values. R^2_{CV} reflects the contrast of interest since our goal is accurate prediction at unmeasured locations, and it is typically lower

than R^2_{CVreg}. Although R^2_{CV} was the primary metric for our model evaluation, we present R^2_{CVreg} for comparison with published results from other authors. Because we are most interested in spatial contrasts between individual exposures within each region, we prioritized home-site over fixed-site R^2_{CV} and RMSE in the model selection process.

In the context of the hierarchical model [1], it is challenging to separate the spatial and temporal contributions to R^2_{CV} for cross-validation of temporally sparse datasets like the home sites and NYCCAS distributed sites. Lindström et al. (2013) proposed three temporally adjusted adaptations of R^2_{CV} that use data from the AQS and fixed sites as the reference MSE instead of MSE_{obs} in [5] in order to focus on spatial prediction accuracy. R^2_{Avg} uses the average values at AQS and fixed sites within that region. R^2_{Close} uses the closest (in absolute distance) AQS or fixed site. R^2_{Smooth} uses the smoothed time trend at the closest AQS or fixed site.

Prediction at participant locations

Using the best models from each metropolitan region, predictions of pollutant log-concentrations at participant residences were made on a two-week scale from January 1999 through March 2012. We back-transformed these predictions using exponentiation to return them to the original scale of concentration measurements and computed averages of two-week predictions over the study period.

Results

Model structure

Table 3 provides an overview of the model structure selected for each metropolitan region and pollutant. Most models have only one time trend, although all the New York models have two.

The two smoothed trends for the NO_2 model in Los Angeles are shown in Figure 2. Figure 2 also includes plots of the fitted trends for a selected AQS site and fixed site.

For $PM_{2.5}$, there was noticeable heterogeneity of the best models across metropolitan regions (Table 3). New York and Winston-Salem had no spatial smoothing in the long-term $PM_{2.5}$ average (β_0(s)) and no model had spatial smoothing in the $PM_{2.5}$ time trend coefficients ($\beta_i(s)$). Half of the regions had two time trends, while the other half had only a single time trend for $PM_{2.5}$. For NO_2, all of the models had spatial smoothing for the long-term average, and the same was true for NO_X except in New York.

The relative contribution of geographic covariates to the PLS scores varied by region and pollutant. Figure S1 in the Supplemental Material shows the correlations between covariates and PLS scores for the NO_2 model in Chicago, which are representative of the general patterns in the other regions (data not shown). Overall, the distance-to-feature covariates and vegetation measures (Normalized Difference Vegetation Index [NDVI] and low development, open development, forest and wetland land use) tended to have the opposite correlation from emissions and traffic measures (A1, A2/A3, and truck route lengths and intersection counts) within buffers.

Model results

Table 4 shows the cross-validation metrics for all models, broken down by pollutant and region. These metrics assess how well the site means are modeled, incorporating both the spatial and temporal components of the predictions. Scatterplots of predictions and observed values are provided in Figure 3 for AQS and fixed sites and Figure S2 in the Supplemental Material for

home sites. Metrics for the snapshot sites (for NO_2 and NO_X) and for AQS and fixed sites (all four pollutants) on the two-week scale are reported in Supplemental Material, Tables S5 and S6.

Predictive accuracy was generally good (R^2_{CV} >0.6) to excellent (R^2_{CV} >0.8) in almost all regions for each pollutant. NO_X models in Baltimore and Los Angeles had the best performance at MESA home sites (R^2_{CV} of 0.92 and 0.91, respectively) (Table 4). The lowest R^2_{CV} for home sites was in the Chicago NO_2 model (0.45), but its RMSE (3.31 ppb) was comparable to that in New York and Los Angeles (3.82 and 3.13 ppb, respectively). At AQS and fixed sites, R^2_{CV} was very good (0.70 for LAC in Los Angeles) to excellent (0.98 for NOx in St. Paul), with two notable exceptions: Winston-Salem NO_X (0.00) and St. Paul $PM_{2.5}$ (0.45). However, in both cases the RMSE was comparable to the corresponding RMSE for models in other regions. The small range of observed data (9.8 to 22.4 ppb) (Supplemental Material, Table S2) and the small number of monitors in Winston-Salem (Table 1) explain the low R^2_{CV} for NO_X in that city.

Table 5 provides three versions of temporally adjusted R^2_{CV} at home sites (R^2_{Avg}, R^2_{Close}, and R^2_{Smooth}). For NO_2 and NO_X, these temporally adjusted R^2_{CV} are fairly similar to the unadjusted R^2_{CV} reported in Table 4, suggesting that we are predicting spatial differences well. For $PM_{2.5}$, however, the temporally adjusted R^2_{CV} are consistently lower than the unadjusted R^2_{CV}.

Boxplots of the long-term averages of predictions at participant residences are provided in Figure 4. On average, predicted concentrations tended to be higher in New York and Los Angeles, consistent with the higher observed monitoring values in those regions. Variability in predictions is also greatest in these two cities, especially in the tails of the distributions.

Table S7 in the Supplemental Material provides performance metrics for the New York models when the NYCCAS data were excluded from the modeling process. Without the NYCCAS data, R^2_{CV} was noticeably lower for $PM_{2.5}$ (0.79 versus 0.91 at AQS and fixed sites, and 0.36 versus 0.54 at home sites) and LAC (0.55 versus 0.75 at AQS and fixed sites, and 0.43 versus 0.51 at home sites).

Discussion

We present here a complex and successful approach to predicting long-term air pollution concentrations for application in a cohort study. While this approach was tailored to this particular well-characterized cohort study—taking advantage of cohort-specific monitoring for example—the success of the approach demonstrates modeling improvements that can be adopted for application in future population-based research on spatially varying pollutants. We believe that this approach to capturing variation in within-region pollution highlights advances that should be adopted in the next generation of air pollution cohort studies, both for understanding contrasts at relatively low concentrations in the U.S. and at the higher concentrations experienced globally.

We describe a unified framework for implementing exposure prediction models of four air pollutants in six metropolitan regions that easily incorporates spatially- and temporally unbalanced monitoring data. The application of a consistent modeling framework to all regions and pollutants is important for studies like MESA Air that use exposure estimates from multiple sub-cohorts together in health analyses. Although we applied the same approach in all regions, we varied the model structure to best fit the data for each region and pollutant. This unified modeling approach was shown to have very good model performance ($R^2_{CV} > 0.70$) for almost all

of the pollutants and regions. The architecture for this modeling approach is publicly available through the SpatioTemporal R package (Lindström et al. 2012).

As a result of the success of our spatiotemporal modeling approaches, we are confident in using these approaches to model outdoor pollutant concentrations in epidemiological analyses in this cohort and in other populations residing in these same communities. We have also found that implementation of portions of this approach, such as PLS regression of a large set of geographic covariates combined with spatial smoothing via universal kriging, can be used with good success in other regions to predict pollutant concentrations without the same level of small-area monitoring (Sampson et al. 2013). The NYCCAS data increased the spatial density of the monitoring data in New York, which was likely one reason for the improved model performance when the data were included. For LAC, the NYCCAS data provided particular benefit because they allowed the model to be extended through 2010, which would not have been possible with only the MESA Air data.

A majority of models included spatial smoothing in the long-term average. This suggests that while PLS scores derived from geographic covariates can predict much of the spatial variation in the data, benefit is gained from borrowing strength across observations nearby in space.

Differences in the underlying pollutant variability likely caused some of the differences seen in temporally adjusted R^2_{CV}. $PM_{2.5}$ tended to exhibit less small-scale spatial variation and greater temporal variability, leading to temporally adjusted R^2_{CV} that are noticeably lower than the unadjusted measures. The NO_2 and NO_X data tended to exhibit greater spatial variability, and the similarity of the unadjusted and temporally adjusted R^2_{CV} values suggest that the unadjusted R^2_{CV} are not overly inflated by well-predicted temporal variation.

The modeling approach presented here does have several limitations. First, we used geographic covariates that were constant in time (although the modeling framework readily extends to spatiotemporal covariates; see Lindström et al. (2013)). Changes in these variables likely occurred during the study decade, but we nonetheless believe that the time-constant geographic variables still provided a useful means to predict long-term pollutant concentrations. Secondly, the calculation of PLS scores was limited to AQS and fixed sites because they had long time series. For LAC in particular, this means that the scores were based on a very small number of locations because the LAC model relied only on MESA Air data (plus NYCCAS data for New York). Thirdly, the cross-validation model selection procedure conditioned on the time trends and PLS scores. Overfitting may have occurred in the cross-validation of the AQS and fixed sites, since the left-out observations were used to estimate the time trends and PLS scores. However, since the home sites were not used in time trend estimation or in defining the PLS scores, any overfitting was restricted to the AQS and fixed site cross-validation. This provides further motivation for prioritization of cross-validation metrics from home sites when selecting the best models.

Conclusions

Our unified spatiotemporal modeling method successfully characterized outdoor concentrations of multiple air pollutants at the homes of cohort members in multiple metropolitan regions. This flexible and powerful modeling approach can incorporate an unbalanced monitoring data structure, leveraging data from supplemental monitoring campaigns that increase the spatial coverage of monitoring data. The method was easily transferred between regions and pollutants, allowing for straightforward comparison between model fits across regions. While aspects of our techniques are particularly tailored to the unique data and resources of MESA Air, lessons

learned here can be applied to understand the spatial and temporal variation of pollutants in future cohort studies. Advances in fine-scale modeling resolved in both space and time are important for the next generation of cohort studies assessing health effects of environmental agents.

References

Abdi H. 2010. Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdiscip Rev Comput Stat 2:97–106; doi:10.1002/wics.51.

Allen RW, Amram O, Wheeler AJ, Brauer M. 2011. The transferability of NO and NO2 land use regression models between cities and pollutants. Atmos Environ 45:369–378; doi:10.1016/j.atmosenv.2010.10.002.

Beckerman BS, Jerrett M, Serre M, Martin R V, Lee S-J, van Donkelaar A, et al. 2013. A hybrid approach to estimating national scale spatiotemporal variability of PM2.5 in the contiguous United States. Environ Sci Technol 47:7233–41; doi:10.1021/es400039u.

Beelen R, Hoek G, Pebesma E, Vienneau D, de Hoogh K, Briggs DJ. 2009. Mapping of background air pollution at a fine spatial scale across the European Union. Sci Total Environ 407:1852–67; doi:10.1016/j.scitotenv.2008.11.048.

Bergen S, Sheppard L, Sampson PD, Kim S-Y, Richards M, Vedal S, et al. 2013. A National Prediction Model for PM2.5 Component Exposures and Measurement Error-Corrected Heatlh Effect Inference. Environ Health Perspect 121:1017–1025.

Brook RD, Rajagopalan S, Pope CA, Brook JR, Bhatnagar A, Diez-Roux A V, et al. 2010. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. Circulation 121:2331–78; doi:10.1161/CIR.0b013e3181dbece1.

Cohen MA, Adar SD, Allen RW, Avol E, Curl CL, Gould T, et al. 2009. Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Environ Sci Technol 43:4687–93.

Cressie N. 1993. Statistics for Spatial Data. Revised Ed. John Wiley & Sons.

Eckhoff P, Braverman T. 1995. Addendum to the user's guide to CAL3QHC version 2.0 (CAL3QHCR user's guide).

Franklin M, Vora H, Avol E, McConnell R, Lurmann F, Liu F, et al. 2012. Predictors of intra-community variation in air quality. J Expo Sci Environ Epidemiol 22:135–47; doi:10.1038/jes.2011.45.

Fuentes M, Guttorp P, Sampson PD. 2007. Using transforms to analyze space-time processes. In *Statistics of Spatio-Temporal Systems* (Held and Finkenstadt eds.), pp. 77–151, RC/Chapman Hall.

Hart J, Yanosky J, Puett R, Ryan L, Dockery D, Smith T, et al. 2009. Spatial Modeling of PM10 and NO2 1985-2000 in the Continental US. Environ Health Perspect 117:1690–1697; doi:10.1289/ehp.0900840.

Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, et al. 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 42:7561–7578; doi:10.1016/j.atmosenv.2008.05.057.

Jerrett M, Arain M, Kanaroglou P, Beckerman B, Crouse D, Gilbert NL, et al. 2007. Modeling the intraurban variability of ambient traffic pollution in Toronto, Canada. J Toxicol Environ Health A 70:200–12; doi:10.1080/15287390600883018.

Kaufman JD, Adar SD, Allen RW, Barr RG, Budoff MJ, Burke GL, et al. 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 176:825–37; doi:10.1093/aje/kws169.

Kloog I, Koutrakis P, Coull BA, Lee HJ, Schwartz J. 2011. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmos Environ 45:6267–6275; doi:10.1016/j.atmosenv.2011.08.066.

Lindström J, Szpiro AA, Sampson PD, Bergen S, Oron AP. 2012. SpatioTemporal: Spatio-Temporal Model Estimation. R package version 1.1.1.

Lindström J, Szpiro AA, Sampson PD, Oron AP, Richards M, Larson T V, et al. 2013. A flexible spatio-temporal model for air pollution with spatio-temporal covariates. Environ Ecol Stat; doi:10.1007/s10651-013-0261-4.

Matte TD, Ross Z, Kheirbek I, Eisl H, Johnson S, Gorczynski JE, et al. 2013. Monitoring intraurban spatial patterns of multiple combustion air pollutants in New York City: Design and implementation. J Expo Sci Environ Epidemiol 23:223–31; doi:10.1038/jes.2012.126.

Mercer LD, Szpiro AA, Sheppard L, Lindström J, Adar SD, Allen RW, et al. 2011. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Atmos Environ 45:4412–4420; doi:10.1016/j.atmosenv.2011.05.043.

Mevik B-H, Wehrens R, Liland KH. 2011. pls: Partial Least Squares and Principal Component regression. R package version 2.3-0.

Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, et al. 2007. Long-term exposure to air pollution and incidence of cardiovascular events in women. N Engl J Med 356:447–458.

NYC Department of Health. 2014. New York City Community Air Survey. Available: http://www.nyc.gov/health/nyccas [accessed 13 January 2014].

Pope CA, Dockery DW. 2006. Health Effects of Fine Particulate Air Pollution: Lines that Connect. J Air Waste Manage Assoc 56:709–742; doi:10.1080/10473289.2006.10464485.

Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson T V, et al. 2013. A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. Atmos Environ 75:383–392; doi:10.1016/j.atmosenv.2013.04.015.

Sampson PD, Szpiro AA, Sheppard L, Lindström J, Kaufman JD. 2011. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. Atmos Environ 45:6593–6606; doi:10.1016/j.atmosenv.2011.04.073.

Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD. 2010. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21:606–631; doi:10.1002/env.

U.S. EPA. 2013. Air Quality SyStem Data: Query AQS Data [digital data set]. Available: http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm

Table 1. Number of monitors by site type, region, and pollutant.

Site type	$PM_{2.5}$	NO_2	NO_X	LAC
Baltimore, MD				
AQS	29	11	8	
MESA Fixed	5	5	5	5
MESA Home	86	87	87	86
MESA Snapshot		104	104	
Chicago, IL				
AQS	20	7	6	
MESA Fixed	6	6	6	6
MESA Home	136	113	113	136
MESA Snapshot		129	129	
Los Angeles, CA				
AQS	23	29	30	
MESA Fixed	7	7	7	7
MESA Home	113	120	120	113
MESA Snapshot		252	250	
New York, NY				
AQS	45	17	11	
MESA Fixed	3	3	3	3
MESA Home	107	119	118	107
MESA Snapshot		157	157	
NYCCAS Reference	5	5	5	5
NYCCAS Distributed	150	150	150	150
St. Paul, MN				
AQS	13	5	5	
MESA Fixed	3	4	4	3
MESA Home	126	132	132	129
MESA Snapshot		107	107	
Winston-Salem, NC				
AQS	16	2	2	
MESA Fixed	4	4	4	4
MESA Home	114	117	117	114
MESA Snapshot		121	121	

Table 2. Summary of $PM_{2.5}$ monitoring data.

Site type	Number of observations per site		Site means ($\mu g/m^3$)		
	Min	Max	Min	Max	Mean ± SD
Baltimore, MD					
AQS	64	345	10.9	16.9	13.4 ± 1.4
MESA Fixed	18	92	12.1	15.4	13.7 ± 1.25
MESA Home	1	3	7.3	22.7	14.3 ± 3.1
Chicago, IL					
AQS	71	320	11.7	16.4	14.0 ± 1.3
MESA Fixed	6	87	12.2	14.0	13.1 ± 0.75
MESA Home	1	4	5.2	19.5	11.5 ± 3.2
Los Angeles, CA					
AQS	82	345	10.7	22.8	16.2 ± 3.5
MESA Fixed	76	85	13.7	19.3	16.2 ± 2.0
MESA Home	1	2	0.7	42.6	16.9 ± 6.1
New York, NY					
AQS	51	342	9.3	17.1	12.5 ± 1.8
MESA Fixed	49	83	11.5	15.7	13.7 ± 2.1
MESA Home	1	3	3.5	41.6	15.1 ± 4.9
NYCCAS Reference	51	52	8.8	9.9	9.4 ± 0.42
NYCCAS Distributed	6	8	6.8	19.8	11.0 ± 11.0
St. Paul, MN					
AQS	55	305	7.9	11.6	10.0 ± 0.91
MESA Fixed	81	89	9.6	10.5	10.0 ± 0.46
MESA Home	1	5	5.0	27.4	10.3 ± 3.8
Winston-Salem, NC					
AQS	86	346	10.3	15.9	13.4 ± 1.5
MESA Fixed	80	93	13.0	13.8	13.4 ± 0.35
MESA Home	1	4	9.0	22.8	14.3 ± 2.6

Table 3. Model structure for the best model (selected by cross-validation) for each pollutant and metropolitan region.

Model	No. of time trends[a]	No. of PLS scores[b]	df/year in time trend[c]	Spatial smoothing[d]	
				Long-term average (β_0)	Time trend coefficients (β_i)
Baltimore, MD					
$PM_{2.5}$	1	3	4	Yes	No
NO_2	1	2	8	Yes	No
NO_X	1	2	8	Yes	Yes
LAC	1	3	8	No	No
Chicago, IL					
$PM_{2.5}$	1	3	8	Yes	No
NO_2	2	2	4	Yes	Yes
NO_X	2	2	8	Yes	No
LAC	1	2	8	Yes	Yes
Los Angeles, CA					
$PM_{2.5}$	2	3	8	Yes	No
NO_2	2	3	8	Yes	Yes
NO_X	1	3	4	Yes	Yes
LAC	1	2	4	Yes	No
New York, NY					
$PM_{2.5}$	2	3	8	No	No
NO_2	2	3	4	Yes	Yes
NO_X	2	2	8	No	No
LAC	2	2	4	Yes	No
St. Paul, MN					
$PM_{2.5}$	1	3	4	Yes	No
NO_2	1	3	4	Yes	No
NO_X	1	3	4	Yes	No
LAC	1	2	8	Yes	No
Winston-Salem, NC					
$PM_{2.5}$	2	2	4	No	No
NO_2	1	3	8	Yes	Yes
NO_X	1	2	8	Yes	Yes
LAC	1	2	8	Yes	No

[a]Selected from either 1 or 2 time trends. [b]PLS, Partial Least Squares. Selected from either 2 or 3 PLS scores. Scores were covariates in the mean component of the long-term average (β_0) and time trend (β_i) fields (denoted by $\boldsymbol{X}_i(s)$ in [3]). [c]Selected from either 4 or 8 degrees of freedom (df) per year. Controls smoothness of estimated time trends. [c]Yes, exponential covariance structure. No, independent covariance structure.

Table 4. Cross-validation measures of predictive accuracy for site means at monitoring locations.

Region	AQS and MESA Fixed sites			MESA Home sites		
	RMSE[a]	R^2_{CV}	R^2_{CVreg}	RMSE[a]	R^2_{CV}	R^2_{CVreg}
$PM_{2.5}$						
Baltimore	0.42	0.90	0.90	1.24	0.84	0.86
Chicago	0.59	0.78	0.82	1.43	0.80	0.80
Los Angeles	1.28	0.83	0.84	2.92	0.77	0.78
New York[b]	0.59	0.91	0.91	2.80	0.54	0.56
St. Paul	0.60	0.45	0.84	1.78	0.78	0.79
Winston-Salem	0.44	0.89	0.90	1.00	0.85	0.86
NO_2						
Baltimore	0.76	0.96	0.97	1.47	0.90	0.90
Chicago	1.51	0.87	0.97	3.31	0.45	0.48
Los Angeles	2.23	0.88	0.89	3.13	0.77	0.78
New York[b]	1.86	0.92	0.93	3.82	0.78	0.78
St. Paul	1.27	0.93	0.94	1.24	0.87	0.87
Winston-Salem	0.95	0.85	0.98	1.41	0.74	0.75
NO_X						
Baltimore	3.32	0.92	0.96	3.98	0.92	0.92
Chicago	3.88	0.87	0.91	6.08	0.59	0.59
Los Angeles	6.74	0.87	0.87	5.69	0.91	0.92
New York[b]	8.85	0.61	0.89	16.66	0.50	0.50
St. Paul	1.69	0.98	0.98	3.58	0.83	0.84
Winston-Salem	5.46	0.00	0.94	3.74	0.60	0.63
LAC						
Baltimore	0.096	0.87	0.91	0.127	0.78	0.79
Chicago	0.045	0.86	0.92	0.108	0.61	0.62
Los Angeles	0.114	0.70	0.93	0.266	0.69	0.71
New York[b]	0.147	0.75	0.79	0.329	0.51	0.52
St. Paul	0.043	0.91	0.92	0.074	0.69	0.69
Winston-Salem	0.020	0.94	0.99	0.088	0.64	0.65

[a]Units for RMSE are $\mu g/m^3$ ($PM_{2.5}$), ppb (NO_2 and NO_X), and 10^{-5}/m (LAC). [b]New York models include NYCCAS reference sites with AQS and Fixed sites, and NYCCAS distributed sites with Home sites.

Table 5. Temporally adjusted cross-validation measures of predictive accuracy for home site means.

Region	R^2_{Avg}	R^2_{Close}	R^2_{Smooth}
$PM_{2.5}$			
Baltimore	0.45	0.52	0.72
Chicago	0.23	0.33	0.64
Los Angeles	0.40	0.23	0.43
New York[a]	0.48	0.36	0.38
St. Paul	0.23	0.29	0.62
Winston-Salem	0.29	0.60	0.77
NO_2			
Baltimore	0.92	0.79	0.74
Chicago	0.73	0.64	0.78
Los Angeles	0.63	0.66	0.66
New York[a]	0.89	0.78	0.64
St. Paul	0.77	0.89	0.90
Winston-Salem	0.73	0.79	0.81
NO_X			
Baltimore	0.86	0.70	0.65
Chicago	0.76	0.73	0.69
Los Angeles	0.81	0.85	0.88
New York[a]	0.72	0.64	0.52
St. Paul	0.79	0.88	0.85
Winston-Salem	0.43	0.62	0.64
LAC			
Baltimore	0.78	0.67	0.32
Chicago	0.56	0.45	0.36
Los Angeles	0.28	0.34	0.48
New York[a]	0.59	0.65	0.53
St. Paul	0.67	0.80	0.84
Winston-Salem	0.37	0.56	0.59

General formula for R^2 measures is $R^2 = max(0, 1 - RMSE^2/MSE_{obs})$. R^2_{Avg} uses the mean-squared error of the average observed values at AQS and Fixed sites within the region as MSE_{obs}. R^2_{Close} uses the mean-squared error of the observed values at the closest AQS or Fixed site as MSE_{obs}. R^2_{Smooth} uses the mean squared error of the values of the smoothed time trend at the nearest AQS or Fixed site as MSE_{obs}.

[a]Includes NYCCAS distributed sites.

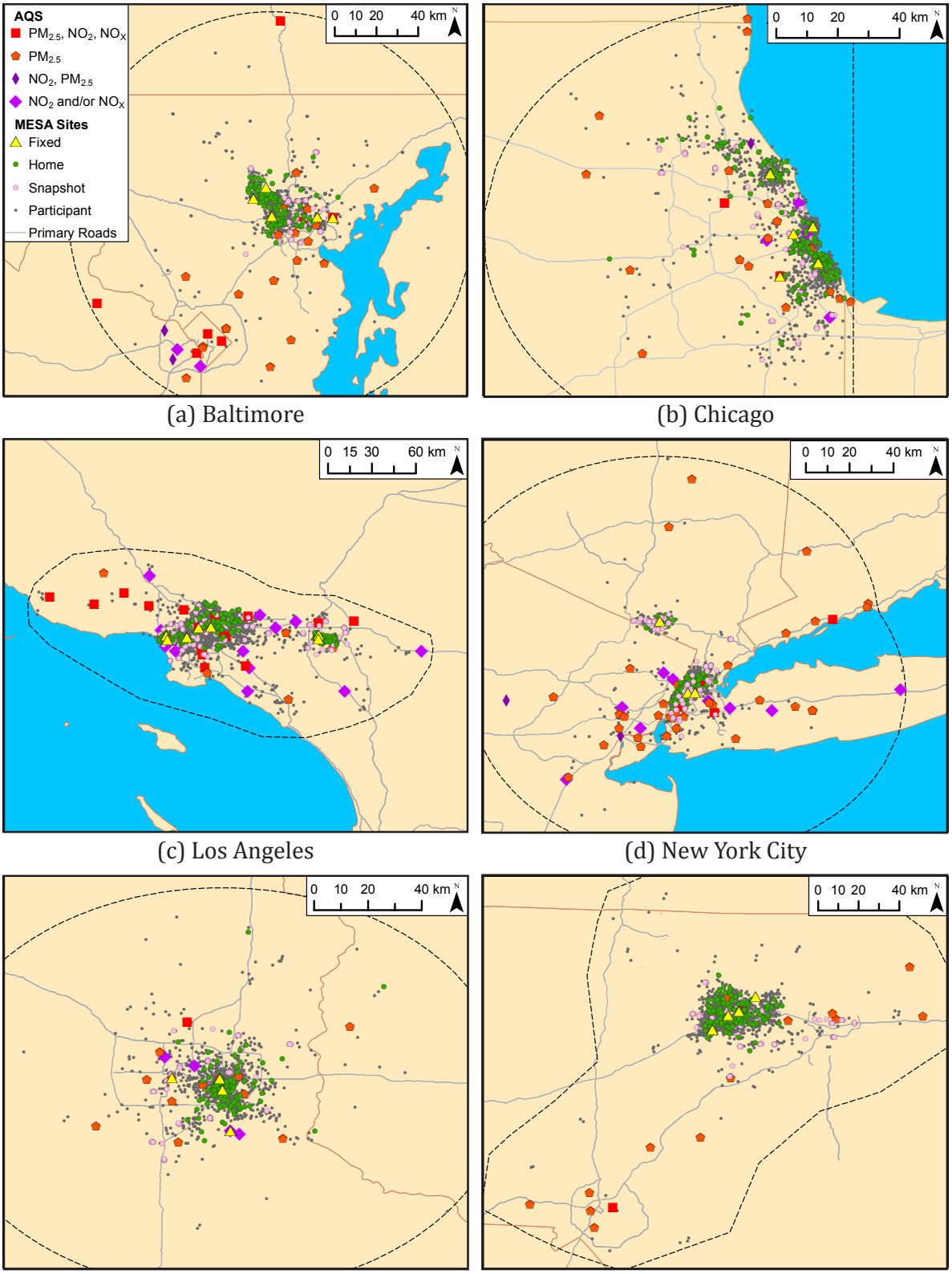
Figure 1. Maps of the modeling areas (denoted by dashed black line) in the six metropolitan regions, including monitor and subject locations. Abbreviations: Fixed, MESA Air Fixed monitoring sites; Home, MESA Air Home monitoring sites; Snapshot, MESA Air Snapshot monitoring sites; Participant, MESA Air participant residence location (moved slightly to protect confidentiality).

Figure 2. Time trends for the NO_2 model in Los Angeles. The top panel shows the smooth time trends calculated from AQS and Fixed sites. The middle and bottom panels show the observed data and fitted trends at an AQS site and Fixed site, respectively.

Figure 3. Long-term averages of cross-validated predictions and observations for AQS and fixed monitoring locations for each pollutant.

Figure 4. Pollutant- and region-specific boxplots of long-term averages of predictions from 1999 through early 2012 at participant residence locations. Metropolitan region abbreviations: Bal, Baltimore; Chi, Chicago; LA, Los Angeles; NY, New York; SP, St. Paul; W-S, Winston-Salem. Boxes extend from the 25th to the 75th percentile, horizontal bars represent the median, whisker extend 1.5 times the length of the interquartile range (IQR) above and below the 75th and 25th percentiles, respectively, and outliers are presented as points.

Figure 1.



(a) Baltimore (b) Chicago

(c) Los Angeles (d) New York City

(e) St. Paul (f) Winston-Salem

Figure 2.

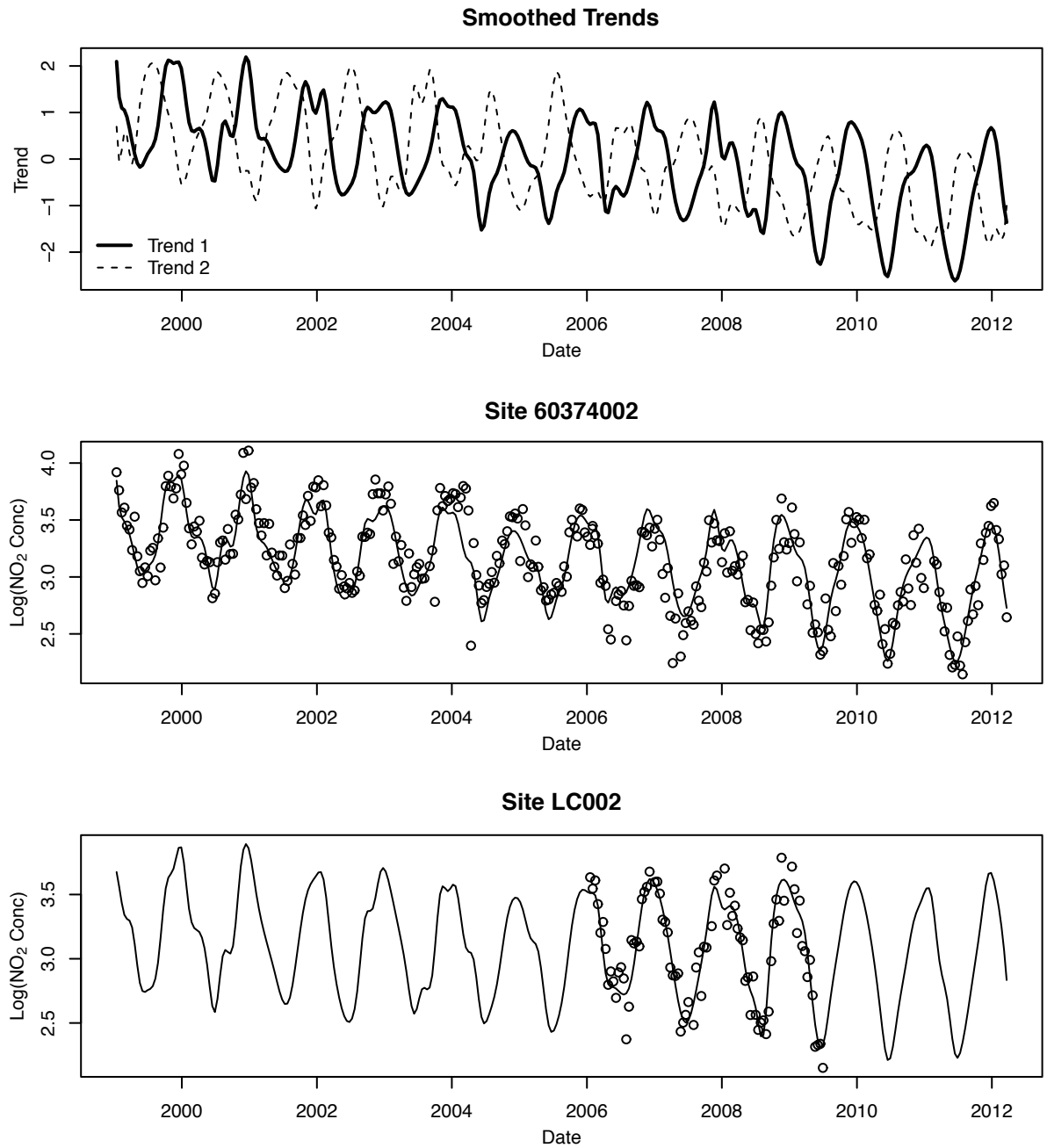

Smoothed Trends
Trend
Trend 1
Trend 2
Date
Site 60374002
Log(NO2 Conc)
Date
Site LC002
Log(NO2 Conc)
Date

Figure 3.

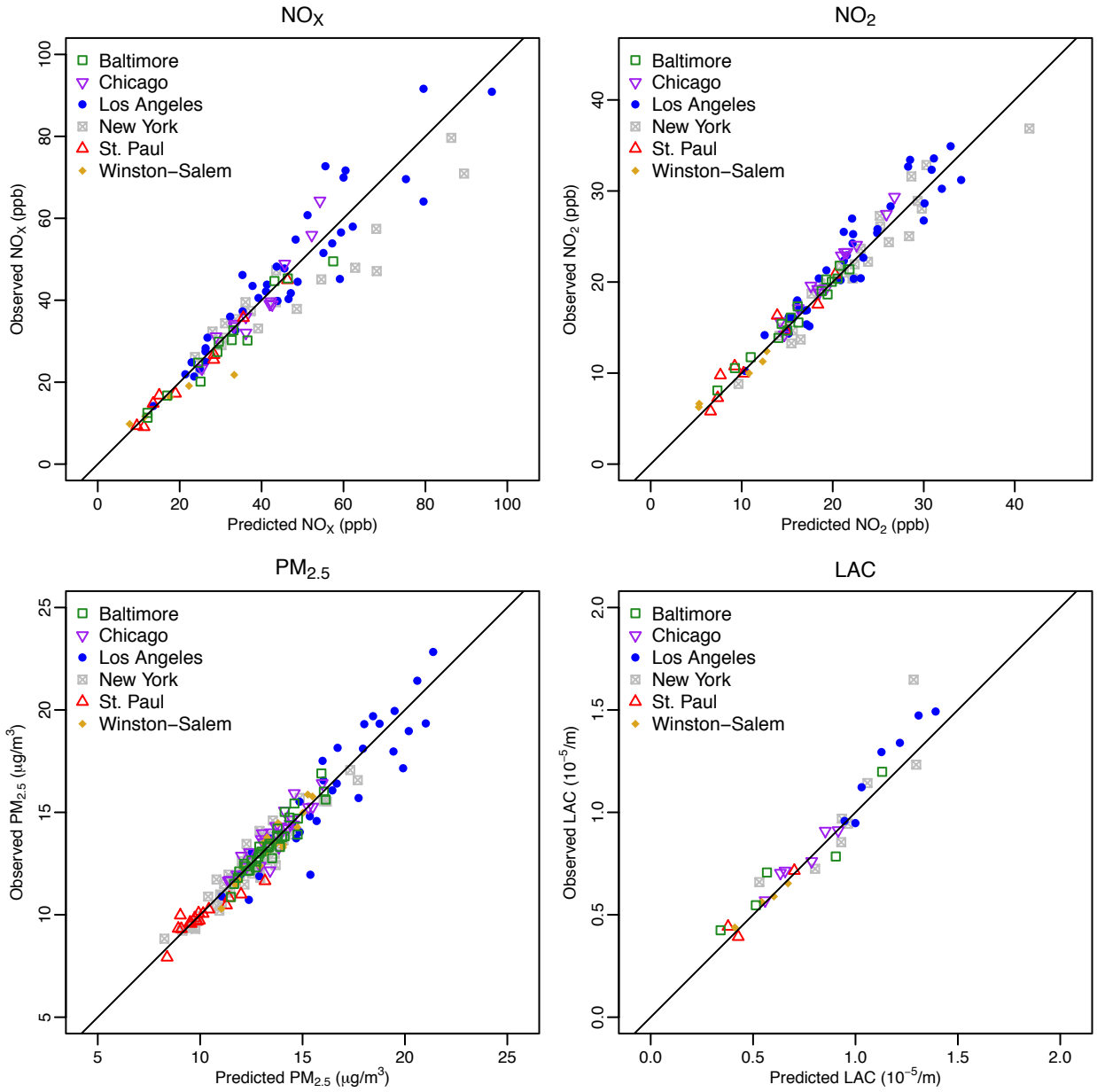
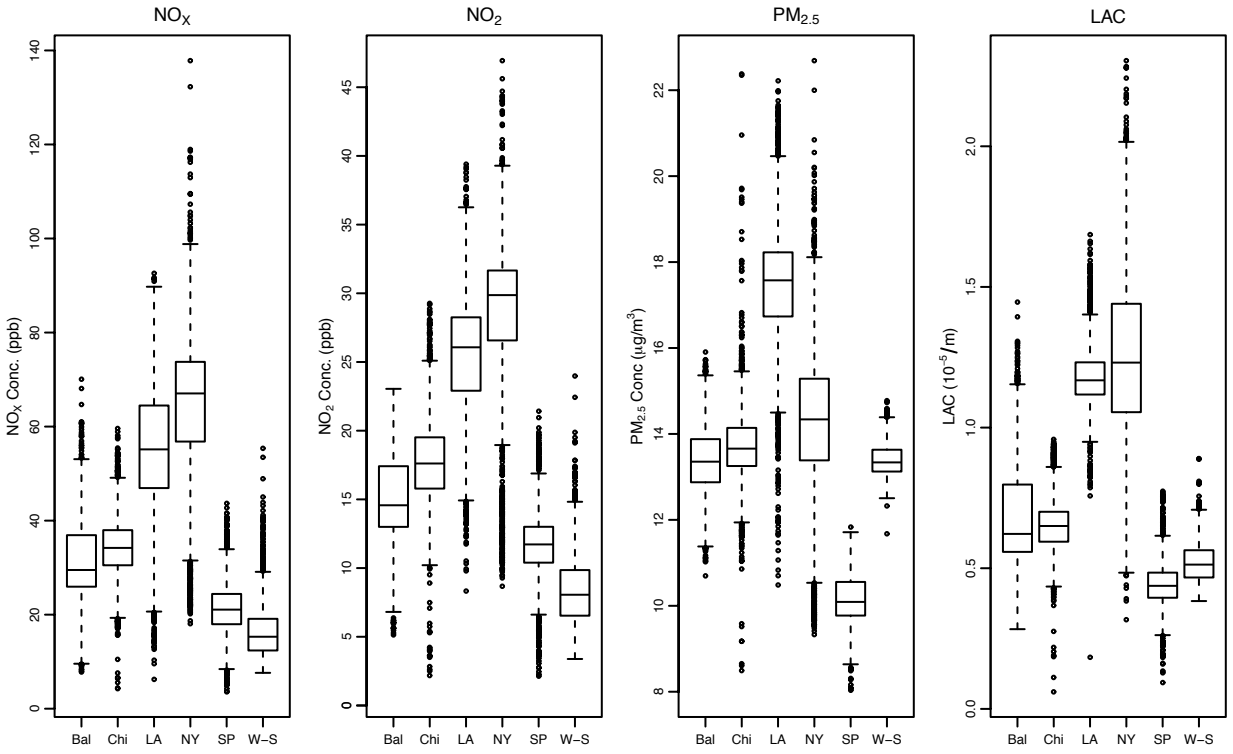

NOX
Baltimore
Chicago
Los Angeles
New York
St. Paul
Winston–Salem
Observed NOX (ppb)
Predicted NOX (ppb)
NO2
Observed NO2 (ppb)
Predicted NO2 (ppb)
PM2.5
Observed PM2.5 (μg/m3)
Predicted PM2.5 (μg/m3)
LAC
Observed LAC (10−5/m)
Predicted LAC (10−5/m)

Figure 4.


NOX
NOX Conc. (ppb)
NO2
NO2 Conc. (ppb)
PM2.5
PM2.5 Conc (µg/m3)
LAC
LAC (10−5/m)
Bal
Chi
LA
NY
SP
W−S