NIH Public Access
Author Manuscript
Environ Sci Technol. Author manuscript; available in PMC 2014 July 02.

Published in final edited form as:
Environ Sci Technol. 2013 July 2; 47(13): 7233–7241. doi:10.1021/es400039u.

A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of $PM_{2.5}$ in the Contiguous United States

Bernardo S. Beckerman[*], **Michael Jerrett**, **Marc Serre**, **Randall V. Martin**, **Seung-Jae Lee**, **Aaron van Donkelaar**, **Zev Ross**, **Jason Su**, and **Richard T. Burnett**

Abstract

Airborne fine particulate matter exhibits spatiotemporal variability at multiple scales, which presents challenges to estimating exposures for health effects assessment. Here we created an model to predict ambient particulate matter less than 2.5 microns in aerodynamic diameter ($PM_{2.5}$) across the contiguous United States to be applied to health effects modeling. We developed a hybrid approach combining a land use regression model (LUR) selected with a machine learning method, and Bayesian Maximum Entropy (BME) interpolation of the LUR space-time residuals. The $PM_{2.5}$ dataset included 104,172 monthly observations at 1,464 monitoring locations with approximately 10% of locations reserved for cross-validation. LUR models were based on remote sensing estimates of $PM_{2.5}$, land use and traffic indicators. Normalized cross-validated R^2 values for LUR were 0.63 and 0.11 with and without remote sensing, respectively, suggesting remote sensing is a strong predictor of ground-level concentrations. In the models including the BME interpolation of the residuals, cross-validated R^2 were 0.79 for both configurations; the model without remotely sensed data described more fine-scale variation than the model including remote sensing. Our results suggest that our modeling framework can predict ground-level concentrations of $PM_{2.5}$ at multiple scales over the contiguous U.S.

INTRODUCTION

Epidemiologic studies aimed at estimating the health effects of ambient air pollution require exposure assessments that accurately describe the expected variability in the pollutant of interest. Air pollution varies at multiple spatiotemporal scales[1]. Spatially, this occurs at the local scale from immediate sources and over larger spatial areas from secondary reactions and transport mechanisms. Temporally, the majority of variation in air pollution comes from changing traffic patterns and meteorological conditions. Improving exposure assessment techniques to estimate this kind of variability is important for deriving defensible health effect estimates. Potential misclassification in the expected spatiotemporal variability of the exposure will likely introduce bias in health effects estimation. Exposure misclassification is a critical concern for health researchers at any time[2] but especially when a large number of subjects reside across many metropolitan areas or states. The ability to leverage statistical power from a large sample size will be reduced by exposure misclassification.

Air pollution exposure models used to derive estimates for health studies fall into several classes: (1) nearest monitor value assignments[3–4]; (2) weighted averages of proximate monitoring locations [5–6]; (3) geostatistical methods such as kriging[7–8]; (4) both

Corresponding Author: Bernardo S. Beckerman, < beckerman@berkeley.edu>, (510) 643-9257.

Supporting Information
Additional figures and a more detailed description of some of the methods are given in the Supporting Information. Also, instructions on how to obtain the DSA package and BMElib are also provided. This material is available free of charge via the Internet at http://pubs.acs.org

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Bayesian[9–12] and Non-Bayesian multivariate models[13–16], which include land use regression; (5) point, line and area source dispersion models[17–18]; (6) estimates derived from remote sensing data[19–20]; and (7) hybrid approaches that may utilize two or more of the previously mentioned approaches [5, 15, 19]. Depending on the quality of the data support, each of these different exposure assessment approaches may have varying levels of exposure misclassification. Methods 1 to 3 are often employed when there are especially sparse monitoring networks over very large areas. The exposure misclassification in these kinds of exposure assessments can typically be classified as Berkson errors[21] because large groups of individuals, living near each other, may all receive the same or similar exposure values. This is unlikely to bias health effect estimates, but will reduce the power to detect an effect. By their nature, these estimation methods fail to capture the multiple scales of variability inherent in traffic-related air pollution[19, 22]. In contrast, the other exposure modeling methods – while potentially able to estimate exposures at more resolved scales – are prone to random error that may bias the health effects estimates downward.

As higher quality and quantity of point exposure observations become available, statistical techniques can be used to take advantage of assumed or observed dependencies in the data, and these methods may include auxiliary information. Techniques including both Bayesian[9–12, 23] and non-Bayesian [13–16, 22] multivariate statistics aim to estimate ambient air pollution exposure by deriving a set of relationships between auxiliary information and the pollutant of interest. In many cases, information about proximate land use, traffic and other contributing factors appear to be significant predictors of observed pollution levels. Models in this class are typically referred to as land use regression (LUR)[17]. These may[12, 15–16, 22, 24–25] or may not[10, 13–14] account for spatiotemporal dependencies in the data. Baxter and colleagues[26] reaffirmed that creating refined air pollution exposure models using auxiliary data increases the power to detect health effects, when the auxiliary data is strongly correlated with the exposure of interest.

Hybrid approaches attempt to combine the best qualities of the above mentioned methods. Most are implemented by using multivariate regression[5] or some form of generalized additive model[15, 19]. These hybrid approaches are particularly useful when the exposure may vary at multiple scales and no single method is capable of describing the expected variability. Building on previous air pollution exposure modeling efforts in California[27] we develop spatiotemporal models at the national scale to predict ambient $PM_{2.5}$ for the purpose of estimating exposure at residential locations for health effects analyses. The model uses a hybrid approach that combines land use regression (LUR) and Bayesian maximum entropy (BME) interpolation. As part of the hybridization approach – using the LUR – we integrated remote sensing estimates that used state-of-the-art atmospheric chemistry and transport models. The supplementary goal of this integration is to capture variability within the remote sensing grid cell using land use information as the remote sensing estimates are at a coarse resolution (~8.9 km). By introducing spatiotemporal interpolation of the residuals we aimed to predict on finer temporal scales than are typically available through LUR or remote sensing methods alone.

MATERIALS AND METHODS

Monthly $PM_{2.5}$ data preparation for the United States

The raw $PM_{2.5}$ data were acquired from the US Environmental Protection Agency (EPA)'s Air Quality System (AQS) and included only measurements from Federal Reference Method (FRM) monitors over the period January 1999 through December 2008. After processing the monitoring data, there were a total of 1,464 $PM_{2.5}$ monitoring locations across the contiguous United States. As these were government monitors, there tended to be slight under-representation of near traffic environments, but there was good representation of most

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

other environments where study subjects likely reside for our secondary health effects analysis (see online supplemental material for details [Section 1]).

The dataset was partitioned into a training set for model-fitting purposes and a cross-validation dataset using a uniform random variable procedure in Stata (Stata Corp, College Station, TX) (see online supplemental material [Section 1] for details on data collection and preparation for modeling and cross-validation). A purely random method, i.e., non-stratified or unconditional, was used to ensure that cross-validated performance measures were representative over the entire study area. The training dataset included 1,317 monitors with 93641 monthly observations; the cross-validation dataset included about 10% of the data, or 147 monitors with 10,531 monthly observations.

Land use, traffic and remotely sensed data

Prior to the beginning of the analysis, we decided that for any given class of LUR input data (i.e. land use, road network, traffic, etc), that all spatial data was completely represented over the contiguous United States and came from a sole source. For example, land use data sourced from different municipalities would not be combined to create a larger land use dataset. This was meant to ensure consistency in the classification and regression parameter estimates. While these criteria limited the variety of potential covariates, it ensured our ability to provide predictions at any location over the contiguous United States.

Generation of land use data—Land use data were based on United States Geological Survey (USGS) National Land Cover Database (NLCD) for 2001. The original 20 classes were aggregated into the following seven categories: agricultural, barren, all developed land, high-density development, green space, water, and wetland. Measures of land use proximate to the monitor were derived with Euclidean buffers of sizes: 100, 200, 300, 400, 500, and 1000 meters.

$PM_{2.5}$ estimates derived from remotely sensed data—Ground-level concentrations of $PM_{2.5}$ were estimated using satellite atmospheric composition data combined with local, coincident scaling factors from a chemical transport model[28]. Specifically, $PM_{2.5}$ estimates were derived from aerosol optical depth (AOD) data from the Terra satellite, in combination with output from GEOS-Chem simulations to estimate the relationship between aerosol optical depth over the atmospheric column and ground-level $PM_{2.5}$[20]. $PM_{2.5}$ was estimated at a 0.1 × 0.1-degree resolution. These were later projected into a Euclidean space using the Albers projection. The resulting grid resolution was approximately 8.9km × 8.9km. These estimates of $PM_{2.5}$ represent an average over the period 2001–2006; this ensured sufficient observations for stable estimation. Remote sensing estimates were integrated into the LUR method as a potential covariate.

Generation of traffic counts for roadways

Although actual traffic on a particular road segment is determined by a wide variety of factors, many of which cannot be captured by available geographic information systems (GIS) layers, we have developed weights that can be applied to road segments near air monitoring locations. To develop these weights we used more than 1.2 million spot traffic counts obtained from the Traffic Metrix dataset, purchased from MPSI (Tulsa, Oklahoma, US) limited to the years 1991–2009. Weights were assigned as the median count within a broad category of road type (i.e., expressway, major road, local road, etc.) stratified by the following classifications: urban, urban county, urban county size of population, rural county and rural county size of population. This enabled us to calculate traffic-weighted road density for those monitoring locations within the same set of Euclidean buffers used for the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

estimation of proximate land use (See online supplemental material [Section 2] for more detail on the derivation of the traffic weights).

Traffic weights showed good large-area correspondence with traffic counts. Within the larger buffer sizes, i.e., 500 m and 1000 m, the Spearman correlation between the estimated traffic weights and the Metropolitan Planning Organization (MPO) traffic data were 0.87 and 0.91, respectively[29].

Modeling approach overview

We used a multi-stage modeling strategy that included LUR and BME interpolation. In the first stage, we constructed the LUR model attributing measures of surrounding traffic, land use, and other auxiliary data to predict the $PM_{2.5}$ based on their empirical relationships. We expected that the LUR method alone would be insufficient to describe the variability in the space-time domain, as we were modeling spatiotemporal data and the LUR model only included spatial data. To capture the residual spatiotemporal variation, we then employed the BME methodology in the second stage. The BME theory, methodological characteristics, and computational techniques are discussed in detail elsewhere[30–33]. The BME technique interpolated the LUR spatiotemporal residuals using the expected distribution of the predicted values as prior information to inform the prediction. The 'residual surface' describes the spatiotemporal variability of the pollution estimates that could not be described by the LUR model. The final predicted estimate of $PM_{2.5}$ has the following form (Eq. 1):

$$\widehat{G}(s,t)=\Psi\left(\mathbf{L}(s)\right)+B\left(\overline{O}(s,t)-\Psi\left(\mathbf{L}(s)\right)\right) \quad \text{Eq. 1}$$

where the estimator of $PM_{2.5}$, $\hat{G}$, is indexed at location *s*, in Euclidean space, at time *t* by ordinal month; the basis function $\boldsymbol{\Psi}$ describes the land use regression function containing the subset of land use measures $\boldsymbol{L}$ used to estimate $PM_{2.5}$ at location *s*. The BME interpolator error model *B*, at location *s* and time *t*, is estimated as a function of the set of observations $\bar{O}$, near *s* at time *t*, that are be used to estimate at location *s* as specified by the BME covariance estimator.

Machine learning for land use regression—Under the framework of the LUR methodology, we employed a machine learning approach using the Deletion/Substitution/Addition algorithm (DSA)[34] when selecting the LUR models. DSA uses an aggressive covariate search algorithm to fit a generalized linear model, with polynomial basis functions, predicated entirely on the power of cross-validation (CV) to select the best predictive model. The CV method used for this research has been shown to have optimal properties when applied to the selection of prediction models[35]. This method attempts to minimize the CV risk, where the CV risk is the v-fold CV mean squared error. Details on the implementation of this method are contained in the Online Supplement [Section 3]. Two different model configurations were run using this method. The first configuration allowed the DSA to choose amongst all of the auxiliary data and included traffic, land use and remote sensing; the second did not include remote sensing.

Bayesian maximum entropy methods—This exposure modeling method relies on BME and its numerical implementation, *BMElib*[30, 36–37], to describe the residual spatiotemporal variability. BME is a non-linear estimator and integrates several important components of information: (*i*) composite space-time metrics (space/time analysis rather than purely spatial or purely temporal); (*ii*) data fluctuation (noise); (*iii*) data uncertainty (i.e., inaccurately modeled data, extrapolation, stochastic empirical laws, missing records, disaggregation or downscaling, and measurement errors); and (*iv*) secondary information

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

correlated with a primary variable in a mathematically rigorous and unified manner. It derives spatial regression techniques involving ordinary, simple, and intrinsic kriging as its limiting cases under restrictive assumptions about spatial or temporal correlation structures, types of data, and mean trend models considered. In addition, it deals with rationally non-stationary mapping situations and incorporates higher order statistical moments, (i.e., skewness) modeling even the coincidence of extreme events. Additional description of the methods is outlined in the online supplemental material [Section 4]. Similar to kriging, the covariance is estimated from the observations. Details of the general form of the model in this research are provided below (Eq. 2).

$$cov(r=|s-s'|,\tau=|t-t'|)=c_1 e^{\frac{-3r}{a_{r1}}} e^{\frac{-3\tau}{a_{t1}}}+c_2 e^{\frac{-3r^{\gamma}}{a_{r2}}} e^{\frac{-3\tau}{a_{t2}}}+c_3 e^{\frac{-3r^2}{a_{r3}{}^2}} \cos\left(\frac{2\pi\tau}{a_{t3}}\right)+c_4 e^{\frac{-3r}{a_{r4}}} \frac{\sin\left(\frac{2\pi\tau}{a_{t4}}\right)}{\frac{2\pi\tau}{a_{t4}}} \quad \text{(Eq. 2)}$$

where, s=location of observation of interest, s′=a proximate location to s, t=time point of interest, t′=proximate time point, c_i and $a_{ri/ti}$ are coefficients to be estimates from the data and γ is a scaling factor which differs between the two model configurations.

RESULTS AND DISCUSSION

Land use regression of $PM_{2.5}$

The distribution of the $PM_{2.5}$ measurements for the USA did not appear to be very skewed, so untransformed observations were modeled. The cross-validation risk plots (Figure 1) show the expected prediction error for each of the best models chosen by DSA as a function of model size. This plot illustrates that as the two models grow in size by the number of covariates, a point exists where an increase in model size does not increase predictive capacity. We chose models that maximize prediction and minimize complexity even if there appeared to be a small increase in predictive capacity as reported by the estimate of the cross-validated risk. In the model including remote sensing, after inclusion of three variables (i.e., remote sensing squared, remote sensing cubed and developed land at 200 m) there was minimal gain by including more (see Figure 1 and Table 1 for model results). For the model without remote sensing, there is little gain after the inclusion of two variables (i.e., traffic at 1000 m and green space at 100 m raised to the third power). The average cross-validated risks of the chosen LUR models with and without remote sensing are 18.25 and 22.44, respectively.

The models chosen by the DSA conform to the assumption of homoscedastic errors — and the errors appear to be close to normally distributed (results not shown). Given these results, the standard errors reported in Table 1 are likely to be plausible.

Bayesian maximum entropy estimation of spatiotemporal residuals

Table 2 reports the estimated parameters of the covariance functions for the residuals of the two LUR models. The parameter c_x represents a covariance-scaling coefficient; a_{rx} and a_{tx} are the range coefficients representing the scale of autocorrelation in space and time, respectively. We identified a phase shift in the temporal periodicity for observations on east and west coasts. It was determined that this was inconsequential to the estimation process as the largest average covariance weighted distance of the range parameters was less than 2000km. This indicated that there would be very little influence from observations on opposite coasts.

Given the smallest range in the covariance estimators was 70 km (Table 2: model without remote sensing), we decided that BME models would be predicted at the same grid

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

resolution and receptor locations as the remote sensing data (i.e., approximately 8.9 km grids). This was done, in part, for spatial consistency in the input and output datasets. The large range in the covariance estimator was also indicative of large-area structure that would not benefit from a highly resolved prediction grid.

Results from the BME analysis showed that the model residuals were fit well by the BME approach (see online supplemental material [Section 5]). Online materials show plots of the DSA model residuals versus the BME estimated residuals at the 8.9km resolution grid based on 93,486 observations.

Combined LUR-BME predictions, fit and cross-validation—The LUR-BME estimates are the sum of the DSA derived LUR model estimates and the BME estimates of the residuals. This section describes results of both how well the LUR-BME approach fit the observations used during the model-training phase and how well it predicted on the ten percent reserved cross-validation dataset. Mean squared error (MSE) in addition to R^2 and normalized R^2 (denoted as ${}_{\text{norm}}R^2$ in Table 3) are reported as measures of model fit and predictive capacity. As the LUR model is cross-sectional in nature, while the observed data is longitudinal, the R^2 value is unable to quantify how well the LUR model describes the chronic exposure trend in the data. Therefore, we generated a by-site mean model which perfectly predicts the monitoring site mean — the best possible result given the limitations of the LUR framework as applied to this research. The R^2 of the by-site mean model is used as the standard to compare the R^2. Normalized R^2 values are R^2 values for the LUR models divided by the R^2 of the by-site mean model (Eq3):

$$_{\text{norm}}R^2 = \frac{R^2}{\text{By-site mean model } R^2} \quad \text{Eq. 3}$$

The fit of the combined LUR-BME model performed very well. There was a substantial reduction in residual error indicating the residual error from the DSA models had spatiotemporal structure.

A large reduction in MSE (seen in Table 3) is observed on the data fit for the LUR-BME model. See Figure b in Section 5 of the online supplemental for plots of the model fits. Unlike the noticeable difference in MSE between the two configurations for the LUR models where the difference in MSE is approximately four, there is not much difference in the MSEs of the LUR-BME models – both are approximately 0.4. R^2 values for LUR models are 0.21 and 0.03 with and without remote sensing, respectively, but appear to perform better as a chronic exposure model as measured by the normalized R^2 of 0.51 and 0.08, respectively. In contrast, the LUR-BME model fit appears to be excellent with R^2 of 0.98 for both models.

Cross-validation analysis—An examination of the reserved 10% cross-validation dataset showed strong agreement between observed data and LUR-BME predictions with no indications of bias or exceptional outliers. The randomly selected cross-validation dataset did not include any observations larger than 35.7 $\mu g/m^3$, whereas the training data had 135 observations larger than 35.7 $\mu g/m^3$ with a maximum observed concentration of 74.7 $\mu g/m^3$.

Figure 2 compares cross-validation predictions of both LUR and LUR-BME models against the observed data. Plots indicate that the combined LUR-BME model is better at predicting observations.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In Table 3, the variance of $PM_{2.5}$ is provided as a reference value to compare against model MSEs. In the training dataset, modest reductions in MSE are observed with the LUR models for both model configurations, with the larger reduction in MSE observed in the model using remote sensing data. Unsurprisingly, the MSE becomes very small – near zero – after incorporating the spatiotemporal BME error model. In the training dataset, there is very little difference in MSE between the two configurations once the BME model is integrated into the estimates.

In the cross-validation dataset, the MSEs for the LUR models appear to be noticeably smaller than in the training dataset; however, the cross-validation dataset did not include observations larger than 35.7 $\mu g/m^3$. If these extreme values are excluded (for comparability between training and CV datasets), the MSEs for the LUR models appear to be more similar between training and cross-validation datasets. After integrating the spatiotemporal BME error model, a substantial reduction in MSE is observed with both model configurations. There appears to be little difference in predictive capacity between the two model configurations once the BME estimates are integrated; the CV MSE for both LUR-BME models is 4.63, which translates to an R^2 of 0.79.

Figure 3 shows side-by-side maps of the LUR-BME models averaged over the entire study period. The top map shows the model configuration using the remote sensing variable; the bottom map shows the model configuration without remote sensing. Additional maps of the prediction surfaces (including temporal estimates) can be found in the online supplemental material [Section 6].

Implications

There is only one other recent paper that has attempted to develop a national-scale spatiotemporal model to estimate ambient air pollution concentrations. In this research – conducted in Canada by Hystad and colleagues[38] – their aim was to create predictive models for nitrogen dioxide (NO_2), $PM_{2.5}$ and ozone (O_3). There have also been several other papers which have predicted a single annual average surface at the national scale, one in the United States for NO_2[39] and another in Canada for $PM_{2.5}$, NO_2 and several volatile organic compounds[40].

In creating a spatiotemporal predictor, Hystad et al.[38] took two approaches. The first used a scaling calibration method to offset a single-year remote sensing estimate based on the ratio of co-located current and historical monitors. The second used a regression calibration technique to adjust the remote sensing by year and by census population for a respective year. No other auxiliary data is considered; this significantly reduced their capacity to predict smaller scale variability. Additionally, monitoring locations are sparse in Canada. They used 177 total suspended particulate locations that were calibrated from 25 $PM_{2.5}$ monitors. For NO_2, there were 120 monitoring locations, and 187 locations were available for O_3. This small number of monitors raises the possibility of insufficient data support to inform predictions over large areas.

While these models represent substantial contributions to the literature, our models represent an advance in several important ways. First, our model is informed by 1464 monitoring locations that were used for model calibration and cross-validation. Second, none of the existing national models used actual traffic data, but relied on basic road networks or deterministic distance estimates to inform predictions for fine-scale variation. In our model we derived national traffic coverage from more than 1.2 million actual traffic counts, and these data significantly improved the fit of the model and the capacity of the model to predict fine-scale variation. Third, we introduced a machine-learning method that prevents over-fitting and derives the best prediction using an aggressive search algorithm. We

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

recently implement DSA at the state level in California to predict fine-scale variability in NO_2 and $PM_{2.5}$[27]; the research presented here has generalized its application to the national scale. Beyond the California model, our new modeling framework also allows for a flexible incorporation of the space-time dimensions so predictions can easily be made by the month in a theoretically sound manner. In sum, our model represents an advance because of the high level of data support, the innovative means of model selection, and the theoretically informed framework for integrating the space-time dimensions of the prediction.

When compared to the 2-stage approach implemented by Liu and others[19] across the state of Massachusetts our results of a CV-R^2 = 0.79 are in part similar to theirs, where they reported CV-R^2 = 0.78 for their model using remotely sensed data. However, in their model that did not use remote sensing, CV-R^2 = 0.46. As their research was estimating daily concentrations, differences in the study design make the results not entirely comparable. There also exists the possibility that spatiotemporal dependency across the nation may appear more stable and predictable than at the state scale. However, the combined results suggest that the framework of combining land use, remote sensing and spatial models in a 2-stage approach is effective for predicting observed $PM_{2.5}$ across multiple spatial and temporal scales.

At the national scale, a LUR type model alone is insufficient to describe the complex spatiotemporal variability in ambient $PM_{2.5}$. Here, only a small explanation of the spatial variability resulted from the first stage of the modeling. In the model with remote sensing, developed space was included. The model without remote sensing included a traffic and green space variable. Even though DSA selected variables that are ‘best’ predictors with no a priori specification on the direction of association, their estimated coefficients are intuitively congruent with our expectation of how they behave in the environment. Developed space and traffic act as sources of pollution (positive coefficient), while green space possesses below average pollution (negative coefficient). The apparent predictive capacity of the model configuration without remote sensing is small – the CV-R^2 is 0.05 – these results are in agreement with our understanding of primary on-road emissions where they contribute to only 3% of the overall ambient variability in $PM_{2.5}$[41]. Notwithstanding their small influence on the observed variability of $PM_{2.5}$, primary emissions such as fine particle, ultrafine particles, specific metals, and elemental carbon may be responsible for a large portion of the observed health impacts[42]. This adds to the importance of ensuring that these components of the ambient air pollution mix are described in exposure assessment tools intended for health effects assessments. The maps in Figure 3 illustrate the difference in the fine scale variability of $PM_{2.5}$ when remote sensing is not part of the model. The LUR models without remote sensing data includes traffic estimates that allow for prediction of $PM_{2.5}$ at scales much finer than either the remote sensing alone or the BME interpolator; this may be describing the portion of the air pollution mixture of interest to health researchers. Additionally, the BME method implicitly accounts for the resulting effects of meteorology and secondary transport that were not possible in other ways for this analysis.

A main assertion of this research was that after accounting for local scale variation, the residual variation would operate over larger areas. The range of the estimated covariance for the BME models indicates this long-range effect. When compared against research that used the BME method to interpolate observed $PM_{2.5}$ in a dataset nearly identical to that used here, but did not use auxiliary data[43], spatial dependence on the small scale was 20 km. The models presented here show the residual spatial dependence to operate at between 70 km and 120 km at the small scales. This is consistent with the premise that LUR would capture variability on a smaller spatial scale and the remaining spatial variability would have larger scale structure.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Limitations in auxiliary data used for the LUR models are likely responsible for their restricted predictive capacity. These limitations are a function of the cross-sectional nature of the land use, remote sensing and available traffic data. Although all data used here represent the best available information, they provided only a snapshot of these environmental contributors. Additionally, during the derivation of the traffic data, it was identified that the traffic-weighted estimates were likely descriptive of larger spatial scale variability in traffic and not necessarily the expected local traffic conditions. Still, given these limitations, the results appear to suggest that our LUR modeling efforts were effective in predicting the intended scale of variability consistent with our understanding of exposures of interest for health effects analyses. Remote sensing supplied important predictive information as shown by its selection by DSA, but the spatial resolution is coarse. The BME interpolation predicted the residual spatiotemporal variability very well, but the interpolation is influenced by the topology and density of the network. Over-smoothing of the estimates may happen where the network is sparse; this is typically where few people reside.

Given the highly predictive results, we expect that our estimates will have utility for public health and environmental science researchers. To facilitate use, temporal exposures (monthly) will be made available at census tracts representing the exposure at the centroids for both models with and without remote sensing (see Online Supplemental Information [Section 8]).

Future work should be directed towards refining the data inputs for the land use regression model that describe both local traffic and temporal variability in contributing environmental and land use factors. Opportunities to improve the accuracy of the remote sensing estimates should be pursued. In addition, downscaling the remote sensing estimates merits further investigation. Finally, explicit comparisons of the hybrid approach to other large-area exposure models in health effects modeling may help to understand the marginal benefit of including auxiliary data at multiple scales over purely in-situ or remotely sensed models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to acknowledge funding support from Health Canada and the following grants from CDC (200-2010-37394) and NIH (R01 ES019573-01). We would also like to thank David Holstius for Python code that made visualization of the exposure maps possible.

References

1. Gilliland F, Avol E, Kinney P, Jerrett M, Dvonch T, Lurmann F, Buckley T, Breysse P, Keeler G, de Villiers T, McConnell R. Air pollution exposure assessment for epidemiologic studies of pregnant women and children: lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research. Environ Health Perspect. 2005; 113(10):1447–54. [PubMed: 16203261]
2. Committee on, H. Exposure Science in the 21st Century: A Vision and a Strategy. The National Academies Press; 2012. Environmental Exposure Science in the 21st, C.; Board on Environmental, S.; Toxicology; Division on, E.; Life, S.; National Research, C.
3. Pope CA 3rd, Burnett RT, Thurston GD, Thun MJ, Calle EE, Krewski D, Godleski JJ. Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. Circulation. 2004; 109(1):71–7. [PubMed: 14676145]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4. Gauderman WJ, Avol E, Gilliland F, Vora H, Thomas D, Berhane K, McConnell R, Kuenzli N, Lurmann F, Rappaport E, Margolis H, Bates D, Peters J. The effect of air pollution on lung development from 10 to 18 years of age. The New England journal of medicine. 2004; 351(11): 1057–67. [PubMed: 15356303]
5. Jerrett, M.; Burnett, RT.; Pope, A., III; Krewski, D.; Thurston, G.; Christakos, G.; Hughes, E.; Ross, Z.; Shi, Y.; Thun, M.; Beckerman, B.; Turner, MC.; Su, J.; Lee, SJ. Spatiotemporal Analysis of Air Pollution and Mortality in California Based on the American Cancer Society Cohort: Final Report; California Air Resources Board. 2011. p. 145http://www.arb.ca.gov/research/rsc/06-09-11/agenda4_contract06-332_draft_report_cynthia_0520_v2.pdf
6. Wu JM, Winer AJ, Delfino R. Exposure assessment of particulate matter air pollution before, during, and after the 2003 Southern California wildfires. Atmospheric Environment. 2006; 40(18): 3333–3348.
7. Kunzli N, Jerrett M, Mack WJ, Beckerman B, LaBree L, Gilliland F, Thomas D, Peters J, Hodis HN. Ambient air pollution and atherosclerosis in Los Angeles. Environ Health Perspect. 2005; 113(2):201–6. [PubMed: 15687058]
8. Jerrett M, Burnett RT, Ma R, Pope CA 3rd, Krewski D, Newbold KB, Thurston G, Shi Y, Finkelstein N, Calle EE, Thun MJ. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology. 2005; 16(6):727–36. [PubMed: 16222161]
9. Fasbender D, Brasseur O, Bogaert P. Bayesian data fusion for space–time prediction of air pollutants: The case of NO2 in Belgium. Atmospheric Environment. 2009; 43(30):4632–4645.
10. Liu Y, Guo H, Mao G, Yang P. A Bayesian hierarchical model for urban air quality prediction under uncertainty. Atmospheric Environment. 2008; 42(36):8464–8469.
11. McBride SJ, Williams RW, Creason J. Bayesian hierarchical modeling of personal exposure to particulate matter. Atmospheric Environment. 2007; 41(29):6143–6155.
12. Riccio A, Barone G, Chianese E, Giunta G. A hierarchical Bayesian approach to the spatio-temporal modeling of air quality data. Atmospheric Environment. 2006; 40(3):554–566.
13. Henderson SB, Beckerman B, Jerrett M, Brauer M. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. Environmental science & technology. 2007; 41(7):2422–8. [PubMed: 17438795]
14. Jerrett M, Arain MA, Kanaroglou P, Beckerman B, Crouse D, Gilbert NL, Brook JR, Finkelstein N, Finkelstein MM. Modeling the intraurban variability of ambient traffic pollution in Toronto, Canada. Journal of toxicology and environmental health Part A. 2007; 70(3–4):200–12. [PubMed: 17365582]
15. Lindstrom, J.; Szpiro, AA.; Sampson, PD.; Sheppard, L.; Oron, AP.; Richards, M.; Larson, T. A FLEXIBLE SPATIO-TEMPORAL MODEL FOR AIR POLLUTION: ALLOWING FOR SPATIO-TEMPORAL COVARIATES. Berkeley Electronics Press; 2011. (http://www.bepress.com/uwbiostat/paper370)
16. Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics. 2010; 21(6):606–631.
17. Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J, Giovis C. A review and evaluation of intraurban air pollution exposure models. Journal of exposure analysis and environmental epidemiology. 2005; 15(2):185–204. [PubMed: 15292906]
18. Keuken M, Zandveld P, van den Elshout S, Janssen NAH, Hoek G. Air quality and health impact of PM10 and EC in the city of Rotterdam, the Netherlands in 1985–2008. Atmospheric Environment. 2011; 45(30):5294–5301.
19. Liu Y, Paciorek CJ, Koutrakis P. Estimating regional spatial and temporal variability of PM(2.5) concentrations using satellite data, meteorology, and land use information. Environ Health Perspect. 2009; 117(6):886–92. [PubMed: 19590678]
20. van Donkelaar A, Martin RV, Brauer M, Kahn R, Levy R, Verduzco C, Villeneuve PJ. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ Health Perspect. 2010; 118(6):847–55. [PubMed: 20519161]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

21. Berkson J. Are There Two Regressions? Journal of the American Statistical Association. 1950; 45(250):164–180.
22. Paciorek CJ, Yanosky JD, Puett RC, Laden F, Suh HH. Practical Large-Scale Spatio-Temporal Modeling of Particulate Matter Concentrations. Ann Appl Stat. 2009; 3(1):370–397.
23. Alan, G.; Sudipto, B. Handbook of Spatial Statistics. CRC Press; 2010. Multivariate Spatial Process Models; p. 495-515.
24. Sampson PD, Szpiro AA, Sheppard L, Lindström J, Kaufman JD. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. Atmospheric Environment. 2011; 45(36):6593–6606.
25. Zwack LM, Paciorek CJ, Spengler JD, Levy JI. Modeling Spatial Patterns of Traffic-Related Air Pollutants in Complex Urban Terrain. Environ Health Perspect. 2011; 119(6)
26. Baxter LK, Wright RJ, Paciorek CJ, Laden F, Suh HH, Levy JI. Effects of exposure measurement error in the analysis of health effects from traffic-related air pollution (vol 20, pg 101, 2010). J Expo Sci Env Epid. 2010; 20(5):486–486.
27. Beckerman BS, Jerrett M, Martin RV, van Donkelaar A, Ross Z, Burnett RT. Application of the Deletion/Substitution/Addition Algorithm to Selecting Land Use Regression Models for Interpolating Air Pollution Measurements in California. Atmospheric Environment. 2012 Accepted 8-Apr-2013.
28. Bey I, Jacob DJ, Yantosca RM, Logan JA, Field BD, Fiore AM, Li QB, Liu HGY, Mickley LJ, Schultz MG. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. J Geophys Res-Atmos. 2001; 106(D19):23073–23095.
29. Jerrett M. Assessing Effects of Ambient Air Pollution on Mortality Using the Cancer Prevention Study II: Development of Advanced Geostatistical Exposure Models. Health Canada. Mar 17.2011 2011:104.
30. Christakos, G.; Bogaert, P.; Serre, ML. Temporal GIS. Springer-Verlag; New York, NY: 2002. (with CD-ROM) ed
31. Christakos, G.; Olea, ML.; Serre, ML.; Yu, HL.; Wang, L-L. Interdisciplinary Public Health Reasoning and Epidemic Modelling: The Case of Black Death. Springer-Verlag; New York, NY: 2005.
32. Christakos G, Serre ML. A spatiotemporal study of exposure-health effect associations. Journal of exposure analysis and environmental epidemiology. 2000; 10(2):168–187. [PubMed: 10791598]
33. Yu, HL.; Kolovos, A.; Christakos, G.; Chen, J-C.; Warmerdam, S.; Dev, B. SERRA Special Issue on Medical Geography as a Science of Interdisciplinary Knowledge Synthesis under Conditions of Uncertainty. 2006. Interactive Spatiotemporal Modelling of Health Systems: The SEKS-GUI Framework.
34. Sinisi S, van der Laan M. Deletion/Substitution/Addition Algorithm in Learning with Applications in Genomics. Statistical Applications in Genetics and Molecular Biology. 2004; 3(1) Article 18.
35. Davies, MM.; van der Laan, MJ. UC Berkeley Division of Biostatistics Working Paper Series. 2013. Optimal Spatial Prediction Using Ensemble Machine Learning. (Working Paper 305)
36. Christakos G. A Bayesian Maximum-Entropy View to the Spatial Estimation Problem. Math Geol. 1990; 22(7):763–777.
37. Christakos G, Serre ML. BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmospheric Environment. 2000; 34(20):3393–3406.
38. Hystad P, Demers PA, Johnson KC, Brook J, van Donkelaar A, Lamsal L, Martin R, Brauer M. Spatiotemporal air pollution exposure assessment for a Canadian population-based lung cancer case-control study. Environ Health-Glob. 2012; 11
39. Novotny EV, Bechle MJ, Millet DB, Marshall JD. National Satellite-Based Land-Use Regression: NO2 in the United States. Environmental science & technology. 2011; 45(10):4407–4414. [PubMed: 21520942]
40. Hystad P, Setton E, Cervantes A, Poplawski K, Deschenes S, Brauer M, van Donkelaar A, Lamsal L, Martin R, Jerrett M, Demers P. Creating National Air Pollution Models for Population Exposure Assessment in Canada. Environ Health Perspect. 2011; 119(8):1123–1129. [PubMed: 21454147]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

41. Traffic related air pollution: A critical review of the literature on emissions, exposure and health effects. Health Effects Institute; Boston, Massachusetts: 2010. http://pubs.healtheffects.org/getfile.php?u=553

42. Kelly FJ, Fuller GW, Walton HA, Fussell JC. Monitoring air pollution: Use of early warning systems for public health. Respirology. 2012; 17(1):7–19. [PubMed: 21942967]

43. Lee SJ, Serre M, van Donkelaar A, Martin RV, Burnett RT, Jerrett M. Comparison of Geostatistical Interpolation and Remote Sensing Techniques for Estimating Long-Term Exposure to Ambient PM2.5 Concentrations across the Continental United States. Environ Health Perspect. 2012

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

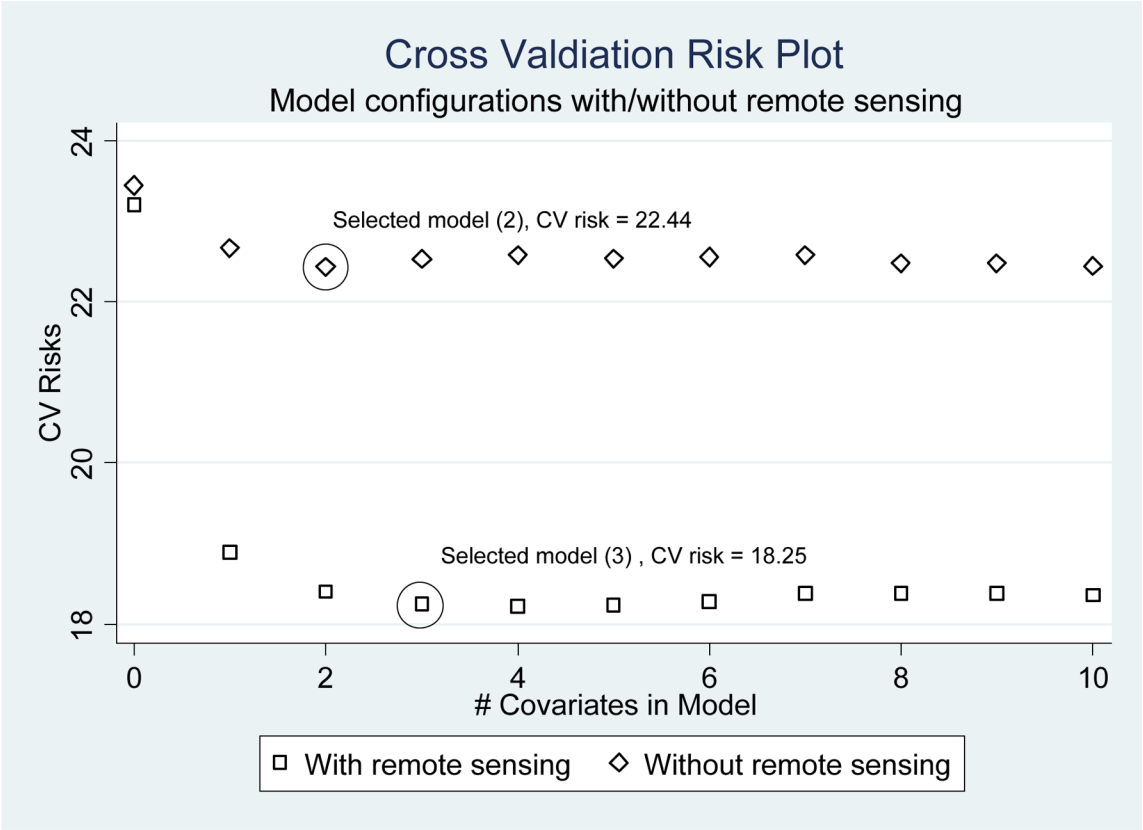

Figure 1.
Cross-validation (CV) risk plots as a function of model size

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

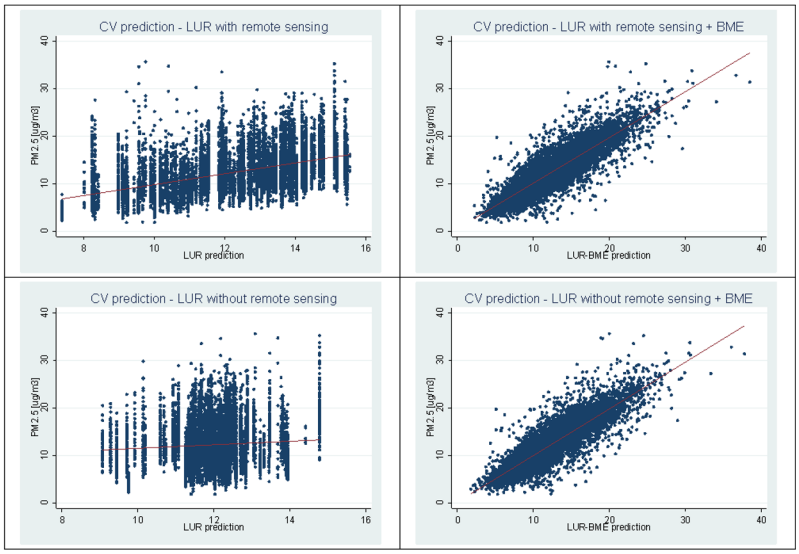

Figure 2.
Comparison of cross-validation prediction for LUR models and combined LUR-BME models

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

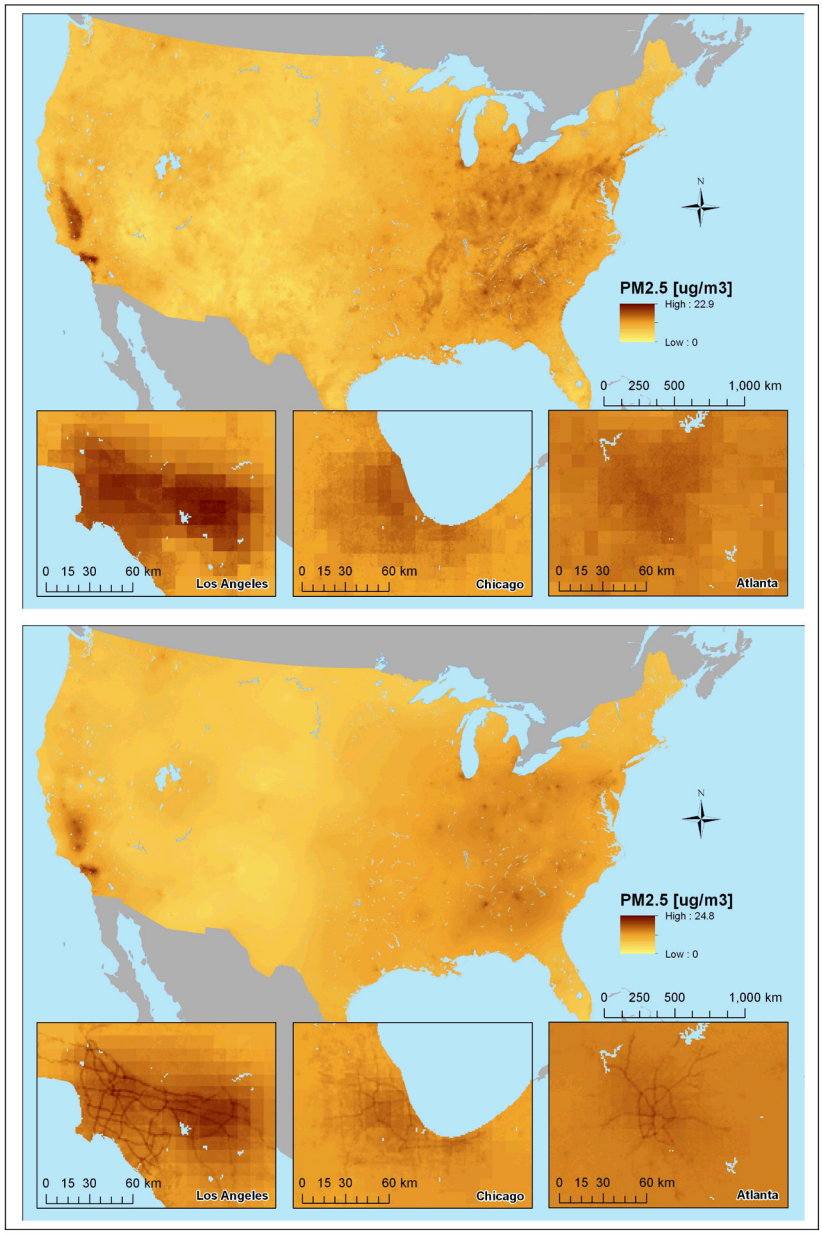

Figure 3.
Maps illustrating differences in spatial variability of LUR-BME exposure models averaged over the study time period: model using remote sensing (top), model without remote sensing (bottom).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Summary model estimates for the two LUR models

Configuration	Variables	Coef.	Robust Std. Err.	z	P>z*	[95% Conf. Interval]	
With Remote Sensing	(Remote Sensing $PM_{2.5}$)^2	0.0701	0.0030	23.37	0.0000	0.064	0.076
	(Remote Sensing $PM_{2.5}$)^3	−0.0024	1.31E-04	−18.69	0.0000	−0.003	−0.002
	Developed Land - Acres [200 m]	0.0404	0.0061	6.62	0.0000	0.028	0.052
	Intercept	5.9251	0.2222	26.66	0.0000	5.490	6.361
Without Remote Sensing	Traffic Weighted Rds (Veh.-km) [1000 m]	1.06E-04	1.37E-05	7.72	0.0000	7.9E-05	1.3E-04
	(Green space - Acres [100 m])^3	−0.0049	9.90E-04	−4.92	0.0000	−0.007	−0.003
	Intercept	11.3593	0.1465	77.56	0.0000	11.072	11.646

*p-values reported as 0.0000 are less than 0.00005

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 2

Parameter estimates for covariance functions of BME estimators.

	With Remote Sensing [γ = 1]			Without Remote Sensing [γ = 2]		
Parameter subscript [x]	**c_x**	**a_{rx} (km)**	**a_{tx} (months)**	**c_x**	**a_{rx} (km)**	**a_{tx} (months)**
1	8.2415	120	1	7.1354	70	1
2	4.9449	5500	5000	8.9192	1900	6000
3	2.5640	1500	12	3.1217	700	12
4	2.5640	1500	12	3.1217	700	12

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 3

Mean squared error (MSE) estimates for different modeling configurations with both the training and cross-validation datasets

	Model	Obs.	MSE	R^2	$_{norm}R^2$
All data	$PM_{2.5}$ [Variance]	104172	23.02	---	---
	$PM_{2.5}$ remote sensing [raw values]	104172	18.67	0.19	0.46
Training dataset	LUR with remote sensing	93641	18.26	0.21	0.51
	LUR-BME with remote sensing	93641	0.39	0.98	---
	LUR without remote sensing	93641	22.26	0.03	0.08
	LUR-BME without remote sensing	93641	0.43	0.98	---
All data ($PM_{2.5}$ < 35.7)	$PM_{2.5}$ [Variance]	104037	21.79	---	---
Training dataset with high $PM_{2.5}$ values removed ($PM_{2.5}$ < 35.7)	LUR with remote sensing	93506	16.97	0.22	0.54
	LUR-BME with remote sensing	93506	0.39	0.98	---
	LUR without remote sensing	93506	20.95	0.04	0.09
	LUR-BME without remote sensing	93506	0.42	0.98	---
Cross-validation dataset	LUR with remote sensing	10531	15.91	0.27	0.63
	LUR-BME with remote sensing	10531	4.63	0.79	---
	LUR without remote sensing	10531	20.77	0.05	0.11
	LUR-BME without remote sensing	10531	4.63	0.79	---