Supporting Information

An LUR/BME framework to estimate PM2.5 explained by on road mobile and stationary sources

For the Journal of

Environmental Science and Technology

Jeanette Reyes[†]*, Marc L. Serre*[†]***

[†]Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina

*Corresponding Author: Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599

Phone: (919) 966-7014 Email: marc_serre@unc.edu Fax: (919) 966-7911

Prepared September 10th 2013

27 pages: 4 Tables, 5 Figures

Table of Contents

Summary of PM2.5 Estimation Methods

CTMs are among the most sophisticated methods to estimate levels of PM2.5 at a fine temporal resolution. They aim at modeling the atmospheric and chemical processes involved in the production of an air pollutant from pollution sources. They represent the ultimate solution to estimating PM2.5 levels and have already proven to be highly valuable in improving our

understanding of air pollution. However, CTMs are complex to implement as they require detailed information on meteorology, emissions inventories and an advanced understanding of chemical processes. Collecting this information is a complex task and running a model over a large domain at a fine resolution can be computationally intensive, which makes CTMs a highly specialized tool and reduces its accessibility to potential users. CTMs can also have high errors because, even in its most refined implementations, they still rely on approximations and simplifications of the true underlying air pollution processes (e.g. completely mixed volumes, simplifying complex chemical reactions into a few key chemical mechanisms, etc.) and are therefore often used in a relative manner. For example, station implementation plans predict future air pollution levels by multiplying the ratios of the CTM's future to current predictions by an observationally based estimate on the current levels[1].

LUR models use a linear regression framework to predict PM2.5 levels based on spatial predictors that include land use, elevation and major emission sources. LURs can be tailored to use readily available data sources, implemented at low computational costs and the regression coefficients can be easily interpretable (e.g. providing an estimate of the increase of PM2.5 resulting from various major contributing sources). Hence LURs provide an attractive alternative to CTMs by being a simple tool that can be used by a wide audience to estimate PM2.5 levels and by quantifying how much total PM2.5 comes from various emission sources. Like CTMs, LURs may lead to large estimation errors. However, that weakness can be alleviated when used in a relative fashion paired with observationally driven estimation methods.

Satellite data provide observationally driven estimates of PM2.5. Satellite data might be considered the converse of CTMs in terms of providing sizable observationally driven approaches to estimating PM2.5. Its attractiveness comes in the large coverage it provides. One Satellite data is available, it can be easily used by a wide audience. However, due to the difficulty in estimating ground levels of air pollutions from remote sensing data, PM2.5 estimates calculated from satellite data are currently subject to high error compared to ground measurements of PM2.5 and they are most useful when estimating PM2.5 levels where monitoring stations are lacking[2].

A Review of LUR Domain Size

The ability of an LUR to explain the variability of an air pollutant is given by its r^2 statistics. The r^2 of previous LUR models vary widely depending on the type of air pollutant modeled and the spatial size of the study domain (Table S1).

A trend emerges when we plot the LUR r^2 as a function of the radius of the spatial domain (Figure S1). LUR models developed over small domains achieve the highest r^2. These models maximize predictability. But they are only valid over small areas, so they lack generalizability (i.e. they are not widely applicable outside the area for which they were developed). They provide site specific knowledge (i.e. knowledge that is highly predictive, but specific to a given area). The high predictability is achieved by sacrificing generalizability. LURs developed for large spatial domains on the other hand are forced to generate regression coefficients that are valid for a large region, which increases their physical meaning and plausibility. What these

models lose in predictability, they gain in generalizability. They provide general knowledge (i.e. knowledge that describes general characteristics of the air pollutant).

In Figure S1 we use the labels “site specific” and “general” to distinguish LUR models developed over small versus large domain sizes. This view allows us to craft a strategy on how the BME knowledge synthesis framework can be used to integrate the knowledge generated by LUR models. Site specific LUR models could be used to generate soft data that are part of the site specific knowledge processed at the integration stage of the BME procedure. On the other hand large area variability LUR models can be used as part of the general knowledge processed at the prior stage of the BME procedure. To our knowledge very few studies have used either strategy. In this work we use the latter[3,4].

Explanation and Choice of Hyperparameters

Many have used differed scales by creating multiple sized buffer zones around locations of interest as part of their variables in their LUR models[5–7]. In this work, we call these buffer zones hyperparameters. To best gauge which input hyperparameters would be a good initial estimate in the multivariate regression, univariate regression hyperparameters from $0.1\ km$ to $1000\ km$ were exhaustively tested using a few criteria. If there was an absolute r^2 maximum at less than $950\ km$ it was chosen as the best univariate hyperparameter. If the hyperparmeter plot resembles a plateau, the beginning of the plateau (working backwards, where r^2 reduced at least 5%) is chosen as the best univariate hyperparameter. Otherwise, a best univariate hyperparameter is not

chosen. The best hyperparameters becomes an input in the search routine to find the optimum LUR model.

Independent Variables: Source, Calculation and Coefficient Sign

Elevation

Elevation data came in the form of a raster file of North America. The raster was converted to points. The final elevation file contained over 1.1 million finely resoluted elevation data across the United States.

The elevation variable $V_{Elev,\boldsymbol{p}}$ is a purely spatial variable which is approximated for each location by using the closest known elevation location. Following past models[8], the β in the LUR model for elevation should be negative.

Total Traffic

Total traffic and average congestion came from traffic data obtained from the Bureau of Transportation Statistics' (BTS) National Transportation Atlas Database (NTAD)[9]. This contains GIS data about all major highways segments in the United States, including road length and Annual Average Daily Traffic (AADT) count. Because the AADT was only extensively available for 2009 at the time of this analysis, AADT is estimated for specific days by scaling 2009 data by the sum of national level "Highway Vehicles" emissions of CO and NO_X from 1998 to 2008 given by the EPA. An example calculation can be found below.

The following steps are used to calculate AADT across the country for June 28, 2003. 1) A linear interpolation is performed to calculate national traffic emissions on June 28, 2003. This involves a linear interpolation between national traffic emissions on December 31, 2002 and December 31, 2003. 2) AADTs are scaled to the national emissions for June 28, 2003. The scaling ratio is $\frac{national\ traffic\ emissions\ for\ June\ 28{,}2003}{national\ traffic\ emissions\ for\ December\ 31{,}2009}$. This ratio is multiplied by every available AADT.

The total traffic $V_{TT,\boldsymbol{p}}$ at $\boldsymbol{p} = (\boldsymbol{s}, t)$ has units of vehicular km driven per km^2, and is calculated by dividing the total number of vehicular miles driven (in year t) within a circular buffer as defined by the hyperperameter centered on $\boldsymbol{s}$, by the area (in km^2) of that buffer. The total traffic measures the areal density of traffic, which positively affects on road mobile emissions, and therefore its β should be positive.

Average Congestion

The average congestion $V_{AC,\boldsymbol{p}}$ has units of vehicular km driven per road km. It is calculated by dividing the number vehicular miles driven within a given buffer by the cumulative length of roads within that buffer. This variable measures the traffic throughput. For a given traffic density, higher throughput will lead to higher traffic congestion. A higher congestion will lead to inefficient traffic idling which will increase on road mobile emission, and therefore its β should also be positive.

Emissions Efficiency

Emission efficiency comes from population data. Population data were collected at the block group level in 2000 and county level from 1998-2009 by the US Census[10]. Data were also collected from the Mexican census for 2000 and 2005[11] and the Canadian census for 2001 and 2006[12]. Because block group populations are only available for 2000, block group population was estimated for specific days by interpolating block groups that were scaled using county level estimates. An example of a block group estimate is provided below. For Canada and Mexico, a linear interpolation is performed for the two years where data exist in order to calculate population estimates between those two given years. To estimate outside this window, the linear relationship is extrapolated for any particular day where population estimates needed to be calculated.

The following steps are used to calculate block group estimates for November 19, 2002. 1) The block group is estimated for July 1 for the year of interest by assuming the following ratio: $\frac{block\ group\ population\ for\ July\ 1{,}2002}{county\ population\ estimates\ for\ July\ 1{,}2002} = \frac{block\ group\ population\ for\ July\ 1{,}2000}{county\ population\ for\ July\ 1{,}2000}$. Note that the ratio was compared to July 1, 2000. This is due to having exact estimates for this year because it is a census year. The county in question is the county in which the block group resides. The block group population is known for July 1, 2000, the county population is known for July 1, 2000, and the county population estimates are known for July 1, 2002. These values can be plugged into the equation to solve for the block group population for July 1, 2002. This equation is performed for every block group for July 1 from 1998-2010. 2) A linear interpolation is done for November 19, 2002. A linear interpolation is done for each block group between July 1, 2002 and July 1, 2003 for every day of interest.

When calculating population density, a buffer is extended from each PM2.5 space/time location. If the centriod of the block group is included within the buffer, the total population of the block group is included in the calculation for population density. This is explained visually in Figure S2. This centriod method shows good agreement with the exact area of the buffer. Thus it can be said that the centriod method is an appropriate approximation for calculating population density. Figure S3 shows how these methods compare with different buffer sizes. This method provides computational efficiency.

Emission efficiency $V_{EE,\boldsymbol{p}}$ is measured from the population density (people per km^2) within a given buffer. We include this variable due to the discrepancy between the data available to estimate TT and the estimation needed of PM2.5. Data used in TT are traffic counts – not emissions. This assumes that every mile driven produces the same amount of emissions regardless of vehicle type. Emission efficiency corrects for this assumption by hypothesizing that areas with high population density tend to have vehicles better suited for urbanized environments, which (in general) are more fuel efficient (i.e. smaller cars, hybrid or electric vehicles, etc). Thus, for a given value for TT (traffic), PM2.5 should decrease with increasing EE (emission efficiency). As a result, plausible β values for emission efficiency should be negative when the regression model contains both the TT and EE variables, in support of our hypothesis that the TT variable over estimates emission from traffic in urbanized environment where cars are presumably more fuel efficient.

Stationary Sources

Point source emissions data were obtained from the EPA's National Emissions Inventory (NEI) website[13]. Data were separated by year and pollutant (i.e. SO_2 stationary source emissions, NH3 stationary source emissions, PM2.5-primary stationary source emissions, and NOX stationary source emissions). The NEI is published every three years (i.e. 1999, 2002, 2005) with some summary files for years in-between (i.e. 1998, 2000, and 2001). For this study, total non-mobile stationary source emissions at the county-level were used. Between NEIs, the EPA also publishes national level annual emissions broken down by pollutant and tier one type. Because NEI data only exist for a few years, NEI emissions are scaled by these national level emissions to establish stationary source values for specific days. An example of this calculation is given below.

The following steps are used to calculate point source data for March 18, 2004. 1) The last available NEI data is chosen. In this case the last available NEI is December 31, 2002. 2) A linear interpolation is done to determine national emissions on March 18, 2004. This involves performing a linear interpolation between national emissions from December 31, 2003 and December 31, 2004. 3) Points are then scaled by national emissions to determine stationary source values for March 18, 2004. Thus the scaling ratio is $\frac{national\ emisssions\ for\ March\ 18,2004}{national\ emissions\ for\ December\ 31,2002}$. This ratio is multiplied to all December 31, 2002 stationary values.

The emissions were modeled as a single stack being emitted from the centriod of each county for the years data exist. National annual emissions in 2009 were not available and thus were estimated. Calculations for these estimates are given below.

It is assumed that the ratio of national emissions between years remains constant. That is, the following equation is assumed:

$\frac{national\ NH_3\ emissions\ for\ December\ 31{,}2008}{national\ NH_3 emissions\ for\ December\ 31{,}2007} = \frac{national\ NH_3\ emissions\ for\ December\ 31{,}2007}{national\ NH_3 emissions\ for\ December\ 31{,}2006}$. The national NH3 emissions for December 31, 2006 and December 31, 2007 are known. These values can be plugged in to solve for national NH3 emissions for December 31, 2008.

Like above, it is assumed that the ratio of national emissions between years remains constant. That is, the following equation is assumed: $\frac{national\ X\ emissions\ for\ December\ 31{,}2009}{national\ X\ emissions\ for\ December\ 31{,}2008} = \frac{national\ X\ emissions\ for\ December\ 31{,}2008}{national\ X emissions\ for\ December\ 31{,}2007}$, where X = NH3, NOX, PM2.5, and SO2. Emissions for all national stationary sources are known for December 31, 2007 and December 31, 2008. These values can be plugged in to solve for national stationary source emissions.

The effect that a stationary source emission has on the concentration at space/time location $\boldsymbol{p}$ is a function of the distance between the stationary source and $\boldsymbol{p}$. We assume that the effects of stationary source emissions decrease exponentially with distance and are additive. Thus, the predictor variable describing the effect of stationary source emissions of PM2.5 constituents i, $i = SO2, NH3, PM2.5, NOX$, is

$$V_{i,\boldsymbol{p}} = \sum_{n=1}^{N} em_i(\boldsymbol{e}_n, t)\exp(\frac{-3\|\boldsymbol{e}_n - \boldsymbol{p}\|}{d_{r_i}}) \quad \text{(Eq. S1)}$$

where $em_i(\boldsymbol{e}_n, t)$ is the emissions in tons/year of constituent i at stationary source emissions location $\boldsymbol{e}_n$ and time t and d_{r_i} is the exponential decay range in km (i.e. the hyperparameter) for each pollutant i.

LUR Offset and Covariance Models

Two different models were used for the global offset $o_z(\boldsymbol{p})$ defined in Equation (2): a constant offset and an LUR offset. The global offset was calculated as the average of all PM2.5 concentrations, which we found to be 11.72 $\mu g/m^3$. The LUR offset values ranged from about 5.7 $\mu g/m^3$ to about 18.4 $\mu g/m^3$.

For each space/time point $\boldsymbol{p}_i$ where an annual average PM2.5 concentration value z_i exists, we use Equation (2) to calculate the corresponding transformed value $x_i = z_i - o_Z(\boldsymbol{p}_i)$. We then model the uncertainty and variability of the transformed PM2.5 values using a homogeneous/stationary S/TRF $X(\boldsymbol{p})$ for which a realization corresponds to the observed values. The G-KB for the transformed S/TRF $X(\boldsymbol{p})$ consists of its expected value

$$m_X(\boldsymbol{p}) = E[X(\boldsymbol{p})], \quad \text{(Eq. S2)}$$

where $E[\cdot]$is the stochastic expectation operator, and its covariance function

$$c_X(\boldsymbol{p}, \boldsymbol{p}') = E[\big(X(\boldsymbol{p}) - m_X(\boldsymbol{p})\big)\Big(X(\boldsymbol{p}') - m_X(\boldsymbol{p}')\Big)]. \quad \text{(Eq. S3)}$$

The expected value for the homogeneous/stationary S/TRF $X(\boldsymbol{p})$ is constant and we set it to zero since the offset captures the general mean of the data. The covariance of $X(\boldsymbol{p})$ is modeled by first calculating experimental covariance values as a function of spatial and temporal lags, and fitting a permissible covariance model to these experimental values. Other models where attempted; however, the sum of exponential models had the best fit. This model captures variability at multiple spatial and temporal ranges in space/time. This model is highly interpretable and

consistent with previous studies[14,15]. The model is expressed by the following equation

$$C_X(r,\tau) = c_1 \exp\left(-\frac{3r}{a_{r_1}}\right)\exp\left(-\frac{3\tau}{a_{\tau_1}}\right) + c_2 \exp\left(-\frac{3r}{a_{r_2}}\right)\exp\left(-\frac{3\tau}{a_{\tau_2}}\right) + c_3 \exp\left(-\frac{3r}{a_{r_3}}\right)\exp\left(-\frac{3\tau}{a_{\tau_3}}\right)$$

(Eq. S4)

When using the LUR offset we obtain transformed data that result in the experimental covariance shown in figure S4. Using least square fitting we obtain the following parameters for its covariance model (Eq. S4): $c_1 = 1.18$, $a_{r_1} = 0.60\ degrees$, $a_{\tau_1} = 351\ days$, $c_2 = 1.99$, $a_{r_2} = 0.60\ degrees$, $a_{\tau_2} = 49951\ days$, $c_3 = 3.05$, $a_{r_3} = 9.60\ degrees$, $a_{\tau_3} = 49951\ days$. This covariance model is the sum of three space/time covariance structures. The first structure models variability occurring over short spatial and temporal ranges, while the third structure models variability at longer spatial and temporal ranges.

We followed the same procedure for the constant offset (figure not shown), which resulted in the following parameters $c_1 = 1.74$, $a_{r_1} = 0.5\ degrees$, $a_{\tau_1} = 251\ days$, $c_2 = 3.82$, $a_{r_2} = 0.5\ degrees$, $a_{\tau_2} = 43501\ days$, $c_3 = 6.02$, $a_{r_3} = 16.5\ degrees$, $a_{\tau_3} = 43501\ days$.

Fundamental Set of BME Equations

In this study the G-KB and S-KB can therefore be written as $G = \{m_X(\boldsymbol{p}), c_X(\boldsymbol{p}, \boldsymbol{p}')\}$ and $S = \{\boldsymbol{x}_h, f_s(\boldsymbol{x}_s)\}$, and in this case the BME equation is[1,16]

$$f_K(x_k) = A^{-1} \int d\mathbf{x}\ f_S(\mathbf{x_s}) f_G(\mathbf{x}) \qquad \text{(Eq. S5)}$$

where $\boldsymbol{x} = (x_k, \boldsymbol{x}_h, \boldsymbol{x}_s)$ is a realization of $\boldsymbol{X}$ at points $\boldsymbol{p} = (\boldsymbol{p}_k, \boldsymbol{p}_h, \boldsymbol{p}_s)$, $f_S(\boldsymbol{x})$ is a PDF characterizing the knowledge and uncertainty associated with the S-KB, $f_G(\boldsymbol{x}) = e^{\mu^T \boldsymbol{g}(\boldsymbol{x})}$ is the

Gaussian PDF for $\boldsymbol{X}$ with mean and covariance matrix obtained from the G-KB, A is a normalization constant, and f_K is the BME posterior PDF describing residual PM2.5 at the estimation point $\boldsymbol{p}_k$. The S-KB can either be considered hard or soft. Hard data are data measured without error while soft data are considered to have some uncertainty associated with its measurements. Soft data can be characterized by any PDF (Gaussian, interval, uniform, triangle, etc.). In this work, soft data come in the form of truncated Gaussian distributions.

Cross validation equations/Results

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(\mu_k - \hat{\mu}_k)^2 \quad \text{(Eq. S6)}$$

$$RMSE = \sqrt{MSE} \quad \text{(Eq. S7)}$$

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|\mu_k - \hat{\mu}_k| \quad \text{(Eq. S8)}$$

$$MR = \frac{1}{n}\sum_{k=1}^{n}\sigma_k \quad \text{(Eq. S9)}$$

Below are results of the 10-fold cross-validation for each fold.

	r^2
overall	0.78
fold 1	0.79
fold 2	0.83
fold 3	0.78
fold 4	0.75
fold 5	0.69
fold 6	0.77
fold 7	0.83
fold 8	0.76
fold 9	0.80
fold 10	0.78

Supporting Table 1

Supporting Information, Table S1: Literature review of past LUR models. BS is black smoke. *Domain size was not explicitly stated and was estimated.

source	spatial domain size (km^2)	pollu-tant	LUR variables	r^2
(Molter et al. 2010)[17]	2,102*	PM_{10}	light goods vehicles on Motorways (200 m), light goods vehicles on A roads (50 m), buses on Motorways (50 m), Y coordinate, PM10 emissions (800 m), roadside (50 m), buses on A roads (50 m), distance to nearst A road, dummy variable	0.97
(Eeftens et al. 2012)[18]	4,435*	$PM_{2.5}$	heavy traffic (1000 m), industry (5000 m), population (1000 m), x-coordinate	0.88
(Dijkema et al. 2011)[19]	6,000	NO_2	background concentration, traffic volume at nearest road, distance to nearest busy road, residential land use (5000 m)	0.87
(Melymuk et al. 2013)[20]	7,124*	PCB	Sum of PCBs in use/storage/building sealants, Metal industries, 1971 population density	0.85
(Beelen et al. 2007)[21]	41,562	NO_2	n/w/s/e indicator, population (5000 m), non-rural area, center of town, traffic intensity (100 m), truck traffic intensity on nearest motorway, distance to nearest motorway (< 100 m, 100-300 m, > 300 m)	0.84
(Mavko et al. 2008)[22]	100*	NO_2	traffic volume on major roads (300 m wedge), length of railroads (100 m), traffic volume on major roads (50 m wedge), parks and open space (50 m)	0.81
(Carr et al. 2002)[23]	600*	ben-zene	sum of daily traffic volume (50 m), sum of daily traffic volume (50-300m), sum of daily traffic volume if traffic jam > 5% (50m), sum of daily traffic volume if traffic jam > 5% (50-300m)	0.803
(Su et al. 2009)[24]	633	NO_2	x coordinate, traffic within 24 hours (650 m), expressway casement (400 m), open land use (1400 m), railway (1400 m), major road (50 m), slope (1400 m), population density (1350 m), distance to coast	0.794
(Ross et al. 2006)[7]	2,800*	NO_2	road length (40 m), traffic density (40 to 300 m), traffic density (300 to 1000 m), distance to coast	0.79
(Clougherty et al. 2008)[25]	170*	$PM_{2.5}$	central site concentration, roadway length (100 m), smoking or grilling, population density	0.76
(Hoek et al. 2011)[26]	324	soot	traffic intensity*inverse distance to roads, urban green area (3000 m), traffic (100 m)	0.76

(Ryan et al. 2007)[27]	2,400*	EC	elevation, average daily traffic count on major roads (400 m), length of bus routes (100 m)	0.75
(Mercer et al. 2011)[28]	6,400*	NO_X	distance to nearest A1 road, distance to coast, distance to commerical source, length of A2 and A3 roads (50 m), length of A1 roads (50 m)	0.74
(Brauer et al. 2003)[29]	41,543	$PM_{2.5}$	number of high traffic roads (250 m), address density (300 m), region west, region middle	0.73
(Briggs et al. 1997)[30]	48	NO_2	traffic volume (60 m), traffic volume (120 m), land cover (60 m), altitude	0.72
(Smith et al. 2011)[31]	1,500*	ben-zene	distance to nearest road with 10,000 to 20,000 vehicles per day, distance to nearest road with 70,000 to 80,000 vehicles per day, distance to nearest road with 100,000 to 120,000 vehicles per day, traffic intensity (1000 m)	0.72
(Dadvand et al. 2011)[32]	2,055	BS	distance to motorway, distance to A/B roads, length of local roads/streets (250 m), easting, northing	0.7
(Arain et al. 2007)[33]	22,500	NO_2	density of roads, land use, population, physical geography, meteorology	0.695
(Gehring et al. 2011)[34]	170	NO_2	traffic volume on the nearest busy road (50 m), distance to nearest main road, percentage of greens space (250 m), percentage of water (100 m)	0.692
(Jerrett et al. 2007)[35]	633	NO_2	ln(NO2), expressway (200 m), major roads (50 m), industrial land use (750 m), density of dwellings (2000 m), x coordinate, downwind, density of 24 hour traffic counts (500 m)	0.69
(Moore et al. 2007)[36]	98,500	$PM_{2.5}$	traffic density (300 m), industrial land area (5000 m), government land area (5000 m)	0.69
(Rosenlund et al. 2008)[6]	1,290	NO_2	location of site (inside limited traffic zone, inner ring, green strip, main ring road, outside main ring road), distance to busy road (less than 150 m, between 150-500 m, greater than 500 m), size of census block, inverse population density of the census block, altitude	0.686
(Ross et al. 2007)[37]	17,404.7*	$PM_{2.5}$	traffic (500 m), population (1000 m), industrial land use (300 m)	0.642
(Kanaroglou et al. 2005)[38]	633	NO_2	distance to nearest expressway, measure of expressway (200 m), measure of local roads within a donut-shaped area (inner radius 300 m-500 m), measure of major roads (500 m), measure of park, open, recreational, or water body land (400 m), density of dwellings (2000 m)	0.633

(Henderson et al. 2007)[39]	2,200	NO	length of highways (100 m), length of highways (1000 m), length of major roads (100 m), population density (2500 m), elevation, longitude, latitude	0.62
(Brauer et al. 2008)[40]	2,200	NO	number of major roads (100 m), number of major roads (1000 m), number of secondary roads (100 m), population density (2500 m), elevation	0.62
(Marshall et al. 2008)[41]	2,844	NO	road density (100-200), elevation, population density (2500 m), latitude, road density (750 m)	0.62
(Gan et al. 2011)[42]	2,200	NO	length of highways (100 m), length of highways (1000 m), the length of major roads (100 m), the population density (2,500 m)	0.62
(Gulliver et al. 2011)[43]	229,848*	SO_2	x coordinate, y coordinate, x^2 coordinate, y^2 coordinate, xy coordinate, minor road length (3000 m), low density urban (10000 m), high density urban (1000 m)	0.605
(Su et al. 2008)[44]	2,877*	NO	highways (100 m), highways (1000 m), major roads (200 m), commerical land use (750 m), elevation, x coordinate, y coordinate	0.6
(Gilbert et al. 2005)[5]	365*	NO_2	distance to nearest highway, traffic count on nearest highway, length of highways (100 m), length of major roads (100 m), length of minor roads (500 m), area of open space (100 m), population density (2000 m)	0.54
(Hart et al. 2009)[45]	7,663,942*	PM_{10}	elevation, percent low-intensity residential land use (1000 m), percent high-intensity residential land use (1000 m), percent ICT land use (1000 m), distance of A1 roads, distance to A2 roads, distance to A3 roads	0.49
(Beelen et al. 2009)[46]	2,854,116*	PM_{10}	hot year round and windy (1000 m), hot summers and cool winters and calm (1000 m), cold calm winter and warm windy summers (1000 m), altitude (1000 m), agriculture (5000 m), low density residential (5000 m), high density residential (5000 m), topex (5000 m)	0.41
(Pearce et al. 2009)[47]	33	$PM_{2.5}$	elevation, direct connectivity of surrounding intersections	0.34
(Adam-kiewicz et al. 2010)[48]	20*	NO_2	Session, distance to terminal, total traffic density (100 m), total length of class 1 roads (100 m), total length of class 2 roads (200 m), distance to nearest class 3 road	0.32

(Beckerman et al. 2013)[4]	7,663,942*	$PM_{2.5}$	remote sensing PM2.5[2], remote sensing PM2.5[3], developed land (200 m)	0.21

Supporting Table 2

Supporting Information, Table S2. Tons of emissions by pollutant for selected years for stationary and mobile sources. The NEI exists for the years 1999, 2002, and 2005 for major contributing sources. Estimation for non-NEI years was performed by calculating a linear interpolation for years in between NEIs and extrapolated for 2006 and 2007 by extending the trend between 2002 and 2005.

		1999	2002	2005
stationary-point	SO2	15,481,401.88	12,637,795.89	12,497,868.41
	NOX	8,634,591.82	7,081,275.68	6,145,388.64
	NH3	87,444.75	184,880.78	181,609.44
	PM2.5-pri	1,323,759.85	883,454.69	963,634.27
	total	25,527,198.30	20,787,407.04	19,788,500.76
	PM2.5-fil	509,454.42	413,437.22	306,340.47
stationary-area	SO2	1,288,281.05	1,371,710.06	1,374,311.84
	NOX	1,754,518.62	1,684,373.32	1,746,920.98
	NH3	4,499,827.90	3,651,324.97	3,601,368.44
	PM2.5-pri	5,518,185.26	4,253,538.09	4,043,500.49
	total	13,060,812.82	10,960,946.45	10,766,101.76
	PM2.5-fil	5,071,063.91	2,248,777.90	2,249,371.69
mobile (on road)	SO2	300,430.69	245,314.06	145,965.66
	NOX	8,371,337.43	7,922,527.06	6,491,820.47
	NH3	266,820.73	288,421.52	298,856.12
	PM2.5-pri	183,895.09	156,142.58	136,179.71
	total	9,122,483.94	8,612,405.22	7,072,821.96
	PM2.5-fil	NA	NA	NA
stationary		**1999-2007**		
	SO2	130,919,278.97		
	NOX	77,771,066.99		
	NH3	35,761,846.24		
	PM2.5-pri	48,993,549.60		
	total	293,445,741.80		
	PM2.5-fil	29,264,027.70		
mobile (on road)	SO2	1,821,317.80		
	NOX	65,046,831.33		
	NH3	2,604,765.10		
	PM2.5-pri	1,360,973.89		
	total	70,833,888.12		

	PM2.5-fil	NA

Supporting Table 3

Supporting Information, Table S3: Cross validation results of all hard data points from 1999. Units: [a] $(\mu g/m^3)^2$, [b] $\mu g/m^3$, [c] unitless.

Method	LUR only	(a) hard/ constant	(b) hard/LUR	(c) hard and soft/LUR	% change from (a) to (b)	% change from (b) to (c)
MSE[a]		6.927	6.020	5.618	-13.09	-6.69
RMSE[b]		2.632	2.454	2.370	-6.77	-3.40
MAE[b]		1.383	1.171	0.990	-15.37	-15.40
MR[b]		1.792	1.150	0.947	-35.82	-17.68
Pearson's Corr.[c]		0.820	0.844	0.852	2.97	0.89
Spearman's Corr.[c]		0.868	0.892	0.900	2.70	0.93

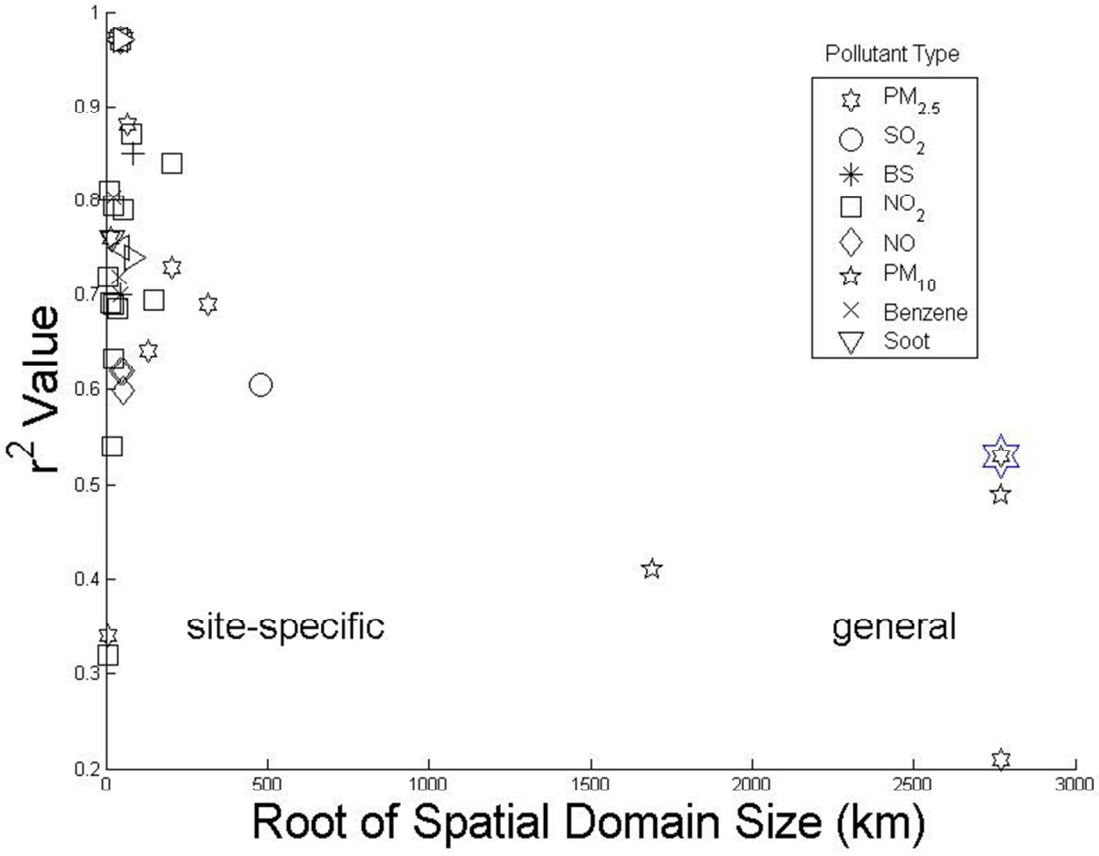
Supporting Table 4

Supporting Information, Table S4: Cross validation results of 20,000 randomly selected hard data points whose closest five neighbors includes at least two soft data points. For units, see above.

Method	LUR only	(a) hard/ constant	(b) hard/LUR	(c) hard and soft/LUR	% change from (a) to (b)	% change from (b) to (c)
MSE[a]		2.877	2.227	1.994	-22.57	-10.49
RMSE[b]		1.696	1.492	1.412	-12.01	-5.39
MAE[b]		1.119	0.921	0.819	-17.71	-11.08
MR[b]		2.024	1.349	1.217	-33.34	-9.82

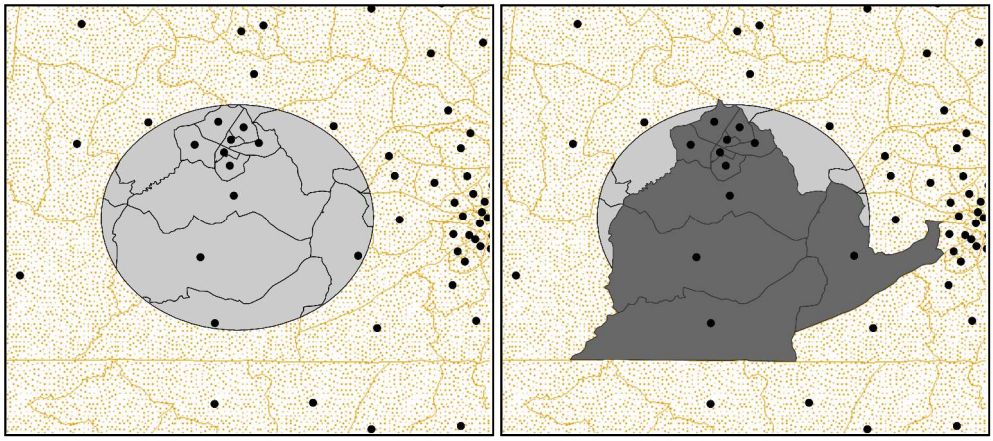
Pearson's Corr.[c]	0.864	0.894	0.906	3.49	1.30
Spearman's Corr.[c]	0.886	0.913	0.923	3.05	1.10

Supporting Figure 1



Supporting Information, Figure S1: Domain size versus r^2 for past LUR models distiguished by pollutant type. The largest shape is the LUR model found in this work. BS is Black Smoke.

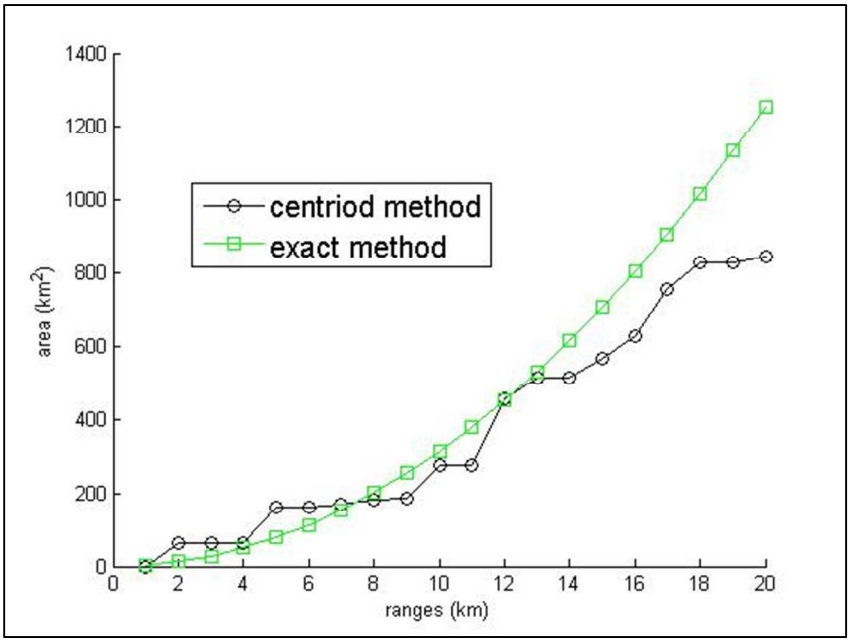
Supporting Figure 2



(a) (b)

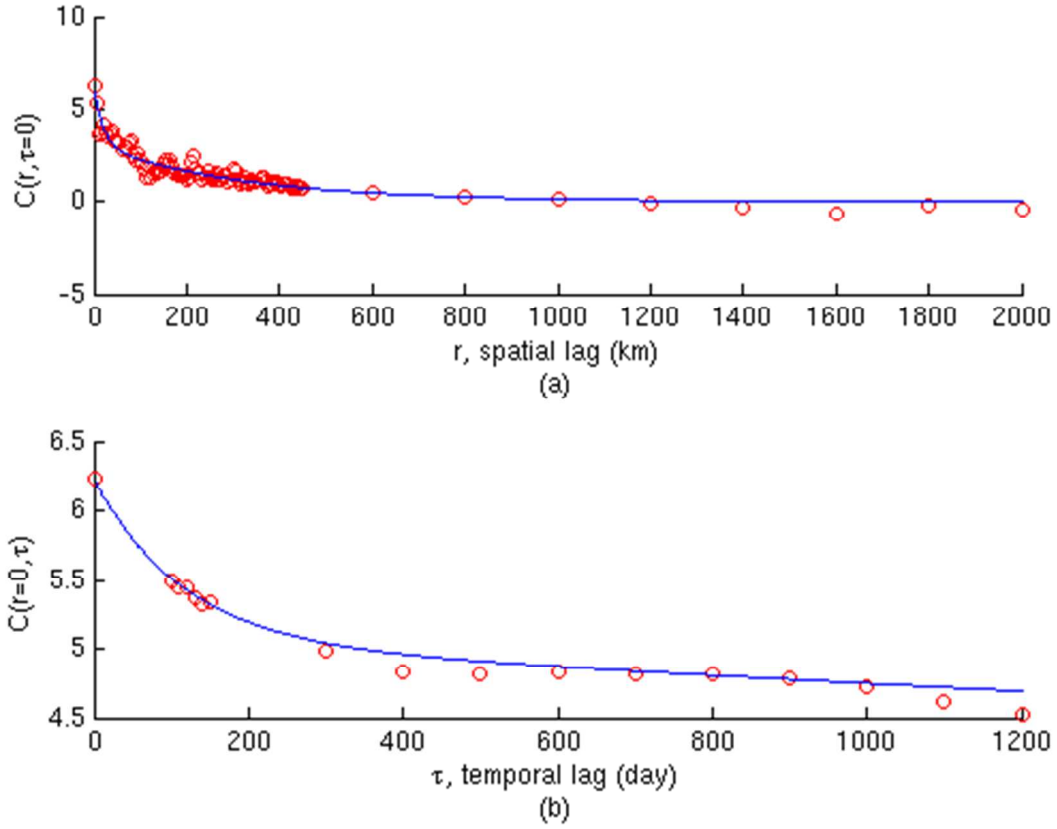
Supporting Information, Figure S2: The exact buffer size (a) when calculating population density compared with the approximate centriod inclusion method (b) for an arbitrary PM2.5 location. The black dots are the centriods of the block groups. The light gray area is a buffer with a 10 km radius. The dark gray area on the right is the total population that will be included in a given population density calculation.

Supporting Figure 3



Supporting Information, Figure S3: Comparing the areas of the buffer around a PM2.5 value to calculate population density by comparing the centriod method to the exact area for up to 20 km for the PM2.5 station shown in Figure S2. The centriod method includes the entire area of a block group if the centriod of that block group in within the buffer.

Supporting Figure 4



Supporting Information, Figure S4: Experimental (red circles) and modeled covariance (blue line) of the LUR offset removed PM2.5 concentrations shown (a) as a function of spatial lag *r* for a temporal lag $\tau = 0$, and (b) as a function of temporal lag τ for a spatial lag $r = 0$.

Supporting Figure 5

Supporting Information, Figure S5: Movies of BME predicted PM2.5 ($\mu g/m^3$) concentration estimation maps across the contiguous U.S. on the first day of every month from 1/1/1999-12/1/2002 obtained from the following methods (a) constant offset / hard data, (b) LUR offset /hard data, (c) LUR offset /hard and soft data.

http://www.unc.edu/depts/case/BMElab/studies/JR_yPM25_US/index.htm

References

(1) De Nazelle, A.; Arunachalam, S.; Serre, M. L. Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions□: An Application for Attainment Demonstration in North Carolina. *Environmental Science & Technology* **2010**, *44* (15), 5707–5713.

(2) Lee, S.; Serre, M. L.; Donkelaar, A. Van; Martin, R. V; Burnett, R. T.; Jerrett, M. Comparison of Geostatistical Interpolation and Remote Sensing Techniques for Estimating Long-Term Exposure to Ambient PM 2 . 5 Concentrations across the Continental United States. *Environmental Health Perspectives* **2012**, *120* (12), 1727–1732.

(3) Yu, H.; Wang, C.; Liu, M.; Kuo, Y. Estimation of Fine Particulate Matter in Taipei Using Landuse Regression and Bayesian Maximum Entropy Methods. *International Journal of Environmental Research and Public Health* **2011**, *8* (6), 2153–2169.

(4) Beckerman, B.; Jerrett, M.; Martin, R. V; Lee, S.; Donkelaar, A. Van; Ross, Z.; Su, J.; Burnett, R. A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM 2.5 in the Contiguous United States. *Environmental Science & Technology* **2013**, *47* (13), 7233–41.

(5) Gilbert, N. L.; Goldberg, M. S.; Beckerman, B.; Brook, J. R.; Jerrett, M. Assessing Spatial Variability of Ambient Nitrogen Dioxide in Montreal, Canada, with a Land-Use Regression Model. *Journal of Air & Waster Management* **2005**, *55* (8), 1059–1063.

(6) Rosenlund, M.; Forastiere, F.; Stafoggia, M.; Porta, D.; Perucci, M.; Ranzi, A.; Nussio, F.; Perucci, C. A. Comparison of regression models with land-use and emissions data to predict the spatial distribution of traffic-related air pollution in Rome. *Journal of Exposure Science and Environmental Epidemiology* **2008**, *18* (2), 192–199.

(7) Ross, Z. E. V; English, P. B.; Scalf, R.; Gunier, R.; Smorodinsky, S.; Wall, S.; Jerrett, M. Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *Journal of Exposure Science and Environmental Epidemiology* **2006**, *16* (2), 106–114.

(8) Kleeman, M. J.; Schauer, J. J.; Cass, G. R. Size and Composition Distribution of Fine Particulate Matter Emitted from Motor Vehicles. *Environmental Science & Technology* **2000**, *34* (7), 1132–1142.

(9) RITA National Transportation Atlas Database http://www.bts.gov/publications/national_transportation_atlas_database (accessed Jul 8, 2010).

(10) Population Estimates http://www.census.gov/popest/estimates.html (accessed Nov 20, 2010).

(11) Censo General de Población y Vivienda 2000 http://www.inegi.org.mx/default.aspx (accessed Jan 9, 2011).

(12) 1996 and 2001 Canadian Census http://www12.statcan.gc.ca/census-recensement/index-eng.cfm (accessed Jan 9, 2011).

(13) National Emissions Inventory http://www.epa.gov/ttn/chief/eiinformation.html (accessed Jan 26, 2011).

(14) Christakos, G.; Serre, M. L. BME analysis of spatiotemporal particulate matter distributions in North Carolina. *Atmospheric Environment* **2000**, *34* (20), 3393–3406.

(15) Akita, Y.; Chen, J.-C.; Serre, M. L. The moving-window Bayesian maximum entropy framework: estimation of PM(2.5) yearly average concentration across the contiguous United States. *Journal of exposure science & environmental epidemiology* **2012**, *22* (5), 496–501.

(16) Christakos, G. *Modern spatiotemporal geostatistics*; Oxford University Press: New York, 2000.

(17) Mölter, A.; Lindley, S.; Vocht, F. De; Simpson, A.; Agius, R. Science of the Total Environment Modelling air pollution for epidemiologic research – Part II : Predicting temporal variation through land use regression. *Science of the Total Environment* **2010**, *409* (1), 211–217.

(18) Eeftens, M.; Beelen, R.; de Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dėdelė, A.; Dons, E.; de Nazelle, A.; Dimakopoulou, K.; Eriksen, K.; Falq, G.; Fischer, P.; Galassi, C.; Gražulevičienė, R.; Heinrich, J.; Hoffmann, B.; Jerrett, M.; Keidel, D.; Korek, M.; Lanki, T.; Lindley, S.; Madsen, C.; Mölter, A.; Nádor, G.; Nieuwenhuijsen, M.; Nonnemacher, M.; Pedeli, X.; Raaschou-Nielsen, O.; Patelarou, E.; Quass, U.; Ranzi, A.; Schindler, C.; Stempfelet, M.; Stephanou, E.; Sugiri, D.; Tsai, M.-Y.; Yli-Tuomi, T.; Varró, M. J.; Vienneau, D.; Klot, S. Von; Wolf, K.; Brunekreef, B.; Hoek, G. Development of Land Use Regression models for PM2.5, PM2.5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. *Environmental science & technology* **2012**, *46* (20), 11195–205.

(19) Dijkema, M. B.; Gehring, U.; Strien, R. T. Van; Zee, S. C. Van Der; Fischer, P.; Hoek, G.; Brunekreef, B. A Comparison of Different Approaches to Estimate Small-Scale Spatial Variation in Outdoor NO2 Concentrations. *Environmental Health Perspectives* **2011**, *119* (5), 670–675.

(20) Melymuk, L.; Robson, M.; Helm, P. a; Diamond, M. L. Application of land use regression to identify sources and assess spatial variation in urban SVOC concentrations. *Environmental science & technology* **2013**, *47* (4), 1885–95.

(21) Beelen, R.; Hoek, G.; Fischer, P.; Brandt, P. A. Van Den; Brunekreef, B. Estimated long-term outdoor air pollution concentrations in a cohort study. *Atmospheric Environment* **2007**, *41* (7), 1343–1358.

(22) Mavko, M. E.; Tang, B.; George, L. A. A sub-neighborhood scale land use regression model for predicting NO2. *Science of the Total Environment* **2008**, *398* (1-3), 68–75.

(23) Carr, D.; Ehrenstein, O. Von; Weiland, S.; Wagner, C.; Wellie, O.; Nicolai, T.; Mutius, E. Von Modeling Annual Benzene, Toluene, NO2, and Soot Concentrations on the Basis of Road Traffic Characteristics. *Environmental Research* **2002**, *90* (2), 111–118.

(24) Su, J. G.; Jerrett, M.; Beckerman, B. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. *Science of the Total Environment* **2009**, *407* (12), 3890–3898.

(25) Clougherty, J. E.; Wright, R. J.; Baxter, L. K.; Levy, J. I. Land use regression modeling of intra-urban residential variability in multiple traffic-related air pollutants. *Environmental Health* **2008**, *7* (17).

(26) Hoek, G.; Beelen, R.; Kos, G.; Dijkema, M.; Zee, S.; Fischer, P. H.; Brunekreef, B. Land Use Regression Model for Ultrafine Particles in Amsterdam. *Environmental Science & Technology* **2011**, *45* (2), 622–628.

(27) Ryan, P. H.; Lemasters, G. K.; Biswas, P.; Levin, L.; Hu, S.; Lindsey, M.; Bernstein, D. I.; Lockey, J.; Villareal, M.; Hershey, G. K. K.; Grinshpun, S. A. A Comparison of Proximity and Land Use Regression Traffic Exposure Models and Wheezing in Infants. *Environmental Health Perspectives* **2007**, *115* (2), 278–84.

(28) Mercer, L. D.; Szpiro, A. A.; Sheppard, L.; Lindström, J.; Adar, S. D.; Allen, R. W.; Avol, E. L.; Oron, A. P.; Larson, T.; Liu, L. S.; Kaufman, J. D. 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). *Atmospheric Environment* **2011**, *45* (26), 4412–4420.

(29) Brauer, M.; Hoek, G.; Vliet, P. Van; Meliefste, K.; Fischer, P.; Gehring, U.; Heinrich, J.; Cyrys, J.; Bellander, T.; Lewne, M.; Brunekreef, B. Estimating Long-Term Average Particulate Air Pollution Concentrations: Application of Traffic Indicators and Geographic Information Systems. *Epidemiology* **2003**, *14* (2), 228–239.

(30) Briggs, D. J.; Collins, S.; Elliott, P.; Fischer, P.; Kingham, S.; Lebret, E.; Pryl, K.; Reeuwijk, H. V. A. N.; Smallbone, K.; van der Veen, A. Mapping urban air pollution

using GIS□: a regression-based approach. *International Journal of Geographical Information Science* **1997**, *11* (7), 699–718.

(31) Smith, L. A.; Mukerjee, S.; Chung, K. C.; Afghani, J. Spatial analysis and land use regression of VOCs and NO2 in Dallas, Texas during two seasons. *Journal of Environmental Monitoring* **2011**, *13* (4), 999–1007.

(32) Dadvand, P.; Rushton, S.; Diggle, P. J.; Goffe, L.; Rankin, J.; Pless-mulloli, T. Using spatio-temporal modeling to predict long-term exposure to black smoke at fine spatial and temporal scale. *Atmospheric Environment* **2011**, *45* (3), 659–664.

(33) Arain, M.; Blair, R.; Finkelstein, N.; Brook, J.; Sahsuvaroglu, T.; Beckerman, B.; Zhang, L.; Jerrett, M. The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. *Atmospheric Environment* **2007**, *41* (16), 3453–3464.

(34) Gehring, U.; Eijsden, M. Van; Dijkema, M. B. A.; Wal, M. F. Van Der; Fischer, P.; Brunekreef, B. Traffic-related air pollution and pregnancy outcomes in the Dutch ABCD birth cohort study. *Occupational and Environmental Medicine* **2011**, *68* (1), 36–43.

(35) Jerrett, M.; Arain, M. A.; Kanaroglou, P.; Beckerman, B.; Crouse, D.; Gilbert, N. L.; Brook, J. R.; Finkelstein, N.; Finkelstein, M. M. 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.

(36) Moore, D. K.; Jerrett, M.; Mack, W. J.; Kunzli, N. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. *Journal of Environmental Monitoring* **2007**, *9* (3), 246–252.

(37) Ross, Z.; Jerrett, M.; Ito, K.; Tempalski, B.; Thurston, G. D. A land use regression for predicting fine particulate matter concentrations in the New York City region. *Atmospheric Environment* **2007**, *41* (11), 2255–2269.

(38) Kanaroglou, P. S.; Jerrett, M.; Morrison, J.; Beckerman, B.; Arain, M. A.; Gilbert, N. L.; Brook, J. R. Establishing an air pollution monitoring network for intra- urban population exposure assessment□: A location-allocation approach. *Atmospheric Environment* **2005**, *39* (13), 2399–2409.

(39) Henderson, S. B.; 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), 2422–2428.

(40) Brauer, M.; Lencar, C.; Tamburic, L.; Koehoorn, M.; Demers, P.; Karr, C. A Cohort Study of Traffic-Related Air Pollution Impacts on Birth Outcomes. *Environmental Health Perspectives* **2008**, *116* (5), 680–686.

(41) Marshall, J. D.; Nethery, E.; Brauer, M. Within-urban variability in ambient air pollution□: Comparison of estimation methods. *Atmospheric Environment* **2008**, *42* (6), 1359–1369.

(42) Gan, W. Q.; Koehoorn, M.; Davies, H. W.; Demers, P. A.; Tamburic, L.; Brauer, M. Long-Term Exposure to Traffic-Related Air Pollution and the Risk of Coronary Heart Disease Hospitalization and Mortality. *Environmental Health Perspectives* **2011**, *119* (4), 501–507.

(43) Gulliver, J.; Morris, C.; Lee, K.; Vienneau, D.; Briggs, D.; Hansell, A. Land Use Regression Modeling To Estimate Historic (1962-1991) Concentrations of Black Smoke and Sulfur Dioxide for Great Britain. *Environmental Science & Technology* **2011**, *45* (8), 3526–3532.

(44) Su, J. G.; Brauer, M.; Ainslie, B.; Steyn, D.; Larson, T.; Buzzelli, M. An innovative land use regression model incorporating meteorology for exposure analysis. *Science of the Total Environment* **2008**, *390* (2-3), 520–529.

(45) Hart, J. E.; Yanosky, J. D.; Puett, R. C.; Ryan, L.; Dockery, D. W.; Smith, T. J.; Garshick, E.; Laden, F. Spatial Modeling of PM10 and NO2 in the Continental United States, 1985–2000. *Environmental Health Perspectives* **2009**, *117* (11), 1690–6.

(46) Beelen, R.; Hoek, G.; Pebesma, E.; Vienneau, D.; Hoogh, K. De; Briggs, D. J. Mapping of background air pollution at a fine spatial scale across the European Union. *Science of the Total Environment* **2009**, *407* (6), 1852–67.

(47) Pearce, J. L.; Rathbun, S. L.; Aguilar-villalobos, M.; Naeher, L. P. Characterizing the spatiotemporal variability of PM 2.5 in Cusco, Peru using kriging with external drift. *Atmospheric Environment* **2009**, *43* (12), 2060–2069.

(48) Adamkiewicz, G.; Hsu, H.; Vallarino, J.; Melly, S. J.; Spengler, J. D.; Levy, J. I. Nitrogen dioxide concentrations in neighborhoods adjacent to a commercial airport: a land use regression modeling study. *Environmental Health* **2010**, *9* (73).