GLOBAL LAND USE REGRESSION AND BAYESIAN MAXIMUM ENTROPY
SPATIOTEMPORAL ESTIMATION OF PM$_{2.5}$ YEARLY AVERAGE CONCENTRATIONS
ACROSS THE UNITED STATES

Jeanette M. Reyes

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in the Department of Environmental Sciences and Engineering

Chapel Hill
2011

Approved by:
Marc Serre
Michael Flynn
William Vizuete
ABSTRACT

Jeanette M. Reyes: Global Land Use Regression and Bayesian Maximum Entropy Spatiotemporal Estimation of PM$_{2.5}$ Yearly Averaged Concentrations across the United States
(Under the direction of Marc L. Serre)

Knowledge of PM$_{2.5}$ concentrations across the United States is limited due to sparse monitoring across space and time. This work incorporates a land use regression (LUR) mean trend into the Bayesian Maximum Entropy (BME) framework along with Gaussian-truncated soft data that accounts for sampling incompleteness to provide estimations in the contiguous United States from 1999 to 2009. The LUR model was optimized to explain the most variability as possible given variable hyperparameters. Variables in the final model included elevation, average car miles driven, average traffic through-put, population density, SO$_2$ point source emissions, and NH$_3$ point source emissions. Compared to a kriging method with a constant mean trend this method showed a mean squared error reduction of over 35%. This is one of the few works to successfully develop a LUR model on a domain of this magnitude across space and time and incorporate the BME estimation methodology.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................... v

LIST OF FIGURES ........................................................................................................ vi

LIST OF ABBREVIATIONS ............................................................................................ vii

CHAPTER

1. INTRODUCTION ......................................................................................................... 1

2. MATERIALS AND METHODS .................................................................................... 4
   2.1 PM\textsubscript{2.5} data ................................................................................................. 4
   2.2 Variable Data ........................................................................................................ 6
   2.3 Model Development .............................................................................................. 8
   2.4 BME Methodology ............................................................................................... 13
   2.5 Cross-Validation Approach ................................................................................. 16
   2.6 Soft Data Construction ......................................................................................... 17

3. RESULTS .................................................................................................................... 17
   3.1 Rolling Yearly PM\textsubscript{2.5} Values ......................................................................... 18
   3.2 Land Use Regression Model ................................................................................. 19
   3.3 Covariance and Mean Trend Model ...................................................................... 22
   3.4 Estimation Maps .................................................................................................. 23
   3.5 Cross-Validation Statistics ................................................................................... 25

4. CONCLUSIONS ........................................................................................................... 26

5. SUPPORTING INFORMATION .................................................................................. 31

6. REFERENCES ............................................................................................................. 39
LIST OF TABLES

Table 1: Optimum univariate ranges and corresponding $\beta$ for each variable……………… 20

Table 2: Variables, ranges, $\beta$, and $r^2$ optimum model for each number of variables ........ 21

Table 3: Statistical measures for each of the four estimation methods ………………………… 25

Table 4: Table 5: Percent reduction compared to the hard/constant method ………………. 26
LIST OF FIGURES

Figure 1: PM$_{2.5}$ unique monitoring stations from 1999-2009 .......................... 19

Figure 2: Plot of $r^2$ versus the hyperparameter range for each univariate model ........20

Figure 3: Experimental (red circles) and modeled (blue lines) covariance of the LUR mean-
trend removed PM$_{2.5}$ yearly average concentrations shown (a) as a function of
spatial lag $r$ for a temporal lag $\tau = 0$, and (b) as a function of temporal lag $\tau$ for a
spatial lag $r = 0$ ...........................................................................................................23

Figure 4: BME estimation map across the contiguous US on January 1, 20002 for
methods (a) hard data/constant mean trend, (b) hard and soft data/LUR mean
trend ................................................................................................................................24

Figure 5: BME estimates using a LUR mean trend across the United States on January 1,
2003 overlaid with hard data (squares) and soft data (circles) ................................. 25
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT</td>
<td>Average Annual Daily Traffic</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criteria</td>
</tr>
<tr>
<td>AQS</td>
<td>Air Quality Systems</td>
</tr>
<tr>
<td>ASE</td>
<td>Absolute Standard Error</td>
</tr>
<tr>
<td>BME</td>
<td>Bayesian Maximum Entropy</td>
</tr>
<tr>
<td>BTS</td>
<td>Bureau of Transportation Statistics</td>
</tr>
<tr>
<td>CAA</td>
<td>Clean Air Act</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>EPA</td>
<td>Environmental Protection Agency</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographical Information Systems</td>
</tr>
<tr>
<td>INEGI</td>
<td>Instituto Nacional de Estadística Y Geografía</td>
</tr>
<tr>
<td>LUR</td>
<td>Land Use Regression</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>ME</td>
<td>Mean Error</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>NEI</td>
<td>National Emissions Inventory</td>
</tr>
<tr>
<td>NTAD</td>
<td>National Transportation Atlas Database</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Distribution Function</td>
</tr>
<tr>
<td>S/TRF</td>
<td>Space/Time Random Field</td>
</tr>
<tr>
<td>SE</td>
<td>Standard Error</td>
</tr>
<tr>
<td>VIF</td>
<td>Variance Inflation Factor</td>
</tr>
</tbody>
</table>
1. INTRODUCTION:

The Clean Air Act (CAA) was first established in 1970 as a means to control pollution from industries and from mobile sources across the United States (U.S. EPA, 2010). One of the pollutants which the EPA monitors is PM$_{2.5}$. PM$_{2.5}$ is particulate matter that has an aerodynamic diameter is 2.5 microns or smaller. Standards have been reduced significantly over the years due to an amalgamation of studies all geared towards human health. Regulation for particles was first established in 1971. However, the differentiation for particle size was first recognized in 1987 with PM$_{10}$ (particulate matter 10 microns in diameter or smaller). In 1997 the differentiation between PM$_{10}$ and PM$_{2.5}$ was established with a daily standard of 65 µg/m$^3$ and an annual standard of 15 µg/m$^3$. In 2006 regulations were tightened even further with a daily standard of 35 µg/m$^3$ and maintaining annual standard of 15 µg/m$^3$. Being able to meet these standards on a state and national level requires a large monitoring network which may be adaptive and collect data on a continuous basis. The PM$_{2.5}$ monitoring network was established on a national scale in 1999. Since then, it has added monitors, taken monitors away in some places, changed the sampling frequency, and sampling duration for the purpose of determining the most representative sample to meet regulatory standards and better protect human health. However, no matter how adaptive a network might be, it will have limitations due to lacking of resources and funding.

Standards of PM$_{2.5}$ were first established in light of human health. Health effects of PM$_{2.5}$ include asthma, respiratory infections, heart attacks, and premature death (U.S. EPA, 2010). Susceptible populations include infants, young children, the elderly, and the chronically ill.
A major hurdle in epidemiological research is assigning accurate exposure levels to general geographical areas or specific locations where subjects reside. Earlier epidemiological studies focused on short term (e.g. daily) exposure to ambient air pollution and its health effects. Short-term health effects include increased hospital admissions especially cardiac admissions, ischaemic heart disease, and premature death (Le Tertre, 2002, Ren, 2008). Short term exposure exhibits short range fluctuations across space and time, hence these earlier studies focused on short range fluctuations in the exposed population. However, more recently there has been an increased interest in the health effects caused by long term exposure to ambient air pollution. For example novel long term exposure health effects that are being investigated include neurocognitive, developmental, and sleep disturbances (Wang, 2007, Zanobetti, 2010). An important indicator of long term exposure to ambient air pollution is the yearly concentration of PM$_{2.5}$, which exhibits fluctuations over long spatial and temporal ranges. While several measurement and modeling studies have focused on daily PM$_{2.5}$, there have been less studies focusing on estimating yearly PM$_{2.5}$ concentration. In order to fill this gap the purpose of this paper is the accurate space/time estimation of the long range fluctuations of PM$_{2.5}$ yearly averaged concentrations.

Land-use regression (LUR) takes characteristics from the study area (traffic count, road length, road length weighted by road type, distance to nearest road, elevation, distance to nearest ocean, land cover, household density, number of trucks, wind, number of buildings, etc.) and develops a multiple linear regression model which aims at describing the variables associated with the pollutant of interest. LUR was first employed by Briggs (Briggs, 1997). This study measured NO$_2$ in 2 week chunks over the course of 4 seasons in Amsterdam, Huddersfield, and Prague as part of the larger SAVIAH study. Variables used included measures of altitude, land use, population density, and traffic. The final LUR model explained 61%-72% of the variability for the cities in question. Subsequent studies have
focused on other metropolitan areas like Montreal (Gilbert, 2005), Rome (Rosenlund 2008), Los Angeles (Moore, 2007), New York City (Ross, 2007), Toronto (Jerrett, 2007), etc. over varying time periods. These LURs aim at achieving high coefficient of determination $r^2$ over local study domains, and can therefore be classified as local LUR models. These local LUR models are generally adequate to represent short range variability in short term (e.g. daily) averages of ambient air pollution concentrations. However, these local LUR models are not well suited to estimate long-term exposure because they do not focus on describing variability over long ranges. In order to achieve high $r^2$ over local study domains, local LUR models tend to sacrifice the generalizability of these local models over larger study domains. It is therefore convenient to distinguish these local LUR models from global LUR models that we define as (1) describing long range variability in air pollutants and (2) de-emphasizing the $r^2$ achieved over a local domain in favor of the generalizability obtained over a large study domain. Since the purpose of our work is to model the long range fluctuations of long term average of ambient air pollutant over a large study domain, we will focus on the development of a global LUR model for yearly PM$_{2.5}$ over the entire USA and over multiple years, which characterizes a large space/time study domain. While Beelen looked at 15 European countries and developed a model for a single year (Beelen, 2009), and others have developed systematic methods for distributing monitors during campaigns over a geographical area (Kanaroglou, 2005), or addressed the discrepancy between large scale and small scale models (Dijkema, 2010), very few have addressed and implemented a LUR over a large spatial domain with temporal variation. The global LUR that we develop in this work will provide general knowledge about the large range trends of PM$_{2.5}$ yearly average concentration across space and time. The Bayesian Maximum Entropy (BME) method of modern spatiotemporal geostatistics (Puangthongthub, 2007) provides an ideal framework to blend the general knowledge provided by that global LUR model together with non-Gaussian data on the PM$_{2.5}$ yearly average concentration measured at specific monitoring
stations. Indeed, a challenge in measuring PM\textsubscript{2.5} yearly averaged concentrations is that some monitoring stations have missing daily PM\textsubscript{2.5} concentrations due to discontinued operations, resulting in incomplete sets of daily concentrations from which to calculate the corresponding yearly average concentration at those monitoring stations. Akita et al. (in preparation) developed a non-Gaussian soft data model that rigorously accounts for the uncertainty associated with yearly concentrations calculated from an incomplete set of daily concentrations. BME offers the flexibility to integrate this non-Gaussian soft data model together with the global LUR to yield accurate estimates of PM\textsubscript{2.5} yearly averaged concentrations at any unsampled locations.

Hence, in this work we use a global LUR to model the long range fluctuations of PM\textsubscript{2.5} yearly average concentration across the United States from 1999-2009 and we integrate this global LUR as a mean trend in the BME space/time estimation framework to calculate accurate estimates of the PM\textsubscript{2.5} yearly average concentration. Variables for the global LUR model in this study include elevation, miles driven (traffic I), miles driven per road segment (traffic II), population density, SO\textsubscript{2} point source emissions, NH\textsubscript{3} point source emissions, PM\textsubscript{2.5}-primary point source emissions, and NO\textsubscript{X} point source emissions. Variables will be explained in greater detail later. These predictor variables can be thought of as comprising three groups: elevation, emissions due to point sources (SO\textsubscript{2} point source emissions, NH\textsubscript{3} point source emissions, PM\textsubscript{2.5}-primary point source emissions, and NO\textsubscript{X} point source emissions), and emissions due to mobile sources (traffic I, traffic II, and population density). The hard and soft data for PM\textsubscript{2.5} yearly average concentration accounts for uncertainty associated with the calculation of a yearly concentration from an incomplete set of daily PM\textsubscript{2.5} concentrations. While few studies have focused on global LUR or on BME modeling of PM\textsubscript{2.5} yearly average concentration, this is one of the first studies to our knowledge that combines the global LUR and BME methods in the estimation of PM\textsubscript{2.5} yearly average concentration across a large space/time domain.
2. MATERIALS AND METHODS:

2.1 PM$_{2.5}$ Data

Daily PM$_{2.5}$ data were collected from the Environmental Protection Agency (EPA) Air Quality Systems (AQS) data base from 1999 to 2009 across the contiguous United States (US EPA, 2010). Data were modified to include only relevant fields pertinent to this study. These relevant fields include location information, sampling day, sampling frequency, sampling methodology, and measured values. Using the sampling methodology code, it was found that all sampling methods have a detection limit of 2 $\mu g/m^3$. A scaled histogram was created of all below detect values. A lognormal curve was fit to them. The mean of this curve was assigned to all values below the detection limit. All collocated daily values were combined using the arithmetic average. Although 5% of data were below detect, it is fruitful to keep these values. By omitting these values, one would lose the information that daily PM$_{2.5}$ values were low in certain areas ultimately causing an over estimating of PM$_{2.5}$ yearly averages.

Because the ultimate goal of this study is to understand long-term effects of PM$_{2.5}$, a more appropriate measure of long-term exposure given daily values would be a yearly average. Yearly average concentrations were constructed for (nearly) every instance a measured value was recorded. For every measured value, all values 365 days prior for that same station were averaged. For example, to construct a yearly average for a particular monitor on January 15, 1999, all daily values collected from January 15, 2008 through January 14, 1999 would be averaged for that given monitor. However, due to data starting on January 1, 1999, this “yearly” average may only consists of 14 daily values. Although 14 daily values may seem artificially low for constructing yearly averages, these averages will provide fruitful when constructed BME soft data explained in greater detail later. Because
this was done for every measured value, there is overlap between yearly averages. Thus, one can think of these averages as the PM$_{2.5}$ yearly average concentrations. The intended sampling frequency of a given daily monitoring station gave information about approximately how many daily values should have been recorded in a given year. Sampling frequencies were daily, every three days, every six days, or not defined. Thus, the expected number of measurements for one year prior was 365, 121, 60, and 25, respectively. (If the sampling frequency was undefined, then a value of 25 was assigned to the expected number of measurements.) The sampling frequency will become part of the BME soft data construction explained in section 2.5. Percent completeness is defined as the total number of daily value averaged divided by the expected number of daily values. If the percent completeness of a PM$_{2.5}$ yearly average is at least 75%, the value is considered hard (where there is no measurement error associated with the value). If the completeness is below 75%, the value is considered soft (there is measurement error associated with the value).

2.2 Variable Data

For the LUR model, possible contender variables included: elevation, total number of cars (traffic I), total number of cars per road length (traffic II), population density, SO$_2$ point source emissions, NH$_3$ point source emissions, PM$_{2.5}$-primary point source emissions, and NO$_x$ point source emissions.

Elevation data came in the form of a raster file of North America. The raster was clipped using an outline the United States and was converted to points. The mod function in ArcGIS™ was used to pick every-other point until the size of the shape file was small enough to be imported into MATLAB. The final elevation file contained over 1.1 million finely resoluted elevation data across the United States.

Point source emission data were obtained from the EPA’s National Emissions Inventory (NEI) website (US EPA, 2010). Data were separated by year and pollutant (i.e. SO$_2$ point source emissions, NH$_3$ point source emissions, PM$_{2.5}$-primary point source
emissions, and NO\textsubscript{X} point 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 point source emissions at the county-level were used. The emissions were modeled as a single stack being emitted from the centroid of each county for the years data exist. The data in this study ranges from January 2, 1999 to December 31, 2009. However, NEI data only exists for five time points (years with NEI data are assigned to December 31 of that year). Between NEIs, the EPA also publishes national annual emissions broken down by pollutant and tier one type. The total national non-mobile emissions were summed for each pollutant. The NEI point data using the last available NEI is then scaled by these national emissions to establish point source values for any given day. An example of this calculation is given in the supporting information. National annual emissions in 2008 were not available for NH_3 point source emissions and 2009 for SO_2 point source emissions, NH_3 point source emissions, PM\textsubscript{2.5}-primary point source emissions, and NO\textsubscript{X} point source emissions. Estimates of national emissions for these years are given in the supporting information.

Traffic data were obtained from the Bureau of Transportation Statistics's (BTS) National Transportation Atlas Database (NTAD) (RITA BTS, 2009). This contains GIS data about all major highways segments in the United States, including road length and Annual Average Daily Traffic (AADT) count. AADT is the number of vehicles that pass through a given road segment on average everyday (i.e. number of cars through a given road segment in a given year divided by 365). Because the AADT was only extensively available for 2009, AADT is estimated for other years by scaling 2009 data by national traffic emissions given by the EPA. (We assume the traffic data characterizes December 31, 2009.) National traffic emissions used are the sum of Tier One “Highway Vehicles” emissions for CO and NO\textsubscript{X} from 1998 to 2008. An example calculation for a given day can be found in the supporting information.
Population data were collected at the block group level in 2000 and county level estimates from 1998-2009 from the US Census website (US Census, 2009). Data were also collected from the Mexican census for 2000 and 2005 (INEGI, 2011) and the Canadian census for 2001 and 2006 (Canadian Census, 2011). Because the most accurate population estimate is one that can be estimated at the finest scale, block group population was estimated for every year. Thus, block group population is scaled to county level estimates in order to estimate block group population for any given day. We assign a date of July 1, 2000 to the 2000 census information and assign July 1 to county estimates. An example of a block group estimate is provided in the supporting information. 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 extended for any particular day where population estimates need to be calculated.

2.3 Model Development

The LUR model predicts PM$_{2.5}$ yearly average concentration given a group of variables. PM$_{2.5}$ yearly average concentration can be expressed using the following equation,

$$Z_{LUR,p} = \beta_0 + \beta_1 V_{1,p} + \beta_2 V_{2,p} + \cdots + \beta_n V_{n,p} + \epsilon_p$$  \hspace{1cm} (3)

where $Z_{LUR,p}$ is PM$_{2.5}$ yearly averaged concentrations, $V_{1,p}, ..., V_{n,p}$ are the predictor variables for PM$_{2.5}$ yearly average concentration for the space/time location $p$, $\epsilon_p$ is the error term, and $\beta_1, ..., \beta_n$ are the coefficients for each predictor variable.

Due to the large overlap in the PM$_{2.5}$ yearly average concentration, developing a model with all of the yearly values would be subject to potential collinearity. Because of this, only a subset of the data was used to construct the LUR model; namely, only the last PM$_{2.5}$ yearly average in a calendar year was used for each station. This constituted approximately
11,000 points. Along with the optimization of given variables, hyperparameters also need to be optimized. Many (e.g. Gilbert et. All, 2005, Rosenlund, 2008, Ross, 2006, etc.) have used differed scales by creating multiple sized buffer zones around the location of interest. The hyperparameter which results in the highest $r^2$ value is the one chosen for the multiple linear regression model. A hyperparameter is a physical parameter within nearly each variable that is allowed to adjust based on predictability of PM$_{2.5}$ yearly average concentrations. Hyperparameters include buffer sizes for traffic I, traffic II, and population density, and exponential decay ranges for SO$_2$ point source emissions, NH$_3$ point source emissions, PM$_{2.5}$-primary point source emissions, and NO$_X$ point source emissions. The hyperparameters range from 0.1 km to 1000 km. An exhaustive hyperparameter search is performed in the univariate case for each variable. The hyperparameter that results in the highest $r^2$ becomes an input parameter in the search routine for all the other models.

As mentioned earlier, conceptually variables fall into three groups: elevation, emissions due to mobile sources (traffic I, traffic II, and population density), and emissions from point sources (SO$_2$ point source emissions, NH$_3$ point source emissions, PM$_{2.5}$-primary point source emissions, and NO$_X$ point source emissions).

For each point source pollutant we construct a predictor representing the PM$_{2.5}$ yearly average concentration resulting from the point source emissions of that pollutant for the year prior on a given day. This is done by assuming that the PM$_{2.5}$ yearly average concentration resulting from a point source emission of that pollutant decreases exponentially away from that point source. Thus, the predictor variable describing the PM$_{2.5}$ yearly average concentration resulting from the point source emission of pollutant $i$ is

$$V_{i,p} = \sum_{n=1}^{N} em(s_n, t) \exp \left( -\frac{||s_n - s||}{\alpha_r} \right)$$  \hspace{1cm} (4)$$

where the predictor variable $V_{i,p}$ is the sum of exponentially decaying contributions from each point source emissions to a space/time location $p = (s, t)$, where $s = (s_1, s_2)$ is a
spatial location and $t$ is a time $t$, $em_i(s_n, t)$ is the emissions in tons/year of pollutant $i$ at location $s_n$ and time $t$, $i = NH_3, NO_x, PM_{2.5}, SO_2$, $s_n$ is the location in space for each point source $n$, $\|s_n - s_i\|$ is the spatial distance between locations $s_n$ and $s_i$, and $a_r$ is the exponential decay range in $km$ for the point source. Thus it is assumed that concentrations are additive.

Looking back at equation (3), $\beta$s multiplied by equation (4) for each point source emissions summed would represent the $PM_{2.5}$ yearly average concentration contribution due to point source emissions.

One variable contributing to the $PM_{2.5}$ yearly average concentration due to traffic is the traffic I variable. Traffic data come in the form of traffic counts for approximately 300,000 given road segments throughout the county. Traffic I describes the number of miles driven within a given buffer defined by the hyperparameter. This is calculated by multiplying the total number of cars that pass through a given road segment by the length of that road segment. The following equation is used to determine the total number of miles driven within a given buffer.

$$Traffic I(s, t) = \sum_{n=1}^{N} rdlen(s_n) * aadt(s_n, t)$$

where $s_n$ is the centroid of the $n^{th}$ road segment such that $\|s_n - s\| \leq a_r$, $rdlen(s_n)$ is the road length (in miles) of road segment $n$, $aadt(s_n, t)$ is the corresponding annual average daily traffic (in average number of cars per day), and $a_r$ is the radial buffer length in kilometers.

The next variable contribution of mobile sources related $PM_{2.5}$ yearly average concentration is traffic II. Traffic II describes the number of cars per mile on average within a given buffer. This is calculated by dividing traffic I by the sum of the road lengths within the buffer. Knowing only miles driven (traffic I) does not fully characterize emissions due to traffic. Traffic I neglects traffic congestion. By scaling traffic I by road length (traffic II), traffic
Il becomes an intrinsic variable that represents mobile emissions due to traffic congestion. The equation for average traffic (traffic II) in a given buffer is similar to total traffic.

\[ \text{Traffic II}(s, t) = \frac{\sum_{n=1}^{N} r \text{den}(s_n, t) \cdot \text{AADT}(s_n, t)}{\sum_{n=1}^{N} r \text{den}(s_n, t)} \] (6)

where \( ||s_n - s_i|| \leq a_r \), where \( a_r \) is the radial buffer length in kilometers.

The final variable describing total mobile source emissions of PM\(_{2.5}\) yearly average concentration is population density. Population density is the number of people within a given buffer divided by the area (\( km^2 \)) of that buffer. This variable is needed to correct for the over estimation of traffic I and traffic II. Traffic I and traffic II give equal weight to miles driven. This means every mile driven regardless of car type contributes the same amount of emissions. However, it is hypothesized that in highly populated areas, cars are generally smaller and contribute fewer emissions for every mile driven compared to less highly populated areas. Population density is similar to traffic in the sense that population is found within a given buffer and if the buffer contains the centroid of the block group, the entire area and population of the block group is included. This is explained in more detail in the supporting information. The equation for population density can be characterized by the following equation:

\[ \text{Pop Den}(s, t) = \frac{\sum_{n=1}^{N} \text{pop}(s_n, t)}{\pi a_r^2} \] (7)

where \( \text{pop}(\cdot) \) is the population within a given buffer, and \( ||s_n - s_i|| \leq a_r \), where \( a_r \) is the radial buffer length in kilometers. Notice the discrepancy between the areas of the block groups included in the calculation versus the area of the buffer. As the buffer size increases, the percent difference between the buffer size and the area of the block groups included decreases. This is further explained in the supporting information.

The sum of the \( \beta \)'s multiplied by the mobile source variables (traffic I, traffic II, and population density) would represent a variable of PM\(_{2.5}\) yearly average concentration due to mobile emissions.
Unlike the other variables, elevation is different in that it does not have a hyperparameter and is a spatial variable only. Elevation is approximated for each PM$_{2.5}$ yearly average concentration space/time location by using the closest known elevation point.

To best gauge which input hyperparameters would be a good initial estimate in the multivariate regression, univariate regressions were performed on each variable testing a wide range of buffer sizes and decay ranges. The hyperparameter that produced the highest $r^2$ was used as an input in the multivariate regression tests. The hyperparameter was chosen using a few criteria. First, if there was an absolute $r^2$ maximum at less than 950 km, it was chosen. If the maximum $r^2$ did occur at a hyperparameter value of 1000 km, then selected the largest hyperparameter value that resulted in a 95% or greater change in $r^2$. If the hyperparameter plot looks like a plateau, this will pick the point where the plateau starts. Otherwise, a hyperparameter is not chosen.

To determine which model resulted in the most predictive power, every combination of variables was tested using a search routine mentioned earlier. This includes eight univariate models, 28 bivariate models, 56 trivariate models, 70 four-variate models, 56 five-variate models, 28 six-variate models, 8 seven-variate models, and one eight-variate model, for a total of 255 possible models. The Akaike Information Criteria (AIC) and all Variance Inflation Factor (VIF) values were found for each model. The AIC value is a measure of parsimony in light of Mean Squared Error (MSE).

$$AIC = n \log \left( \frac{SSE}{n} \right) + 2(m + 1) \quad (8)$$

where $n$ is the number of unique PM$_{2.5}$ yearly averaged station/years (approximately 11,000 points) in the regression, $SSE$ is the error sum of squares, and $m$ is the number of predictor variables in a model. The VIF value is a measure of collinearity. If a variable within a model has a VIF value greater than 10, that particular variable is considered to be collinear with the remaining variables in a model. The VIF for the $j^{th}$ variable in a model of $m$ variables is
\[ VIF_j = \frac{1}{1-r_j^2} \] (10)

where \( r_j^2 \) is the correlation coefficient from regressing the \( j^{th} \) variable on the remaining \( m - 1 \) variables. Note that each model produces \( m \) VIF values. If any one of them is greater than 10, the model is said to have collinearity and is not considered as the final optimum LUR model. The model with the highest \( r^2 \) and a VIF value less than 10, for models with \( m = 1 \) variable, 2 variables, …, 8 variables are more closely inspected. Out of these eight possible models, the optimum LUR model was the model with the lowest AIC value given physically meaningful \( \beta \)s.

2.4 BME Methodology

BME is a mathematically rigorous geostatistical space/time framework which can incorporate information from many different sources and is implemented using the BMElib suite of functions in MATLAB\textsuperscript{TM} (BMElib, Serre and Christakos, 1999, Christakos, Bogaert, and Serre, 2002). This methodology is able to combine general knowledge about a process with site-specific knowledge in the form of measured values or other types of specific observations. In this work, general knowledge comes in the form of a mean trend and a covariance model and site-specific knowledge comes from PM\textsubscript{2.5} yearly average concentration. Site-specific knowledge can either be considered hard or soft. Hard data is site-specific knowledge with little to no uncertainty associated with the measured value. Soft data is site-specific knowledge with uncertainty. The power of BME comes in the form of the treatment of soft data. Soft data can be characterized by any Probability Distribution Function (PDF) (Gaussian, interval, uniform, triangle, etc.). In this work, soft data come in the form of truncated Gaussian PDFs.

The buttress of BME has been detailed in other works (Christakos 1990, Christakos, 2001, Serre, 1999), and can be summarized as performing the following steps: 1) gathering general and site-specific knowledge, 2) using the Maximum Entropy principle of information
theory to process the general knowledge in the form of a prior PDF, 3) integrating site-specific knowledge using an epistemic Bayesian conditionalization rule on both hard and soft data, and 4) creating space/time estimates based on the analysis.

The theory of space/time random field (S/TRF) is used to model the distribution of the PM$_{2.5}$ yearly average concentration across space and time. Our notation for variables will consist of denoting a single random variable $Z$ in capital letter, its realization, $z$, in lower case; and vectors and matrices in bold faces, e.g. $Z = [Z_1, ..., Z_n]^T$ and $z = [z_1, ..., z_n]^T$. Let $Z(p) = Z(s, t)$ be a S/TRF representing PM$_{2.5}$ yearly average concentration at spatial location $s = (s_1, s_2)$ and time $t$, and let the residual PM$_{2.5}$ yearly averaged concentration S/TRF $X(p)$ be defined as

$$X(p) = Z(p) - m_Z(p)$$

where $m_Z(s)$ is a global geographical trend that can be modeled using various deterministic models. In this work, we first set $m_Z(p)$ to a constant global geographical trend, and we then compare that approach with setting $m_Z(p)$ equal to the LUR model (Eq. 3). Equation (11) then expresses that the S/TRF $X(p)$ models the space/time variability and uncertainty associated with the difference between the S/TRF $Z(p)$ and a deterministic geographical trend model. Ultimately, the goal is to calculate $z_k$, the PM$_{2.5}$ yearly average concentration at some unsampled space/time location $p_k$. This is done by obtaining the BME estimate $\hat{x}_k$ for the residual PM$_{2.5}$ yearly average concentration at the estimation point $p_k$, given the general and site specific knowledge about the residual S/TRF $X(p)$, and adding back $m_z(p_k)$, the global mean trend calculated at $p_k$.

The BME fundamental set of equations for modeling the residual S/TRF $X(p)$ is (Christakos 2000, 2008, De Nazelle et al. 2010)

$$\left\{ \begin{align*}
\int d\mathbf{x} \ (\mathbf{g}(\mathbf{x}) - E[\mathbf{g}])e^{\mu^T g(\mathbf{x})} &= 0 \\
\int d\mathbf{x} f_\delta(\mathbf{x})e^{\mu^T g(\mathbf{x})} - Af_k(x_k) &= 0
\end{align*} \right.$$  

(12)
where \( x \) is a vector of residual PM\(_{2.5}\) yearly average concentrations at mapping points \( p \) consisting of the union of the data points \( p_d \) and the estimation point \( p_K \). \( g \) is a vector of functions selected such that their expected values \( E[g] \) is known from the general knowledge base G-KB, \( f_S(x) \) is a PDF characterizing the knowledge and uncertainty associated with the site specific knowledge base S-KB, \( A \) is a normalization constant, and \( f_K \) is the BME posterior probability density function describing residual PM\(_{2.5}\) yearly average concentration at the estimation point \( p_K \), where the subscript \( K = G \cup S \) means that \( f_K \) is based on the blending of the G- and S-KB.

The G-KB for the residual S/TRF \( X(p) \) consists of the local space/time mean trend function

\[
m_x(p) = E[X(p)]
\]

and the covariance function

\[
c_x(p, p') = E[(X(p) - m_x(p))(X(p') - m_x(p'))].
\]

The S-KB comprises the hard and soft data. The hard data \( x_h = z_h - m_Z(p_h) \) consist of values measured without significant errors of the residual PM\(_{2.5}\) yearly average concentration at the set of hard data points \( p_h \). On the other hand, the soft data \( x_s \) available at points \( p_s \) have an associated uncertainty described by the PDF \( f_s(x_s) \).

In this study the G-KB and S-KB can therefore be written as \( G = \{m_x(p), c_x(p, p')\} \) and \( S = \{x_h, x_s\} \), and in this case the BME fundamental set of equations reduces to (Christakos, 2000)

\[
f_K(x_K) = A^{-1} \int dx \ f_S(x_s) f_G(x)
\]

where \( x = (x_K, x_h, x_s) \) is a realization of \( X \) at points \( p = (p_K, p_h, p_s) \), \( f_G(x) = e^{u^T g(x)} \) is the Gaussian PDF for \( X \) with mean and covariance matrix obtained from the G-KB, and \( A \) is a normalization constant.
As described in previous works (Serre, 1999, Christakos, 2000), in the case that the S-KB is reduced to hard data and soft data with Gaussian distributions, then BME reduces to the classical kriging method with measurement errors. A key conceptual difference between classical kriging and this work is that the PDF $f_x(x_s)$ describing the uncertainty associated with the soft data is not restricted to the Gaussian distribution. This allows us to rigorously integrate any non-Gaussian soft data, such as soft data with a truncated Gaussian distribution, as described next.

2.5 Hard and Soft Data

The PM$_{2.5}$ yearly average concentration at some time $t$ is defined as the average of PM$_{2.5}$ daily concentrations over the 365 days preceding time $t$. In practice the collection frequency of daily measurements changes from one monitoring station to another. Let $n_i^*$ be the intended number of daily measurements at station $i$ under normal operating conditions. In this study, $n_i^*$ is typically equal to 365, 121, 60 or 25, while the actual number of PM$_{2.5}$ daily concentrations recorded at that station for the 365 days prior to some time $t$ is $n_{i,t} \leq n_i^*$. An assumption that underlies previous works (Akita et al., in preparation) is that the $n_i^*$ have been selected by monitoring agencies so that if $n_{i,t} \geq 0.75n_i^*$, then the arithmetic average of the $n_{i,t}$ PM$_{2.5}$ daily concentrations $y_{i,j}$, ($j = 1 \ldots, n_{i,t}$) taken prior to time $t$, i.e.

$$\mu_{i,t} = \frac{\sum_{j=1}^{n_{i,t}} y_{i,j}}{n_{i,t}}$$

is a hard datum for the PM$_{2.5}$ yearly average concentration at site $i$ and over a year span ending at time $t$. In this work we therefore calculate a hard datum using Eq.16 for each site $i$ and time $t_h$ for which a daily concentration was measured, and for which $n_{i,t_h} \geq 0.75n_i^*$, i.e. for which the set of intended PM$_{2.5}$ daily concentrations for the 365 days prior to $t_h$ is at least 75% complete.
Furthermore, we construct a soft datum that models the PM$_{2.5}$ yearly average concentration at each site $i$ and time $t_s$ for which a daily concentration was measured, and for which $0 < n_{i,t_s} < 0.75n_{i}^*$, in other words whenever the set of intended PM$_{2.5}$ daily concentrations is incomplete. Consistent with the assumption underlying the hard data, we set the expected value of the soft datum to the arithmetic average $\mu_{i,t_s}$ calculated from Eq. 16, and its variance to

$$\sigma_{i,t_s}^2 = \frac{n_i^* - n_{i,t_s}}{n_i^*} \frac{\sum_{j=1}^{n_{i,t_s}} (y_{i,j} - \mu_{i,t_s})^2}{n_{i,t_s}}$$

(17)

The first term of this equation is a finite population correction factor that measures the incompleteness of the intended set of daily concentrations, and the second term quantifies the variability of measured daily concentrations within the yearly period. Because PM$_{2.5}$ yearly average concentrations cannot be negative, the PDF describing the soft data are truncated below zero. In this work, we use a Gaussian PDF truncated below zero, with mean and variance calculated using Eq. 16 and 17, respectively.

2.6 Cross-Validation Approach

In order to test the estimation improvement of LUR and BME, a cross-validation was performed to compare four different methods: 1) setting the global mean trend $m_Z(\mathbf{p})$ to a constant value and considering all data as hard, 2) using this constant global mean trend with both hard and soft data, 3) setting the global mean trend to the LUR model and considering all data as hard, and 4) using this LUR global mean trend with both hard and soft data. For each of these methods, the cross validation procedure consists of randomly selecting 20,000 hard data points, removing each one at a time, and re-estimating it from the remaining PM$_{2.5}$ yearly average concentration values. In this manner we get for each method a set of true and cross-estimated values, from which we can calculate estimation errors, and the corresponding mean square error (MSE) as well as any other statistics. The
estimation method with the lowest MSE is the most accurate amongst all four methods compared.

3. RESULTS

3.1 PM$_{2.5}$ Yearly Average Concentration

Out of the 1,587,223 original PM$_{2.5}$ daily concentration values measured from 1999 to 2009, we obtained 1,479,759 daily values after collocated measurements were averaged, and we calculated 1,478,149 PM$_{2.5}$ yearly average concentrations using eq. 16. These yearly values came from 1,576 unique monitoring sites predominantly located along populated areas across the contiguous United States (Figure 1). As described earlier, the below detect values were assigned the mean of the lognormal distribution they formed, which was calculated to be 1.63 $\mu g/m^3$. Out of 1,478,149 PM$_{2.5}$ yearly average concentration values, 438,966 (29.69%) were below the 75% completeness threshold (i.e. 0 < $n_{i,t} < 0.75n_i$), and therefore modeled as soft data. The mean of the PM$_{2.5}$ yearly average concentration is 12.44 $\mu g/m^3$, the variance is 11.57, the skewness is 0.56, and the kurtosis is 5.75. The minimum value is 1.63 $\mu g/m^3$ and the maximum value is 75.40 $\mu g/m^3$. The histogram of the data appears slightly normally distributed (Figure S3).
3.2 Land Use Regression Model

Initial building of the model began with looking at the univariate regression of a subset of the PM$_{2.5}$ yearly average concentrations (consisting of only the last yearly average calculated for each monitoring station) with respect to each of the eight variables: elevation, traffic I, traffic II, population density, SO$_2$ point source emissions, NH$_3$ point source emissions, PM$_{2.5}$-primary point source emissions, and NO$_x$ point source emissions. Except for elevation, each variable is controlled by a hyperparameter range, which consists of the radial buffer size for traffic I, traffic II, and population density, and the exponential decay range for the point source emissions of SO$_2$, NH$_3$, PM$_{2.5}$-primary, and NO$_x$ for the year prior. For each univariate model, the optimal hyperparameter range was obtained by plotting the $r^2$ versus the hyperparameter range (Figure 2), and selecting the hyperparameter value that resulted in the highest $r^2$ (or that was at the beginning of a plateau, as explained earlier). Note that elevation is not included in the table because elevation has no hyperparameter.
In the univariate case, traffic I has the highest $r^2$, i.e. it explained the most variation observed in the PM$_{2.5}$ yearly average concentrations. All $\beta$s are positive except for elevation (Table 1).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Hyperparameter Range (km)</th>
<th>$\beta$</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>0.0000</td>
<td>-0.0033799</td>
<td>0.1864</td>
</tr>
<tr>
<td>Traffic I</td>
<td>695.4968</td>
<td>0.00376441</td>
<td>0.4483</td>
</tr>
<tr>
<td>Traffic II</td>
<td>30.5035</td>
<td>3.64E-05</td>
<td>0.1196</td>
</tr>
<tr>
<td>Population Density</td>
<td>1046.2500</td>
<td>0.1473</td>
<td>0.2820</td>
</tr>
<tr>
<td>SO$_2$ point source emission</td>
<td>304.6362</td>
<td>3.77E-06</td>
<td>0.2752</td>
</tr>
<tr>
<td>NH$_3$ point source emission</td>
<td>93.0182</td>
<td>3.51E-05</td>
<td>0.0453</td>
</tr>
<tr>
<td>NO$_x$ point source emission</td>
<td>322.7500</td>
<td>7.27E-06</td>
<td>0.2355</td>
</tr>
<tr>
<td>PM$_{2.5}$-primary point source emission</td>
<td>190.0020</td>
<td>2.85E-05</td>
<td>0.1551</td>
</tr>
</tbody>
</table>
Table 1: Optimum hyperparameter range and corresponding $\beta$ and $r^2$ for each univariate regression model

We then investigated $m$-variate regression models with $m = 1$ to 8 variables. For each $m$, we considered each combination of $m$ variables out of the total of 8, and selected the optimal combination, defined as that which resulted in the highest $r^2$ given a maximum VIF value less than 10 (except in the 8-variate model). In all multi-variate cases, the hyperparameters were selected such that the $r^2$ for the model was maximized. For the final six-variate model, however, the hyperparameters were similar to those selected in the univariate case (Table S1). Table 2 lists the optimal combination of variables for each $m$-variate model, with the corresponding $r^2$, VIF and AIC. Note that a model with $m$ variables produces $m$ VIF values, of which the maximum is reported. We note that even though for a given $m$-variate model all possible combinations of $m$ variables are tested, the combination of $m$ variables that results in the highest $r^2$ seems (mostly) to be built off the variables selected for the optimal $m - 1$ model. For example, the optimum 4-variable model contained the variables from the optimum 3-variable model (traffic I, traffic II, and $SO_2$) plus elevation.

<table>
<thead>
<tr>
<th>Var #</th>
<th>Variables</th>
<th>$r^2$</th>
<th>VIF</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Traffic I</td>
<td>0.44828</td>
<td>0</td>
<td>125716</td>
</tr>
<tr>
<td>2</td>
<td>Traffic I, Traffic II</td>
<td>0.50282</td>
<td>1.03</td>
<td>124548</td>
</tr>
<tr>
<td>3</td>
<td>Traffic I, Traffic II, $SO_2$</td>
<td>0.51625</td>
<td>2.36</td>
<td>124242</td>
</tr>
<tr>
<td>4</td>
<td>Traffic I, Traffic II, $SO_2$, Elevation</td>
<td>0.5262</td>
<td>2.51</td>
<td>124011</td>
</tr>
<tr>
<td>5</td>
<td>Traffic I, Traffic II, $SO_2$, Elevation, $NH_3$</td>
<td>0.52952</td>
<td>2.67</td>
<td>123935</td>
</tr>
<tr>
<td>6</td>
<td>Traffic I, Traffic II, $SO_2$, Elevation, $NH_3$, Population Density</td>
<td>0.53088</td>
<td>4.15</td>
<td>123904</td>
</tr>
<tr>
<td>7</td>
<td>Traffic I, Traffic II, $SO_2$, Elevation, $NH_3$, Population Density, PM$_{2.5}$-primary</td>
<td>0.53584</td>
<td>6.01</td>
<td>123786</td>
</tr>
<tr>
<td>8</td>
<td>Traffic I, Traffic II, $SO_2$, Elevation, $NH_3$, Population Density, PM$_{2.5}$-primary, NO$_X$</td>
<td>0.55738</td>
<td>41.93224</td>
<td>123255</td>
</tr>
</tbody>
</table>

Table 2: Variables, ranges, $\beta$'s, and $r^2$ optimum model for each number of variables
Out of the eight contender models (Table 2), we selected as our final model the optimal 6-variable model, which uses the following 6 variables: elevation, traffic I, traffic II, population density, SO$_2$ point source emissions, and NH$_3$ point source emissions. The 8-variate model was not chosen because it has at least one VIF value greater than 10. The 7-variate model was not chosen because the $\beta$ for PM$_{2.5}$-primary was negative, causing it not to be physically meaningful.

### 3.3 Mean Trend and Covariance Models

Two different models were used for the global mean trend function $m_z(p)$ defined in Eq (11): a constant global mean trend value and the LUR model. The constant global mean trend was calculated as the average of all PM$_{2.5}$ yearly average concentration values, which we found to be 11.72 ($\mu g/m^3$). On the other hand, the LUR global mean trend values ranged from about 6 ($\mu g/m^3$) to about 18 ($\mu g/m^3$).

Values for the residual PM$_{2.5}$ yearly average concentration S/TRF $X(p)$ were calculated using Eq. (11) by subtracting the LUR global mean trend from the PM$_{2.5}$ yearly average concentration values. Experimental covariance values where then calculated for various spatial and temporal lags. The experimental covariance value obtained for the LUR mean trend removed data decreases as the spatial or temporal lag increases (Figure 3). We then fit a permissible covariance model to the experimental covariance values. We found through least square fitting that the covariance model with the best fit was

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

where $c_1 = 2.3780$, $a_{r_1} = 0.5$ degrees, $a_{r_1} = 900$ days, $c_2 = 3.5670$, $a_{r_2} = 6.5$ degrees, $a_{r_2} = 500000$ days. This covariance model is the sum of two space/time covariance structures. The first structures models variability occurring over short spatial and temporal ranges, while the second structure models variability at longer spatial and temporal ranges.
The constant, mean-trend removed covariance model explains short, medium, and long range transport associated with PM$_{2.5}$ (Figure S5) and is more fully explained in the supporting information.

![Image]

**Figure 3:** Experimental (red circles) and modeled covariance (blue line) of the LUR mean-trend removed PM$_{2.5}$ yearly average concentrations shown (a) as a function of spatial lag $r$ for a temporal lag $\tau = 0$, and (b) as a function of temporal lag $\tau$ for a spatial lag $r = 0$.

### 3.4 Estimation Maps

Figure 4 shows the United States for January 1, 2002. The left figure shows the BME estimates using only hard data and a constant mean trend and the right figure shows the BME estimates using hard and soft data and a LUR mean trend.
Figure 4: BME estimation map across the contiguous US on January 1, 2002 for methods (a) hard data/constant mean trend, (b) hard and soft data/LUR mean trend

Without the LUR model, the “hot spots” in Alabama, Ohio, and the top of West Virginia would have been missed. The LUR mean trend is able to capture more small scale variability due to the availability of LUR variables on a smaller scale.

Figure 5 shows a BME estimation map overlaid with the hard (squares) and soft (circles) data points for January 1, 2003. In general, there is good agreement between the BME estimates and the yearly averaged hard and soft data. There is a hot spot in California and higher values on the eastern half of the United States that are shown with the hard and soft data and picked up by the BME estimates. This closely resembles the estimated LUR mean trend (Figure S3).
3.5 Cross-Validation Statistics

To test the effectiveness of the LUR model along with the addition of soft data, a number of statistical measures were calculated using a cross-validation. Due to the large number of yearly values, only a randomly generated subset of 20,000 points was used for the cross-validation statistics. The subset was randomly generated to characterize the entire data set as a whole. Table 3 shows the results for the four methods (i.e. constant mean trend with hard data, constant mean trend with hard/soft data, LUR mean trend with hard data, LUR mean trend with hard/soft data) used for each cross-validation statistic. Table 4 shows the percent reduction (or in the case of the Pearson and Spearman correlation, the increase) compared to the BME estimates using only hard data and a constant mean trend.

<table>
<thead>
<tr>
<th>Method</th>
<th>hard/constant</th>
<th>hard and soft/constant</th>
<th>hard/LUR</th>
<th>hard and soft/LUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>2.3209</td>
<td>2.0886</td>
<td>1.643</td>
<td>1.5004</td>
</tr>
<tr>
<td>Method</td>
<td>hard/constant</td>
<td>hard and soft/constant</td>
<td>hard/LUR</td>
<td>hard and soft/LUR</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>------------------------</td>
<td>------------</td>
<td>------------------</td>
</tr>
<tr>
<td>MSE</td>
<td>0</td>
<td>10.0090</td>
<td>29.2085</td>
<td>35.3527</td>
</tr>
<tr>
<td>ME</td>
<td>0</td>
<td>33.5851</td>
<td>52.7019</td>
<td>77.5131</td>
</tr>
<tr>
<td>MAE</td>
<td>0</td>
<td>5.9347</td>
<td>25.2275</td>
<td>30.7072</td>
</tr>
<tr>
<td>SE</td>
<td>0</td>
<td>4.6729</td>
<td>15.8879</td>
<td>18.6916</td>
</tr>
<tr>
<td>ASE</td>
<td>0</td>
<td>4.8193</td>
<td>9.6386</td>
<td>12.0482</td>
</tr>
<tr>
<td>Pearson's Corr.</td>
<td>0</td>
<td>2.6004</td>
<td>7.8649</td>
<td>9.4965</td>
</tr>
<tr>
<td>Spearman's Corr.</td>
<td>0</td>
<td>0.8355</td>
<td>2.8914</td>
<td>3.5180</td>
</tr>
</tbody>
</table>

Table 3: Statistical measures for each of the four estimation methods

Table 4: Percent reduction compared to the hard/constant method

The biggest reduction in MSE is seen incorporating the LUR mean trend using the hard and soft data. Incorporating soft data with a constant mean trend shows a modest percent decrease except for the ME where there is an increase. There was a larger percent decrease using hard data with a LUR mean trend in every measure. Most notably the ME had a large percent decrease and MSE decrease was sizable. Incorporating the soft data with a LUR mean trend performed better than having hard data with a LUR mean trend.

4. DISCUSSION

The difficulty in regulatory settings and epidemiological settings is to ensure pollutant levels are below some regulatory standard and accurately determine a person’s exposure. Many epidemiology studies look at historical data to assign exposure. Due to sparseness of the available data set, however, misclassification can occur. The BME methodological
framework combines multiple knowledge sources to better estimate exposure. One type of general knowledge sources which can be incorporated into BME is the mean trend. The mean trend can be constant, linear, or some polynomial form of order two or higher. This work uses a LUR model as a mean trend. LUR models take geographical characteristics of the surrounding area where data exist to predict the pollutant of interest. The final model predicted PM$_{2.5}$ yearly average concentration using elevation, a traffic variable describing the number of miles driven, another traffic variable describing average rate of traffic flow, SO$_2$ point source emissions, and NH$_3$ point source emissions.

Conceptually the independent variables for the LUR model can be divided into three groups: elevation, emissions due to point sources, and emissions due to traffic. By grouping the variables in this manner, the $\beta$'s make physical sense. Historically, elevation has a negative $\beta$ in LUR models (Beelen, 2009, Ryan, 2007, Rosenlund, 2008, Poplawski, 2009). This is due to how PM$_{2.5}$ components change to differences in elevation. Places at higher elevations have lower pressures. This causes an increase in CO and hydrocarbons but a decrease in NO$_x$ (EPA, 1976). Thus if a parcel of air were to move to a higher elevation where air is cleaner, the volume of the parcel would increase by the ideal gas law ($PV = nRT$) resulting in a lower concentration ($\text{concentration} = \text{mass}/\text{volume}$). Both NH$_3$ and SO$_2$ have positive $\beta$'s: as NH$_3$ and SO$_2$ levels increase, so does predicted PM$_{2.5}$ yearly average concentration. For traffic, there is a discrepancy between the data available and the estimation needed. Traffic data collected are traffic counts – not traffic emissions. This assumes that every mile driven produces the same amount of emissions regardless of car type. Population density corrects for this assumption. That is, the traffic variables alone overestimate PM$_{2.5}$ yearly average concentration. However, we hypothesize that areas with high population density tend to have smaller cars, which (in general) are more fuel efficient. Thus the $\beta$ for population density is negative, in support of our hypothesis, to correct for this overestimation of traffic emissions.
Although the optimum seven-variable model had a permissible VIF value and a higher $r^2$ than the optimum six-variable model, the six-variable model was chosen. This was due to the grouping of the variables. In the seven-variable model, grouping $\text{SO}_2$ point source emissions, $\text{NH}_3$ point source emissions, and $\text{PM}_{2.5}$-primary point source emissions leads to negative predicted $\text{PM}_{2.5}$ yearly average concentration. The $\beta$ for $\text{PM}_{2.5}$-primary point source emissions in the seven-variate model was negative, making no physical sense. Mathematically, the regression equation was able to increase the $r^2$ of the seven-variate model; however, the increase is an artifact of the regression calculation and not of product of having the variables be physically meaningful. Thus, the seven-variate model was not the final optimum LUR model and the six-variate model was chosen. Out of the total tons of emissions, $\text{NH}_3$ makes up the smallest amount and $\text{SO}_2$ made up the largest amount (Figure S5a). In the univariate cases for all variables (except elevation), $\beta$s were positive. When more variables are added, certain $\beta$s will decrease to compensate for the contributions from other predictors. Figure S4a shows predicted $\text{PM}_{2.5}$ yearly average concentration using the univariate Traffic II model. Figure S4b shows predicted $\text{PM}_{2.5}$ yearly average concentration using Traffic II and the $\beta$ from the final six-variate model. Using the final $\beta$s for Traffic II predicts a smaller concentration than the univariate case due to ignoring the contributions from the remaining five variables. Mobile sources also contribute more to $\text{PM}_{2.5}$ yearly average concentration in the final model than point sources (Figure S5b).

The statistical measures overall showed the biggest percent decrease (from the hard data/constant mean trend method) came from using hard and soft data with a LUR mean trend. Incorporating soft data with a constant mean showed only modest improvement. This may be due to the number of soft data points selected. About 30% of the data set was soft data, but 20,000 points were randomly selected from the entire data set. Perhaps if the
cross-validation set contained a large percentage of soft data, the percent decrease in the statistical measures would have been larger.

Although this study used a substantial amount of data, there were some notable gaps. For the traffic variables, traffic was calculated within given buffers. However, traffic data were available only for the United States. Traffic is not characterized accurately when the buffer extends into Mexico and Canada. Because the domain size is so large, assumptions were made about the data. For example, population data at the block group level were only available for 2000 in the United States. For all other years, population was scaled according to county estimates. Point source data were only available for 1998-2002 and 2005. For all other years, point sources were scaled according to national levels. From the literature a few LUR models incorporate wind direction. That would not have been appropriate for that work simply due to lack of nationally known wind directions at a fine temporal and spatial scale at point source emission locations and PM$_{2.5}$ monitoring stations. Because this work investigated yearly averages, prevailing winds over an entire year may average to approximate a circular shape. Thus incorporating wind direction may not prove fruitful in estimating PM$_{2.5}$ yearly average concentrations.

This work was able to successfully incorporate a LUR mean trend model into the BME estimation method. The statistical measures calculated show a large percent decrease when using a LUR mean trend. The LUR mean trend is also able to pick up more of small scale variability and shows good agreement with measured values. This is one of the first known studies to implement a LUR model on such a large spatial and temporal domain. This is one of the few studies which have incorporated a temporal component into an LUR model, given physical meaning to the $\beta$ parameters in the model, grouped variables in a physically meaningful way, and incorporated a LUR model into the BME estimation framework.
Future work could include adding more sources that could be protective such as parks, green space, or lack of road connectivity. This work focused mostly on anthropogenic sources. Other potential variables could include biogenic sources. All variables were evaluated separately. Perhaps investigating interaction between variables could improve the LUR model. Because new data are always available, new data could be incorporated when they become available. This could include adding 2010 census information or 2010 NTAD traffic data. Data were calculated using a rolling yearly average. If the averaging period were shorter (e.g. six months or one month) more temporal variables (e.g. temperature, precipitation, etc.) could accurately be incorporated into the LUR model.
5. SUPPORTING INFORMATION

Example of scaling point sources to national emissions to establish point source values for any given day

The following steps would be used to calculate point source data from March 18, 2004. 1) The last available NEI data would be chosen. In this case the last available NEI is December 31, 2002. 2) A linear interpolation would be 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 point source values for March 18, 2004. Thus the scaling ratio is \[ \frac{\text{national emissions for March 18, 2004}}{\text{national emissions for December 31, 2002}} \]. This ratio is multiplied to all December 31, 2002 point values.

Estimating national \( \text{NH}_3 \) emissions for December 31, 2008

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

\[
\frac{\text{national } \text{NH}_3 \text{ emissions for December 31, 2008}}{\text{national } \text{NH}_3 \text{ emissions for December 31, 2007}} = \frac{\text{national } \text{NH}_3 \text{ emissions for December 31, 2007}}{\text{national } \text{NH}_3 \text{ emissions for December 31, 2006}}.
\]

The national \( \text{NH}_3 \) emissions for December 31, 2006 and December 31, 2007 are known. These values can be plugged in to solve for national \( \text{NH}_3 \) emissions for December 31, 2008.

Estimating national emissions for December 31, 2009 for \( \text{NH}_3 \), \( \text{NO}_x \), \( \text{PM}_{2.5} \), and \( \text{SO}_2 \)

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

\[
\frac{\text{national } X \text{ emissions for December 31, 2009}}{\text{national } X \text{ emissions for December 31, 2008}} = \frac{\text{national } X \text{ emissions for December 31, 2008}}{\text{national } X \text{ emissions for December 31, 2007}}.
\]
where \( X = \text{NH}_3, \text{NO}_X, \text{PM}_{2.5}, \text{and SO}_2 \). Emissions for all national point sources are known for December 31, 2007 and December 31, 2008. These values can be plugged in to solve for national point source emissions.

**Example of scaling traffic counts to traffic emissions to estimate traffic count for any given day**

The following steps would be 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 \( \frac{\text{national traffic emissions for June 28, 2003}}{\text{national traffic emissions for December 31, 2002}} \). This ratio is multiplied by every AADT.

**Example of scaling block group population to estimate block group population for any given day**

The following steps would be 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:

\[
\frac{\text{block group population for July 1, 2002}}{\text{county population estimates for July 1, 2002}} = \frac{\text{block group population for July 1, 2000}}{\text{county population for July 1, 2000}}.
\]

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.

**Calculating population density**

Like stated earlier, when calculating population density, a buffer is extended from each PM$_{2.5}$ space/time location. If the centroid 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 below.

![Figure S1: The exact buffer size (a) when calculating population density compared with the approximate centroid inclusion method (b) for an arbitrary PM$_{2.5}$ location](image)

The black dots are the centroids 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. Below shows how these methods compare with different buffer sizes.
Figure S2: Comparing the areas of the buffer using the centroid method to the exact area for up to 20 km.

Figure S2 shows the change in area as a function of buffer size for the PM$_{2.5}$ station shown in Figure S1. This centroid method shows good agreement with the exact area of the buffer. Thus it can be said that the centroid method is an appropriate approximation for calculating population density.

Figure S3: Histogram of all PM$_{2.5}$ yearly average concentration.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Table S1: Optimized hyperparameter for the final six-variate LUR model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic I</td>
<td>689.9873</td>
</tr>
<tr>
<td>Traffic II</td>
<td>32.9566</td>
</tr>
<tr>
<td>Population</td>
<td>730.0000</td>
</tr>
<tr>
<td>Density</td>
<td></td>
</tr>
<tr>
<td>NH₃</td>
<td>10.4990</td>
</tr>
<tr>
<td>SO₂</td>
<td>210.3490</td>
</tr>
</tbody>
</table>

Figure S4: Histogram of the mean trend calculated using the LUR method

Further explanation of the constant, mean-trend covariance model

The constant, mean-trend removed covariance model was found to be

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

where \( c_1 = 4.6293 \), \( a_{r_1} = 0.75 \text{ degrees} \), \( a_{\tau_1} = 2000 \text{ days} \), \( c_2 = 4.6293 \), \( a_{r_2} = 8 \text{ degrees} \), \( a_{\tau_2} = 10000 \text{ days} \), \( c_3 = 2.3146 \), \( a_{r_3} = 30 \text{ degrees} \), \( a_{\tau_3} = 10000 \text{ days} \). This covariance model has three exponential structures. In space, there is a short, medium, and long range. This describes the small scale spatial variability to the larger scale variability of PM$_{2.5}$. The ranges
in time are all long. This corresponds to historical levels in a given location. That is, PM$_{2.5}$ levels are consistently low in areas with historically low levels.

Figure S5: Experimental and modeled spatial and temporal covariance for a constant mean trend removed data

Figure S6: BME estimates using a LUR mean trend across the United States on January 1, 2003 overlaid with hard data (squares) and soft data (circles)
Figure S7: Predicted PM$_{2.5}$ using the LUR model (a)
Figure S8: Predicted PM$_{2.5}$ using (a) the univariate Traffic II model and (b) Traffic II data using the Traffic II $\beta$ from the final six-variate model.

Figure S9: Total contribution in tones from (a) point sources (tons/year) and (b) predicted PM$_{2.5}$ contributions from point sources and mobile emissions ($\mu$g/m$^3$).
6. REFERENCES


term effects of particulate air pollution on cardiovascular diseases in eight European cities.” J Epidemiol Community Health. 56: 773-779.


