Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yuqi Sun Date

Development and Application of a Statistical Emulator for Estimating Personal Exposure to Ambient Air Pollution

By

Yuqi Sun

MPH, Emory University

Rollins School of Public Health

Department of Biostatistics

_____ **[Chair's Signature]**

Howard H. Chang

_____ **[Member's Signature]**

Stefanie Ebelt Sarnat

Development and Application of a Statistical Emulator for Estimating Personal Exposure to Ambient Air Pollution

By

Yuqi Sun

Bachelor of Medicine, Anhui Medical University

MPH, Emory University

Rollins School of Public Health

2015

Thesis Committee Chair: Howard H. Chang, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics

2015

Abstract

Development and Application of a Statistical Emulator for Estimating Personal Exposure to Ambient Air Pollution

By

Yuqi Sun

$PM_{2.5}$ are particles that has aerodynamic diameter equal or smaller than 2.5 μm and can penetrate into the region where lungs exchange gas. Current evidence indicates that $PM_{2.5}$ can result in adverse respiratory symptoms, reduction in lung function, hospital admissions, physician visits for respiratory illness, chronic cough and asthma. The majority of health studies on $PM_{2.5}$ use measurements from fixed location monitors. However, these measurements are only for outdoor levels which may not reflect human exposure to pollution from outdoor sources and they cannot capture variations in exposure between people. This study develops an emulator for estimating population exposure to $PM_{2.5}$ and its variance using a Bayesian hierarchical model. The emulator will contribute to the relevance of large population-based health studies by producing an exposure metric with greater biological relevance than the traditional use of ambient concentrations.

Development and Application of a Statistical Emulator for Estimating Personal Exposure to Ambient Air Pollution

By

Yuqi Sun

MPH, Emory University

Rollins School of Public Health

2015

Advisor: Howard H. Chang

Acknowledgements

I thank the faculty, advisors, and staff of the Biostatistics Department at Rollins School of Public Health for the dynamic two years of learning that I have had. This thesis is a sample of the vast knowledge that was attained and applied through my two years here at Rollins. I would especially like to thank Professor Howard H. Chang for all of his advice and support to help me write this thesis. Also I would like to give a special thanks to Professor Stefanie Ebelt Sarnat for taking the time to read my thesis.

Contents

1 Introduction

1.1 Backgound

Particulate air pollution represents a mixture of solid and liquid particles suspended in the air. The particles' sizes vary from several nm to tens of μm. Among them, smaller particles (μm scale) are commonly generated from combustion, condensation or chemical reactions. $PM_{2.5}$ are particles that have aerodynamic diameters equal or smaller than 2.5 μm and they can penetrate into the region where lungs exchange gas [1−2]. Based on previous epidemiological and toxicological studies on $PM_{2.5}$, these particles have greater toxicity than larger particles since they can penetrate into the lungs easier and contain more chemically active species [3−4]. Studies have shown that increased rates of mortality and morbidity for respiratory and cardiovascular diseases are associated with exposure to high level of $PM_{2.5}$ [5]. Experimental studies have also found that inhalable particles will aggravate airway pathology by inducing inflammation. Specifically current evidence indicates that $PM_{2.5}$ can result in adverse respiratory symptoms, reduction in lung function, hospital admissions, physician visits for respiratory illness, chronic cough and asthma [6−8]. Hence associations between air pollution and adverse health outcomes from epidemiological studies have played a major role in public health by informing regulatory standards [9].

1.2 Problem statement

The majority of health studies on $PM_{2.5}$ use measurements from fixed-location outdoor monitors. However, the monitors have limited spatial coverage and more of them are placed in urban areas [10]. Moreover, these measurements are only for outdoor levels which may not reflect human exposure to pollution from outdoor sources since people stay indoors the majority of time [11]. Also (1) exposure error may arise from unobserved spatial variation in air pollution concentrations and (2) spatial variations in population or environmental characteristics that contribute to different exposures. The use of outdoor pollution levels also means that previous studies cannot incorporate variations in exposure between people in the health analysis [12].

1.3 Purpose statement

The overarching goal of this project is to develop spatial statistical methods to estimate daily personal exposure to $PM_{2.5}$ and apply it in an air pollution and health study. Since knowledge of the variance of population exposure is useful in reducing ecological bias [13], our approach will consider both the daily mean and variance of population exposure to ambient air pollution. We will then examine the short-term associations between $PM_{2.5}$ levels and daily emergency department visits for respiratory diseases and asthma using different exposure metrics in a time-series analysis.

1.4 Significance statement

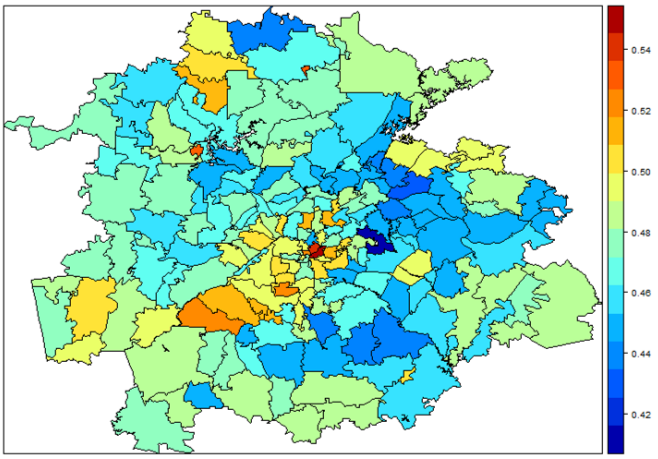
This study develops an approach to account for exposure measurement error in time-series study of air pollution and health. Methods to account for exposure uncertainty will contribute to the reliability and relevance of large population-based health studies.

1.5 Data

For developing the statistical model, we utilized an existing dataset of daily simulated personal exposures to ambient $PM_{2.5}$ from the Stochastic Human Exposure and Dose Simulation (SHEDS) in Atlanta from years 1999 to 2002 [21]. SHEDS is a probabilistic model that estimates the population distribution of total PM exposures. Exposure simulation were conducted by a two stage Monte Carlo approach. Simulated exposures reflect demographic proportions of sex, age, and residential housing condition as well as air exchange rate (AER) and time spent in different places. For our data, 100 hypothetical individuals were simulated for each census tract on each day in order to estimate the daily distribution of population exposures. These simulations were then aggregated to the zipcode level. Daily zipcode level outdoor concentrations entered into the simulation were hybrid of data obtained from monitors and output from dispersion model which uses emission sources and meteorological information. Indoor $PM_{2.5}$ concentrations for residential microenvironment and physical factor data are calculated by mass balance equation. $PM_{2.5}$ concentrations in nonresidential microenvironments

are calculated using equations developed from indoor and outdoor measurement data for different microenvironments. Demographic data for population are also included in the input [22–24]. Figure 1 shows average daily exposure/concentration ratios aggregated at the zipcode level, illustrating the spatial pattern in Atlanta. The ratio ranges from 0.42 to 0.54. Several factors influenced the association between indoor and outdoor $PM_{2.5}$ levels. Examples include: temperature, wind speed and infiltration factor.

Figure 1: Average daily exposure/concentration ratios by zipcode (data 1999 to 2002)



For years 2002 to 2008, also obtained daily $PM_{2.5}$ ambient concentrations, AER, wind speed, temperature, population, meterological observation. These variables were compiled similarly to those used to run the SHEDs simulations from the US Census 2000 and Census 2010 demographic proportions of sex and age were obtained. For the health analysis, individual-level records of emergency department (ED) visits from 41 of 42 hospitals during the same period were obtained from individual hospitals and Georgia Hospital Association for the 20-county metropolitan Atlanta area.

2 Methods

2.1 Model development

Our goal is to develop a statistical model to predict personal exposure mean and variance estimated by SHEDS. We first carried out a model selection procedure for identifying useful predictors that are inputs and parameters of the SHEDS algorithm.

When fitting the model, mean personal exposure and variance of personal exposure were transformed to the log scale because of right skewness. Since we aimed to link outdoor pollutant levels to personal exposure, all predictors were interacted with log outdoor $PM_{2.5}$ levels. The selection of final model was based on prediction performance. Akaike information criterion (AIC), root-mean-square error (RMSE), mean absolute error (MAE), R-square and 95% coverage were calculated to compare different models. For calculating RMSE, MAE and 95% coverage, we treated year 1999 to 2001 as training dataset and 2002 as testing dataset. Model 1 included $PM_{2.5}$ and weekend interacted with $PM_{2.5}$. Model 2 included $PM_{2.5}$, weekend interacted with $PM_{2.5}$ and wind speed interacted with $PM_{2.5}$. Model 3 included $PM_{2.5}$, weekend interacted with $PM_{2.5}$ and temperature interacted with $PM_{2.5}$. Model 4 included $PM_{2.5}$, weekend interacted with $PM_{2.5}$, temperature interacted with $PM_{2.5}$ and wind speed interacted with $PM_{2.5}$. Model 5 included $PM_{2.5}$, weekend interacted with $PM_{2.5}$ and AER interacted with $PM_{2.5}$. All models were fitted separately for each zipcode.

Based on prediction performance (see appendix), Model 5 was selected as the model we used in next step. We also identified evidence of heterogeneity in the regression coefficients across zipcodes.

To account for spatial heterogeneity across zipcodes, random intercepts and random slopes were introduced into the folowing hierarchical models. For mean personal exposure model we have:

$$y_{s,t}^{(1)} = \beta_{0,s}^{(1)} + \beta_{1,s,t}^{(1)} x_{1,s,t} + \beta_{2,s,t}^{(1)} x_{2,s,t} + \varepsilon_{s,t}^{(1)} \tag{1a}$$

$$\beta_{0,s}^{(1)} = \gamma_{0,0}^{(1)} + \alpha_s^{(1)}; \quad \beta_{1,s,t}^{(1)} = \gamma_{1,0}^{(1)} + \boldsymbol{Z_{1,s,t}^{(1)} \gamma_1^{(1)}} + \psi_s^{(1)}; \quad \beta_{2,s,t}^{(1)} = \gamma_{2,0}^{(1)} + \boldsymbol{Z_{2,s,t}^{(1)} \gamma_2^{(1)}} + \phi_s^{(1)}$$

Similarly, for variance of personal exposure model:

$$y_{s,t}^{(2)} = \beta_{0,s}^{(2)} + \beta_{1,s,t}^{(2)} x_{1,s,t} + \beta_{2,s,t}^{(2)} x_{2,s,t} + \varepsilon_{s,t}^{(2)} \tag{1b}$$

$$\beta_{0,s}^{(2)} = \gamma_{0,0}^{(2)} + \alpha_s^{(2)}; \quad \beta_{1,s,t}^{(2)} = \gamma_{1,0}^{(2)} + \boldsymbol{Z_{1,s,t}^{(2)} \gamma_1^{(2)}} + \psi_s^{(2)}; \quad \beta_{2,s,t}^{(2)} = \gamma_{2,0}^{(2)} + \boldsymbol{Z_{2,s,t}^{(2)} \gamma_2^{(2)}} + \phi_s^{(2)}$$

where $y_{s,t}^{(1)}$ denotes the log mean of personal exposure and $y_{s,t}^{(2)}$ denotes the log

variance of personal exposure in zipcode s and on day t. $x_{1,s,t}$ is the outdoor $PM_{2.5}$ concentration on log scale. $x_{2,s,t}$ is the interaction term between AER and log $PM_{2.5}$ concentration. $\boldsymbol{Z^{(i)}_{j,s,t}}$ is a vector of covariates used in SHEDS simulations which can be spatially-varying or time-varying. We considered percent male, percent of population from 15 to 65, temperature and wind speed. Specifically, $\boldsymbol{Z^{(1)}_{1,s,t}}$ included percent of male, temperature, weekend, wind speed,percent of male interacted with weekend, wind speed interacted with AER and temperature interacted with weekend. $\boldsymbol{Z^{(1)}_{2,s,t}}$ included temperature. $\boldsymbol{Z^{(2)}_{1,s,t}}$ included percent of people whose age from 15 to 65, temperature, weekend, wind speed, temperature percent of male interacted with weekend and percent of people whose age from 15 to 65 interacted with weekend. $\boldsymbol{Z^{(2)}_{2,s,t}}$ included temperature. For the random effects, we assumed $\alpha_s^{(1)} \sim N(0, \sigma^2_{u^{(1)}}\boldsymbol{I})$, $\psi_s^{(1)} \sim N(0, \sigma^2_{g^{(1)}}\boldsymbol{I})$, $\phi_s^{(1)} \sim N(0, \sigma^2_{d^{(1)}}\boldsymbol{I})$, $\alpha_s^{(2)} \sim N(0, \sigma^2_{u^{(2)}}\boldsymbol{I})$, $\psi_s^{(2)} \sim N(0, \sigma^2_{g^{(2)}}\boldsymbol{I})$ and $\phi_s^{(2)} \sim N(0, \sigma^2_{d^{(2)}}\boldsymbol{I})$. Finally, assumed $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon^{(1)}_{s,t} \\ \varepsilon^{(2)}_{s,t} \end{bmatrix} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$.

The exposure model was estimated under a Bayesian framework and used Markov chain Monte Carlo (MCMC) with the following priors. $\boldsymbol{\Sigma} \sim inv - wish(4\widehat{\boldsymbol{\Sigma}}, 4)$. We estimated $\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$ by fitting Model 1a and Model 1b without random intercept and random effects. We further assumed $\sigma^2_{u^{(1)}} \sim Inv - gamma(\alpha_{u^{(1)}}, \beta_{u^{(1)}})$, $\sigma^2_{g^{(1)}} \sim Inv - gamma(\alpha_{g^{(1)}}, \beta_{g^{(1)}})$, $\sigma^2_{d^{(1)}} \sim Inv - gamma(\alpha_{d^{(1)}}, \beta_{d^{(1)}})$, $\sigma^2_{u^{(2)}} \sim Inv - gamma(\alpha_{u^{(2)}}, \beta_{u^{(2)}})$, $\sigma^2_{g^{(2)}} \sim Inv - gamma(\alpha_{g^{(2)}}, \beta_{g^{(2)}})$ and $\sigma^2_{d^{(2)}} \sim Inv - gamma(\alpha_{d^{(2)}}, \beta_{d^{(2)}})$. All hyper-parameters α and β were chosen to be 0.01. For $\mu_0^{(i)}, \gamma_{i,0}^{(j)}$ and each parameter in vectors of $\boldsymbol{\gamma^{(j)}_{i,s,t}}$: i=1,2 and j=1,2, we assumed they followed $N(0, 100^2)$. The MCMC was run with 15,000 iterations with the first 5000 iterations treated as burn-in. The MCMC samples were thinned by 10.

Since zipcodes are contiguous areal units, a conditional autoregressive (CAR) model was also considered in order to account for spatial correlation [25]. We considered 2 additional models for random effects shown in Table 1. For the CAR models, let matrix $\boldsymbol{W}$ be an $n * n$ spatial weight matrix that captures the dependence structure between areas. If areas i,j share the same boundary then $w_{i,j} = 1$. Also let $\boldsymbol{N}$ be a diagonal matrix where $n_{i,i}$ equals to the total number of neighbors for area i. For the CAR model, we assumed the random effects followed a mean-zero normal distribution

with covariance $\tau^2(\boldsymbol{N}-\boldsymbol{W})^{-1}$.

Table 1: Three models for comparison

	Random Intercept	Random slope(PM2.5)	Random slope(AER)
Model 1	Normal	Normal	Normal
Model 2	Spatial	Normal	Normal
Model 3	Spatial	Spatial	Spatial

To compare different random effect specifications, a cross-validation experiment was performed. Using data from 1999 to 2000 as training data and data from 2001 to 2002 as testing data, deviance information criterion (DIC), R-square, RMSE, MAE, average standard deviation and 95% coverage were calculated for comparison.

We used the posterior samples of the final model to predict mean personal exposure and variance of personal exposure for the period 2002-2008 in each zipcode and on each day. We then aggregated the exposure to the 20-county Atlanta area by using population size in zipcode s on day t. Since we only had data on population size for 2000 and 2010 from census, we assumed a linear trend between these two years.

2.2 Health effect model

Let y_t denote the total number of ED visits for the outcome of interest on day t. In our study the outcome was respiratory disease. Since the health outcome was aggregated over zipcodes, the desired exposure should represent the average exposure experienced by the at-risk population. We consider a Poisson log-linear model given by [26]:

$$logE(y_t) = \beta\mu_{t-q} + \frac{1}{2}\beta^2\sigma^2_{t-q} + \boldsymbol{V_t}\boldsymbol{\gamma} \tag{2}$$

Here β is the parameter of interest and corresponds to the log relative risk assocaited with $PM_{2.5}$ level on q^{th} lagged day. From the exposure model, μ_{t-q} is the predicted overall population mean exposure on day $t-q$. σ^2_{t-q} is the corresponding overall variance on day t. $\boldsymbol{V_t}$ is a vector of covariates including smooth functions of daily minimum temperature natural cubic spline with daily dew point temperature (degree of freedom = 3), weekend, federal holidays, long-term trend (12 degree of freedom per year) and indicator of hospital exits and entrance in the dataset [21].

Metropolis-Hastings algorithm was used to estimate parameters in the health effect model. The MCMC had 20,000 iterations with the first 5000 iterations as burn-in. We took the every 10^{th} sample. To update μ_t and σ_t^2, we randomly drew $\mu_t^{(i)}$ and $\sigma_t^{2(i)}$ from their posterior samples predicted by the exposure models. To update β, we used a proposal distribution $\beta^{(i)} \sim N\left(\beta^{(i-1)}, \tau^2\right)$ where τ^2 is a tuning parameter which controls the acceptance rate to 60%. We updated $\boldsymbol{\gamma}$ with proposal $\boldsymbol{\gamma^{(i)}} \sim MVN\left(\boldsymbol{\gamma^{(i-1)}}, \zeta\Sigma_{\widehat{\gamma}}\right)$ where $\Sigma_{\widehat{\gamma}}$ is the covariance matrix derived from initial maximum-likelihood estimation (MLE) fit and ζ is the tuning parameter which controls the acceptance rate at about 30%.

Using the health model, we also examined associations using population-coverage ambient concentration of $PM_{2.5}$, as well as using personal exposures but without considering variance of mean exposure. Lagged effects from of day 0 to day 3 were examined.

3 Results

Table 2 summaries the results from different random effect specifications. The DIC from Model 1 to Model 3 were -348367.1,-348367.9, -348404.5 respectively. The decreasing DIC trend indicates that the spatial model gave a better fit than the independent normal model.

Table 2: Result of comparison for the three models

	Population exposure model			Exposure variance model		
	Model (1)	Model (2)	Model (3)	Model (1)	Model (2)	Model (3)
R-square	0.97394	0.97396	0.97396	0.89177	0.89183	0.89195
RMSE	0.55748	0.55732	0.55725	1.93724	1.93667	1.93563
MAE	0.38282	0.38285	0.38276	0.89550	0.89529	0.89508
Average sd	3.42805	3.42779	3.42751	5.52422	5.52328	5.52050
95% coverage	0.94807	0.94804	0.94812	0.94037	0.94043	0.94028

Table 3: Posterior of estimates in final model by using equation 2

Population exposure model				Exposure variance model			
Parameter	Posterior mean	2.5% quantile	97.5% quantile	Parameter	Posterior mean	2.5% quantile	97.5% quantile
Intercept	-0.752	-0.754	-0.750	Intercept	-4.117	-4.123	-4.111
log PM2.5 (Main)	0.900	0.876	0.919	log PM2.5 (Main)	2.372	2.300	2.451
*AER	0.493	0.484	0.502	*AER	-0.304	-0.330	-0.280
*Weekend	-0.101	-0.103	-0.098	*Weekend	-0.145	-0.155	-0.135
*Percent of Male	-0.057	-0.094	-0.009	*Percent Age 15 to 65	-0.203	-0.324	-0.103
*Temperature (*100)	-0.024	-0.022	-0.027	*Temperature (*100)	-0.087	-0.096	-0.079
*AER interacted with Temperature (*100)	-0.078	-0.086	-0.068	*AER interacted with Temperature (*100)	0.509	0.482	0.538
*Weekend with percent of male	0.048	0.044	0.052	*Weekend interacted with percent of male	-0.058	-0.073	-0.042
*Weekend interacted with temperature (*100)	0.022	0.021	0.024	*Weekend interacted with percent age 15 to 65	0.073	0.060	0.086
*Wind speed (*100)	0.486	0.463	0.507	*Wind speed	-0.022	-0.022	0.021
*Wind speed interacted with AER (*100)	-0.923	-0.954	-0.891	σ^2 (*100)	0.070	0.069	0.070
σ^2 (*100)	0.533	0.530	0.536				

The spatial models do increase the prediction performance for our exposure model. The spatial model, had larger r-square and smaller RMSE, MAE, average standard deviation. However, the improvements were very minor. So the normal model (Model 1) was chosen as the final model to predict personal mean exposure and variance for 2002-2008. The results also show that r-squares in exposure variance model are less than those in population exposure model. This means it is more difficult to predict variance of exposure.

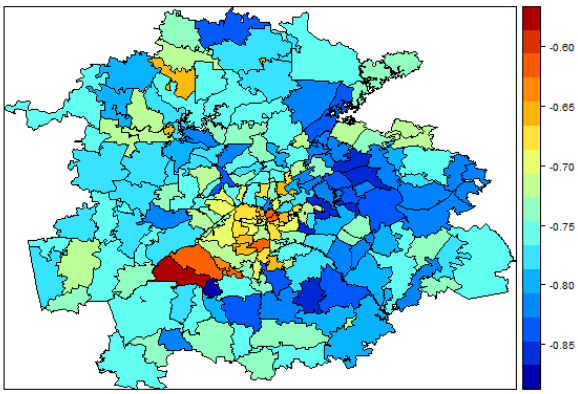
Results of parameter estimates for the final model fitted using the complete data (1999-2000) are shown in Table 3. In this table, we treat log $PM_{2.5}$ as main effect and all other predictors are interacted with main effect indicated by $*$.

In the population exposure model, the coefficient of AER is positive which means higher AER leads to higher personal exposure given the same outdoor concentration. Wind speed is also positively associated with personal exposure. This is probably because the higher AER results in more similar pollution concentration levels in both outdoor and indoor. And wind speed is an important predictor for AER. The coefficient of weekend is negative, meaning that lower personal exposure happens during weekend. The reason of this may be that people spend more time at home during weekend. In the exposure variance model, AER has a negative coefficient. Wind speed is also negatively associated with variance of personal exposure. Weekend also has a negative association with variance of personal exposure.

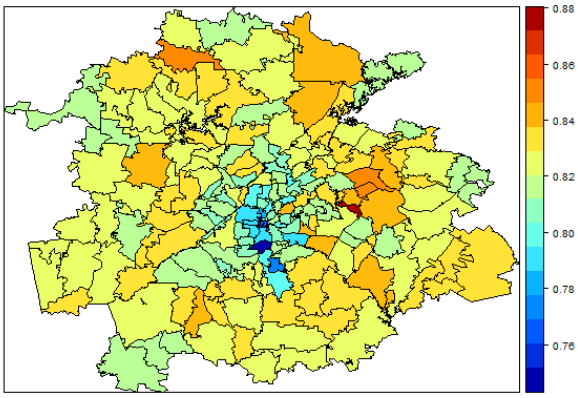
Maps of posterior means of intercepts and main effects in the final models are shown from Figure 1 to Figure 2. The figures indicate that the impacts of main effects of outdoor $PM_{2.5}$ and AER * log($PM_{2.5}$) are less in the center of Atlanta in mean exposure model. And in the variance of exposure model, the impact from main effect $PM_{2.5}$ is less in the central area. The reason of these results may be that there are more buildings in the central area of Atlanta and people tend to spend more time indoor in these areas.

Figure 2: Intercept and main effect across zipcode of mean exposure model

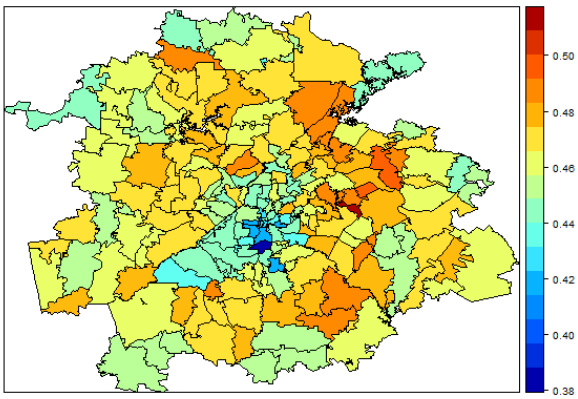
$$y_{s,t}^{(1)} = \beta_{0,s}^{(1)} + \beta_{1,s,t}^{(1)} log(PM2.5) + \beta_{2,s,t}^{(1)} AER * log(PM2.5) + \varepsilon_{s,t}^{(1)} \text{ (equation 1)}$$



(a) Posterior mean of $\beta_{0,s}^{(1)}$ across zipcodes



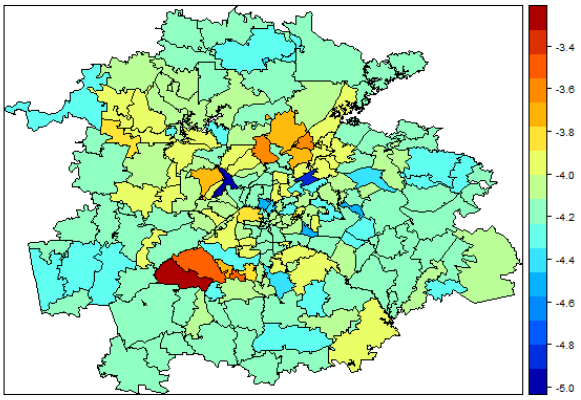
(b) Posterior mean of $\beta_{1,s,t}^{(1)}$ across zipcodes



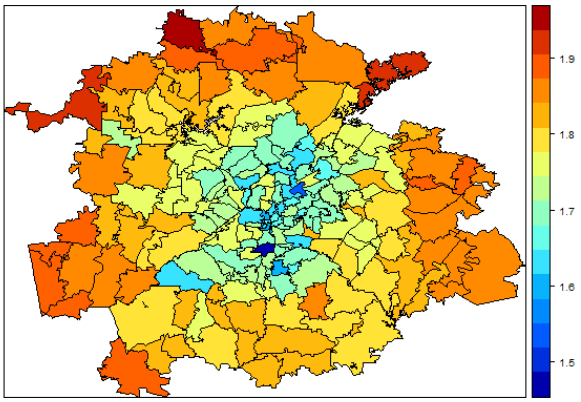
(c) Posterior mean of $\beta_{2,s,t}^{(1)}$ across zipcodes

Figure 3: Intercept and main effect across zipcode of variance Of mean exposure model

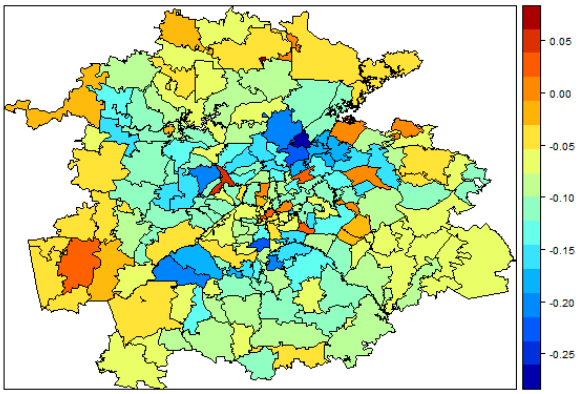
$$y_{s,t}^{(2)} = \beta_{0,s}^{(2)} + \beta_{1,s,t}^{(2)} log(PM2.5) + \beta_{2,s,t}^{(2)} AER * log(PM2.5) + \varepsilon_{s,t}^{(2)} \text{ (equation 1)}$$



(a) Posterior mean of $\beta_{0,s}^{(2)}$ across zipcodes



(b) Posterior mean of $\beta_{1,s,t}^{(2)}$ across zipcodes



(c) Posterior mean of $\beta_{2,s,t}^{(2)}$ across zipcodes

Table 4 shows the summary statistics of different exposures aggregated over the 20-county Atlanta region which were used in health effect model. The mean personal exposure is about half of the outdoor concentration. Tables 5-8 show the results from the three different health posterior models for lagged effect from day 0 to day 3. Relative risk (RR), 95% confidence interval of RR and RR per interquartile range with 95% posterior interval are given.

Table 4: Exposure summary used in health effect model (ug /m^3)

Model	Mean	Median	SD	IQR
Ambient concentration (population weighted)	15.365	14.271	7.097	(10.112, 19.055)
Mean Exposure	7.250	6.691	3.239	(4.881, 8.827)
Variance of Mean Exposure	17.359	7.588	32.521	(4.751, 14.626)

Table 5: Lagged day 0 effect of three health effect model

Model	Relative Risk	95% PI	RR*IQR (95% PI)
Ambient concentration (population weighted)	1.00092	(1.000510, 1.001331)	1.008258 (1.004570,1.011966)
Population exposure (Assume variance=0)	1.00100	(1.000015,1.001882)	1.003952 (1.000059,1.007447)
Population exposure	1.00101	(1.000161,1.001873)	1.003991 (1.000635,1.007411)

Table 6: Lagged day 1 effect of three health effect model

Model	Relative Risk	95% PI	RR*IQR (95% PI)
Ambient concentration (population weighted)	1.00102	(1.000638,1.001408)	1.009159 (1.005720,1.012662)
Population exposure (Assume variance=0)	1.00169	(1.000806,1.002466)	1.006685 (1.003184,1.009766)
Population exposure	1.00168	(1.000760,1.002521)	1.006646 (1.003002,1.009985)

Table 7: Lagged day 2 effect of three health effect model

Model	Relative Risk	95% PI	RR*IQR (95% PI)
Ambient concentration	1.00106	(1.000679,1.001439)	1.009520
(population weighted)			(1.006080,1.012934)
Population exposure	1.00198	(1.001156,1.002810)	1.007836
(Assume variance=0)			(1.004569,1.011134)
Population exposure	1.00191	(1.001121,1.002787)	1.007558
			(1.004431,1.011043)

Table 8: Lagged day 3 effect of three health effect model

Model	Relative Risk	95% PI	RR*IQR (95% PI)
Ambient concentration	1.00081	(1.000439,1.001187)	1.007267
(population weighted)			(1.003933,1.010666)
Population exposure	1.00184	(1.000883,1.002474)	1.007280
(Assume variance=0)			(1.003489,1.009798)
Population exposure	1.00171	(1.000773,1.002562)	1.006765
			(1.003054,1.010148)

When using ambient concentration of $PM_{2.5}$ in health model, the relative risks are around 1.00100 for all lagged days. When using population exposure and its variance, the relative risks which vary from 1.00101 to 1.0191 and are much higher than that derived from ambient concentration. Also the differences in risks across different lag days are more apparent compared to using ambient concentration. The results using population exposure and variance will give a more direct way to describe the relationship between personal exposure and health effect. Results of using mean of population exposure in the models are similar to the result of using both mean and variance of population exposure. But the results of using mean in the model is slightly higher than using both mean and variance. This means using the mean exposure-only model may overestimate the relative risk. Also across results for different exposure metrics, there is evidence that the relative risk at 2 day lag is the highest. Tables 4-8 also provide the relative risks per IQR range increase in the exposure metrics. Overall

risks associated with per IQR increase in personal exposure are smaller than that from ambient.

4 Discussion

Our research addresses exposure error in air pollution epidemiology which arises from spatial variation in exposure concentration as well as the descrepancy between outdoor concentration and personal exposure. The Bayesian hierarchical framework is used to model the spatial and temporal variability in human exposure to $PM_{2.5}$. This may result in better exposure assessment for health study. The statistical approaches developed in this project can be applied to other important air pollutants like carbon monoxide, sulfate, and ozone.

There are some limitations in our research. First we used estimated populations in predicting exposure. And meterological variables was only available at a single monitor, instead of for each zipcode. These may result in inaccuracy in predictions. Second, before fitting the Bayesian hierarchical model, we have coarsened some predictors levels since they have similar impact. For examples, we use weekend instead of individual day of week and, proportion of age group for 15-65 which is a very broad range group. And we did not try the non-linear term of different variables. Another limitation is may have captured additional information that we did not directly model seasonality which on predicting personal exposures.

5 References

[1] Brunekreef, Bert, and Stephen T. Holgate. "Air pollution and health." The lancet 360.9341 (2002): 1233-1242.

[2] Querol X, Alastuey A, Rodriguez S, et al. Monitoring of PM10 and PM2.5 around primary particulate anthropogenic emission sources[J]. Atmospheric Environment, 2001, 35(5): 845-858.

[3] Burnett R., et al. Association between particulate- and gas-phase components of urban air pollution and daily mortality in eight Canadian cities. Inhal Toxicol 2000: 12(Suppl 4): 1539.

[4] Franklin M, Zeka A, Schwartz J. Association between PM2. 5 and all-cause and specific-cause mortality in 27 US communities[J]. Journal of Exposure Science and Environmental Epidemiology, 2007, 17(3): 279-287.

[5] Dominici F., McDermott A., Zeger S., and Samet J. National maps of the effects of particulate matter on mortality: exploring geographical variation. Environ Health Perspect 2003: 111(1): 3944.

[6] Chauhan , A. J. and Johnston , S. L. 2003. Air pollution and infection in respiratory illness. Br. Med. Bull., 68: 95112.

[7] Milligan , P. J. M. , Brabin , B. J. , Kelly , Y. J. , Pearson , M. G. , Mahoney , G. Dunne , E. , Heaf , D. and Reid , J. 1998. Association of spatial distribution of childhood respiratory morbidity with environmental dust pollution. J. Toxicol. Environ. Health, 55: 169184.

[8] Yang , C.-Y. , Hsieh , H.-J. , Tsai , S.-S. , Wu , T.-N. and Chiu , H.-F. 2006. Correlation between air pollution and postneonatal mortality in a subtropical city: Taipei, Taiwan. J. Toxicol. Environ. Health A, 69: 20332040.

[9] Pope III, C. Arden. "Review: epidemiological basis for particulate air pollution health standards." Aerosol Science & Technology 32.1 (2000): 4-14.

[10] Hoff, Raymond M., and Sundar A. Christopher. "Remote sensing of particulate pollution from space: have we reached the promised land?." Journal of the Air & Waste Management Association 59.6 (2009): 645-675.

[11] Burke, JANET M., Maria J. Zufall, and Haluk Ozkaynak. "A population exposure model for particulate matter: case study results for PM2. 5 in Philadelphia, PA."

Journal of Exposure Analysis and Environmental Epidemiology 11.6 (2001): 470-489.

[12] Liu, Yang, Christopher Joseph Paciorek, and Petros Koutrakis. "Estimating regional spatial and temporal variability of PM2. 5 concentrations using satellite data, meteorology, and land use information." (2009).

[13] Salway, Ruth, and Jon Wakefield. "A hybrid model for reducing ecological bias." Biostatistics 9.1 (2008): 1-17.

[14] Janssen NA, Schwartz J, Zanobetti A, Suh HH (2002). Air conditioning and source-specific particles as modifiers of the effects of PM 10on hospital admission for heart and lung disease. Environmental Health Perspectives 110, 43-49.

[15] Bell ML, Ebisu K, Peng RD, Dominici F (2009). Adverse health effects of particulate air pollution: modification by air conditioning. Epidemiology 20, 682-686.

[16] Weisel CP, Zhang J, Turpin BJ, Morandi MT, Colome S, Stock TH, Spektor DM, Korn L, Winer AM, Kwon J, Meng QY, Zhang L, Harrington R, Liu W, Reff A, Lee JH, Alimokhtari S, Mohan K, Shendell D, Jones J, Farrar L, Maberti S, Fan T (2005). Relationships of indoor, outdoor, and personal air (RIOPA). Part I. collection methods and descriptive analyses. Research Report Health Effects Institute 130: 1-107.

[17] Williams, Ron, et al. "The 1998 Baltimore Particulate Matter Epidemiology-Exposure Study: part 2. Personal exposure assessment associated with an elderly study population." Journal of exposure analysis and environmental epidemiology 10.6 Pt 1 (1999): 533-543.

[18] McCurdy T, Glen G, Smith L, Lakkadi Y (2000). The national exposure research laboratorys consolidated human activity database. Jour nal of Exposure Analysis and Environmental Epidemiology 10, 566-578.

[19] Burke JM, Zufall MJ, Ozkaynak H (2001). A population exposure model for particulate matter: case studyresults for PM 2 . 5 in Philadelphia, PA. Jour nal of Exposure Analysis and Environmental Epidemiology 11 ,470-489.

[20] Zidek J, Shaddick G, White R, Meloche J, Chatfield C (2005). Using a probabilistic model (pCNEM) to estimate personal exposure to air pollution. Environmetr ics 16, 481-493.

[21] Sarnat S E, Sarnat J A, Mulholland J, et al. Application of alternative spatiotemporal metrics of ambient air pollution exposure in a time-series epidemiological study in Atlanta[J]. Journal of Exposure Science and Environmental Epidemiology, 2013,

23(6): 593-605.

[22] Burke JM, Zufall MJ, Ozkaynak H (2001). A population exposure model for particulate matter: case study results for PM 2 . 5 in Philadelphia, PA. Jour nal of Exposure Analysis and Environmental Epidemiology 11 ,470-489.

[23] Ivy D, Mulholland JA, Russel AG (2006). Development of ambient air quality population-weighted metrics for use in time-series health studies. Jour nal of the Air & Waste Management Association 58, 711-720.

[24] AERMOD: Description of Model Formulation (2004). US Environmental Protection Agency,EPA-454/R-03-004.

[25] Gelfand AE, Vounatsou P (2003). Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics 4, 11-25

[26] Shaddick G, Lee D, Zidek J V, et al. Estimating exposure response functions using ambient pollution concentrations[J]. The Annals of Applied Statistics, 2008, 2(4): 1249-1270.

6 Appendix

Figure 4: Two zipcodes mean and variance of exposure temporal trends of SHEDS data

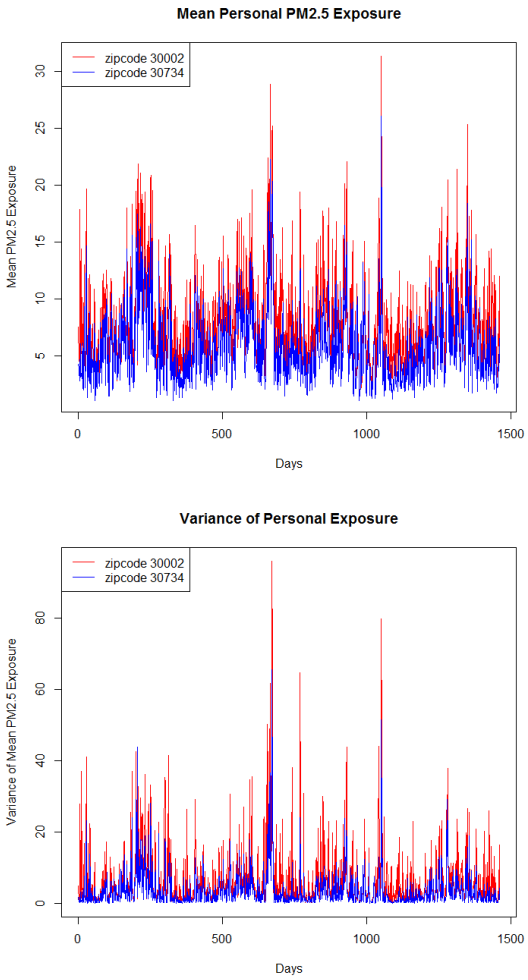


Figure 5: Daily ED counts temporal trend for second data

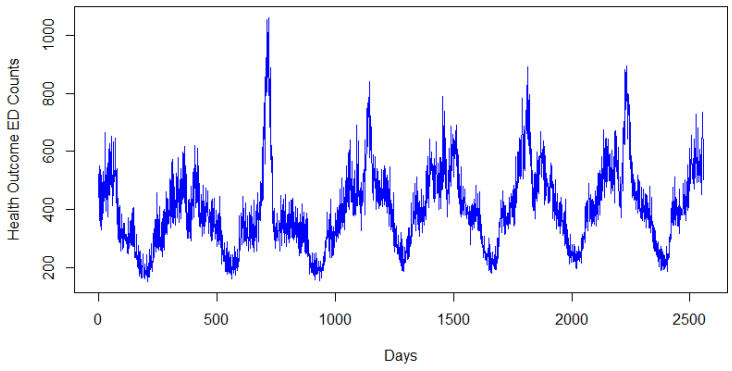


Table 9: Mean personal exposure linear model selection

	Model 1	Model 2	Model 3	Model 4	Model 5
AIC	-2040.902	-3292.225	-2043.194	-3296.289	-3347.333
(median,sd)	(176.065)	(326.208)	(175.730)	(322.413)	(256.578)
RMSE	0.808	0.525	0.837	0.526	0.504
(median,sd in original scale)	(0.152)	(0.144)	(0.144)	(0.134)	(0.106)
MAE	0.578	0.374	0.598	0.374	0.364
(median,sd in original scale)	(0.108)	(0.099)	(0.103)	(0.093)	(0.072)
R-square	0.947	0.978	0.946	0.978	0.979
(median,sd in original scale)	(0.013)	(0.010)	(0.012)	(0.001)	(0.005)
95% coverage	0.961	0.949	0.961	0.952	0.952
(median,sd)	(0.011)	(0.021)	(0.013)	(0.021)	(0.022)

Table 10: Variance of personal exposure linear model selection

	Model 1	Model 2	Model 3	Model 4	Model 5
AIC	545.759	81.304	513.499	59.698	101.876
(median,sd)	(364.314)	(417.001)	(361.785)	(410.068)	(368.606)
RMSE	1.720	1.470	1.721	1.446	1.484
(median,sd in original scale)	(0.632)	(0.634)	(0.633)	(0.630)	(0.570)
MAE	0.955	0.814	0.953	0.804	0.806
(median,sd in original scale)	(0.268)	(0.264)	(0.268)	(0.262)	(0.230)
R-square	0.901	0.927	0.900	0.926	0.925
(median,sd in original scale)	(0.050)	(0.050)	(0.049)	(0.048)	(0.033)
95% coverage	0.943	0.941	0.944	0.944	0.944
(median,sd)	(0.038)	(0.045)	(0.038)	(0.047)	(0.045)