SUPPLEMENTAL MATERIAL

EXPANDED METHODS

Study Population

This analysis is based on data from the American Cancer Society (ACS) Cancer Prevention Study II (CPS-II) cohort. This cohort has been used previously to study the health effects of air pollution.[1,2] The ACS CPS-II prospective cohort included 1,184,587 participants who were enrolled by over 77,000 volunteers between September 1982 and February 1983. Participants were largely friends and family members of the volunteers, and were recruited from all 50 states, the District of Columbia, and Puerto Rico. Participants in the ACS CPS-II cohort were at least 30 years of age and had at least one family member aged 45 years or older. The Emory University School of Medicine Human Investigations Committee provided ethics approval for the ACS CPS-II; ethics approval for the present analysis was obtained from the Ottawa Hospital Research Ethics Board.

At time of enrollment, participants completed a four-page, self-administered questionnaire providing their residential address and information on birth date, age, height and weight, and marital status. Additionally the questionnaire collected information on family history in relation to cancer, history of diseases, current physical condition, smoking history and habits, diet (including consumption of beer, wine, and hard liquor), use of medications and vitamins, occupational history and exposures, and additional miscellaneous information. Copies of the full ACS CPS-II questionnaires for both men and women are available at (http://www.cancer.org/research/researchprogramsfunding/epidemiology-cancerpreventionstudies/studyquestionnaires/index).

Vital status follow-up was conducted every two years. For the years 1984, 1986, and 1988, vital status was obtained by the study volunteers and confirmed by obtaining death certificates. Subsequent vital status follow-ups were conducted using computerized record linkage to the National Death Index.[3] In this analysis participants were followed up from the time of enrollment (between September 1982 and February 1983) through 2004. During the first six years of follow-up, cause of death was coded using a 2-digit ACS CPS-II code that was a consolidation of International Classification of Diseases, Ninth Revision (ICD-9) codes. Subsequent cause of death coding used ICD-9 or ICD-10 codes. Cause of death was captured for the underlying or primary cause of death, and the next two contributing causes of death, unless a cancer was reported later in the death certificate. In the event of a cancer reported later on a death certificate it would be captured instead of the second contributing cause of death.

Although the total ACS CPS-II cohort included nearly 1.2 million participants, approximately 385,000 individuals with invalid home address information and 130,000 individuals with missing individual-level data were excluded. The final analytic cohort used in this analysis included 669,046 participants. A total of 237,201 participants, approximately 35% of the initial study cohort, died during the 22-year follow-up period.

Geocoding of Residential Locations

In preparation for geocoding of the ACS-CPS-II cohort, Dr. Michael Jerrett, UC Berkeley, accompanied Mr. Zev Ross, ZevRoss Spatial Analysis, to Atlanta and met with ACS professional staff to ensure smooth access to the data and correct geocoding procedures. Geocoding was performed on the ACS CPS-II cohort data with a Dell Precision M65 laptop at the ACS offices in Atlanta. The laptop included ArcGIS 9.3.1 and ZP4 (Semaphore Corporation) software for use in the geocoding (expiration date of July 2010). The road data and geocoding locators were prepared in advance and TeleAtlas-based

StreetMap data from ESRI were used for all geocoding. A single composite locator made up of two, very similar street locators was created. Both individual locators used a side offset of 15 meters.[4] The default end offset of 3 m, spelling sensitivity of 80%, a minimum candidate score of 10 and a minimum match score of 60 were used and was set to assign a non-match to addresses that matched more than one address. The only difference between the two locators used was that the primary locator ("Street_Address") used 2005 TeleAtlas roads (this is the ESRI 9.3.1 Update to the Data & Maps) while the other ("Street_Addres1") used 2003 TeleAtlas roads. Each one of these road files includes about 40 million road segments. In our testing, we noticed that a small percentage of addresses (<1%) were properly geocoded with the older road network only and decided that it would be valuable to include a locator based on this older street network to capture this small percentage of additional matches.

Testing in advance of geocoding. We tested for address correction and geocoding speed and accuracy using voter registration data from three states (NY, WA and PA). Voter registration data provides ideal testing data in the sense that addresses are likely to be a little messy and it represents one of the few publicly available datasets with millions of records of real residential addresses. The drawback of voter registration data is that only a limited number of states make the data available for free and, as a result, the accuracy and speed tests will not necessarily be representative of a national dataset. From a speed perspective, voter registration data was likely to be deceptively fast because it consists of geographically close addresses reducing the amount of search time needed, in particular, the address correction software. We tested the address processing on a sample of 3 million records from the states above. In addition, we conducted a qualitative comparison of geocoding accuracy using 100 manually selected addresses in New York City for which we could identify the actual building/address on an aerial photo that has been orthographically corrected. We chose to look at NYC because of the high number of verifiable addresses, high-resolution aerial photography available for the area and because Queens County has an unusual address numbering system (hyphenated) that we wanted to test with the address correction and geocoding methodology. Overall, we found that the locators based on TeleAtlas-based StreetMap data performed extremely well on voter registration data. Match rates were high, matches appeared to be highly accurate and the speed was impressive. Both address correction and geocoding could be completed less than 2 hours each.

Address raw data. ACS professional staff provided tab-delimited text files for geocoding. These include a file for all CPS-II baseline cohort participants (all82.txt) with 1,175,991 records. The file included the variables ID1, street, city, state and zip code. The address history dataset also includes a field each for the day, month and year of date archive.

"Manual" address processing. The files were read into R statistical software (version 2.9.0) and reviewed using regular expressions scripting to identify all characters that were not numbers or letters. Based on this review, several characters were identified for removal in advance of address correction (these include *, %, ", \, _, /,), (, ^, `, $, /). A unique sequential ID (called zID2) was added because the ID1 is not unique in the address history file. Finally, all the addresses with "C/O" at the beginning were identified. These likely represent "Care Of" and would not be a home address. The "C/O" was stripped from the address and a TRUE/FALSE field to identify these later if necessary was appended.

Address correction. ZP4 software from Semaphore Corporation was used to conduct address correction. We used the April 2010 release (expiration date of July 2010) and included the add-on route-to-street conversion data (called "LACS" by Semaphore) to convert rural-style addresses to street-style addresses for improved geocoding accuracy. The program was run in batch mode using the GUI and a CSV input file. The address correction took more than 10 hours and was allowed to run overnight. The resulting address corrected file was imported into R and stripped of non-essential fields for geocoding.

Geocoding performance. As mentioned above, the geocoding was performed using a composite locator based on TeleAtlas street data available in ArcGIS Data & Maps. The geocoding was executed using a Python script run from the Command Prompt. Geocoding the ACS data took somewhat longer than our test data (approximately 2.5 hrs). To get a sense of which addresses should have been geocodeable, the geocodeable addresses were identified using the criteria below. In all cases, these are case insensitive. This is an approximate estimate of an address's ability to be geocoded, some of these might actually be geocodeable and other addresses might not be geocodeable. A non-geocodeable address was defined as an address with one or more of the following:

1. Street address includes the word "box"
2. No letters in address
3. No numbers in address
4. An address beginning with "RR" followed by whitespace followed by a number.
5. An address with "RT" followed by whitespace followed by a number.
6. An address with no zip AND no city
7. An address with no zip AND no state

None of these were excluded from the geocoding but were identified to assess match rates. In total 84% for all 82 addresses were geocodeable. Among geocodeable addresses at least 89% of addresses were geocoded at a score of 80 or above. Positional match rates are assigned a score from 0 to 100. The score is generated from various address elements. Scores will be lower if the candidate address contains misspellings, incorrect information, addresses outside the range on a street segment, or missing elements of the address. Detailed information on match rates is presented elsewhere.[5]

Exposure estimates

Exposure model estimations. Exposure to $PM_{2.5}$ was estimated by using the geocoded home addresses of the study participants and linking them to ambient $PM_{2.5}$ concentrations derived using a national-level hybrid land use regression (LUR) and Bayesian Maximum Entropy (BME) interpolation model (LUR-BME) more fully documented elsewhere.[6] Monthly $PM_{2.5}$ data for a total of 1,464 monitoring sites from 1999 to 2008 were used to estimate the LUR-BME model in two stages. A training data set of 1329 monitors was used to select the variables, and approximately 10% or 135 monitors were retained for cross-validation. In the first stage, the model was fit with a deterministic LUR with monthly pollution averages as the dependent variable and land use information as predictors. The LUR model used a deletion/substitution/addition algorithm with v-fold cross-validation to select predictor variables. This method reduces the chance of over fitting by continuously predicting on leave-out folds, meaning selected variables minimize the mean square error of predictions on data that are not used to fit the model. Based on this method two variables were selected: traffic-weighted roads within 1000m of a monitor (based on modeled traffic counts) and the cube of percentage of green space within a 100m buffer around the monitor. In the second stage, a BME kriging interpolation model was used to capture the residual spatiotemporal variation in $PM_{2.5}$ concentrations not predicted in the first stage using LUR. Monthly values from 1999 to 2004 were averaged and assigned to study participants using their geocoded addresses. The estimated overall mean $PM_{2.5}$ exposure concentration was 12.6 (SD=2.9) $\mu g/m^3$, with a range from 1 to 28 $\mu g/m^3$.

Cross-validation. The randomly selected cross-validation data of 135 monitors that were not used in the variable selection or model prediction were used to evaluate how well the model predicted at locations not used to calibrate the model. Supplemental Figure 1 compares cross-validation predictions of the LUR-BME plotted against the observed monthly data. This analysis showed good agreement between observed data and LUR-BME predictions with no significant bias or outliers. The model R^2 is 0.79, which suggests the model predicts well at locations different from those used to calibrate the model.

Supplemental Figure 2 shows a map of the LUR-BME models averaged over the entire study period. The patterns present have spatial patterns consistent with well-known regional patterns of $PM_{2.5}$ pollution, with areas of Central and Southern California having the highest levels, followed by parts of the Industrial Midwest and the Southeast. Supplemental Figure 3 shows a zoom in of the Los Angeles Metropolitan Area. Here areas with large highways are clearly visible indicating the ability of the model to predict small-area variation in $PM_{2.5}$.

Statistical Analysis

Adjusted mortality hazard ratios (HRs) were estimated using the Cox proportional hazards regression models. Follow-up time in days since enrollment was used as the time axis. Survival times of those still alive at end of follow-up were censored and, in analyses of cause-specific mortality, if death occurred for another cause, survival times were censored at the time of death. All models were stratified by one-year age categories, sex, and race (white, black, other), allowing each category to have its own baseline hazard. The age, sex, and race information was taken directly from the ACS CPS-II enrollment questionnaire. The models included the estimated $PM_{2.5}$ exposure concentration as a continuous variable, based on the LUR-BME exposure model described above.

The models also controlled for multiple individual-level covariates as follows. Tobacco smoke covariates included 13 variables that characterized current and former smoking habits (including smoking status of never, former, or current smoker, linear and squared terms for years smoked and cigarettes smoked per day, indicator for starting smoking at younger than 18 years of age, and pipe/cigar smoker) and one continuous variable that assessed exposure to second hand cigarette smoke (hours/day exposed). All of the tobacco smoke related covariates were obtained directly from self-reported information collected from the ACS CPS-II enrollment questionnaire. Covariates that controlled for occupational exposures included a variable that indicated regular occupational exposure to dust and fumes (including asbestos, chemical/acids/solvents, coal or stone dusts, coal tar/pitch/asphalt, diesel engine exhaust, or formaldehyde) as self-reported on the ACS CPS-II enrollment questionnaire and seven variables that reflected workplace exposure in each subject's main lifetime occupation. The seven workplace exposure variables were indicator variables that indicated rankings of occupational exposures derived from occupational information from the ACS CPS-II enrollment questionnaire and documented elsewhere.[7,8] Variables that represented marital status (separated/divorced/widowed or single versus married), variables that characterized levels of education (high school, more than high school versus less than high school), two body mass index (BMI) variables (linear and squared terms for BMI), and variables characterizing consumption of alcohol (beer, missing beer, wine, missing wine, liquor, missing liquor) were also included as covariates in the models and were obtained directly from self-reported information collected from the ACS CPS-II enrollment questionnaire. Dietary covariates included indicator variables for quartile ranges of a dietary fat index and quartile ranges of a dietary vegetable/fruit/fiber index. These diet indices were derived based on diet information provided in the ACS CPS-II enrollment questionnaire as documented elsewhere.[9]

To evaluate the sensitivity of the results to control for geographical, social, economic, and environmental settings (contextual conditions), some of the Cox Proportional Hazards models also included ecologic covariates obtained from the 1990 Census of Population Long-Form for the subjects' residential zip code area.[10] These contextual variables included: median household income; percentage of people with < 125% of poverty-level income; percentage of unemployed persons over the age of 16 years; percentage of adults with less than 12^{th} grade education; and percentage of the population who were Black or Hispanic. These ecological covariates were included in the models using both zip-code level data as well as zip-code deviations from the county means.

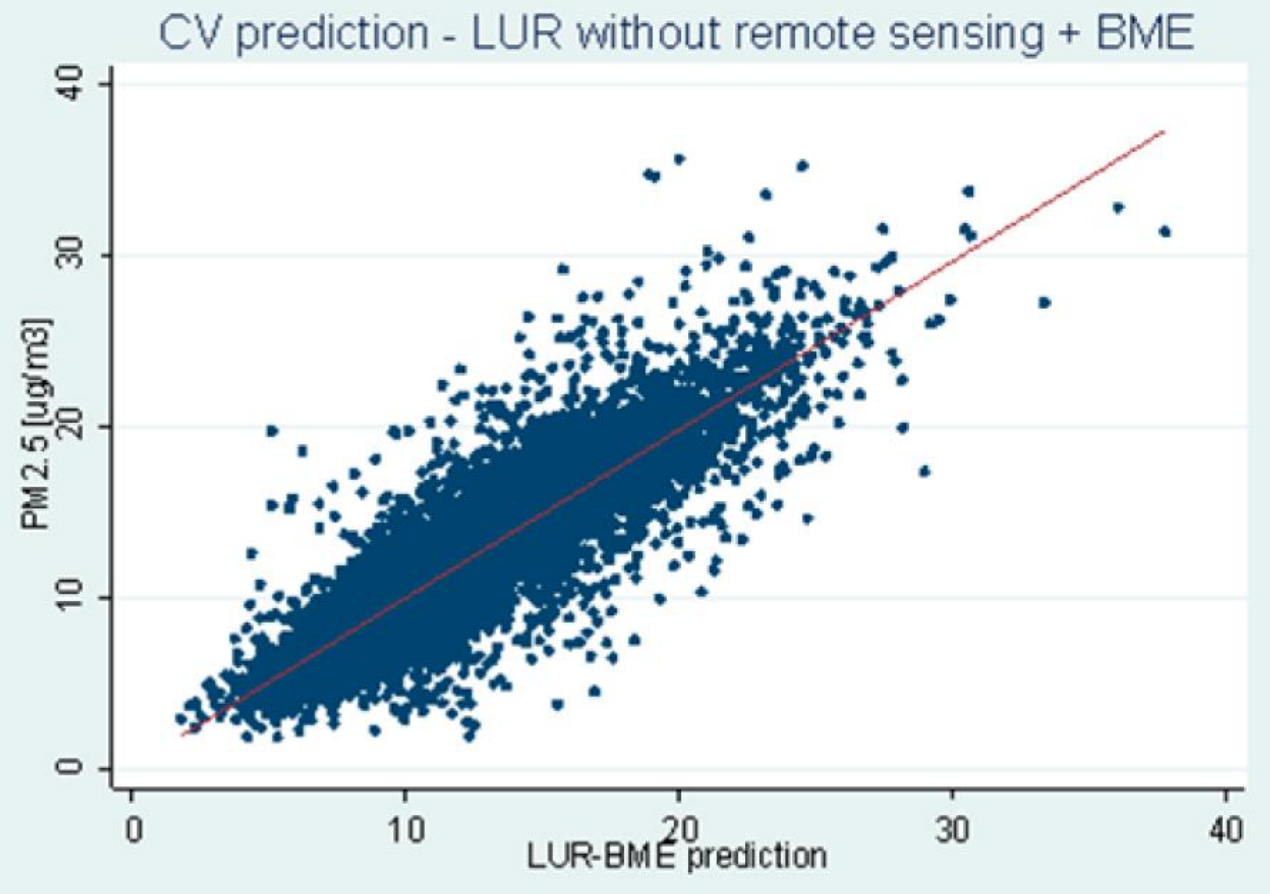
Baseline HRs associated with an increment of 10 $\mu g/m^3$ of $PM_{2.5}$ were estimated for all-cause, CVD, hypertension and diabetes mortality. Two approaches were used to evaluate effect modification of cardiometabolic risk factors at time of enrollment on the $PM_{2.5}$-CVD mortality association. First, adjusted HRs (and 95% CIs) for CVD mortality were estimated in relation to three key categorical indicators of cardiometabolic risk (diabetes, doctor diagnosed high blood pressure, and heart disease at time of enrollment) and categorically high and low $PM_{2.5}$ concentrations (> 75^{th} percentile and < 25 percentile). To formally test for additive interactions between $PM_{2.5}$ exposure and key cardiometabolic risk factors, the relative excess risk due to interaction, the attributable proportion due to interaction, and the synergy index were calculated using the MOVER method for the analysis of 4 x 2 tables as documented elsewhere.[11]

The second approach to evaluate effect modification of cardiometabolic risk factor estimated adjusted HRs associated with increases in $PM_{2.5}$ (using $PM_{2.5}$ as a continuous variable) for cardiovascular mortality, while stratifying by all cardiometabolic risk factors that were available based on information from the ACS CPS-II enrollment questionnaire. These risk factors include: BMI levels, doctor diagnosed high blood pressure, heart disease, and diabetes, exercise levels; vegetable/fruit/fiber and fat intake; and use of medications including aspirin, heart medications, and diuretics. Specifically, different ranges of BMI (<25, 25-30, 30-35, 35+) were determined using height and weight information provided on the enrollment questionnaire. Doctor diagnosed blood pressure, heart disease, and diabetes along with exercise levels and medication use were all self-reported on the enrollment questionnaire. Quartile ranges of a dietary fat index and quartile ranges of a dietary vegetable/fruit/fiber index were derived based on diet information provided in enrollment questionnaire as noted above and more fully documented elsewhere (Chao et al. 2000). Because the likelihood of any individual in the cohort having any of the key risk factors at enrollment depends partially on age and smoking status at enrollment, indicators of cardiometabolic risk were cross-stratified with four age-at-enrollment and smoking status strata (never smokers, age < 60; never smokers, age >= 60; ever smokers, age < 60; ever smokers, age >= 60). In order to evaluate if the associations differ for different follow-up times, we conducted the analysis, stratified across strata relating to cardiometabolic risk factors for three different follow-up periods: 0-7, 7-14, and 14-22 years.

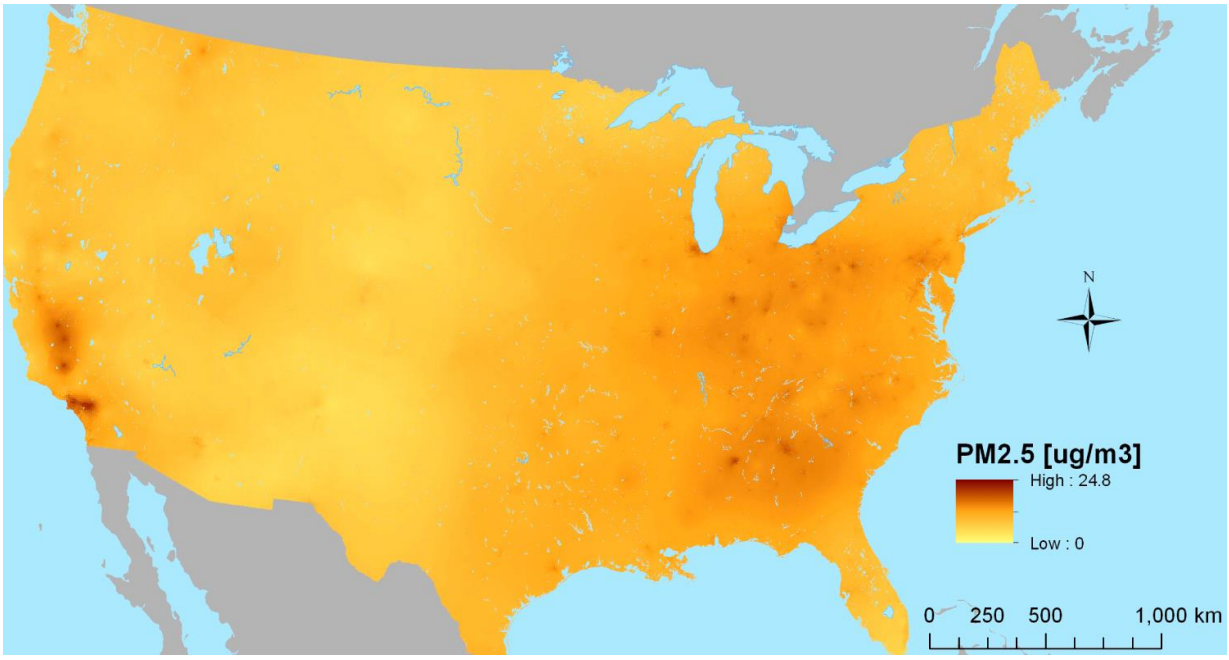
Specific cause-of-death analyses were conducted using primary and contributing cause-of-death information provided on the death certificate focusing on hypertensive disease, diabetes and interactions with other cardiovascular causes of death.

Supplemental References

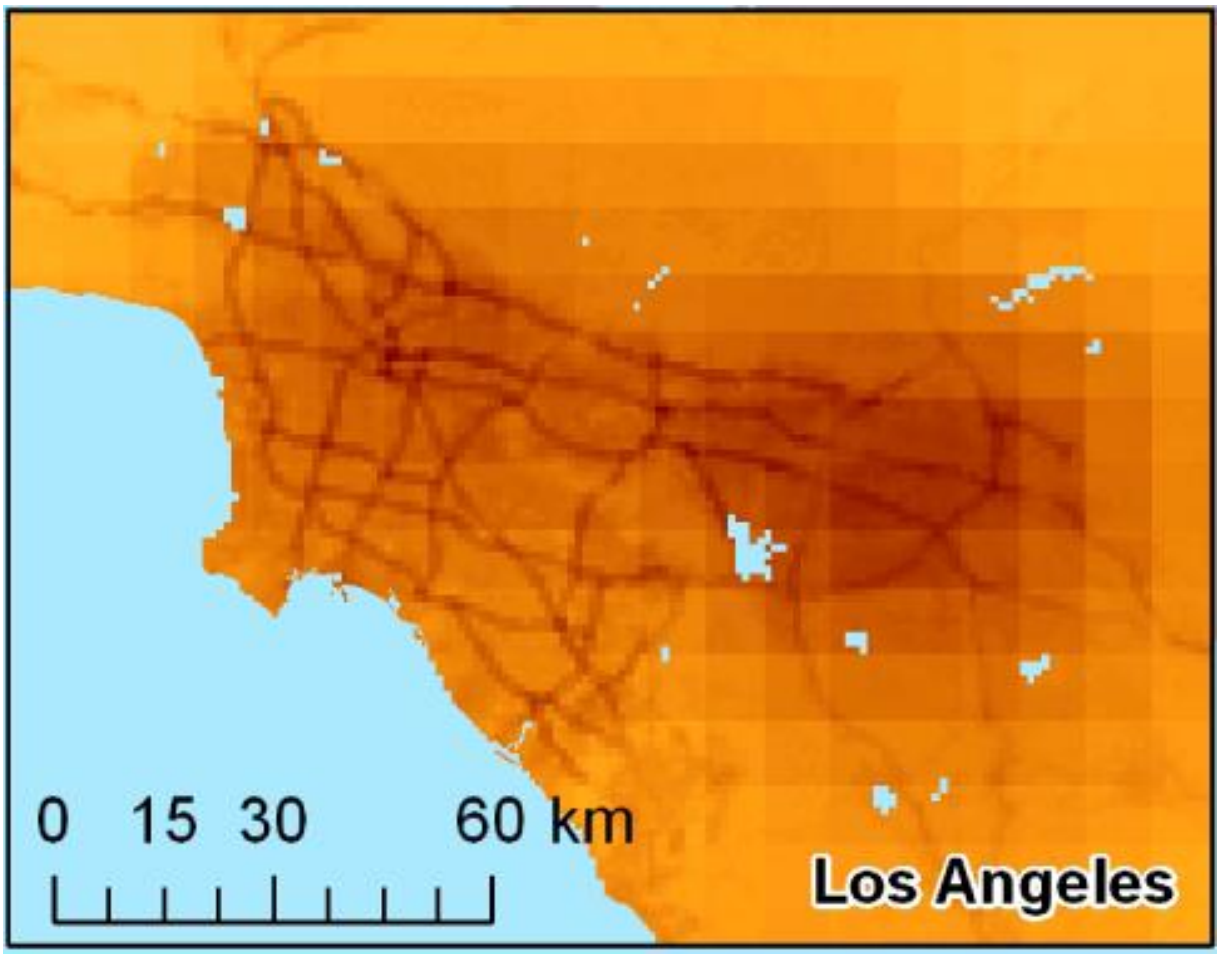
1. Pope CA III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA* 2002;287:1132-1141.
2. Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, Turner MC, Pope CA III, Thurston G, Calle EE, Thun MJ. Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. HEI Research Report 140, *Health Effects Institute*, Boston, MA. 2009; 140:5-114, discussion 115-136.
3. Calle EE, Terrell DD. Utility of the National Death Index for ascertainment of mortality among Cancer Prevention Study II participants. *Am J Epidemiol*. 1993;137:235-241.
4. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* 2003;2:10 doi:10.1186/1476-072X-2-10.
5. Goldberg DW. Improving geocoding match rates with spatially-varying block metrics. *Transactions in GIS* 2011;15:829-850.
6. Beckerman BS, Jerrett M, Serre M, Martin RV, Lee SJ, van Donkelaar A, Ross Z, Su J, Burnett RT. A hybrid approach to estimating national scale spatiotemporal variability of $PM_{2.5}$ in the contiguous United States. *Environ Sci Technol.* 2013;47:7233-7241.
7. Siemiatycki J, Nadon L, Lakhani R, Beegin D, Geerin M. Exposure assessment. In: Siemiatycki J, ed. *Risk Factors for Cancer in the Workplace*. Baton Rouge, La: CRC Press; 1991:45-114.
8. Krewski D, Burnett RT, Goldberg MS, et al. Reanalysis of the Harvard Six Cities Study and the American Cancer society Study of Particulate Air Pollution and Mortality. Special Report. Cambridge, Mass: Health Effects Institute; 2000.
9. Chao A, Thun MJ, Jacobs EJ, Henely SJ, Rodriquez C, Calle EE. Cigarette smoking and colorectal cancer mortality in the Cancer Prevention Study II. J Natl Cancer Inst. 2000;92:1888-1896.
10. U.S. Department of Commerce, Bureau of the Census. Census of population and housing, 1990 (United States): Summary tape file 3B. ICPSR version. Washington, DC; 1993.
11. Zou GY. On the estimation of additive interaction by use of the four-by-two table and beyond. *Am J Epidemiol* 2008;168(2):212-224.



Supplemental Figure I. Observed on Predicted Plot from $PM_{2.5}$ levels based on the LUR-BME Model; Data from 135 randomly selected cross-validation sites with monthly averages. Adapted from Beckerman et al. 2012.[6]



Supplemental Figure II: Coterminous United States with Predicted $PM_{2.5}$ levels based on the LUR-BME Model. Adapted from Beckerman et al. 2012.[6]



Supplemental Figure III: Los Angeles Metropolitan Area with Predicted $PM_{2.5}$ levels based on the LUR-BME Model. Adapted from Beckerman et al. 2012.[6]