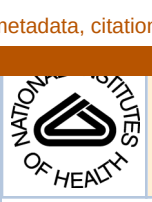

Published in final edited form as:

J Expo Sci Environ Epidemiol. 2012 September ; 22(5): 496–501. doi:10.1038/jes.2012.57.

The moving-window Bayesian Maximum Entropy framework: Estimation of $PM_{2.5}$ yearly average concentration across the contiguous United States

Yasuyuki Akita, Ph.D.[1], **Jiu-Chiuan Chen, M.D., Sc.D.**[2], and **Marc L. Serre, Ph.D.**[1,*]

[1]Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina at Chapel Hill

[2]Division of Environmental Health, Department of Preventive Medicine, University of Southern California Keck School of Medicine

Abstract

Geostatistical methods are widely used in estimating long-term exposures for air pollution epidemiological studies, despite their limited capabilities to handle spatial non-stationarity over large geographic domains and uncertainty associated with missing monitoring data. We developed a moving-window (MW) Bayesian Maximum Entropy (BME) method and applied this framework to estimate fine particulate matter ($PM_{2.5}$) yearly average concentrations over the contiguous U.S. The MW approach accounts for the spatial non-stationarity, while the BME method rigorously processes the uncertainty associated with data missingnees in the air monitoring system. In the cross-validation analyses conducted on a set of randomly selected complete $PM_{2.5}$ data in 2003 and on simulated data with different degrees of missing data, we demonstrate that the MW approach alone leads to at least 17.8% reduction in mean square error (MSE) in estimating the yearly $PM_{2.5}$. Moreover, the MWBME method further reduces the MSE by 8.4% to 43.7% with the proportion of incomplete data increased from 18.3% to 82.0%. The MWBME approach leads to significant reductions in estimation error and thus is recommended for epidemiological studies investigating the effect of long-term exposure to $PM_{2.5}$ across large geographical domains with expected spatial non-stationarity.

Keywords

long-term exposure; geostatistics; moving-window; Bayesian Maximum Entropy; $PM_{2.5}$

Introduction

Several epidemiological studies have demonstrated that long-term exposure to fine-particulate matter ($PM_{2.5}$) is associated with increased morbidity and mortality (Boldo et al., 2006; Pope et al., 2009). In most epidemiological studies, long-term concentrations are estimated by conventional distance-based approaches, such as aggregating the air pollution data from the nearest monitoring station or taking the areal average concentrations in the county, census block, or zip code area where the study participants reside. These methods implicitly assume a uniform concentration within the defined spatial unit surrounding the study subject, without being able to capture the local-scale concentration gradient. Recent

[*]**Corresponding author** Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599-7431 marc_serre@unc.edu Phone: (919) 966 7014, Fax: (919) 966 7911 .

studies have attempted to account for the local-scale spatial variability of ambient concentrations by applying geostatistical methods (Liao et al., 2006) or land use regression (LUR) models (Hoek et al., 2008). Geostatistical techniques, in particular, have been widely used in air pollution epidemiologic studies. For instance, a stronger association between long-term exposure to PM 2.5 and chronic health effects relative to previous studies was found by estimating within-city exposure using a kriging geostatistical approach over the Los Angeles metropolitan area (Jerrett et al. 2005).

Although geostatistical methods represent a major improvement over the conventional distance-based approaches, other methodological issues arise when these geostatistical methods are applied to estimate the long-term $PM_{2.5}$ concentration in epidemiologic studies with a large geographic coverage. One major concern is that the spatial correlation pattern of ambient concentrations is assumed to be stationary across the entire spatial domain (Liao et al., 2009; Zhang et al., 2009). In other words, in these studies, spatial autocorrelation of the concentration is assumed to be unchanged across the entire study domain, and a single semivariogram model obtained from all the observations is used for the estimation across the domain. In a national-scale study on ambient air pollution, however, spatial correlation patterns are expected to vary across regions. For example, in the U.S., $PM_{2.5}$ concentrations show high spatial variability. In California higher concentrations were clustered in small areas whereas higher concentrations were more widely spread on the east coast. In addition, the level is generally low in the central and northwestern U.S. (Bell et al., 2007). Thus, in order to estimate the national-scale long-term $PM_{2.5}$ concentration, a framework that accounts for the non-stationarity of spatial variability is needed.

Another issue pertaining to the estimation of the long-term $PM_{2.5}$ concentration is the data completeness criterion used to define the long-term ambient concentration at a given monitoring station. The long-term $PM_{2.5}$ concentration is generally approximated by taking the average of daily $PM_{2.5}$ concentrations observed over some time period of exposure (e.g., one year average), only if there are enough daily measurements (e.g., 75% of intended samples) to represent the long-term exposure within the time period of interest (Miller et al., 2007). All average concentrations not satisfying the data completeness criteria, although potentially informative for understanding the spatiotemporal process of air pollution exposures, are simply discarded from the subsequent analysis, due to the lack of appropriate methodological framework to handle the uncertainty associated with yearly averages calculated from an incomplete set of daily concentrations.

The overall goal of this study is, therefore, to estimate the national-scale long-term $PM_{2.5}$ concentration that addresses all of the aforementioned issues. We achieve this goal by estimating $PM_{2.5}$ yearly average concentrations over the contiguous U.S. using a moving-window (MW) implementation of a geostatistical estimation framework based on the Bayesian Maximum Entropy (BME) method. The MW approach provides an efficient framework to account for the non-stationarity of spatial processes (Haas, 1990), while the BME method (Christakos, 2000; Yu et al., 2009) processes the uncertainty of the $PM_{2.5}$ yearly average concentration due to the incompleteness of $PM_{2.5}$ daily concentrations. In order to evaluate model performance, a cross-validation analysis was conducted to compare the domain wide stationary kriging (SK) method with the proposed moving-window kriging (MWK) and moving-window BME (MWBME) methods.

Materials and Methods

$PM_{2.5}$ Monitoring Data

$PM_{2.5}$ daily concentrations measured from 2002 to 2003 were obtained from the Air Quality System (AQS) maintained by the U.S. Environmental Protection Agency (EPA) (US EPA,

2009). The daily concentrations exceeding the federal maximum sample value (500$\mu g/m^3$) were regarded as outliers and removed from the data (US EPA, 2008). If multiple monitors were operated at the same monitoring site on the same day, the resulting co-located daily concentrations were averaged. In total, 297,297 daily concentrations were observed at 1177 monitoring sites during the period. The mean (±SD) of the daily concentrations is 12.44±8.29 $\mu g/m^3$. The daily concentrations were then used to construct the $PM_{2.5}$ yearly average concentration in 2003.

The Bayesian Maximum Entropy Method

The BME method introduced by Christakos (Christakos, 1990; Christakos, 2000) provides a mathematically rigorous framework that integrates a variety of available knowledge bases with data having varying levels of epistemic uncertainty. These data are categorized in hard data corresponding to exact measurements, and soft data having an uncertainty characterized by a probability density function (PDF) of any type. A full description of the BME method can be found elsewhere (Christakos et al., 2001; Serre and Christakos, 1999). In brief the BME method can be viewed as a two-stage knowledge processing procedure: At the prior stage, maximum entropy theory is used to process the general knowledge base at hand and produce a prior PDF describing the spatial process. Then at the posterior stage, an operational Bayesian conditionalization rule is used to update this prior PDF with respect to the site specific hard and soft data available, which produces a BME posterior PDF describing the value of the spatial process at any estimation point of interest.

Let $Z(\mathbf{s})$ be a spatial random field (SRF) representing the $PM_{2.5}$ yearly average concentration at some spatial location $\mathbf{s}$. We will denote as Z_k the random variable representing the SRF at estimation point $\mathbf{s}_k$ (i.e., $Z_k=Z(\mathbf{s}_k)$), and similarly Z_h and Z_s are vectors of random variables representing the SRF at the hard data points $\{\mathbf{s}_h\}$ and the soft data points $\{\mathbf{s}_s\}$, respectively. By convention, lower case variables (e.g. z_h, z_s, or z_k) will denote realizations or deterministic values taken by their corresponding upper case random variables (e.g. Z_h, Z_s or Z_k) In the case that the general knowledge base $\boldsymbol{G}$ about the SRF $Z(\mathbf{s})$ consists in its mean trend $m_Z(\mathbf{s})=E[Z(\mathbf{s})]$ and covariance function $c_Z(\mathbf{s},\mathbf{s}')$, then the BME fundamental equation reduces to

$$f_K(Z_K) = A^{-1} \int dz_s f_G(z_h, z_s, z_k) f_{S(z_s)}$$

where A is a normalization constant, the prior PDF f_G obtained from entropy maximization on $\boldsymbol{G}=\{m_Z(.),c_Z(.)\}$ is multivariate normal with mean and covariance given by $m_Z(.)$ and $c_Z(.)$, respectively, the vector of deterministic values $\mathbf{z}_h$ corresponds to the hard data, and f_S is a PDF characterizing the epistemic uncertainty of the soft data. The BME posterior PDF is denoted with a subscript $\boldsymbol{K}=\boldsymbol{G}\cup\boldsymbol{S}$ representing the union of the general knowledge $\boldsymbol{G}=\{m_Z(.),c_Z(.)\}$ and site specific knowledge $\boldsymbol{S}=\{z_h, f_S(.)\}$. The expected value of the BME posterior PDF provides an estimate of the $PM_{2.5}$ yearly average concentration at the estimation point, and the corresponding BME posterior standard deviation provides a useful characterization of the associated estimation uncertainty. In the limiting case where only hard data are included in the estimation process, the BME estimator is simply the kriging estimator. This makes BME a consistent extension of the widely used kriging estimator when one needs to integrate non-Gaussian soft data, as is the case in this work.

The $PM_{2.5}$ yearly average concentration data

We defined the $PM_{2.5}$ yearly average concentration at any date t in 2003 as the average of the $PM_{2.5}$ daily concentrations over the 365 days preceding date t. However, an exact $PM_{2.5}$ yearly average concentration is rarely obtained, since at most of the monitoring sites the

$PM_{2.5}$ daily concentrations were collected on a three-day cycle during the study period. In most epidemiological studies the $PM_{2.5}$ yearly average concentrations satisfying some acceptable data completeness criterion are treated as the exact yearly average concentration. In this study, we used the completeness criterion that there must be more than 75% of intended measurements in each quarters of the year prior to *t*. If the completeness criterion was satisfied, the hard datum for the $PM_{2.5}$ yearly average concentration at monitoring site *i* and date *t* is simply defined as the mean of the daily concentrations observed over one year preceding date *t*.

If the completeness criterion was not met, then the $PM_{2.5}$ yearly average concentration was treated as a soft data if there were at least one measurement in each quarter. In the BME method, the epistemic uncertainty associated with incomplete $PM_{2.5}$ daily concentrations is characterized by a PDF. In this work, we assume that an adequate approximation for the PDF at monitoring station *i* and date *t* is a normal distribution with the mean $\mu_{s,i}$ and the standard deviation $\sigma_{s,i}$ truncated below zero, since concentrations cannot be negative. The mean $\mu_{s,i}$ is simply set to the sample mean of the n_i daily concentrations measured at station *i* over one year preceding date *t*. The epistemic uncertainty associated with this soft datum arises from the difference between the true mean of all 365 daily concentrations, and the sample mean $\mu_{s,i}$ calculated from an incomplete sample of size n_i selected from a finite population of size 365. Therefore, a reasonable value for the standard deviation $\sigma_{s,i}$ is

$$\sigma_{s,i} = \sqrt{\frac{\sum_{j=1}^{n_i} \left(y_{i,j} - \mu_{s,i}\right)^2 / (n_i - 1)}{n_i}} \times \sqrt{\frac{365 - n_i}{365}}$$

where the first term of this equation is the standard deviation of the sample mean and the second term is a finite population correction factor to account for the finite population size.

The Moving-window Approach

The MW approach described by Haas (1990) accounts for the non-stationarity of a spatial process over a large geographic domain by localizing the estimation procedure to small estimation neighborhoods where the spatial process can be assumed stationary within the neighborhood. Our implementation of the MW approach consists in calculating a semivariogram at each estimation point of interest using only the data points within the spatial neighborhood around that estimation point. Then the geostatistical analysis for that estimation point is conducted using the location-specific semivariogram and the data within the neighborhood.

The size of the window has to be small enough to assure stationarity of the spatial process within the window, but also large enough so that it contains enough data points to model a reliable semivariogram. In this study, we used a window containing 50 monitoring sites, based on the minimum sample size expected to produce a stable sample semivariogram estimate (Olea, 2006). The sample semivariogram was then used to fit a negative definite semivariogram model using an automated weighted least square procedure. In this study, the following three parametric semivariogram models were tested: (1) exponential, (2) Gaussian, and (3) spherical model.

Cross-validation analysis and Estimation Maps

In order to evaluate model performance of the SK, MWK and MWBME methods, a leave-one-out cross-validation analysis was conducted. We randomly selected 30 dates in 2003. We removed one at a time each observed yearly average concentration that met the

completeness criterion for these 30 dates (n=24544), and re-estimated that value based on its neighboring hard and soft data. The differences between observed and re-estimated values are the n cross validation errors, from which cross validation statistics can be calculated. The SK method assumes nationwide stationarity and therefore uses, for a given estimation date t, a single semivariogram throughout the U.S. On the contrary, in MWK and MWBME the semivariogram is calculated at each estimation point using only the data within the corresponding estimation window. In SK and MWK, only the hard data points were considered for the estimation, whereas in MWBME both hard and soft data were used for the estimation.

Model performance was evaluated using the following cross-validation statistics: mean square estimation error (MSE), mean estimation error (ME), mean standardized estimation error (MS), root mean square standardized error (RMSS), and mean of the root of estimation error variance (MR). The estimation error is the difference between estimated and observed values, and the standardized estimation error is equal to the estimation error divided by the square root of the estimation error variance. The MSE, ME, and MS should ideally be as close to zero as possible. The RMSS measures the standard deviation of standardized errors and should ideally be equal to one. The MR should ideally be as small as possible. In addition, the Pearson correlation coefficient and Spearman's rank correlation were calculated to evaluate the linear correlation and rank order of the estimated and observed yearly concentrations. Furthermore, in order to visually compare the estimation result, maps of the estimated $PM_{2.5}$ yearly average concentration were produced over California based on the three aforementioned methods.

Simulation

In this study only a small fraction (~18.3%) of $PM_{2.5}$ yearly average concentrations did not meet the completeness criterion over the study period which leads to a small ratio of soft to hard data points. In order to explore the performance of the aforementioned estimation methods under the situation where there are more frequent missing data, four simulated $PM_{2.5}$ daily concentration data sets were constructed by randomly removing 5%, 10%, 15%, and 20 % of daily concentrations from the original data set. Using these realistic simulated data sets, the hard and soft data for $PM_{2.5}$ yearly concentrations were re-constructed, which resulted in a substantially larger fraction of soft to hard data points. Finally, these simulated yearly average concentrations were used to re-run the cross-validation analysis to evaluate model performance.

All analyses were conducted using Matlab R2010a (MathWorks Inc., Natick, MA, USA) and BMElib version 2.0b (Christakos et al., 2001; Serre and Christakos, 1999).

Result and Discussion

The $PM_{2.5}$ yearly average concentrations on December 31, 2003, which uses all the $PM_{2.5}$ daily measurements observed during 2003, are shown in Figure 1 (a). The yearly average concentrations that met the completeness criterion are shown with circles. They were treated as exact measurements (hard data) in the estimation process. By contrast, those shown with triangles did not meet the completeness criterion and were discarded by the conventional SK and MWK methods as unreliable yearly average concentrations. In the MWBME analysis, however, those were treated as soft data and used in the estimation process. Figure 1 (b) shows time series of the $PM_{2.5}$ daily and corresponding yearly average concentrations in 2003 at the monitoring site 06-079-2002. The solid line shows the $PM_{2.5}$ yearly average concentrations that met the completeness criterion, whereas the dotted lines show the yearly average concentration which did not meet the criterion and the corresponding 95% upper and lower confidence bounds. At this monitoring site, the $PM_{2.5}$ yearly average

concentrations were treated as soft data from January to October because of the missing data in September and October 2002.

There are obvious geographical patterns in the spatial distribution of $PM_{2.5}$ yearly average concentration across the contiguous U.S in 2003. Figure 1 (a) reveals that $PM_{2.5}$ yearly average concentrations were generally high in the Midwestern U.S and low in the Central U.S. In addition, the concentration changed drastically over California, where high concentrations were confined in Los Angeles and the Central Valley region. This result indicates that the variability of the concentration is expected to be higher in California than in the Midwestern and Central U.S. and consequently that assuming stationarity across the entire contiguous U.S. is inappropriate. The spatial distribution of the semivariogram parameters calculated in the MW methods can be used to evaluate the geographical change in spatial variability (See Supplemental Material, p. 2). Analysis of these plots in this work indicated that there are clear variation in these parameters (See Supplemental Material, Figure 1), indicating that the total sill is relatively high in California, whereas the value is low in the Midwestern and Central U.S., which is expected from Figure 1 (a). The semivariogram range is short in the Midwestern U.S. and California, whereas the range is generally long in the Central U.S. These results confirm the non-stationarity of the spatial variability of concentrations, and imply that the MW approach is an appropriate choice when estimating the $PM_{2.5}$ yearly average concentration over the contiguous U.S.

Table 1 shows the cross-validation statistics obtained by SK, MWK, and MWBME based on the exponential semivariogram model. MWK reduced the MSE by 17.8% relative to the SK. This indicates that using the MW approach to account for the non-stationarity of spatial variability leads to a 17.8% improvement in estimation performance over a method that assumes a country wide semivariogram. The MWBME further reduced the MSE by 8.4% relative to MWK, which indicates that there was a cumulative improvement in estimation performance when using the MW approach and accounting for the soft data. The ME and MS statistics were slightly higher for MWK than for SK and MWBME, but generally all values were close to 0, indicating that all estimation methods are reasonably unbiased and therefore bias played little role in model performance. The RMSS were generally slightly greater than one, indicating that each method reports an estimation error standard deviation that is bias low. Similarly to the MSE, the MR for MWBME is the smallest. Indeed, MWBME reduced the MR by 10% relative to that of SK. The Pearson's correlation and Spearman's rank correlation obtained with MWBME were also the highest among all methods. Hence overall, MWBME performs the best among all three methods. The cross-validation statistics based on the other semivariogram models are listed in Supplemental Material (See Supplemental Material, Table 1 and 2). In terms of MSE, the exponential semivariogram model outperformed the other two semivariogram models.

In order to furthermore understand which method produced better estimation error, we show in Figure 2 the spatial distribution of the absolute value of the standardized estimation errors on December 31, 2003 in the contiguous U.S. and in California. These standardized errors should be normally distributed across the U.S. However, the maps obtained with SK (Figures 2a and 2b) show clear spatial clustering, with for example high standardized errors seen in California. On the other hand MWBME results in standardized errors with reduced spatial clustering (Figures 2c and 2d). Thus, by accounting for the geographical changes in spatial variability, the MW approach results in a better assessment of the uncertainty associated with the estimation of $PM_{2.5}$ yearly average concentrations.

Figure 3 shows maps of the estimated $PM_{2.5}$ yearly average concentration in California on December 31st, 2003 obtained by (a) SK, (b) MWK, and (c) MWBME. The circles show hard data points, whereas triangles indicate soft data points used in MWBME. The estimated

concentration map created by SK uses a nationwide semivariogram model with a range of 13,117Km. On the other hand MWK and MWBME use ranges that vary between 177Km to 1344Km across California (See Supplemental Material, Figure 1), which better captures the high spatial variability of $PM_{2.5}$ in California. As a result SK produces map displaying a smooth spatial distribution of concentration across California, while the MW approach produce maps with more spatial variability. These maps illustrate how the MW approach allows to capture strong spatial gradients of the concentration in areas of the country that are characterized by high local-scale spatial variability. Finally, the MWBME map describes further spatial details by accounting for the additional information provided by the soft data shown as triangles. In addition, Supplemental Material, Figure 3 shows the distribution of the estimation error variance over the U.S.

MSEs obtained with the MWK and MWBME methods using the true and simulated datasets are shown in Table2. The fraction of soft data points increased from 18.3% (True Data) to 82.0%, when 20% of the daily measurements were removed (Simulation Data 4). The MSE obtained with MWK, which relied only on hard data, increased as the fraction of the soft data increased, while that obtained with MWBME, which processes both hard and soft data, remained almost the same. Thus, the relative improvement in MSE obtained by MWBME over MWK increases as the fraction of the soft data increased. Even though the relative improvement was relatively small (8.4%) for the true data, the improvement increased up to 43.7% as the number of missing daily concentrations increased. The completeness of intended samples at a given station varies because of intermittent malfunctioning or because the station is only in operation for part of the year. We find that between 1999 and 2007 the average completeness of intended samples varies from 68.0% to 88.7% (See Supplemental Material, Table 3). This range of completeness matches that of our simulated datasets (Table 2). Thus, using the MWBME method is expected to improve the mapping accuracy for $PM_{2.5}$ in the US from 1999 to 2007, as well as for other air pollutants or other countries with comparable or lower completeness of intended samples.

In this study we used a moving window calculated so that it encompasses 50 monitoring stations, which is the smallest number of stations needed to obtain a stable sample semivariogram. As a result the radius of the moving window ranged from 163km where the density of the monitoring stations was highest, to 1118km where the density was lowest (see Supplemental Material, p. 8).

Several other approaches have been used recently to estimate the long-term concentration over large geographic domains including the land use regression (LUR) approach (Hystad et al., 2011) and Geographic Information System (GIS)-based spatial smoothing (Hart et al., 2009, Yanosky et al., 2009). These approaches use land use covariates, such as land cover, road type, traffic count, and meteorological data to capture the effect of the pollutant sources and to detect small-scale variation in concentration. Although obtaining such land use covariate over large geographic domain is challenging, their inclusion in the modeling approach might solve much of the non stationarity of the problem. This is a limitation of this study which should be explored in future works. However, like the conventional kriging approach, these land use methods rely only on the average concentrations satisfying a completeness criterion. Therefore, the estimation quality of these approaches is expected to deteriorate as the completeness of the intended samples decreases.

To the best of our knowledge, this work is amongst the first U.S. nationwide studies of $PM_{2.5}$ yearly average concentrations that present an estimation method that both (a) accounts for the non-stationarity of the spatial variability of $PM_{2.5}$, and (b) incorporate $PM_{2.5}$ yearly average concentrations not meeting the completeness criterion. Our cross-validation analysis indicates that the MWBME method presented here performed better than

more conventional approaches, and also minimized the exposure misclassification for our cross validation dataset. As can be seen from Table 1, MWBME better preserved exposure rank order. MWBME also unraveled exposure gradients that are not as discernable using other methods. This is illustrated in Figure 3 (see also the movies of Supplementary Material) where the MWBME map reveals gradients in long-term $PM_{2.5}$ concentrations that are not visible in the map produced by the conventional SK method. In conclusion, epidemiological studies investigating the health effect of $PM_{2.5}$ yearly average concentrations over a large spatial domain should account for the non-stationarity of the environmental processes, and should use a methodological framework that can incorporate yearly concentrations that fail a completeness criterion, rather than entirely discarding these data. In this work we demonstrate that the MWBME method, based on a moving-window implementation of the BME framework, addresses these issues.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in part by grants R21HL89422 and R01AG033078 from the National Institutes of Health.

References

Bell ML, Dominici F, Ebisu K, Zeger SL, Samet JM. Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies. Environ Health Perspect. 2007; 115(7):989–995. [PubMed: 17637911]

Boldo E, Medina S, LeTertre A, Hurley F, Mucke HG, Ballester F, et al. Apheis: Health impact assessment of long-term exposure to PM2.5 in 23 European cities. Eur J Epidemiol. 2006; 21(6): 449–458. [PubMed: 16826453]

Christakos G. A Bayesian Maximum-entropy View To The Spatial Estimation Problem. Math Geol. 1990; 22(7):763–777.

Christakos, G. Modern spatiotemporal geostatistics. Oxford University Press; Oxford ;New York: 2000.

Christakos, G.; Bogaert, P.; Serre, ML. Temporal GIS : advanced functions for field-based applications. Springer; Berlin ;New York: 2001.

Haas TC. Lognormal And Moving Window Methods Of Estimating Acid Deposition. J Am Stat Assoc. 1990; 85(412):950–963.

Hart JE, Yanosky JD, Puett RC, Ryan L, Dockery DW, Smith TJ, et al. Spatial modeling of PM10 and NO2 in the continental United States, 1985-2000. Environ Health Perspect. 2009; 117(11):1690–1696. [PubMed: 20049118]

Hystad P, Setton E, Cervantes A, Poplawski K, Deschenes S, Brauer M, et al. Creating National Air Pollution Models for Population Exposure Assessment in Canada. Environ Health Perspect. 2011; 31:31.

Jerrett M, Burnett RT, Ma RJ, Pope CA, Krewski D, Newbold KB, et al. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology. 2005; 16(6):727–736. [PubMed: 16222161]

Hoek G, Beelen R, Dehoogh K, Vienneau D, Gulliver J, Fischer P, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atomospheric Environment. 2008; 42:7561–7578.

Liao D, Peuquet DJ, Duan Y, Whitsel EA, Dou J, Smith RL, et al. GIS approaches for the estimation of residential-level ambient PM concentrations. Environ Health Perspect. 2006; 114(9):1374–1380. [PubMed: 16966091]

Liao D, Whitsel EA, Duan Y, Lin HM, Quibrera PM, Smith R, et al. Ambient particulate air pollution and ectopy--the environmental epidemiology of arrhythmogenesis in Women's Health Initiative Study, 1999-2004. J Toxicol Environ Health A. 2009; 72(1):30–38. [PubMed: 18979352]

Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, et al. Long-term exposure to air pollution and incidence of cardiovascular events in women. N Engl J Med. 2007; 356(5):447–458. [PubMed: 17267905]

Olea RA. A six-step practical approach to semivariogram modeling. Stoch Env Res Risk A. 2006; 20(5):307–318.

Pope CA, Ezzati M, Dockery DW. Fine-Particulate Air Pollution and Life Expectancy in the United States. N Engl J Med. 2009; 360(4):376–386. [PubMed: 19164188]

Serre ML, Christakos G. Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study. Stoch Env Res Risk A. 1999; 13(1-2):1–26.

U.S. EPA. AQS Data Coding Manual v2.33. U.S. Environmental Protection Agency; Research Triangle Park, NC: 2008.

U.S. EPA. Air Quality System. U.S. Environmental Protection Agency; Research Triangle Park, NC: 2009. Available: http://www.epa.gov/ttn/airs/airsaqs/

Yanosky JD, Paciorek CJ, Suh HH. Predicting chronic fine and coarse particulate exposures using spatiotemporal models for the Northeastern and Midwestern United States. Environ Health Perspect. 2009; 117(4):522–529. [PubMed: 19440489]

Yu HL, Chen JC, Christakos G, Jerrett M. BME Estimation of Residential Exposure to Ambient PM10 and Ozone at Multiple Time Scales. Environ Health Perspect. 2009; 117(4):537–544. [PubMed: 19440491]

Zhang ZM, Whitsel EA, Quibrera PM, Smith RL, Liao D, Anderson GL, et al. Ambient fine particulate matter exposure and myocardial ischemia in the Environmental Epidemiology of Arrhythmogenesis in the Women's Health Initiative (EEAWHI) study. Environ Health Perspect. 2009; 117(5):751–756. [PubMed: 19479017]

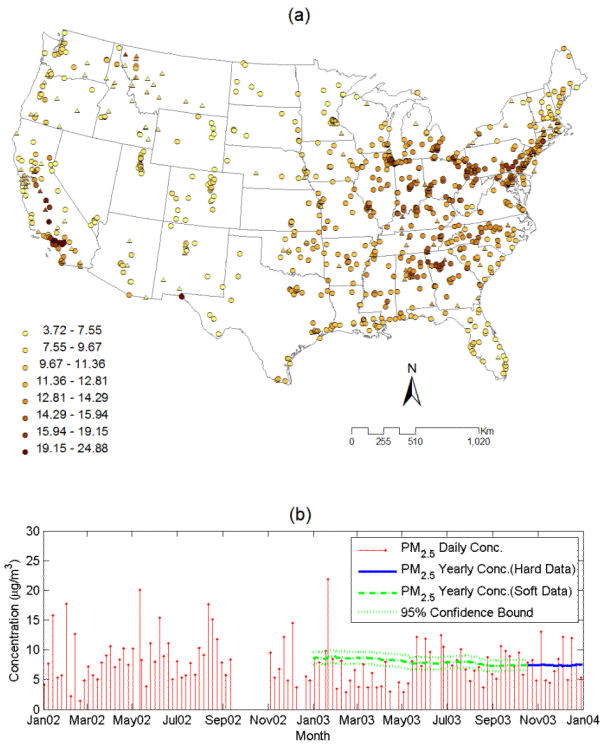

Figure1.

(a) $PM_{2.5}$ yearly average concentration ($\mu g/m^3$) on December 31, 2003 and (b) time series of $PM_{2.5}$ daily and yearly average concentrations at monitoring site 06-079-2002.

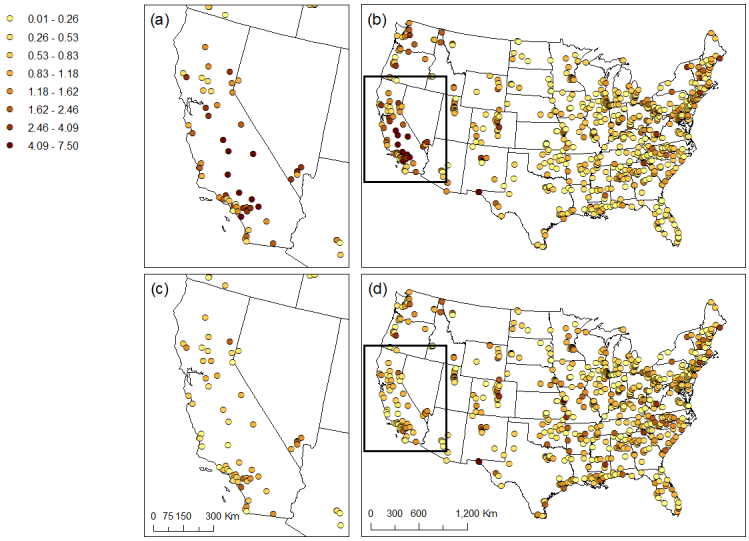

Figure2.

Absolute value of the standardized estimation error in the contiguous US (right) and in California (left) on December 31st, 2003 obtained by the SK ((a) and (b)), and the MWBME ((c) and (d)) estimation methods.

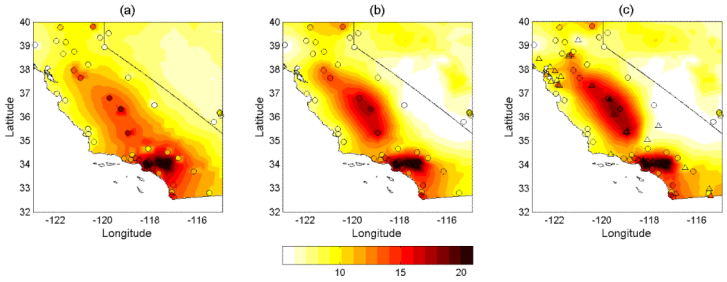

Figure3.
Maps of the estimated $PM_{2.5}$ yearly average concentrations (μg/m3) in California on December 31st, 2003 obtained with (a) SK, (b) MWK, and (c) MWBME.

Table 1

Cross-validation statistics obtained with the SK, MWK, and MWBME estimation methods based on an exponential semivariogram model (Size of validation set: n = 24544)

Method	SK	MWK	MWBME
MSE $((\mu g/m^3)^2)$	2.691	2.212	2.028
ME $(\mu g/m^3)$	0.0990	0.129	0.108
MS	0.0513	0.0531	0.0429
RMSS	1.094	1.074	1.068
MR $(\mu g/m^3)$	1.485	1.354	1.327
Pearson's Corr.	0.864	0.890	0.900
Spearman's Rank Corr.	0.872	0.890	0.899

Table 2

MSE obtained with the MWK and MWBME methods using the true and simulated datasets

	MWK	MWBME	Comp[a]	Soft Data[b]	Improvement[c]
True Data	2.21	2.03	84.5	18.3	8.4
Simulated Data 1 (5%)	2.34	2.04	80.6	25.9	12.8
Simulated Data 2 (10%)	2.64	2.09	76.6	39.9	20.9
Simulated Data 3 (15%)	3.13	2.16	72.5	62.0	31.2
Simulated Data 4 (20%)	3.97	2.24	68.6	82.0	43.7

[a] Average of a completeness of intended samples (%)

[b] Fraction of the soft data (%)

[c] Relative improvement (%) in MSE obtained by MWBME over MWK