A New Hybrid Spatio-Temporal Model For Estimating Daily Multi-Year PM$_{2.5}$ Concentrations Across Northeastern USA Using High Resolution Aerosol Optical Depth Data

Itai Kloog[1], **Alexandra A. Chudnovsky**[2], **Allan C. Just**[3], **Francesco Nordio**[3], **Petros Koutrakis**[3], **Brent A. Coull**[3], **Alexei Lyapustin**[5], **Yujie Wang**[6], and **Joel Schwartz**[3]

[1]Department of Geography and Environmental Development, Ben-Gurion University, Israel

[2]Department of Geography and Human Environment, Tel-Aviv University, Israel

[3]Department of Environmental Health - Exposure, Epidemiology and Risk Program, Harvard School of Public Health, Landmark Center 401 Park Drive West, Boston MA USA 02215

[4]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02215

[5]GEST/UMBC, NASA Goddard Space Flight Center, Baltimore, Maryland, USA

[6]University of Maryland Baltimore County, Baltimore, MD, USA

Abstract

Background—The use of satellite-based aerosol optical depth (AOD) to estimate fine particulate matter (PM$_{2.5}$) for epidemiology studies has increased substantially over the past few years. These recent studies often report moderate predictive power, which can generate downward bias in effect estimates. In addition, AOD measurements have only moderate spatial resolution, and have substantial missing data.

Methods—We make use of recent advances in MODIS satellite data processing algorithms (Multi-Angle Implementation of Atmospheric Correction (MAIAC), which allow us to use 1 km (versus currently available 10 km) resolution AOD data. We developed and cross validated models to predict daily PM$_{2.5}$ at a 1×1km resolution across the northeastern USA (New England, New York and New Jersey) for the years 2003–2011, allowing us to better differentiate daily and long term exposure between urban, suburban, and rural areas. Additionally, we developed an approach that allows us to generate daily high-resolution 200 m localized predictions representing deviations from the area 1×1 km grid predictions. We used mixed models regressing PM$_{2.5}$ measurements against day-specific random intercepts, and fixed and random AOD and temperature slopes. We then use generalized additive mixed models with spatial smoothing to generate grid cell predictions when AOD was missing. Finally, to get 200 m localized predictions,

Address correspondence to: Itai Kloog, Department of Geography and Environmental Development, Ben-Gurion University of the Negev, P.O.B. 653 Beer Sheva, Israel, ikloog@bgu.ac.il.

we regressed the residuals from the final model for each monitor against the local spatial and temporal variables at each monitoring site.

Results—Our model performance was excellent (mean out-of-sample R^2=0.88). The spatial and temporal components of the out-of-sample results also presented very good fits to the withheld data (R^2=0.87, R^2=0.87). In addition, our results revealed very little bias in the predicted concentrations (Slope of predictions versus withheld observations = 0.99).

Conclusion—Our daily model results show high predictive accuracy at high spatial resolutions and will be useful in reconstructing exposure histories for epidemiological studies across this region.

Keywords

Air pollution; Aerosol Optical Depth (AOD); Epidemiology; $PM_{2.5}$; Exposure error; High resolution aerosol retrieval; MAIAC

1. Introduction

The use of satellite-based aerosol optical depth (AOD) to estimate exposure to fine particulate matter ($PM_{2.5}$-particulate matter <2.53μm in aerodynamic diameter) concentrations for epidemiologic studies has increased substantially over the past few years (Chang et al., 2013; A. A. Chudnovsky et al., 2013; Chudnovsky et al., 2012; Kim et al., 2013; Kloog et al., 2012b, 2011; Lee et al., 2011; Lin et al., 2013; Nordio et al., 2013). Both short term (hours, days) and chronic (months, years) exposure to $PM_{2.5}$ has been extensively associated with detrimental human health effects (Halonen et al., 2008; Kloog et al., 2012a; Peacock et al., 2011; Zanobetti and Schwartz, 2009). Fine PM exposures were found to be associated with respiratory and cardiovascular morbidity, mortality from cardiovascular and respiratory diseases and an increase in hospital admissions (Dominici et al., 2006; Kloog et al., 2012a; Schwartz, 1996)

Traditionally, estimation of $PM_{2.5}$ exposures have been based on using measurements of a central ground monitor, and assigning these measurement to populations within a specified distance of the monitor (Laden et al., 2006; Samet et al., 2000). This approach introduces exposure error (also known as exposure misclassification), and likely biases the effect estimates downward due to spatial misalignment (Zeger et al., 2000). In recent years, to avoid this exposure misclassification, many studies have used regression models based on geographic covariates (termed "land use" regressions- LUR) to expand in situ measurements of $PM_{2.5}$ concentrations to large areas. LUR is essentially an interpolation technique that employs the $PM_{2.5}$ concentrations as the dependent variable, with proximate land use, traffic and physical environmental variables used as independent predictors (Beckerman et al., 2013; Gryparis et al., 2009; Hoek et al., 2008; Vienneau et al., 2010). Since the geographic covariates are generally not time varying, the temporal resolution of the LU model predictions is limited and thus can only be employed to assess long-term exposures for chronic health effects studies. Finally, current exposure assessments are limited to urban areas due to the paucity of monitoring sites in rural locations.

Satellite-based AOD provides physical daily measurements that can be used to estimate air quality and pollution due to its extensive spatial coverage and repeated observations of the earth surface and atmosphere. AOD is a measure of the extinction of electromagnetic radiation at a given wavelength due to the presence of aerosols in an atmospheric column. However, satellite-based AOD is a measure of light attenuation in the column which is affected by ambient conditions (such as vertical profile, chemical composition and humidity). In contrast, $PM_{2.5}$ is a measure of dry particle mass near the surface and thus we do not expect them to be strongly correlated (Chudnovsky et al., 2012).

In recent years many studies have used various statistical methods to establish quantitative relationships between AOD and $PM_{2.5}$ (Chang et al., 2013; A. Chudnovsky et al., 2013; Cordero et al., 2013; Gupta et al., 2013; Kim et al., 2013). Traditionally, the health exposure studies have used the standard MODIS (Moderate Resolution Imaging Spectroradiometer) AOD product of the "Dark Target" algorithm (Levy et al., 2007) which has a nadir resolution of 10×10 km^2. Lately, AOD at significantly higher spatial resolution (1×1 km2) has been offered by a new Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm. Chang and colleagues (Chang et al., 2013) used some novel ideas such as a statistical downscaling and data fusion techniques to predict $PM_{2.5}$ concentrations at spatial point locations in the southeastern United States during the period 2003–2005. Their model showed relatively high cross-validated predictions (R^2=0.78 and a root mean-squared error (RMSE) of 3.61 μg/m^3). Chudnovsky and colleagues (Chudnovsky et al., 2014) used one year of observations from MODIS based Aqua at a 1 km spatial resolution to obtain daily $PM_{2.5}$ estimates for AOD retrieval days. Their model results indicated that high resolution MAIAC (Multi-Angle Implementation of Atmospheric Correction) data better explains the variations in $PM_{2.5}$ However, the developed model was restricted to retrieval days, and the out-of-sample predictive performance spatially was similar to that estimated by Chang et al. (spatial R^2=0.79)

These recent studies, while generally showing better fits than previous models, still leave room for improvement in terms of reducing exposure error. In addition, they all lack detailed high resolution predictions across large space-time domains (especially in rural areas) which is critical for acute exposure epidemiologic studies.

In this paper, we incorporate MAIAC-based AOD satellite data which allows us to present a much simplified model, while still improving significantly over currently available models. We developed and validated models to predict daily $PM_{2.5}$ at a 1×1 km resolution across the northeastern USA (New England, New York and New Jersey) for the years 2003–2011, allowing us to better differentiate daily and long term exposure between urban, suburban, and rural areas. Additionally, we developed an approach that allows us to generate daily high-resolution 200×200 m localized predictions that are separate from the area 1×1 km grid predictions.

2. Material and Methods

2.1 Study domain

The spatial domain of our region included the Northeastern part of the USA (Figure 1), and includes the states of Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Rhode Island and Vermont. Many urban areas (notably New York and Boston) are included as well as rural hill towns, large forested regions, water bodies, mountains and the Atlantic sea shoreline. The study area included 285,284 discrete 1×1km satellite grid cells.

2.2 AOD Data

One of the fundamental aerosol products from MODIS is spectral AOD (sometime referred to as Aerosol Optical Thickness or AOT). This is a global level 2 daily MODIS product. The details of the standard "Dark Target" (DT) MODIS algorithm over land providing 10×10 km2 resolution AOD have been broadly reported (Levy et al., 2007; Remer et al., 2005). Recently a new processing algorithm (MAIAC) has been developed that provides a high 1km resolution AOD from MODIS data (Chudnovsky et al., 2014; Lyapustin et al., 2011a, 2011b). MAIAC processing begins with the gridding of MODIS L1B data to a fixed 1km grid, and accumulation of up to 16 days of measurements in the memory. Then it uses time series analysis and processing of groups of pixels to derive surface bidirectional reflectance distribution function (BRDF) and aerosol parameters without assumptions typical of current MODIS operational processing algorithms. The spatio-temporal analysis also helps MAIAC's cloud mask augmenting traditional pixel-level cloud detection techniques. In this analysis we used MAIAC AOD based on collection 6 MODIS Aqua L1B data for the years 2003–2011.

2.3 Monitoring data

Data for daily $PM_{2.5}$ mass concentrations across the Northeast region (see Figure 1) for the years 2003–2011 were obtained from the U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) database as well as the IMPROVE (Interagency Monitoring of Protected Visual Environments) network. IMPROVE monitor sites are located in national parks and wilderness areas while EPA monitoring sites are located across the Mid-Atlantic including urban areas such as New York City, Boston, etc. There were 161 monitors with unique locations operating in the Northeast during the study period. The Mean $PM_{2.5}$ across the Northeast during the study period was 11.7 μg/m3 with a standard deviation of 7.8 μg/m3 and an interquartile range (IQR) of 9.1 μg/m3.

2.4 Spatial Predictors of $PM_{2.5}$

We used a hybrid model, containing both AOD, land use, and meteorological predictors. The following spatial predictors were used: population density, elevation, traffic density, percentages of land use according to 12 land use categories based on United States Geological Survey (USGS) National Land Cover Dataset (NLCD), point emissions and total area-source emissions (tons per year) for $PM_{2.5}$, PM_{10}, SO_2, and NO_x.

Elevation—Elevation data were added through a satellite-based digital elevation model from the USGS National Elevation Dataset (NED) covering the United States at a spatial

resolution of 1 arc sec (Maune, 2007). There are sharp elevation contrasts across such a large study area and thus we used elevation as a spatial predictor (generally higher elevations are associated with lower air temperatures).

Traffic Density—Road data were obtained through the U.S. Census 2000 topologically integrated geographic encoding and referencing system (TIGER, 2006). We calculated the total A1 road length (class 1 roads that are hard surface highways including Interstate and U.S. numbered highways, primary State routes, and all controlled access highways) across the study area. The A1 roads were intersected with the 1×1 km grid and the resulting attribute tables contained the density of all A1 road segment lengths in each 1 km^2 grid cells.

Population Density—Population density data were obtained through the U.S. Census 2000 dataset (Census, 2000). We calculated the weight-averaged population for each 1×1 km grid cell based on the tracts intersecting these grid cells.

Point emissions—Additional point emissions data for $PM_{2.5}$, PM_{10}, SO_2, and NO_x were obtained through the 2005 U.S. EPA National Emissions Inventory (NEI) facility emissions report (EPA, 2010). Locations reporting zero emissions within the appropriate grid cell were assigned a value of one-half of the minimum value among all monitoring locations.

Area-source emissions—Area-source $PM_{2.5}$, PM_{10}, SO_2, and NO_x emissions data were obtained through the 2005 U.S. EPA-NEI tiered emissions reports (EPA, 2010), which provide estimates of total area-source emissions by county and year. Intersecting source emission areas for each 1×1 km grid were calculated by weight averaging the values from each intersecting areas intersecting n each 1×1km grid cell.

Percentages of land use—We used the NLCD from 2001 (Homer et al., 2004), available as raster files with a 30 m spatial resolution. We calculate the percentages of mixed forest, deciduous forest, evergreen, crop, pasture, grass, shrub, water, high development, medium development, low development and open development areas in each 1×1 km grid across the study area.

2.5 Temporal Predictors of PM$_{2.5}$

Meteorological data—All meteorological variables used in the analysis were obtained through the National Climatic Data Center (NCDC, 2010). Only continuous operating stations with daily data running from 2003–2011 were used (26 stations spread across the study area). Grid cells were matched to the closest weather station with available meteorological variables (24h means). We used the following meteorological variables: air temperature, wind speed, daily visibility, sea land pressure (SLP) and relative humidity. All the variables represent 24 hour averages except visibility, which is only computed during daylight hours.

NDVI—We used the publicly available monthly MODIS NDVI (Normalized Difference Vegetation Index) product (MOD13A3) at 1 km spatial resolution. The monthly resolution was chosen since NDVI values do not change considerably within a month except periods of spring green-up and fall senescence.

PBL—We used publicly available daily data on the height of the planetary boundary layer (PBL) obtained from the NOAA Reanalysis Data (NOAA, 2010). The spatial scale of the data was 32×32 km. The height of the boundary layer may vary with wind speed (Oke, 1987), influencing the concentration and vertical profile of pollutants. The boundary layer not only controls transport and location of pollutants and aerosols but also their concentrations would be different in variable boundary layer structures (Angevine et al., 2013).

2.6 Statistical Methods

All modeling was done using the R statistical software version 3.02 and SAS (Statistical Analysis System) version 9.3.

As we have shown in previous studies (Chudnovsky et al., 2014; Kloog et al., 2012b, 2011) there is a varying spatial relationship between AOD and $PM_{2.5}$ on a single day due to differences in factors including particle composition, PBL, relative humidity, vertical profiles. Thus, we use a mixed effects model incorporating spatial and temporal predictors and day-specific random-effects to take into account these temporal variations in the PM2.5–AOD relationship. Since the exposure predictions generated by our model are primarily intended for use in epidemiologic health effect studies, we generate exposure predictions at two different spatial scales at which health data are typically collected: small area (zip code, census tract, etc) geocoded data (SAGD) and address-specific geocoded data (RGD). When only SAGD health data is available, we use predictions at the 1×1km grid cell level, whereas when RGD are available, we add an additional local daily estimation component at a very high resolution (200×200m) to the grid-level predictions. Using higher resolution MAIAC data compared to our previous 10×10 km MODIS AOD data allowed us to simplify our approach in this way.

All models were fit to data from each year (2003–2011) separately. To generate the daily 1×1km $PM_{2.5}$ predictions in each grid cell for the entire 2003–2011 period, we developed a prediction process using a series of models: The first model calibrates the AOD grid-level observations to the $PM_{2.5}$ monitoring data collected within 1km of an AOD value, while adjusting for the land use and meteorological variables. In the second model we predict daily $PM_{2.5}$ concentrations in grid cells without monitors but with available AOD measurements using the model 1 fit. Then in the third model, to estimate daily $PM_{2.5}$ in grid cells with no AOD on that day, we take advantage of the region-specific association between grid-cell AOD and $PM_{2.5}$ levels, and the association between $PM_{2.5}$ level in a given grid with that in neighboring grid cells.

To generate the daily 200×200m $PM_{2.5}$ local predictions for studies using RGD, we take the residuals of model 3 at each monitoring site (that is the ground level $PM_{2.5}$ data minus the model 3 predictions) and regress them against very fine (200×200) monitor-specific spatial and temporal predictors of $PM_{2.5}$ which included traffic density, population density, elevation, percent urban, distance to major roads (A1), distance to source emission points, PBL and visibility. This stage allows us to predict local PM throughout the study area on any given day.

To accommodate the fact that daily AOD data missingness is not random, the first stage model incorporated inverse probability weighting (IPW) to potentially avoid bias in the regression coefficient estimates and thus in the resulting predictions. This approach effectively up-weights dates and grid cells which are under-represented due to a large degree of missing data. To obtain the weights that account for the non-random missingness in AOD values, we fit the following logistic regression model for the probability (p) of observing an AOD value in cell i on day j:

$$ln(\tfrac{\dot{p}}{1-\dot{p}})=\beta_0+\beta_1\text{Elevation}_i+\beta_2\text{SLP}_{ij}+\beta_3\text{Temperature}_{ij}+\text{Month}$$
$$IPW=(\tfrac{1}{\dot{p}})$$

(Equation 1)

where (p) is the probability for availability of AOD in each day in each grid cell in each year. There were no observations which had a disproportionate influence in the yearly models.

Finally, since the daily PM-AOD calibration factors can vary spatially, the study area is divided into 7 regions. The day-specific intercept, AOD, and temperature random effects in the model are nested within regions of the study. Specifically, the first model can be written as

$$PM_{ij}=(\alpha+u_j+g_{j(reg)})+(\beta_1+v_j+h_{j(reg)})\,AOD_{ij}+(\beta_2+k_j)\,\text{Temperature}_{ij}+\sum_{m=1}^{5}\gamma_{1m}X_{1mi}+\sum_{m=1}^{28}\gamma_{2m}X_{2mj}+\sum_{m=1}^{2}\gamma_{3m}X_{3mij}+\varepsilon_{ij}$$
$$(u_jv_jk_j)\sim N[(0\,0\,0),\textstyle\sum]$$
$$(g_{j(reg)},h_{j(reg)})\sim N[(0\,0),\textstyle\sum_{REG}]$$

(Equation 2)

where PM_{ij} is the measured PM2.5 concentration at a spatial site i on day j; α and u_j are the fixed and random (day-specific) intercepts, respectively, AOD_{ij} is the AOD value in the grid cell corresponding to site i on day j; β_1 and v_j are the fixed and day-specific random slopes, respectively. Temperature$_{ij}$ is the temperature value in the grid cell corresponding to site i on day j; and, β_2 and k_j are the fixed and random slopes for temperature. X_{1mi} is the value of the mth spatial predictor at site i, X_{2mj} is the value of the mth temporal predictor on day j, and X_{3mij} is the value of the mth spatial-temporal predictor at site i on day j. $g_{j(reg)}$ and $h_{j(reg)}$ are the daily random intercepts and AOD slopes specific to each study area region, nested within the overall random effects u_j and v_j. Here, we assume Σ is a 3×3 diagonal matrix with diagonal elements σ^2_u, σ^2_v, σ^2_k, and Σ_{REG} is a 2×2 diagonal matrix with diagonal elements σ^2_g, σ^2_h.

We used ten-folds out of sample cross validation (CV) to validate our model 1 predictions. We randomly divide our data into 90 and 10 percent splits ten times. We predict for the 10% data sets using the model fitted from the remaining 90% of the data. We then report these computed R^2 values. To test our results for bias we regress the measured *PM* value for a given site and day against the corresponding predicted value. We estimated the model prediction precision by taking the square root of the mean squared prediction errors

(RMSPE). In addition we calculated prediction errors from a model that contained the spatial components only, to make it more comparable to the commonly used monthly/yearly prediction models available such as Yanosky (Yanosky et al., 2008). Temporal R^2 was calculated by regressing Delta *PM* against Delta predicted where: Delta *PM* is the difference between the actual *PM* in place *i* at time *j* and the annual mean *PM* at that location, and Delta predicted is defined similarly for the predicted values generated from the model. Spatial R^2 was calculated by regressing the site-specific annual means in observed PM versus the same annual means for predicted *PM*.

The next model (model 2), uses the fit of model 1 to predict a $PM_{2.5}$ concentration for each day and grid cell at which we have an observed AOD value. This resulted in yearly datasets with $PM_{2.5}$ prediction for all day-grid cell combinations with available AOD.

In model 3 of the sequence, we estimate daily $PM_{2.5}$ for all grid cells in the study area even in days when no AOD data are present. We fit a generalized additive model with a smooth function of latitude and longitude (using the grid cell centroids) and a random intercept for each cell. This is similar to other interpolation techniques (universal kriging, etc.) to use nearby grid cells to help fill in the missing, however we also construct a 100 km buffer around each grid cell and fit a regression analysis relating the predicted PM at each grid cell to the the daily mean $PM_{2.5}$ from the stations in each buffer around that grid cell, which also aids in predicting the value on missing days. The 100 km buffer size was chosen since we wanted a small enough buffer to ensure relevance and hence improve the R^2 of the monitored value on the grid cell values, but large enough a buffer to include multiple PM monitors to produce more stable estimates (again improving the prediction R^2). To allow for temporal variations in the spatial correlation, we fit a separate spatial surface for each two-month period of each year. Using this method provides additional information about the concentration in the missing grid cells that classic interpolation would not provide. Specifically, we fit the following semiparametric regression model:

$$\text{PredPM}_{ij} = (\alpha + u_i) + (\beta_1 + v_i)\,\text{MPM}_{ij} + s(X_i, Y_i)_{k(j)} + \varepsilon_{ij},$$
$$(u_i v_i) \sim [(\mathbf{0}\,\mathbf{0}), \Omega_\beta] \qquad \text{(Equation 3)}$$

where Pred *PM*$_{ij}$ is the predicted $PM_{2.5}$ concentration at a grid cell i on a day *j* from the model 2; *MPM*$_{ij}$ is the mean PM in the relevant 100km buffer for site i on a day *j*; *α* and *u$_i$* are the fixed and grid-cell specific random intercepts, respectively; *β$_1$* and *v$_i$* are the fixed and random slopes, respectively. The smooth X_i, Y_i are the latitude and longitude, respectively, of the centroid of grid cell i, and s(X_i, Y_i) $_{k(j)}$ is the a smooth function of location (modeled by thin plate splines) specific to the two-month period k(*j*) in which day j falls (that is, a separate spatial smooth was fit for each two-month period).

To estimate the goodness of fit and due to computational limits, we used a "leave 10 out" approach where we randomly selected 10 monitors to leave out of the model 3 predictions. We the test the fit and bias (regressing the measured *PM* values against the predicted values and predicting PM levels at the left out monitoring locations).

Finally, we run a forth model where we take the residuals constructed by taking the difference between a given daily monitored $PM_{2.5}$ concentration and the 1km × 1km corresponding daily model 3 prediction, and regress that against spatial and temporal predictors of $PM_{2.5}$ at each monitor. Specifically, we fit the following model:

$$ResidPM_{ij} = f_1(Traffic\ Density_i, Population\ Density_i) + f_2(Elevation_i) + f_3(Percent\ Urban_i) +$$
$$f_4(Distance\ to\ A1\ roads_i) + f_5(Distance\ to\ Point\ emissions_i) + f_6(PBL_{ij}) + f_7(Traffic\ Density_i, PBL_{ij}) +$$
$$+ f_{8ij}(Traffic\ Density_i, Visibility_{ij}) + \varepsilon_{ij}$$

(Equation 4)

Where *ResidPM$_{ij}$* is the residual at a spatial monitor site *i on day j; f$_1$* denotes a penalized spline for an interaction between traffic density and population density at a spatial monitor site *i,, f$_2$–f$_6$* denote (potentially nonlinear) effects of elevation, percent urban land use, distance to A1 road, distance to point emissions and PBL respectively at a spatial monitor site *i (*and day *j* for PBL*). f$_7$* denotes a penalized spline for an interaction between traffic density and PBL at a spatial monitor site *i on day j* and *f$_8$* denotes a penalized spline for an interaction between traffic density and visibility at a spatial monitor site *i on day j.* Finally ε_{ij} is the error.

4. Results

Overall mean for MAIAC AOD was 0.25. Figure 2 presents the mean ground $PM_{2.5}$ measurements across all years (2003–2011) and the difference between monitor and predicted $PM_{2.5}$ concentrations at each monitoring site. The figure shows how measured $PM_{2.5}$ concentrations at monitors corresponds well with our predicted concentrations, and the differences at 95% of the monitoring sites are within ±1.3 μg/m3, clearly showing excellent agreement between observed and predicted values. Figure 3a presents a density plot exhibiting the daily variation of AOD slopes between 2003 and 2011 during model 1 calibrations, while Figure 3b presents a density plot exhibiting the daily variation of temperature slopes for the same period. The figure shows there is considerable day-to-day variability in these slopes.

Table 1 presents results from our model 1 calibration analysis. The yearly predictive power of our model is extremely high, with an overall "*out of sample*"R^2 for daily values of 0.88 (year-to-year variation 0.82–0.90), with a highly significant association between PM $_{2.5}$ and the main explanatory variable-AOD. The models yield almost no bias in our cross validation results as the slope of observed versus predicted =0.99 (year-to-year variation 0.98–1.01). The spatial and temporal out-of-sample results also presented excellent fits (Table 1): For the temporal model, the mean "*out of sample*"R^2 was 0.87 (year-to-year variation 0.81–0.90) and for the spatial model the mean "*out of sample*"R^2 was 0.87 (year-to-year variation 0.80–0.93). When looking at Root Mean Square Prediction Error (RMSPE), our models exhibit very low RMSPE values of 2.33 μg/m3 (year-to-year variation 1.95–2.89 μg/m3).

Our "spatial" component only RMSPE is much lower at 0.82 μg/m3. The final prediction model (model 3) presents very similar results and is presented in table 2.

Figure 4 shows the spatial pattern of predicted PM$_{2.5}$ concentrations in Boston from the AOD models, averaged over the entire study period. The spatial variation in these long term PM2.5 predictions ranges from 2.36 μg/m3 to 40.12 μg/m3 showing a good range of variability for our model.

Figure 5 shows the deviation of the estimated local PM$_{2.5}$ concentrations (model 4) from the average 1×1km PM$_{2.5}$ concentrations at a very fine resolution (200×200m) aggregated over a year (2003) in Boston.

4. Discussion

In this paper we used novel 1 km AOD MODIS data based on the MAIAC algorithm to predict PM$_{2.5}$ concentrations across the Northeastern USA. Using the newly available MAIAC data allows us to simplify our previously described models while still gaining much better predictive power and reducing the exposure error. In addition, in our updated model we introduce significant methodological improvements with the predictions for daily fine resolution (200×200m) PM exposure, which can better capture pollution attributed to local sources such as traffic. Our models yield extremely good model fits and observed versus predicted slopes that were practically '1', which clearly show that there is no bias in the resulting exposure predictions.

It is important to compare the predictions of our models to the other 'state of the art' prediction models that are available today. While the various existing models have all advanced in the past few years (including some models that use MAIAC data as well), they all still lack the ability to generate daily predictions for studies of the acute effects (short term) of air pollution as well as chronic (long term) effects, which are useful in epidemiologic studies aiming to estimate both acute and chronic effects of exposure. Sampson and colleagues (Sampson et al., 2013) recently published a universal kriging model using Partial Least Squares regression for estimating annual PM$_{2.5}$ concentrations. They present a high accuracy of prediction with an overall R^2 of 0.88 and well-calibrated predictive intervals. While the prediction accuracy of this model for long-term exposures is high, this model (1) does not predict daily and (2) is complex and requires an enormous number of spatial variables that are not always freely-available to researchers. In contrast our model presents similar spatial accuracy (R^2=0.87), but with the additional advantage of yielding daily predictions also with very good prediction accuracy (temporal R^2=0.87). Hu and colleagues (Hu et al., 2014) published a study estimating ground-level PM$_{2.5}$ concentrations in the Southeastern United States. Similar to our model they used MODIS based MAIAC data with a two-stage spatial statistical model with meteorological fields and land use parameters as ancillary variables to estimate daily mean PM$_{2.5}$ concentrations. The model was run in southeastern USA, for the year 2003. They used cross validation and predicted for only days with available MAIAC data resulting in an R^2 of 0.67 and RMSPE of 3.88 μg/m3. Their models do not yield daily predictions (in every grid cell on every day) and are less accurate compared to our presented model (such as cross validated RMSPE of

3.88 μg/m3 compared to 2.33 μg/m3 in our model). Other models published recently have all similar issues with lack of highly accurate *daily* predictions across the study area (Kim et al., 2013; Lin et al., 2013; Liu et al., 2009; Yanosky et al., 2008; Yap and Hashim, 2013). It is important to note that Beckerman (Beckerman et al., 2013) recently published results showing that in a Hybrid model estimating national scale spatiotemporal variability of pm$_{2.5}$ the inclusion of satellite based AOD into the LUR model did not improve the model. The study though used the older 10×10 MODIS data and generally the cross validated results were lower than our presented models (R^2=0.79).

Although the new 1×1 km MAIAC data is a considerable improvement over the previous MODIS 10×10 km data (for example, R^2=0.83 for our previously published 10 km model versus R^2=0.88 in the current MAIAC model, spatial R^2=0.78 versus R^2=0.87), finer scale or less noisy future satellite-based data could further reduce exposure error bias resulting in larger and more accurate health effects estimates. Jerrett and colleagues have previously demonstrated how fine-scale variations in PM$_{2.5}$ are associated with larger health effects than those that vary regionally (Jerrett et al., 2005). We partly address this limitation in our model by adding our new local stage model 4 predictions at very fine 200×200 m scale that complements our 1×1 km predictions.

Acknowledgments

References

Angevine WM, Brioude J, McKeen S, Holloway JS, Lerner BM, Goldstein AH, Guha A, Andrews A, Nowak JB, Evan S, et al. Pollutant transport among California regions. J Geophys Res Atmospheres. 2013; 118:6750–6763.

Beckerman BS, Jerrett M, Martin RV, van Donkelaar A, Ross Z, Burnett RT. Application of the Deletion/Substitution/Addition Algorithm to Selecting Land Use Regression Models for Interpolating Air Pollution Measurements in California. Atmos Environ. 2013:172–177.

Census, U.S. US Census of Population and Housing. 2000.

Chang HH, Hu X, Liu Y. Calibrating MODIS aerosol optical depth for predicting daily PM2. 5 concentrations via statistical downscaling. J Expo Sci Environ Epidemiol. 2013:398–404. [PubMed: 24368510]

Chudnovsky A, Lyapustin A, Wang Y, Schwartz J, Koutrakis P. Analyses of high resolution aerosol data from MODIS satellite: a MAIAC retrieval, southern New England, US, in: First International Conference on Remote Sensing and Geoinformation of Environment. International Society for Optics and Photonics. 2013:87951E–87951E.

Chudnovsky AA, Kostinski A, Lyapustin A, Koutrakis P. Spatial scales of pollution from variable resolution satellite imaging. Environ Pollut. 2013; 172:131–138. [PubMed: 23026774]

Chudnovsky AA, Koutrakis P, Kloog I, Melly S, Nordio F, Lyapustin A, Wang Y, Schwartz J. Fine particulate matter predictions using high resolution aerosol optical depth (AOD) retrievals. Atmos Environ. 2014:189–198.

Chudnovsky AA, Lee HJ, Kostinski A, Kotlov T, Koutrakis P. Prediction of daily fine particulate matter concentrations using aerosol optical depth retrievals from the Geostationary Operational Environmental Satellite (GOES). J Air Waste Manag Assoc. 2012; 62:1022–1031. DOI: 10.1080/10962247.2012.695321 [PubMed: 23019816]

Cordero L, Wu Y, Gross BM, Moshary F. Assessing satellite AOD based and WRF/CMAQ output PM2.5 estimators. 2013; :872319–18. DOI: 10.1117/12.2027430

Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, Samet JM. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. Jama. 2006; 295:1127–34. [PubMed: 16522832]

EPA. USEPA National Emissions Inventory. 2010.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. Measurement error caused by spatial misalignment in environmental epidemiology. Biostat Oxf Engl. 2009; 10:258–274. DOI: 10.1093/biostatistics/kxn033

Gupta P, Khan MN, da Silva A, Patadia F. MODIS aerosol optical depth observations over urban areas in. Atmos Environ. 2013; 45:27.

Halonen JI, Lanki T, Yli-Tuomi T, Kulmala M, Tiittanen P, Pekkanen J. Urban air pollution, and asthma and COPD hospital emergency room visits. Thorax. 2008; 63:635–641. [PubMed: 18267984]

Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, Briggs D. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ. 2008; 42:7561–7578. DOI: 10.1016/j.atmosenv.2008.05.057

Homer C, Huang C, Yang L, Wylie B, Coan M. Development of a 2001 national landcover database for the United States. Photogramm Eng Remote Sens. 2004; 70:829–840.

Hu X, Waller LA, Lyapustin A, Wang Y, Al-Hamdan MZ, Crosson WL, Estes MG Jr, Estes SM, Quattrochi DA, Puttaswamy SJ, Liu Y. Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sens Environ. 2014; 140:220–232. DOI: 10.1016/j.rse.2013.08.032

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J, Giovis C. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol. 2005; 15:185–204. DOI: 10.1038/sj.jea.7500388 [PubMed: 15292906]

Kim M, Zhang X, Holt JB, Liu Y. Spatio-Temporal Variations in the Associations between Hourly PM2. 5 and Aerosol Optical Depth (AOD) from MODIS Sensors on Terra and Aqua. Health (N Y). 2013; 5:8–13.

Kloog I, Coull BA, Zanobetti A, Koutrakis P, Schwartz JD. Acute and Chronic Effects of Particles on Hospital Admissions in New-England. PLoS ONE. 2012a; 7:e34664.doi: 10.1371/journal.pone.0034664 [PubMed: 22529923]

Kloog I, Koutrakis P, Coull BA, Lee HJ, Schwartz J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmos Environ. 2011; 45:6267–6275. DOI: 10.1016/j.atmosenv.2011.08.066

Kloog I, Nordio F, Coull BA, Schwartz J. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2. 5 exposures in the Mid-Atlantic states. Environ Sci Technol. 2012b; 46:11913–11921. [PubMed: 23013112]

Laden F, Schwartz J, Speizer FE, Dockery DW. Reduction in fine particulate air pollution and mortality: Extended follow-up of the Harvard Six Cities study. Am J Respir Crit Care Med. 2006; 173:667–72. [PubMed: 16424447]

Lee HJ, Liu Y, Coull BA, Schwartz J, Koutrakis P. A novel calibration approach of MODIS AOD data to predict PM 2.5 concentrations. Atmos Chem Phys. 2011; 11:7991–8002.

Levy R, Remer L, Mattoo S, Vermote E, Kaufman Y. Second-generation operational algorithm: Retrieval of aerosol properties over land from inversion of Moderate Resolution Imaging Spectroradiometer spectral reflectance. J Geophys Res. 2007; 112:1984–2012. DOI: 10.1029/2006JD007811

Lin G, Fu J, Jiang D, Hu W, Dong D, Huang Y, Zhao M. Spatio-Temporal Variation of PM2. 5 Concentrations and Their Relationship with Geographic and Socioeconomic Factors in China. Int J Environ Res Public Health. 2013; 11:173–186. [PubMed: 24362546]

Liu Y, Paciorek CJ, Koutrakis P. Estimating Regional Spatial and Temporal Variability of PM2.5 Concentrations Using Satellite Data, Meteorology, and Land Use Information. Env Health Perspect. 2009; 117:886–892. DOI: 10.1289/ehp.0800123 [PubMed: 19590678]

Lyapustin A, Martonchik J, Wang Y, Laszlo I, Korkin S. Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables. J Geophys Res Atmosheres 1984–2012. 2011a:116.

Lyapustin A, Wang Y, Laszlo I, Kahn R, Korkin S, Remer L, Levy R, Reid JS. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. J Geophys Res Atmospheres 1984–2012. 2011b:116.

Maune, D. Digital elevation model technologies and applications: The DEM users manual. American Society for Photogrammetry and Remote Sensing; Maune, DF: 2007.

NOAA. [accessed 1.29.14] NCEP/NCAR Reanalysis 1 [WWW Document]. 2010. http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html

Nordio F, Kloog I, Coull BA, Chudnovsky A, Grillo P, Bertazzi PA, Baccarelli AA, Schwartz J. Estimating spatio-temporal resolved PM10 aerosol mass concentrations using MODIS satellite data and land use regression over Lombardy, Italy. Atmos Environ. 2013; 74:227–236. DOI: 10.1016/j.atmosenv.2013.03.043

Oke, TR. Boundary layer climates. Psychology Press; 1987.

Peacock JL, Anderson HR, Bremner SA, Marston L, Seemungal TA, Strachan DP, Wedzicha JA. Outdoor air pollution and respiratory health in patients with COPD. Thorax. 2011; 66:591–596. [PubMed: 21459856]

Remer LA, Kaufman YJ, Tanré D, Mattoo S, Chu DA, Martins JV, Li R-R, Ichoku C, Levy RC, Kleidman RG, Eck TF, Vermote E, Holben BN. The MODIS Aerosol Algorithm, Products, and Validation. J Atmospheric Sci. 2005; 62:947–973.

Samet J, Zeger S, Dominici F, Curriero F, Coursac I, Dockery D, Schwartz J, Zanobetti A. The national morbidity, mortality, and air pollution study. Part II Morb Mortal Air Pollut U S Res Rep Health Eff Inst. 2000; 94:5–70.

Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson TV, Kaufman JD. A Regionalized National Universal Kriging Model using Partial Least Squares Regression for Estimating Annual PM$_{2.5}$ Concentrations in Epidemiology. Atmos Environ. 2013

Schwartz J. Air pollution and hospital admissions for respiratory disease. Epidemiology. 1996; 7:20–8. [PubMed: 8664396]

TIGER. Topologically Integrated Geographic Encoding and Referencing system. 2006.

Vienneau D, De Hoogh K, Beelen R, Fischer P, Hoek G, Briggs D. Comparison of land-use regression models between Great Britain and the Netherlands. Atmos Environ. 2010; 44:688–696.

Yanosky JD, Paciorek CJ, Suh HH. Predicting Chronic Fine and Coarse Particulate Exposures Using Spatio-temporal Models for the Northeastern and Midwestern US. Environ Health Perspect. 2008; :522–529. DOI: 10.1289/ehp.11692 [PubMed: 19440489]

Yap XQ, Hashim M. A robust calibration approach for PM 10 prediction from MODIS aerosol optical depth. Atmospheric Chem Phys. 2013; 13:3517–3526.

Zanobetti A, Schwartz J. The Effect of Fine and Coarse Particulate Air Pollution on Mortality: A National Analysis. Environ Health Perspect. 2009; 117:898–903. DOI: 10.1289/ehp.0800108 [PubMed: 19590680]

Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, Cohen A. Exposure measurement error in time-series studies of air pollution: concepts and consequences. Env Health Perspect. 2000; 108:419–26. [PubMed: 10811568]

Research highlights

- Our Models resulted in very high out-of-sample R^2 (R^2=0.88)

- Our model performed well both spatially and temporally (R^2=0.87, R^2=0.87)

- Our results revealed very little bias (Slope of predictions versus withheld observations = 0.99)

- Importantly, these R^2 are for daily, rather than monthly or yearly, values.
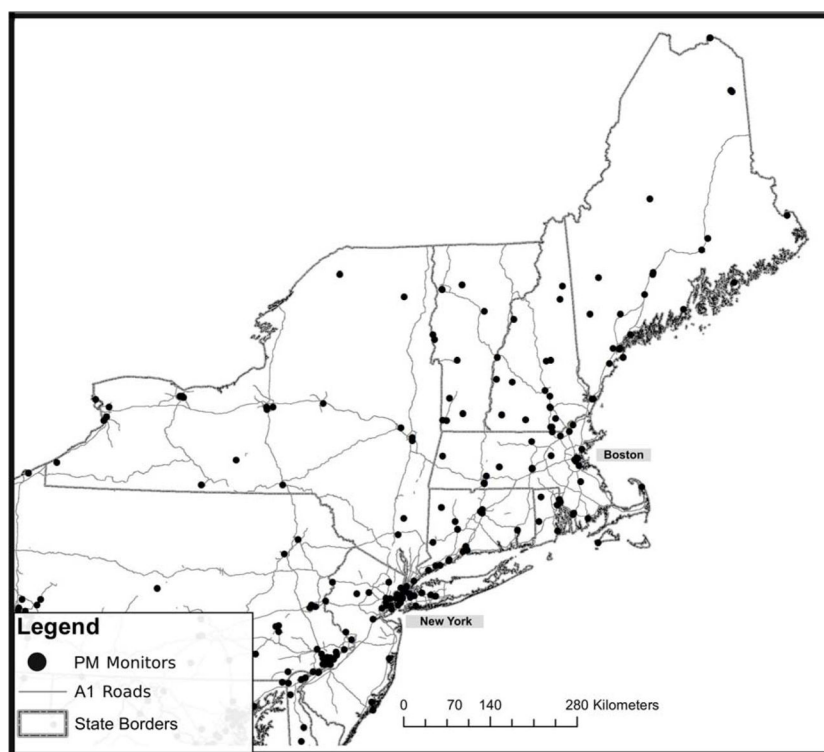
Figure 1.
Map of the study area showing the location of EPA and IMPROVE monitoring sites across the Northeastern USA.
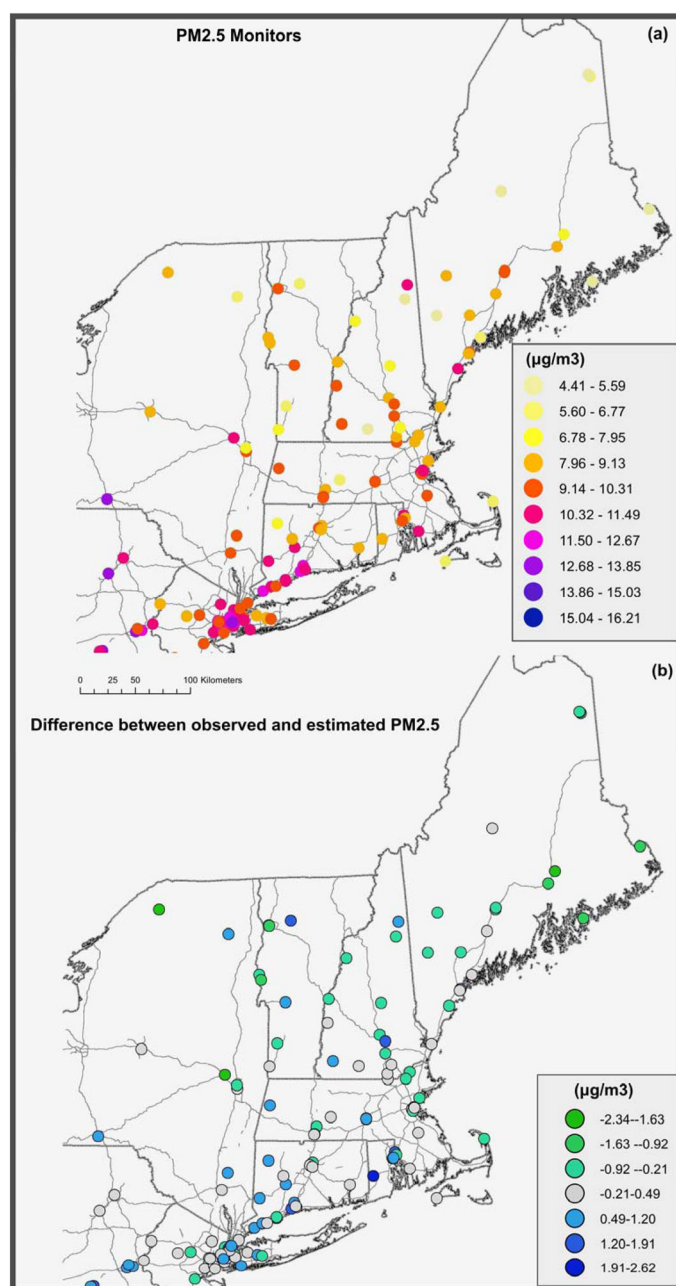
Figure 2.
EPA and IMPROVE PM$_{2.5}$ annual means and the difference between measured and estimated PM$_{2.5}$ concentrations at each PM$_{2.5}$ monitor.
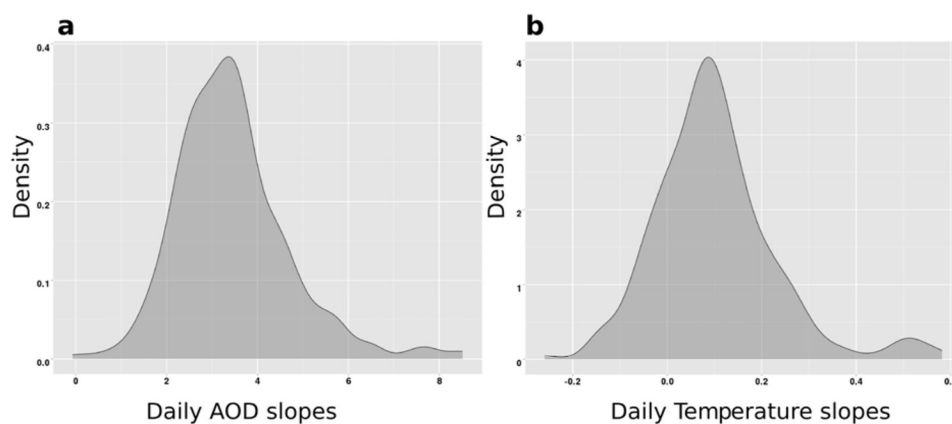
Figure 3.
Density plots exhibiting the daily variation of AOD slopes (a) and temperature slopes (b) between 2003 and 2011 during the stage 1 calibrations.
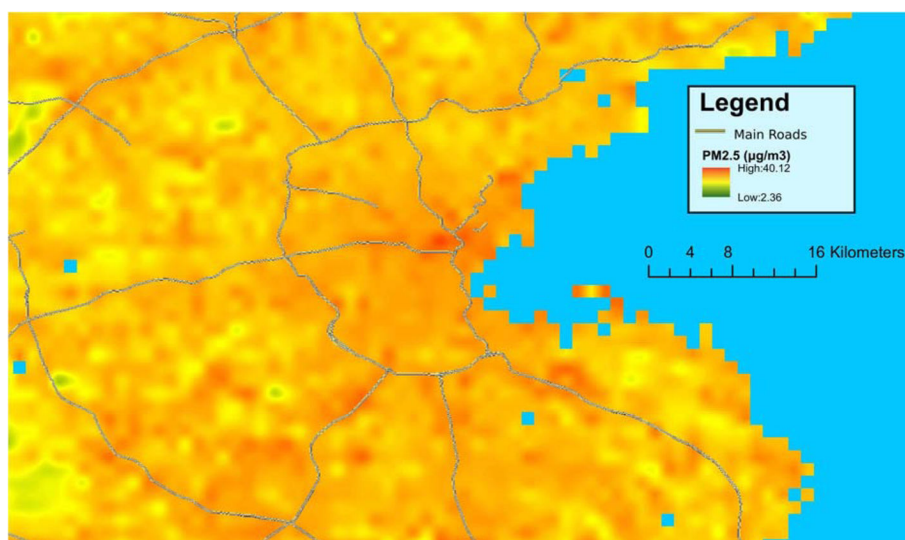
Figure 4.
Mean PM$_{2.5}$ concentrations in each 1×1 km grid cell in Boston during 2003 predicted by the AOD models
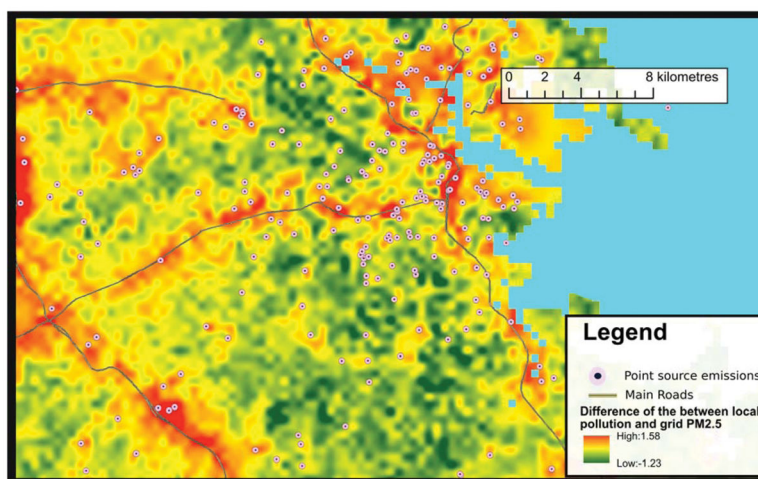
Figure 5.
The deviations of the estimated local pollution concentrations at a very fine resolution (200×200 m) from the average 1×1km grid PM$_{2.5}$ aggregated over a year (2003) in Boston

Table 1

Prediction accuracy: Ten-fold cross-validated R^2 for $PM_{2.5}$ stage 1 predictions (Calibration stage for 2003–2011).

year	R^2	Slope	Spatial R^2	Temporal R^2	RMSPE (µg/m³)	Spatial RMSPE (µg/m³)
2003	0.89	0.98	0.87	0.88	2.89	1.05
2004	0.89	0.99	0.90	0.88	2.34	0.85
2005	0.88	0.98	0.87	0.87	2.85	1.00
2006	0.89	1.00	0.85	0.88	2.30	0.81
2007	0.90	1.01	0.93	0.90	2.34	0.73
2008	0.88	0.99	0.92	0.86	2.22	0.72
2009	0.86	0.98	0.86	0.84	1.97	0.71
2010	0.90	0.99	0.84	0.89	1.85	0.69
2011	0.82	0.99	0.80	0.81	2.22	0.82
Overall Mean 2003–2011	**0.88**	**0.99**	**0.87**	**0.87**	**2.33**	**0.82**

Table 2

Prediction accuracy: R^2 for Stage 3 PM$_{2.5}$ predictions (final prediction model including locations without AOD for 2003–2011).

year	R^2	Slope	Spatial R^2	Temporal R^2	RMSPE (µg/m³)	Spatial RMSPE (µg/m³)	Slope for leave 10-out
2003	0.91	1.02	0.88	0.91	2.69	0.65	1.09
2004	0.89	1.01	0.93	0.88	2.47	0.60	1.09
2005	0.88	1.01	0.93	0.88	2.79	0.63	1.10
2006	0.89	1.02	0.93	0.89	2.37	0.53	1.08
2007	0.91	1.01	0.96	0.91	2.26	0.47	1.04
2008	0.87	0.99	0.96	0.86	2.30	0.42	1.03
2009	0.87	1.01	0.94	0.86	1.95	0.37	1.06
2010	0.89	1.00	0.96	0.88	1.91	0.29	1.04
2011	0.84	1.01	0.92	0.83	2.09	0.39	1.05
Overall Mean	**0.88**	**1.01**	**0.93**	**0.88**	**2.32**	**0.48**	**1.07**