Available online at www.sciencedirect.com
ScienceDirect
Atmospheric Environment 41 (2007) 2255–2269
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

A land use regression for predicting fine particulate matter concentrations in the New York City region

Zev Ross[a,*], Michael Jerrett[b], Kazuhiko Ito[c], Barbara Tempalski[d], George D. Thurston[c]

[a]ZevRoss Spatial Analysis, Ithaca, NY 14850, USA
[b]Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA 94720, USA
[c]Nelson Institute of Environmental Medicine, New York University School of Medicine, Tuxedo Park, NY 10987, USA
[d]Department of Geography, University of Washington, Seattle, WA 98195-2350, USA

Received 23 June 2006; received in revised form 7 November 2006; accepted 8 November 2006

Abstract

We developed regression equations to predict fine particulate matter ($PM_{2.5}$) at air monitoring locations in the New York City region using data on nearby traffic and land use patterns. Three-year averages (1999–2001) of $PM_{2.5}$ at US Environmental Protection Agency (EPA) monitors in the 28 counties including and surrounding New York City were calculated using daily data from the EPA's Air Quality Subsystem. As the secondary contribution to $PM_{2.5}$ concentrations is lowest in the winter, we also calculated and modeled average winter 2000 $PM_{2.5}$ to conduct a preliminary evaluation of model sensitivity to source contribution. Candidate predictor variables included traffic, land use, census and emissions data from local, state and national sources and were tabulated for a series of circular buffer regions at varying distances around the monitors using a geographic information system. In total, more than 25 variables at 5 different buffer distances were considered for inclusion in the model. Before evaluating the variables we removed several samples from the modeling for validation. For comparison and validation purposes we computed both a model using data for the full 28-county region as well as a more urbanized 9-county region. We found that traffic within a buffer of 300 or 500 m explains the greatest proportion of variance (37–44%) in all 3 models. Measures of urbanization, specifically population density, explain a significant amount of the residual variation (7–18%) after including a traffic variable. Finally, a measure of industrial land use further improves the 28-county and 9-county models based on the 3-yr annual averages, explaining an additional 4% and 11% of the variation, respectively, while vegetative land use improves the winter model explaining an additional 6%. The final models predicted well at validation locations. In total, the final land use regression models explain between 61% and 64% of the variation in $PM_{2.5}$.

Keywords: Land use regression; Particulate matter; New York; Traffic; Exposure

1. Background

Although exposure to $PM_{2.5}$ (particles up to 2.5 μm in diameter) has been linked to a wide array of health effects such as aggravation of existing

*Corresponding author. Tel.: +1 607 277 0004; fax: +1 866 877 3690.
E-mail addresses: zev@zevross.com (Z. Ross), jerrett@berkeley.edu (M. Jerrett), kaz@env.med.nyu.edu (K. Ito), bjtemp@u.washington.edu (B. Tempalski), thurston@env.med.nyu.edu (G.D. Thurston).

1352-2310/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2006.11.012

heart and lung disease and premature mortality (Pope et al., 2002), health studies often rely on relative between-city central monitor estimates that assign entire metropolitan areas the same level of exposure. Recent studies of $PM_{2.5}$ have shown that within-city or intra-urban exposure gradients can also be associated with atherosclerosis (Künzli et al., 2005) and high risks of premature mortality (Jerrett et al., 2005a). These studies have used geostatistical interpolation models that capture regional patterns of pollution well, but often fail to account for near-source impacts from local traffic and industry. Given the large health effects reported in these and other European studies (Hoek et al., 2002; Nafstad et al., 2003) a need exists to refine these estimates of intra-urban exposure to reduce uncertainties potentially associated with measurement error.

Several recent studies have demonstrated the potential of land use regression to supply accurate, small-area estimates of air pollution concentrations without the expense of dispersion or exposure modeling (Brauer et al., 2003; Briggs et al., 2000). The goal of land use regression is to explain as much of the variation in existing air quality data for a given pollutant using data on nearby traffic, land use and other variables. Using, in most cases, multiple linear regression, a model is developed with existing monitors that can then be applied to unmonitored locations provided the appropriate geographic data is available.

Ross et al. (2006) developed land use regression models using traffic, distance to the coast and road length measures that explained nearly 80% of the variation in nitrogen dioxide levels in San Diego, California and were able to predict validation locations—locations that were not included in the modeling—to within, on average, 2.1 ppb. Land use regression models predicting nitrogen dioxide using traffic and other variables in Montréal and several European cities also produced accurate predictions (Jerrett et al., 2005b).

In contrast to more localized gases such as nitrogen dioxide, however, particulate matter mass has a significant regional secondary component with smaller contributions from local sources (Bari et al., 2003). This complicates estimating intra-urban exposure with land use regression. Models of $PM_{2.5}$ in three European cities produced mixed results. One of the only studies to attempt prediction of fine particle concentrations with the land use regression methods to date was undertaken in Europe as part of the Traffic Related Air Pollution and Childhood Asthma study (TRAPCA). Researchers measured $PM_{2.5}$ for representative temporal periods over 1 yr in the Netherlands, Munich, Germany and Stockholm County, Sweden. They found significant differences in model explanatory power from region to region with R^2 values ranging from 73% (the Netherlands) down to 56% and 50% (Munich and Stockholm, respectively). The limited variability of the monitoring sites in Stockholm County was a suggested explanation for this difference (Brauer et al., 2003). Other health analyses have shown the model predicts some childhood respiratory outcomes (Brauer et al., 2002). More recently, in Germany, Hochadel et al. found that land use regression predicted $PM_{2.5}$ absorbance ($R^2 = 65$–82%), but failed to predict $PM_{2.5}$ mass well ($R^2 = 9$–17%) (Hochadel et al., 2006).

North American cities have vastly different transportation and land use patterns and the applicability of land use regression to predict $PM_{2.5}$ is unknown (Gilbert et al., 2005). To our knowledge this is the first attempt to apply land use regression to the analysis of $PM_{2.5}$ in North America.

2. Methods

2.1. Overview

Land use regression uses concentrations of ambient pollution at the monitoring location as the dependent variable. Surrounding land use, transportation and other data are extracted using geographic information systems (GIS) and included in a regression equation as predictor variables. In this study we have assembled a database of information on land use and transportation around the $PM_{2.5}$ monitors in the New York City region. After describing the data, we present our statistical modeling strategy.

2.2. Dependent variables: Ambient particulate matter data

Three-year averages (1999–2001) of $PM_{2.5}$ for monitors in the 28-county area of New York City (Fig. 1) were calculated using data from the US EPA's Air Quality Subsystem (AQS). We first computed quarterly averages for each year for monitors that had at least 8 observations (i.e., at

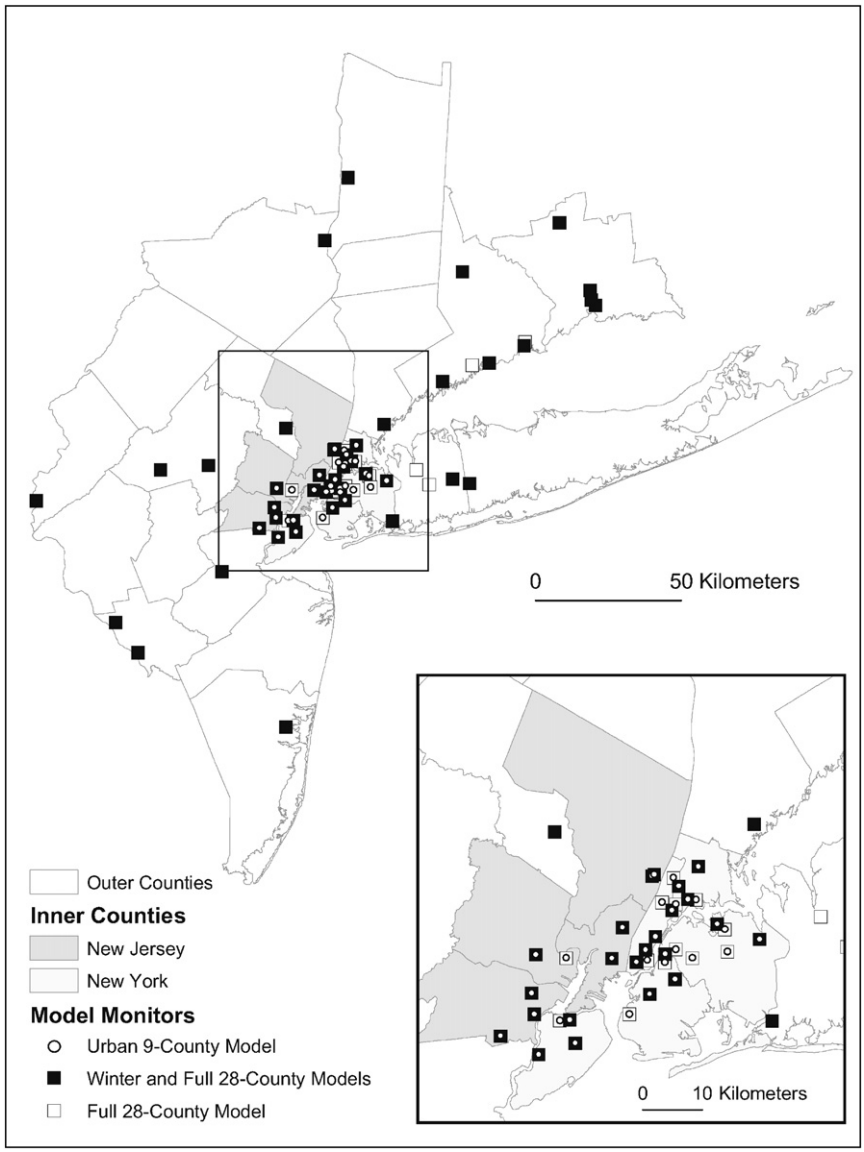

Fig. 1. $PM_{2.5}$ samples in the New York City region.

least half of the scheduled monitoring days). We then computed 3-yr averages by taking an average of the quarterly means when a complete set of 4 quarterly means existed. All particulate matter concentrations are in $\mu g\,m^{-3}$. Winter 2000 averages were computed using data from January, February and March 2000. Monitors were included if they had at least 8 observations for the quarter.

At several monitoring locations both monitors using the Federal Reference Method (FRM) and the tapered-element oscillating microbalance continuous monitoring method (TEOM) were used. When both TEOM and FRM data were collected at the same site, we used FRM data. We limited the winter 2000 analysis to those sites that had both 3-yr and winter 2000 averages. In total 62, 36 and 45 monitors were included in the analyses of the 28-county (1999–2001), 9-county (1999–2001) and winter (2000) models, respectively. The 28 counties were chosen to mirror the counties included in the Best Practice Model of the New York Metropolitan Transportation Council (NYMTC) for regional travel demand forecasting in the New York metropolitan area. The 9 counties were selected to represent the more urbanized area of New York City's five boroughs and adjacent New Jersey

counties. In general, the 28 counties range from very urban to relatively rural. Twelve counties have fewer than 500,000 people (with a minimum population of 96,000 in Putnam County), while six counties have more than 1 million people (with a maximum population of 2,470,000 in Kings County). Most counties have heavily urban populations—the populations are greater than 90% urban in 22 counties and greater than 99% in 12. The 6 least urban counties (Hunterdon, Warren, Sussex, Orange, Dutchess and Putnam), all of which are more than 50 km west or north of downtown Manhattan, range between 40–76% urban. For comparison purposes, the average county population for the 3 states is 340,000 while the average urban percentage is 64%.

Although AQS includes, in most cases, monitor latitudes and longitudes, we found these data were inaccurate in some locations. We therefore used geolocations provided directly by the appropriate state agencies (the Department of Environmental Conservation in NY and the Departments of Environmental Protection in NJ and CT). In most cases, the locations were identified by the agencies with either a GPS or orthophotos in a GIS. For a limited number of older sites, these more accurate methods were not used by the agencies so we manually verified their locations in a GIS using a combination of road layers, orthophotos, New York State Department of Transportation Raster Quadrangles and United States Geological Survey (USGS) Digital Raster Graphics.

2.3. Independent variables: Traffic, land use, population and emissions data

At each air monitoring location in AQS with adequate data on $PM_{2.5}$ we generated circular buffers around the monitor location with varying radii (50, 100, 300, 500 and 1000 m). All of the GIS layers were then intersected with the circular buffers and the traffic and land use composition of all buffers at each air monitoring location was calculated. Calculations were performed using a vector data structure using ArcGIS 9.1 (ESRI, Redlands, CA).

Traffic and road data: We used 2002 traffic estimates for New York City and surrounding counties provided by the NYMTC, the regional council of governments established to help make transportation-related decisions in the region. Traffic estimates were generated by the NYMTC for approximately 40,000 road links in 28 counties using the Best Practice Model (BPM), a tool used for regional travel demand forecasting in the metropolitan area. The BPM makes use of traffic data from more than a dozen municipal and other sources. The road network used in the BPM was developed to include all minor arterial roads or higher using five separate network databases conflated to match the LION street centerline file in New York's five boroughs and the US Census Bureau's Topologically Integrated Geographic Encoding and Referencing database elsewhere. A complete description of the model and line network development can be found in Parsons Brinckerhoff Quade & Douglas, Inc. (2005). The traffic estimates, divided by vehicle type, are available for four time periods (morning, midday, afternoon and night). Estimates for afternoon flows were used for the analysis as a complete set of estimates (for all 40,000 links) for the other time periods was not yet available. We calculated buffer totals for truck traffic and total traffic. Traffic data were scaled to 1000 s of vehicle km^{-1}. In addition to the traffic calculations, we used this data source to calculate road density (in kilometers) in each buffer area.

Land use data: Land use data were assembled from several sources. These include extremely detailed (1 in to approximately 250 ft) tax lot data from the New York City Department of City Planning covering all tax lots in the five boroughs (2003), medium-scaled (1:40,000) land use data for New Jersey from the New Jersey Department of Environmental Conservation (1995/1997) and coarse scale (1:100,000) land use data (National Land Cover Data) from the USGS (1997, based on Landsat images from 1992 and confirmed using aerial photos). For each land use layer the total area of individual land use categories (e.g., industrial, forest, etc) in each buffer was calculated. Although the tax lot and NJ land use layers were more detailed, only the data from the USGS covered the entire area of interest.

Similar land uses within each land use layer were grouped to create aggregate variables. The land use categories within the three layers were not identical but comparable groupings were tabulated. We generated an industrial and residential category from all three land use layers. In addition, for the NJ and National Land Cover Data we created variables representing water, vegetation and barren land use types.

The industrial category discussed throughout most of this paper comprises the industrial/commercial/transportation category from USGS data and the NJ land use data (this will be called "industrial"). We created a comparable grouping using the NYC tax lot data by combining the categories denoted "Industrial and Manufacturing", "Transportation and Utility" and "Commercial and Office Buildings." The NYC tax lot data were used for sensitivity analysis as this data did not cover the entire region. Land use areas were converted to acres.

Census population data: Census data (2000) at the block group level was acquired from the US Census Bureau's Summary File 1 (for population and housing data) and Summary File 3 (for income data). All population and housing data are in 1000s, income data was not scaled (US Census Bureau, 2000).

Emissions data: We calculated $PM_{2.5}$ primary emissions (area, point, off-road mobile and on-road mobile) for each county in the analysis and each monitor was assigned the value associated with its county. We also calculated the number of point sources and amount of $PM_{2.5}$ emissions from point sources within each of the buffers using data for 1999 from the US EPA's National Emissions Inventory.

Statistical data analysis: In total, more than 25 variables at 5 different buffer distances were considered for inclusion in the final model. These include total traffic, total truck traffic, industrial land use, residential land use, total county-wide emissions, point emissions, total number of point sources of $PM_{2.5}$, total population, number of households, number of housing units, median income of population and percent nonwhite population.

We developed 3 separate models, 2 that included data from the full 28-county region, one of which used the 3-yr average $PM_{2.5}$ as the response (the "28-county model") and one that used the winter 2000 $PM_{2.5}$ as the response (the "winter model"); and a third (the "9-county model") using 3-yr average $PM_{2.5}$ as the response but was geographically limited to 9 more urbanized counties. Before evaluating the variables we removed several samples from the modeling for validation. Samples were removed from each model separately using the same criteria. We divided the samples based on their land use code in the AQS into Agricultural-Forest, Commercial, Industrial-Mobile and Residential categories. Twenty percent of each group was removed for validation. The modeling subsample for the 28-county model included 49 samples (13 for validation), for the winter model 9 samples were withheld and the model building was based on 36 samples and for the 9-county model we used 29 samples (7 for validation) (Table 1).

All of the variables discussed above were included in our variable selection procedures. We used a combination of forward, stepwise and all-subsets selection and included variables in the model, as usual, based on the sums of squares explained, Mallow's Cp, variance inflation factors and other diagnostics. In the forward selection process, as the strongest variables are added to the model the remaining variables are re-evaluated for inclusion. As such, the same variable at different buffer distances may be included if it outperforms other variables. The sensitivity of model parameters to the sample selection was evaluated with a bootstrap in which 5 random samples were excluded, the model was run, the coefficients were recorded, the 5 samples were returned to the pool and another random sample of 5 was removed. This process was repeated 10,000 times.

Table 1
Modeling and validation samples by land use category

Land use category[a]	3-yr average (1999–2001)				Winter 2000 average	
	28-county total	28-county validation	9-county total	9-county validation	28-county total	28-county validation
Industrial/mobile	8	2	6	1	4	1
Residential	32	6	19	4	19	5
Commercial	18	4	11	2	10	2
Agricultural/forest	4	1	0	0	3	1
Total	62	13	36	7	36	9

[a]Based on the codes the EPA's Air Quality Subsystem.

To evaluate the independence assumption, we tested the spatial autocorrelation in the particulate matter values themselves and then in the residuals from our final models using the Moran's *I* statistic (Bailey and Gatrell, 1995) and two neighborhood constructions—a Queen's contiguity matrix based on Thiessen polygons and a nearest neighbor approach where we limited the analysis to the 3 nearest neighbors. Statistical significance was tested using a permutation test with 999 iterations. We also qualitatively assessed the extent of spatial autocorrelation using variograms of the residuals from our final models.

For visualization purposes, we created a smooth surface of interpolated predictions by predicting at 5600 random point locations (200/county) with the land use regression formula and kriging these predictions. For health analyses we would generate point predictions at the actual subject residential or work locations, but the visualization supplies useful information for assessing the validity of the predicted pollution surface.

To compare our LUR estimates to another method commonly used in health studies, we also performed kriging on the raw $PM_{2.5}$ values. All the statistical analysis was conducted using R statistical software and the GSTAT library (Pebesma, 2004; R Development Core Team, 2005).

We evaluated the quality of the predictions at validation locations by calculating the root mean squared error at validation locations. We assessed the quality of the final models (those generated after returning the validation locations to the pool of samples) by calculating the root mean squared error based on fitted values in a leave-one-out cross-validation.

3. Results

3.1. Descriptive statistics for $PM_{2.5}$

Three-year average $PM_{2.5}$ values for both the 28- and 9-county regions were approximately normally distributed (Fig. 2) with a mean for the 28-county region of $14.3\,\mu g\,m^{-3}$ (median = 14.3, standard deviation = 1.78) and for the 9-county region of $15.3\,\mu g\,m^{-3}$ (median = 15.1, standard deviation = 1.42). The highest concentrations are located in several different counties including New York (Manhattan), New York; Bergen and Hudson, New Jersey; and New Haven, Connecticut. Most of these locations are situated in proximity to major highways. Locations with higher concentrations tend to be located closer to Manhattan. Winter 2000 averages were also approximately normally distributed with a mean of $14.0\,\mu g\,m^{-3}$ (median = 14.3, standard deviation = 2.55) with 3 of the 4 highest concentrations located in Manhattan and one in New Haven, Connecticut. All statistics above are based on the full set of samples with no validation samples removed.

3.2. Regression model building and results

3.2.1. Models based on 3-yr average concentrations

For both the 28-county region and the 9-county region, the total traffic variables were the strongest predictors (see Table 2 for distributional statistics

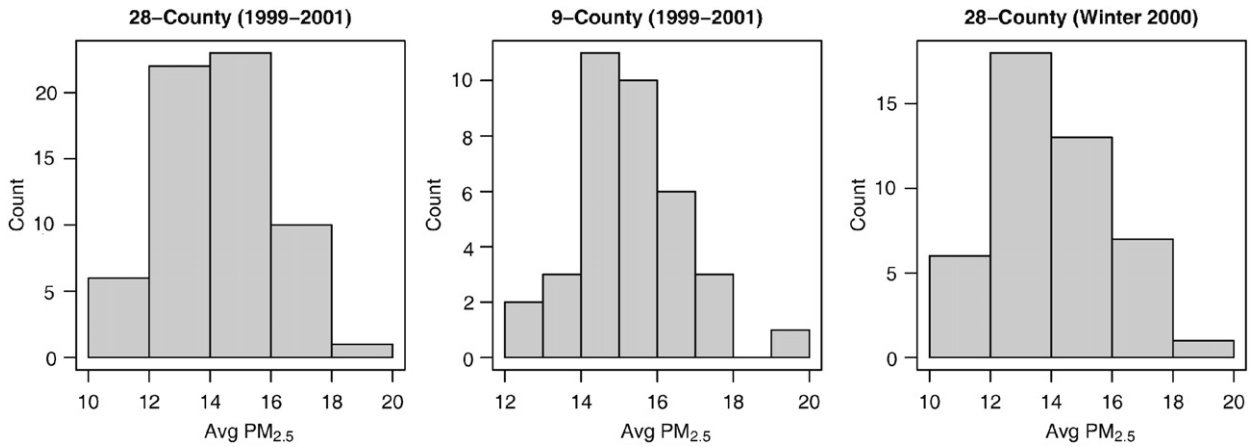

Fig. 2. Distributions of $PM_{2.5}$ concentrations.

Table 2
Distribution statistics for final predictor variables

	28-county (1999–2001)			9-county (1999–2001)			28-county (winter)		
	Traffic (500 m)	Population (1000 m)	Industrial land use (300 m)	Traffic (500 m)	Population (1000 m)	Industrial land use (300 m)	Traffic (300 m)	Population (1000 m)	Vegetation land use (1000 m)
Minimum	0.00	0.21	0.00	0.00	0.21	0.00	0.00	0.21	4.92
1st quartile	1.46	6.31	1.17	2.41	16.75	3.40	0.33	6.09	31.36
Median	4.55	14.53	9.23	5.38	36.47	7.87	1.07	11.19	96.66
Mean	6.41	27.52	14.26	8.03	42.60	12.85	2.21	22.97	125.90
3rd quartile	9.80	42.28	20.34	13.48	55.01	16.59	3.83	27.22	122.10
Maximum	24.06	119.40	52.70	24.06	119.40	45.77	8.35	119.40	697.10

Traffic Units: 1000s of vehicle $km\,h^{-1}$.
Population units: 1000s.
Land use units: Acres.

on final predictors). Total traffic in the 500 m buffer (traffic500) and total traffic in the 300 m buffer, in particular, led all other variables in explanatory power. Urbanization-related variables, primarily in the 500 and 1000 m buffers, are also strong predictors both with and without traffic500 in the model. These include total population in both the 500 and 1000 m buffers and both households and housing units in the 1000 m buffer.

With traffic in the 500 m buffer and the total population variable in the model, the next strongest predictor is industrial land use. Industrial land use in the 300 m buffer is strongest in both the 28-county and 9-county regions.

The final models for both the 28-county region and the 9-county region include traffic in the 500 m buffer, industrial land use in the 300 m buffer, and total population in the 1000 m buffer (Table 3).

The 28-county model predicted validation locations to within, on average, 0.93 $\mu g\,m^{-3}$ (6.5% of actual concentrations) with a root mean squared error of 1.10. The 9-county model predicts validation locations with an average absolute value residual of 0.77 $\mu g\,m^{-3}$ (5.0% of actual concentrations) with a root mean squared error of 0.87. There was some bias in the validation predictions, particularly in the 28-county model where the model was more likely to overpredict $PM_{2.5}$ values, though no bias is apparent in the models after including the validation samples (Fig. 3).

Within the 28-county model the predictions are better for those validation sites located in the more urbanized 9-county area. The 9 validation sites located within the 9 counties have a root mean squared error of 0.90 whereas the 4 sites located in the other 19 counties have a root mean squared error of 1.45. This distinction, however, is due primarily to a single large residual in Waterbury, CT at a site with the second highest industrial land use among all the 62 locations and relatively high traffic. Among the 49 modeling samples in the 28-county area, 27 were inside the 9 more urbanized counties and 22 outside.

Table 3
Final model results

	Value	SE	t	p	VIF
Model: 28 county (1999–2001)					
Intercept	12.273	0.261	46.965	0.000	—
Traffic (500 m)	0.121	0.027	4.530	0.000	1.344
Population (1000 m)	0.031	0.006	5.704	0.000	1.378
Industrial land use (300 m)	0.028	0.010	2.721	0.009	1.253
Multiple *R*-Squared	*0.642*				
Model *p*-value	*0.000*				
Model: 9-County (1999–2001)					
Intercept	13.171	0.364	36.232	0.000	—
Traffic (500 m)	0.098	0.025	3.967	0.000	1.196
Population (1000 m)	0.020	0.006	3.547	0.001	1.359
Industrial land use (300 m)	0.040	0.013	3.005	0.005	1.321
Multiple *R*-Squared	*0.617*				
Model *p*-value	*0.000*				
Model: 28-county (winter 2000)					
Intercept	12.841	0.509	25.214	0.000	—
Traffic (300 m)	0.463	0.106	4.370	0.000	1.106
Population (1000 m)	0.033	0.010	3.355	0.002	1.181
Vegetative land use (1000 m)	−0.005	0.002	−2.453	0.019	1.200
Multiple *R*-Squared	0.607				
Model *p*-value	0.000				

When the validation samples are returned to the pool of modeling samples and the models are re-run, the parameters remain similar with the greatest change occurring in the industrial land use variable. The average absolute value residuals based on the *fits* (note that this is based on the fitted values for observations included in the modeling rather than the prediction of validation locations) from the full models (no excluded samples) are $0.85\,\mu g\,m^{-3}$ or 6.0% of actual concentrations (28-county model) with a root mean squared error based on a leave-one out cross-validation of 1.15, and $0.69\,\mu g\,m^{-3}$ or 4.5% of actual concentrations (9-county model), root mean squared error of 1.00. The final 28-county model based on 3-yr averages explains 64% of the variation, and the 9-county model explains 62% of the variation. Plots of the fits on observed values are pictured in Fig. 4.

Although plots of the nine sampling locations with both FRM and TEOM monitors showed little or no bias (four locations had higher TEOM values and five had higher FRM values), we tested the inclusion of an indicator variable in the models discussed above to control for a possible effect. This indicator was not significant in any of the models and was excluded.

3.2.2. *Model based on winter 2000 concentrations*

The modeling of winter 2000 concentrations led to a very similar model as those based on 3-yr average $PM_{2.5}$. Length of road in the 1000 m buffer, amount of high density residential land use and traffic within the 1000 and 500 buffers are, singly, the top predictors of winter $PM_{2.5}$ after validation samples are removed. Although these variables appeared to predict well individually, collinearity

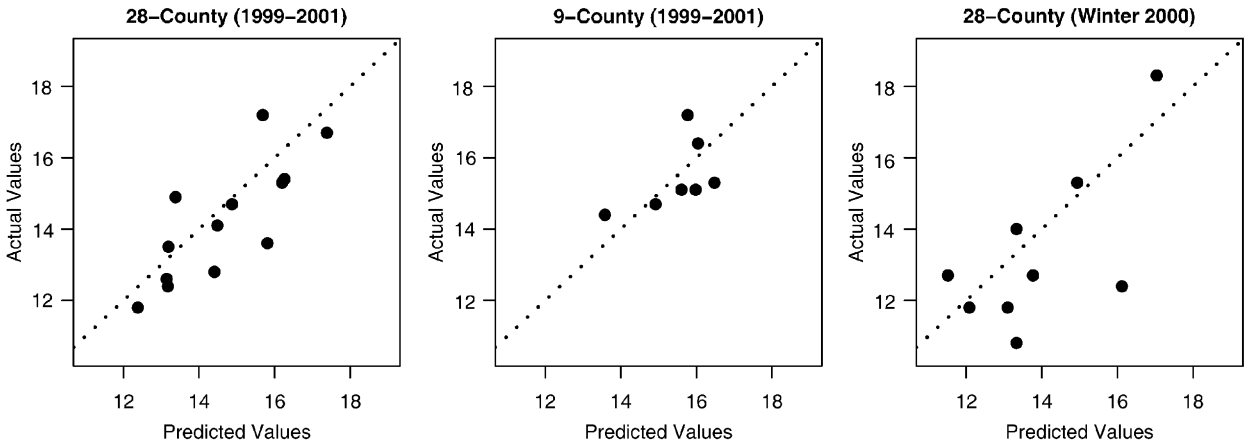

Fig. 3. Predicted vs. actual values for the validation samples.

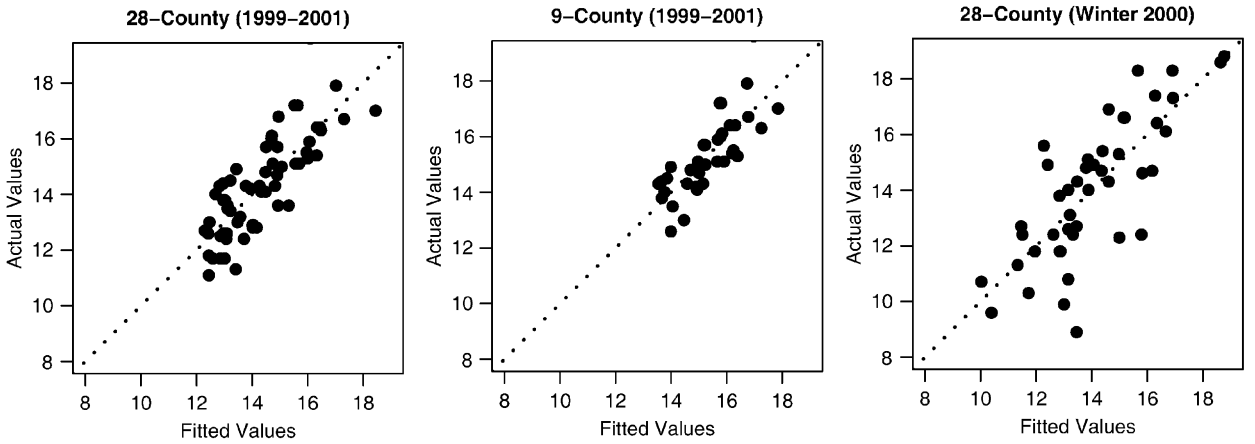

Fig. 4. Fitted vs. actual values for final models.

with other strong candidate predictors and highly influential observations pointed us to more robust variables with slightly less predictive power.

The final, three-variable model includes traffic within 300 m, total population within 1000 m and vegetative land use within 1000 m (Table 3). The inclusion of the vegetative land use variable is strongly influenced by two rural monitors in New Jersey. Removing these two points reduces the level of statistical significance to $p<0.15$. The parameter value, however, remains very similar (-0.00457 vs. -0.00460) with and without these points and we opted to retain the variable as it helps to account for rural–urban differences. The average absolute value residuals from the predictions at validation locations is $1.37\,\mu g\,m^{-3}$ (within an average of 11% of actual concentrations) with a root mean squared error of 1.72. These numbers are inflated by one large residual—an overprediction at a Brooklyn location with an unusually high population density. Without this one location, the average absolute value residual is $1.08\,\mu g\,m^{-3}$ (9% of actual) and a root mean squared error of 1.27.

For the winter model the predictions are, on average, better at sites outside of the nine urbanized counties with a root mean squared error of 1.30 for these five locations and 2.14 at four validation sites inside these urbanized counties, but the distinction between highly urbanized and less urbanized areas, again, is primarily a product of the single large residual mentioned above. Among the 36 modeling samples in the 28-county area, 19 were inside the nine more urbanized counties and 17 outside.

When the validation samples are returned to the pool of modeling samples and the models are re-run, the parameters remain similar with the greatest change occurring in vegetative land use (an approximate 15% change). The average absolute value residuals based on the *fits* from the full models (no excluded samples) are $1.19\,\mu g\,m^{-3}$ or, on average, 9% of actual concentrations with a root mean squared error based on a leave-one out cross-validation of 1.69. This final model explains 61% of the variation in winter 2000 $PM_{2.5}$ concentrations. Interpolated surfaces of land use regression predictions are shown in Figs. 5(a)–(c).

3.3. Bootstrap

The bootstrap results for all models show that the models are relatively stable to the choice of samples. We observe a slight bimodal shape in the industrial variable in the 9-county models. In all cases, alterations from normality are caused by a single sample (AIRS ID 340030004), located near Fort Lee, NJ, just 1.5 mile from Manhattan and extremely close to Interstate 95 and the George Washington Bridge. This one point had the highest $PM_{2.5}$ of all samples. We have no reason to believe that the underlying data for this station is inaccurate and the monitor was not excluded (Figs. 6(a)–(c)).

3.4. Spatial autocorrelation

As expected, we find that the $PM_{2.5}$ values are highly autocorrelated with Moran's I values of 0.52 ($p=0.001$), 0.27 ($p=0.003$) and 0.43 ($p=0.001$) for the 28-county, 9-county and winter models, respectively, using the Queen's contiguity matrix and 0.49 ($p=0.001$), 0.38 ($p=0.002$) and 0.37 ($p=0.002$) for the same models using a 3 nearest-neighbor approach. The spatial autocorrelation in our residuals from the full models was considerably diminished and generally nonsignificant suggesting that the models did not violate the independence assumption and that the included covariates account for the autocorrelation. The residuals produced Moran's I values of 0.21 ($p=0.003$), 0.09 ($p=0.108$) and 0.08 ($p=0.105$) for the 28-county, 9-county and winter models, respectively, using the Queen's contiguity matrix and 0.10 ($p=0.101$), 0.01 ($p=0.324$) and -0.03 ($p=0.496$) using the three nearest neighbors. Variograms of the raw $PM_{2.5}$ values show considerable spatial autocorrelation but the pattern is significantly diminished and barely visible in variograms of the residuals from the full models (variograms not shown).

3.5. Kriging of raw $PM_{2.5}$ values

Despite the limited number of samples (49, 29 and 36 for the 28-county-3-yr, 9-county-3-yr and 28-county, winter-2000 models, respectively), kriging based on exponential models performed surprisingly well—outperforming land use regression at most locations. For the 28-county model using 3-yr averages we calculated an average absolute value residual of 0.68 and a root mean squared error on these predictions of 0.90 (compared with 1.10 for LUR). For the 9-county model the mean absolute value residual was 0.48, with a root mean squared error of 0.61 (compared with 0.87 for LUR). And kriging the winter 2000 $PM_{2.5}$ resulted in an average

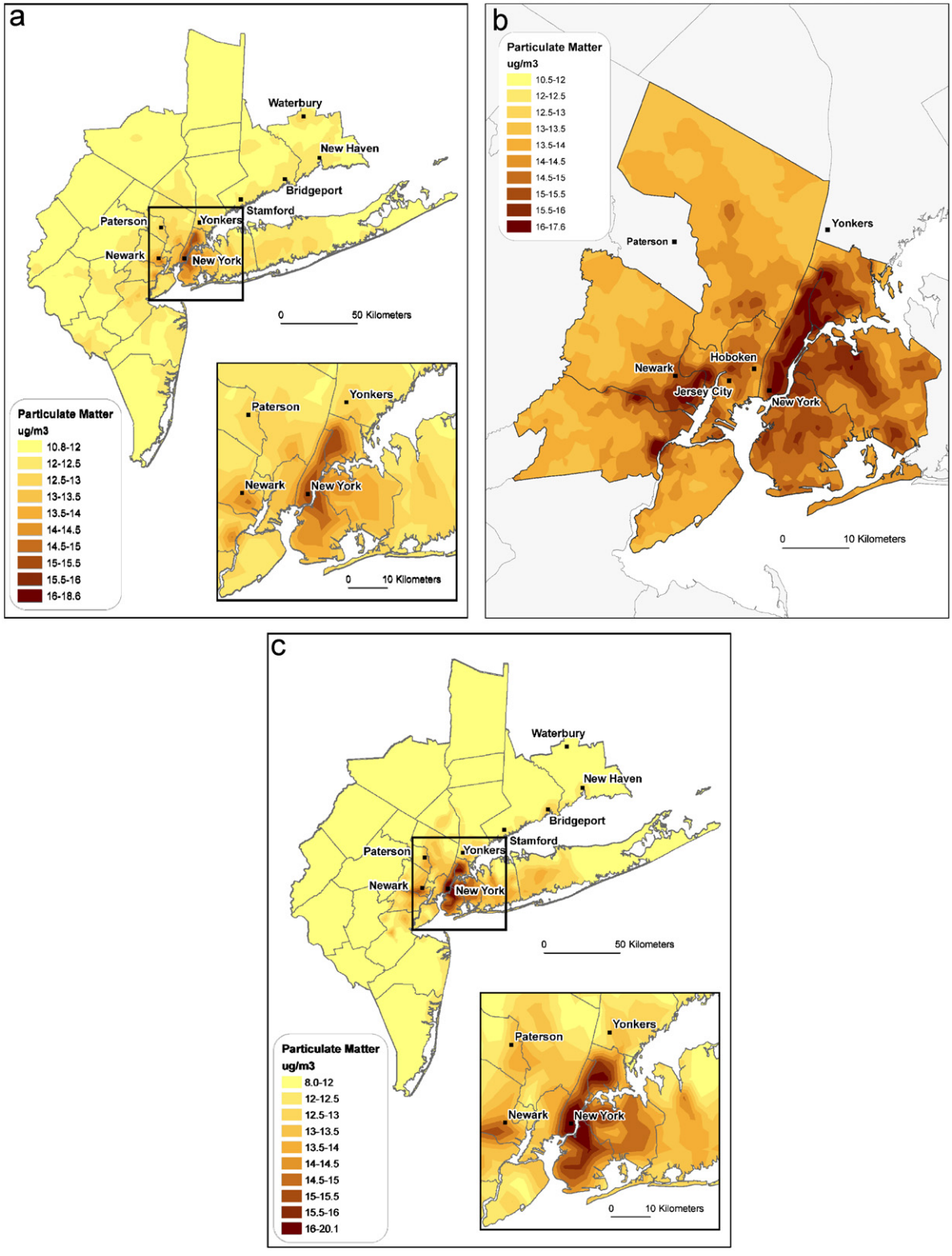

Fig. 5. (a) Interpolated LUR predictions: 28-county (1999–2001). (b) Interpolated LUR predictions: 9-county (1999–2001). (c) Interpolated LUR predictions: 28-county (winter 2000).

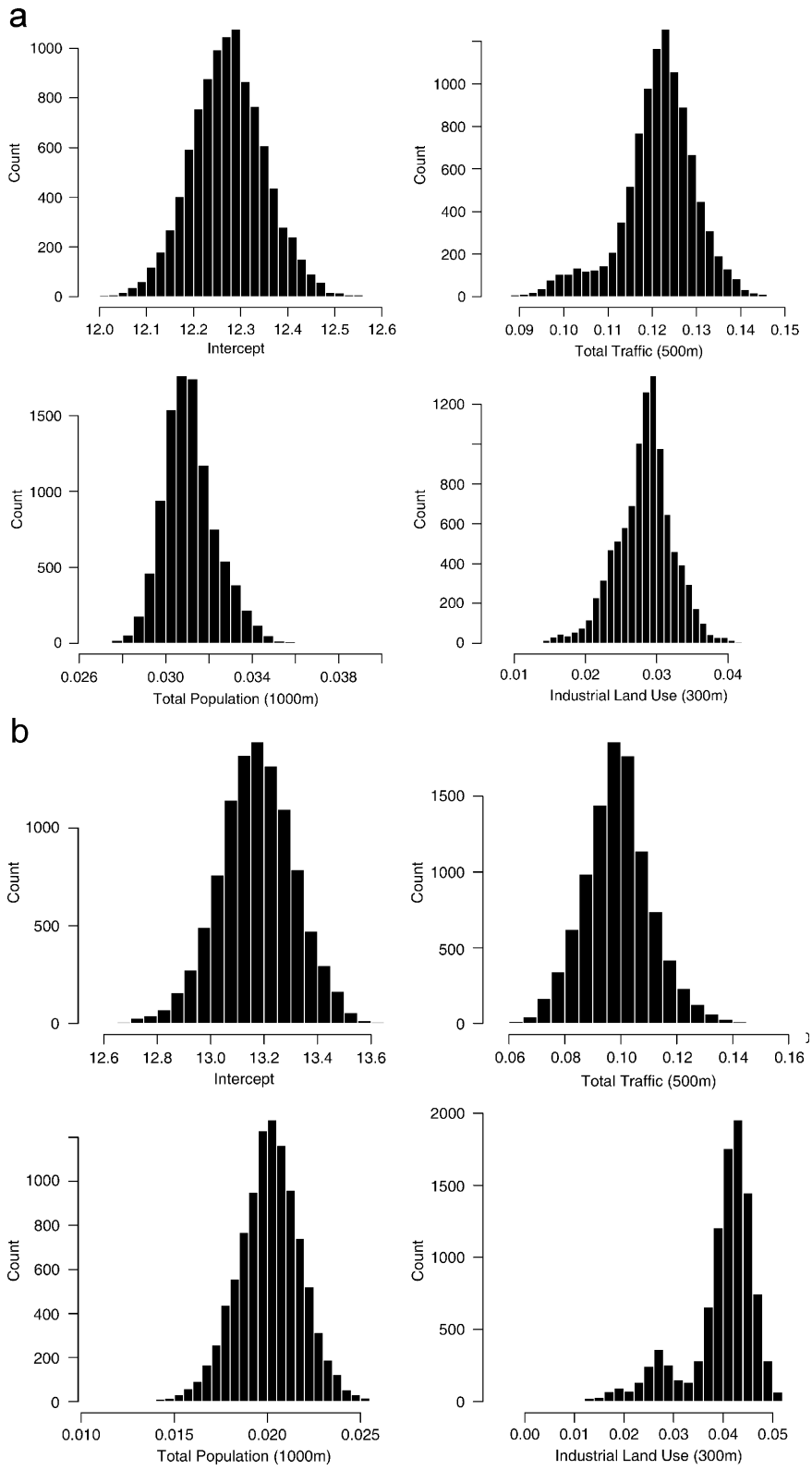

Fig. 6. (a) Bootstrap of final parameters, 28-county (1999–2001). (b) Bootstrap of final parameters, 9-county (1999–2001). (c) Bootstrap of final parameters, 28-county (winter, 2000).

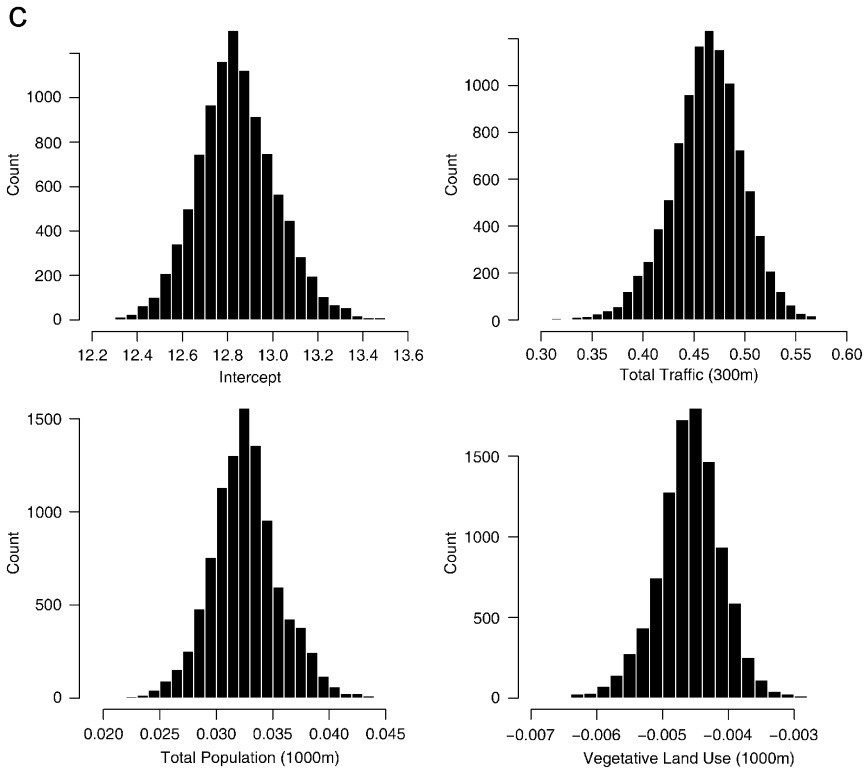

Fig. 6. (*Continued*)

absolute value residual of 1.39 and a root mean squared error of 1.55 (compared with 1.72 for kriging).

Although predicting at validation locations based on kriging the *modeling* samples performed well, land use regression outperforms kriging based on cross-validation results using the full set of samples. The root mean squared errors of cross-validation predictions, for example, are 1.15, 1.00 and 1.69 for the land use regression compared with 1.30, 1.47 and 2.04 for the kriging (28-county-3-yr, 9-county-3-yr and 28-county-winter-2000 models, respectively).

4. Discussion

We developed land use regression models for predicting $PM_{2.5}$ in New York City and surrounding counties using a combination of traffic, land use and census variables. These models explained more than 60% of the variability in $PM_{2.5}$ concentrations and predicted validation locations well with predictions at validation locations generally within 10% of actual values.

All models included traffic within the 500 m buffer or 300 m buffer and total population within 1000 m. The 28-county and 9-county models based on 3-yr averages also included an industrial land use variable (300 m) while the 28-county model based on winter 2000 concentrations included a variable representing vegetative land use (1000 m). There was little difference between the 28-county models based on winter 2000 concentrations and a 3-yr average. The relative strength of a smaller buffer for traffic (300 vs. 500 m) in the winter model does potentially suggest a stronger local influence but otherwise, the difference in regional and local contributions does not appear to strongly influence the model.

Given the localized impact of traffic, we would have expected variables in smaller buffers to be the strongest predictors. Although variables in smaller buffers were indeed good predictors (traffic within 100 m, for example, explains approximately 25% of variation in 3-yr, 28-county concentrations), they did not perform as well as variables in larger buffers. This is likely attributable to a number of factors. As a pollutant with both a regional and local contributions it is possible that strong

predictor variables are able to explain some of both components. Predictors based on relatively small buffers may also perform better when modeling highly concentrated samples (which was not the case in this analysis). Finally, as demonstrated through European studies (Hochadel et al., 2006) $PM_{2.5}$ varies more gradually over space than elemental carbon and as such the strength of the larger buffers may reflect this larger area variation in the pollutant.

In general, these variables were robust to sample selection, though we found that the industrial land use variable was affected by the inclusion or exclusion of a New Jersey sample that had the highest $PM_{2.5}$ concentrations. Nevertheless, the fact that industrial land use (300 m) in the five New York City boroughs—using an entirely different GIS layer and without the influence of the New Jersey sample—was also significant lends support to the important role played by industrial land use.

The 28-county model without validation samples based on 3-yr averages exhibited some bias, over-predicting 10 values while under-predicting 3. The bias is less evident, but still exists, in the full model that includes the validation samples (34 are over-predicted and 28 are underpredicted). The under-prediction appears to occur in the more urbanized New Jersey counties and overprediction appears to occur in areas distant from New York City and, to a lesser extent, in eastern New York City and Long Island. The bias does not exist in the 9-county model. Although slight bias is present, the models performed well overall on cross-validation.

While the root mean squared errors for the 3-yr average models and the winter 2000 model suggests that predictions at validation locations are more precise for the 3-yr averages compared to the winter values, the RMSE disguises differences in variance and spread. The winter 2000 $PM_{2.5}$ values have approximately twice the variance and a 24% greater inter-quartile range than the 3-yr, 28-county $PM_{2.5}$ values. The mean absolute percentage error (MAPE), may more appropriately allow comparison and reveals that the predictions are more similar in percentage terms. The MAPE for the predictions at validation locations is 6.5% for the 3-yr, 28-county model and 6.2% for the winter 2000, 28-county model (and 5.0% for the 3-yr, 9 county model).

We found that without the validation samples (about 20% of the total) kriging outperformed land use regression, but when these samples are returned to the pool of modeling samples land use regression performs better. As ordinary kriging employs only data on the variable of interest, it is not vulnerable to unusual values in potential predictor variables. These unusual values in traffic, for example, could inflate (or deflate) land use regression parameters particularly in models based on a limited number of observations. At the same time reliance of kriging on raw monitoring data limits its potential for capturing small area variation such as an intersection of major highways. Kriging, as a result, may "oversmooth" and is likely to miss areas with unusually high particulate matter. It is also possible that kriging in this context violates the fundamental assumption of stationarity. Points close to high-traffic freeways, for example, may indeed exhibit a different relationship between $PM_{2.5}$ and distance than points in, for example, rural areas far from any major source of $PM_{2.5}$.

Particulate matter exhibits strong seasonal and diurnal patterns in New York City with higher concentrations in the summer months and higher concentrations during the morning (6 am–9 am) and late evening (5 pm–10 pm) (DeGaetano and Doherty, 2004). Although this analysis makes use of particulate matter concentrations averaged over time, future analyses may wish to consider these variations and develop, for instance, daily or seasonal maps that could highlight short-term variability (cf. Christakos and Serre, 2003).

As mentioned in the methods section, incomplete data provided by the NYMTC precluded our use of traffic data for all time periods. It is likely that models based on average daily traffic rather than average daily afternoon rush hour traffic might lead to different parameters. It is also possible that an analysis such as ours that relies on rush hour traffic could lead to a wider spread of $PM_{2.5}$ predictions, as we might expect high traffic areas in off-peak times to have a significantly larger increase in traffic during rush hour. Nevertheless, we would expect relative rates of traffic for different areas to remain very similar. We found, for example, the Spearman rank correlation coefficient for evening rush hour traffic compared with nighttime (8 pm–6 am) traffic for the same road segments is 0.93 suggesting that the results would likely have been similar had we used total average daily traffic.

These models demonstrate that $PM_{2.5}$ can be predicted using land use regression in a North American context. Using a combination of traffic, census and land use variables we were able to

predict more than 60% of the variation in $PM_{2.5}$ over a wide area and predict well at validation locations. Although each of the three models discussed above are not identical in variables or parameters, their similarity reinforces the relationship between $PM_{2.5}$ and land use variables nearby. Given the strong predictive power of both land use regression and kriging, we also applied kriging with external drift, a technique that combines these two methods, to the data. The limited residual autocorrelation from the land use regression, however, precluded adequate variograms fitting and predictions were not improved using this technique.

These models and land use regression models in general, hold particular promise in epidemiological settings where small area variations can be associated with significant health effects. While these models appear to predict well, data limitations hampered our ability to investigate all potentially useful predictors. In particular, further research is needed on the possible effects of street canyons and of seasonal variations in $PM_{2.5}$.

Acknowledgments

We thank anonymous reviewers for their helpful suggestions. We also thank Dr. John Mbwana, Cornell University, for help with traffic data analysis; Michael Chiume, Sangeeta Bhowmick and Alan Gershowitz, NYMTC, for traffic data preparation. Funding sources: Health Effects Institute, the National Institute of Environmental Health Sciences (Grants 5 P30 ES07048 5P01ES011627), Environmental Protection Agency STAR (RD-83184501-0), the Verna Richter Chair in Cancer Research, and the NSERC/SSHRC/McLaughlin Chair in Population Health Risk Assessment at the University of Ottawa.

References

Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Wiley, New York, NY.

Bari, A., Ferraro, V., Wilson, L., Luttinger, D., Husain, L., 2003. Measurement of gaseous HONO, HNO_3, SO_2, HCl, NH_3, particulate sulfate and PM2.5 in New York, NY. Atmospheric Environment 37, 2825–2835.

Brauer, M., Hoek, G., Val Vliet, P., Meliefste, K., Fischer, P., Brunekreef, B., 2002. Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children. American Journal of Respiratory and Critical Care Medicine 166, 1092–1098.

Brauer, M., Hoek, G., van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys, J., Bellander, T., Lewne, M., Brunekreef, B., 2003. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. Epidemiology 14, 228–239.

Briggs, D., deHoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., Smallbone, K., 2000. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Science of the Total Environment 253 (1–3), 151–167.

Christakos, G., Serre, M.L., 2003. Efficient mapping of California mortality fields at different spatial scales. Journal of Exposure Analysis and Environmental Epidemiology 13, 120–133.

DeGaetano, A., Doherty, O., 2004. Temporal, spatial and meteorological variations in hourly $PM_{2.5}$ concentration extremes in New York City. Atmospheric Environment 38, 1547–1558.

Gilbert, N.L., Goldberg, M.S., Beckerman, B., Brook, J.R., Jerrett, M., 2005. Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land-use regression model. Journal of the Air and Waste Management Association 55, 1059–1063.

Hochadel, M., Heinrich, J., Gehring, U., Morgenstern, V., Kuhlbusch, T., Link, E., Wichmann, H.E., Krämer, U., 2006. Predicting long-term average concentrations of traffic related air pollutants using GIS-based information. Atmospheric Environment 40, 542–553.

Hoek, G., Brunekreef, B., Goldbohm, S., Fischer, P., van den Brandt, P.A., 2002. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. Lancet 360, 1203–1209.

Jerrett, M., Burnett, R.T., Ma, R.J., Pope, C.A., Krewski, D., Newbold, K.B., Thurston, G., Shi, Y.L., Finkelstein, N., Calle, E.E., Thun, M.J., 2005a. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology 16 (6), 727–736.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C., 2005b. A review and evaluation of intraurban air pollution exposure models. Journal of Exposure Analysis and Environmental Epidemiology 15, 185–204.

Künzli, N., Jerrett, M., Mack, WJ., Beckerman, B., LaBree, L., Gilliland, F., Thomas, D., Peters, J., Hodis, H.N., 2005. Ambient air pollution and atherosclerosis in Los Angeles. Environmental Health Perspectives 113, 201–206.

Nafstad, P., Haheim, L.L., Oftedal, B., Gram, F., Holme, I., Hjermann, I., Leren, P., 2003. Lung cancer and air pollution: a 27 year follow up of 16 209 Norwegian men. Thorax 58, 1071–1076.

Parsons Brinckerhoff Quade & Douglas, Inc., 2005. Transportation Models and Data Initiative: General Final Report, New York Best Practice Model. January 30, New York.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers and Geosciences 30, 683–691.

Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Journal of the American Medical Association 287, 1132–1141.

R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ross, Z., English, P., Jerrett, M., Scalf, R., Gunier, R.B., Smorodinsky, S., Wall, S., 2006. Nitrogen dioxide prediction in southern California using land use regression modeling: potential for environmental health analyses. Journal of Exposure Science and Environmental Epidemiology 16 (2), 106–114.

US Census Bureau., 2000. Census 2000, Summary File 1 (SF 1) and Summary File 3 (SF 3). Using American Factfinder, vol. 2005.