Published in final edited form as:
Environ Res. 2009 August ; 109(6): . doi:10.1016/j.envres.2009.06.001.

Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy

Jason G. Su[a], **Michael Jerrett**[a,*], **Bernardo Beckerman**[a], **Michelle Wilhelm**[b,c], **Jo Kay Ghosh**[b], and **Beate Ritz**[b,c]

[a]Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, USA

[b]Department of Epidemiology, School of Public Health, University of California, Los Angeles, USA

[c]Center for Occupational and Environmental Health, School of Public Health, University of California, Los Angeles, USA

Abstract

Land use regression (LUR) has emerged as an effective means of estimating exposure to air pollution in epidemiological studies. We created the first LUR models of nitric oxide (NO), nitrogen dioxide (NO_2) and nitrogen oxides (NO_x) for the complex megalopolis of Los Angeles (LA), California. Two-hundred and one sampling sites (the largest sampling design to date for LUR estimation) for two seasons were selected using a location-allocation algorithm that maximized the potential variability in measured pollutant concentrations and represented populations in the health study. Traffic volumes, truck routes and road networks, land use data, satellite-derived vegetation greenness and soil brightness, and truck route slope gradients were used for predicting NO_x concentrations. A novel model selection strategy known as “ADDRESS” (A Distance Decay REgression Selection Strategy) was used to select optimized buffer distances for potential predictor variables and maximize model performance.

Final regression models explained 81%, 86% and 85% of the variance in measured NO, NO_2 and NO_x concentrations, respectively. Cross-validation analyses suggested a prediction accuracy of 87–91%. Remote sensing-derived variables were significantly correlated with NO_x concentrations, suggesting these data are useful surrogates for modeling traffic-related pollution when certain land use data are unavailable. Our study also demonstrated that reactive pollutants such as NO and NO_2 could have high spatial extents of influence (e.g., > 5000 m from expressway) and high background concentrations in certain geographic areas. This paper represents the first attempt to model traffic-related air pollutants at a fine scale within such a complex and large urban region.

Keywords

Nitrogen oxides; Air pollution; Traffic; Land use regression; GIS; Remote sensing; Los Angeles

[*]Corresponding author. Fax: +15106425815., jerrett@berkeley.edu (M. Jerrett).

Appendix A. Supporting Information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.envres.2009.06.001.

1. Introduction

With a population of 16.7 million, the Los Angeles Metropolitan Area (hereafter referred to as LA) is the largest urban area in the state of California and the second-largest in the United States. LA is also consistently ranked as one of the most polluted metropolitan areas in the US, partially due to heavy reliance on automobiles for transportation, and because more than 40% of all goods transported into the United States move through the Long Beach/Los Angeles port complex, which generates thousands of truck trips per day (Hricko, 2008). The LA Basin is also susceptible to atmospheric inversions (Bytnerowicz and Fenn, 1996), which trap exhaust from on-road vehicles, airplanes, and locomotives, as well as from shipping, manufacturing and other industrial sources in the troposphere. Additionally, with an average rainfall of only 381 mm/year, Los Angeles experiences minimal pollution removal by precipitation. The high level of auto-dependence in LA is replicating in other major conurbations around the world, particularly in newly industrialized countries (Walsh, 2008). Lessons learned from studying air pollution exposures and subsequent health effects in Los Angeles may therefore have widespread applicability to other large urban areas.

This paper reports the first attempt to model NO, NO_2 and NO_x concentrations in LA (Fig. 1) using a land use regression (LUR) approach. The LUR was developed as part of a study sponsored by the California Air Resources Board (CARB) to examine the impacts of outdoor air pollution on respiratory health in children living in LA. The LUR method seeks to predict pollution concentrations at a given site based on surrounding land use, road network, traffic, physical environment and population characteristics using a series of buffers (Su et al., 2009). Geospatial modeling techniques including LUR are an attractive alternative to ambient (government) air monitoring data for assessing traffic pollutant exposures, since they can be applied to large populations and account for neighborhood-scale variations in pollutant concentrations. To our knowledge, LUR to date has not been applied to large urban areas with population over 10 million for traffic-related air pollution (NO_x) but only to estimate more spatially homogenous air pollutant concentrations such as fine particulate matter (Moore et al., 2006; Ross et al., 2007).

2. Materials and methods

2.1. Sampling location determination

We selected neighborhood monitoring locations ($n = 201$) using a location-allocation algorithm (Kanaroglou et al., 2005) that took into account variability in traffic pollution and the spatial distribution of our childhood respiratory health study population, specifically, participants in the Los Angeles Family and Neighborhood Study (LA FANS). The estimation domain for locating optimal sampling sites covered more than 10,000 km^2 of LA. Briefly, the location-allocation algorithm involves a two-step algorithm that (1) builds a demand surface of spatial variation (i.e., semi-variance) and (2) solves a constrained spatial optimization problem to determine locations for a pre-specified number of samplers.

The demand surface was created using two criteria: first, samplers should be placed where the pollution surface is expected to exhibit high spatial variability, and second, population density of subjects in our health study should be relatively high. To create an initial pollution surface across the LA metropolitan area, an LUR model was applied to predict NO_2 concentrations by adapting the LA land use and transportation data to the regression coefficients previously derived for the San Diego area (Ross et al., 2006). Given a first estimate of the pollution surface, spatial variability of pollution $Z(x, h)$ at location x with a distance h is determined by the following adaptation of the semivariogram equation (Cressie, 1993):

$$\gamma(x,h)=\frac{1}{2N}\sum_{i=1}^{n}(Z(x)-Z(x+h_i))^2 \quad (1)$$

This creates the demand surface that satisfies the first criterion noted above. To satisfy the second criterion, we appropriately modified the demand surface achieved through Eq. (1) by intensifying the demand for pollution monitors in areas with high densities of populations. To achieve this effect, a weighting scheme is implemented according to

$$W_R=\frac{P_R/P_T}{\widehat{\gamma}_R/\widehat{\gamma}_T} \quad (2)$$

where P_R is the population of interest in region R within the study area and P_T is the population for the entire study area.

The task was to place all monitors within 500 m of the census tracts (CT) of residences included in LA FANS and, at the same time, in areas with the most spatial variation in traffic-related air pollution. Thus, CTs that included LA FANS residences were buffered to a distance of 500 m and the eligible census blocks were assigned a population-weighted semi-variance value based on the specific population counts from LA FANS. The CTs were widely dispersed throughout the urbanized area of LA. Irregular lattice points were created using street block centroids for locating sampling sites. Finally, monitoring locations were selected using a maximum attendance location-allocation algorithm Eq. (3) based on the population-weighted semi-variance.

$$\sum_{i=1}^{k}\sum_{j=1}^{m}w_i(1-bd_{ij})x_{ij} \quad (3)$$

where k is the number of demand locations and m is the number of candidate locations. In our case, k = 201, including 15 co-located sites with the governmental monitoring stations. The weight w_i at location i represents demand, while d_{ij} is the distance between locations i and j. X_{ij}, is the allocation decision variable attaining the value of 1 if demand location i is served by a station in j and 0 otherwise. Attendance linearly decreases with distance at the rate of parameter b, a value determined by the maximum distance of influence.

2.2. Sampler preparation and field deployment

Each air sampler was loaded with two pre-coated collection pads—one pad to measure NO_2 (part number PS-134) and the other for NO_x (part number PS-124). NO concentrations were derived as the difference between NO_2 and NO_x concentrations. To protect against rain damage, each sampler was placed inside a plastic shelter that shielded the sampler from the top but allowed air to flow freely through the bottom. The shelter was constructed from a 4-in diameter PVC plumbing cap, with eye bolts on the sides and top of the shelter, which were used to secure the shelter to the mounting locations. This was the same shelter methodology utilized in the East Bay Children's Respiratory Health Study (Singer et al., 2004).

Field staff followed a strict protocol for installing and removing the samplers in the field. We identified the exact location of installation for each sampler, using GPS readings, text descriptions and photographs of the exact pole, fence or tree on which the sampler was to be hung. Samplers were installed approximately 8–10 ft above the ground level (the height of each sampler was recorded on the log sheet). Duplicate samplers were installed either side-by-side or on opposite sides of the pole ("back-to-back"), ensuring that duplicates were hung at the same height. During installation and collection efforts, two additional GPS measurements were taken at each location (a total of four measurements per site, per season). The majority of neighborhood sites where the samplers were installed were within 50 m of the exact location selected by the location-allocation model; all sites were within 200 m of the selected location.

2.3. Pollution sampling

We used passive air samplers from Ogawa & Company USA, Inc. (Pompano Beach, FL) to conduct monitoring in two seasons during September 2006 (late summer warm season) and February 2007 (mid-winter rainy season). Each air sampler measuring concentrations of NO_2 and NO_x was deployed for a 2-week period (maximum variability was ± 2 h). NO was derived as the difference between NO_2 and NO_x. Measurement periods were specifically selected after reviewing historical monitoring data to characterize long-term average concentrations most accurately. In total, from the 201 samplers deployed in each season measurements were obtained for 183 sites in September 2006 and for 181 sites in February 2007 (some monitors were stolen or vandalized, and some sites were relocated in February 2007 because access to the original site was unavailable). This left 181 sites (with monitoring data and not relocated) available for analyses that averaged concentration data from both seasons. In addition to these 181 samplers, we co-located 15 samplers at Southern California Air Quality Management District (SCAQMD) air monitoring stations, deployed 50 duplicates, and collected data from 30 field blanks.

2.4. Traffic data

Three types of roadway configuration and traffic volume data were analyzed for their ability to predict traffic-related pollution. They included Dynamap data from TeleAtlas (Global Crossroads, Boston, MA), Highway Performance Monitoring System (HPMS) data from the National Transportation Atlas Database and Metropolitan Planning Organization (MPO) data from the Southern California Association of Governments (SCAG). Since each of these data sources has some limitations and because traffic was of great importance for the exposure contrasts, extensive efforts were made to derive the most comprehensive traffic data available. Traffic data imputed from the Dynamap data are discussed below. A description of the HPMS data, the traffic data conflated from HPMS to Dynamap and the MPO data is included in Supplementary Information.

We used Teleatlas' Dynamap 2000 as our base roadway configuration and traffic volume data because the underlying road network had the most accurate spatial representation when compared to digital orthophotos. The Dynamap data were combined into a mosaic from individual county files with repeated road segments removed before the analysis. The complete Dynamap physical coverage provided traffic volumes (i.e., 24-h annual average daily traffic (AADT) counts) for 2.5% of the road network for LA (18 504 out of 740 047 roadway segments) during the period from 1987 to 2005 (Table 1). The median AADT value from measured road segments within a road category (e.g., highway with or without limited access) was assigned to road segments of the same category; i.e., to impute traffic data to road segments without measurements. The circular area distances (buffers) we chose for LUR model development ran from 50 to 5000 m at an interval of 100 m. Such large buffer sizes were selected because previous studies in LA indicated influence from land use over this extended spatial range (Moore et al., 2006). Buffer statistics included total vehicle miles traveled (VMT) (count*km) for (1) highways (including primary roads with limited access or interstate highways (A1) and primary roads without limited access or state highways (A2)); (2) major roads (i.e., secondary and connecting roads (A3)); (3) highways and major roads (A1+A2+A3); and (4) all roads (A1+A2+ •+A7, A4 = local roads, A5 =

one-way vehicle dirt trails, A6 = road ramps, A7 — bicycle or pedestrian trails) (Table 1). Within a circular distance of *j* of sampler *i*, total vehicle miles traveled T_{ij}^{v} was estimated by summing all (*k*) traffic volumes ($V_{i,j,k}$) of a road segment (*l*) within that search distance:

$$T_{ij}^{v}=\sum_{k=1}^{m}(V_{i,j,k}l_{k}) \quad (4)$$

VMT estimated in Eq. (4) thus include statistics for highways, major roads, highways+major roads and all road traffic categories.

2.5. Road network and slope gradient

We used road network (including highway, major and local roads) from Dynamap as surrogates for traffic-related pollution. The total length L_{ij}^{c}(m) of all road segments *(k)* of road category c within a circular search distance *j* of sampler *i* was estimated by

$$L_{ij}^{c}=\sum_{k=1}^{n}L_{i,j,k}^{c} \quad (5)$$

Slope of a truck route was defined as an angle in degrees. We first converted the Dynamap roadway network into raster cells and assigned each raster cell a slope derived from a digital elevation model (DEM) produced by the US Geological Survey (USGS, 1999). The average slope M_{ij} of all the truck route segments *(k)* within a circular search distance *j* of sampler *i* was estimated by

$$M_{ij}=\frac{1}{n}\sum_{k=1}^{n}m_{i,j,k} \quad (6)$$

We also included distance to truck routes as a potential explanatory variable during model selection. Truck routes were extracted from HPMS data for 2007 and straight distances to truck routes were created for the study region. Also, the number of major road intersections inside each circular buffer area was calculated to identify whether areas with more intersections have higher NO_x concentrations.

2.6. Tasseled-cap transformation

Current LUR models use road network information as a surrogate for levels of traffic-related pollution; however, influences from some land use types such as parking lots, which have similar spectral reflectance as roads, are usually unavailable in road network data and thus, unaccounted for in LUR models. In addition, most land use variables such as industrial, commercial and open land use applied in LUR models to date were originally derived from remote sensed data. The most comprehensive, high-resolution (six bands finer than 30 m) and freely available global coverage remote sensing data are Landsat Enhanced Thematic Mapper Plus (ETM+) data. Because of the complexities involved with display and extraction of information contained in the Landsat ETM+ data, a tasseled-cap transformation was used to reduce the number of channels to be considered, and to provide a more direct association between signal response and physical processes on the ground (Crist and Cicone, 1984). Analyses of simulated and actual TM data have revealed that vegetation and soils data in the six reflective TM bands (excluding the thermal band) primarily occupy three dimensions. Within these three dimensions two planes are defined, which are occupied by fully vegetated and bare soil samples, along with a “transition zone” between the two which is occupied by partially vegetated samples. The three features corresponding to the three data dimensions are named brightness, greenness and wetness. Brightness, a weighted sum of all six bands, is a measure of overall reflectance (e.g., differentiating light from dark soils). Greenness is a contrast between near-infrared and visible reflectance, and is thus a measure of the presence and density of green vegetation (Crist and Cicone, 1984).

Tasseled-cap indices for LA were derived from the ETM+ data collected from a nominal altitude of 705 km in a near-polar, near-circular, sun-synchronous orbit at an inclination of 98.2°, imaging the same 183-km swath of the Earth's surface every 16 days (http://landsat.gsfc.nasa.gov/). The ETM+ imagery we acquired included three visible (resolution 30 m), three infrared (30 m), two thermal (60 m) and a panchromatic (15 m) band. The scenes for LA were at path 41/row 36,41/37 and 40/37, all captured on June 21, 2001. These images were orthorectified by the United States Geological Survey (USGS) and projected to Universal Transverse Mercator (UTM) zone 10N coordinate system with a WGS84 datum (World Geodetic System of 1984).

Orthorectification is the process by which the geometric distortions of the image are modeled and accounted for, resulting in a planimetricly correct image. Because the Earth is in 3D while most sensors are in 2D, orthorectification corrects for many of the anomalies resulting from this conversion. The success of the orthorectification process depends on the accuracy of the digital elevation map and the correction formulae (Bolstad, 2005). Because the root-mean-square error is less than 30 m, the EMT+ data were not atmospherically and topographically corrected (GLCF, 2009).

Greenness is independent of brightness and increases with increasing proportions of green vegetation. Thus, greenness might be a better, though inversely related, surrogate for the degree of influence of traffic or lack of stationary industrial sources, and a better marker for variables such as open space used in earlier studies.

2.7. Land use characteristics

Land use data for LA were acquired from the Southern California Association of Governments for the year 2000. Major land use types included commercial, residential, industrial and open land use. The total area A_{ij} of a land use type within a circular buffer search distance *j* of sampler *i* was estimated by summing over all (*k*) areas of a land use type inside the buffer:

$$A_{ij}=\sum_{k=1}^{m}(S_{i,j,k}) \quad (7)$$

Additionally, physical geographic variables such as distance to coast, elevation and coordinates of latitude and longitude were calculated for each sampler location and used as covariates for our ADDRESS (A Distance Decay REgression Selection Strategy) modeling process.

2.8. Model selection and diagnostics

For the model selection process we used the ADDRESS selection strategy (Su et al., 2009) based on series of distance decay curves created by correlation coefficients of spatial covariates with the measured (first step) or the predicted residuals (from the second step onward) of exposure concentrations. Guided by the visual distance decay curves of correlation coefficients, with ADDRESS we identified the optimized distances of influence for all variables during the model selection process. Using a forward manual stepwise procedure, a spatial covariate with the highest correlation at each search distance was added to the LUR model based on its statistical significance level (*p* = 0.05). The variable selection process continued until no further spatial covariates were selected for inclusion in the model. Because primary and secondary highways and major roads are densely distributed throughout LA, any subject's exposure at their residence is influenced not only by the nearest roadways and traffic but likely also by urban-scale traffic patterns that vary over ranges of 5 km or more. Thus, the maximum distance of a buffer was set to 5000 m. However, if a distance decay curve still showed an upward trend at 5000 m, the maximum buffer distance was further extended until a downward trend was identified. In an urban environment, correlations might not be zero even at very high buffer distances because of the influence of background pollutant concentrations; however, we expected to see a decrease in the influence of certain emission sources, such as emissions from a roadway, after a large enough distance.

To test the efficacy of the prediction models, model diagnostics included (1) evaluating whether selected variables were collinear based on variance inflation factors (VIFs); (2) outliers assessment, i.e., determining Cook's distances to assess whether a single observation changed the regression estimates; (3) examining whether spatial autocorrelations of the prediction residuals in our final optimized models existed based on the Moran's *I* statistic (Bailey and Gatrell, 1995); (4) assessing whether predictions satisfy the US EPA (Environmental Protection Agency) requirements of a prediction model (US EPA, 1991) by adopting normalized mean bias (NMB) (see Eq. (5)) and normalized mean error (NME) (see Eq. (6)) tests below; (5) conducting a Chow (1960) test to assess whether large sample sites would benefit LUR modeling results; and (6) applying cross-validation techniques to 16 random samples for model reliability tests.

$$NMB=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{C_i^{*}-C_i}{C_i}\right)\times 100\% \quad (8)$$

$$NME=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{|C_i^{*}-C_i|}{C_i}\right)\times 100\% \quad (9)$$

where C_i^{*} and C_{i} refer to the predicted and observed pollutant concentrations at sampler *i*.

Moran's *I* statistics were conducted on a first- and second-order Rook contiguity matrix to the Thiessen polygons created from the sampler sites. Statistical significance was tested using a permutation test with 999 iterations. Although a location-allocation algorithm was applied to locate samplers, the samplers were also restricted to residential census tracts of subjects in the LA FANS health study. If samplers were clustered within census tracts of residence, it would be difficult to remove the near-range autocorrelation with fixed effects or with standard autoregressive techniques. Therefore all models were refit in STATA 8.0 adding a cluster parameter (based on census tract ID) to our ADDRESS models and also using generalized estimating equations (GEE) (Liang and Zeger, 1986) clustered on the census tracts where the monitors were located. Sensitivity analyses were also done with robust standard error estimation using census tract as the cluster unit.

The EPA's suggested performance criterion for NMB is ± 5% to ±15%, and for NME 30–35% of pollutant levels greater than 60 ppb (C_i≥60 ppb) (US EPA, 1991). For the Chow test, half of the total sites available were randomly selected from each of four quartile groups of NO_x concentrations and the remaining measurements were used separately to model NO, NO_2 and NO_x using the same spatial covariates from the full dataset. The Chow test identifies structural stability of regression models of the two subsets compared to the pooled full dataset (Gujarati, 1995). To test the transferability of the models, we also standardized prediction across the three models: using variables common to the three models and a variable across the three LUR models was set to have the same buffer distance.

To identify similarities and differences in spatial distribution patterns of NO, NO_2 and NO_x as well as in their emissions sources, the prediction coefficients used to model the three pollutants were also compared for significant predictors.

3. Results

3.1. NO_x measurements

Based on the field blank measures, sampler detection limits were <0.14μg for NO_2 and <0.76μg for NO_x. Our duplicate measurements indicated that the average coefficient of variation was low (3.3% for NO_2 and 2.1% for NO_x). Measured pollutant levels ranged from 5.33 to 42.73 ppb for NO_2 (median = 27.29 ppb) and from 8.08 to 156.95 ppb for NO_x (median = 60.88 ppb), after correcting for blank concentrations. Annual means measured by 14 monitors in the Southern California Air Quality Management District regulatory network showed strong relationships with the campaign-specific means. We calculated the intraclass correlation coefficient (ICC) using a generalized linear model with compound symmetry covariance, clustered by site, and found high correlations between our 2, 2-week average measurements and annual averages based on measurements at the government monitoring stations (ICC = 0.93 for NO, 0.87 for NO_2, and 0.96 for NO_x). Comparison of February 2007 measurements made by Ogawa samplers collocated at government sites (which are chemiluminescence samplers) produced slopes of 0.82 (R^2 = 0.91), 0.72 (R^2 = 0.64) and 0.81 (R^2 = 0.92) for NO and NO_2 and NO_x, respectively. The relationships between collocated measurements in September 2006 were also strong, with slopes of 1.16 (R^2 = 0.82), 0.82 (R^2 = 0.84) and 1.01 (R^2 = 0.89) for NO, NO_2 and NO_x, respectively. To compare the duplicate measurements for both seasons combined, we calculated the ICC as above, clustering by site and season, and found high correlations for all three pollutants of interest (ICC = 0.97 for NO, 0.92 for NO_2, and 0.98 for NO*x*). These results suggest that the short-term monitoring captures the longer term spatial pattern of exposure well in the Los Angeles area.

3.2. Distance decay curves

Because distance decay curves for NO, NO_2 and NO_x showed an upward trend for all traffic estimates at the pre-specified 5000 m buffer distance, we extended the maximum buffer distance to 15 km to identify the optimal distance of influence from traffic sources. Among the four types of traffic density estimates we considered, traffic measures imputed from TeleAtlas Dynamap data were found to have the highest correlations with measured NO, NO_2 and NO_x concentrations and were therefore used in the ADDRESS model selection process. As illustrated in the distance decay curves presented in Fig. 2a, correlations between VMT and traffic-related air pollution concentrations increased steeply out to 10 km and reached a peak around 11 km, especially for NO_x. Highway vehicle density within 11 km alone explained 46.6% of the model variance for NO_x. After a buffer distance of 11 km, correlations trended to slope downward. The explanatory power of highway vehicle density within 11 km suggests the increasing contribution from background NO_x transport within that distance, a finding consistent with earlier integrated meteorological models calibrated for LA (Wu et al., 2005). Since previous studies identified the maximum distance of influence from a roadway to be 1500 m (Jerrett et al., 2005), we created semivariograms of NO and NO_2 concentrations (Fig. 3) using the 201 monitoring site measurements to identify distances of spatial dependency for the two pollutants. Based on these analyses, the distance of spatial dependency for NO was about 11 km, while NO_2 concentrations experienced an even slower decrease, reaching the range where there was no spatial association until about 20 km.

For local effects, we saw a sharp drop of influence of major road from near source (100 m) to a distance of 500–700 m; however, for highways, highways and major roads combined, and all road categories, we saw a sharp increase of influence from near source to a 500 m buffer distance (not shown on figures).

Compared to all other variables we explored, distance to truck routes (not shown on figures) correlated most strongly with NO, NO_2 and NO_X measures (correlation coefficient = 0.57–0.67), and explained 44.2% of the model variance for NO_2. Remote sensing-derived greenness (Fig. 2b – d) also correlated highly with the three pollutant concentrations and was seen as a much better predictor than open space, even though soil brightness (Fig. 2b – d) correlations were lower than those for greenness. The greenness and soil brightness images, shown in Fig. 4a and b, demonstrate that they could be used as surrogates for road networks when such information is unavailable and have the potential to identify off-road vehicle usage such as traffic in parking lots. Truck route slope gradient explained up to 9% of NO_x concentration variance.

Based on the distance decay correlation curves, we found that the majority of the variables analyzed influenced concentrations at a spatial extent greater than 3000 m, especially traffic-related variables. This is inconsistent with previous research findings from medium- or smaller-size urban areas (e.g., traffic influence < 1500m). Using the highway network in LA as an example, Figs. 5 and 6 show the distance decay of NO and NO_2 concentrations with increasing distance from highways (the bar charts show average measured pollutant concentrations with a bin size 200 m). Based on the trend curves, if we use a 50% reduction in concentration as a standard for spatial extent identification, then the spatial extents for NO and NO_2 would be 3000 and 5000 m from highways, respectively. If 10% were used as a threshold, then the spatial extent for both pollutants would be greater than 10 000 m. When treating the 15 SCAQMD monitoring sites as background, the so-called background concentrations are 24.7 and 24.3 ppb for NO and NO_2, and they correspond to 45.8% and 73.2% of the near road concentrations (NO 53.916 ppb and NO_2 33.2 ppb) using the first bin in Figs. 5 and 6.

3.3. ADDRESS modeling results

Before we modeled the NO, NO_2 and NO_x concentrations, we chose 16 monitoring sites randomly for cross-validation, four sites from each of four quartile groups of NO_x concentrations. The remaining 167 sites were used in ADDRESS to model annual concentrations of NO, NO_2 and NO_x, based on the mean concentrations from the two measurement periods.

The final optimized models explained 81, 86 and 85% model variance for NO, NO_2 and NO_x, respectively (Table 2). All three models had near source traffic influence with a buffer distance of 100 m from major road. The local traffic influence was also found for the three models, such as the influence of buffer distance of 400 m for traffic on all the roadways. Strong presence of background concentrations were also captured by the three models, for example, the influence of buffer distance of 11 km for highway and major roads. Closer distance to truck routes was also associated with higher traffic-related pollutant concentrations. Industrial and commercial land usage were also positively correlated with pollutant concentrations (industrial buffer distance 1700 m for NO_2 and 2700 m for NO and NO_x; for commercial land use they were 1200 m for NO and 1000 m for NO_2 and NO_x). In addition, *X* coordinate and soil brightness were also found to be significant in predicting NO_x concentrations.

Because ADDRESS assumed independency of each sampling measurement, spatial clustering was not considered. By contrast, the correlation of samplers within a census tract was taken into consideration on the GEE models; the standard errors were therefore smaller than with ADDRESS.

All the chosen spatial covariates were statistically significant at a 0.05 level with expected signs. The average VIF for NO, NO_2 and NO_x were 1.25, 1.31 and 1.22, respectively, with the maximum VIF being 1.56, demonstrating a lack of significant collinearity between the chosen spatial covariates. The prediction scatter plots displayed in Fig. 7a (for NO), b (NO_2) and c (NO_x) further demonstrate that the prediction models were not influenced by significant outliers and the model prediction residuals were normally distributed.

The maximum Cook's distance for NO, NO_2 and NO_x was 0.10, 0.36 and 0.23, respectively, also confirming the absence of influential outliers in each model. After selecting these parsimonious models, we tested for spatial autocorrelation based on first-and second-order Thiessen polygon connectivity matrices with Moran's *I*. There was significant autocorrelation with the first-order matrix in all models, but not in the second-order tests (with *I* being equal to −0.008 ($p = 0.48$), −0.023 ($p = 0.35$) and 0.031 ($p = 0.18$) for NO, NO_2 and NO_x model residuals). The ADDRESS models with an extra cluster parameter and the GEE models (Table 2 second and third columns) showed similar results to corresponding ADDRESS models that did not incorporate adjustment for clustering, except some small changes to the standard errors.

In our cross-validation models (Fig. 7d, e and f), the 16 randomly picked samples explained 91%, 87% and 92% of the model variances for NO, NO_2 and NO_x, respectively. The Chow test showed that there was no significant difference between the models with the full and half dataset for NO, NO_2 and NO_x, indicating model stability to subset selection.

The prediction coefficients in Table 3 reflected that a spatial covariate of the same buffer distance showed great similarity in significant digits in prediction of NO, NO_2 and NO_x. Generally, land use variables had smaller variation compared with traffic-related ones.

The final prediction surfaces for NO, NO_2 and NO_x are displayed in from Fig. 8a–d. Fig. 8a (for NO) and 8d (for NO_x) show similar concentration patterns, i.e., higher NO and NO_x levels near to emission sources such as highways and industrial land use. Similarly, in calculating the correlation coefficients between NO, NO_2 and NO_x prediction surfaces, we found that NO has the highest correlation with NO_x ($r = 0.93$), while the correlation between NO and NO_2 was smaller ($r = 0.85$). There were greater differences in spatial patterns between NO and NO_2 than between NO and NO_x. Fig. 8b is the predicted NO_2 concentration surfaces with and Fig. 8c without highway and major road traffic at a buffer distance of 11 km included in the model. The NO_2 prediction surface without the 11 km spatial variable for traffic predicts higher NO_2 gradients, even though the total concentrations are slightly lower.

4. Discussion and conclusions

In this paper, we modeled NO, NO_2 and NO_x concentrations for the LA metropolitan area using the ADDRESS modeling strategy. Our final three prediction models explained 81%, 86% and 85% of NO, NO_2 and NO_x variances, respectively. The models presented here have a higher power of prediction (R^2) than a large majority of previously developed LUR models (Jerrett et al., 2007; Henderson et al., 2007; Hoek et al., 2008). To our knowledge, this is the first application of an intensive air pollution monitoring campaign (with 201 samplers) to model traffic-related air pollution in a mega city. Our LUR models and semivariograms suggest that the distance of influence for highway and major roads density around the monitors is greater than 10 km. Earlier regional studies in England similarly reported even larger influence areas for NO_2—suggesting that regional patterns are an important contributor to NO_2 levels in this locale (Stedman et al., 1997). Because of LA's geography, infrastructure, road network and population characteristics, background concentrations were high even though both NO and NO_2 are generally considered as reactive pollutants. As the background concentration increases, the spatial extent of influence from the source increases correspondingly (Zhou and Levy, 2007). In addition, high emission rates also increase the spatial extent of impact for absolute comparisons (Zhou and Levy, 2007). Land use regression models based on a limited number of buffer distances (up to 5000 m) were found to be less predictive of the spatial distribution of traffic-related air pollution in the LA metropolitan area. Most previous LUR models considered circular area buffers of less than 1500m for examining roadway and traffic variables, and land use and population density within a maximum distance of 3000m (Jerrett et al., 2005; Henderson et al., 2007). Typically, for a medium–large size city (2–5 million population), the influence of traffic diminishes with increasing distance nearest roads, and local sources of pollution dominate over background effects (Jerrett et al., 2005, 2007). Highways (including primary and secondary) have a total length of more than 3000 km and highway and major road densities are 435 and 1669m/km^2, respectively, in the LA Basin, much higher than corresponding road categories for a typical medium size city. Levels of NO_x pollution at a residence in a mega city may be influenced by local sources, and in addition may depend strongly on urban-scale and even regional background sources. These results indicate that beyond the distance decay gradients suggested in recent studies of between 300 and 1500m (see, e.g., Jerrett et al., 2007; Zhou and Levy, 2007) that there is a cumulative impact of traffic and density around the monitors and that the range of influences is probably much greater in mega cities than in smaller urban areas. These findings suggest that for mega cities, larger zones of influence may be necessary to predict traffic pollution accurately in future studies.

It is also important to note that a higher correlation (R) between a given explanatory variable within a given buffer distance and a pollutant concentration does not correspond to a higher pollutant concentration at that buffer distance. A correlation coefficient of a buffer distance reflects the influence of total traffic or land use within that buffer, not at that buffer distance. The pollutant concentration at that buffer distance could be much smaller than concentrations at locations closer to the source of pollution. In land use regression modeling and exposure analysis, it is of interest for potential health effects to include all information within the optimized buffer distances. The LUR models were optimized to give a best estimate of total NO and NO_2 concentrations regardless of source. Regional, urban and local contributions were all taken into consideration. This is consistent with current exposure analysis that treats the scope of influence as a function of the regional background, an urban background and the local traffic (e.g., near a major road) (Hoek et al., 2001). When near a roadway, the influence of local traffic is the dominant source and the contribution from background might be negligible; however, further away from the roadway, the influence of local traffic decreases relative to the background. It is in our study that the background concentrations had a higher proportion of contribution to the overall concentration of a point of interest.

Satellite remote sensing of trace gases and aerosols for air quality applications has a rich history. Three major applications are analyses and forecasts of events that affect air quality, inference of surface air quality itself and estimates of emissions. The pollutants monitored include particular matter (PM_{10} and $PM_{2.5}$, aerodynamic diameter <10 and 2.5μm, respectively), O_3, NO_x, CO (carbon monoxide), HCHO (formaldehyde), and SO_2 (sulfur dioxide) and VOCs (volatile organic compounds). However, the spatial resolution of satellite remote sensing is usually coarser than 1 km (see Martin, 2008 for a detailed review of satellite remote sensing of surface air quality). Su et al. (2008, 2009) found that remote sensed data such as those from Landsat ETM+ (spatial resolution 30 m) can be used to enhance land use regression models in identifying small area variation of ambient pollutant concentrations. The Landsat-derived satellite data may overestimate roadway emissions to some degree, but will not be overwhelmingly biased because the spectral information between a tar-roofed building and busy vehicle-traveled roadway/parking lot are different. Places with high vegetation cover have the effect of reducing pollutant concentrations, while roadways and tar-roofed buildings do not. The slight over-estimation of pollutant concentrations at places with tar-roofed buildings might better represent the spatial pattern of pollutant concentrations at those places. Compared to open land use, the degree of greenness or soil brightness from remote sensed data should be more accurately characterizing ground land use. Overall, we demonstrated that remote sensing-derived data such as vegetation greenness and soil brightness can be useful model inputs that will improve the estimation of spatial variability in NO, NO_2 and NO_x concentrations, especially greenness which correlated highly with these concentrations ($r = 0.40–0.50$). Advantage of using ETM+ data for LUR include its global coverage and free access (http://landsat.gsfc.nasa.gov/). In locations where other spatial covariates are not readily available such as certain land use data, Landsat ETM+ data might provide effective surrogate measures. Many of the world's megacities are in poorer countries with less-developed geospatial infrastructure, making this method for deriving land cover information potentially useful in many other regions. Remote sensing-derived data could also help identify other environments with traffic, such as parking lots. The other advantage is the comparability between places with regard to land use classification. The model we developed here provides a relatively easy and feasible way to improve exposure analysis.

We also found that at increasing distances from steeper gradient truck routes, the influence of the slope gradients decreases suggesting that steeper gradient truck routes exerted higher NO_x emissions and had a positive influence on concentrations. However, the maximum correlation between truck route slope gradients and NO_x concentrations was 0.27, while statistically significant, a correlation much smaller than we previously identified, i.e., 0.50 (Su et al., 2009). When using slope gradients as variables in the LUR model, we assumed that all highways with steeper slope gradients have higher emissions (because of the acceleration effect) but that the transportation of goods and services is roughly equal in each direction of a highway. Because slope gradient is a proxy for hard acceleration driving patterns, it may supply a useful approximation for elevated emissions in some study areas. In LA, the biggest contributor to air pollution is traffic, and NO was found to be highly concentrated near highways and major roadways, especially near truck routes on which most of the goods and services vehicles travel. Because of the onshore sea breeze and secondary formation of NO_2 associated with inland air transportation processes, NO_2 concentrations were found to be higher in the northern and eastern parts of the metropolitan area, and lower in the western areas. Concentrations of NO_x (i.e., NO+NO_2) were also higher near emission sources due to the higher proportion of NO compared to NO_2 contributing to the NO_x composition (22.37 ppb vs 18.96 ppb for the basin). With the inclusion of highway traffic within a buffer distance of 11 km, the modeled NO_2 concentration surface was higher and at the same time smoother (Fig. 8b) compared to the model that did not consider the influence of highway traffic within this distance (Fig. 8c).

Most previous land use regression models (Jerrett et al., 2005; Henderson et al., 2007) have included population density as a predictor. However, in our modeling process, population density was not included for all three prediction models because of the high VIFs (VIF>2).

Our sensitivity analyses demonstrated that when population density was included, the model prediction power increased only by 0.50%, 1.28% and 0.71%, respectively, for NO, NO_2 and NO_x; thus, omitting population density from our LUR did not substantially decrease the predictive power of our models.

This study illustrated that enhanced LUR model selection and data input can produce much higher prediction powers when modeling traffic-related air pollution in the extraordinarily complex cityscape of Los Angeles. The models developed here and the substantive insights gained into the relevance of factors for the modeling process may supply guidance for future studies in other mega cities. For rapidly industrializing areas with increasing traffic and relatively sparse land use or traffic data, the remote sensing data we included in our model promises improvements in exposure assessment. Similarly, depending on the topology of an area, roadway slope gradients can be derived from readily available DEMs and can either augment or partially replace ground-level data. The ADDRESS modeling strategy provides a feasible method to evaluate models across a range of complex urban environments (Jerrett et al., 2005, 2007). The enhanced exposure modeling strategy may increase our prediction of human exposure to traffic-related pollution for epidemiologic studies in many urban locales.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by the California Air Resources Board (CARB) under contract 04–323 “Traffic-Related Air Pollution and Asthma in Economically Disadvantaged and High Traffic Density Neighborhoods in Los Angeles, California”, CARB 06–332 “Spatial-temporal Analysis of Air Pollution and Mortality in California Based on the American Cancer Society Cohort” and by the California Office of Environmental Health Hazard Assessment 07-E0009 “Understanding and Acting on Cumulative Impacts on California Communities”. We thank the research team at UCLA who conducted the Ogawa NO_x measurements in Los Angeles.

References

Bailey, TC.; Gatrell, AC. Interactive Spatial Data Analysis. Harlow, Essex: Addison Wesley Longman; 1995.

Bolstad, P. GIS Fundamentals: A First Text on Geographic Information Systems. second ed.. White Bear Lake, MN: Eider Press; 2005.

Bytnerowicz A, Fenn ME. Nitrogen deposition in California forests: a review. Environ. Pollut. 1996; 92:127–146. [PubMed: 15091393]

Cressie, N. Statistics for Spatial Data. New York: Wiley; 1993. Revised ed.

Crist EP, Cicone RC. A physically-based transformation of thematic mapper data—the TM tasseled cap. IEEE Trans. Geosci. 1984:256–263. Remote Sensing GE-22.

GLCF (Global Land Cover Facility). [Accessed March 5, 2009] 2009. ⟨http://glcf.umiacs.umd.edu/index.shtml⟩

Gujarati, D. Basic Econometrics. third ed.. New York, NY: McGraw-Hill; 1995.

Henderson SB, Beckerman B, Jerrett M, Brauer M. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. Environ. Sci. Technol. 2007; 41:2422–2428. [PubMed: 17438795]

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.

Hoek G, Fischer P, Van Den Brandt P, Goldbohm S, Brunekreef B. Estimation of long-term average exposure to outdoor air pollution for a cohort study on mortality. J. Expo. Anal. Environ. Epidemiol. 2001; 11:459–469. [PubMed: 11791163]

Hricko A. Global trade comes home: community impacts of goods movement. Environ. Health Perspect. 2008; 116:A78–A81. [PubMed: 18288306]

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. [PubMed: 15292906]

Jerrett M, Arain MA, Kanaroglou P, Beckerman B, Crouse D, Gilbert NL, Brook JR, Finkelstein N, Finkelstein MM. Modelling the intra-urban variability of ambient traffic pollution in Toronto, Canada. J. Toxicol. Environ. Health. 2007; 70:200–212.

Kanaroglou P, Jerrett M, Morrison J, Beckerman B, Arain MA, Gilbert N, Brook J. Establishing an air pollution monitoring network for intraurban population exposure assessment; a location-allocation approach. Atmos. Environ. 2005; 39:2399–2409.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

Martin RV. Satellite remote sensing of surface air quality. Atmos. Environ. 2008; 42:7823–7843.

Moore DK, Jerrett M, Mack WJ, Kunzli N. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. J. Environ. Monit. 2006; 9:246–252. [PubMed: 17344950]

USGS. National Elevation Dataset from the US Geological Survey. Sioux Falls, SD: 1999. ⟨http://gisdata.usgs.net/ned/⟩ Last online [access September 25, 2008]

Ross Z, Jerrett M, Ito K, Tempalski B, Thurston GD. A land use regression for predicting fine particulate matter concentrations in the New York City region. Atmos. Environ. 2007; 41:2255–2269.

Ross Z, English P, Jerrett M, Scalf R, Gunier RB, Smorodinsky S, Wall S. Nitrogen dioxide prediction in southern California using land use regression modeling: potential for environmental health analyses. J. Expo. Sci. Environ. Epidemiol. 2006; 16:106–114. [PubMed: 16047040]

Singer BC, Hodgson AT, Hotchi T, Kim JJ. Passive measurement of nitrogen oxides to assess traffic-related pollutant exposure for the east bay children's respiratory health study. Atmos. Environ. 2004; 38:393–403.

Stedman JR, Vincent KJ, Campbell GW, Goodwin JWL, Downing CEH. New high resolution maps of estimated background ambient NO_x and NO_2 concentrations in the UK. Atmos. Environ. 1997; 31:3591–3602.

Su JG, Berkerman B, Jerrett M. A distance decay variable selection strategy for optimized land use regression modeling. Sci. Total Environ. 2009; 407:3890–3898. [PubMed: 19304313]

Su JG, Brauer M, Buzzelli M. Estimating urban morphometry at the neighborhood scale for improvement in modeling long-term average air pollution concentrations. Atmos. Environ. 2008; 42:7884–7893.

US EPA (US Environmental Protection Agency). Guideline for regulatory application of the Urban Airshed Model. Report No. EPA-450/4-91-013. 1991.

Walsh MP. Ancillary benefits for climate change mitigation and air pollution control in the world's motor vehicle fleets. Annu. Rev. Public Health. 2008; 29:1–9. [PubMed: 18173380]

Wu J, Lurmann F, Winer A, Lu R, Turco R, Funk T. Development of an individual exposure model for application to the Southern California children's health study. Atmos. Environ. 2005; 39:259–273.

Zhou Y, Levy JI. Factors influencing the spatial extent of mobile source air pollution impacts: a meta-analysis. BMC Public Health. 2007; 7:89. [PubMed: 17519039]

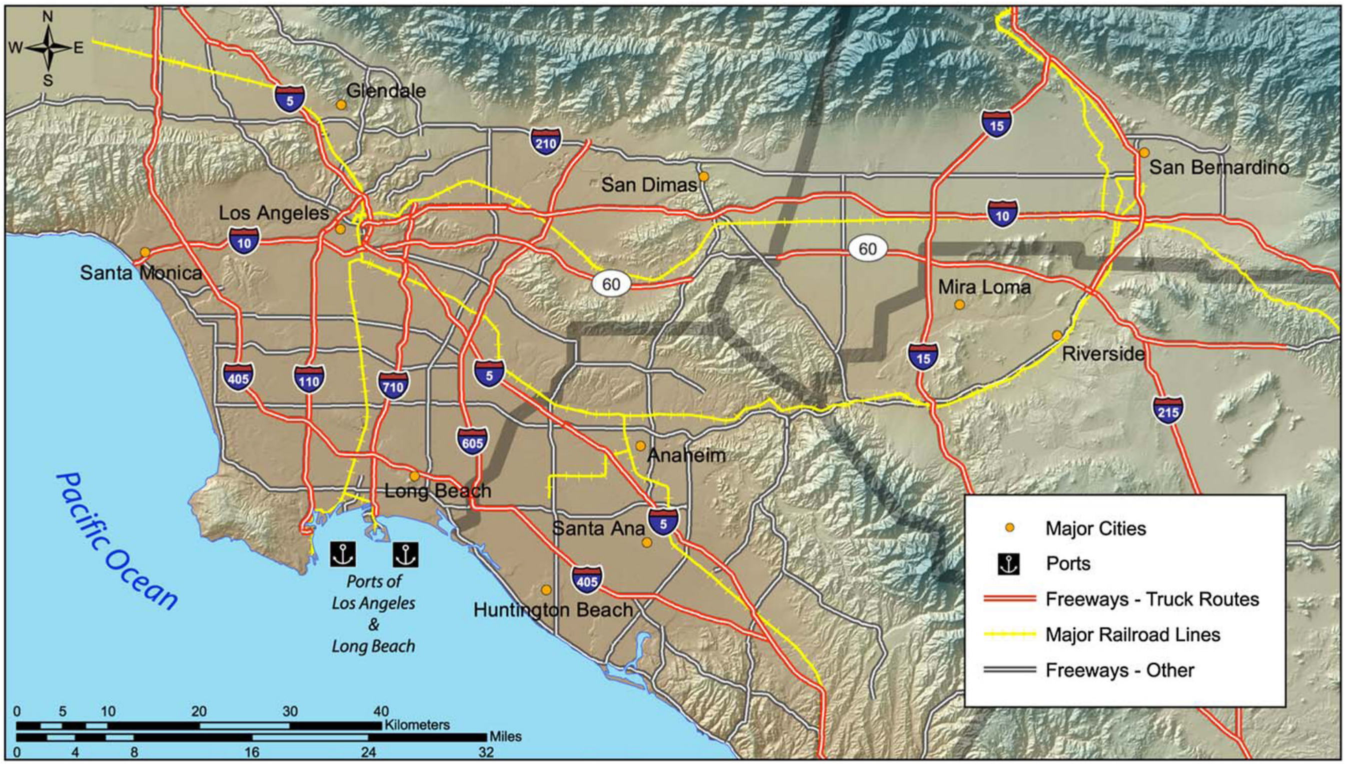
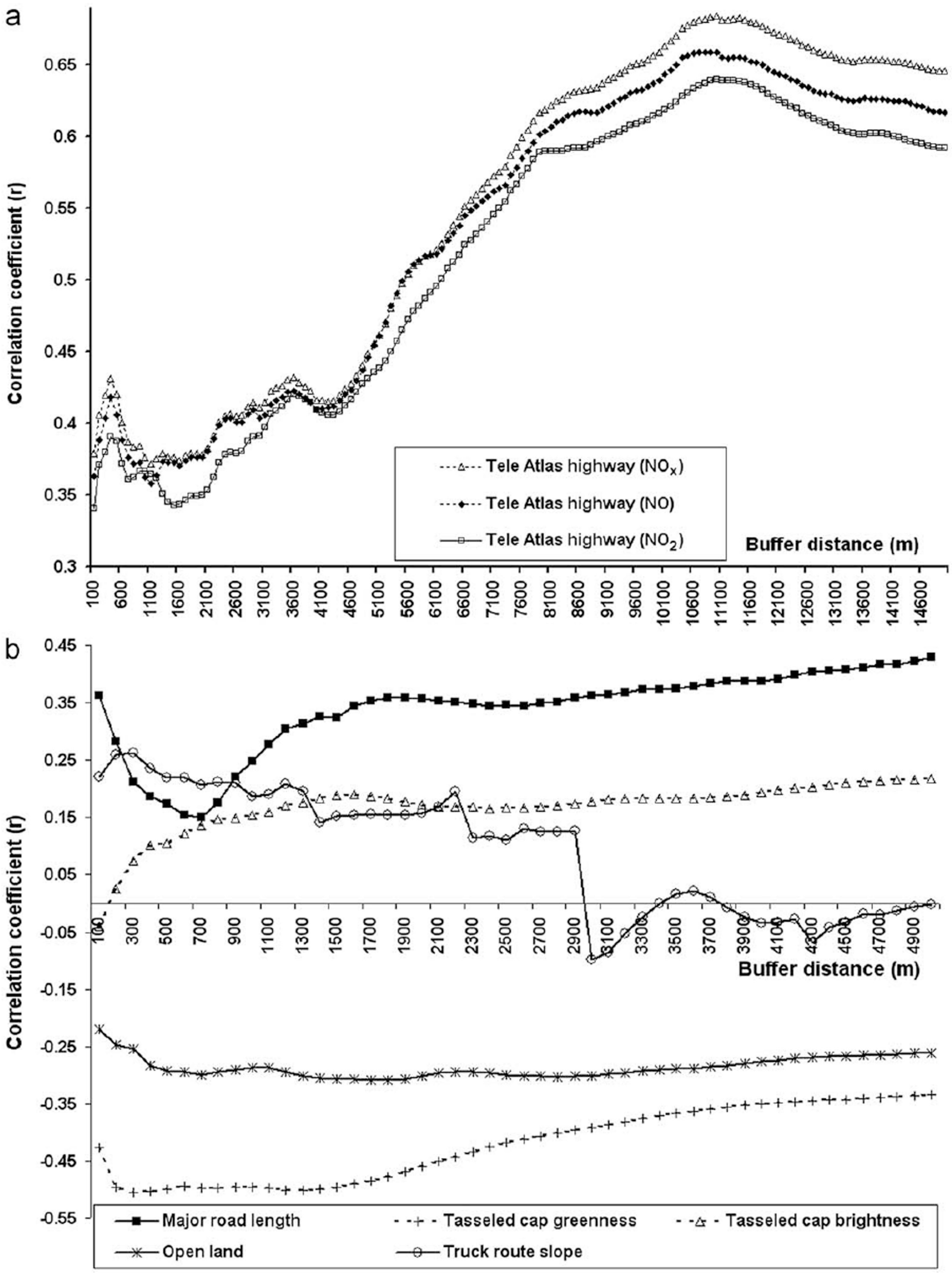


Fig. 1.
The Los Angeles Study Area.


a
Correlation coefficient (r)
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
Tele Atlas highway (NO$_X$)
Tele Atlas highway (NO)
Tele Atlas highway (NO$_2$)
Buffer distance (m)
100
600
1100
1600
2100
2600
3100
3600
4100
4600
5100
5600
6100
6600
7100
7600
8100
8600
9100
9600
10100
10600
11100
11600
12100
12600
13100
13600
14100
14600
b
Correlation coefficient (r)
0.45
0.35
0.25
0.15
0.05
-0.05
-0.15
-0.25
-0.35
-0.45
-0.55
100
300
500
700
900
1100
1300
1500
1700
1900
2100
2300
2500
2700
2900
3100
3300
3500
3700
3900
4100
4300
4500
4700
4900
Buffer distance (m)
Major road length
Tasseled cap greenness
Tasseled cap brightness
Open land
Truck route slope

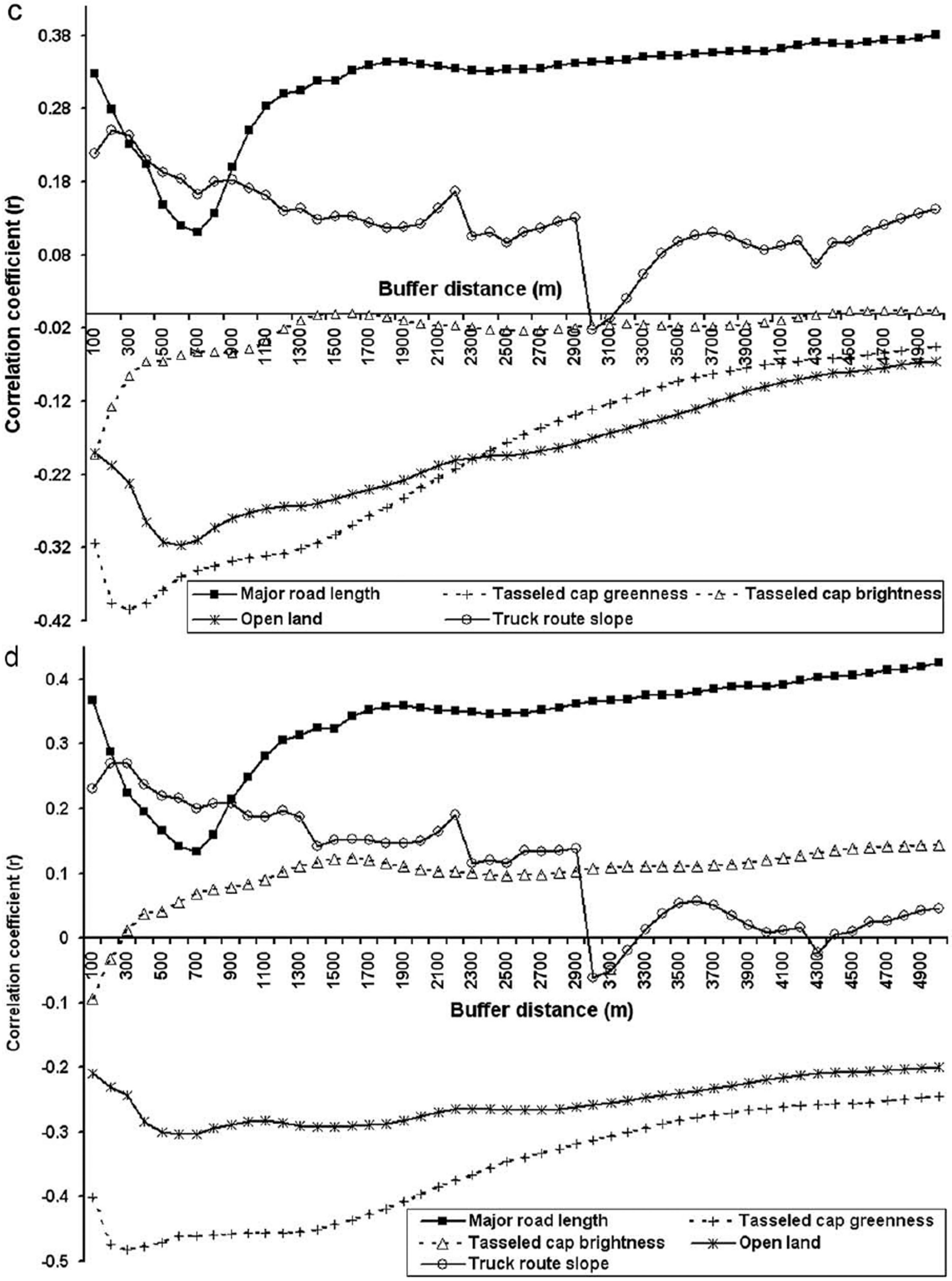


Fig. 2.

Distance decay curves of correlation between selected spatial covariates and measured air pollution concentrations (a) for traffic volumes-total vehicle miles traveled, (b) for NO, (c) for NO_2 and (d) for NO_x.

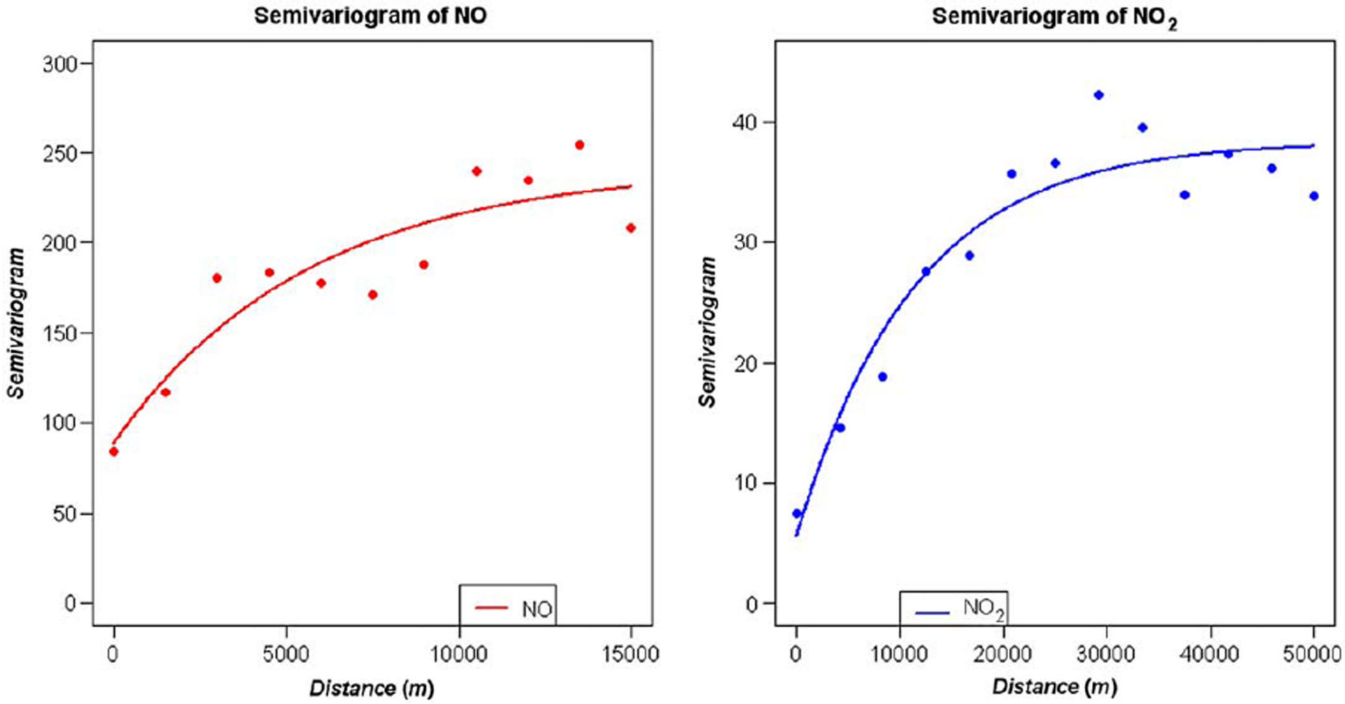


Fig. 3.

The semivariograms of NO and NO_2 based on measurements from the 201 monitoring sites.

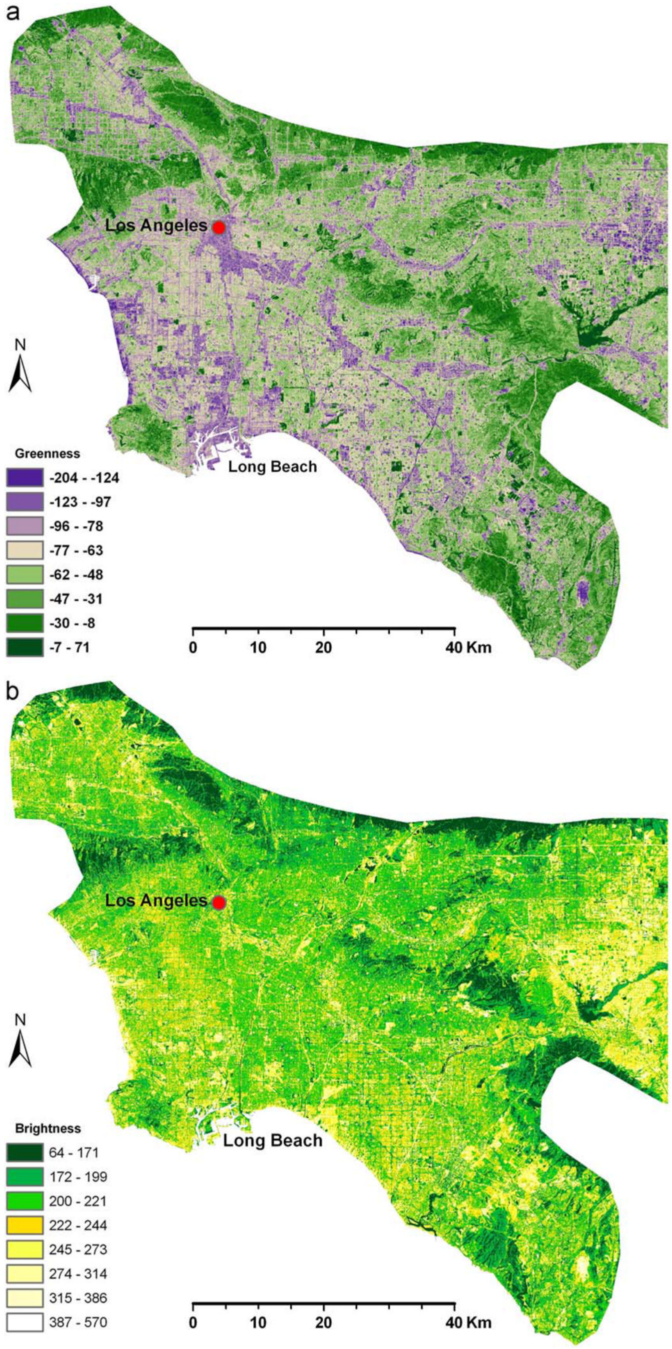


Fig. 4.

Tasseled-cap greenness (a) and soil brightness (b).

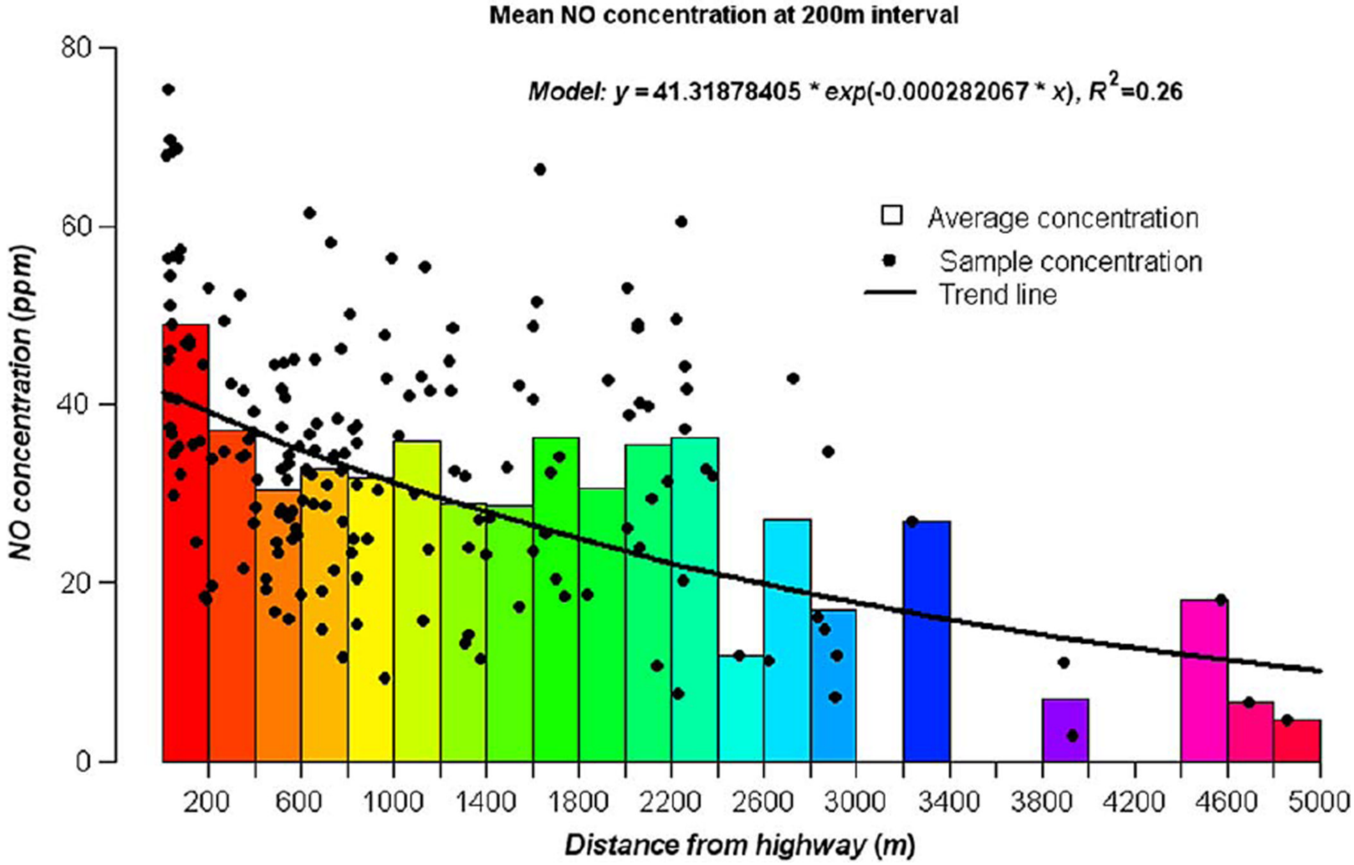


Fig. 5.

The distance decay of NO concentrations further away from highway (Al and A2) based on the 201 monitoring sites in the LA metropolitan area.

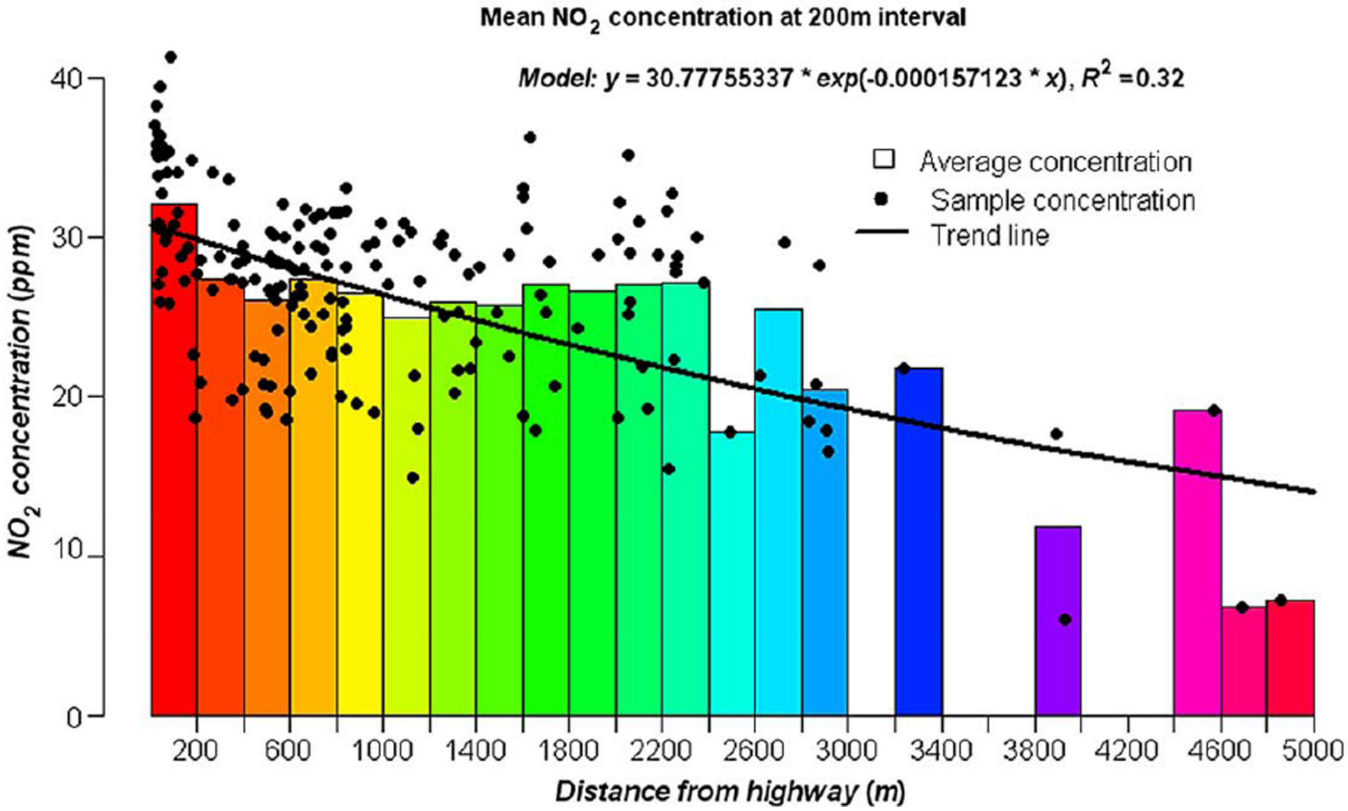


Fig. 6.

The distance decay of NO_2 concentrations further away from highway (Al and A2) based on the 201 monitoring sites in the LA metropolitan area.

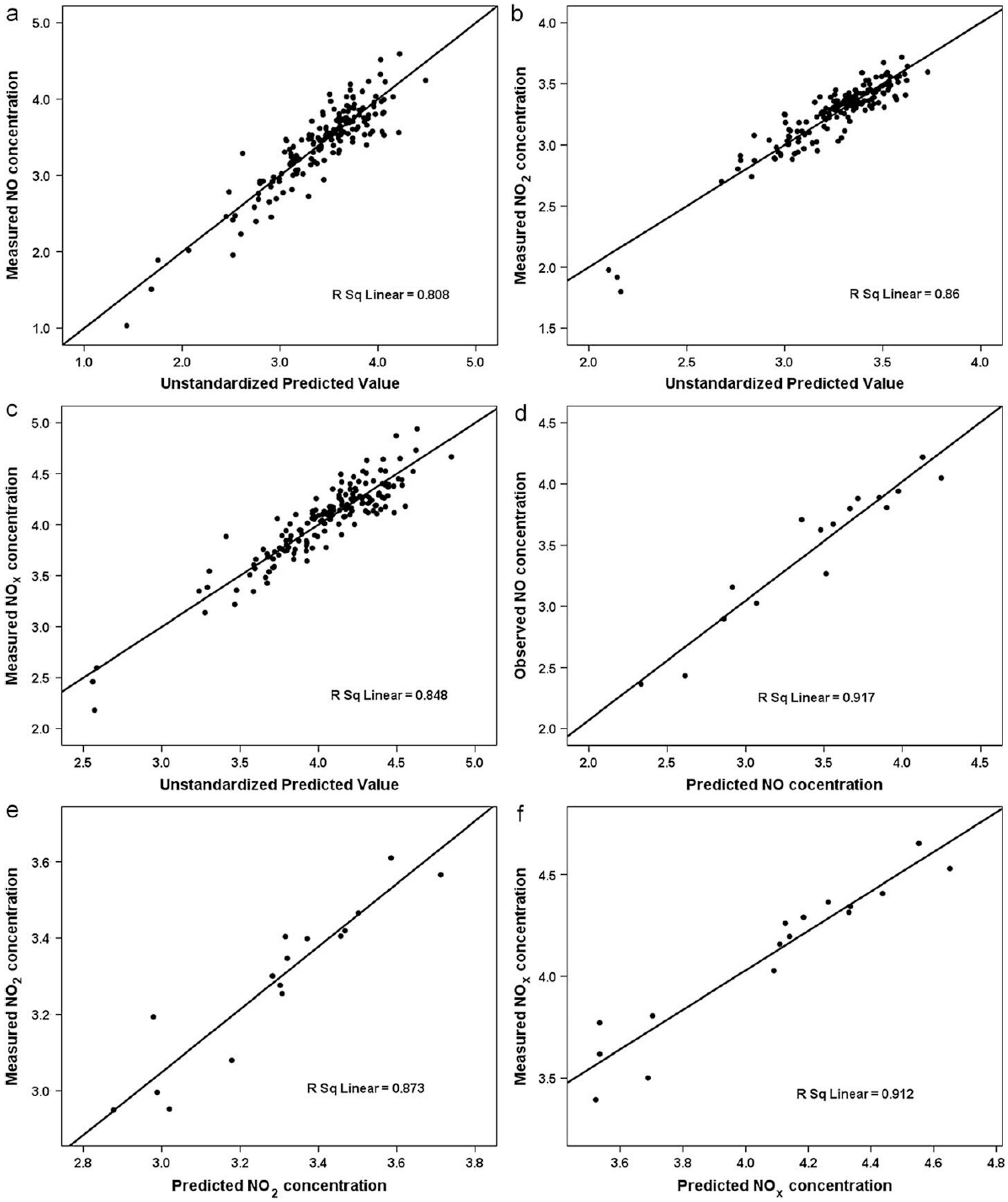


Fig. 7.

Model predictions of natural log-transformed NO, NO$_2$ and NO$_x$ (a, b and c) and corresponding cross-validation results (d, e and f).

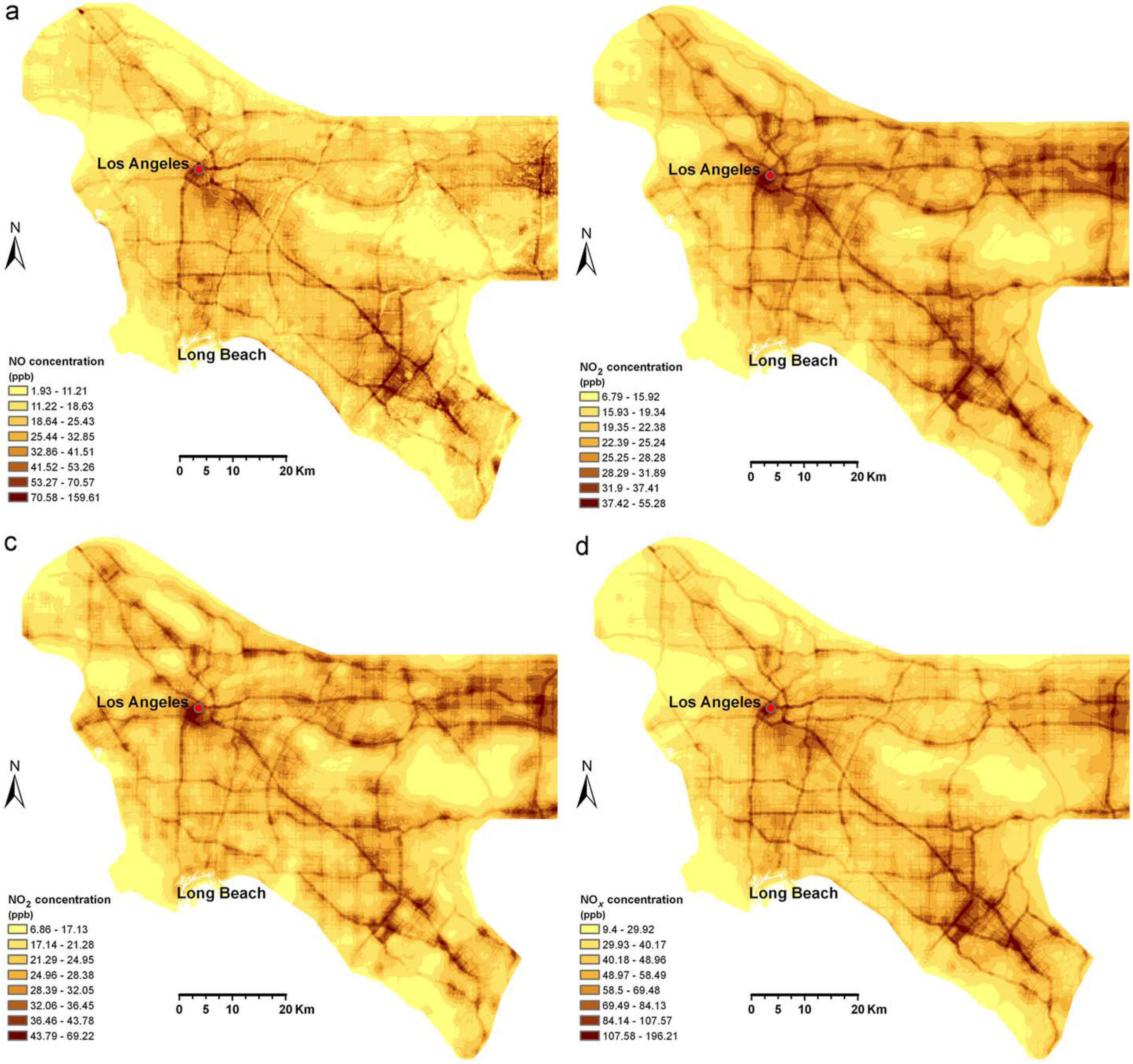


Fig. 8.

Model prediction surfaces of NO (a), NO_2 (b, c) and NO_x (d) through an ADDRESS selection process. The difference between 5b and 5c is that 5c represents the modeling result omitting highway buffer distance 11 km as a predictor.

Table 1

Traffic statistics for measured road segments in LA and the proportion of roads with measurements.

Road category[a]	**Traffic volume measurements**						**TeleAtlas dynamap data coverage**	
	# roads	**Minimum**	**Maximum**	**Mean**	**Median**	**Std**	**# roads**	**% measured**
Al	1658	3100	190000	92969.10	**104000**	39860.36	24508	6.77
A2	799	1885	76000	21458.43	**18750**	10756.69	9004	8.87
A3	8045	21	114622	15847.57	**13500**	11119.84	124391	6.47
A4	7954	1	129500	4611.44	**2383**	6093.39	500302	1.59
A5	3	564	2100	1092.67	**614**	872.73	9105	0.03
A6	40	556	148500	40407.45	**17000**	45863.49	37 866	0.11
A7	5	488	26950	9500.60	**6780**	11007.81	34 871	0.01
Total:	18504						740047	2.50

[a] Al: Primary highways with limited access; A2: primary highways without limited access; A3: secondary and connecting roads; A4: local, neighborhood and rural roads; A5: vehicular trails; A6: road access ramp; A7: roads as other thoroughfares.

Table 2

Model prediction results based on the ADDRESS model, the ADDRESS model with clustering taken into account and GEE models for NO, NO_2 and NO_x.

Pollutant	Variable	ADDRESS model[a]					ADDRESS model with clustering[b]				GEE model[c]			
		Coef.	Std. err.	*t*	*P*>|*t*|	VIF[a]	Coef.	Std. err.	*t*	*P*>|*t*|	Coef.	Std. Err.	*t*	*P*>|*t*|
(2a) NO	Intercept	−2.6407240	0.654861	−4.03	0.000		−2.6407240	0.792661	−3.33	0.001	−2.6407240	0.634952	−4.16	0.000
	TeleAtlas traffic highway and major roads (11000m)	0.00000003	0.000000	11.93	0.000	1.35	0.00000003	0.000000	10.58	0.000	0.00000003	0.000000	12.30	0.000
	TeleAtlas traffic all roads (400 m)	0.00000210	0.000000	6.75	0.000	1.27	0.00000210	0.000000	5.91	0.000	0.00000210	0.000000	6.96	0.000
	Distance to truck routes (m)	−0.00003880	0.000013	−3.06	0.003	1.49	−0.00003880	0.000014	−2.87	0.005	−0.00003880	0.000012	−3.15	0.002
	Major road (100m)	0.00053750	0.000143	3.77	0.000	1.13	0.00053750	0.000178	3.01	0.004	0.00053750	0.000138	3.89	0.000
	Industrial (2700 m)	0.00036130	0.000084	4.29	0.000	1.20	0.00036130	0.000089	4.08	0.000	0.00036130	0.000082	4.42	0.000
	Commercial (1200 m)	0.00277730	0.000533	5.21	0.000	1.19	0.00277730	0.000514	5.40	0.000	0.00277730	0.000517	5.37	0.000
	Soil brightness (700 m)	0.01005310	0.001727	5.82	0.000	1.32	0.01005310	0.001935	5.20	0.000	0.01005310	0.001674	6.00	0.000
	X coordinate (m)	0.00000660	0.000001	5.76	0.000	1.09	0.00000660	0.000001	4.80	0.000	0.00000660	0.000001	5.94	0.000
	Open (100 m)	−0.1542625	0.049998	3.09	0.002	1.20	−0.1542625	0.047074	−3.28	0.002	−0.1542625	0.048478	−3.18	0.001
	R^2 (*p*)|R^2 (*p*)[d]		0.81 (<0.0001)|0.92 (<0.0001)					0.81 (<0.0001)				NA		
(2b) NO_2	Intercept	−11.282530	2.443303	−4.62	0.000		−11.282530	3.725334	−3.03	0.003	−11.2825300	2.369021	−4.76	0.000
	TeleAtlas traffic highway and major roads (11000 m)	0.00000001	0.000000	9.72	0.000	1.44	0.00000001	0.000000	7.23	0.000	0.00000001	0.000000	10.03	0.000
	TeleAtlas traffic all roads (400 m)	0.00000072	0.000000	5.26	0.000	1.28	0.00000072	0.000000	4.04	0.000	0.00000072	0.000000	5.43	0.000
	Distance to truck routes (m)	−0.00004390	0.000006	−7.90	0.000	1.49	−0.00004390	0.000014	−3.04	0.003	−0.00004390	0.000005	−8.15	0.000
	Major road (100)	0.00018990	0.000063	3.01	0.003	1.15	0.00018990	0.000070	2.72	0.008	0.00018990	0.000061	3.11	0.002
	Local road (1400)	0.00000234	0.000001	2.82	0.005	1.56	0.00000234	0.000001	2.73	0.008	0.00000234	0.000001	2.91	0.004
	Industrial (1700m)	0.00059240	0.000096	6.16	0.000	1.36	0.00059240	0.000145	4.10	0.000	0.00059240	0.000093	6.35	0.000
	Commercial (1000 m)	0.00261960	0.000308	8.50	0.000	1.16	0.00261960	0.000376	6.96	0.000	0.00261960	0.000299	8.77	0.000
	X coordinate (m)	0.00000515	0.000001	10.11	0.000	1.12	0.00000515	0.000001	7.38	0.000	0.00000515	0.000000	10.43	0.000
	Y coordinate (m)	0.00000316	0.000001	4.98	0.000	1.20	0.00000316	0.000001	3.26	0.002	0.00000316	0.000001	5.14	0.000

Pollutant	Variable	ADDRESS model[a]					ADDRESS model with clustering[b]				GEE model[c]			
		Coef.	Std. err.	t	$P>\|t\|$	VIF[a]	Coef.	Std. err.	t	$P>\|t\|$	Coef.	Std. Err.	t	$P>\|t\|$
	R^2 (p)|R^2 (p)		0.86 (<0.0001)|0.87 (<0.0001)					0.86(<0.0001)				NA		
(2c) No_x	Intercept	−0.0590325	0.429163	−0.14	0.891		−0.0590325	0.543035	−0.11	0.914	−0.05903250	0.417438	−0.14	0.888
	TeleAtlas traffic highway and major roads (11000 m)	0.00000002	0.000000	13.13	0.000	1.30	0.00000002	0.000000	10.83	0.000	0.00000002	0.000000	13.5	0.000
	TeleAtlas traffic all roads (400 m)	0.00000144	0.000000	7.11	0.000	1.27	0.00000144	0.000000	5.20	0.000	0.00000144	0.000000	7.31	0.000
	TeleAtlas traffic major road (100 m)	0.00002710	0.000007	3.92	0.000	1.14	0.00002710	0.000008	3.22	0.002	0.00002710	0.000007	4.03	0.000
	Distance to truck routes (m)	−0.0000450	0.000008	−5.47	0.000	1.47	−.0000450	0.000011	−3.94	0.000	−0.0000450	0.000008	−5.63	0.000
	Industrial (2700 m)	0.00029010	0.000054	5.38	0.000	1.16	0.00029010	0.000070	4.15	0.000	0.00029010	0.000053	5.53	0.000
	Commercial (1000 m)	0.00328070	0.000450	7.29	0.000	1.12	0.00328070	0.000488	6.72	0.000	0.00328070	0.000438	7.49	0.000
	Soil brightness (1700 m)	0.00442720	0.001104	4.01	0.000	1.19	0.00442720	0.001500	2.95	0.004	0.00442720	0.001074	4.12	0.000
	X coordinate (m)	0.00000572	0.000001	7.62	0.000	1.10	0.00000572	0.000001	6.67	0.000	0.00000572	0.000001	7.83	0.000
	R^2 (p)|R^2 (p)		0.85 (<0.0001)|0.92 (<0.0001)					0.85(<0.0001)				NA		

[a] ADDRESS model: an optimized distance decay model selection strategy for our land use regression models. VIF = variance inflation factor.

[b] For clustering analysis, observations were grouped using census tract. We assumed that measurements from multiple sites within a census tract might be correlated but not across census tracts.

[c] GEE model: generalized estimation equation model to analyze correlated data within census tracts.

[d] R^2 (p)|R^2 (p): the left side part is for model prediction power and right side for cross-validation result.

Table 3

Prediction coefficients of significant spatial covariates used to predict pollutant concentrations of NO, NO_2 and NO_x.

Variable	Pollutant[a]		
	NO	NO_2	NO_x
Intercept	−2.64072400	−11.28253000	−0.05903250
TeleAtlas traffic highway and major roads (11000 m)	0.00000003	0.00000001	0.00000002
TeleAtlas traffic all roads (400 m)	0.00000210	0.00000072	0.00000144
Distance to truck routes (m)	−0.00003880	−0.00004390	−0.00004500
Major road (100 m)	0.00053750	0.00018990	0.00002710
Local road (1400)		0.00000234	
Industrial (1700m)		0.00059240	
Industrial (2700 m)	0.00036130		0.00029010
Commercial (1000 m)		0.00261960	0.00328070
Commercial (1200m)	0.00277730		
Soil brightness (700 m)	0.01005310		
Soil brightness (1700m)			0.00442720
X coordinate (m)	0.00000660	0.00000515	0.00000572
Y coordinate (m)		0.00000316	
Open (100 m)	−0.15426250		

[a] NO, NO_2 and NO_x are natural log transformed.