Title
Modeling the Concentrations of On-Road Air Pollutants in Southern California

Journal
Environmental Science and Technology, 47(16)

ISSN
0013-936X

Authors
Li, Lianfa
Wu, Jun
Hudda, Neelakshi
et al.

Publication Date
2013-08-20

DOI
10.1021/es401281r

Peer reviewed

Environmental Science & Technology

Article

pubs.acs.org/est

Modeling the Concentrations of On-Road Air Pollutants in Southern California

Lianfa Li,[†,‡] Jun Wu,[*,†,§] Neelakshi Hudda,[∥] Constantinos Sioutas,[∥] Scott A. Fruin,[⊥] and Ralph J. Delfino[§]

[†]Program in Public Health, College of Health Sciences, and [§]Department of Epidemiology, School of Medicine, University of California, Irvine, California, United States

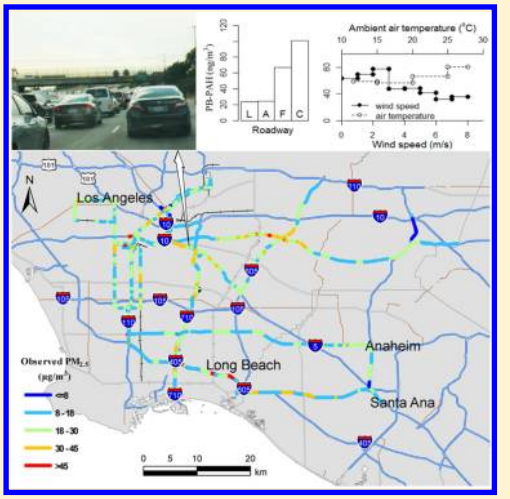
[‡]State Key Lab of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

[∥]Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, California 90089, United States

[⊥]Keck School of Medicine, Environmental Health Division, University of Southern California, Los Angeles, California 90033, United States

Supporting Information

ABSTRACT: High concentrations of air pollutants on roadways, relative to ambient concentrations, contribute significantly to total personal exposure. Estimation of these exposures requires measurements or prediction of roadway concentrations. Our study develops, compares, and evaluates linear regression and nonlinear generalized additive models (GAMs) to estimate on-road concentrations of four key air pollutants, particle-bound polycyclic aromatic hydrocarbons (PB-PAH), particle number count (PNC), nitrogen oxides (NO_x), and particulate matter with diameter <2.5 μm ($PM_{2.5}$) using traffic, meteorology, and elevation variables. Critical predictors included wind speed and direction for all the pollutants, traffic-related variables for PB-PAH, PNC, and NO_x, and air temperatures and relative humidity for $PM_{2.5}$. GAMs explained 50%, 55%, 46%, and 71% of the variance for log or square-root transformed concentrations of PB-PAH, PNC, NO_x, and $PM_{2.5}$, respectively, an improvement of 5% to over 15% over the linear models. Accounting for temporal autocorrelation in the GAMs further improved the prediction, explaining 57−89% of the variance. We concluded that traffic and meteorological data are good predictors in estimating on-road traffic-related air pollutant concentrations and GAMs perform better for nonlinear variables, such as meteorological parameters.



I. INTRODUCTION

Numerous studies have linked traffic-related air pollutant exposures to adverse health effects including respiratory illnesses, cardiovascular diseases,[1] pregnancy outcomes, and mortality.[2,3] Exposures to traffic-related pollutants are strongly influenced by time spent near traffic emission sources, such as in-vehicle travel, because in the commuting environment, concentrations of traffic related pollutants like ultrafine particles (UFP) and volatile organic compounds can be as much as an order of magnitude higher than in ambient outdoor environments.[4−6] It has been estimated that around 33−45% of UFP[7] and 30−55% of diesel particulate matter (PM)[8] exposure for nonsmoking urbanites in Los Angeles comes from population average time in vehicles. Our previous work indicated that in-vehicle travel time explained approximately 48% of the variance in daily exposure to particle-bound polycyclic aromatic hydrocarbons (PB-PAH) using personal measurements.[9]

Only a limited number of epidemiological studies[10−13] have specifically examined exposure to traffic-related air pollutants from commuting, including two in Southern California.[11,12] Ritz and Yu[12] found an increased risk of low birth weight for women who traveled more than 60 min to work [unadjusted odds ratio (OR): 5.57; 95% confidence interval (CI): 1.16−26.8] using a census-based measure of commuting level. McConnell et al.[11] reported an association of severe wheeze with commuting time in asthmatic children and the association was stronger in analysis restricted to children with commuting times 5 min or longer (adjusted OR: 1.97; 95% CI: 1.02−3.77).

Accurate exposure assessment during commute requires measurements or predictions of on-road concentrations. In our previous work, we developed the models for linking roadway concentrations to in-vehicle concentrations.[14,15] These models can predict in-vehicle particle number concentrations based on driving and vehicle characteristics and ventilation setting, if roadway concentrations are known. However, few on-road

Received: March 22, 2013
Revised: July 4, 2013
Accepted: July 16, 2013
Published: July 16, 2013

concentration models have been developed. Of the most relevant studies, Fruin et al.[7] developed multiple linear regression models that explained up to 60−70% of the variance in the concentrations of particle number (PNC), black carbon (BC), nitric oxide (NO), and PB-PAH on the arterial roads and freeways in Los Angeles. Recently, Aggarwal et al.[16] used two-way stratified multilinear regression to predict UFP number concentrations on Minnesota freeways with a varying performance (R^2: 0.41−0.89) across different size distributions. The previous studies were generally based on linear models and limited by the sampling time of day and sampling routes (mainly on freeways).

Although nonlinear relationships may exist between pollutant concentrations and predictor variables (e.g., meteorology),[17] multiple linear regression has been mostly used in ambient[18,19] and on-road[7,16] air pollution exposure assessment except for a few studies. Singh et al.[20] modeled ambient nitrogen dioxide (NO_2) and sulfur dioxide (SO_2) concentrations using polynomial regression and artificial neural network. Several studies obtained improved results using generalized additive models (GAMs) to predict ambient concentration of UFP and $PM_{2.5}$[21,22] as well as NO_x and NO_2.[23,24] Compared with the other nonlinear models, GAMs provide a more flexible modeling framework because of their capabilities of utilizing both quantitative and qualitative variables, and using a semiparametric rather than parametric approach to capture the nonlinear relationship. This provides the potential for better fits to measurements.[25]

The aim of this study was to examine the associations between on-road pollutant concentrations and predictive variables (traffic, meteorology, and elevation), and to develop robust models to estimate on-road concentrations of four important air pollutants, namely, PB-PAH, PNC, NO_x, and $PM_{2.5}$. The on-road models developed in this study can be combined with air exchange rates and inside-to-outside ratios to estimate in-cabin concentrations and personal exposures in commuting.[14]

II. MATERIALS AND METHODS

Study Region. The study region (Figure S1 of Supporting Information) in the metropolitan Los Angeles area covers 3120 km^2 and includes Los Angeles and Orange counties in Southern California. This region has a high population density of 2702 inhabitants per square kilometer and is one of the most densely populated urbanized areas in the United States.[26] It encompasses a high density of complex roadway networks and has high levels of traffic congestion,[27] which contributes notably to the air pollution problem in the region.

Measurement of Air Pollutant Concentrations. A hybrid vehicle (2010 Honda Insight) was used as a mobile measurement platform that was operated in "green mode" which shuts off the engine when stationary. This generally removes any chances of sampling platform exhaust (Section 2.1 of Supporting Information for details). The instruments were powered using marine batteries and drew air samples from a fan-driven sampling duct installed across the rear windows to effectively decrease instrument response time. For consistency, the mobile monitoring platform was driven in the central freeway lane, when possible. The sampling routes (Figure S1 of Supporting Information) included six major commuter and truck transport freeways, and some arterial and local roads, totally covering over 210 miles of roads (approximately 75% on freeways and 25% on surface streets) during 20 days ranging from March 25 to June 16, 2011 (5:00 a.m. to 23:00 p.m.). Most of the measurements were conducted periodically (4−10 h a day with a run on a sampling route lasting about 4−5 h) during 18 weekdays and 2 weekends.

On-road concentrations were measured for four key air pollutants: PB-PAH was measured using EcoChem PAH Analyzer (Model PAS 2000; detection limit: 3 ng/m^3); PNC was measured using a condensation particle counter (CPC, TSI Inc. model 3007; detection size range 10−1000 nm and detection limit <0.01 particles/cm^3); NO_x was measured using 2-B Technology NOx analyzer (model 401−410; resolution: higher of 1.5 ppb or 2% of reading); $PM_{2.5}$ was measured using TSI Dust-Trak DRX (Model 8533; detection range: 0.001−100 mg/m^3). Regular flow and zero reading checks were conducted to ensure data quality (Section 2.2 of Supporting Information). Instruments were periodically calibrated and time was synchronized to be within 1 s with the Global Positioning System (GPS) device (Garmin GPSMAP 76CSC, position: <10 m, typical; velocity: 0.05 m/s steady state). The GPS device also recorded speed of the vehicle and elevation. Data were recorded by instruments at 1−10 s intervals, which were visually aligned to adjust for instrument response time and then time-averaged into one-minute concentrations for model development.

Predictor Variables. On-road pollutant concentrations are affected by on-road emission sources, regional background concentrations, meteorology, and elevation.

Traffic variables tested included the following [(1)−(4)]:

(1) Roadway type serves as an indicator for traffic volume and roadway configuration.[7,16] Road data were extracted from the ESRI street database based on the 2003 TeleAtlas roadway network (http://www.esri.com). Roadway types were recorded based on GPS field observations and calibrated against the classification of ESRI street database (Section 3.1 of Supporting Information). In model development, we classified four types of roadways as dummy variables: freeway/highway connectors (the roads connecting to different freeways/highways), freeways/highways, major arterials, or local roads.

(2) Real-time traffic and diesel truck counts were compiled from the comprehensive database on freeways and highways at a 5-min resolution based on measured total traffic counts and estimated truck counts from the California Department of Transportation (Caltrans) Performance Measurement System (PeMS) (http://pems.dot.ca.gov/) (1720 counters in total). Since the real-time PeMS measurements only covered 64−69% (in length) of the freeway/highway roads of the sampling routes, these variables may be unsuitable for locations without these traffic data, but are of interest to assess their predictive power. Therefore, as an alternative to limited PeMS data, we also obtained the segment-level 2002 annual average daily traffic (AADT) counts (the latest year available) that were produced by Caltrans staff based on a combination of measurements and modeled values. The AADT data covered continuous road segments for freeways/highways and major surface streets. Length-weighted AADT was calculated as [sum (AADT X road length on each segment)/sum (road length for all segments)]. We selected the 500 m buffer a priori because we wanted to incorporate local traffic

impacts while avoiding influences from background and regional sources. We also tested the influence of different buffer distances and 500 m seemed a reasonable choice (Section 3.2 and Figure S2 of Supporting Information).

(3) The number of roadway lanes was derived from the Caltrans roadway data as an indicator of design volume/capacity of the roadways.

(4) Traffic speed (miles/hour) was derived from the vehicle speed based on the GPS device on the mobile platform. In the field measurement, the driver always attempted to follow traffic (compared to speeding or too slow), thus vehicle speed generally reflected surrounding traffic speed. Traffic speed varied by roadway type (local roads/arterial vs freeways/highways)[15] and traffic condition (e.g., congestion).

Because pollutant concentrations are also strongly influenced by meteorological parameters such as air temperature, wind, and humidity,[28] we also examined the following meteorological parameters [(5)−(6)] in the models:

(5) Hourly ambient meteorological parameters (air temperature, °C; relative humidity, %; wind speed, meters/second abbreviated as m/s; and wind direction) were obtained from the nearest 14 weather monitoring stations operated by National Weather Service and South Coast Air Management District (Figure S1 of Supporting Information). Wind variables were incorporated into the models as the product terms by multiplying wind speed with sine and cosine functions of wind direction, with positive sine value representing wind from the east and positive cosine value representing wind from the north.[29] Some studies[29,30] have shown that using the product terms of wind speed by direction is a good way to incorporate both wind variables in the GAM. Additionally, we tested the predictive power of the product terms of wind speed multiplying sine and cosine functions of the angle between wind direction and roadway orientation.

(6) On-road air temperature (°C) and relative humidity (%) differed from ambient temperature and humidity and were collected simultaneously with pollutant measures at 10-s temporal resolution, and averaged over 1 min. On-road meteorological parameters were recorded using the TSI Qtrak monitor.

(7) Although elevation has been used to model ambient air pollution,[31−33] no studies have incorporated it in modeling on-road pollutant concentrations. We examined the elevation (meter) for each sampling location based on the 10-m resolution remote sensing images from the U.S. National Elevation Data (NED) set (http://nationalmap.gov/), as a potential predictor.

Data Analysis and Selection of Predictors. Exploratory data analysis was conducted, i.e., summary statistics, box plots for identifying outliers, $Q-Q$ plots for normal transformations, correlation analysis (including correlation coefficients and scatter plots) for examining the linear or nonlinear relationships between predictive variables and concentrations (or their transformations), as well as comparison by groups of roadway type, ambient wind speed, and air temperature to investigate their respective influence on variation of on-road concentrations. We used R 2.11.1 (Bell Laboratories, New Jersey, USA) for all the analysis. Section 4.1 of Supporting Information presents technical details for the data analysis.

Correlation analysis was used for variable screening. To avoid multicollinearity issues, variance inflation factors (VIFs) were then used to identify the weakly correlated variables (VIF < 10) and highly correlated (VIF ≥ 10) groups of variables (traffic group and meteorology group). Backward-selection was iteratively conducted until the optimal set of variables were selected with the maximum R^2 or minimum Akaike's information criterion (AIC)[24] (more details in Section 4.2 of Supporting Information).

Predictive Models. We examined and compared three models: linear regression, nonlinear generalized additive models (GAMs), and autoregressive nonlinear models.

Linear regression model is a widely used regression model in estimating ambient air pollution. Its version with factor variables as dummy variables was detailed in Munro.[34] We also compared the relative effects of predictable variables on different scales on the on-road pollutant concentrations (the outcome variable) in linear regression by standardizing the predictive and outcome variables as the standard score (z-score), i.e., the number of standard deviations an observation or datum is off its mean over the valid measurement periods (thus removing the difference in units).[35]

Multivariable GAMs incorporate both continuous (quantitative) and categorical (qualitative) variables, as well as linear and nonlinear relationships. The models specify a distribution (e.g., normal or binomial) of the dependent variable and a link function, g relating the expected value of the distribution to the m predictor variables, and attempt to fit functions $f_i(x_i)$ to satisfy:

$$g(\hat{\mu}_u) = \beta_0 + \sum_{i=1}^{q} f_i(x_u^i) + f_w(w_{s_u}, w_{c_u}) + \sum_{j=q+1}^{q+p} \beta_j x_u^j + \sum_{k=q+p+1}^{q+p+m} \text{factor}(x_u^k) + \varepsilon \quad (1)$$

where $\hat{\mu}_u$ is the estimate of the expected on-road concentration at the location, u; β_0 is the model's intercept; w_{s_u} ([wind speed]·sine([wind direction]) and w_{c_u} ([wind speed] · cosine([wind direction]) represent the product terms of wind speed by direction, at u; and x_u^i, x_u^j, and x_u^k are other independent variables among which x_u^i are q continuous nonlinear variables, x_u^j are p continuous linear variables, and x_u^k are m dummy variables as factors. $fw(w_{s_u},w_{c_u})$ and $f_i(...)$ are the nonparameter smooth functions used to construct the nonlinear relationships between (w_{s_u}, w_{c_u}) or x_u^i and $g(\hat{\mu}_u)$, β_j is the linear coefficient for x_u^j, and $g(...)$ is the link function of the expected value and the independent variables. Here, we assumed that the log or square-root transformation of the average concentration was normally distributed based on the normality test ($Q-Q$ plot) under which assumption, $\hat{\mu}_u = g(\hat{\mu}_u)$. In GAMs, each categorical variable (x_u^k, e.g., roadway type) was transformed by the factor functions to dummy variables with their differential intercept coefficients solved.[36] ε was the normal random error term ($\varepsilon \sim N(0, \sigma^2)$). Since the smooth functions may be fit using parametric or nonparametric means, the GAM provides the potential for more adaptive fits to data than other methods. We used Wood's integrated approach[25] (mgcv package for R) for model selection and automatic smoothing parameter selection with the generalized cross-validation (GCV) criterion to

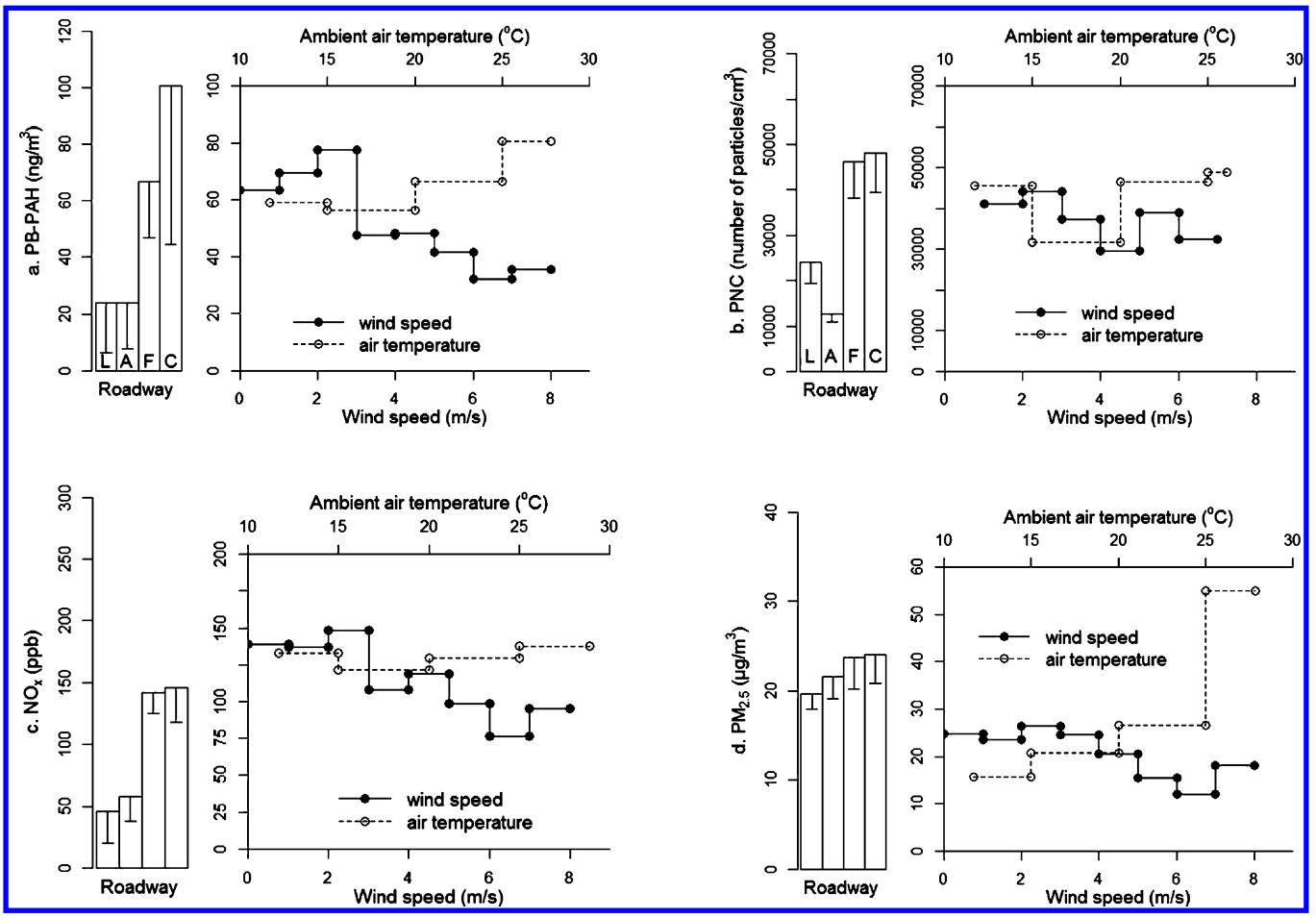


Figure 1. Comparison of air pollutant concentrations by roadway (L: local roads, A: arterial roads, F: freeways and highways, C: freeway/highway connectors; mean: the bar, median: the short line close to the bar's top), and by ambient wind speed and air temperature (mean for the interval).

determine the smoothing parameters (Section 5 of Supporting Information). This approach selects the optimal degrees of freedom for the derivative-based penalized thin plate splines, and thus the smoothed splines can properly represent the observed correlation trend while minimizing overfitting.

Autoregressive nonlinear models incorporate significant temporal autocorrelation that the continuous measurement data may have due to the short averaging time (one minute). With significant temporal autocorrelation, the uncertainty of the predictive coefficients may be underestimated. We used the autocorrelation function (ACF) and partial autocorrelation function (PACF) to measure temporal autocorrelation and developed autoregressive models based on eq 1. Here, ε includes serially correlated errors and is not negligible. The errors from regression models were assumed to be not independent in the time series data and that the process generating the regression errors was stationary. That is, all of the errors had the same expectation and the same variance (σ^2), and the covariance of two errors depended only upon their separation s in time. After the empirical test of the residuals from the independent model (eq 1), we found autocorrelated regression errors were sufficiently described with a term for the first-order autoregressive process, AR(1):

$$\varepsilon_t = \phi\varepsilon_{t-1} + \nu_t \quad (2)$$

where the 'random shocks' ν_t are assumed to be Gaussian white noise, $\nu_t \in N(0, \sigma)$. Similar to Zwack et al.,[21] we used GAMM to conduct autoregressive nonlinear modeling in *R*.

Model Validation. We used the 10 times × 10 folds cross validation (CV) procedure proposed by Arlot et al.[37] that is suitable for time-series analysis. In the CV, the sampling data on each day were evenly divided into 10 segments by time (more details in Section 6 of Supporting Information). One segment was selected as test data (this was repeated 10 times so that each segment was used once as test data). For each segment of test data, training data came from the remaining 9 segments with the constraint that an interval of at least 10 min between the measurement time of the training samples and that of the test samples was maintained to avoid temporal autocorrelation between the test and training data. The above procedures were repeated 10 times to derive the mean R^2 of CV results. Section 6 of Supporting Information presents the specific procedures for the CV. Our data showed little temporal autocorrelation (<0.2) for measurements at 10 min or longer time intervals. We examined the CV *R*-squared (R^2) between measured values and predicted values. We also evaluated the generalizability of the model by conducting independent tests for each day using the data of all the other days to train the model and for each freeway or highway using the data of all the other routes to train the model.

Additional Analysis. Since real-time total traffic and estimated truck counts were only available on 64−69% of the

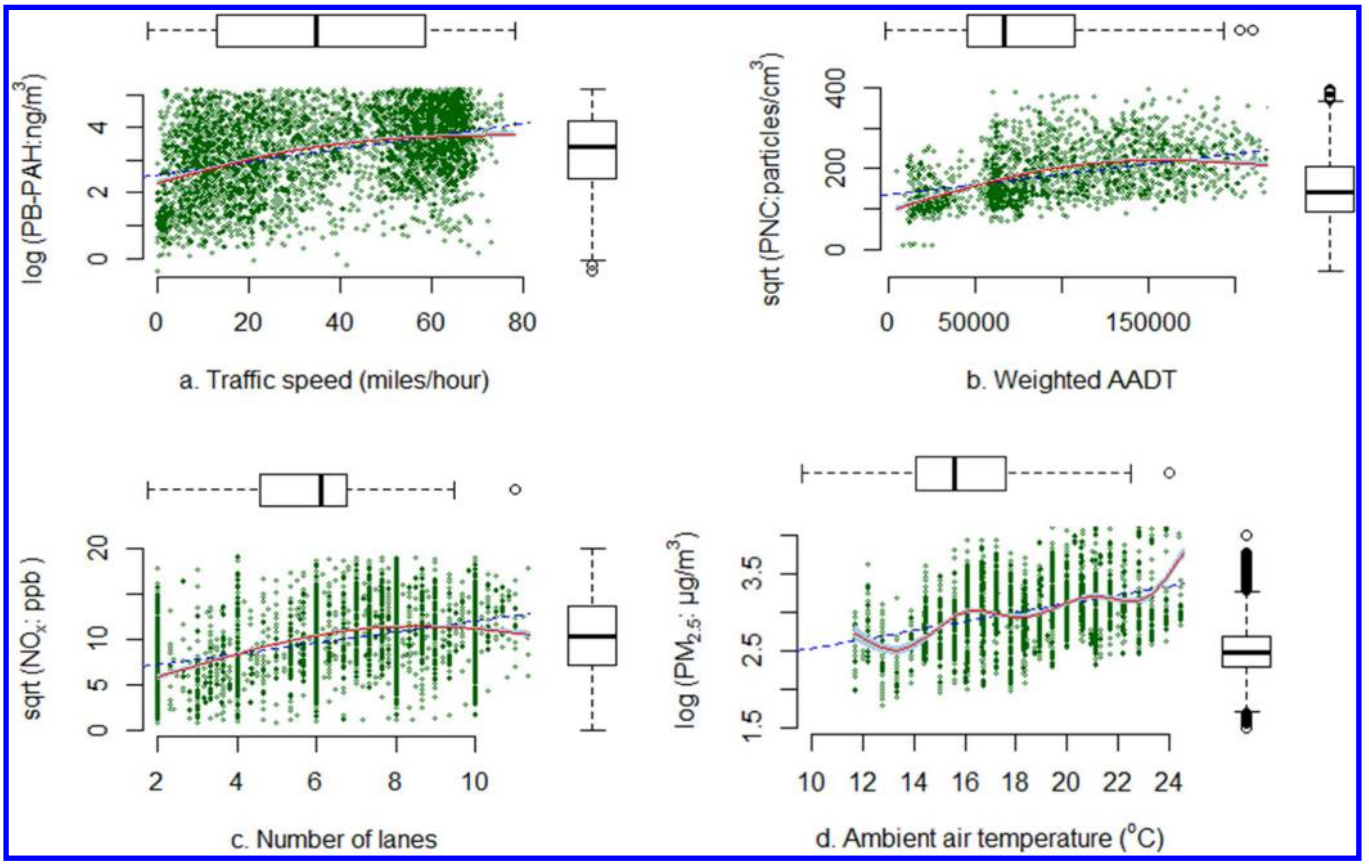


Figure 2. Scatter plots of four covariates with log or square-root (sqrt) transformed pollutant concentrations with GAM-fitted lines (red solid curve) and linear regression lines (blue dashed line).

freeway/highway routes surveyed, we examined the effectiveness of the two variables for modeling traffic-related pollutants on freeways/highways. Further, we examined the predictive power of five ambient pollutants, i.e., NO_2, NO_x, carbon monoxide (CO), $PM_{2.5}$, and SO_2 from government-operated air monitoring stations as extra predictors. We matched the minute-level on-road samples with hourly ambient pollutant concentrations by time and the shortest distance to the monitoring stations.

III. RESULTS AND DISCUSSION

On-Road Pollutant Concentrations. The average on-road concentration was 57.7 ng/m^3 for PB-PAH, 35010 particles/cm^3 for PNC, 118.9 ppb for NO_x, and 23.1 μg/m^3 for $PM_{2.5}$. Summary statistics for one-minute average concentrations are listed in Table S1 of Supporting Information. Data loss was 12.9% for PB-PAH, 55.4% for PNC, 0.4% for NO_x, and 25% for $PM_{2.5}$, mostly due to low reliability or complete loss due to instrument malfunction. PB-PAH and $PM_{2.5}$ were log-transformed and PNC and NO_x were square-root transformed to be normal. The $Q-Q$ plots (Figure S3 of Supporting Information) showed that the transformations of concentrations were normally distributed.

The measurements of air pollutant concentrations were consistent with the previous studies except seasonal and regional differences (Table S2 of Supporting Information). The average PNC, NO_x, and $PM_{2.5}$ concentrations of two earlier studies in Los Angeles[6,7] were higher than ours since their measurements were mainly in winter and spring 2003 and stable atmospheric conditions occur more often in the cool season than in the warm season.[24] Further, the implementation of the air quality regulations, especially for the goods movement corridors, also led to lowering of pollutant concentrations in more recent years.[38]

Relationship between Air Pollutant Concentrations and Predictor Variables. Higher concentrations were generally observed on freeways/highways and their connectors, and at lower wind speed (Figure 1). All differences were statistically significant by Student t and Wilcoxon statistics (Table S3 of Supporting Information). Moderate to strong linear Pearson's or Spearman's correlation (>0.35) was observed between traffic-related variables and transformed concentrations of PB-PAH, PNC, and NO_x (Table S4 of Supporting Information). Among the traffic-related variables, real-time total traffic and estimated truck counts were only weakly positively correlated (0.1−0.3) with pollutant concentrations, likely due to partial spatial coverage of the traffic data and uncertainty in estimated truck counts. In addition, large differences in truck counts were accounted for by the road type variable. Road-length weighted AADT had slightly higher correlation with transformations of concentrations than AADT. Scatter plots (Figure 2 a−c) also showed linearly increasing trends between traffic variables and transformed concentrations. Unlike the other three pollutants, $PM_{2.5}$ correlations with air temperatures and wind speed were stronger than that with traffic variables, reflecting the secondary photochemical origins of much of $PM_{2.5}$ in Southern California.

Linear correlation may obscure significant contributions of nonlinear variables to the prediction of concentrations.[21,24] In Figure 2, we paired typical linear (blue dashed) and nonlinear

dx.doi.org/10.1021/es401281r | Environ. Sci. Technol. 2013, 47, 9291−9299

(red solid) regression lines in the scatter plots. For ambient meteorological variables such as wind product terms (Figure S4 of Supporting Information) and air temperature (Figure 2 d), nonlinear correlative patterns could more objectively represent such a nonmonotonic trend even though its Pearson's linear correlation was not high.

Predictor Variables. We found that traffic-related variables and wind product terms were the most important contributors to PB-PAH, PNC, and NO_x, while air temperatures, wind product terms, and relative humidity were the most important predictors for $PM_{2.5}$. Table 1 lists the final models, including the variance explained by each variable and the coefficients for linear regressors or the degrees of freedom for nonlinear regressors (in the paired parentheses) and Table S5 of Supporting Information lists differential intercept coefficients for the factor variable, roadway type. Figure S5−S8 of Supporting Information show the fitted spline plots of the variables in GAM and the curves show the associated trends between predictor variables and the transformed concentrations.

Traffic Predictors. Traffic variables (including roadway type, weighted AADT, traffic speed, and number of lanes) together accounted for a significant portion of the variance explained in both linear regression (35−40%) and GAM models (12−23%) for traffic-related pollutants. The traffic variables presented a closely linear correlation (increasing trend) with concentrations. In the standardized linear regression, the influence of each traffic-related variable was similar across the traffic-related pollutants, PB-PAH, PNC, and NO_x, i.e., coefficients for standardized independent traffic variables only differed slightly among different pollutants (Table S6 of Supporting Information). The trends of associations between predictors and concentrations were also similar (increasing) in GAM for PB-PAH, PNC, and NO_x (Figure S5−S8 of Supporting Information). This is expected since these variables indicate traffic emission sources whose strength is likely linearly related with concentrations.

Among the traffic-related variables, roadway type was a significant predictor, accounting for about 10.2−11.8% of the total variance in linear regression and 5.0−12.2% of the total variance in GAM. Differential intercept coefficients for freeways, highways, and their connectors were much higher than those for local roads and arterials, indicating several-fold higher pollutant concentrations on freeways and highways than arterial and local roads (approximately 4.5 times for PB-PAH, 3.6 times for PNC, 3.0 times for NO_x) (Figure S4-a, b, c of Supporting Information). Our result is consistent with the previous studies: Westerdahl et al.[6] reported that roadway type strongly influenced variations of on-road concentrations of PB-PAH, BC, and NO_x; Fruin et al.[7] also found that PNC concentrations on arterial roads were roughly one-third of those on freeways. However, roadway type was not used as a predictor in previous studies that focused mainly on freeways or highways.

Due to insufficient spatial and temporal coverage (64−69%) on the freeways/highways, the PeMS 5 min average total traffic and estimated truck counts were statistically insignificant and thus not selected in the final models. In comparison with the study of Fruin et al.,[7] we discussed use of traffic and truck counts based on the sensitivity test on the freeways/highways in Section 7 of Supporting Information (Table S7 and S8).

Traffic speed, as indicator for traffic emission sources, was positively associated with concentrations and was the only real-

Table 1. Variables Selected in Regression Models for Prediction of Concentrations

Covariates	PB-PAH (log ng/m^3)		PNC (square-root number of particles/cm^3)		NO_x (square-root ppb)		$PM_{2.5}$ (log μg/m)	
Models	LR[a] V(C)[c]	GAM[b] V(D)[d]	LR[a] V(C)[c]	GAM[b] V(D)[d]	LR[a] V(C)[c]	GAM[b] V(D)[d]	LR[a] V(C)[c]	GAM[b] V(D)[d]
Intercept	1.9	2.5	143.1	151.5	4.9	6.7	−1.4	2.9
Traffic speed (miles/hour)	14.3% (0.006)	1.6% (8)	17.1% (0.31)	3.9% (6)	15.3% (0.03)	1.9% (6)	-	-
Weighted AADT (traffic count)	9.9% (2.5×10^{-6})	2.4% (8)	11.0% (1.7×10^{-6})	5.3% (6)	9.3% (1.45×10^{-5})	3.8% (8)	-	1.9% (6)
Number of lanes	2.3% (0.03)	0.8% (6)	-	1.9% (4)	0.35% (0.06)	1.4% (3)	-	2.0% (5)
Ambient air temperature (°C)	0.1% (0.011)	8.1% (8)	0.4% (0.63)	8.3% (8)	0.2% (0.06)	12.0% (6)	34.4% (0.09)	24.9% (8)
On-road air temperature (°C)	-[e]	-	-	-	-	-	4.8% (0. 08)	11.8% (8)
On-road relative humidity (%)	-	-	-	-	-	-	10.8% (0.02)	13.2% (8)
Wind speed and wind direction (m/s)	2.1% (−0.13)	21.2% (10)	4.3% (−8.89)	17.8% (10)	1.2% (−0.62)	17.5% (10)	1.3% (−0.10)	14.8% (10)
Elevation (m)	0.9% (-7.4×10^{-3})	3.7% (8)	0.2% (−0.41)	4.5% (8)	0.5% (−0.02)	4.0% (4)	-	0.4% (3)
Roadway type[f]	11.8%	12.2%	11.8%	12.3%	10.2%	5.0%	-	2.4%
Total variance explained	41.5%	50.0%	44.9%	54.5%	37.2%	45.6%	51.3%	71.0%

[a]LR: linear regression. [b]GAM: generalized additive regression. [c]V(C): V is the variance explained; C is the linear coefficient for changes in transformed concentration of air pollutant for one unit change in the variable predictor. [d]V(D): V is the variance explained; D is the degrees of freedom, see Figure S4-S8 of Supporting Information for a graphic depiction of linear and nonlinear relationship between various predictors and air pollutant concentrations. [e]− indicates not used in the model due to multicollinearity or lack of statistical significance. [f]Roadway type is a qualitative variable. Table S5 of Supporting Information lists the intercept coefficients for different types.

dx.doi.org/10.1021/es401281r | Environ. Sci. Technol. 2013, 47, 9291−9299

time on-road variable selected in the models for traffic-related pollutants. The removal of traffic speed (the other variables remained unchanged) only slightly influenced the prediction performance of the models of the traffic-related pollutants (Table S9 of Supporting Information). In linear regression, the removal of traffic speed was compensated by more variance explained by weighted AADT in the absence of the traffic speed variable (20.7−23.6% vs 9.3−11.0%).

Meteorological Predictors. Compared with traffic variables, meteorological variables had nonlinear relationships with the pollutant concentrations (Figure S4−S8 of Supporting Information). For example, ambient air temperature presented a nonmonotonic trend with $PM_{2.5}$ (Figure 2-d). In particular, the product terms of wind speed by direction presented a more complex relationship (varying surfaces, Figure S4 of Supporting Information) with the concentrations.

Among the meteorological variables, wind speed and direction were important predictors for all the pollutants. Strong winds were associated with lower pollutant concentrations (Figure 1) but pollutant concentrations were not linearly correlated with wind speed and thus the contribution of wind speed and direction was much higher in GAM than linear regression (14.8−21.2% vs 1.2−4.3%). Further, we found that the product terms of wind speed by direction and those of wind speed by angle to roadway generated similar results in model performance (Table S10 of Supporting Information).

Meteorological rather than traffic-related variables contributed considerably to $PM_{2.5}$ concentrations (overall R^2: 0.66−0.71), which agrees with previous literature indicating $PM_{2.5}$ is a regional pollutant[39] that is more affected by regional or background concentrations than by local traffic contributions.[40] Particularly, hourly ambient air temperature had a stronger influence on $PM_{2.5}$ than the other three pollutants (accounting for 24.9−34.5% of variance for $PM_{2.5}$ vs 0.1−12.0% for the other three pollutants). Real-time on-road air temperature and relative humidity were also significant predictors for $PM_{2.5}$ (Table 1), but not for the other pollutants. The removal of on-road air temperature and relative humidity from the models decreased the variance explained by approximately 10.0−14.0% for $PM_{2.5}$ (Table S9 of Supporting Information). The significant contribution of air temperatures on $PM_{2.5}$ concentration is expected since there was a positive correlation between air temperature and photochemical conversion and oxidation of gaseous PM precursors to PM mass, which was higher in the summer.[41] Interestingly, ambient and on-road air temperatures were just moderately correlated (Pearson's correlation: 0.49) and did not produce multicollinearity (VIFs in linear regression were <10: 3.8 and 2.3, respectively), allowing both variables to be used in the $PM_{2.5}$ model. Other than the difference in the temporal resolution of measurements (by hour vs minute), on-road temperature measurements may reflect the combined effects of ambient temperature and waste engine heat, hot pipe emissions, and the warm-up of roads and asphalt.

Elevation. In our models, elevation had a small contribution. Although having limited variation in our measurement data, elevation may likely influence on-road pollutant concentrations because of different local emissions and pollutant dispersion patterns in hilly areas.[28,31]

Model Performance. Cross validation results are shown in Table 2. Linear regression had moderate predictive power (CV R^2: 0.36−0.51). GAM had moderately better predictive power in general (CV R^2: 0.46 for PB-PAH, 0.50 for PNC, 0.43 for NO_x, 0.66 for $PM_{2.5}$), improving the variance explained by about 7% to 15% over linear regression. Overall, the R^2 for the independent tests of model generalizability by day and by freeway/highway (Table S11 of Supporting Information), although slightly lower, were similar to the results of the cross validation tests.

Table 2. Cross Validation Results for Linear Regression and GAM

		linear regression		generalized additive model	
		General[a]	CV[b]	General[a]	CV[b]
PB-PAH	Samples	3740	3336	3740	3336
	R^2	0.42	0.38	0.50	0.46
PNC	Samples	1778	1570	1778	1570
	R^2	0.45	0.43	0.55	0.50
NO_x	Samples	4434	3950	4434	3950
	R^2	0.37	0.36	0.46	0.43
$PM_{2.5}$	Samples	3405	3041	3405	3001
	R^2	0.51	0.51	0.71	0.66

[a]General: no cross validation. [b]CV: cross validation.

For PB-PAH, PNC and NO_x, the traffic variables (such as traffic speed and weighted AADT) accounted for less variance in the GAM than in linear regression but meteorological variables (ambient air temperature and the wind product terms) accounted for more variance in the GAM. As a nonparametric approach, GAMs can more efficiently model nonlinear relationships (such as those between meteorological variables and the concentrations). But for a predictor (such as traffic speed) closely linearly related to the target variable, GAMs may not achieve significant gains over linear regression, as demonstrated in our test of univariate models (Table S12 of Supporting Information).[42,43] Further, in a multivariate GAM, the predictive power may not be simply an additive function of the contribution of each variable.[25] In other words, the effect of a predictor depends on the other predictors that may be potential confounding indicators.[44] In GAMs, the addition of nonlinear meteorological predictors that improved the predictive power adversely affected (confounded) the predictive power of the traffic variables. We also tested the overall predictive power of the combined set of traffic variables (without meteorological variables included) in the multivariate models and the result (Table S13 of Supporting Information) showed that the GAM had slight improvement (by 3−5% in the variance explained) over linear regression for traffic-related pollutants (PB-PAH, PNC, and NO_x). The above comparisons show that choice of the models (linear regression vs GAMs) is important for predictive power of the nonlinear variables such as meteorological ones.

Temporal autocorrelation (based on ACF) of 1 lag (one minute) was 0.63−0.70, indicating strong temporal autocorrelation. The model that incorporated lag 1 (one minute) temporal autocorrelation (AR1 = 0.63−0.70) had better CV R^2 (0.57 for PB-PAH, 0.68 for PNC, 0.72 for NO_x, 0.89 for $PM_{2.5}$), a significant improvement over the GAM in the R^2 by 11% to more than 20% for the four pollutants (Table S14 of Supporting Information). The application of autoregressive models may be unpractical in epidemiological studies where measurement data of time series are usually difficult to acquire.

Influence of Ambient Air Pollutant Concentrations. The significant contribution of ambient meteorological variables in the GAM was attributed to their nonlinear

relationship with the on-road pollutant concentrations that was influenced by urban-scale meteorological and air pollutant phenomena.[24] The sensitivity test (Table S13 of Supporting Information) using ambient air pollutants as predictors shows that the ambient concentrations performed similarly to the meteorological variables. The incorporation of ambient air pollutant concentrations along with meteorological variables in the models slightly to moderately improved the model performance. For $PM_{2.5}$, a regional pollutant, addition of ambient pollutant concentrations had the highest improvement (20% for linear regression and 10% for GAM).

Limitations. This study has several limitations. First, the models were based on 1 min average concentrations. Compared to the models using a longer averaging time, the residuals of our models were temporally autocorrelated and produced overly small confidence limits in the linear regression model. Longer averaging times (e.g., 5 min), however, reduced the sample size and increased uncertainties in the variables, particularly roadway and traffic variables, as 5 min travel on freeways can be 10 km in distance. With 5 min averaging time, model performance was not as good (e.g., R^2 ranged 0.23−0.30 for linear regression and 0.35−0.44 for GAM). Second, due to limited spatiotemporal coverage for total traffic counts and uncertainty in truck count estimates, traffic and truck counts were not directly used in our final models although they were significant regressors in the freeway/highway models. Their related alternatives such as roadway type used in the final models captured most but not all of the spatial variability in the counts. Third, 2002 AADT was used with 2010 on-road concentrations and other predictor variables to train the model. This temporal nonalignment may have produced some bias, although AADT explained only a small fraction of the observed variance (2.4−11.0%). Finally, overfitting is always a risk in nonlinear GAM. In our case, the degree of freedom in our GAM was 5−12, generally considered acceptable to ensure that overfitting will not occur with more than 1500 samples with a large variance.[25] Furthermore, the mgcv package used for modeling controlled the complexity of the splines by imposing a penalty on the parameters of the splines, lowering the overfitting risk.[25]

IV. IMPLICATIONS

In a metropolitan area with a high density of population and complex roadway networks, we found that traffic variables (traffic speed and weighted AADT) were linearly correlated with traffic-related pollutants (PB-PAH, PNC, and NO_x) and explained most of the total variance in linear regression for these pollutants. Compared to linear regression, the nonparametric GAM more adequately captured the nonlinear relationship between meteorological variables (e.g., the product terms of wind speed by direction, air temperatures) and air pollutant concentrations, thus improving the total variance explained by 19−23% over linear regression for traffic-related pollutants and 39% for $PM_{2.5}$. For future studies, traffic variables (e.g., at least roadway type) should be examined in models for traffic-related air pollutants, while meteorological variables should be examined for regional pollutants such as $PM_{2.5}$. Short-term exposure assessment and health effects studies may require similar exposure estimates at a high temporal resolution (e.g., daily or even hourly). In this study, we suggest the use of GAM rather than linear regression since it would be favorable to incorporate meteorological impacts, and as demonstrated in this paper, the relationship between air pollutant concentrations and meteorological parameters are likely nonlinear.

The measurements of ambient pollutant concentrations, if added into the model as predictors, had a slight or moderate improvement in the prediction. Therefore, ambient air pollutant variables, if available, should also be examined and used in models for future studies.

Our study is one of the first studies on the prediction of on-road pollutant concentrations. Among the few published studies, Fruin et al.[7] was based on arterial roads (2.5 h total over two days) and freeways (12 h total over four days) in Los Angeles (R^2: 0.60−0.70), while Aggarwal et al.[16] was based on Minnesota freeways (40 h total over 19 days in summer) (R^2: 0.41−0.89). Compared with the previous studies, our measurements covered a much longer time (approximately 112 h total over 20 days) and longer and more diverse routes (approximately 210 miles including local roads, arterial, and freeways/highways). Although the previous models had a good performance, they are limited to specific conditions with narrower applications, whereas our models have more general applications to other locations, times, and air pollutants. Further, the two previous studies were based on linear models, while our study demonstrated the usefulness of the GAM approach in modeling nonlinear variables such as meteorological parameters.

Our study identified linear relationships between traffic variables and on-road concentrations of traffic-related air pollutants, and nonlinear relationships between meteorological variables and the on-road concentrations. The inherit relationship (linear vs nonlinear) between predictors and the air pollutant dependent variable determines the utility of linear regression or GAM for the exposure modeling. In this study, GAM performed better for nonlinear variables (e.g., meteorological variables) and for the prediction of $PM_{2.5}$, the on-road concentration of which was more greatly influenced by meteorology and regional background particle concentrations rather than local traffic.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional information regarding materials and methods used, technical details, and results. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: 949-824-0548, Fax: 949-824-0529, e-mail: junwu@uci.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was supported by the California Air Resources Board through contract number #07-310.

■ REFERENCES

(1) Brugge, D.; Durant, L. J.; Rioux, C. Near-highway pollutants in motor vehicle exhaust: a review of epidemiologic evidence of cardiac and pulmonary health risks. *Environ. Health* **2007**, *6*, 23.

(2) Chen, H.; Goldberg, M. S.; Villeneuve, P. J. A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases. *Rev. Environ. Health* **2008**, *23*, 243−297.

(3) Stanek, W. L.; Brown, S. J.; Stanek, J.; Gift, J.; Costa, L. D. Air pollution toxicology-a brief review of role of the science in shaping the current understanding of air pollution health risks. *Toxicol. Sci.* **2011**, *120*, S8−S27.
(4) Jo, W. K.; Park, K. H. Commuter exposure to volatile organic compounds under different driving conditions. *Atmos. Environ.* **1999**, *33*, 409−417.
(5) Zhu, Y.; Eiguren-Fernandez, A.; Hinds, W. C.; Miguel, A. H. In-cabin commuter exposure to ultrafine particles on Los Angeles freeways. *Environ. Sci. Technol.* **2007**, *41*, 2138−2145.
(6) Westerdahl, D.; Fruin, S.; Sax, T.; Fine, P. M.; Sioutas, C. Mobile platform measurements of ultrafine particles and associated pollutant concentrations on freeways and residential streets in Los Angeles. *Atmos. Environ.* **2005**, *39*, 3597−3610.
(7) Fruin, S.; Westerdahl, D.; Sax, T.; Sioutas, C.; Fine, P. M. Measurements and predictors of on-road ultrafine particle concentrations and associated pollutants in Los Angeles. *Atmos. Environ.* **2008**, *42*, 207−219.
(8) Fruin, S.; Winer, A.; Rodes, C. Black carbon concentrations in California vehicles and estimation of in-vehicle diesel exhaust particulate matter exposures. *Atmos. Environ.* **2004**, *38*, 4123−4133.
(9) Wu, J.; Tjoa, T.; Li, L.; Jaimes, G.; Delfino, R. Modeling personal polycyclic aromatic hydrocarbon (PAH) exposure in human subjects in southern California. *Environ. Health* **2012**, *11*, 47.
(10) Greenwald, R.; Sarnat, J.; Sarnat, S.; Yip, F.; Boehmer, T. *In-cabin air pollution exposure and acute respiratory response in healthy and asthmatic automobile commuters*; B17. How Bad is Traffic Pollution? American Thoracic Society 2012 International Conference, The American Thoracic Society; San Francisco, CA, 2012.
(11) McConnell, R.; Liu, F.; Wu, J.; Lurmann, F.; Peters, J.; Berhane, K. Asthma and school commuting time. *J. Occup. Environ. Med.* **2010**, *52*, 827−828.
(12) Ritz, B.; Yu, F. The effect of ambient carbon monoxide on low birth weight among children born in southern California between 1989 and 1993. *Environ. Health Perspect.* **1999**, *107*, 17−25.
(13) Zuurbier, M.; Hoek, G.; Oldenwening, M.; Meliefste, K.; Hazel, P.; Brunekreef, B. Respiratory effects of commuters' exposure to air pullution in traffic. *Epidemiology* **2011**, *22*, 219−227.
(14) Fruin, S.; Hudda, N.; Sioutas, C.; Delfino, R. Predictive model for vehicle air exchange rates based on a large, representative sample. *Environ. Sci. Technol.* **2011**, *45*, 3569−3575.
(15) Hudda, N.; Eckel, S. P.; Knibbs, L. D.; Sioutas, C.; Delfino, R.; Fruin, S. A. Linking in-vehicle ultrafine particle exposures to on-road concentrations. *Atmos. Environ.* **2012**, *59*, 578−586.
(16) Aggarwal, S.; Jain, R.; Marshall, D. J. Real-time prediction of size-resolved ultrafine particulate matter on freeways. *Environ. Sci. Technol.* **2012**, *46*, 2234−2241.
(17) Bloomfield, P. J.; Royle, J. A.; Steinber, L. J.; Yang, Q. Accounting for meteorological effects in measuring urban ozone levels and trends. *Atmos. Environ.* **1996**, *30*, 3067−3077.
(18) Thompson, L. M.; Reynolds, J.; Cox, H. L.; Guttorp, P.; Sampson, D. P. A review of statistical methods for meteorological adjustment of troposheric ozone. *Atmos. Environ.* **2001**, *35*, 617−630.
(19) Hoek, G.; Beelen, R.; 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.
(20) Singh, P. K.; Gupta, S.; Kemar, A.; Shukla, P. S. Linear and nonlinear modeling approaches for urban air quality prediction. *Sci. Total Environ.* **2012**, *426*, 244−255.
(21) Zwack, M. L.; Paciorek, J. C.; Spengler, D. J.; Levy, I. J. Modeling spatial patterns of traffic-related air pollutants in complex urban terrain. *Environ. Health Perspect.* **2011**, *119*, 852−859.
(22) Liu, Y.; Paciorek, J. C.; Koutrakis, P. Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology and land use information. *Environ. Health Perspect.* **2009**, *117*, 886−892.
(23) Hart, E. J.; Yanosky, D. J.; Puett, R.; Ryan, J.; Dockery, W. D.; Smith, J. T.; Garshick, E.; Laden, F. Spatial modeling of PM_{10} and NO_2 in the continental United States, 1985−2000. *Environ. Health Perspect.* **2009**, *117*, 1690−1696.
(24) Li, L.; Wu, J.; Wilhelm, M.; Ritz, B. Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in southern California. *Atmos. Environ.* **2012**, *55*, 220−228.
(25) Wood, S. *Generalized Additive Models: An Introduction with R*; Taylor & Francis: Florida, 2006.
(26) *List of United States Urban Areas*; US Census Bureau, USA, 2010; http://www.census.gov/geo/www/ua/uafacts.html.
(27) Schrank, D.; Lomax, R.; Eisele, B. *TTI's 2011 Urban Mobility Report*; Texas A&M Transportation Institute, TX, 2011.
(28) Admassu, M.; Wubeshet, M. *Air Pollution*; University of Gondar: Gondar, 2006.
(29) Carslaw, C. D.; Beevers, D. S.; Tate, E. J. Modeling and assessing trends in traffic-related emissions using a generalized additive model. *Atmos. Environ.* **2007**, *41*, 5289−5299.
(30) Li, L.; Wu, J.; Ghosh, K. J.; Ritz, B. Estimating spatiotemporal variability of ambient air pollutant concentrations with a hierarchical model. *Atmos. Environ.* **2013**, *71*, 54−63.
(31) Bishop, A. G.; Morris, A. J.; Stedman, H. D. The effects of altitude on heavy-duty diesel truck on-road emissions. *Environ. Sci. Technol.* **2001**, *35*, 1574−1578.
(32) Briggs, D.; Hoogh, C.; Gulliver, J.; Wills, J.; Elliott, P.; Kingham, S.; Smallnone, K. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *Sci. Total Environ.* **2000**, *253*, 151−167.
(33) Larson, T.; Henderson, S.; Brauer, M. Mobile monitoring of particle light absorption coefficient in an urban area as a basis for land use regression. *Environ. Sci. Technol.* **2009**, *43*, 4672−4678.
(34) Munro, B. H. *Statistical Methods for Health Care Research*, 4thed.; Lippincott Williams & Wilkins: Philadelphia, 2001.
(35) Kreyszig, E. *Applied Mathmatics*, 4th ed.; Wiley Press: New York, 1979.
(36) Wooldridge, M. J. *Introductory Econometrics: A Modern Approach*; South-Western Cengage Learning: Mason, OH, 2009.
(37) Arlot, S. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40−79.
(38) *Annual Summary Report Calendar Year 2012*; Port of Long Beach Air Quality Monitoring Program, 2012; http://caap.airsis.com/ReportsPOLA.aspx.
(39) Russell, M.; Allen, D. T.; Collins, D. R.; Fraser, M. P. Daily, seasonal, and spatial trends in $PM_{2.5}$ mass and composition in Southeast Texas. *Aerosol Sci. Technol.* **2004**, *38*, 14−26.
(40) Zhou, Y.; Levy, J. I. Factors influencing the spatial extent of mobile source air pollution impacts: a meta-analysis. *BMC Public Health* **2007**, *7*, 89.
(41) *The Particle Pollution Report: Current Understanding of Air Quality and Emissions through 2003*; US Environmental Protection Agency, NC, 2004; http://www.epa.gov/airtrends/aqtrnd04/pmreport03/report_2405.pdf#page=1.
(42) Hastie, T. J. Generalized additive models. In *Statistical Models in S*, Chambers, J. M., Hastie, T. J., Eds.; Wadsworth and Brooks: Pacific Grove, CA, 1992; pp 249−308.
(43) Vislocky, L. R.; Fritsch, M. Generalized additive models versus linear regression in generating probabilistic MOS forests of aviation weather parameters. *Weather Forecast* **1995**, *10*, 669−680.
(44) Jbilou, J.; Adlounin, S. Generalized additive models in environmental health: a literature review, In *Novel Approaches and Their Applications in Risk Assessment*; Luo, Y., Ed.; InTech: Croatia, 2012; DOI: 10.5772/2548.