Published in final edited form as:
Atmos Environ. 2012 August 1; 55: 220–228. doi:10.1016/j.atmosenv.2012.03.035.

Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in Southern California

Lianfa Li[a,b], **Jun Wu**[a,*], **Michelle Wilhelm**[c], and **Beate Ritz**[c]

[a]Program in Public Health, College of Health Sciences, University of California, Irvine, USA

[b]State Key Lab of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, China

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

Abstract

Land-use regression (LUR) models have been developed to estimate spatial distributions of traffic-related pollutants. Several studies have examined spatial autocorrelation among residuals in LUR models, but few utilized spatial residual information in model prediction, or examined the impact of modeling methods, monitoring site selection, or traffic data quality on LUR performance. This study aims to improve spatial models for traffic-related pollutants using generalized additive models (GAM) combined with cokriging of spatial residuals. Specifically, we developed spatial models for nitrogen dioxide (NO_2) and nitrogen oxides (NO_x) concentrations in Southern California separately for two seasons (summer and winter) based on over 240 sampling locations. Pollutant concentrations were disaggregated into three components: local means, spatial residuals, and normal random residuals. Local means were modeled by GAM. Spatial residuals were cokriged with global residuals at nearby sampling locations that were spatially auto-correlated. We compared this two-stage approach with four commonly-used spatial models: universal kriging, multiple linear LUR and GAM with and without a spatial smoothing term. Leave-one-out cross validation was conducted for model validation and comparison purposes. The results show that our GAM plus cokriging models predicted summer and winter NO_2 and NO_x concentration surfaces well, with cross validation R^2 values ranging from 0.88 to 0.92. While local covariates accounted for partial variance of the measured NO_2 and NO_x concentrations, spatial autocorrelation accounted for about 20% of the variance. Our spatial GAM model improved R^2 considerably compared to the other four approaches. Conclusively, our two-stage model captured summer and winter differences in NO_2 and NO_x spatial distributions in Southern California well. When sampling location selection cannot be optimized for the intended model and fewer covariates are available as predictors for the model, the two-stage model is more robust compared to multiple linear regression models.

Keywords

Land-use regression; Spatial residuals; Generalized additive model; Cokriging; Traffic air pollution

[*]Corresponding author. Program in Public Health & Department of Epidemiology, Anteater Instruction & Research Bldg (AIRB) # 2034, University of California, Irvine, CA 92697-3957, USA. Tel.: +1 949 824 0548; fax: +1 949 824 0529. junwu@uci.edu. .

Appendix. Supplementary data Supplementary data related to this article can be found online at doi:10.1016/j.atmosenv.2012.03.035.

1. Introduction

An increasing body of literature links exposure to traffic-related air pollutants to mortality and morbidity (Aguilera et al., 2008;Bassok et al., 2010; Iniguez et al., 2009). Air pollutant concentrations are influenced by both local sources (e.g., traffic exhaust and industrial emissions) and regional/background pollutant contributions from atmospheric transport and chemistry. Developing adequate models for traffic exposures is essential for health effects studies.

Kriging has been used widely to model spatial dependence of air pollutant concentrations (Beelen et al., 2009). Traditional kriging estimates pollutant concentrations at un-sampled locations as the combination of mean predictions from the nearby samples or polynomial trend models of spatial coordinates with smoothing of the residuals (Johnston et al., 2003). An increasing number of studies recently have developed land-use regression (LUR) models to estimate local variability in air pollutant concentrations. These models estimated local pollutant concentrations using spatially-resolved variables reflecting emission intensity, proximity to sources, and atmospheric dispersion (Hart et al., 2009; Hoek et al., 2008). Unlike kriging that takes into account spatial autocorrelation of residuals, most of the previous LUR models assumed independence of residuals. Many kriging and LUR models have performed moderately well as assessed by *r*-square (R^2) values, ranging from 0.52 to 0.76 (Hart et al., 2009).

In terms of sampling approach for spatial model development, some studies employed a location-allocation algorithm for locating monitoring sites (Su et al., 2009a, 2009b). This approach, in combination with development of a distance decay regression selection strategy, resulted in a much improved LUR model performance [R^2 of 0.88 for nitrogen dioxide (NO_2) and 0.91 for nitrogen oxides (NO_x) for models developed in Los Angeles, California] (Su et al., 2009b). However, the location-allocation approach may be impractical for many air pollution epidemiological studies due to logistical and funding constraints. The Multi-Ethnic Study of Atherosclerosis and Air Pollution study arranged the majority of monitors in clusters of six, with three on either side of a major road at distances of approximately 50, 100, and 300 m from the road (Mercer et al., 2011). This approach maximizes the variability of measured concentrations near roadways but may be insufficient to capture spatial variability of concentrations across the entire study region.

Most existing LUR models were developed using multiple linear regression. Generalized additive models (GAM) can capture both linear and non-linear relations between covariates and air pollution concentrations, are semi-parametric and multiple-dimensional, and use penalized splines (Hastie, 1990). GAM has previously been used to predict spatially-resolved concentrations of fine particulate matter and NO_2 based on land-use, traffic, satellite, and/or meteorological data in two studies, and both included spatial smoothing terms of sampling site coordinates as predictor variables in the models (Hart et al., 2009; Liu et al., 2009).

Although several studies reported moderate spatial autocorrelation in residuals (Hystad et al., 2011; Liu et al., 2009; Su et al., 2009a), most previous LUR models did not take into account spatial autocorrelation of residuals, which may moderately or substantially affect the predicted values for regional pollution surfaces (Mercer et al., 2011; Paciorek, 2010). Furthermore, most previous LUR models focused on annual average concentrations. An exception is Mercer et al.’s recent study (2011) that predicted spatial variability of NO_x in three seasons using a two-stage model combining linear LUR with universal kriging although their model performance was no better than that of most previous studies (R^2 up to

0.75). Some studies added temporal profiles using data from nearby continuous monitoring stations (Aguilera et al., 2009;Brauer et al., 2010), but the assumption of stable spatial variability over time may be invalid (Wu et al., 2011). The estimation of seasonally changing exposures is important when examining health outcomes with shorter periods of vulnerability (e.g. pregnancy outcomes). In addition, there may be considerable variability between seasons for coefficient estimates, thus it may be inappropriate to combine seasonal data to develop a single annual average model to account for spatial variability of pollutant concentrations over seasons.

In this study, we combined GAM with cokriging of spatial residuals to estimate summer and winter NO_2 and NO_x pollution surfaces in Southern California, assuming that spatial variations in concentrations are substantially influenced by both local variation due to proximity to emission sources and by global variation due to atmospheric transport (Ainslie et al., 2008;Beelen et al., 2009; Liu et al., 2009). Local means were predicted by GAM, while spatial residuals from GAM were assumed to be second-order stationary (Gartan and Guyon, 2010) and modeled through cokriging with global residuals (representing global variations) at sampling locations nearby. We compared the GAM plus cokriging model with four other methods: universal kriging, multiple linear LUR, and GAM with and without spatial smoothing of coordinates. We also examined the impact of sampling location selection and different types of predictor variables on the model performance.

2. Materials and methods

2.1. Study domain

The study domain covered Los Angeles and Orange counties in Southern California, with an area of 160 × 161 km^2 and over 12 million residents in 2008. The urban core of the area (Los Angeles-Long Beach-Santa Ana) had a population density of 2729 inhabitants per km^2 and was the most densely-populated urbanized area in the United States (U.S. Census Bureau, 2000). The region has been one of the most polluted places in the country (American Lung Association, 2011). Port emissions, trucks, automobiles airports, and industry contributed to local air pollution problems (Kunzli et al., 2003). The region encompassed the nation's largest marine port complex (American Association of Port Authorities, 2010) and had six major commuter and truck transport freeways. It is also bordered by mountain ranges and frequently experiences surface inversion with limited vertical mixing (Lu and Turco, 1994). All of these factors contributed to high pollutant concentrations in the region.

2.2. Measured NO_2 and NO_x concentrations

2.2.1. Measurements collected by University of California, Los Angeles (UCLA)—All NO_2 and NO_x samples (N = 201) were collected in Los Angeles County using passive air samplers from Ogawa & Company USA, Inc. (Pompano Beach, FL) in two continuous weeks in a late summer warm season (September 9–22, 2006) and a mid-winter season (February 10–23, 2007). The sampling locations were selected using a location-allocation algorithm that maximized the potential variability in measured concentrations and the spatial distribution of the health study population (i.e., participants in the Los Angeles Family and Neighborhood Study) (Su et al., 2009b). Each sampler was deployed for a 2-week period (maximum variability was 72 h). There were a total of 181 valid measurements in each season. Co-located samples were collected at 14 South Coast Air Management District (SCAQMD) stations.

2.2.2. Measurements collected by University of California, Irvine (UCI)—All NO_2 and NO_x samples were collected using the same Ogawa samplers as above in south Los

Angeles County and Orange County in two alternate weeks in the summer (July 10–18 and July 24-August 1) and the winter (November 13–21 and December 4–12) in 2009. The sampling sites were residential outdoor locations of 45 pregnant women who enrolled in our prospective Air Pollution and Birth Outcomes Study and 9 volunteers who were employees or students at UCI. Co-located samples were collected at 11 SCAQMD stations. Each sampler was deployed for a one-week period (maximum variability was 24 h). Overall, we obtained valid measurements at 53 sites in the summer and 64 sites in the winter.

2.2.3. Government-based monitoring data from SCAQMD—Hourly NO_2 and NO_x measures were obtained from the SCAQMD monitoring network at 31 sites (including sites collocated with passive samplers) in Southern California. The measurements were conducted by federal designated automated chemiluminescence methods using active instruments. To remove systematic bias between different types of instruments (passive vs. active), we aggregated hourly active concentrations measures and regressed them against integrated passive measurements at the collocated SCAQMD sites during the same sampling periods. We then converted all passive measurements to concentrations based on the chemiluminescence methods. The correlation coefficients between the passive and active measurements were all above 0.9 (Supplemental Materials Table S1).

2.3. Spatial data and covariates

2.3.1. Roadway data—We obtained roadway data from the ESRI StreetMap™ North America 9.3 (http://www.esri.com). This dataset included 2003 TeleAtlas® street polylines, which – as we previously demonstrated – are more accurate than TIGER 2000-based streets (Wu et al., 2005). We classified roadways into four categories based on the U.S. Census Feature Class Code (U.S. Census Bureau, 1993): primary highways, typically interstates, with limited access (A1); primary roads without limited access, non-interstate roads (A2); smaller, secondary or connecting roads, usually with more than two lanes (A3); and local, neighborhood and rural roads, usually with a single lane of traffic in each direction (A4). We calculated the nearest distance from a sampling site to each roadway type and the total roadway length within different buffer sizes around the sampling site.

2.3.2. Total traffic and truck volumes—We compiled a comprehensive traffic database for freeways and major surface streets in the study region based on measurements and estimated values. Hourly total traffic and truck counts on freeways and highways were obtained from the California Department of Transportation (Caltrans) Performance Measurement System (PeMS) (http://pems.dot.ca.gov/). The truck counts from the PeMS algorithm were not measured but estimated from 5-min aggregated count and occupancy data (Urban Crossroads, Inc., 2006). We averaged the total and truck traffic counts at the PeMS monitoring sites for each of our sampling periods and assigned the point PeMS data to adjacent roadway segments (<300 m) with matching street names, which were then extended to contiguous road segments (within 15 km) with matching street names. For surface streets without hourly measurements, we used Caltrans annual average daily traffic (AADT) counts derived from a combination of tri-annual measurements and estimated values. We calculated inverse distance-weighted vehicle and truck counts within different buffer sizes around each sampling location (Wilhelm and Ritz, 2003).

2.3.3. Population density—We obtained block level population data from U.S. Census 2000 (U.S. Census Bureau, 2004). Population kernel density was calculated in ArcGIS (version 9.3; ESRI, Redlands, CA) using a 5000 m search radius and a 30 m resolution.

2.3.4. Land-use type—We obtained the 2001 land-use data from the Southern California Association of Government (SCAG) (Aerial Information Systems, 1996; Park and

Stenstrom, 2008). The SCAG land-use data were first developed using 1990 aerial and photographs and updated using computer interactive photo interpretation techniques and digital orthophotography with 1 m resolution (Park and Stenstrom, 2008). We classified the original 108 land-use types into four major categories: transportation; agriculture, open space and vacant; industrial; and residential. We calculated the percentage of area for each land-use category within different buffer sizes (50 m–15 km) around each sampling location.

2.3.5. Remote Sensing data—We obtained 30 m × 30 m ETM+data from Landsat's thematic mapper (http://landsat.gsfc.nasa.gov/) for 3–5 cloud-free days in each sampling period. From the ETM+ data, we extracted land surface temperature (LST) and a normalized difference vegetation index (NDVI) using Environment for Visualizing Images (ENVI) software (ITT Visual Information Solutions, Boulder, CO). LST was calculated using the transform equation between temperature and radiance (YCEO, 2010); NDVI was calculated using a standard algorithm and the data near infrared band and red band (ENVI, 2011).

2.3.6. Atmospheric stability—We obtained Pasquill atmospheric stability classes every 3 h at approximately 40 km by 40 km spatial resolution from the National Oceanic and Atmospheric Administration (NOAA) AIR Resources Laboratory archive of the Eta 4-D Data Assimilation System (EDAS) (http://www.arl.noaa.gov/ready.html). We assigned the atmospheric stability from the nearest modeling grid to each sampling site. We classified stability classes E, F and G as stable, A, B, and C unstable, and D as neutral. The percentage of time with stable air conditions was calculated for each sampling period.

2.4. Modeling approach

2.4.1. Selection of spatial covariates for prediction of local means—The decay buffering and correlation analysis method (Su et al., 2009a) was used to select optimized buffer between 50 m and 15 km for all spatial variables besides nearest distance to the roadways and atmospheric stability. A variable was dropped from further analysis if the absolute correlation coefficient with the measured concentrations was less than 0.3.

2.4.2. GAM: local mean modeling—The general equation for predicting air pollution concentrations at the location, u was:

$$\widehat{y}_{\mathrm{u}}=\widehat{\mu}_{\mathrm{u}}(X)+\widehat{\varepsilon}_{\mathrm{us}}(Z)+\varepsilon_{\mathrm{un}}\mathrm{with}\widehat{y}_{\mathrm{u}} \geq 0 \quad (1)$$

where $\widehat{y}_{\mathrm{u}}$ is the estimated concentration at location u, $\widehat{\mu}_{\mathrm{u}}$ is the estimate of local mean at u, determined by spatial covariates, X, $\widehat{\varepsilon}_{\mathrm{us}}$ is the estimate of the spatial residual at u, determined by spatial residuals of neighborhood samples around u, $Z\sum\mathrm{Nb}(\mathrm{u})$, and $\varepsilon_{\mathrm{un}}$ is a random residual at u, with normal distribution, $\varepsilon_{\mathrm{n}}\sim N(0,1)$, and ignored.

We used the GAM package in R statistical software (R version 2.11.1) to conduct the GAM part (local mean modeling) of our model. Since both NO_2 and NO_x concentrations followed a normal distribution in this study (−1 < skewness < 1, Supplemental Materials Fig. S1), no log transform was required. The following is the GAM equation for local mean:

$$\mathrm{g}\ (\widehat{\mu}_{\mathrm{u}})=\mu_0+\sum_{i=1}^{m} f_i\left(\mathscr{X}_{\mathrm{u}}^{i},\mathrm{df}\right)+\sum_{j=m+1}^{n}\beta_j\mathscr{X}_{\mathrm{u}}^{j}\mathrm{with}\mathrm{g}\ (\widehat{\mu}_{\mathrm{u}})\geq 0 \quad (2)$$

where μ_0 is the model intercept, x_{u}^{i} or $x_{\mathrm{u}}^{j} \in X$ are local covariates, $f_{\mathrm{i}}(\ldots)$ is the smooth function consisting of series basis functions (representing the non-linear relationship), df is degrees of freedom that controls the smooth degree of the curve fit, β_{j} are the linear parameters used to construct the linear relationship between x_{u}^{j} and $\mathrm{g}\ (\widehat{\mu}_{\mathrm{u}})$, and n is the number of covariates. For normally distribution, the link function is $g(\mu_{\mathrm{u}}) = \mu_{\mathrm{u}}$. For log transformed concentrations, we can remove the constraint in [2], $\mathrm{g}\ (\widehat{\mu}_{\mathrm{u}}) \leq 0$. To avoid the problem of over-fitting, we used small degrees of freedom (2–8) in our models. Smooth terms of spatial location were not incorporated in the model because we considered spatial variability in the second stage by cokriging spatial residuals.

We selected the covariates in three steps. To avoid multi-collinearity, we first divided the covariates into two groups: one group of weakly correlated covariates with variance inflation factors (VIFs) < 10 and independent groups of highly correlated covariates with VIF ≥ 10 (O'Brien, 2007). From each group of the highly correlated covariates we selected one variable at a time and combined them with all the weakly correlated covariates to construct a combination of covariates for the model. Then Akaike's information criterion (AIC) or R^2 was used to further backward-select the variables in each combination: the covariates with p values ≥0.1 were removed until R^2 remained the same, improved, or decreased least when all possible combinations of the remaining covariates were considered. Finally, the covariate combination with the maximum R^2 or minimum AIC was selected as optimal inputs in the model.

2.4.3. Cokriging of spatial residuals to minimize error variance—Spatial residuals from GAM were cokriged with global residuals at nearby sampling locations, assuming that after removal of local means, the space domain is stable:

$$\widehat{\varepsilon}_{\mathrm{us}} = \sum_{i=1}^{n_{\mathrm{u}}} \lambda_{\mathrm{u}_i}^{\mathrm{u}} \widehat{\varepsilon}_{\mathrm{us}}\ (Z_{\mathrm{u}_i}) + \sum_{i=1}^{n_u} \lambda_{\mathrm{u}_i}^{\mathrm{g}} \varepsilon_{\mathrm{g\ s}}\ (Z_{\mathrm{u}_i}) \quad (3)$$

where 3 $\boldsymbol{\varepsilon} = [\varepsilon_{\mathrm{us}}]\sim N(0,\mathrm{V}(\theta))$, θ is the vector of variogram parameters,$Z = [Z_{\mathrm{u}i}]$ is the set of neighborhood samples around u, $\varepsilon = [\varepsilon_{\mathrm{us}}]\tilde{}N\ (0, V\ (\theta))$, θ is the estimate of spatial residuals at $z_{\mathrm{u}i}$, and is derived by subtracting the GAM-predicted local mean from the measured or observed concentration at each sampling site, $\varepsilon_{\mathrm{gs}}(\mathrm{z}_{\mathrm{u}i})$ is the estimate of the global residual or the total variation of the measured values on a regional scale, deriving by subtracting the global mean concentration (average of measured concentrations at all sites) from the measured concentration at each sampling site. $\lambda_{\mathrm{u}_i}^{\mathrm{u}}$ and $\lambda_{\mathrm{u}_i}^{\mathrm{u}}$ are the optimal weights generated by maximum likelihood based on the estimates of θ.

Spatial residuals are influenced by both local variability (affected by local covariates) and background/regional variability (affected by global trend). Variogram was used to model the spatial residuals and regional variability not captured by the GAM. Variogram reflects a feature's variability along a certain distance in a spatial field. According to the optimal principle of unbiased estimation and minimal error variance of cokriging, error variance of spatial residuals will decrease substantially if variogram of spatial and global residuals is precisely captured (Goovaerts, 1997). We used theoretical variogram to fit the experimental variogram of spatial and global residuals and the cross covariance between them. Two parameters were used to describe the variogram: sill (the maximum value of the variogram as the lag distance approaches infinity) and range (the distance where the difference of the variogram from the sill becomes negligible). A longer range and a smaller sill indicate a more continuous spatial surface. We examined the semi-variogram cloud, tested different lag sizes and number of lags, and different variogram models to find the best reasonable fit using ArcGIS's Geostatistical Analyst.

2.5. Model comparison

We compared our GAM plus cokriging model with four other methods: universal kriging, multiple linear LUR, and GAM with and without spatial smoothing terms of coordinates. Universal kriging estimates local means based on coordinates and incorporates spatial residuals without use of supplemental information such as local covariate. The multiple linear LUR assumes a linearly additive relationship between concentrations and spatial covariates. GAM incorporates non-linear relationships (Hart et al., 2009; Liu et al., 2009). We also examined the contribution of cokriging to model improvement by comparing the GAM plus cokriging model with the GAM with and without spatial smoothing terms of coordinates.

Since certain models may be more vulnerable to less-thanoptimal siting of monitors than others, we compared the models using all the data and the UCLA data alone since the UCLA sites were optimally sited through a location-allocation algorithm (no separate models were built for the UCI data because of the small number of sites). We examined whether simple models (e.g. multiple linear LUR) perform as well as more complex models (e.g. GAM plus cokriging) in the context of an optimized monitoring protocol and which model(s) perform better with more arbitrary placing of monitors due to study restrictions. In addition, we compared our modeled annual average concentrations with those from previously published LUR models based on the same UCLA data (Su et al., 2009b). Finally, we examined the contribution of different types of variables on model performance, particularly the factors that may not be easily available in other parts of the world such as traffic counts and atmospheric stability.

2.6. Cross validation

We used leave-one-out cross validation (LOOCV) for model evaluation. LOOCV uses a single observation from the original sample as the validation data and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. We used four measures to compare the performance of the models: R^2, inter-quartile range (IQR) of prediction errors, the square root of the mean of the squared prediction errors (RMSPE), and box plots of precision errors. A higher R^2 value, a smaller IQR and RMSPE, a mean and IQR close to 0, and a narrower confidence interval (CI) were used to identify a better model.

3. Results

3.1. Distance decay correlation analysis

Five types of covariates were selected for training the model: atmospheric stability, NDVI and land surface temperature, distance-weighted vehicle and truck counts, three land-use types, and roadway variables (nearest distance to roadways and road lengths) (see Supplemental Materials Table S2 for Pearson's correlations of all the covariates and their optimal buffering distances).

The four models for summer and winter NO_2 and NO_x each had different sets of predictors (Table 1). NDVI had a moderate predictability in all models, explaining 9.5e25.4% of the variances in measured concentrations. Atmospheric stability was a statistically significant predictor for summer NO_2 and NO_x (variance explained: 13.8–14.3%) but not for winter concentrations. Residential land-use for summer NO_x and winter NO_2 and NO_x had moderate predictive power (variance explained: 7.1e11.8%). Other land-use patterns (transportation, and farm and open fields), length of local roads (A4) within 15 km, and nearest distance to major freeways (A1) and local streets (A4) were included in different models resulting in a range of variances explained. As expected, annual average daily traffic counts had less predictive power than the average daily traffic volume integrated over each sampling period from hourly measurements on freeways and highways (R^2: 0.35–0.40 vs. 0.59–0.78 for NO_2; 0.35–0.41 vs. 0.52–0.74 for NO_x).

3.2. Variogram modeling of spatial residuals

Global residuals had longer ranges (8969–17,105 m vs. 2199–3020 m) and higher sills (15.7–25.6 ppb vs. 5.1–8.5 ppb for NO_2 and 269.5–320.0 ppb vs. 84.0–137.5 ppb for NO_x) than local spatial residuals, indicating a global spatial variability at a regional scale (Supplemental Materials Table S3). Cross covariance between spatial and global residuals was moderate. Variogram of global residuals in the summer had a shorter range (8969 m vs. 10,526 m for NO_2; 10,861 m vs. 17,105 m for NO_x) and higher sills (25.6 ppb vs. 15.7 ppb for NO_2; 320.0 ppb vs. 269.5 ppb for NO_x) than those in the winter (Supplemental Materials Fig. S2 and Table S3). Similarly, variogram of spatial residuals in the summer had a slightly shorter range and slightly higher sills than those in the winter, indicating a higher spatial heterogeneity in the summer than in the winter.

3.3. Comparison of different models

Among the five models, the GAM plus cokriging model had the highest cross validation R^2 (0.88–0.92), IQR closest to zero (−0.31–0.09 ppb), and the smallest RMSPE (1.67–7.83 ppb) (Table 2). For prediction errors, this model also had a narrower 95% CI, closer-to-zero median, and fewer outliers (Fig. 1). The GAM plus cokriging model improved the cross validation R^2 for about 16.7–26.0% over GAM with spatial smoothing, 22.2–35.3% over GAM without spatial smoothing, 33.8–58.6% over universal kriging, and 39.7–53.3% over multiple linear LUR (with coordinates as additional predictors) (Supplemental Materials Fig. S3). GAM with spatial smoothing outperformed GAM without spatial smoothing with 7.0–21.0% higher R^2 and smaller RMSPE. In addition, we found that incorporating the coordinates moderately to substantially improved the performance of the linear LUR (Table 2), likely because the coordinates accounted for partial spatial variability not captured by other variables included in the model. Besides seasonal predictions (Table 2), the GAM plus cokriging model also had the best performance in comparison with the other four models (Supplemental Materials Table S4) for annual average concentrations using all the measurement data.

3.4. Prediction of summer and winter pollution surfaces

Noticeable differences were observed across season for the spatial patterns of pollution surfaces (1 km × 1 km) predicted by the GAM plus cokriging model. For both NO_2 and NO_x, pollution surfaces were steeper in the summer and more continuous in the winter (Fig. 2). In addition, the concentrations were substantially lower in Orange County than Los Angeles in the summer but not in the winter (Supplemental Materials Fig. S4). Despite the heterogeneity of spatial distributions by season, we obtained high levels of agreement between the predictive and measured values in the two sub-regions (Supplemental Materials Fig. S4).

3.5. GAM plus cokriging model vs. multiple linear LUR (UCLA data only)

For seasonal predictions, the GAM plus cokriging model was slightly or moderately better than multiple linear LUR with coordinates (0.88–0.94 vs. 0.79–0.83) (Supplemental Materials Table S5). For annual average predictions, despite differences in the selection of spatial covariates in our study and Su et al. (2009b) study, similar prediction performance was observed between our GAM plus cokriging model, our multiple linear model, and Su et al.'s multiple linear model (R^2: 0.92 vs. 0.87 vs. 0.87 for NO_2; 0.91 vs. 0.87 vs. 0.92 for NO_x) (Supplemental Materials Table S5).

4. Discussion

We developed two-stage models to estimate NO_2 and NO_x concentrations in Southern California. Our GAM plus cokriging model performed well in predicting summer and winter concentrations (R^2 = 0.88–0.92). This study adds to the literature on air pollution exposure assessment in several ways: (1) it is one of the first studies to account for spatial residuals by cokriging them with global residuals; (2) we compared the performance of the two-stage model with four other commonly-used methods; (3) we demonstrated the importance of developing models to capture substantial differences in spatial distributions of pollutants between seasons (summer and winter), which is important for studies of short-term or sub-chronic health effects; and (4) we examined the influence of sampling site selection and different types of spatial covariates on model performance.

Cokriging spatial residuals with global residuals moderately to substantially improved the predictions over the other four methods. Several previous LUR studies examined spatial autocorrelation of residuals. Liu et al. (2009) observed residual spatial auto-correlation using semi-variogram. Su et al. (2009b) and Hystad et al. (2011) tested spatial auto-correlation of the residuals using Moran's I. Paciorek (2010) summarized the general phenomenon of spatial auto-correlation in residuals of regression models. However, most LUR models did not utilize spatial residual information, with an exception of the study of Mercer et al. (2011).Su et al. (2009a, 2009b) partially accounted for spatial autocorrelation by adding a census tract cluster variable or using coordinates as predictors in the models. However, their models did not directly incorporate spatial residuals and did not consider spatial variability of global residuals, which could influence predicted concentrations with regional contributions (e.g. NO_2 and fine particulate matter). Traditional universal kriging models residuals with variogram, but it relies on coordinates to predict local means and does not incorporate potentially important covariates such as traffic, land-use, and meteorology. Mercer et al. (2011) developed a two-stage model combining prediction of local means by linear LUR with universal kriging of the residuals. Our two-stage models achieved higher R^2 than that reported by Mercer et al. (2011) (0.88–0.92 vs. 0.75) for both seasonal and annual NO_x predictions. The consideration of non-linearity and large-scale variability may at least partially account for the better performance of our model.

Interestingly, the cross validation R^2 for the GAM plus cokriging models were similar to the multiple linear LUR results reported by Su et al. (2009b) using only UCLA data (Supplemental Materials Table S5). The use of a location-allocation algorithm approach for site selection might partially account for the good performance of the simple linear LUR. However, it may be difficult to generalize the location-allocation sampling approach because it requires intensive effort in data collection and sometimes may not be practical (Mercer et al., 2011). For example, because of study constraints, the UCI study collected samples at subjects' homes with less-than-ideal placement of sampling sites for spatial model development. Nevertheless, our two-stage model seems to have been able to overcome some of the problems of non-optimal sampling locations and achieved good performance for the entire study region (R^2: 0.88–0.92). The two-stage model may provide better results than the other models when resources are limited or when investigators have to rely on existing data or convenience and non-optimal sampling locations.

We found the spatial distribution of NO_2 and NO_x being more continuous in winter than in summer, similar to the study of Mercer et al. (2011). In this region the atmosphere is less stable in the summer than in the winter, which leads to more rapid disperse of pollutants in the summer and consequently faster decrease in concentrations within a short distance of sources. Our results are also consistent with previous studies that showed narrower impact zones (e.g., 300 m downwind) of primary traffic emissions in the daytime with good mixing

(Zhu et al., 2002) and a wider impact zone (e.g., as far as 2600 m) before sunrise with stable atmosphere (Hu et al., 2009). The NO_2/NO_x ratios predicted by our models were expected: lower ratios near roadway sources and higher ratios in areas with high ozone concentrations (Supplemental Materials Fig. S5).

Variables that may not be readily available in other regions (e.g. total traffic counts, truck counts, NDVI, and LST) individually increased R^2 by about 6–16% in multiple linear LUR and 1–7% in the two-stage models (Supplemental Materials Table S6, S7 and S8). These covariates did not contribute substantially to the performance of GAM plus cokriging model (Table S8), partly because the two-stage model incorporated spatial autocorrelation of residuals that accounted for a significant portion of variance explained (>22%). Accounting for the spatial autocorrelation of residuals somewhat compensates for the loss in prediction power due to missing covariates. This has significant implications because in many locations around the world, researchers may not have access to the variables we employed. Our results suggest that one can probably still develop reliable spatial models for NO_2 and NO_x as long as sufficient sampling data and some limited covariate data are available. Although incorporating the influence of residuals may mask the importance of certain prediction variables not accounted for in the models, it is a practical approach for improving exposure assessment in epidemiologic studies when data on important prediction variables are not available (Mercer et al., 2011).

When possible, we used spatial variables specific to each sampling period. Likely due to data limitations, few previous investigators included hourly vehicle and truck count data in their models. We found the variables 'vehicle and truck counts on free-ways and highways' based on hourly averages to be stronger predictors than 'annual average daily traffic counts' (r: 0.52–0.78 vs. 0.35–0.41). Estimated hourly truck counts had similar correlation with pollutant concentrations as hourly vehicle counts (r: 0.60–0.68 vs. 0.59–0.65). NDVI and LST were found to be good predictors for local means of pollutant concentrations. LST may reflect population density and emission sources (e.g. traffic exhaust) while NDVI may reflect less urbanized areas with fewer emission sources.

Atmospheric stability in the summer explained 16.3% and 17.2% of variation in NO_2 and NO_x concentrations, respectively, but had less predictive ability in the winter. Only a few previous studies explored the usefulness of these data for predicting exposure to traffic-related air pollutants (Kwon et al., 2006). We acknowledge that there may be substantial uncertainties in the atmospheric stability data since they were derived from model outputs rather than upper air measurements. In addition, there were less differences in the stability class estimates for the limited number of EDAS modeling grids ($N = 12$) in our study area in the winter than in the summer, which partially explained the poor predictive power of stability class in the winter.

This study has several limitations. First, the measured concentrations were derived from slightly different approaches (i.e. two alternate one-week measurements at UCI and two-week measurements at UCLA in two seasons) and from different years (2006, 2007 and 2009). To minimize potential bias from different monitoring period lengths, we used SCAQMD measurement data as the benchmark and standardized the UCLA and UCI measurements to SCAQMD values. Second, the 2001 SCAG land-use data may be associated with uncertainty and errors. Park and Stenstrom (2008) reported that the SCAG data may contain varying degrees of mixed land-use information and thus may not be as precise as the high resolution Landsat data. However, in this study we only used four land-use categories (i.e. transportation; agriculture, open space and vacant; industrial; and residential) and the buffer size of the land-use covariates in the final models ranged from 2 km to 15 km. We expected that the uncertainty of the SCAG land-use data had limited effects on the covariates of area ratios for different land-use types within the relatively large area. Third, although measures were taken to minimize over-fitting (e.g. restricting the degrees of freedom for each covariate to 2–8 in GAM and no use of the coordinates of the sampling sites in GAM), our two-stage model may still potentially over-fit the data. Finally, we modeled summer and winter seasons separately rather than systematically integrating temporal and spatial variability in the models using more advanced techniques such as those used in Szpiro et al. (2010). This is partly because of the limited number of monitoring sites with long-term NO_2 and NO_x measurements in our study region. More importantly, we focused on temporal variability over longer periods (e.g. season and trimesters during pregnancy), thus the more comprehensive spatial-temporal modeling with a higher temporal resolution is beyond the scope of the current paper.

5. Conclusion

We developed a two-stage model that combined GAM with cokriging of spatial residuals to estimate spatial variability of NO_2 and NO_x in summer and winter seasons in Southern California. Our models generated reasonably accurate predictions for NO_2 and NO_x pollution surfaces and captured summer and winter differences in NO_2 and NO_x spatial distributions. Compared to multiple linear regression models, the two-stage model was more robust in predicting concentration measurements when sampling locations may have been selected in a less-than-optimal manner and when fewer spatial covariates were available as predictors for the model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The study was supported by the National Institute of Environmental Health Sciences (NIEHS R21ES016379) and the Health Effect Institute (HEI 4787-RFA09-4110-3 WU). Collection of the NO_2 and NO_x data by UCLA researchers at the 181 sampling sites was funded by California Air Resources Board Contract No. 04-323. We thank the study subjects and volunteers who participated in the air pollution exposure sampling. We also thank Jane Berner from California Department of Transportation who helped us retrieve the PeMS traffic count data.

References

Aerial Information Systems. Aerial Information Systems. Redlands, California: 1996. Southern California 1990 Aerial Land Use Study: Land Use Level III/IV Classification.

Aguilera I, Sunver J, Fernandez-patier R, Hoek G, Aguirre-alfaro A, Nieuwenhuijsen JM, Herce-garraleta D, Brunekreef B. Estimation of outdoor NO_x, NO_2, and BTEX exposure in a cohort of pregnant women using land use regression modeling. Environmental Science & Technology. 2008; 42:815–821. [PubMed: 18323107]

Aguilera I, Guxens M, Garcia-Esteban R, Corbella T, Nieuwenhuijsen MJ, Foradada CM, Sunyer J. Association between GIS-based exposure to urban air pollution during pregnancy and birth weight in the INMA sabadell cohort. Environmental Health Perspective. 2009; 117:1322–1327.

Ainslie B, Steyn D, Su GJ, Buzzelli M, Brauer M, Larson T, Rucker M. A source area model incorporating simplified atmospheric dispersion and advection at fine scale for population air pollution exposure assessment. Atmospheric Environment. 2008; 42:2394–2404.

American Association of Port Authorities. Port Industry Statistics: North American Port Container Traffic (1990-2009). 2010

American Lung Association. State of the AIR 2011. American Lung Association; Washington, DC: 2011.

Bassok A, Hurvitz MP, Christine Bae C-H, Larson T. Measuring neighbourhood air pollution: the case of Seattle's international district. Journal of Environmental Planning and Management. 2010; 53:23–39.

Beelen R, Hoek G, Pebesma E, Vienneau D, de Hoogh K, Briggs D. Mapping of background air pollution at a fine spatial scale across the European Union. Science of the Total Environment. 2009; 407:1852–1867. [PubMed: 19152957]

Brauer M, Lencar C, Tamburic L, Koehoorn M, Demers P, Karr C. A cohort study of traffic-related air pollution impacts on birth outcomes. Environmental Health Perspective. 2010; 116:680–686.

ENVI. ENVI-User's Guide: Transform. 2011.

Gartan, C.; Guyon, X. Spatial Statistics and Modeling. Springer Secience & Business; New York: 2010.

Goovaerts, P. Geostatistics for Natural Resources Evaluation. Oxford University Press; New York: 1997.

Hart EJ, Yanosky DJ, Puett R, Ryan J, Dockery WD, Smith JT, Garshick E, Laden F. Spatial modeling of PM10 and NO_2 in the Continental United States, 1985–2000. Environmental Health Perspectives. 2009; 117:1690–1696. [PubMed: 20049118]

Hastie, TJ. Generalized Additive Models. Chapman and Hall; New York: 1990.

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. Atmospheric Environment. 2008; 42:7561–7578.

Hu SS, Fruin S, Kozawa K, Mara S, Paulson SE, Winer AM. A wide area of air pollutant impact downwind of a freeway during pre-sunrise hours. Atmospheric Environment. 2009; 43:2541–2549.

Hystad P, Setton E, Cervantes A, Poplawski K, Deschenes S, Brauer M, Donkelaar A, Lamsal L, Martin R, Jerrett M, Demers P. Creating national air pollution models for population exposure assessment in Canada. Environmental Health Perspectives. 2011; 119:1123–1129. [PubMed: 21454147]

Iniguez C, Ballester F, Estarlich M, Llop S, Fernandez-patier R, Agirre-Alfaro A, Esplugues A, Valencia. Estimation of personal NO_2 exposure in a cohort of pregnant women. Science of the Total Environment. 2009; 407:6093–6099. [PubMed: 19740523]

Johnston, K.; Hoef, MJ.; Krivoruchko, K.; Lucas, N. ESRI. , editor. ArcGIS 9: Using ArcGIS Geostatistical Analyst. 2003.

Kunzli N, McConnell R, Bates D, Bastain T, Hricko A, Lurmann F, Avol E, Gilliland F, Peters J. Breathless in Los Angeles: the exhausting search for clean air. American Journal of Public Health. 2003; 93:1494–1499. [PubMed: 12948969]

Kwon J, Weisel C, Turpin B, Zhang J, Korn L, Mordandi M, Stock T, Colome S. Source proximity and outdoor-residential VOC concentrations: results from ROIPA study. Environmental Science & Technology. 2006; 40:4074–4082. [PubMed: 16856719]

Liu Y, Paciorek JC, Koutrakis P. Estimating regional spatial and temporal variability of PM2.5 concentrations using satellite data, meteorology and land use information. Environmental Health Perspectives. 2009; 117:886–892. [PubMed: 19590678]

Lu R, Turco RP. Air pollutant transport in a coastal environment .1. 2-Dimensional Simulations of sea-breeze and mountain effects. Journal of the Atmospheric Sciences. 1994; 51:2285–2308.

Mercer DL, Szpiro AA, Sheppard L, Lindstrom J, Adar DS, Allen WR, Avol LE, Oron PA, Larson T, Liu L, Kaufman DJ. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NO_x) for Multi-Ethnic Study of Attherosclerosis and Air Pollution (MESA AIR). Atmospheric Environment. 2011; 45:4412–4420. [PubMed: 21808599]

O'Brien MR. A caution regarding rules of thumb for variance inflation factors. Quality & Quantity. 2007; 41:673–690.

Paciorek JC. The importance of scale for spatial-counfound bias and precision of spatial regression estimators. Statistical Science. 2010; 25:107–125. [PubMed: 21528104]

Park M, Stenstrom KM. Classifying environmentally significant urban land uses with satellite imagery. Journal of Environmental Management. 2008; 86:181–192. [PubMed: 17291679]

Su GJ, Jerrett M, Beckerman B. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. Science of the Total Environment. 2009a; 407:3890–3898. [PubMed: 19304313]

Su GJ, Jerrett M, Beckerman B, Wilhelm M, Ghosh KJ, Ritz B. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. Environmental Research. 2009b; 109:657–670. [PubMed: 19540476]

Szpiro AA, Sampson DP, Sheppard L, Lumley T, Adar DS, Kaufman DJ. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics. 2010; 21:606–631.

Urban Crossroads, Inc.. PeMS Data Extraction Methodology and Execution Technical Memorandum for the Southern California Association of Governments Southern. California Association of Government; Irvine: 2006.

U.S. Census Bureau. A Guide to State and Local Census Geography. Association of Public Data User; Princeton, NJ: 1993.

U.S. Census Bureau. [Retrieved February 10, 2012] List of United States urban areas. 2000. from: http://www.census.gov/geo/www/ua/ua2k.txt

U.S. Census Bureau. 2000 Census of Population and Housing, Summary Tape File 3A. U.S. Census Bureau; Washing, DC: 2004.

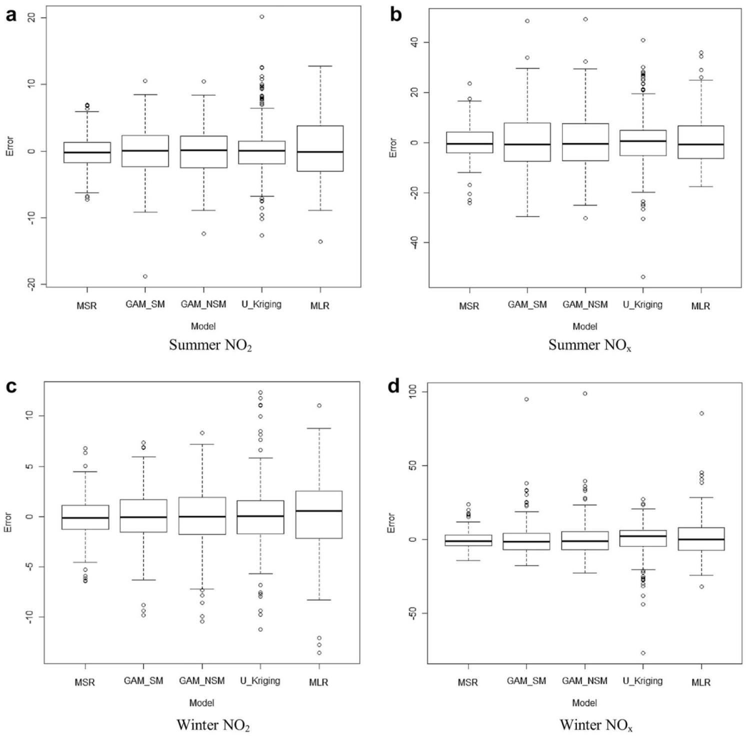
Wilhelm M, Ritz B. Residential proximity to traffic and adverse birth outcomes in Los Angeles County, California, 1994–1996. Environmental Health Perspective. 2003; 111:207–216.

Wu J, Funk TH, Lurmann FW, Winer AM. Improving spatial accuracy of roadway networks and geocoded addresses. Transactions in GIS. 2005:585–601.

Wu J, Wilhelm M, Chung J, Ritz B. Comparing exposure assessment methods for traffic-related air pollution in an adverse pregnancy outcome study. Environmental Research. 2011; 111:685–692. doi:10.1016/j.envres.2011.1003.1008. [PubMed: 21453913]

YCEO. Converting Landsat TM and ETM+ Thermal Bands to Temperature. Yale Center for Earth Observation; 2010.

Zhu YF, W.C. H, Kim S, Shen S, C. S. Study of ultrafine particles near a major highway with heavy-duty diesel traffic. Atmospheric Environment. 2002; 36:4323–4335.



Summer NO_2

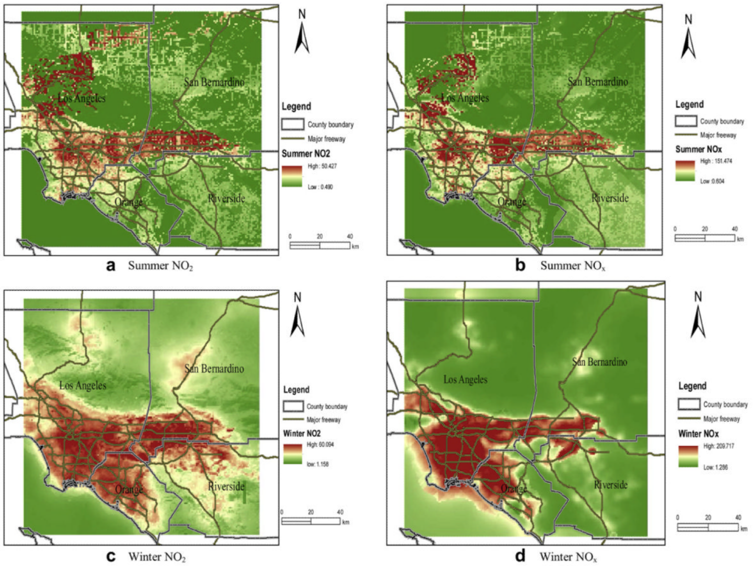
Summer NO_x

Winter NO_2

Winter NO_x

Fig. 1.

Box plots of precision errors for five models of NO_2 (a) and NO_x (b) in summer and NO_2 (c) and NO_x (d) in winter (error bars indicate 95% confidence intervals; circles indicate outliers) (MSR: GAM plus cokriging of spatial residuals; GAM_SM: GAM with spatial spline term for coordinates; GAM_NSM: GAM without spatial spline term for coordinates; U_Kriging: universal kriging; MLR: multiple linear LUR).



a Summer NO_2

b Summer NO_x

c Winter NO_2

d Winter NO_x

Fig. 2.

Prediction of NO_2 and NO_x distribution (1 km × 1 km) in summer and winter using GAM plus cokriging of spatial residuals.

Table 1

Spatial covariates selected for GAM after removing collinearity.

Smooth factors	NO_2						NO_x					
	Summer			Winter			Summer			Winter		
	V.P[a]	Dis[b]	*p*-value	V.P[a]	Dis[b]	*p*-value	V.P[a]	Dis[b]	*p*-value	V.P[a]	Dis[b]	*p*-value
NDVI	13.65	0.25	2.73e-7	9.46	0.20	7.2e-4	25.37	0.25	6.33e-10	12.83	0.50	6.15e-6
Land surface temperature	1.38	1.10	0.013	6.56	0.30	0.03	–	–	–	12.89	0.50	0.026
Weighted traffic flow	4.48	13.00	5.1e-4	–	–	–	2.76	13.00	0.006	8.01	15.00	5.35e-3
Weighted truck flow	13.87	11.00	9.84e-7	–	–	–	6.69	11.00	8.89e-4	5.24	11.00	0.0014
Atmospheric stability	13.79	–	1.27e-7	–	–	–	14.29	–	4.07e-6		–	–
Residential land-use proportion	–	–	–	7.53	15.00	7.69e-4	11.81	15.00	8.92e-4	7.11	15.00	0.012
Transportation land-use proportion	–	–	–	7.37	15.00	5.08e-7	–	–	–	15.96	15.00	3.9e-10
Farm and open land-use proportion	–	–	–	10.52	2.50	7.23e-4	–	–	–	1.92	3.00	0.0036
Local road length (A4)	5.97	15.00	1.87e-4	–	–	–	–	–	–	–	–	–
Minimal distance to major freeways (A1)	–	–	–	16.54	–	6.16e-8	–	–	–	–	–	–
Minimal distance to local streets (A4)	14.85	–	7.24e-7	–	–	–	5.12	–	1.21e-4	–	–	–

[a] V.P: the percent of variance explained by the variable (%).

[b] Dis: optimal buffering distance (unit: km).

Table 2

Comparison of predictive errors in five models by leave-one-out cross validation.

Models	**NO_2**						**NO_x**					
	Summer			**Winter**			**Summer**			**Winter**		
	CV R^2[a]	***M***[b]	**RMSPE**[c]	**CV R^2**[a]	***M***[b]	**RMSPE**[c]	**CV R^2**[a]	***M***[b]	**RMSPE**[c]	**CV R^2**[a]	***M***[b]	**RMSPE**[c]
Multiple linear LUR (without *x* and *y* coordinates)	0.56	0.54	5.17	0.55	−0.98	13.45	0.42	0.97	4.45	0.55	−2.17	14.83
Multiple linear LUR (with *x* and *y* coordinates)	0.64	−0.27	4.71	0.60	0.54	3.69	0.63	−0.98	12.45	0.63	2.17	13.46
Universal kriging	0.75	0.16	3.94	0.68	0.082	3.31	0.70	0.96	11.63	0.72	2.04	12.12
GAM without spatial spline term	0.68	0.14	4.42	0.58	−0.42	3.79	0.66	−0.76	12.43	0.64	−0.37	13.38
GAM with spatial spline term	0.78	0.26	3.65	0.79	−0.06	2.68	0.73	−0.39	10.97	0.75	−0.51	10.91
GAM plus cokriging	0.91	−0.02	2.40	0.92	0.05	1.67	0.92	0.09	5.92	0.88	−0.31	7.83

[a] CV R^2: cross validation R^2.

[b] *M*: Median (IQR) of prediction error in ppb.

[c] RMSPE: the squared root of the mean of the squared prediction errors in ppb.