Atmospheric Environment 45 (2011) 4412–4420

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NO_x) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air)

Laina D. Mercer [a], Adam A. Szpiro [a,*], Lianne Sheppard [a,b], Johan Lindström [c,d], Sara D. Adar [e], Ryan W. Allen [f], Edward L. Avol [g], Assaf P. Oron [b], Timothy Larson [h], L.-J. Sally Liu [b,1], Joel D. Kaufman [b]

[a] Department of Biostatistics, University of Washington, WA, USA
[b] Department of Environmental and Occupational Health Sciences, University of Washington, WA, USA
[c] Centre for Mathematical Sciences, Lund University, Sweden
[d] Department of Statistics, University of Washington, WA, USA
[e] Department of Epidemiology, University of Michigan, MI, USA
[f] Simon Fraser University, Faculty of Health Sciences, Canada
[g] Department of Preventive Medicine, University of Southern California, CA, USA
[h] Department of Civil and Environmental Engineering, University of Washington, WA, USA

ARTICLE INFO

Article history:
Received 30 December 2010
Received in revised form
14 May 2011
Accepted 16 May 2011

Keywords:
Universal kriging
Land-use regression
Spatial modeling
Air pollution
Exposure assessment
Los Angeles

ABSTRACT

Background: Epidemiological studies that assess the health effects of long-term exposure to ambient air pollution are used to inform public policy. These studies rely on exposure models that use data collected from pollution monitoring sites to predict exposures at subject locations. Land-use regression (LUR) and universal kriging (UK) have been suggested as potential prediction methods. We evaluate these approaches on a dataset including measurements from three seasons in Los Angeles, CA.

Methods: The measurements of gaseous oxides of nitrogen (NO_x) used in this study are from a "snapshot" sampling campaign that is part of the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). The measurements in Los Angeles were collected during three two-week periods in the summer, autumn, and winter, each with about 150 sites. The design included clusters of monitors on either side of busy roads to capture near-field gradients of traffic-related pollution.

LUR and UK prediction models were created using geographic information system (GIS)-based covariates. Selection of covariates was based on 10-fold cross-validated (CV) R^2 and root mean square error (RMSE). Since UK requires specialized software, a computationally simpler two-step procedure was also employed to approximate fitting the UK model using readily available regression and GIS software.

Results: UK models consistently performed as well as or better than the analogous LUR models. The best CV R^2 values for season-specific UK models predicting $\log(NO_x)$ were 0.75, 0.72, and 0.74 (CV RMSE 0.20, 0.17, and 0.15) for summer, autumn, and winter, respectively. The best CV R^2 values for season-specific LUR models predicting $\log(NO_x)$ were 0.74, 0.60, and 0.67 (CV RMSE 0.20, 0.20, and 0.17). The two-stage approximation to UK also performed better than LUR and nearly as well as the full UK model with CV R^2 values 0.75, 0.70, and 0.70 (CV RMSE 0.20, 0.17, and 0.17) for summer, autumn, and winter, respectively.

Conclusion: High quality LUR and UK prediction models for NO_x in Los Angeles were developed for the three seasons based on data collected for MESA Air. In our study, UK consistently outperformed LUR. Similarly, the 2-step approach was more effective than the LUR models, with performance equal to or slightly worse than UK.

1. Introduction

A central challenge in epidemiological analyses of the long-term impact of air pollution exposure on health is assigning exposures to individual subjects in cohort studies. Seminal studies in the field

Abbreviations: UK, universal kriging; LUR, land-use regression; CV, cross validated; RMSE, root mean square error.

* Corresponding author. Department of Biostatistics, University of Washington, F-600, Health Sciences Building, Seattle, WA 98195-7232, USA. Tel.: +1 206 616 6846.
E-mail address: aszpiro@u.washington.edu (A.A. Szpiro).
[1] Present address: Swiss Tropical and Public Health Institute, Switzerland.

1352-2310/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2011.05.043

assigned exposures using area-wide monitored concentrations in different geographic regions (Dockery et al., 1993; Pope et al., 2002), but this approach ignores small-scale variation between individuals living within the same geographic region. More recent studies have incorporated intra-urban variation in ambient concentrations, with the exposures estimated by a variety of methods ranging from assigning the value measured at the nearest monitor to the subject's home, to using distance to traffic as a proxy for traffic-related pollution, to predicting concentrations from complex spatial or spatio-temporal statistical models (Miller et al., 2007; Basu et al., 2000; Ritz et al., 2006; Puett et al., 2009; Adar et al., 2010).

Exposure studies have shown that levels of road traffic pollutants vary substantially within cities, often on the scale of meters (Briggs et al., 2000; Zhu et al., 2002; Gilbert et al., 2003). As such, studies have modeled intra-urban variation using multivariate regression based on Geographic Information System (GIS) covariates, often referred to as "land-use" regression (LUR) (Briggs et al., 1997; Hoek et al., 2008; Brauer et al., 2003; Jerrett et al., 2005a; Su et al., 2009), and the geostatistical method kriging (Jerrett et al., 2005b; Künzli et al., 2005; Beelen et al., 2009). Little has been done, however, to explore the degree to which the predictive power of these models can be improved by combining predictions from LUR with smoothing of correlated residuals through universal kriging (UK).

The current study is motivated by the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air), a prospective cohort study funded by the Environmental Protection Agency (EPA) to assess the associations between chronic exposure to air pollution and changes in sub-clinical markers of cardiovascular disease. Exposure assessment is based on existing regulatory monitoring networks as well as supplemental monitoring campaigns that include data collected at irregularly spaced times and locations (Cohen et al., 2009). Here we focus on spatial analysis of one of these supplementary monitoring campaigns, the "snapshot" campaign, which was designed to characterize spatial variability and roadway gradients using samples collected simultaneously at multiple locations across each city during three seasons. Ultimately, exposures for MESA Air will be predicted using a spatio-temporal model that incorporates all of the available monitoring data (Szpiro et al., 2010; Lindström et al., 2011; Sampson et al., 2011). The spatial analysis we present here provides insights into variable selection for the spatio-temporal model.

This paper focuses on evaluating spatial methods for predicting concentrations of gaseous oxides of nitrogen (NO_x) from the MESA Air snapshot campaign in Los Angeles, CA. We compare the performance of LUR with UK, which can be viewed as a generalization of LUR that also incorporates kriging in order to exploit spatial correlation. We also consider a computationally simpler 2-step approximation to UK.

2. Methods

2.1. Study design and data

The Ogawa method was used to measure NO_x concentrations (*NO*, NO_2, NO_x, and SO_2 Sampling Protocol Using the Ogawa Sampler; Ogawa and Company, U.S.A., Inc.: Pompano Beach, FL, 1998). Concurrently deployed duplicate samples were used to assess precision. The mean relative percent difference for these samples in MESA Air was 5.2% with an RMSE of 6.1 ppb.

The snapshot campaign in downtown and coastal Los Angeles consists of 449 2-week average concentration measurements of NO_x. These samples were collected during three 2-week periods in June 2006, October 2006, and January 2007. The sampling periods were chosen to capture seasonal differences and the sampling locations were chosen to maximize the variability of measured concentrations, characterize different land-use categories, and cover the geographic region as broadly as possible. This campaign is described in greater detail by Cohen et al. (2009). In each round of monitoring, the majority of monitors were arranged 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. An example of an idealized roadway gradient cluster is shown in Fig. 1.

Over 300 GIS-based covariates were calculated for MESA Air. As described below, some were combined into composite covariates, and others were excluded because they did not have sufficient variability in the dataset. The approximately 65 covariates considered for model development are summarized in Table 1. All of the land-use covariates were derived using the ArcGIS (Version 9.3, ESRI, Redlands, CA) software package based on land-use data from the US Geological Survey (Price et al., 2006) roadway information from Tele Atlas Dynamap 2000 (TeleAtlas®, Lebanon, NH), and census data from the US Department of Commerce (2001). Census feature class codes (CFCC) were used to categorize roadways. The largest roads are designated as A1 (primary highways with limited access) while other major roads are designated as A2 (primary roads without limited access) or A3 (secondary and connecting roads). We derived an intense land-use covariate that combines industrial, residential, transportation, commercial and services land-use and an open space covariate that is the sum of bays, croplands and pastures, wetlands, forests, groves, rangeland, and sandy beach areas. These combined covariates were created for parsimony and because many of these land-use and open space covariates did not have sufficient variability to be included in our models alone. The distance to coast variable is based on the Tele Atlas Dynamap 2000 County Boundary defined border of the Pacific Ocean.

2.2. Analysis

2.2.1. Land-use regression and universal kriging

In this section we give a brief description of LUR and UK as well as an ad-hoc 2-step approximation to UK that can be easily fit using widely available software. We separately fit LUR and UK models to the log-transformed NO_x concentration data from each of the three seasons. Log-transformed NO_x concentrations approximated a Normal distribution. Additional details of the LUR, UK, and the 2-step approach can be found in the Appendix.

LUR models are multiple linear regressions that assume independent residuals and use GIS-based covariates to predict concentrations at locations without measurements. UK can also be seen as a regression with geographic covariates but with the addition of correlated residuals. In other words, UK allows

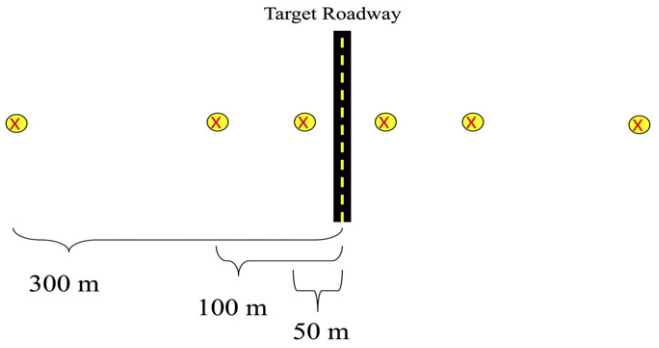

Fig. 1. Idealized road gradient sampling scheme.

Table 1
GIS-based variables considered for model development. A1 roads are primary highways with limited access, A2 roads are primary roads without limited access, and A3 roads are secondary and connecting roads.

Predictor variable	Symbol	Units	Buffer radii	Functional form
Non-roadway				
Population	Pop	Total people within buffer (m)	500, 1000, 1500, 2000, 2500, 3000, 5000, 10,000, 15,000	Scaled by 1/100,000
Intense use land	Int	km^2	50, 100, 150, 300, 500, 750, 1000, 1500, 2000, 3000	Scaled by 1/100
Open space land	Open	km^2	50, 100, 150, 300, 500, 750, 1000, 1500, 2000, 3000	Untransformed
Distance to coast	D2C	Meters	n/a	Trunc. 15 km & 25 km scaled by 1/100,000
Distance to industrial source (e.g. D2Comm: dist. to commercial, D2Lport: dist. to large port, D2Ryard: dist. to rail yard)	D2(variable)	Meters	n/a	
Roadway				
Distance to nearest A1, A2, or A3	D2R	Meters	n/a	Log10[a]
Distance to nearest A1	D2A1	Meters	n/a	Log10[a]
Distance to nearest A2	D2A2	Meters	n/a	Log10[a]
Distance to nearest A3	D2A3	Meters	n/a	Log10[a]
Length of A1 roads within buffer	A1	Meters	50, 100, 150, 300, 400, 500, 750, 1000, 5000, 10,000, 15,000	Scaled by 1/10,000
Length of A2 and A3 roads within buffer	A23	Meters	50, 100, 150, 300, 400, 500, 750, 1000, 5000, 10,000, 15,000	Scaled by 1/100,000

[a] Minimum distance of 10 m.

information about nearby concentration measurements to influence the predictions through the estimated correlation structure.

Kriging is a minimum mean-squared error approach to spatial prediction (Cressie, 1993; Banerjee et al., 2004). Unlike ordinary and simple kriging that assume an underlying constant mean surface, UK can be easily integrated with a regression model. The mean and covariance parameters for UK are estimated together using maximum likelihood through the geoR package in R (Ribeiro and Diggle, 2001). We assumed an exponential correlation structure.

ArcGIS is a widely available tool, but it does not allow UK with an arbitrary set of covariates or using maximum likelihood fitting. The 'universal' kriging option within ArcGIS models the spatial correlation as expected and allows the mean to be defined as a function of the latitude and longitude, but does not allow the use of land-use covariates to be included in the mean model. The 'simple' and 'ordinary' krigings methods in ArcGIS model the spatial correlation and fit a constant mean. With these limitations in mind we also evaluated an ad-hoc 2-step approach to prediction, first estimating the LUR model assuming independent errors and then using simple kriging in ArcGIS to estimate the dependence in the LUR residuals and exploit this dependence to improve the predictions by smoothing. This approach, which is similar but not identical to UK, has some advantages: it allows for spatial structure in the residuals, whereas LUR does not, and it can be implemented with ArcGIS, whereas UK with land-use covariates cannot. The 2-step procedure performs well in our examples, but we note that it may not be optimal because, unlike in UK, the regression coefficients are estimated without accounting for spatial correlation.

Table 2
a: Summary of Los Angeles NO_x snapshot monitoring locations. b: Summary of Los Angeles NO_x snapshot concentrations (ppb). c: Summary of Los Angeles log(NO_x) snapshot concentrations (log(ppb)).

		Summer	Autumn	Winter
a				
Dates	Start	6/27/2006	10/17/2006	1/23/2007
	End	7/12/2006	11/3/2006	2/7/2007
Samples	Total	148	152	149
	Clusters	24	24	23
	Individuals	14	13	20
b				
All sites	Mean	34.2	75.1	95.3
	Variance	132.1	551.0	728.4
Clusters	Mean	34.3	74.7	96.0
	Between variance	89.6	450.5	620.4
	Average w/in variance	65.9	152.5	170.0
Individual sites	Mean	31.2	77.7	91.0
	Between variance	144.8	359.2	571.8
c				
All sites	Mean	3.47	4.27	4.51
	Variance	0.15	0.10	0.09
Clusters	Mean	3.47	4.26	4.52
	Between variance	0.14	0.09	0.08
	Average w/in variance	0.04	0.02	0.02
Individual sites	Mean	3.35	4.32	4.48
	Between variance	0.24	0.06	0.07

2.2.2. Cross-validation

The analysis in this paper employs a 10-fold cross-validation approach (Hastie et al., 2001). The exposure data are split into ten roughly equal size groups. To account for our unique study design, the clusters of six monitors were always kept within the same group. Then one group of data is set aside (test set) and the model is fit on the remaining nine groups (training set). The model estimated using the training set is used to predict values at the monitor locations in the test set. This is repeated until predictions for all groups have been generated. To evaluate the predictive ability of the model, R^2 and RMSE are calculated based on comparing these predicted values to the observations. We refer to these as the cross-validated (CV) R^2 and RMSE.

2.2.3. Model selection

Variable selection was done separately for each season and modeling approach (LUR or UK), all starting with the same set of covariates. We excluded covariates without at least 5% nonzero values in our data. Then the least absolute shrinkage and selection operator (lasso) was implemented as a prescreening tool via the glmnet function found in the R package glmnet (Friedman et al., 2010). The tuning parameter λ was selected to yield 15–20 nonzero regression coefficients from the lasso. We followed the lasso algorithm with a modified exhaustive search of the possible covariate combinations. The exhaustive search approach ensures that the model selection process accounts for the spatial correlation in the data, while the lasso algorithm does not account for this correlation. The final models were selected based on predictive ability as measured by cross-validated R^2 and RMSE.

In our exhaustive search, we considered no more than one buffer size for each covariate in any given model. In a recent paper, Su et al. (2009) reported on a methodology for selecting both a small- and long-range buffers for some geographic covariates. We restricted our models to a single (typically small) buffer size in the interest of parsimony and because coefficients for long-range buffer covariates are difficult to interpret in a model that predicts local concentrations.

In addition to the separate models for each season, a "common" model for all seasons was chosen via an exhaustive search of covariates that performed well in each of the seasons. When multiple models had the same cross-validated R^2 and RMSE, the most parsimonious model was chosen.

3. Results

A total of 148, 152, and 149 monitoring locations were included in the analysis for summer, autumn, and winter, respectively. Sampling dates, cluster, and individual site allocations are summarized in Table 2a. Fig. 2a–c display the locations of the monitors in each season and their measured concentrations. Clustered data locations have been spread on the map for display purposes. Notably, there is evidence of a large-scale spatial gradient, with some of the lowest concentrations observed near the coast and higher values inland. We also see evidence of the roadway gradients, particularly inland where the highest concentrations in a given cluster are found at the monitors closest to the roads. In Table 2b we see that the magnitude of variability differs by season and the majority of the variability in NO_x concentrations is between the individual sites and clusters as opposed to within clusters. Road clusters have similar means to individual sites. Except in summer, road cluster means are more variable than individual sites.

Results for models using variables selected based on LUR are shown in Table 3a. Using these variables, the LUR models explained 74%, 60%, and 67% of the variability with RMSE of 0.20, 0.20, and 0.17 in the summer, autumn, and winter, respectively. In all seasons UK and the 2-step approach had a cross-validated R^2 greater than those of the LUR model (using variables selected for LUR). The distance to nearest A1 roadway and distance to coast covariates were present in the best models for all three seasons.

Results for models using variables selected based on UK are shown in Table 3b. The UK predictions explained 75%, 72%, and 74% of the variability with RMSE of 0.20, 0.17, and 0.15 in the summer, autumn, and winter, respectively. As with the variables selected based on LUR, we see that UK and the 2-step approach consistently

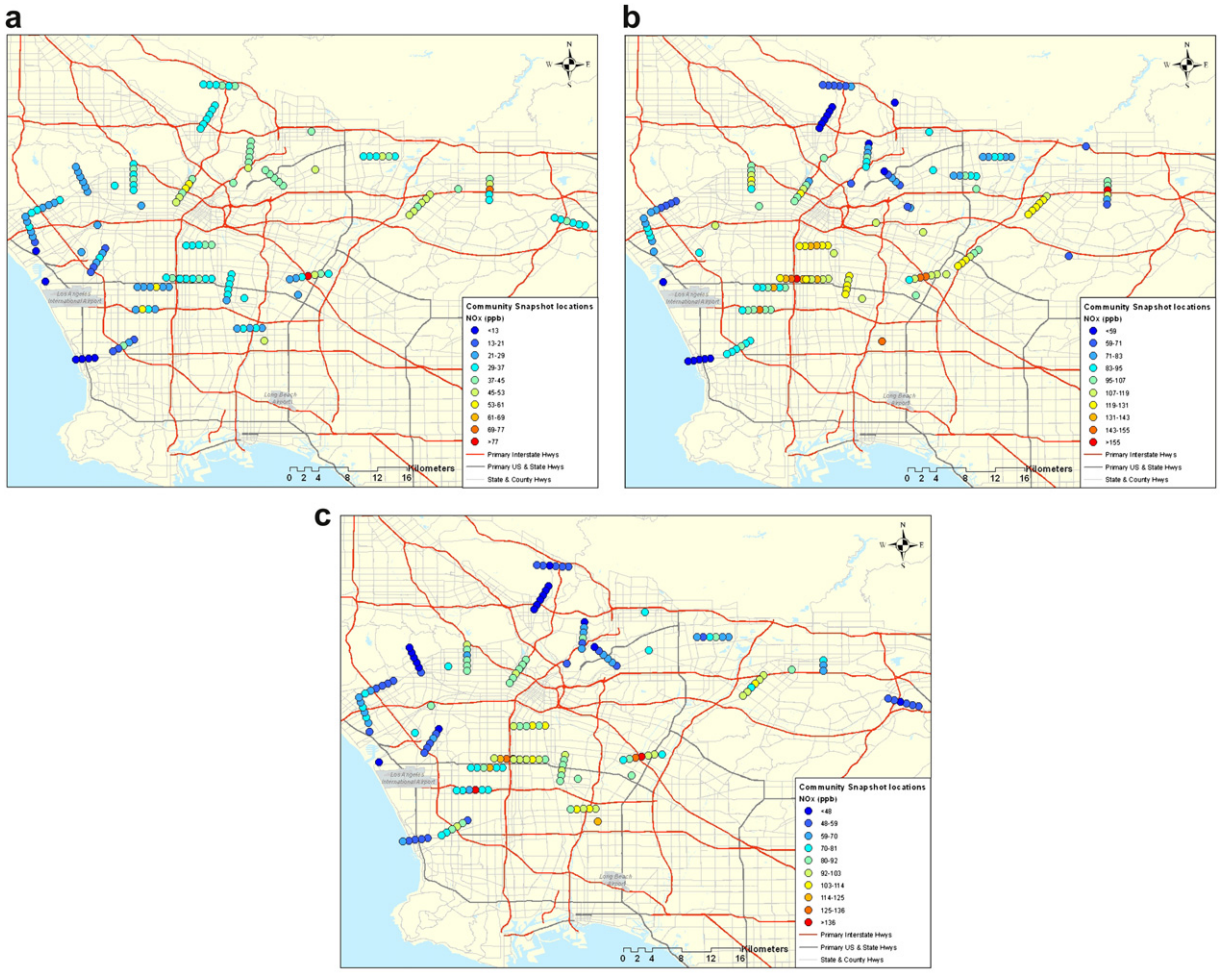

Fig. 2. a: Los Angeles sampling locations and concentrations in Summer 2006. The circles represent monitor locations. The distance between monitors and size of monitors have been exaggerated to make visualization of all locations possible. b: Los Angeles sampling locations and concentrations in Autumn 2006. The circles represent monitor locations. The distance between monitors and size of monitors have been exaggerated to make visualization of all locations possible. c: Los Angeles sampling locations and concentrations in Winter 2007. The circles represent monitor locations. The distance between monitors and size of monitors have been exaggerated to make visualization of all locations possible.

Table 3
a: Cross-validated R^2 (RMSE) calculated for land-use regression, universal kriging, and the 2-step approach for the 'best' model as chosen for predictive ability of land-use regression in each season (on log(ppb) scale) for NO_x. b: Cross-validated R^2 (RMSE) calculated for land-use regression, universal kriging, and the 2-step approach for the 'best' model as chosen for predictive ability of universal kriging in each season (on log(ppb) scale) for NO_x.

Season	Covariates	LUR	UK	2-step
a				
Summer	D2A1, D2C, D2Comm, A23_50 m, A1_50 m, D2Lport, Int_3k	0.74 (0.20)	0.75 (0.19)	0.75 (0.19)
Autumn	D2A1, D2C, D2Comm, A23_50 m, A1_50 m, POP_5k	0.60 (0.20)	0.72 (0.17)	0.70 (0.17)
Winter	D2A1, D2C, A23_500 m, D2R, Int_3k, D2Ryard	0.67 (0.17)	0.67 (0.17)	0.68 (0.17)
b				
Summer	D2A1, D2C, D2Comm, A23_50 m	0.70 (0.21)	0.75 (0.20)	0.75 (0.20)
Autumn	D2A1, D2C, D2Comm, A23_50 m, A1_50 m, POP_5k	0.60 (0.20)	0.72 (0.17)	0.70 (0.17)
Winter	D2A1, D2C, D2Comm, A23_400 m, A1_50 m, Int_3k	0.48 (0.22)	0.74 (0.15)	0.70 (0.17)

outperform LUR in terms of prediction accuracy. Distances to nearest A1 roadway, coast, and commercial or service locations were included in the best model for each season.

The degradation in prediction accuracy for LUR models in winter in Table 3b compared with Table 3a is noteworthy. The difference is likely due to the fact that the winter LUR model selected based on UK fails to account for large-scale spatial structure, while the corresponding model selected based on LUR includes additional covariates (D2R, D2Ryard) that have the potential to capture large-scale spatial structure.

We also identified a model that is common for all seasons. Here common is defined as having the highest average cross-validated R^2. The estimated coefficients and prediction performances for the common model are shown in Table 4. Overall the summer model is the most distinct, in part because large-scale spatial structure is captured in the mean model by the distance to coast covariate, resulting in a smaller estimated range. This may be related to the dominant on-shore wind in the Los Angeles region during the summer season. In the winter and autumn seasons, the wind is more variable so distance to coast is not as predictive of NO_x concentrations, and we find a longer range for the residual spatial correlation.

Fig. 3a–c show scatterplots of the observed and cross-validated predicted NO_x concentrations (including 95% prediction intervals) at all of the monitoring locations in each of the three seasons, based on the common UK model. There are no notable outliers, and the prediction quality and interval coverage appear generally good, consistent with the R^2 values reported above.

The coefficient estimates vary across seasons, although mostly without changing sign. The two exceptions are A1 roads within 50 m and population in a 5 km buffer; both these coefficients have negative signs in summer, although they also have very large standard errors. The common model consistently performed well in all seasons with cross-validated R^2 of 0.70, 0.68, and 0.72 (RMSE 0.21, 0.18, and 0.16) in the summer, autumn, and winter, respectively. As in the case of individual models for each season, UK and the 2-step approach have better predictive ability than LUR.

To further illustrate the advantages of UK, Fig. 4a–c show the variograms for the residuals in the summer, autumn, and winter season, respectively, from ordinary least squares regression models with covariates chosen in the common LUR and UK models. The residuals for both models clearly exhibit spatial structure. The initial drop in semivariance around 100 m is likely due to a combination of wind direction coupled with our clustered design that is keenly sensitive to local traffic effects; Vienneau et al. (2010) observed a similar phenomenon in their LUR models. As an additional test for spatial correlation, we applied a Moran's I test with inverse-distance weighting (Paradis et al., 2004). The Moran's I test was statistically significant with $p < 0.001$ for all six sets of residuals. The LUR approach does not exploit the spatial correlation for predictions, instead assuming independent observations. UK estimates the covariance and incorporates the spatial structure in predictions; the result is improved predictive performance.

Table 4
Estimated parameters and standard errors for common model (on log(ppb) scale) for NO_x. Details about the scaling of the covariates can be found in Table 1.

Parameter	Summer	Autumn	Winter
Regression coefficient estimate (SE)			
Intercept	2.927 (0.25)	4.214 (0.25)	3.684 (0.25)
D2A1	−0.269 (0.04)	−0.161 (0.04)	−0.137 (0.03)
A1_50 m	−4.002 (7.02)	10.972 (5.37)	8.664 (4.26)
A23_400 m	0.403 (0.28)	0.233 (0.22)	0.551 (0.19)
Pop_5k	−0.012 (0.03)	0.112 (0.03)	0.021 (0.03)
D2C	7.424 (0.69)	0.484 (0.92)	1.334 (1.03)
Int_3k	1.754 (0.77)	0.845 (0.76)	3.746 (0.88)
D2Comm	−1.507 (0.97)	−3.846 (0.80)	−2.071 (0.78)
Covariance			
$\hat{\phi}$ (range in meters)	5702	15,432	13,021
$\hat{\sigma}^2$ (partial sill)	0.030	0.016	0.027
$\hat{\tau}^2$ (nugget)	0.011	0.023	0.010
Cross-validated R^2 (RMSE)			
LUR	0.66 (0.22)	0.55 (0.19)	0.50 (0.21)
UK	0.70 (0.21)	0.68 (0.18)	0.72 (0.16)
2-step	0.70 (0.21)	0.66 (0.18)	0.70 (0.17)

4. Discussion

Epidemiological studies of air pollution and health outcomes rely on quality predictions of air pollution concentrations at sites without monitoring data. The modeling approaches taken in this paper have proven to be effective at capturing small-scale variability, as measured by cross-validated R^2, and are reasonably straightforward to implement in readily available software (e.g. ArcGIS, R).

In our study, UK yielded cross-validated R^2 for log-transformed NO_x concentrations ranging from 0.72 to 0.75. The best models selected for LUR were not as good but were still able to achieve reasonable predictive ability (CV R^2 ranged from 0.60 to 0.74). Previous LUR studies of oxides of nitrogen (NO, NO_2, or NO_x) have yielded cross-validated R^2 values ranging from 0.49 to 0.87 (Hoek et al., 2008).

A recent paper that analyzed NO_x concentrations in Los Angeles reported an R^2 value of 0.92 based on LUR, but this number is estimated using a single test set of only 16 locations so it is difficult to assess its generalizability (Su et al., 2009). Other considerations such as the distribution of monitor locations and variability in measured concentration also impact R^2 and may explain the difference between our results and those reported by Su et al. (2009).

Previous studies have shown LUR performing better than ordinary kriging, which relies solely on spatial smoothing without incorporating covariates. Few studies have compared LUR and UK

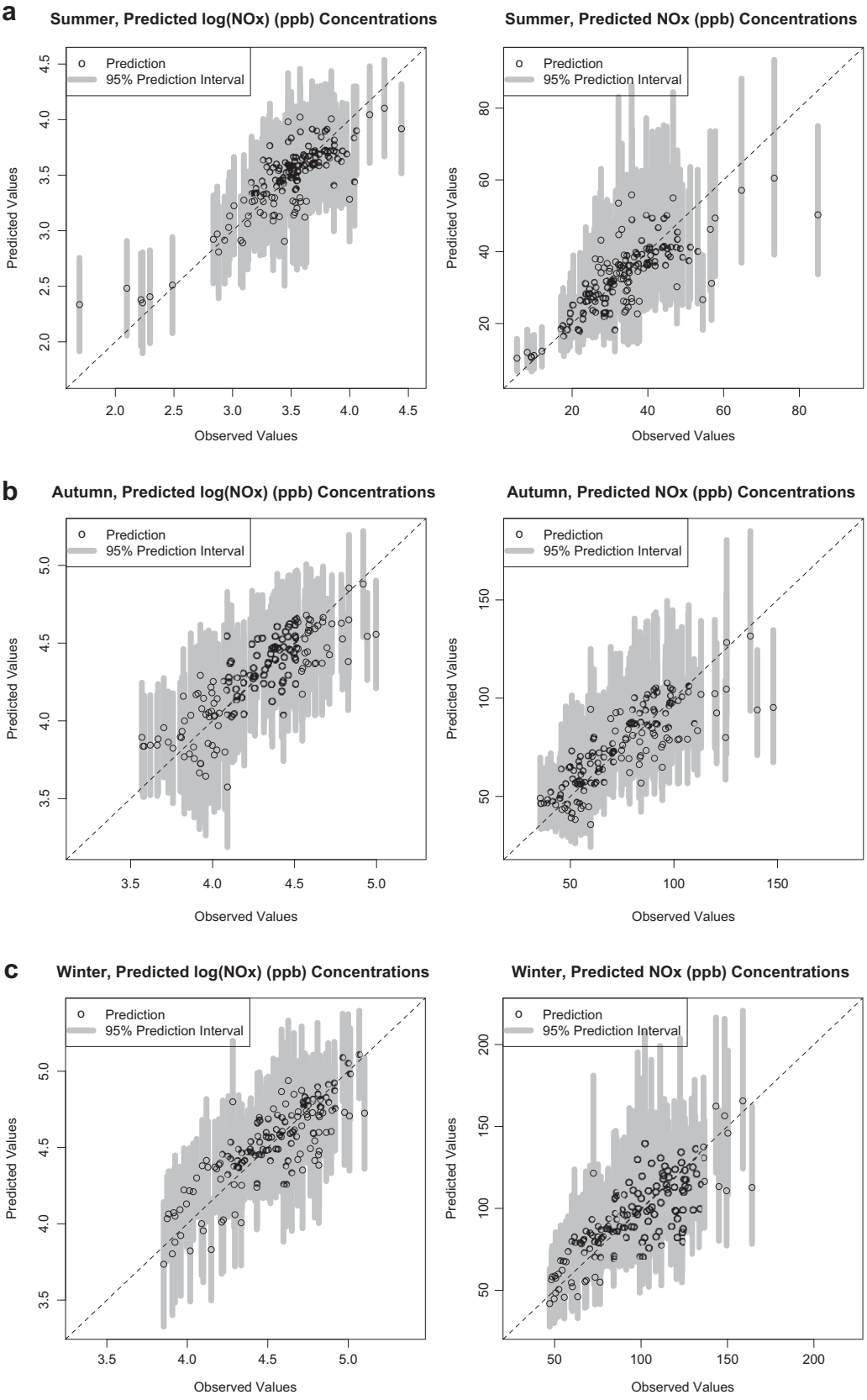

Fig. 3. a: Universal kriging predictions and 95% prediction intervals and observed concentrations in Summer 2006. b: Universal kriging predictions and 95% prediction intervals and observed concentrations in Autumn 2006. c: Universal kriging predictions and 95% prediction intervals and observed concentrations in Winter 2007.

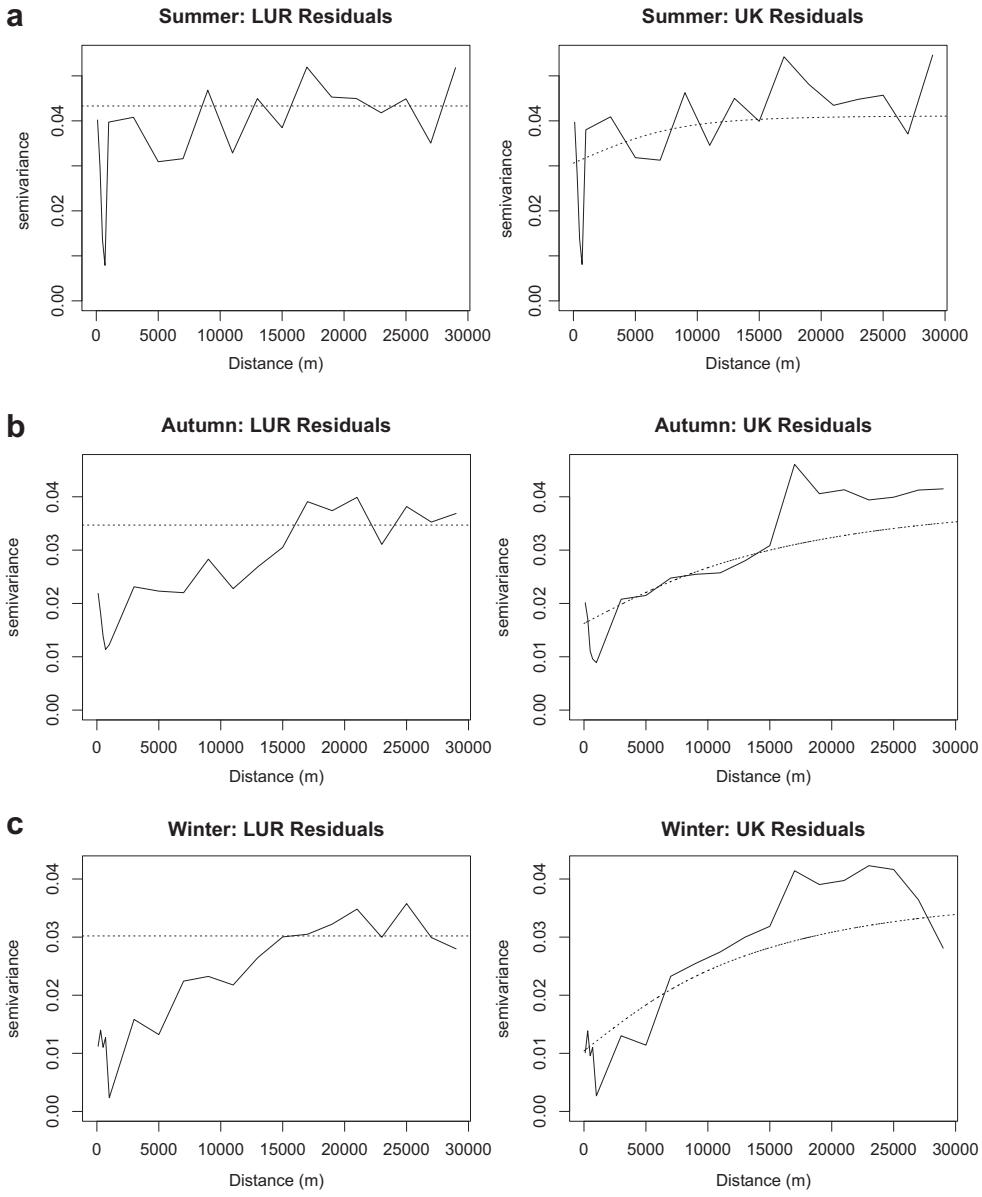

Fig. 4. a: Display of the variograms for land-use regression and universal kriging common model residuals for the summer season. The dotted black line is the covariance function estimated by the model for each approach. b: Display of the variograms for land-use regression and universal kriging common model residuals for the autumn season. The dotted black line is the covariance function estimated by the model for each approach. c: Display of the variograms for land-use regression and universal kriging common model residuals for the winter season. The dotted black line is the covariance function estimated by the model for each approach.

or entertained a 2-step approach along the lines of the one we describe. One exception is a recent paper that demonstrated the superiority of UK over other modeling approaches in mapping background air pollution across the European Union (Beelen et al., 2009). In our study, UK consistently outperformed LUR. Similarly, the 2-step approach was more effective than the LUR models, with performance equal to or slightly worse than UK. Even for the models selected for their LUR performance, the UK and 2-step approaches performed as well or better in all seasons.

Since air pollution emission and dispersion is a deterministic process, one might argue that the improved performance via correlated residuals masks the absence of important variables from our models. However, the inability to collect or precisely quantify all variables affecting pollution should be viewed as an inherent aspect of exposure assessment. Even with MESA Air's highly detailed set of GIS variables, there remained substantial unaccounted-for spatial correlations. Hence, incorporating these correlations in the model is the preferred approach both practically and conceptually.

Our results indicate that the 2-step approach has predictive ability similar to UK. This is encouraging since it provides an additional option for researchers less familiar with using flexible tools to fit UK models (such as geoR). The 2-step approach is based on first fitting a mean model using standard linear regression and then modeling the spatial correlation between residuals using tools available in ArcGIS. This approach is preferable to choosing between ordinary kriging (assuming a constant mean) and land-use regression (assuming independence).

The important model selection stage still needs to be performed separately. At present, there is no single optimal solution for model selection. However, it is important to choose an approach that works well for the research problem at hand, and appropriately

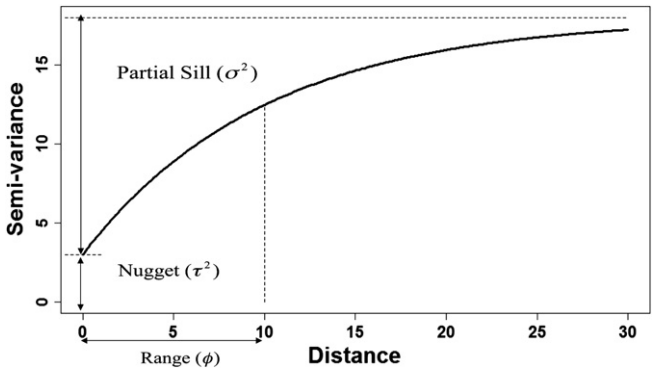

Fig. 5. Theoretical exponential variogram.

handles highly correlated covariates (e.g. traffic density or population in different size buffers).

Validation is an essential part of assessing the performance and robustness of a model, and its design should be considered carefully. The results in this paper use 10-fold cross-validated R^2 (Hastie et al., 2001), with specific consideration given to the clustered design of our monitoring data. Since we used cross-validation as part of the model selection procedure, our reported prediction accuracy may still be somewhat optimistic because the same data were used for selection and final evaluation. As an alternative to cross-validation, some authors (e.g. Su et al., 2009) consider a test dataset that is separate from the training data. This has the virtue of providing independent test data, but if the test dataset is relatively small (which it often must be in order to have sufficient training data available to fit the model) the resultant R^2 value can be less informative than the cross-validated R^2 about sensitivity of the model to possible extreme observations. Moreover, as is the case with cross-validation, typically the test set's locations are similar in nature to the training set, and not to the eventual target locations thus the estimated prediction accuracy may be biased.

There was considerable variability between seasons in the magnitude, and in some cases also the signs, for the common model coefficient estimates and spatial covariance model parameters shown in Table 4. In particular, summer has the shortest range, indicating the spatial correlation is the smallest in this season, while autumn and winter show evidence of longer range correlation. One possible explanation for this discrepancy is that the distance to coast covariate captures the effect of prevailing winds in summer and accounts for the dominant spatial structure in the data, whereas this simple covariate does not account for the more complex circulation patterns in autumn and winter. The variability between seasons suggests that it would be unwise to use our dataset to naively model all three seasons jointly or to combine them in a single model to predict long-term average concentrations. Instead, we incorporate our findings from analyzing the snapshot data into the MESA Air spatio-temporal model (Szpiro et al., 2010; Lindström et al., 2011) to reliably predict NO_x concentrations at subject locations over several years and derive long-term average exposures for the time periods of interest based on these predictions.

Acknowledgments

Although the research described in this presentation has been funded in part by the United States Environmental Protection Agency through grant RD831697 and assistance agreement CR-834077101-0 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. Additional funding was provided by grants P50 ES015915 and T32 ES015459 from the National Institute of Environmental Health Sciences. Travel for Johan Lindström has been paid by STINT (The Swedish Foundation for International Cooperation in Research and Higher Education) Grant IG2005-2047. The authors thank the University of Southern California field team, Suresh Ratnam, Maria Barney, Effie Wu, and Jane Quan, who deployed the Ogawa samplers.

Appendix

Land-use regression model

Take our monitoring data to be $\vec{Y} = (Y(s_1), \ldots, Y(s_n))^T$, where $Y(s_i)$ is the observed concentration at location s_i. We also have a set of k land-use covariates to be used as predictors. We model the data as:

$$Y(s_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N\left(0, \nu^2\right) \quad \text{and} \quad i = 1, \ldots n$$

equivalently,

$$\vec{Y}(s) = \mathbf{X}(s)\vec{\beta} + \vec{\epsilon} \quad \text{where} \quad \vec{\epsilon} \sim N\left(0, \nu^2\right)$$

where $\mathbf{X}(s)$ are the GIS-based covariates from sampled sites, respectively. Let s_0 indicate locations without concentration data (e.g. subject homes) and $Y(s_0)$ be the corresponding unobserved concentration at location s_0. The estimated regression parameters $\vec{\beta}$ are used to predict NO_x concentrations at unmeasured location with:

$$E[\vec{Y}(s_0) | \vec{Y}(s);\ \mathbf{X}(s_0)] = \mathbf{X}(s_0)\vec{\beta}$$

Universal kriging

Regression and covariance parameters are estimated by fitting the following model with the geoR package in R (Ribeiro and Diggle, 2001):

$$\vec{Y}(s) = \mathbf{X}(s)\vec{\beta} + \vec{\epsilon} \quad \text{with} \quad \vec{\epsilon} \sim N\left(0, \Sigma\left(\sigma^2, \tau^2, \phi\right)\right)$$

A covariance function must be specified. We use the exponential covariance model, which models the covariance of two measurements as a function of the distance between them such that

$$C(d) = \begin{cases} \sigma^2 \exp(-d/\phi) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}$$

where d is the distance between observations. A theoretical variogram for the exponential correlation function is shown in Fig. 5.

Assuming $\vec{Y}(s)$ and $\vec{Y}(s_0)$ are jointly normal, from standard multivariate normal theory we have the result:

$$\begin{pmatrix} \vec{Y}(s_0) \\ \vec{Y}(s) \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{X}(s_0)\vec{\beta} \\ \mathbf{X}(s)\vec{\beta} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

where $\Sigma_{12} = \Sigma_{21}^T$. Predictions for unmeasured locations s_0 are generated using the conditional distribution of unobserved data given observed data at locations s:

$$E[\vec{Y}(s_0)|\vec{Y}(s)] = \mathbf{X}(s_0)\vec{\beta} + \Sigma_{12}\Sigma_{22}^{-1}\left(\vec{Y}(s) - \mathbf{X}(s)\vec{\beta}\right)$$

2-step predictions with LUR & simple kriging

As in LUR the multivariate regression model $\vec{Y}(s) = \mathbf{X}(s)\vec{\beta} + \vec{\epsilon}$ is fit, where $\vec{\epsilon} \sim N(0, \nu^2)$. Simple kriging is used with the regression residuals, $\vec{\epsilon}$, to estimate $\Sigma(\sigma^2, \tau^2, \varphi)$. Then, as in universal kriging, predictions are based on the normal conditional expectation of $\vec{Y}(s_0)$ given $\vec{Y}(s)$. Operationally, this involves adding the land-use regression predictors with the simple kriging prediction.

References

Adar, S.D., Klein, R., Klein, B.E.K., Szpiro, A.A., Cotch, M.F., Wong, T.Y., O'Neill, M.S., Shrager, S., Barr, R.G., Siscovick, D., Daviglus, M.L., Sampson, P.D., Kaufman, J.D., 2010. Air pollution and the human microvasculature in vivo assessed via retinal imaging: the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). PLOS Medicine 7 (11).

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall.

Basu, R., Woodruff, T.J., Parker, J.D., Saulnier, L., Schoendorf, K.C., 2000. Particulate air pollution and mortality: findings from 20 U.S. cities. New England Journal of Medicine 343 (24), 1742–1749.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., Briggs, D.J., 2009. Mapping of background air pollution at a fine spatial scale across the European union. Science of the Total Environment 407 (6), 1852–1867.

Brauer, M., Hoek, G., van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys, J., Bellander, T., Lewne, M., Brunekreef, B., 2003. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. Epidemiology 14 (2), 228–239.

Briggs, D.J., Collins, S., Elliot, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Van Reeuwijk, H., Smallbone, K., Van der Veen, A., 1997. Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science 11, 699–718.

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

Cohen, M.A., Adar, S.D., Allen, R.W., Avol, E., Curl, C.L., Gould, T., Hardie, D., Ho, A., Kinney, P., Larson, T.V., Sampson, P.D., Sheppard, L., Stukovsky, K.D., Swan, S.S., Liu, L.J., Kaufman, J.D., July 2009. Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Environmental Science and Technology 43 (13), 4687–4693.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics. Wiley.

Dockery, D.W., Pope, C.A., Xu, X., Spangler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., Speizer, F.E., 1993. An association between air pollution and mortality in six cities. New England Journal of Medicine 329 (24), 1753–1759.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1–22.

Gilbert, N.L., Woodhouse, S., Stieb, D.M., Brook, J.R., 2003. Ambient nitrogen dioxide and distance from a major highway. Science of the Total Environment 312 (1–3), 43–46.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer.

Hoek, G., Beelen, R., de Hoogh, K., Gulliver, J., Fischer, P., Briggs, D.J., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 42, 7561–7578.

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

Jerrett, M., Buzzelli, M., Burnett, R.T., DeLuca, P.F., 2005b. Particulate air pollution, social confounders, and mortality in small areas of an industrial city. Social Science and Medicine 60 (12), 2845–2863.

Künzli, N., Jerrett, M., Mack, W.J., Beckerman, B., LaBree, L., Gilliland, F., Thomas, D., Peters, J., Hodis, H.N., 2005. Ambient air pollution and atherosclerosis in Los Angeles. Environmental Health Perspectives 113 (2), 201–206.

Lindström, J., Szpiro, A., Sampson, P., Sheppard, L., Oron, A., Richards, M., Larson, T., 2011. A Flexible Spatio-temporal Model for Air Pollution: Allowing for Spatio-temporal Covariates. UW Biostatistics. Working Paper Series (Working Paper 370).

Miller, K.A., Siscovick, D.S., Sheppard, L., Shepherd, K., Sullivan, J.H., Anderson, G.L., Kaufman, J.D., 2007. Long-term exposure to air pollution and incidence of cardiovascular events in women. New England Journal of Medicine 356, 447–458.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Ito, K., Krewski, D., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Journal of the American Medical Association 9 (287), 1132–1141.

Price, C., Nakagaki, N., Hitt, K.J., Clawges, R.C., 2006. Enhanced Historical Land-use and Land-cover Data Sets of the U.S. Geological Survey. Tech. Rep., U.S. Geological Survey. Digital Data Series 240. URL: http://pubs.usgs.gov/ds/2006/240.

Puett, R.C., Hart, J.E., Yanosky, J.D., Paciorek, C., Schwartz, J., Suh, H., Speizer, F.E., Laden, F., 2009. Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. Environmental Health Perspectives 117 (11), 1697–1701.

Ribeiro, P.J., Diggle, P.J., 2001. geoR: a package for geostatistical analysis. R-NEWS 1 (2), 14–18.

Ritz, B., Wilhelm, M., Zhao, Y., 2006. Air pollution and infant death in southern California, 1989–2000. Pediatrics 118 (2), 493–502.

Sampson, P.D., Szpiro, A.A., Sheppard, L., Lindström, J., Kaufman, J.D., 2011. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. Atmospheric Environment. doi:10.1016/j.atmosenv.2011.04.073.

Su, J.G., Jerrett, M., Beckerman, B., Wilhelm, M., Ghosh, J.K., Ritz, B., 2009. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. Environmental Research 109 (6), 657–670.

Szpiro, A.A., Sampson, P.D., Sheppard, L., Lumley, T., Adar, S.D., Kaufman, J.D., 2010. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21, 606–631.

US Department of Commerce, W. D., 2001. US Census Bureau, Tiger/line Files Redistricting Census 2000 URL: http://www.census.gov/geo/www/tiger.

Vienneau, D., de Hoogh, K., Beelen, R., Fischer, P., Hoek, G., Briggs, D., 2010. Comparison of land-use regression models between Great Britain and the Netherlands. Atmospheric Environment 44 (5), 688–696.

Zhu, Y.F., Hinds, W.C., Kim, S., Shen, S., Sioutas, C., 2002. Study of ultrafine particles near a major highway with heavy-duty diesel traffic. Atmospheric Environment 36 (27), 4323–4335.