Atmospheric Environment 122 (2015) 477–483

Contents lists available at ScienceDirect

Atmospheric Environment

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

Modeling particulate matter concentrations measured through mobile monitoring in a deletion/substitution/addition approach

Jason G. Su [a, *], Philip K. Hopke [b], Yilin Tian [b], Nichole Baldwin [b], Sally W. Thurston [c], Kristin Evans [d], David Q. Rich [d]

[a] Environmental Health Sciences, School of Public Health, University of California at Berkeley, Berkeley, CA 94720-7360, USA
[b] Institute for a Sustainable Environment, and Center for Air Resources Engineering and Science, Clarkson University, Potsdam, NY 13699-5708, USA
[c] Biostatistics and Computational Biology, School of Medicine and Dentistry, University of Rochester, Rochester, NY 14642, USA
[d] Public Health Sciences, School of Medicine and Dentistry, University of Rochester, Rochester, NY 14642, USA

HIGHLIGHTS

- Topic models aggregating mobile data to pre-designed locations for data reduction.
- Land use regression modeling on mobile measurements through D/S/A algorithm.
- Minimizing out-of-sample prediction error through V-fold cross-validation modeling.
- Elevation contributing greatest to daily pollutant variations.
- Highway and major roadways likely contributing to increase in Delta-C.

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form
29 September 2015
Accepted 2 October 2015
Available online 8 October 2015

Keywords:
Land use regression
Woodsmoke
Aethalometer
D/S/A
Mobile air pollution monitoring

ABSTRACT

Land use regression modeling (LUR) through local scale circular modeling domains has been used to predict traffic-related air pollution such as nitrogen oxides (NO_X). LUR modeling for fine particulate matters (PM), which generally have smaller spatial gradients than NO_X, has been typically applied for studies involving multiple study regions. To increase the spatial coverage for fine PM and key constituent concentrations, we designed a mobile monitoring network in Monroe County, New York to measure pollutant concentrations of black carbon (BC, wavelength at 880 nm), ultraviolet black carbon (UVBC, wavelength at 3700 nm) and Delta-C (the difference between the UVBC and BC concentrations) using the Clarkson University Mobile Air Pollution Monitoring Laboratory (MAPL). A Deletion/Substitution/Addition (D/S/A) algorithm was conducted, which used circular buffers as a basis for statistics. The algorithm maximizes the prediction accuracy for locations without measurements using the V-fold cross-validation technique, and it reduces overfitting compared to other approaches. We found that the D/S/A LUR modeling approach could achieve good results, with prediction powers of 60%, 63%, and 61%, respectively, for BC, UVBC, and Delta-C. The advantage of mobile monitoring is that it can monitor pollutant concentrations at hundreds of spatial points in a region, rather than the typical less than 100 points from a fixed site saturation monitoring network. This research indicates that a mobile saturation sampling network, when combined with proper modeling techniques, can uncover small area variations (e.g., 10 m) in particulate matter concentrations.

1. Introduction

Epidemiological studies examining associations between increased air pollutant concentrations and health endpoints (e.g. mortality, stroke, myocardial infarction) have used both pollutant data directly measured at a monitoring station in the city/region and estimates of pollution generated by various air pollution modeling techniques. These modeling techniques include inverse distance weighting (IDW) and kriging (Brauer et al., 2008; Mercer et al., 2011), land use regression (LUR) modeling (Su et al., 2008a, 2010, 2009a), spatiotemporal models such as Bayesian Maximum

* Corresponding author.
E-mail address: jasonsu@berkeley.edu (J.G. Su).

http://dx.doi.org/10.1016/j.atmosenv.2015.10.002
1352-2310/

Entropy (De Nazelle et al., 2010; de Nazelle and Serre, 2006) and dispersion models (Beevers et al., 2012; Gulliver and Briggs, 2011; Lepeule et al., 2011). LUR modeling has emerged as a method for estimating exposure to air pollution in epidemiological studies that reduces misclassification errors. Data from routine government fixed-site monitoring or special-purpose designed fixed-site networks are usually modeled to derive estimated air pollution exposure distributions across time and space.

Because of much smaller spatial variation gradients in particulate matter (PM) compared to other pollutants such as nitrogen oxides (NO_X), it generally requires greater spatial coverage of monitoring to reflect PM's inherent spatial gradients. Through our previous studies, we found that mobile monitoring and subsequent modeling can predict $PM_{2.5}$ (fine particulate matter of diameter $\leq$2.5 μm) concentrations reasonably well, with prediction powers (variance being explained) reaching 58–84% (Allen et al., 2011a; Larson et al., 2007; Su et al., 2006, 2013, 2007). These studies, however, were conducted in regions with relatively greater variation in surface elevation and mobile monitoring was conducted at night, when the impact from traffic was minimal.

This study takes place in Monroe County, New York that is situated on the southern shore of Lake Ontario in western New York, with elevation ranging from 73 m to 385 m (Fig. 1). In our effort to examine associations between myocardial infarction and acute increases in the concentration of ambient woodsmoke, we developed a Deletion/Substitution/Addition (D/S/A) LUR technique to model PM exposures for residents in the region, including exposures to Black Carbon (BC – optical absorption at 880 nm: a proxy for traffic pollution) (Cheng et al., 2014), Ultraviolet BC (UVBC – optical absorption at 370 nm: including air pollution from residential woodsmoke, traffic, and industrial activities) (Cheng et al., 2014) and Delta-C (enhanced optical absorption at 370 relative to 880 nm: a proxy for industrial activities and woodsmoke) (Allen et al., 2011b; Cheng et al., 2014).

The D/S/A algorithm maximizes the prediction accuracy for locations without measurements using the V-fold cross-validation technique and it reduces overfitting relative to other traditional approaches (see the method section for more details). It has been demonstrated to have optimal properties for deriving and assessing performance of predictive models (Davies and van der Laan, 2012; van der Laan et al., 2004). Traditional LUR modeling techniques use all the data in a linear model without considering cross-validation, and they generally overestimate the accuracy of prediction observations (Subramanian and Simon, 2013). By fitting the model to subsets of the data and examining predictive accuracy on the samples excluded from the model development, it is possible to assess directly the predictive accuracy of a particular statistical model. One of the most common forms of cross-validation is the 'leave-one-out' (LOO) method, in which the model is fit repeatedly leaving out a single observation and validated against the excluded observation. Within the machine learning literature, it is widely accepted that LOO is a suboptimal method for cross-validation since it gives estimates of the prediction error that are more variable than other forms of cross-validation, such as V-fold or bootstrap (Efron, 1983). In contrast to hold-out evaluations (HEV) (i.e., one group of sites being used as an independent validation set) (Johnson et al., 2010; Wang et al., 2013, 2014), the D/S/A modeling technique reduces prediction error in case the HEV group does not represent the population. For repeated random sub-sampling, some observations may never be selected in the validation subsample, whereas others may be selected more than once. In other words, validation subsets may overlap. In D/S/A, all observations in the V-folds are used for both training and validation, and each observation is used for validation exactly once. Ten-fold (V = 10) cross-validation is commonly used. Since each time an independent validation dataset is used to assess the performance of a model

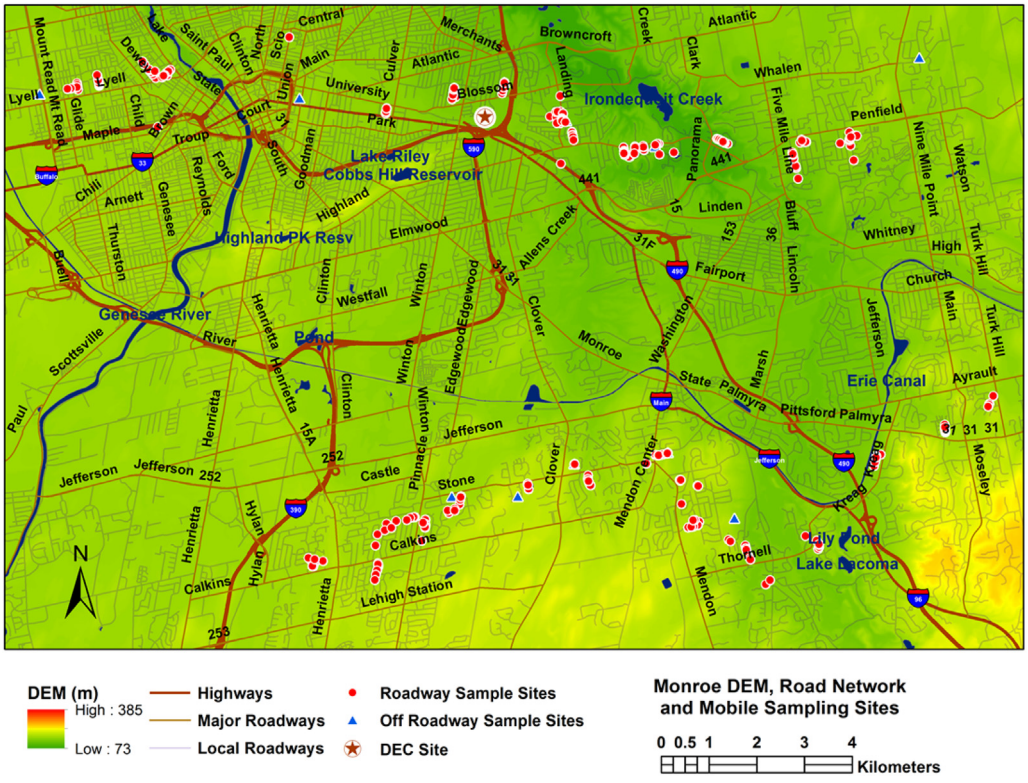

Fig. 1. The Monroe County elevation, road network and mobile sampling sites.

built using a training dataset, the V-fold cross-validation implemented in D/S/A minimizes the chance of over-fitting the model to the data.

In this study, we generated land use regression models using the D/S/A techniques for the Monroe County, New York and the modeling results will be used to estimate the daily concentrations of these pollutants for the few days before the symptom onset at the homes of patients (all Monroe County residents) who experienced a myocardial infarction.

2. Materials and methods

2.1. Design of mobile monitoring

Two routes were established for an "urban" area, and two "suburban" routes were defined (Fig. 1). The "urban" routes were chosen to be close to I-490 with Urban 1 being within the ring road around Rochester (Fig. S3). Urban 2 was further east along the northern side of I-490. (Fig. S4). The two suburban routes were further to the south but always north of I-90 (New York State Thruway) (Figs. S1 and S2). We used the prior measurements to provide data from the more urban areas of the domain and concentrated the additional runs on more suburban to rural areas (Wang et al., 2011a). Further, our sampling was designed to be near the known residential locations of the subjects of the myocardial infarction health study for which these models will provide exposure estimates. Measurements were made on roadways or other public properties such as parking lots. Maps for each of these routes are provided in the Supplemental Material Figures S1–S4. Three mobile monitoring campaigns were conducted in November and December of 2013 and March 2014, in morning (07:00–10:00), afternoon (14:00–17:00) and evening hours (19:00–22:30) using the Clarkson Mobile Air Pollution Monitoring Laboratory (MAPL) (Wang et al., 2011b). The detailed schedules for each of the three measurement campaigns are provided in Tables S1–S3 of the Supplemental Material. The objective of this work was to characterize the exposure to wood smoke and these periods reflect the part of the year when significant amounts of wood smoke have been found in prior studies (Wang et al., 2012a, 2012b).

2.2. Instrumentation

The MAPL is a recreational vehicle equipped with batteries and an inverter system that permits the operation of instruments without external power. The MAPL measures a comprehensive array of pollutants including both gaseous and particulates pollutants. For this study we only discuss its application to identify BC, UVBC and Delta-C through a two-wavelength Aethalometer (Magee Scientific Model AE42) (at a time resolution of 1 min). All of the mobile measurements included location information (i.e., latitude and longitude) and the start and stop time of pollutant measurements based on a GPS data logger. However, the concentrations of the measured pollutants were recorded in an instrument different from the GPS data logger. The GPS data logger recorded time and locational data with an interval of 15 s while the Aethalometer logged time and pollutant concentrations every 1 min. Every measured pollutant concentration was assigned locational data based on matching its time to the nearest time from the GPS data logger. Through this matching technique, all the Aethalometer measurements were assigned corresponding locational data and mean pollutant concentrations were calculated for each location.

2.3. Data preparation

The Aethalometer estimates the BC mass concentrations from the rate of change of light transmission through a filter. A routine method of post-processing was applied to the raw BC (880 nm) and UVBC (370 nm) data that involved the ONA algorithm for reduction of noise and a loading correction factor determined by Wang et al. (2011a). From the estimated BC and UVBC concentrations, the values of Delta-C were calculated as the difference between UVBC and BC (Allen et al., 2011b; Wang et al., 2011a).

2.4. Development of spatial covariates

Prior studies based on receptor modeling of data collected at the central New York State Department of Environmental Conservation (NYS DEC) (Wang et al., 2012a) site have shown that residential woodsmoke and traffic make substantial contributions to $PM_{2.5}$ air pollution in Rochester. Therefore, we developed spatial covariates that included sources of both (e.g., traffic and residences with fireplaces) or lack of both sources (e.g., water and vegetation). Specifically, the spatial predictors we obtained or developed included:

- The number of bedrooms, fireplaces, kitchens, property value and property year built for all residences derived from the Monroe County property assessment data for 2013;
- The geocoded locations of highway, major, and local roadways using road network data provided by the Environmental Science Research Institute (ESRI) Business Analyst 2010 (Redlands, CA);
- Highway and major roadway annual average daily traffic counts (AADT) obtained from the Highway Performance Management System (HPMS) data (U.S. Department of Transportation, DC) for 2011;
- The Digital Elevation Model (DEM; with a spatial resolution 10 m) data obtained from the United States Geological Survey (USGS) (Reston, VA);
- Land use and land cover data derived from the USGS National Land Cover Database (NLCD) for 2011 including (1) natural vegetation for forest, shrub land, and herbaceous plants (%,Veg1); (2) land use of largely vegetation for natural vegetation plus developed open space (%, Veg2); (3) natural environments for all land cover types other than developed low intensity, medium intensity, high intensity and water (%, NE1); (4) NE1 plus water (%, NE2), and (5) developed high intensity land cover (%, DHI);
- Monthly precipitation (kg m-2), specific humidity (10-2 kg/kg), air temperature (K), zonal (east–west) wind speed (m s-1) and Meridional (north–south) wind speed (m s-1) for 2010–2014 obtained from the North America Land Data Assimilation System Phase 2 (NLDAS-2).

2.5. D/S/A LUR modeling approach

2.5.1. Aggregate data using topic models

The location data and corresponding measured concentrations were further aggregated using topic models (Su et al., 2015) by either road segment (with a linear buffer zone of 10 m from a roadway) or property boundary for all of the measurement time periods. In topic models, an individual's exposure space is separated into different spatial locations including stationary (including home, work and other stationary locations) and in-transit (including commute and other travel) locations. Time weighted metrics is then applied to estimate the overall exposure of the individual in the time period of interest. In this application, for those road segments with more than one measurement location, a weighted location on a road segment (mean location along each measured locations on that road segment) was used as the location

for the calculated mean pollutant concentrations for that road segment. The same data aggregation algorithm was used for the measurement locations on public properties other than road segments (e.g., parking lots). After applying the topic models on the data collected in each of the three months, the mobile data were aggregated into 122 spatial points along the mobile unit routes. Also, we separated data into morning, afternoon, and evening measurements.

2.5.2. Identify distance decay functions for PM prediction variables

We mapped the decay of each predictor as a function of distance with respect to each of the three pollutants with distance ranging from 50 m to 5000 m at increments of 50 m. The distance decay curves serve as a visual tool to identify trend of impact at which each predictor has on the pollutant outcome value, the direction of its impact, and the possible maximum distance within which it affects the variable prediction. Correlation coefficients between a pollutant concentration (i.e., corrected BC, UVBC, and Delta-C) and all the identified predictors were calculated, separately, for daily, morning, afternoon, and nighttime data only.

2.5.3. D/S/A LUR modeling approach

We applied the D/S/A modeling technique to conduct LUR for the region. The details of the D/S/A LUR modeling approach can be found in our previous studies in California to predict concentrations of Nitrogen Oxides (NO_X) and Particulate Matter (PM) in California, USA (Beckerman et al., 2013a, 2013b; Su et al., 2015). The D/S/A program is a data-adaptive estimation method from which estimator selection is based on cross-validation under user-specified constraints. D/S/A could be specified as a standard multivariate linear regression with polynomial functions. D/S/A algorithm also supports various number of interaction terms. Based on our previous experience in applying the D/S/A algorithm for LUR modeling, we typically included squared terms (a power of 2) in our modeling process. The cross-validation scheme used here is called V-fold cross-validation. In the modeling process, the original sample is randomly partitioned into V (the number of folds) equal size subsamples. Of the V subsamples, a single subsample is retained as the validation data for testing the model, and the remaining V-1 subsamples are used as training data. The cross-validation process is then repeated V times, with each of the V subsamples used exactly once as the validation data. Ten-fold cross-validation is commonly used. Since each time an independent validation dataset is used to assess the performance of a model built using a training dataset, the V-fold cross-validation implemented in D/S/A minimizes the chance of over-fitting the model to the data. The V fold results can then be averaged (or otherwise combined) to produce a single estimation.

The D/S/A algorithm assumes the existence of a non-linear (i.e., polynomial) relationship between pollutant concentrations and a set of predictors, which is supported by past research (Aldrin and Haff, 2005; Tsiros et al., 2009). In our analysis, we allowed first and second order polynomial functions in the models. The polynomial functions implemented in the D/S/A algorithm might create a situation in which at times the squared term and its linear term for the same variable showed opposite signs. However, they reflected conditional mean-effects on a non-linear relationship between the predictor and the pollutant concentrations. We did not include any interaction in the modeling process because the inclusion of both interaction and squared terms made the models complicated to interpret and at the same time the models had very limited improvements. The software used in this research (version 3.1.4) can be accessed through http://www.stat.berkeley.edu/~laan/Software/DSA/DSA_3.1.4.zip for the Windows version and through http://www.stat.berkeley.edu/~laan/Software/DSA/DSA_3.1.4.tar.gz for the Unix version.

3. Modeling results

3.1. Distance decay functions of prediction variables

There are two types of variables used in the modeling. The buffer variables are local area specific variables that are averaged over a defined buffer distance (e.g., 10, 50, 200 m). Other variables are assumed to be uniform across the domain such as wind speed, temperature, and humidity. Based on the distance decay functions, we found that the relationships between air pollutant concentrations and the predictors were relatively weak, except for pollutant data collected in the morning hours. Therefore, only the data collected in the morning hours were used in the final LUR modeling. We understand that using the data collected from morning hours might not represent the daily means (Baldwin et al., 2015; Choi et al., 2012; Durant et al., 2010; Hu et al., 2009); however, in our study, they represented greater spatial variations compared to the daily means. The purpose of our LUR modeling was to identify the spatial variations of the three pollutants for the subsequent assessment of health outcomes.

Figs. S5–S7 show the distance decay curves for the predictors that had significant associations with corresponding pollutant concentrations or were used in the final models for the morning hours (7:00–10:00). Table 1 provides corresponding statistics for the non-buffer variables. Elevation was found having the highest correlations with BC, UVBC, and Delta-C with all of the relationships being negative. Both higher AADT and higher portions of vegetation were associated with higher BC and UVBC concentrations. For the number of fireplaces, when the buffer distances were below 400 m or above 2550 m, there was a positive association with BC and UVBC concentrations; for distances between 400 and 2550 m, the relationships were negative. For Delta-C, the positive associations were much smaller than those with BC and UVBC. The negative associations with fireplace ranged from 400 m to 2900 m, a much greater scope than those for BC and UVBC. Also, the portion of major roadway showed a positive association with Delta-C.

For non-buffer statistics, the meridional wind and precipitation were negatively associated with pollutant concentrations; while latitude, longitude, humidity, and temperature were positively associated with pollutant concentrations.

3.2. D/S/A LUR modeling results

We have 80 independent sampling points and $V = 10$. For each model run, 72 points were used for model training and 8 points used for model validation. The models for BC, UVBC, and Delta-C are presented in Table 2, and the comparison between the predicted and the measured concentrations are shown in Fig. 2. As expected, elevation was significantly and negatively associated with all three pollutants. Developed high intensity land use, portions of highway roadways and # fireplaces were significantly associated with BC and UVBC concentrations. At a closer distance (900 m), % highway had a significant contribution to the higher levels of Delta-C; while at long distance (~4250 m), due to the conditional means with impacts from its other buffer distances, the contribution was negatively but significantly associated with levels of Delta-C. Local roadways were found negatively and significantly associated with all the three pollutants.

4. Discussions and conclusions

In this paper, we have modeled spatially dispersed measurements of fine particle BC, UVBC, and Delta-C through D/S/A LUR

Table 1
The correlation coefficients between pollutant concentration and non-buffer predictors.

	Latitude (m)	Longitude (m)	Meridional wind (m s^{-1})	Zonal wind (m s^{-1})	Precipitation (kg m^{-2})	Specific humidity (10^{-2} kg/kg)	Temperature (K)
BC	0.47	0.15	−0.24	0.05	−0.11	0.49	0.44
UVBC	0.52	0.15	−0.24	0.06	−0.12	0.54	0.48
Delta-C	0.40	0.11	−0.17	0.05	−0.10	0.42	0.37

Table 2
Associations between BC, UVBC, and Delta-C concentrations (ng/m^3) and each predictor, using the D/S/A LUR model.

		Estimate	Standard error	t value	p-value	R^2
BC	(Intercept)	1780.00	184.00	9.68	<0.001	0.60
	DEM^2(500 m)	−0.03	0.01	−5.93	<0.001	
	% NE1 (100 m)	−5.76	1.11	−5.21	<0.001	
	% DHI (50 m)	16.60	4.37	3.79	<0.001	
	% Highway (200 m)	−309.00	70.40	−4.39	<0.001	
	# Fireplaces^2(150 m)	0.24	0.06	4.26	<0.001	
	% Highway^2 (1550 m)	−66.70	17.30	−3.85	<0.001	
	% Highway (800 m)	204.00	74.70	2.74	0.008	
	% Local roadway (50 m)	−9.96	4.46	−2.23	0.029	
UVBC	(Intercept)	2878.00	353.00	8.16	<0.001	0.63
	DEM (50 m)	−14.40	1.90	−7.53	<0.001	
	% DHI (50 m)	41.50	7.90	5.26	<0.001	
	% Local roadway (50 m)	−28.00	6.15	−4.56	<0.001	
	# Fireplaces (150 m)	11.50	3.66	3.14	<0.001	
	% Highway^2 (250 m)	−80.90	16.10	−5.02	<0.001	
	# Kitchens (400 m)	1.77	0.38	4.64	<0.001	
	% Highway^2 (1550 m)	−105.00	21.40	−4.93	<0.001	
	% Highway (700 m)	551.00	121.00	4.55	<0.001	
Delta-C	(Intercept)	1390.00	169.00	8.21	<0.001	0.61
	DEM (200 m)	−5.44	0.85	−6.42	<0.001	
	% Highway (4250 m)	−292.00	56.9	−5.13	<0.001	
	% DHI (2300 m)	28.60	4.09	6.98	<0.001	
	% Local roadway (50 m)	−35.10	8.50	−4.13	<0.001	
	% Veg2 (150 m)	3.13	0.79	3.94	<0.001	
	% Local roadway^2 (50 m)	0.65	0.26	2.55	0.013	
	AADT^2 (50 m)	−0.0000001	0.00000003	−3.44	0.001	
	% Highway (900 m)	147.00	54.10	2.71	0.009	
	% Highway^2 (1150 m)	−32.50	12.60	−2.57	0.013	

Goodness of fit: BC model: residual standard error = 232.5 on 66 degrees of freedom (DF), F-statistic = 12.18 on 8 and 66 DF, p-value <0.001; UVBC model: residual standard error = 311.8 on 66 degrees of freedom, F-statistic = 13.78 on 8 and 66 DF, p-value <0.001; Delta-C model: residual standard error = 141.2 on 65 degrees of freedom, F-statistic = 11.45 on 9 and 65 DF, p-value <0.001.
DEM = digital elevation model in meters; NE1 = natural environments for all land cover types other than developed low intensity, medium intensity, high intensity and water; DHI = developed high intensity land cover; Veg2 = land use of largely vegetation including forest, shrub land, herbaceous plants and developed open space; AADT = annual average daily traffic.

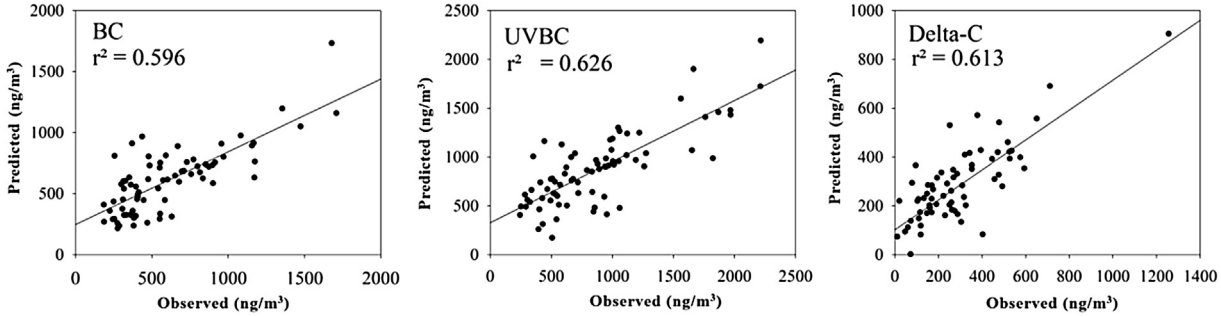

Fig. 2. The comparison between the predicted and the measured concentrations for BC, UVBC, and Delta-C.

modeling techniques. We found that the D/S/A LUR modeling approach can achieve good results. The D/S/A LUR modeling algorithm maximizes accuracy for predictions on out-of-sample observations. The V-fold cross-validation technique applied in D/S/A significantly reduces over-fitting of the model to the data. In addition, the algorithm supports modeling non-linear (i.e., polynomial) relationship between pollutant concentrations and a set of predictors, revealing more realistic associations between them.

In previous studies (Su et al., 2013, 2008b), mobile measurements of fine particles were used as input data to model air pollution from residential woodsmoke. That model was built for late evening periods when the impacts of traffic were minimal. During the day, especially during morning hours, both residential woodsmoke and traffic contribute to ambient PM concentrations.

Delta-C is considered as a marker for residential woodsmoke. However, the distance decay curves and modeling results used in this study found that highways also partially contributed to increases in Delta-C (e.g., highway buffer 900 m). It might be true that Delta-C is generated specifically from residential woodsmoke during nighttime when there is minimum impact from traffic. During daytime, however, traffic has impacts not only on BC and UVBC concentrations but also probably on Delta-C concentrations. The positive associations between BC/UVBC and # fireplaces indicate that usage of fireplace also contributed to the increase of BC and UVBC concentrations. The D/S/A LUR model can be used to integrate the contributions of both traffic and residential woodsmoke into a single modeling framework. Though the daytime exposures included impacts from traffic and residential woodsmoke, our goal was still to model the times of day with maximum spatial variation. This is why only the data from the morning hours were used in the modeling process.

Based on the distance decay curves (Figs. S5–S7), we found that elevation explained more than twice of variance than other variables in pollutant concentrations and the distance decay curves showed that elevation had a different trend to other variables, indicating elevation's independent impact on pollutant concentrations. Although the region has a relatively small range in elevation (about 300 m), we suspect the drainage process, due to changes in elevation, have an impact on levels of concentrations being measured. This has been confirmed in our previous woodsmoke studies (Larson et al., 2007; Su et al., 2008b).

We limited our buffers to a maximum of 2000 m (except for traffic) and this is consistent with the scope of impact identified from our buffer distance decay curves in Figs. S5–S7. It is also consistent with literature on the distance of impact from roadways and land use characteristics. The buffer distances for the variables in the final models were largely less than 1000 m. Previous studies have even identified the distance of impact to be 5 km for $PM_{2.5}$ and NO_X (Beckerman et al., 2013a) or even 11 km for NO_X (Su et al., 2009b), and these distances of impact are considered as background effects.

This study designed two mobile transects, extending from the center of the city to the more rural areas east of the city. The transects crossed various types of land uses and roadways. However, since the mobile monitoring sites extended from east to west, the ideal situation would also to include, if time and financial possible, sampling along the north–south direction.

Similar to other validation techniques, D/S/A V-fold cross-validation only yields meaningful results if the validation set and training set are drawn from the same population. If the structure of the system evolves over time, it can introduce systematic differences between the training and validation sets. For example, meteorological conditions might change from year to year and the parameters selected for one year might not be effective for predicting concentrations for another year. In our study, no meteorological parameters were included in the final models, but the inclusion of traffic and land use characteristics might limit the modeling results to be effective for an extended period of time (e.g., more than 10 years). Given that urban structures are largely established, these changes might be minor for the immediate years and the models are assumed to be effective.

Because the D/S/A technique requires the samples to be divided into V-folds, if the number of sample sites is small, the training and validation data might not be big enough to make correct predictions and evaluations. We recommend having at least 60 sample sites for a D/S/A algorithm to be run. Another limit of the D/S/A algorithm is due to its model flexibility such as allowing both polynomial and interaction functions in the same model. These functions, when used together in one model, might make a model complicated for interpretation. Cautions should be taken if multiple functions are to be applied at the same time.

The advantage of mobile monitoring is that it can monitor pollutant concentrations at hundreds of spatial points in a region, rather than the typical less than 100 points from a fixed site saturation monitoring network. This research indicates that a mobile saturation sampling network, when combined with proper modeling techniques, can uncover small area variations (e.g., 10 m) in particulate matter concentrations.

Acknowledgments

This work was supported by the New York State Energy Research and Development Authority under contract no. #32971. The authors want to thank Dr. Michael Brauer of the University of British Columbia for helping to facilitate the work presented in this paper.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2015.10.002.

References

Aldrin, M., Haff, I.H., 2005. Generalised additive modelling of air pollution, traffic volume and meteorology. Atmos. Environ. 39, 2145–2155.
Allen, G.A., Miller, P.J., Rector, L.J., Brauer, M., Su, J.G., 2011a. Characterization of valley winter woodsmoke concentrations in Northern NY using highly time-resolved measurements. Aerosol Air Qual. Res. 11, 519–530.
Allen, G.A., Miller, P.J., Rector, L.J., Brauer, M., Su, J.G., 2011b. Characterization of valley winter woodsmoke concentrations in Northern NY using highly time-resolved measurements. Aerosol Air Qual. Res. 11 (5), 519–530.
Baldwin, N., Gilani, O., Raja, S., Batterman, S., Ganguly, R., Hopke, P., Berrocal, V., Robins, T., Hoogterp, S., 2015. Factors affecting pollutant concentrations in the near-road environment. Atmos. Environ. 115, 223–235.
Beckerman, B.S., Jerrett, M., Martin, R.V., van Donkelaar, A., Ross, Z., Burnett, R.T., 2013a. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. Atmos. Environ. 77, 172–177.
Beckerman, B.S., Jerrett, M., Serre, M., Martin, R.V., Lee, S.J., van Donkelaar, A., Ross, Z., Su, J.G., Burnett, R.T., 2013b. A hybrid approach to estimating national scale spatiotemporal variability of PM2.5 in the contiguous United States. Environ. Sci. Technol. 47 (13), 7233–7241.
Beevers, S.D., Kitwiroon, N., Williams, M.L., Carslaw, D.C., 2012. One way coupling of CMAQ and a road source dispersion model for fine scale air pollution predictions. Atmos. Environ. 59, 47–58.
Brauer, M., Lencar, C., Tamburic, L., Koehoorn, M., Demers, P., Karr, C., 2008. A cohort study of traffic-related air pollution impacts on birth outcomes. Environ. Health Perspect. 116, 680–686.
Cheng, Y.H., Lin, C.C., Liu, J.J., Hsieh, C.J., 2014. Temporal characteristics of black carbon concentrations and its potential emission sources in a southern Taiwan industrial urban area. Environ. Sci. Pollut. R 21, 3744–3755.
Choi, W., He, M.L., Barbesant, V., Kozawa, K.H., Mara, S., Winer, A.M., Paulson, S.E., 2012. Prevalence of wide area impacts downwind of freeways under pre-sunrise stable atmospheric conditions. Atmos. Environ. 62, 318–327.
Davies, Molly M., van der Laan, Mark J., December 2012. Optimal Spatial Prediction Using Ensemble Machine Learning. In: U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 305. http://biostats.bepress.com/ucbbiostat/paper305.
De Nazelle, A., Arunachalam, S., Serre, M.L., 2010. Bayesian maximum entropy integration of ozone observations and model predictions: an application for attainment demonstration in North Carolina. Environ. Sci. Technol. 44, 5707–5713.
de Nazelle, A., Serre, M.L., 2006. Ozone exposure assessment in North Carolina using Bayesian maximum entropy data integration of space time observations and air quality model prediction. Epidemiology 17. S189–S189.
Durant, J.L., Ash, C.A., Wood, E.C., Herndon, S.C., Jayne, J.T., Knighton, W.B., Canagaratna, M.R., Trull, J.B., Brugge, D., Zamore, W., Kolb, C.E., 2010. Short-term variation in near-highway air pollutant gradients on a winter morning. Atmos. Chem. Phys. 10, 8341–8352.
Efron, B., 1983. Model selection and the bootstrap. Math. Soc. Sci. 5, 236–236.
Gulliver, J., Briggs, D., 2011. STEMS-air: a simple GIS-based air pollution dispersion model for city-wide exposure assessment. Sci. Total Environ. 409, 2419–2429.
Hu, S.S., Fruin, S., Kozawa, K., Mara, S., Paulson, S.E., Winer, A.M., 2009. A wide area of air pollutant impact downwind of a freeway during pre-sunrise hours. Atmos. Environ. 43, 2541–2549.

Johnson, M., Isakov, V., Touma, J.S., Mukerjee, S., Ozkaynak, H., 2010. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmos. Environ. 44, 3660–3668.
Larson, T., Su, J., Baribeau, A.M., Buzzelli, M., Setton, E., Brauer, M., 2007. A spatial model of urban winter woodsmoke concentrations. Environ. Sci. Technol. 41, 2429–2436.
Lepeule, J., Galineau, J., Hulin, A., Bottagisi, S., Marquis, N., Caini, F., Bohet, A., Kaminski, M., Charles, M.A., Slama, R., 2011. Maternal exposure to urban air pollution during pregnancy assessed by a dispersion model and fetal growth. Epidemiology 22. S121–S121.
Mercer, L.D., Szpiro, A.A., Sheppard, L., Lindstrom, J., Adar, S.D., Allen, R.W., Avol, E.L., Oron, A.P., Larson, T., Liu, L.J.S., Kaufman, J.D., 2011. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the multi-ethnic study of atherosclerosis and air pollution (MESA Air). Atmos. Environ. 45, 4412–4420.
Su, J., Jerrett, M., Beckerman, B., 2008a. Modeling intra-urban spatial variability of volatile organic compounds using a land use regression method. Epidemiology 19, S312–S313.
Su, J., Larson, T., Baribeau, A.M., Brauer, M., Setton, E., Buzzelli, M., 2006. Modeling residential woodsmoke with socioeconomic variables extracted from hydrologically based buffers. Epidemiology 17. S483–S483.
Su, J.G., Allen, G., Miller, P., Brauer, M., 2013. Spatial modeling of residential woodsmoke across a non-urban upstate New York region. Air Qual. Atmos. Health 6, 85–94.
Su, J.G., Buzzelli, M., Brauer, M., Gould, T., Larson, T.V., 2008b. Modeling spatial variability of airborne levoglucosan in Seattle, Washington. Atmos. Environ. 42, 5519–5525.
Su, J.G., Jerrett, M., Beckerman, B., Verma, D., Arain, M.A., Kanaroglou, P., Stieb, D., Finkelstein, M., Brook, J., 2010. A land use regression model for predicting ambient volatile organic compound concentrations in Toronto, Canada. Atmos. Environ. 44, 3529–3537.
Su, J.G., Jerrett, M., Beckerman, B., Wilhelm, M., Ghosh, J.K., Ritz, B., 2009a. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. Environ. Res. 109, 657–670.
Su, J.G., Jerrett, M., Beckerman, B., Wilhelm, M., Ghosh, J.K., Ritz, B., 2009b. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. Environ. Res. 109, 657–670.
Su, J.G., Jerrett, M., Meng, Y.-Y., Pickett, M., Ritz, B., 2015. Integrating smart-phone based momentary location tracking with fixed site air quality monitoring for personal exposure assessment. Sci. Total Environ. 506–507, 518–526.
Su, J.G., Larson, T., Baribeau, A.M., Brauer, M., Rensing, M., Buzzelli, M., 2007. Spatial modeling for air pollution monitoring network design: example of residential woodsmoke. J. Air Waste Manag. Assoc. 57, 893–900.
Subramanian, J., Simon, R., 2013. Overfitting in prediction models - is it a problem only in high dimensions? Contemp. Clin. Trials 36, 636–641.
Tsiros, I.X., Dimopoulos, I.F., Chronopoulos, K.I., Chronopoulos, G., 2009. Estimating airborne pollutant concentrations in vegetated urban sites using statistical models with microclimate and urban geometry parameters as predictor variables: a case study in the city of Athens Greece. J. Environ. Sci. Health A 44, 1496–1502.
van der Laan, M.J., Dudoit, S., Keles, S., 2004. Asymptotic optimality of likelihood-based cross-validation. Stat. Appl. Genet. Mol. Biol. 3. Article4.
Wang, M., Beelen, R., Basagana, X., Becker, T., Cesaroni, G., de Hoogh, K., Dedele, A., Declercq, C., Dimakopoulou, K., Eeftens, M., Forastiere, F., Galassi, C., Grazuleviciene, R., Hoffmann, B., Heinrich, J., Iakovides, M., Kunzli, N., Korek, M., Lindley, S., Molter, A., Mosler, G., Madsen, C., Nieuwenhuijsen, M., Phuleria, H., Pedeli, X., Raaschou-Nielsen, O., Ranzi, A., Stehanou, E., Sugiri, D., Stempfelet, M., Tsai, M.Y., Lanki, T., Udvardy, O., Varro, M.J., Wolf, K., Weinmayr, G., Yli-Tuomi, T., Hoek, G., Brunekreef, B., 2013. Evaluation of land use regression models for NO2 and particulate matter in 20 European study areas: the ESCAPE project. Environ. Sci. Technol. 47, 4357–4364.
Wang, M., Beelen, R., Bellander, T., Birk, M., Cesaroni, G., Cirach, M., Cyrys, J., de Hoogh, K., Declercq, C., Dimakopoulou, K., Eeftens, M., Eriksen, K.T., Forastiere, F., Galassi, C., Grivas, G., Heinrich, J., Hoffmann, B., Ineichen, A., Korek, M., Lanki, T., Lindley, S., Modig, L., Molter, A., Nafstad, P., Nieuwenhuijsen, M.J., Nystad, W., Olsson, D., Raaschou-Nielsen, O., Ragettli, M., Ranzi, A., Stempfelet, M., Sugiri, D., Tsai, M.Y., Udvardy, O., Varro, M.J., Vienneau, D., Weinmayr, G., Wolf, K., Yli-Tuomi, T., Hoek, G., Brunekreef, B., 2014. Performance of multi-city land use regression models for nitrogen dioxide and fine particles. Environ. Health Perspect. 122, 843–849.
Wang, Y.G., Hopke, P.K., Rattigan, O.V., Zhu, Y.F., 2011a. Characterization of ambient black carbon and wood burning particles in two urban areas. J. Environ. Monit. 13, 1919–1926.
Wang, Y.G., Hopke, P.K., Utell, M.J., 2011b. Urban-scale spatial-temporal variability of black carbon and winter residential wood combustion particles. Aerosol Air Qual. Res. 11, 473–481.
Wang, Y.G., Hopke, P.K., Utell, M.J., 2012a. Urban-scale seasonal and spatial variability of ultrafine particle number concentrations. Water Air Soil Pollut. 223, 2223–2235.
Wang, Y.G., Hopke, P.K., Xia, X.Y., Rattigan, O.V., Chalupa, D.C., Utell, M.J., 2012b. Source apportionment of airborne particulate matter using inorganic and organic species as tracers. Atmos. Environ. 55, 525–532.