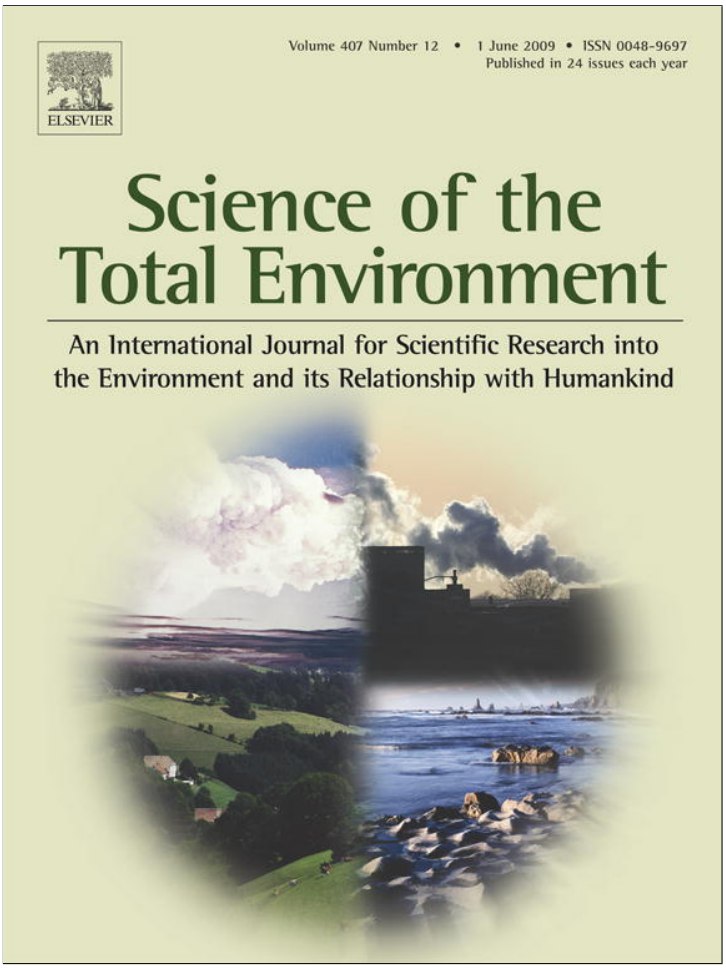

Science of the Total Environment 407 (2009) 3890–3898

Contents lists available at ScienceDirect

Science of the Total Environment

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

A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures

J.G. Su, M. Jerrett *, B. Beckerman

Environmental Health Sciences, School of Public Health, University of California, Berkeley, 50 University Hall, Berkeley, CA, 94720-7360, USA

ARTICLE INFO

Article history:
Received 28 August 2008
Received in revised form 13 January 2009
Accepted 28 January 2009
Available online 21 March 2009

Keywords:
Land use regression
Air pollution
GIS
Spatial distance decay
Model selection

ABSTRACT

Land use regression (LUR) has emerged as an effective and economical means of estimating air pollution exposures for epidemiological studies. To date, no systematic method has been developed for optimizing the variable selection process. Traditionally, a limited number of buffer distances assumed having the highest correlations with measured pollutant concentrations are used in the manual stepwise selection process or a model transferred from another urban area.

In this paper we propose a novel and systematic way of modeling long-term average air pollutant concentrations through "A Distance Decay REgression Selection Strategy" (ADDRESS). The selection process includes multiple steps and, at each step, a full spectrum of correlation coefficients and buffer distance decay curves are used to select a spatial covariate of the highest correlation (compared to other variables) at its optimized buffer distance. At the first step, the series of distance decay curves is constructed using the measured concentrations against the chosen spatial covariates. A variable with the highest correlation to pollutant levels at its optimized buffer distance is chosen as the first predictor of the LUR model from all the distance decay curves. Starting from the second step, the prediction residuals are used to construct new series of distance decay curves and the variable of the highest correlation at its optimized buffer distance is chosen to be added to the model. This process continues until a variable being added does not contribute significantly ($p>0.10$) to the model performance. The distance decay curve yields a visualization of change and trend of correlation between the spatial covariates and air pollution concentrations or their prediction residuals, providing a transparent and efficient means of selecting optimized buffer distances. Empirical comparisons suggested that the ADDRESS method produced better results than a manual stepwise selection process of limited buffer distances. The method also enables researchers to understand the likely scale of variables that influence pollution levels, which has potentially important ramifications for planning and epidemiological studies.

Published by Elsevier B.V.

1. Introduction

Recent studies have shown that the spatial variability of selected air pollutants within urban areas is greater than typically recognized and is associated with previously unaccounted for variability in health impacts (Hoek et al., 2002; Gilbert et al., 2005; Finkelstein and Jerrett, 2007; Miller et al., 2007). The development of models to assess air pollution exposures within cities for assignment to subjects in health studies has therefore been identified as a research priority (Brunekreef and Holgate, 2002; Brauer et al., 2003; Moore et al., 2007). These exposure assessment methods include proximity-based assessments, statistical interpolation, land use regression (LUR) models, new uses of line dispersion models, integrated emission-meteorological models, and hybrid models (Jerrett et al., 2005). While surrogate measures, such as distance to roads, have been related to large health effects (Hoek et al., 2002), these may misclassify exposure because they are not directly estimated from monitored data. Potential alternatives to surrogate measures arise from geographic and dispersion exposure methods. These methods utilize geographic information systems (GIS) to combine available geographic data with short-term monitoring information to develop exposure models capable of identifying small-area variations in pollution. Results from these models can then be overlaid on geo-referenced health data to assign exposure to individuals at their place of residence, work, or some combination of both. Among the assessment methods, LUR is a promising approach that seeks predicting pollution concentrations at a given site based on surrounding land use, traffic, physical geography and population characteristics. The main strength of LUR is the empirical structure (e.g., selection of optimized buffer size) of the regression mapping and its relatively simple inputs and low cost (as compared with dispersion modeling, for example; Jerrett et al., 2005).

LUR was first introduced in the SAVIAH (Small Area Variations In Air quality and Health) study (Briggs et al., 1997) and has been used extensively for exposure analysis and environmental health research (Briggs et al., 1997; Brauer et al., 2003; Jerrett et al., 2005; Bell, 2006;

* Corresponding author. Tel.: +1 5106423960; fax: +1 5106425815.
E-mail address: jerrett@berkeley.edu (M. Jerrett).

0048-9697/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.scitotenv.2009.01.061

Hochadel et al., 2006; Liao et al., 2006; Ross et al., 2006; Sahsuvaroglu et al., 2006; Henderson et al., 2007; Jerrett et al., 2007; Moore et al., 2007; Aguilera et al., 2008). Selection of spatial covariates at appropriate distances of influence (e.g., distance to road, industrial land use, etc.) is important for determining final model performance. To date no systematic approach has been identified on selection of those distances (i.e., circular areas or buffers) of influence for exposure analysis. Typically less than 10 circular buffer distances considered having high correlations with the dependent variable are chosen for a variable (Henderson et al., 2007; Aguilera et al., 2008). Because of the differences in topography, meteorology, land use, traffic and population composition of one urban area to another, the buffer distances suitable for one urban area might not be the best choice for another. In addition, the optimized distance of the highest correlation of a variable might be mistakenly selected. The transferability of LUR from one urban area to another is very limited not only by the availability of equivalent variables but also by the buffer distances used (Briggs et al., 1997; Poplawski et al., 2009; Su et al., in press). Hoek et al. (2008) suggested that differences in variable selection strategy could play an important role in model prediction differences between studies. We suggest that a distance decay curve of correlation should be calculated first on each available variable to identify the optimized distance of corresponding variable and a series of distance decay curves applied in the model optimization process in any given region of research interest.

In selection of a LUR model, current research uses best subsets, manual forward (Jerrett et al., 2007) or automated stepwise (Aguilera et al., 2008) selection process based on the limited number of buffers. Because of limited knowledge of the distance decay of influence of a variable, a LUR may not be optimized to produce the optimal prediction result based on substantive understanding of the physical processes generating the emissions and controlling the transport of the pollutants in the atmosphere. This paper uses A Distance Decay REgression Selection Strategy (ADDRESS) to select variables of the highest correlation (compared to other variables) at optimized buffer distances through a series of distance decay curves in a multi-step selection process. As previous land use regression models were typically calibrated with NO_2 (nitrogen dioxide) as a marker for traffic exposure, this selection process was also demonstrated using NO_2 measured in the spring of 2004 in Toronto, Canada as an example.

2. Materials and methods

2.1. Pollution sampling

NO_2 was measured for a two-week period during the spring of 2004 using duplicate two-sided Ogawa passive diffusion samplers at 100 locations across Toronto. Sampling locations were determined using a location-allocation approach outlined in the paper by Kanaroglou et al. (2005). The outcome of using a location-allocation model is a sampling network that better captures the inherent variability in city-wide exposures (Jerrett et al., 2007).

2.2. Model variable and distance decay curve

The LUR model was developed by regressing the NO_2 measurements on spatial covariates chosen for the City of Toronto. The main spatial covariates included five major categories: (1) Tasseled-cap transformation indices, based on Landsat images that yield measures of greenness and soil brightness (Crist and Cicone, 1984); (2) land use characteristics (commercial, industrial, residential, and open); (3) population density; (4) physical geography such as geographic coordinates, elevation, distance to coast; and (5) transportation systems such as highway (including expressway, primary and secondary highway), major and local road as well as railway lengths, highway and major road slope gradients, and major road traffic density. Expressway casement (in ha), representing both road length and width in the form of polygon, was also included to see if road length plus width could improve model prediction power.

To assist in selecting spatial covariates for LUR, 30 circular area distances (buffers) of interval 50 m were created for each sampler, ranging from 0–50 m, 0–100 m, 0–150 m, and up to 0–1500 m for traffic related sources, and 60 buffers to a maximum distance of 3000 m for land use, Tasseled-cap transformation and population density. The correlation of NO_2 with the covariates at each buffer distance was calculated and the distance decay curves of correlation of all the covariates were displayed in a single chart. The distance decay curves provide us an intuitive view of change and trend of correlation of NO_2 with the selected spatial covariates at those buffer distances and help selection of spatial covariates at buffer distances of the highest significance. Importantly this method also reduces impact of collinearity in the selection of model parameters, which has been a major limitation of the forward selection strategies employed in most land use regression models.

2.3. ADDRESS and model diagnostics

As traffic related concentrations decay from roadway outwards and the assumed maximum distance of influence is 1500 m (Jerrett et al., 2005; Henderson et al., 2007), the maximum buffer distance for traffic related covariates was set to be 1500 m. Similarly, the land use, Tasseled-cap transformation and population density variables were assumed to have an influence up to 3000 m and the maximum buffer distance was therefore restricted to a maximum distance of 3000 m. ADDRESS is illustrated in Fig. 1 as follows: first, the correlations of the measured NO_2 with all the spatial covariates were calculated at all the 30 or 60 buffer distances and displayed in a series of distance decay curves; second, a spatial covariate with the highest correlation (compared to other variables) at an optimized buffer distance was tested first. The optimized buffer distance was at its highest correlation on the distance decay curve if the highest correlation distance was not on a flattened curve. When the highest correlation distance was on a plateau proportion of the distance decay curve, the highest slope change distance was chosen as the optimized buffer distance. Expert judgment was given to choose an optimized buffer distance. If the chosen optimized buffer distance was significant at the entry level ($p = 0.10$), the spatial variable was added to the model, and the prediction residuals of the bivariate model were calculated for all the sampling locations; third, the distance decay curves of correlation of prediction residuals were estimated and the spatial covariate of the highest significance at an optimized search distance was chosen to be added into the model. Though similar to the manual forward selection process where spatial covariates are added one at a time to the model (e.g., Jerrett et al., 2007; Mavko et al., 2008), we used a series of distance decay curves to select a spatial covariate at its optimized buffer distance. If a spatial covariate at its optimized buffer distance is incorporated into the model, the remaining buffer distances of that variable cannot be considered for further selection in ADDRESS unless (a) a correlation between the two buffer distances show no significant correlation, (b) the two distances are non-overlapping concentric bands (e.g., 0–50 m and 1000–3000 m), and (c) the buffer distance to be selected is the optimized distance with a significant correlation to prediction residuals. To facilitate the visual display of distance decay curves and to reduce computing power required to run the selection process, only overlapping distances were calculated first since selection of two buffer distances of the same variable based on observation that conditions (a) and (c) are rare. When satisfied at the same time, condition (b) was then added to further restrict the selection process. In addition, regardless of how many buffers were used to identify the spatial distance decay of correlation of a variable, at most two buffer distances of that variable were allowed to be selected. Fourth, when a spatial covariate at its optimized search distance was added into the model, the least significant variable of a

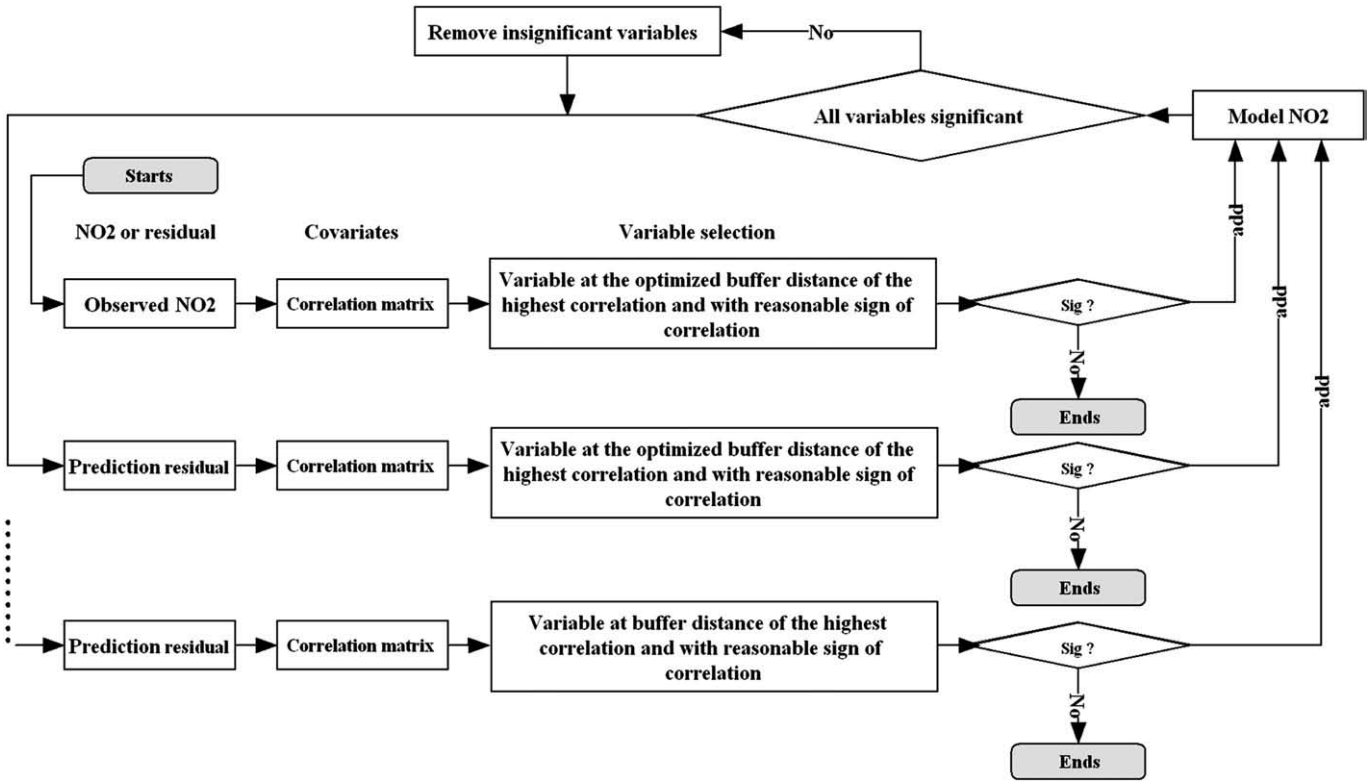

Fig. 1. Concept of a distance decay regression selection strategy based on NO_2 concentrations.

search distance that did not meet the significance level ($p = 0.10$) for staying in the model was removed. The selection process continued until no further spatial covariate of a buffer distance could be added to the model. Similar to other approaches (e.g., Henderson et al., 2007), variables were chosen if they had the expected sign based on correlation with the dependent variable (e.g., traffic density should be positively associated with NO_2 levels).

In model diagnostics, variance inflation factors (VIF) were examined to identify variables that were collinear and could be eliminated. To evaluate the independence assumption, we also tested spatial autocorrelation on the residuals from our final optimized models using the Moran's *I* statistic (Bailey and Gatrell, 1995), the first order contiguity matrix based on Thiessen polygons created from those 100 measurement sites. Statistical significance was tested using a permutation test with 999 iterations. Additional model diagnostics included Cook's distance to examine outliers.

To identify whether ADDRESS was an improvement over the previously developed techniques given the availability of all the buffer distance statistics, the manual stepwise selection process typically applied for land use regression modeling (Briggs et al., 1997; Sahsuvaroglu et al., 2006; Henderson et al., 2007; Jerrett et al., 2007; Morgenstern et al., 2007; Aguilera et al., 2008; Mavko et al., 2008) was used to model NO_2 levels using the same data as in this research. This involved the following steps: (1) identifying the optimized buffer distance for each variable based on the highest correlation with NO_2 levels; (2) running a stepwise selection process using the variables derived from step (1) and those non-buffer variables (e.g., *X* and *Y* coordinates, distance to coast); and (3) removing variables of changing coefficient signs and of high collinearity. The performance of the manual stepwise selection process was compared with ADDRESS. To illustrate whether ADDRESS could benefit land use regression modeling even with limited number of buffers, the ADDRESS model was applied to all the variables used in this research but with buffer distances limited to those typically used for land use regression modeling, including buffer distance 50, 100, 200, 300, 500, 750, 1000, 1250 and 1500 m for traffic related variables and plus 1750, 2000, 2500 and 3000 m for other variables (e.g., population density and land use variables). The ADDRESS model applied for the limited buffer distances was compared with the manual stepwise selection result as well as with the ADDRESS model applied for the full buffer distance statistics (i.e., 50–1500 m for traffic related variables and 50–3000 m for other land use variables, all at an interval of 50 m). Software packages used for ADDRESS modeling and corresponding model diagnostics included SPSS (SPSS Inc., Chicago, IL), Erdas Imagine 8.5 (ERDAS Inc., Atlanta, GA) for Tasseled-cap transformation and ArcGIS 9.2 (ESRI, Redlands, CA) for derivation of spatial covariates. Customized programs were used to conduct the model selection process in ArcGIS 9.2 and in Microsoft Office (Excel).

3. Results

The descriptive statistics of NO_2 sampling over 100 locations in Toronto, Canada are listed in Table 1. The NO_2 samples had a mean concentration of 10.15 ppb, with values ranging from 4.92 to 19.31 ppb and a standard deviation of 3.12 ppb.

3.1. Distance decay curves

The distance decay curves illustrated in Fig. 2a show that *X* coordinate has the highest negative correlation with NO_2, and 42.7% of

Table 1
Descriptive statistics of NO_2 sampling over 100 locations in Toronto, Canada.[a]

Minimum	Maximum	Mean	Std. error	Std. dev
4.92	19.31	10.15	0.31	3.12
1 quartile	Median	3 quartile	IQR	Full range
7.9	10.12	11.43	3.53	14.39

IQR: Inter-quartile range.
[a] NO_2 measurement unit: ppb.

the model variance could be explained by this variable, which is consistent with earlier studies in Toronto (Jerrett et al. 2007). NO_2 levels decreased from west to east were probably because of the decreasing highway density and industrial emissions from west to east. Vegetation greenness and open land use also have high correlations with NO_2 levels, with open land use a steady increase to the maximum buffer distance of 3000 m and greenness a slow increase to reach a peak at buffer distance 1300 m before a slow drop. Through the distance decay curves, we can see that open space does not necessarily represent places with vegetation. When the buffer distance is less than 250 m, the distance decay curve of open space even has positive correlations with NO_2 levels. The positive correlation demonstrates that near road infrastructures (e.g., parks) and possibly near road parking spaces were classified as open space, which contribute to the higher levels of NO_2. The distance decay curves in Fig. 2a show that in most cases, greenness has a higher prediction power than open space; this also demonstrates that open space has the mixed effect of green space (reduced emissions) and impervious surface (increased emissions). The distance decay curve for expressway demonstrates that correlations go up for the first 300 m. When the distance goes beyond 400 m, there is a sharp drop of correlation and it flattens at buffer distance 500 m. This affirms previous assumption that influence from traffic decreased significantly for the first 300–500 m (Henderson et al., 2007; Jerrett et al., 2007; Beckerman et al. 2008). By contrast, the influence of major roads drops to its lowest level when the distance is 200 m; however, the curve goes up again and flattens at buffer distance 700 m, possibly indicating two scales of influence—near source and urban background levels representing larger-area traffic density. With the assistance of distance decay curves, we could identify the distance of influence of

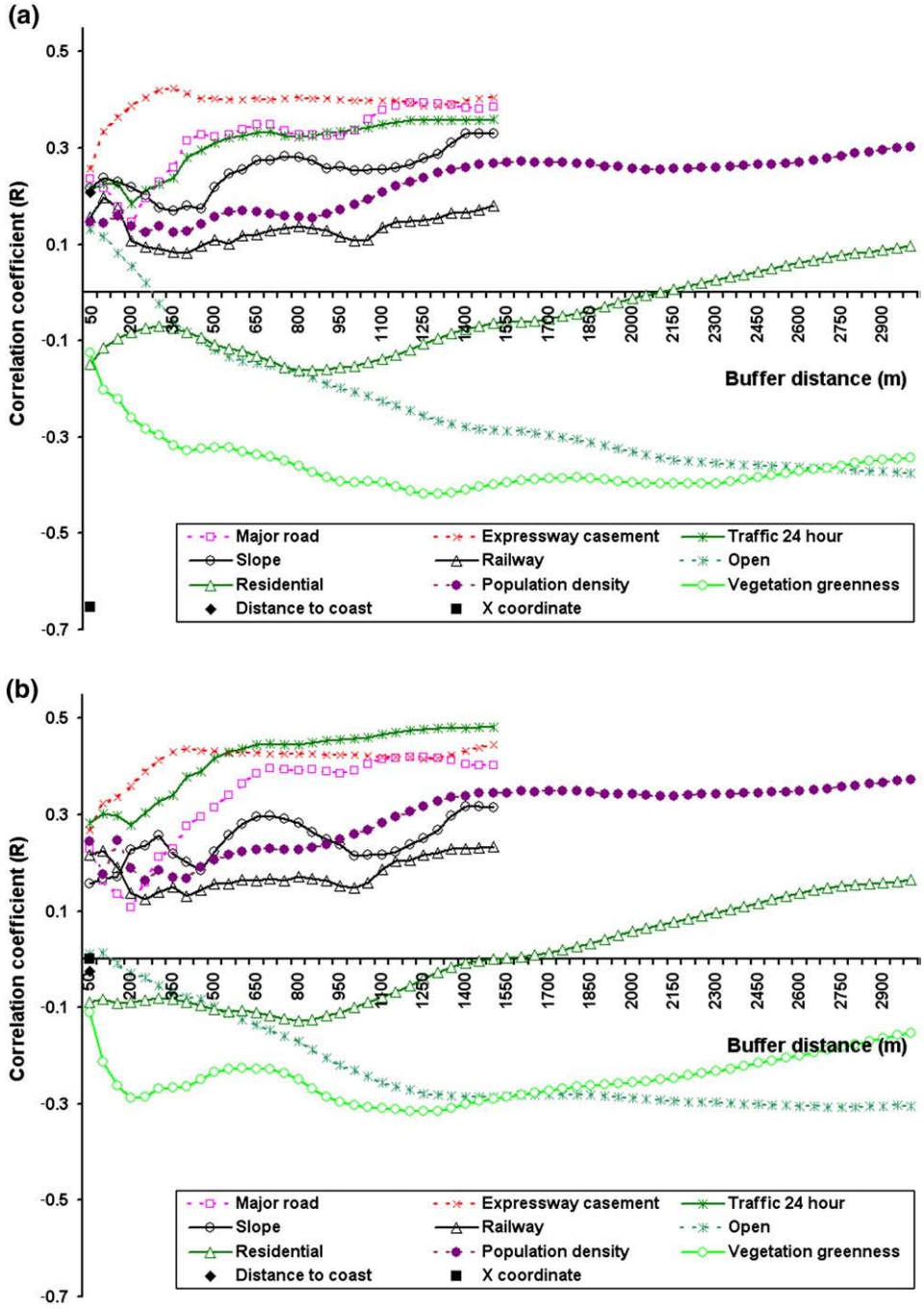

Fig. 2. Distance decay curves of correlation between spatial covariates and the measured (panel a) or predicted residuals of (panels b–f) NO_2 concentrations.

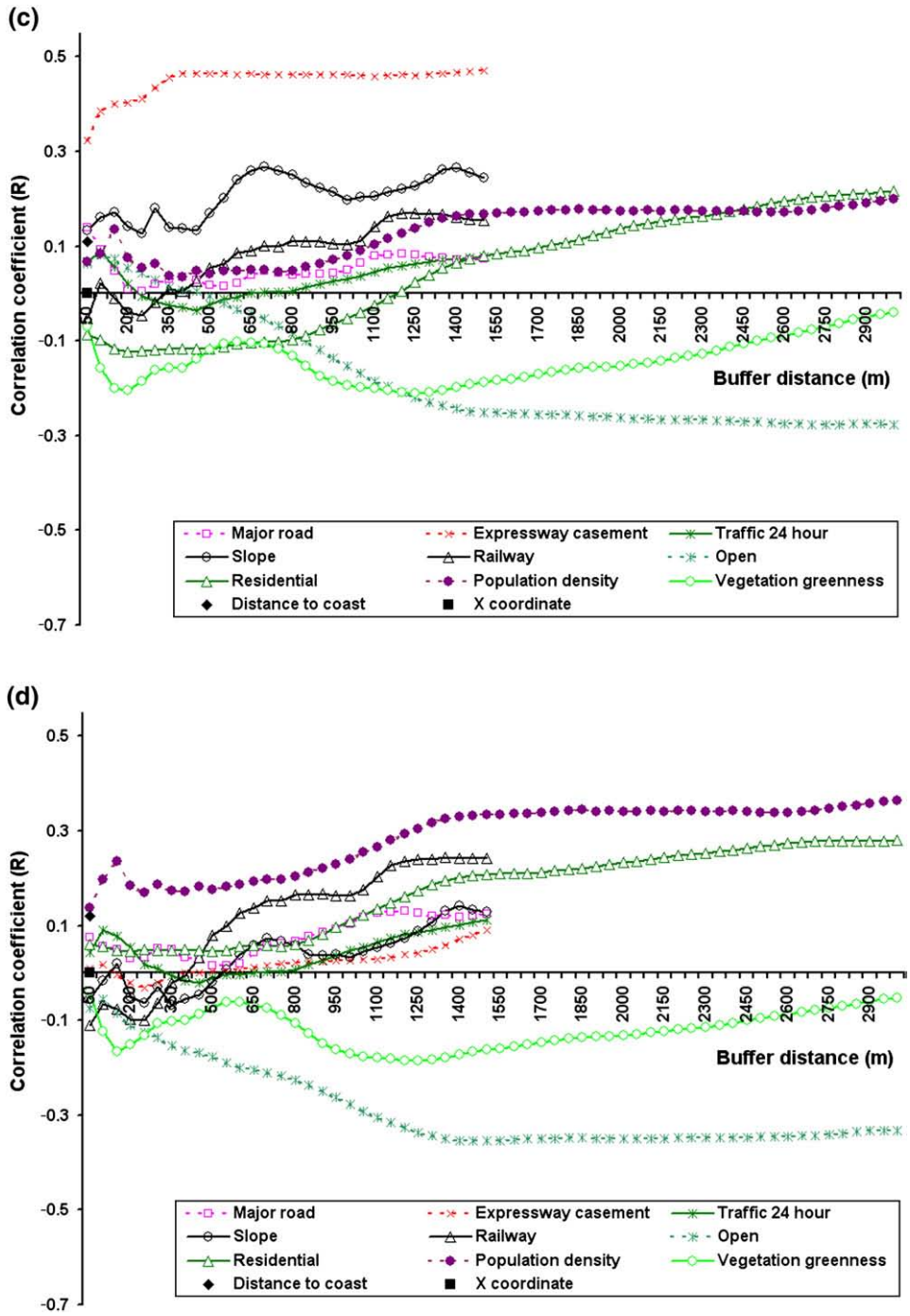

Fig. 2 (*continued*).

various land uses and transportation sources to aid the model selection.

3.2. ADDRESS implementation

The NO_2 measurements were transformed with the natural logarithm to reduce heteroskedasticity. Using the highest correlation coefficient from the distance decay curves of Fig. 2a, *X* coordinate was first used to build a bivariate model. The prediction residuals of the bivariate model were calculated and used to create a series of spatial covariate distance decay curves at a buffer distance of 50–3000 m. The distance decay curves of residuals in Fig. 2b show that 24 h traffic from major roads has the highest correlation compared to other variables and buffer distance 1500 m has the highest correlation on the traffic distance decay curve; however, because the 1500 m buffer is on a flattened curve, the highest slope change buffer distance before the plateau at buffer distance 650 m was chosen and added to the bivariate model to predict NO_2 levels. The prediction residuals from the above two variables were calculated and the distance decay curves of correlation with all the spatial covariates are illustrated in Fig. 2c. Fig. 2c shows that after incorporating both *X* coordinate and 24 h traffic into the prediction model, expressway casement has the highest correlation with the model residuals for an optimized buffer distance of 400 m. Similarly, because the flattened curve for buffer distance was greater than 400 m, the highest correlation at buffer distance 1500 m was

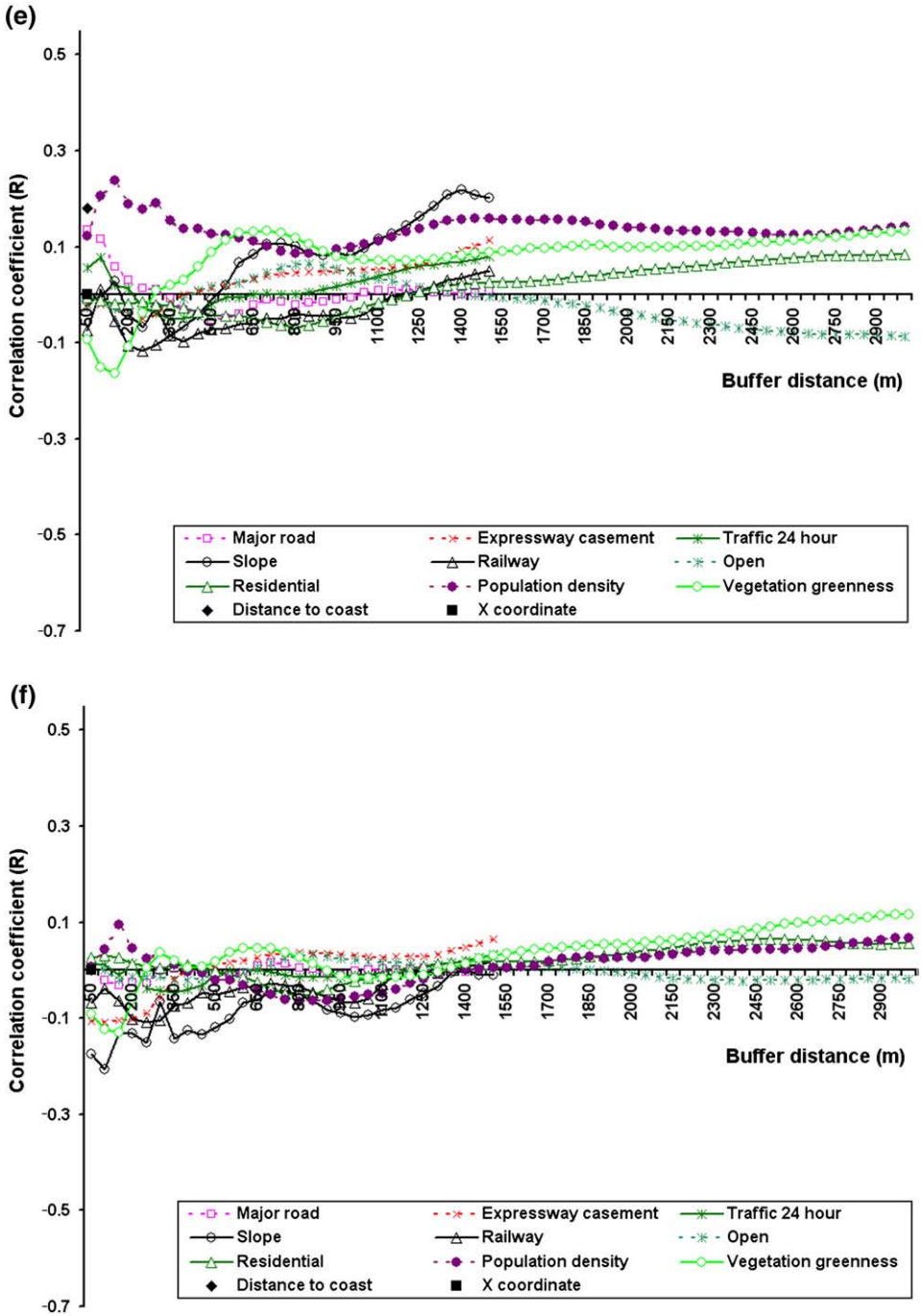

Fig. 2 (*continued*).

not selected. Instead, expressway casement at buffer distance 400 m was added to the prediction model. The *X* coordinate, 24 h traffic and expressway casement were found to explain 64.1% of the model variance. The distance decay curves of correlation for the remaining residuals (Fig. 2d) show that the influence of the open land use has the highest correlation at the optimized buffer distance of 1400 m. Even though residential land use was found to have similar correlations with open land use, residential land use was not chosen as a model predictor because the coefficients changed from negative (Fig. 2a) to positive (Fig. 2d). Residential land use has been linked with lower traffic volume and lower NO_2 levels in previous research (Jerrett et al., 2005; Henderson et al., 2007). The model with the addition of open land use explained 68.9% of the variance and all the four variables were significant ($p<0.05$). The prediction residual distance decay curves (Fig. 2e) from the above four predictors show that population density has the highest correlation at the optimized buffer distance 150 m. After adding population density at buffer distance 150 m into the model, the LUR model explained 70.4% of the variance. Similarly, major road at buffer distance 50 m, slope gradient of highway and major road at distance 1400 m, railway network at distance 1200 m, distance to shoreline and population density at distance 1350 m were added to the prediction model based on the corresponding highest correlations and optimized buffer distances with the prediction residuals. However, population density at 150 m was found not significant and was therefore removed from the model. ADDRESS created a LUR model with *X* coordinate (m), traffic 24 h (650 m), expressway casement (400 m), open land use (1400 m), railway (1200 m),

Table 2
Optimized land use prediction models from ADDRESS using the full buffer distances (A), the manual stepwise selection process (B) and from ADDRESS with limited buffer distances (C).

Modeling technique	Model variable	Unstandardized coefficients		*t*	Significance level	Collinearity statistics		R^2
		B	Std. error			Tolerance	VIF	
A	Intercept	12.212744	1.271	9.608	0.000			0.794
	X coordinate (m)	−0.000017	0.000	−8.433	0.000	0.767	1.304	
	Traffic 24 h (650 m)	0.000067	0.000	2.837	0.006	0.680	1.471	
	Expressway casement (400 m)	0.047685	0.008	6.339	0.000	0.809	1.235	
	Open land use (1400 m)	−0.000710	0.000	−3.075	0.003	0.684	1.461	
	Railway (1200 m)	0.000023	0.000	3.329	0.001	0.781	1.281	
	Major road (50 m)	0.001049	0.000	3.305	0.001	0.794	1.259	
	Slope (1400 m)	0.048814	0.013	3.685	0.000	0.764	1.310	
	Population density (1350 m)	0.002842	0.001	3.545	0.001	0.518	1.932	
	Distance to coast (m)	0.000018	0.000	4.091	0.000	0.572	1.749	
B	Intercept	14.295949	1.302	10.982	0.000			0.710
	X coordinate (m)	−0.000020	0.000	−9.702	0.000	0.956	1.046	
	Expressway casement (350 m)	0.061059	0.010	5.896	0.000	0.828	1.208	
	Population density (3000 m)	0.004671	0.001	3.944	0.000	0.656	1.525	
	Major road (1200 m)	0.000012	0.000	2.659	0.009	0.684	1.462	
	Slope (1400 m)	0.032969	0.014	2.311	0.023	0.888	1.126	
C	Intercept	12.227392	1.291	9.474	0.000			0.789
	X coordinate (m)	−0.000017	0.000	−8.300	0.000	0.761	1.314	
	Traffic 24 h (500 m)	0.000109	0.000	2.560	0.012	0.687	1.455	
	Expressway casement (500 m)	0.037914	0.006	6.630	0.000	0.799	1.251	
	Open land use (1500 m)	−0.000626	0.000	−3.019	0.003	0.648	1.542	
	Railway (1500 m)	0.000017	0.000	3.222	0.002	0.662	1.510	
	Major road (50 m)	0.001033	0.000	3.212	0.002	0.792	1.263	
	Slope (1500 m)	0.042576	0.014	3.030	0.003	0.733	1.365	
	Population density (1500 m)	0.002868	0.001	3.248	0.002	0.490	2.040	
	Distance to coast (m)	0.000017	0.000	3.928	0.000	0.557	1.794	

A: modeling process using ADDRESS with the buffer distances ranging from 0–50 m, 0–100 m, 0–150 m, and up to 0–1500 m for traffic related sources (30 circular buffers), and up to a maximum distance of 3000 m for land use, Tasseled-cap transformation and population density (60 circular buffers); B: a manual stepwise selection process with all the spatial covariates, C: a modeling process using ADDRESS but with limited buffer distances, including 50, 100, 200, 300, 500, 750, 1000, 1250 and 1500 m for traffic related variables and plus 1750, 2000, 2500 and 3000 m for other variables (e.g., population density and land use variables.

major road (50 m), distance to coast (m), population density (1350 m) and slope gradient (1400 m) as predictors and the model explained 79.4% of the variance. The distance decay curves derived from the remaining prediction residuals in Fig. 2f show that all the correlation coefficients are less than 0.15 for those variables with correct signs of correlation and no significant variables and

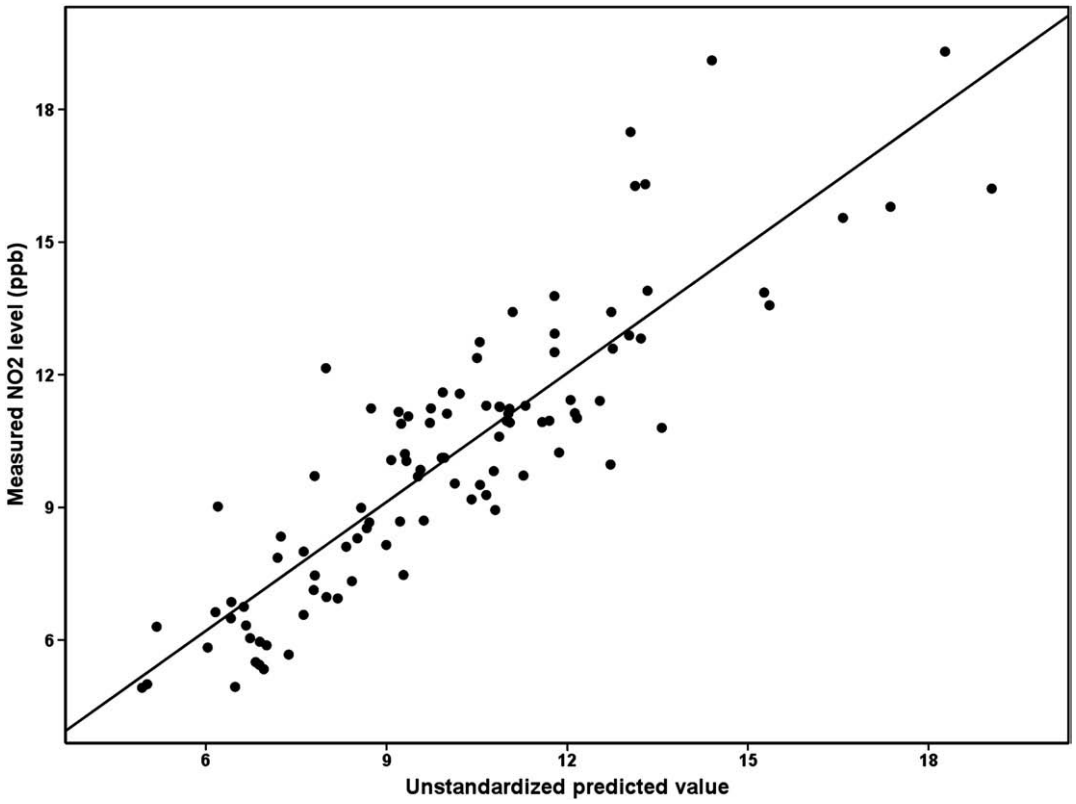

Fig. 3. Observed mean NO_2 on predicted value.

optimized buffer distances could be further added to the model. The final prediction model is listed in Table 2A and Fig. 3 (Fig. converted to natural unit: ppb).

Each individual variable has a significant t score and acceptable multicollinearity, as demonstrated by the average variance inflation factors (VIF) in Table 2A. All of the coefficients have the expected sign. Examination of Cook's distance (maximum = 0.103) and prediction model in Fig. 3 confirmed the absence of significant outliers. Additionally, Moran's I tests ($I = 0.063$ and $p = 0.13$) show that the spatial autocorrelation of residuals is insignificant for the model. This indicates that the model developed does not violate the independence assumption and it included fixed covariates to account for autocorrelation.

After applying the manual selection process to all the available 17 variables with corresponding optimized buffer distances and the four non-buffer variables, and after removing high collinearity and inconsistent coefficient variables, the LUR model from the manual stepwise selection process explains 71.0% of the model variance (Table 2B). This result is consistent with previous modeling results from the same area (Jerrett et al., 2007; Finkelstein and Jerrett, 2007). By contrast, ADDRESS has a higher prediction power ($R^2 = 0.794$) than the manual stepwise selection process. We see a reduction of the mean square error and root mean square error by 27.6% and 25.0%, respectively, from ADDRESS compared to the manual stepwise selection technique. When ADDRESS was applied to the same 17 buffer variables and the four non-buffer variables, but with nine traffic related buffers and 13 land use and other variable buffers, the model has a prediction power of 0.789 (R^2 in Table 2C). However, the VIF for population density is greater than 2.0. After removing population density variable, the prediction power (variance explained) is 76.4%. The slightly lower prediction power is because the limited buffer ADDRESS model tended to simplify the distance decay curves and possibly remove some peak values. However, the limited buffer ADDRESS model is seen as a better model than the manual stepwise regression model in prediction of NO_2 levels.

4. Summary and conclusions

We developed a systematic method for selecting variables in a LUR model by adding spatial covariates one at a time using distance decay selection strategy. Rather than seeking transfer of a LUR model from one urban area to another, ADDRESS applies a full spectrum of correlation coefficients and a series of distance decay curves to select spatial covariates at optimized buffer distances for the modeling process. Each time, a variable of the highest correlation with NO_2 levels or prediction residuals at an optimized buffer distance, identified using distance decay curves, was added to the prediction model. The selection process ensures the best optimized model to be chosen using available spatial variables.

Most of the manual stepwise selection process selected buffer distances of the highest correlations with pollutant levels at the first step and then used those distances to conduct an optimized stepwise regression process. Because of the collinearity, not all the variables at the initial optimized distances held in the final model, so we saw a decreased prediction of power ($R^2 = 0.710$) to our ADDRESS model ($R^2 = 0.794$). Jerrett et al. (2007) and Mavko et al. (2008) applied a manual forward screening selection strategy based on the highest t scores and correlation coefficients r from the regression residuals, respectively; however, those selection processes were more time consuming and only the highest correlation buffer distance was chosen each time. By contrast, ADDRESS selects the highest correlation variable at an optimized buffer distance by visualizing the distance decay curves of all the available variables. The optimized buffer distance of the chosen variable does not necessarily have the highest correlation on its distance decay curve and it might be selected based on the highest slope change if the highest correlation buffer distance is seen as a flattened continuation of the curve. Expert judgment is required to select the highest correlation variable and corresponding optimized buffer distance to identify the maximum distance of influence of a factor during each selection process. Expert judgment is also required to identify the directional change of the curves. If, for example, the greenness to be added to the model is positively correlated with prediction residuals, the greenness variable should not be added to the model even though it has, among the remaining spatial covariates, the highest correlation. The change of direction of correlation for a variable (Fig. 2) during the modeling process is mainly because of the collinearity between variables already chosen and the variable to be added.

By contrast, the ADDRESS model applied to the limited buffer distances produced very similar results to the full ADDRESS model. This was because the variables selected and distances used were quite similar to the full ADDRESS model. As most of the LUR models had less than 10 buffer distances, less than the 15 limited buffer distances used in this research, this makes a transparent identification of the maximum distance of influence of a factor very difficult. With limited buffer distances, peek distance of influence might be flattened and the overall distance decay pattern might be distorted. This is similar to the modified area unit problem: when an analysis changes from a fine scale to a coarse resolution, the prediction accuracy decreases for the analysis and error increases accordingly.

The approach developed here could be used as a guideline for conducting LUR analysis in a systematic and transparent way leading to improved pollution predictions for planning and epidemiological studies. Using this approach also has the benefit of increasing understanding of the inherent scale of correlation between measured pollutants and adjacent land use and traffic variables. Because of the differences in topography, meteorology, land use, traffic and population composition of one urban area compared to another, the optimized distance of the same influential factor (e.g., traffic) might differ from one urban area to another. To avoid subjective selection of buffer distances, the buffer distance decay curves could be used to identify such optimized distances of influence because each distance decay curve is an objective representation of spatial distance decay function of a source of pollution as long as corresponding measurements used to construct such a curve are scientifically sound. Such information ensures objective face validity of the selected variables and more importantly can be interpreted in epidemiological studies attempting to understand associations between fine-scale variations in pollution and health outcomes. The ADDRESS approach may therefore enhance understanding of relationships between pollutant levels and land use variables for planning purposes and improve the interpretation of epidemiologic results.

Acknowledgements

This project was funded by Health Canada, the Canadian Institutes of Health Research, and the California Air Resources Board.

References

Aguilera I, Sunyer J, Fernández-Patier R, Hoek G, Aguirre-Alfaro A, Meliefste K, et al. Estimation of outdoor NOx, NO_2, and BTEX exposure in a cohort of pregnant women using land use regression modeling. Environ Sci & Techno 2008;42:815–21.

Bailey TC, Gatrell AC. Interactive spatial data analysis. Harlow, Essex: Addison Wesley Longman; 1995.

Bell ML. The use of ambient air quality modeling to estimate individual and population exposure for human health research: a case study of ozone in the Northern GA Region of the United States. Environ Int 2006;32:586–93.

Beckerman B, Jerrett M, Brook JR, Verma DK, Arain MA, Finkelstein MM. Correlation of nitrogen dioxide with other traffic pollutants near a major expressway. Atmos Environ 2008;42:275–90.

Brauer M, Hoek G, van Vliet P, Meliefste K, Fischer P, Gehring U, et al. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. Epidemiology 2003;14:228–39.

Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, et al. Mapping urban air pollution using GIS: a regression-based approach. Int J Geographic Inf Sci 1997;11:699–718.
Brunekreef B, Holgate ST. Air pollution and health. Lancet 2002;360:1233–42.
Crist EP, Cicone RC. A physically-based transformation of Thematic Mapper data — the TM Tasseled Cap. IEEE Trans on Geosci and Rem Sens 1984;GE-22:256–63.
Finkelstein M, Jerrett M. A study of the relationships between Parkinson's disease and markers of traffic-derived and environmental manganese air pollution in two Canadian cities. Environ Research 2007;104:420–32.
Gilbert NL, Goldberg MS, Beckerman B, Brook JR, Jerrett M. Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land-use regression model. J Air & Waste Manage Assoc 2005;55:1059–63.
Henderson SB, Beckerman B, Jerrett M, Brauer M. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. Environ Sci Technol 2007;41:2422–8.
Hochadel M, Heinrich J, Gehring U, Morgenstern V, Kuhlbusch T, Link E, et al. Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. Atmos Environ 2006;40:542–53.
Hoek G, Meliefste K, Cyrys J, Lewné M, Bellander T, Brauer M, et al. Spatial variability of fine particle concentrations in three European areas. Atmos Environ 2002;36:4077–88.
Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, Briggs D. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 2008;42:7561–78.
Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. A review and evaluation of intraurban air pollution exposure models. J Exposure Anal Environ Epidemiol 2005;15:185–204.
Jerrett M, Arain MA, Kanaroglou P, Beckerman B, Crouse D, Gilbert NL, et al. Modelling the intra-urban variability of ambient traffic pollution in Toronto, Canada; J. Toxicol Environ Health 2007;70:200–12.
Kanaroglou PS, Jerrett M, Morrison J, Beckerman B, Arain MA, Gilbert NL, et al. Establishing an air pollution monitoring network for intra-urban population exposure assessment: a location-allocation approach. Atom Environ 2005;39:2399–409.
Liao D, Peuquet DJ, Duan Y, Whitsel EA, Dou J, Smith RL, et al. GIS approaches for the estimation of residential-level ambient PM concentrations. Environ Health Perspect 2006;114:1374–80.
Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, et al. Long-term exposure to air pollution and incidence of cardiovascular events in women. N Engl J Med 2007;356:447–58.
Mavko ME, Tang B, George LA. A sub-neighborhood scale land use regression model for predicting NO_2. Sci Tot Environ 2008;398:68–75.
Moore DK, Jerrett M, Mack WJ, Kunzli N. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. J Environ Monit 2007;9:246–52.
Morgenstern V, Zutavern A, Cyrys J, Brockow I, Gehring U, Koletzko S, et al. Respiratory health and individual estimated exposure to traffic-related air pollutants in a cohort of young children. Occup Environ Med 2007;64:8-16.
Poplawski K, Gould T, Setton E, Allen R, Su J, Larson T, et al. Intercity transferability of land use regression models for estimating ambient concentrations of nitrogen dioxide. J Exposure Sci Environ Epidemiol 2009;19:107–17.
Ross Z, English PB, Scalf R, Gunier R, Smorodinsky S, Wall S, et al. Nitrogen dioxide prediction in Southern CA using land use regression modeling: potential for environmental health analyses. J Exposure Sci Environ Epidemiol 2006;16:106–14.
Sahsuvaroglu T, Arain A, Kanaroglou P, Finkelstein N, Newbold B, Jerrett M, et al. A land use regression model for predicting ambient concentrations of nitrogen dioxide in Hamilton, Ontario, Canada. J Air Waste Manag Assoc 2006;56:1059–69.
Su JG, Larson T, Gould T, Cohen M, Buzzelli M. Transboundary air pollution and environmental justice: Vancouver and Seattle compared. GeoJournal. In press.