



Published in final edited form as:

Inhal Toxicol. 2007 ; 19(Suppl 1): 127–133.

A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure

Patrick H. Ryan and Grace K. LeMasters

Division of Epidemiology and Biostatistics, Department of Environmental Health, University of Cincinnati Medical Center, Cincinnati, Ohio, USA

Abstract

Epidemiologic studies of air pollution require accurate exposure assessments at unmonitored locations in order to minimize exposure misclassification. One approach gaining considerable interest is the land-use regression (LUR) model. Generally, the LUR model has been utilized to characterize air pollution exposure and health effects for individuals residing within urban areas. The objective of this article is to briefly summarize the history and application of LUR models to date outlining similarities and differences of the variables included in the model, model development, and model validation. There were 6 studies available for a total of 12 LUR models. Our findings indicated that among these studies, the four primary classes of variables used were road type, traffic count, elevation, and land cover. Of these four, traffic count was generally the most important. The model R^2 explaining the variability in the exposure estimates for these LUR models ranged from .54 to .81. The number of air sampling sites generating the exposure estimates, however, was not correlated with the model R^2 suggesting that the locations of the sampling sites may be of greater importance than the total number of sites. The primary conclusion of this study is that LUR models are an important tool for integrating traffic and geographic information to characterize variability in exposures.

Epidemiologic studies of the health effects of air pollution require accurate exposure assessments at unmonitored locations (e.g., subjects' place of residence) in order to minimize exposure misclassification. Methodologies employed to accomplish this task include, but are not limited to, spatial interpolation (e.g., kriging), proximity models, and dispersion modeling (e.g., California Line Source Dispersion model, CALINE) and these and others have been reviewed (Jerrett et al., 2005). Intraurban air pollution is characterized, however, by high spatial variability of pollutants with rapid decay from the source (Briggs et al., 1997, 2000). For example, nitrogen dioxide (NO₂) has been shown to have two- to threefold differences within 50 m or less (Hewitt, 1991), sulfur concentrations have been demonstrated to decrease by one-half between 50 and 150 m from a highway (Reponen et al., 2003), and ultrafine particles have been shown to be elevated above background concentrations to approximately 300 m from highways (Zhu et al., 2002). These small-scale variations in pollutant concentrations are not identifiable using most interpolation techniques based upon monitoring density and spatial distribution of traffic sources (Brauer et al., 2003). Furthermore, proximity models have a likelihood of exposure misclassification due to the assumption of isotropic dispersion and the use of a categorical exposure designation (e.g., residence <100 m = exposed, residence >100 m = unexposed) (Jerrett et al., 2005). In order to address these limitations, land-use regression (LUR) models have been developed and utilized to model traffic pollutants including NO₂ and PM_{2.5} (Briggs et al., 1997, 2000; Brauer et al., 2003; Ross et al., 2006; Gilbert et al., 2005).

Land-use regression utilizes the monitored levels of the pollutant of interest as the dependent variable and variables such as traffic, topography, and other geographic variables as the independent variables in a multivariate regression model (Gilliland et al., 2005). Levels of pollution may then be predicted for any location, such as individual homes, using the parameter estimates derived from the regression model. The incorporation of site-specific variables into this method detects small area variations more effectively than other methods of interpolation (Briggs et al., 1997; Gilliland et al., 2005). The objective of this article is to briefly summarize the history and application of LUR models through June 2006 by outlining the similarities and differences of the variables included in the model, model development, and model validation.

METHODS

The inclusion criteria for this review included publication in a peer-reviewed journal available in the PubMed database through June 2006 in addition to the authors' recently developed model (Ryan et al., 2007). Those papers with a description of the development of the LUR model have been included. Variables included in the LUR model are described and the various methods for validating the models are presented.

A literature search was conducted using the keywords "air pollution," "exposure," "land use," and "regression." With this search three articles were identified. From these an additional two were identified based on articles cited within the references. Although many of the reviewed models have been subsequently applied to investigate associated health effects, only articles describing the LUR model development have been included in this review. Finally, recommendations are provided based upon these analyses.

RESULTS AND DISCUSSION

There were six publications identified which described the development and validation of a LUR model (Brauer et al., 2003; Briggs et al., 1997, 2000; Gilbert et al., 2005; Ross et al., 2006; Ryan et al., 2007). Three of these (Brauer et al., 2003; Briggs et al., 1997, 2000) described the development of LUR models for more than one location. Thus, in total 12 LUR models have been evaluated.

All models use the measured levels of a pollutant at sampling locations as the dependent variable in a multiple least-squares regression model. The geographic variables of interest within a radius of defined distance(s) are then utilized as independent variables in this regression model. These radii, known as "buffers," are applied in each LUR model, though the length varies among and within each.

Although the independent variables in each least-squares regression model were unique, they may be broadly categorized into four major classes: (1) road type, (2) traffic count, (3) elevation, and (4) land cover. Each LUR model contained unique definitions of one or more of these class variables depending upon the available data and pollutant of interest (Table 1). For example, all models included some form of traffic count or intensity measure while only three contained elevation in their final LUR model. The LUR model derived by Briggs et al. for Prague, Czech Republic, contained traffic volume data while the LUR model for Amsterdam, the Netherlands, depended upon the length of defined road types because no traffic volume data was available (Briggs et al., 1997). Similarly, the distance to the nearest major road was included as an independent variable in three of the LUR models, though "major roads" was defined differently in each (e.g., truck count greater than 1000 or traffic count greater than 50,000, Table 1).

The variables included in the final LUR models for each study are summarized in Table 2. Two of the studies (Briggs et al., 1997,2000) were conducted as part of the Small Area Variations

in Air Quality and Health (SAVIAH) project. The first of these resulted in LUR models for Huddersfield, UK, Amsterdam, the Netherlands, and Prague, Czech Republic, with coefficients of determination (R^2) equal to .61, .62, and .72, respectively. These models included a weighted traffic volume variable, land cover, and altitude as independent predictor variables (Table 2). The second of these studies describes a revised LUR model which includes these three variables for four cities in the United Kingdom (Huddersfield, Hammersmith and Ealing, Northampton, and Sheffield) (Briggs et al., 2000).

The outcome variables in LUR models developed by Brauer et al. in Stockholm County, Sweden, Munich, Germany, and the Netherlands were sampled measurements of $PM_{2.5}$ and $PM_{2.5}$ filter absorbance (a marker of diesel exhaust particulates) (Brauer et al., 2003). In general, these LUR models included as predictor variables a measure of traffic (on the nearest road or the number of roads within a given buffer) and a measure of housing or population density (Table 2). LUR models for the $PM_{2.5}$ filter absorbance were found to have greater R^2 values than those for sampled $PM_{2.5}$ levels (Table 2).

More recent studies have developed LUR models for NO_2 in Montreal, Canada, and San Diego County, California (Gilbert et al., 2005; Ross et al., 2006). The final regression model for Montreal included the distance to the nearest highway, the traffic count on the nearest highway, the length of highways within 100 m of the sampling site, the length of minor roads within 500 m of the sampling site, the area of open space less than 100 m from the sampling site, and the population density within 2000 m of the sampled site (Gilbert et al., 2005). The LUR model explained 54% of the variability seen in the sampled locations. The final LUR model developed by Ross et al. included the length of roads within 40 m of the sampling site, the traffic volume 40 to 1000 m of the sampling site, and the distance to the Pacific coast and had a larger R^2 of 0.77 (Ross et al., 2006).

Finally, the authors developed a LUR model for sampled levels of ECAT (Elemental Carbon Attributable to Traffic, a marker of diesel exhaust) in Cincinnati, OH. Significant predictor variables in this model included elevation, which was the strongest factor, truck counts within 400 m of the sampled site, and the length of bus routes within 100 m of the sampled site. This model explained 75% of the sampled ECAT variability (Ryan et al., 2007). Interestingly, in these studies the number of sampled sites did not correlate with the final model R^2 as there were only 24 sampling sites in Cincinnati, OH, with an R^2 of .75, while studies with 80 sampling sites had an R^2 of .61–.72.

The validity of each LUR model was assessed by either comparing the predicted levels of the modeled pollutant to sampled levels at sites not included in the prediction model (Briggs et al., 1997, 2000; Gilbert et al., 2005; Ross et al., 2006), mean prediction errors (Brauer et al., 2003), and/or a bootstrap to evaluate the sensitivity of the model parameters (Ross et al., 2006, Ryan et al., 2007). (Table 2). Briggs et al. included 8–10 reference sites for the validation of the LUR models for Huddersfield, Prague, and Amsterdam. Comparison of the measured levels of NO_2 at these sites and the levels predicted by the LUR models resulted in R^2 between .79 and .87 indicating good predictions of NO_2 levels at the reference sites (Briggs et al., 1997). Likewise, in the revised model proposed by Briggs et al., R^2 values ranging from .58 to .76 were computed between predicted and measured values at validation sites (Briggs et al., 2000). Gilbert et al. compared the predicted levels of NO_2 to the measured levels at seven sampling sites and found an R^2 of .52 (Gilbert et al., 2005). Ross et al. applied their final model to 12 validation sites and found all the predicted levels to be within 1.5 times the range of the observed levels with an average estimated error of 2.1 ppb (Ross et al., 2006).

Mean prediction errors (calculated as the square root of the sum of the squared differences of the observed concentration at site i and the predicted concentration at site i from a model

developed without site *i*) were calculated to test the validity of the LUR models developed by Brauer et al. The root mean squared errors (RMSE) ranged from 1.1 to 1.6 $\mu\text{g}/\text{m}^3$ for $\text{PM}_{2.5}$ and from 0.22 to 0.31×10^{-5} for $\text{PM}_{2.5}$ filter absorbance (Brauer et al., 2003). A bootstrap analysis was undertaken to evaluate the sensitivity of the estimated model parameters in two models (Ross et al., 2006; Ryan et al., 2007). In each of these studies, five randomly selected samples were removed before running the final LUR model. Histograms of the observed parameter estimates derived for each of 1000–20,000 iterations suggested in both studies that the model parameters were robust.

Land-use regression models offer the advantage of accounting for small scale variability in intraurban pollutant concentrations. The method also accounts for sampled levels of the pollutant of interest. Furthermore, the methodology is transferable, though care must be taken to include the appropriate road, traffic, topographic, and land cover variables for the pollutant and urban area. For example, we found elevation to be the most significant predictor variable in the LUR model for Cincinnati, though elevation was not included in the final model developed for San Diego County. This finding may be related to differing topography of the two study locations. Likewise, the distance from the Pacific Ocean was included in the San Diego model, though this is obviously unnecessary for noncoastal cities. Furthermore, care must be taken to appropriately select the independent variables and buffer radii for the pollutant of interest. Generally, the choice of buffer radii is based upon the decay of the modeled pollutant, e.g., wider for NO_2 and narrower for estimates of diesel, as well as the density of the geographic variables surrounding the sampling location. For example, ultrafine particles have been found to be elevated above background concentrations to 300 m from highways (Zhu et al., 2002). Therefore, buffers within this distance (or slightly greater) are appropriate when modeling ultrafine particles. The size of the buffers chosen is critical as it will determine the amount of surrounding traffic, length of roads and land cover which may explain the variability of the sampled pollutant.

The traffic variables to be included in the model are generally chosen based upon the modeled pollutant and the available data. For example, the inclusion of average daily truck counts (rather than passenger vehicles) and bus routes in our LUR model (Ryan et al., 2007) was appropriate as the pollutant of interest, ECAT, is a marker of diesel combustion. The built environment surrounding the sampling locations has been measured through the use of population and housing density (Brauer et al., 2003; Gilbert et al., 2005; Ross et al., 2006) and aerial photographs and remote sensing data (Briggs et al., 2000; Ryan et al., 2007) The accuracy of these data, therefore, will affect the overall accuracy of the model.

Another potential limitation of LUR models is the amount and quality of the land use and traffic data available. Traffic counts are often dependent on local government agencies and may be collected prior to the sampling period. The most ideal traffic counts would be collected simultaneously at the time of pollutant sampling. Other land-use data including population density, housing density, and remote sensing data may also have been collected prior to the sampling period. Additionally, the LUR model will be dependent upon the quality of the air pollutant sampled. For example, Ryan et al. have used ECAT as a marker of diesel combustion. Thus, the reliability of the LUR model to classify exposure will be dependent upon the specificity of ECAT as a marker of diesel exhaust (Ryan et al., 2007). Others have modeled NO_2 which may not be representative of other air pollutants (Jerrett et al., 2005).

A measure of traffic volume was the most significant independent variable in five of the six models which reported variable significance. Brauer et al. report that the total traffic on the nearest road, the traffic intensity 50–250 m from the sampling site, and the number of high-traffic roads less than 250 m from the sampling sites were the most significant independent variables for the models developed for Stockholm County, Munich, and the Netherlands,

respectively (Table 2). Likewise, the traffic count on the nearest highway and the traffic volume 40–300 m from the sampling sites were the most significant variables in the LUR models derived for Montreal and San Diego County, respectively. In the Cincinnati model, although truck count was a significant independent variable, elevation was the most significant.

Another apparent issue is that the number of sampling locations necessary to provide adequate information for a LUR model may be of less importance than the variability of exposure at various sites. There was, in fact, a weak inverse correlation between the number of measurement sampling sites available to develop the LUR and the variability explained by the model. For example, the number of monitoring sites available to develop the LUR with model R^2 of .79 in San Diego County consisted of 39 sampling sites, while in Montreal there were 67 sites with a model R^2 of .54. These results suggest that the suitability of a monitoring network may depend more on the variability of the land characteristics captured by the network and less upon the total number of sampling sites.

In comparison with proximity models of exposure, LUR models require similar geographic variables (traffic volume, distance to pollutant source) and software, but necessitate sampling data. The added benefit, however, is the ability to differentiate exposure within proximity distances through the use of additional land-use variables. Geostatistical models (e.g., kriging) are similar to LUR models with respect to the need for sampling data. The assumption of spatial autocorrelation between monitoring sites necessary for kriging, however, may not be valid in an urban environment and for some pollutants. The LUR model, in contrast, assumes independence between sampled locations. A complete review and comparison of the various types of intraurban air pollution models is available (Jerrett et al., 2005).

In summary, we have reviewed the land-use regression models developed to date. The integration of road networks, traffic count information, topography, and land cover data within a geographic information system is common throughout all of these models. Although the methodology is suitable for intraurban exposure assessment, the adaptability depends upon the available and relevant local data. Though there is likely to be a minimum number of sampling sites needed to evaluate intraurban exposures, ultimately the precision of the LUR model is dependent upon the location of these sites to optimally characterize the sources and pattern of exposure. Before beginning a study, a pilot study may be warranted and more attention given to the placement of sampling sites in relationship to sources of air pollution. Additionally, the buffer radii and predictor variables should be carefully chosen based upon the sampled pollutant and available data.

Acknowledgements

This study was supported by NIEHS grants ES11170 and ES10957.

References

- Hewitt CN. Spatial variations in nitrogen dioxide concentrations in an urban area. *Atmos Environ* 1991;25B:429–434.
- Brauer M, Hoek G, van Vliet P, Meliefste K, Fischer P, Gehring U, Heinrich J, Cyrus J, Bellander T, Lewne M, Brunekreef B. Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems. *Epidemiology* 2003;14:228–239. [PubMed: 12606891]
- Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, Pryn K, Van Reeuwijk H, Smallbone K, Van der Veen A. Mapping urban air pollution using GIS: a regression-based approach. *Int J Geogr Inform Sci* 1997;11:699–718.
- Briggs DJ, de Hough C, Gulliver J, Wills J, Elliott P, Kingham S, Smallbone K. A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. *Sci Total Environ* 2000;253:151–167. [PubMed: 10843339]

- 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;55:1059–1063.
- Gilliland F, Avol E, Kinney P, Jerrett M, Dvonch T, Lurmann F, Buckley T, Breyse P, Keeler G, McConnell R. Air pollution exposure assessment for epidemiologic studies of pregnant women and children: Lessons learned from the Centers for Children’s Environmental Health and Disease Prevention Research. *Environ Health Perspect* 2005;113:1447–1454. [PubMed: 16203261]
- Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahuvaroglu T, Morrison J, Giovis C. A review and evaluation of intraurban air pollution exposure models. *J Expos Anal Environ Epidemiol* 2005;15:185–204.
- Reponen T, Grinshpun SA, Trakumas S, Martuzevicius D, Wang ZM, LeMasters G, Lockey JE, Biswas P. Concentration gradient patterns of aerosol particles near interstate highways in the Greater Cincinnati airshed. *J Environ Monit* 2003;5:557–562. [PubMed: 12948227]
- Ross Z, English PB, Scalf R, Gunier R, Smorodinsky S, Wall S, Jerrett M. Nitrogen dioxide prediction in Southern California using land use regression modeling: Potential for environmental health analyses. *J Expos Sci Environ Epidemiol* 2006;16:106–114.
- Ryan PH, LeMasters GK, Biswas P, Levin L, Hu S, Lindsey M, Bernstein DI, Lockey J, Villareal M, Hershey GKK, Grinshpun SA. A comparison of proximity and land use regression traffic exposure models and wheezing in infants. *Environ Health Perspect* 2007;115:278–284. [PubMed: 17384778]
- Zhu Y, Hinds WC, Kim S, Sioutas C. Concentration and size distribution of ultrafine particles near a major highway. *J Air Waste Manage Assoc* 2002;52:1032–1042.

TABLE 1

Classes and definitions of common geographic variables included in land use regression models

Class	Variable used	Variable definition	Study	
Road type	Road type 1	Road serving > 25000 people	Briggs et al. (1997)	
	Road type 2	Road serving 5000—25000 people	Briggs et al. (1997)	
	Road type 3	Road serving 1000—5000 people	Briggs et al. (1997)	
	Highway	Undefined	Gilbert et al. (2005)	
	Major road	Undefined	Gilbert et al. (2005)	
	Major road	Average daily traffic count > 50,000	Ross et al. (2006)	
	Major road	Average daily truck count > 1000	Ryan et al. (In press)	
	High traffic road	Road serving > 25000 people	Brauer et al. (2003)	
	Medium traffic road	Road serving 10000—25000 people	Brauer et al. (2003)	
	Minor road	Undefined	Gilbert et al. (2005)	
	Bus route	Public transportation route	Ryan et al. (2007)	
	Traffic count	Weighted traffic volume	$15 * (\text{Traffic volume} < 40 \text{ m}) + (\text{Traffic volume} 40\text{--}300 \text{ m})$	Briggs et al. (1997, 2000)
		Traffic volume	Traffic volume (1000 vehicle km hr ⁻¹)	Briggs et al. (1997)
Traffic count on nearest highway		Undefined	Gilbert et al. (2005)	
Average daily traffic count		Average number of cars traveling in both directions/weekday (vehicle-km/hr)	Ross et al. (2006)	
Traffic intensity		Vehicles/day	Brauer et al. (2003)	
Heavy vehicle traffic intensity		Heavy traffic/day	Brauer et al. (2003)	
Average daily truck count		Average number of trucks traveling in both directions	Ryan et al. (2007)	
Elevation	Altitude	Meters above sea level	Briggs et al. (1997, 2000)	
	Elevation	Meters above sea level	Ross et al. (2006), Ryan et al. (2007)	
Land cover	Land cover factor	Weighted sum of the areas of industrial and high density residential land	Briggs et al. (1997, 2000)	
	Land cover	Area of built up land	Briggs et al. (1997)	
	Industrial use land	Area of land designated for industrial use	Gilbert et al. (2005)	
	Open space land	Area of land designated as open space	Gilbert et al. (2005)	
	Commercial use land	Area of land designated for commercial use	Gilbert et al. (2005)	
	Government/industry land	Area of land designated for government or industrial use	Gilbert et al. (2005)	
	Household density	Number of houses in area	Gilbert et al. (2005), Ross et al. (2005), Brauer et al. (2003)	
	Population density	Population in area	Ross et al. (2006), Brauer et al. (2003)	
	Land use	Area covered by industry, heavy industry, multi-family residential housing	Ross et al. (2006), Ryan et al. (2007)	
	Distance to coast	Distance to Pacific Ocean	Ross et al. (2006)	

TABLE 2

Locations, pollutant, final regression variables, and model R² of previous land use regression models

Paper	Study location(s)	Air pollutant	Sampling sites	Variables (and buffer size) included in final regression model	Model R ²	Model Validation
Briggs et al. (1997)	Huddersfield, U.K.	NO ₂	80	1) Weighted traffic volume 2) Land cover factor <300 m 3) Altitude 4) Sampler height	0.61	Comparison with reference sites n = 8 R ² = 0.82
	Amsterdam, Netherlands	NO ₂	80	1) Road type 1 <50 m 2) Road type 2 <50 m 3) Road type 1 50–200 m 4) Road type 2 50–200 m 5) Road type 3 50–200 m 6) Road type 3 <50 m 7) Land cover <100 m 8) Distance to road type 1	0.62	n = 10 R ² = 0.79
	Prague, Czech Republic	NO ₂	80	1) Traffic volume <60 m 2) Traffic volume 60–120 m 3) Land cover factor <60 m 4) Altitude	0.72	n = 10 R ² = 0.87
Briggs et al. (2000)	Huddersfield, U.K.	NO ₂	20	1) Weighted traffic volume 2) Land cover factor <300 m 3) Altitude	0.76	Comparison with reference sites n = 10 R ² = 0.76
	Hammersmith and Ealing, U.K.	NO ₂	11	1) Weighted traffic volume 2) Land cover factor <300 m 3) Altitude	*	****
	Northampton, U.K.	NO ₂	35	1) Weighted traffic volume 2) Land cover factor <300 m 3) Altitude	0.58	n = 28 R ² = 0.58
	Sheffield, U.K.	NO ₂	28	1) Weighted traffic volume 2) Land cover factor <300 m 3) Altitude	0.73	n = 18 R ² = 0.73
	Stockholm county, Sweden	PM _{2.5}	42	1) Total traffic on nearest road 2) Population density 1000–5000 m	0.50	Cross-validation: RMSE 1.10 µg/m ³
Brauer et al. (2003)	Stockholm county, Sweden	PM _{2.5} filter absorbance	42	1) Total traffic on nearest road 2) Population density 1000–5000 m	0.66	0.22 × 10 ⁻⁵ m ⁻¹
	Munich, Germany	PM _{2.5}	40	1) Traffic intensity <50 m 2) Traffic intensity 50–250 m 3) Population density <300 m	0.56	1.35 µg/m ³
	Munich, Germany	PM _{2.5} filter absorbance	40	1) Housing density <300 m 2) Traffic intensity <50 m 2) Traffic intensity 50–250 m 3) Population density <300 m	0.67	0.31 × 10 ⁻⁵ m ⁻¹
	Netherlands	PM _{2.5} filter absorbance	41	1) Number of high traffic roads <250 m 2) Housing density <300 m	0.73	1.59 µg/m ³
Gilbert et al. (2005)	Netherlands	PM _{2.5} filter absorbance	41	1) Number of high traffic roads <250 m 2) Housing density <300 m	0.81	0.31 × 10 ⁻⁵ m ⁻¹
	Montreal, Canada	NO ₂	67	1) Distance from nearest highway 2) Traffic count on nearest highway 3) Length of highway <100 m 4) Length of major road <100 m 5) Length of minor road <500 m 6) Area of open space <100 m 7) Population density <2000 m	0.54	Comparison with reference sites n = 8 R ² = 0.52

Paper	Study location(s)	Air pollutant	Sampling sites	Variables (and buffer size) included in final regression model	Model R ²	Model Validation
Ross et al. (2006)	San Diego County, U.S.	NO ₂	39	1) Length of road <40 m 2) Traffic volume 40–300 m 3) Traffic volume 300–1000 m 4) Distance to coast	0.77	n = 12 Predicted, on average, <2.1 Ppb, Bootstrap
Ryan et al. (2007)	Cincinnati, U.S.	ECAT**	24	1) Elevation 2) Truck count <400 m 3) Length of bus routes <100 m	0.75	Bootstrap

* Refer to Table 1 for variable definitions.

** Derived from PM_{2.5} concentration, a marker of diesel exhaust particulates.

*** Root mean squared error.

**** No reference sites available.

Boldface indicates most significant independent variable in final LUR model.