Evaluation of Land Use Regression Models Used to Predict Air Quality Concentrations in an Urban Area

Markey Johnson ¹, V. Isakov ², J.S. Touma ², S. Mukerjee ², H. Özkanak ²

¹ Air Health Science Division, Water Air and Climate Change Bureau, Health Canada, 269 Laurier Ave West, Ottawa, Ontario, Canada K1M 2B7
² U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC 27711

Corresponding author: Vlad Isakov, PhD
Address: MD - E243-02, Atmospheric Modeling and Analysis Division, National Exposure Research Laboratory, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711. Telephone: 919-541-2494, Fax: 919-541-1379, Email: isakov.vlad@epa.gov
Abstract

Cohort studies designed to estimate human health effects of exposures to urban pollutants require accurate determination of ambient concentrations in order to minimize exposure misclassification errors. However, it is often difficult to collect concentration information at each study subject location. In the absence of complete subject-specific measurements, land use regression (LUR) models have frequently been used for estimating individual levels of exposures to ambient air pollution. The LUR models, however, have several limitations mainly dealing with extensive monitoring data needs and challenges involved in their broader applicability to other locations. In contrast, air quality models can provide high-resolution source-concentration linkages for multiple pollutants, but require detailed emissions and meteorological information.

In this study, first we predicted air quality concentrations of PM$_{2.5}$, NO$_x$, and benzene in New Haven, CT using hybrid modeling techniques based on CMAQ and AERMOD model results. Next, we used these values as pseudo-observations to develop and evaluate the different LUR models built using alternative numbers of (training) sites (ranging from 25 to 285 locations out of the total 318 receptors). We then evaluated the fitted LUR models using various approaches, including: 1) internal “Leave-One-Out-Cross-Validation” (LOOCV) procedure within the “training” sites selected; and 2) “Hold-Out” evaluation procedure, where we set aside 33-293 tests sites as independent data sets for external model evaluation. LUR models appeared to perform well in the training data sets. However, when these LUR models were tested against independent hold out (test) data sets, their performance diminished considerably. Our results confirm the challenges facing the LUR community in attempting to fit empirical response surfaces to spatially- and temporally-varying pollution levels using LUR techniques that are site dependent. These results also illustrate the potential benefits of enhancing basic LUR models by utilizing air quality modeling tools or concepts in order to improve their reliability or transferability.

Keywords

Air pollution, exposure assessment, intra urban scale, dispersion models, health effects assessment, land-use regression models.
There is a growing body of literature linking proximity to roadways and traffic intensity with adverse respiratory health effects (Brunekreef and Holgate 2002; Gauderman et al. 2005; McConnell et al. 2006; Samet 2007; Samal et al. 2008; Health Effects Institute, 2010). Pollutants of interest include nitrogen dioxide (NO$_2$), fine particular matter (PM$_{2.5}$), and elemental carbon (EC) since each has been linked to vehicular emissions and respiratory health. Meta analysis of pollution gradients around roadways shows variations in concentration levels depending on meteorological conditions, whether the pollutants are reactive or inert, as well as for low and high background concentration levels, and roadway traffic volume. In general, a reasonable distance to consider as the near source zone for traffic impact is about 500m (Zhu et al 2002; Harrison et al. 2003; Kim et al. 2004, Baldauf et al. 2009). In most urban areas, such a distance includes a large population residing near roadways. Thus, there is a need to develop accurate exposure models to assess the health impacts to those highly exposed populations.

Urban air pollution levels exhibit high spatial and temporal variability (Weijers et al., 2004, Westerdahl et al., 2005, Isakov et al., 2007) and to accurately estimate the subtle effects of exposure to these levels, a large number of monitoring sites are required to determine concentrations as accurately as possible and to minimize exposure misclassification. Fixed urban background air quality measurement stations in routine monitoring networks are widely used to represent the exposure of subjects residing within large geographic areas surrounding these monitoring stations. Geographic Information Systems (GIS) based spatial interpolation approaches and statistical techniques have been used to interpolate from ambient monitoring data. However, ambient monitoring data are usually limited in the number of monitoring sites as well as the species of pollutants measured. Thus, these methods do not adequately capture the smaller- scale spatial and temporal variations in pollutant concentrations such as those occurring near roadways. In a few studies, direct personal monitoring at individual residences, augmented with a few additional fixed monitoring stations, is used to provide the needed spatial and temporal resolution. Since this approach is more expensive, measurement campaigns are usually carried out at limited time intervals, e.g., during two-week periods in selected seasons.
Due to the increasing ability of GIS to provide land-use data, a widely used approach in community health studies for capturing the smaller-scale variability has been land use regression (LUR) models. These statistical models regress measured values of a pollutant at sampling locations (the dependent variable) against geographic variables, such as traffic intensity, proximity to roadways, ports, harbors, and industrial sources; and population and housing density (independent variables). Air pollution levels are then predicted for any location, such as individual homes, using the parameter estimates derived from the regression model (Brauer et al., 2003; Ross et al., 2006; Jerrett et al., 2007; Ryan and LeMasters, 2007; Hoek et. al., 2008; Vienneau et. al., 2010). In a recently conducted critical review on exposure and health effects, the Health Effects Institute panel recommended using LUR modeling as a more accurate method for exposure analysis (Health Effects Institute, 2010).

Despite their advantages of being easy to apply, simple, practical, and widely used by researchers, LUR models have several limitations. Among these are that they require accurate monitoring data and at a large number of sites, especially in highly-industrialized urban areas with many types of emission sources. In these areas, monitoring data collection is expensive and time consuming. Other limitations are: 1) these models are typically not transferable from one urban area to the another; 2) they lack the ability to connect specific sources of emissions to concentrations for developing pollution mitigation strategies; and 3) they are not typically designed to address multi-pollutant aspects of air pollution (e.g., they usually deal with one pollutant at a time). LUR methodology, however, is relatively new and many of these issues are currently being examined or addressed.

In contrast, air quality models have been used for many years in air quality management but only more recently applied in exposure assessments to provide improved spatial exposure estimates (Jerrett et al., 2005; Marshall et al., 2008). Recent developments include the hybrid air quality modeling that combines dispersion models to provide local-scale concentrations with photochemical grid models to provide regional background concentrations. Isakov et al. (2009) provided an example of coupling the AERMOD local scale model with the regional CMAQ photochemical grid model results to provide hybrid modeled concentration estimates at the census block group level with time scales ranging from hourly to daily averages.
The current study has the following four main aims:

1) Develop conventional LUR models for New Haven, CT using locally collected demographic, housing, land-use, proximity to traffic and other main source information, along with modeled air quality concentrations for benzene, NO\textsubscript{x} and PM\textsubscript{2.5} (e.g., as pseudo-monitoring data) at numerous receptors (318 total), instead of average pollutant concentration measurements that are typically collected at limited number of monitoring locations.

2) Internally evaluate the performance of fitted LUR models (ascertained by adjusted model R\textsuperscript{2} and other measures) as a function of the number and location of receptor (e.g., pseudo-monitoring) sites selected, through an iterative stratified random sampling procedure.

3)Externally evaluate the predictive power of the fitted LUR models by comparing the LUR model predictions to concentrations at receptors that were not used in fitting the LUR models.

4) Examine and recommend strategies for optimum site selection in order to improve the predictive power and transferability of LUR models.

2. Methods

The air quality and LUR modeling analysis was done in two stages. First, we predicted air quality concentrations of PM\textsubscript{2.5}, NO\textsubscript{x}, and benzene using hybrid modeling techniques based on CMAQ and AERMOD model results. PM\textsubscript{2.5} was selected as a criteria pollutant that is mostly regional and is largely formed in the atmosphere due to secondary reactions. NO\textsubscript{x} was selected because it is strongly influenced by local combustion sources. Benzene was selected as representative of air toxics pollutants that are mostly emitted from mobile sources. We focused on a 20 by 20 km area that includes the impacts from a majority of emission sources. The locations of 318 census block group centroids and the modeling domain are shown in Figure 1.

Next, we used these modeled concentrations to develop and evaluate LUR models. The LUR models were then evaluated to examine the implications of varying the number of training sites (25, 50, 75, 100, 125, 150, 200, and 285 sites) used to build the models on LUR fit and performance.
2.1. Air Quality Modeling (Stage 1)

Detailed, spatially-resolved hourly concentrations for multiple pollutants were estimated based
on a hybrid air quality model approach using the AERMOD dispersion model to provide
predictions of local concentration gradients (e.g., within a few meters to few kilometers from the
source) and the CMAQ regional model to provide concentrations due to photochemical reactions
and transport from sources outside the domain, e.g., background (Isakov et al., 2009, Touma et.
al., 2006). The model results showed that there was a wide range of predicted ambient
concentrations across the study area. Spatial maps of modeled concentrations, shown in Figure 2
for all three pollutants, reveal spatial variability due to impact of mobile and stationary sources.
Predicted concentration levels were highest in the city center and near port areas and highways.

The air quality model estimates have been evaluated against available ambient measurements
(Cook et al., 2008; Isakov, et. al. 2009). Figure 3 displays the observed and predicted PM$_{2.5}$
concentrations at the available three monitors in the area in 2001. For PM$_{2.5}$, there is a
noticeable difference in observed concentrations between the three monitors showing a degree of
spatial variability across the area. The model was able to reasonably simulate these spatial
differences; both the magnitude and distributions of modeled concentrations look similar to
observed concentrations at three monitor locations. For NO$_x$, there was insufficient amount of
observational data to perform a model to measurement comparison in a conventional way.
Therefore, we compared time trends between modeled concentrations and available monitor
data, and found a similar rate of decrease in NO$_x$ concentrations in both modeled and monitor
data (Fig. 4). In this comparison, we used all observed concentrations from monitor 0027
(41.30111°N, -72.902778°W) for the four year period from 2005 to 2008, and daily average
concentrations at the closest receptor to the monitor location coordinates to create distributions
of NO$_x$ concentrations for modeled years 2001 and 2010. The distributions show similar trends
for both mean and median concentrations. Model evaluation for benzene and CO is presented in
Cook et al. (2008). For benzene, hourly model results were compared with existing monitor
data from monitor 9005 (latitude 41.34111 N, longitude 72.921389 W) at a suburban location.
Generally, modeled results agree with observed values to within a factor of two. Both modeled
and observed concentrations vary according to time of day. This is expected because benzene
emissions from automobiles are higher in the morning and afternoon rush-hour periods, and
meteorological conditions are more favorable to less dispersion during these periods.

2.2. Land Use Regression Modeling (Stage 2)

LUR models utilize ambient air quality measurements together with detailed land-use data from
gеographic information systems (GIS) to estimate concentrations throughout an area. To build a
LUR model, the measured values of air quality concentrations at sampling locations are
regressed against GIS variables. Air pollution concentrations may then be predicted for any
location in the study area, such as at individual homes, using the parameter estimates derived
from the regression model. In this study, we developed LUR models for PM$_{2.5}$, NO$_x$, and
benzene by combining land-use data with air pollution concentrations predicted by coupled
regional and local scale air quality models. The dependent variables for LUR development were
modeled pollutant concentrations of NO$_x$, benzene, and PM$_{2.5}$ generated by air quality models.
The independent variables were traffic data, emission source information, and land-use variables.

2.2.1 Air Pollution Data

Two-month (July-August of 2001) summer time average of hourly modeled concentrations of
PM$_{2.5}$, NO$_x$, and benzene were estimated at 318 census block group centroids in New Haven, CT.
These receptors can be viewed as pseudo-monitoring sites (instead of field measurements) for
the purpose of their use in the LUR model.

2.2.2. Land-Use Information

Geographic Information Systems (GIS) were used to generate a broad range of spatial land-use
variables including traffic, industrial point sources, and housing or population density as
potential predictors of estimated air pollutant concentration. Data sources for variables were: 1)$link level road network for the New Haven area from the Connecticut Department of
Transportation (CDOT); 2) traffic volumes from the TRANPLAN TRANsportation PLANNing
four step travel demand integrated model (ref http://www.citilabs.com/tranplan/). Potential
misalignment of road locations was addressed by merging the travel demand model’s dataset
with a more spatially accurate U.S. Census Topographically Integrated Geographic Encoding
and Referencing (TIGER) 2000 road network (Cook et al., 2008); 3) 2000 U.S. Census data; and
4) distances from point source emissions, and port and harbor locations, from the EPA 1999 National Emission Inventory database. Traffic variables included inverse distance to roadways with various traffic densities and traffic intensity within various distances. Point source variables included inverse distance to industrial emitters and inverse distance to the airport, harbor, and seaport. Values for land-use variables were calculated using algorithms consistent with previous LUR model development by EPA (Smith et al., 2006; Mukerjee et al., 2009). From the initial pool of approximately 60 potential predictors, we eliminated variables that were highly correlated ($R^2$~1.0) with other selected predictors or difficult to interpret. For example, inverse distance to the airport was excluded because the small size and location of the airport made this variable difficult to interpret. 19 land-use variables were included in the final model selection process.

2.2.3. Site Selection

Training sites used to develop LUR models were selected via stratified random sampling. Specifically, we divided the study area into 4 geographic regions and randomly selected sites within those areas for inclusion in the training data sets. The number of sites selected from each geographic area was proportional to the number of census block groups in that area. For typical LUR modeling research, site selection is done in a more purposeful manner, e.g., using location allocation modeling or other analysis to determine the representativeness of the sites. However, these site selection approaches are mostly convenience based in terms of representativeness of sites selected near key sources (e.g., near roadways, industrial facilities, etc.) mainly due to practical considerations, such as security, accessibility, power, shelter, approval considerations. For instance, Smith et al., 2006 study in El Paso, TX and few others like that have deployed samplers only on school properties dispersed in the study community. Here, the modeling coverage using 318 receptors was quite extensive and highly resolved spatially. Therefore, we believe our stratified random site selection procedure reasonably corresponds to the monitoring based LUR modeling approaches.

2.2.4. LUR Model Development and Evaluation
We developed LUR models using training datasets with 25, 50, 75, 100, 125, 150, 200, and 285 census block group sites. To be consistent with the typical approach used in the LUR modeling literature, we chose the model specifications that provided best fit to the data for each pollutant and number of training sites selected, instead of applying a pre-specified or constant model that would not have provided the highest predictive power for the fitted LUR models. Based on preliminary analysis, the spatial distribution of pollutants in the study area was characterized using multivariate linear regression models. For model selection, we examined all linear models containing 3-7 independent predictors, the optimal number of predictors based on preliminary analyses. To address potential collinearity and improve interpretability, LUR models were excluded from further consideration if the signs for any of the regression coefficients in the model were not in the expected direction. The best model for each training set was selected using Akaike information criteria (AIC), Mallow’s C(p), the proportion of variance explained (adjusted $R^2$) and variance inflation factor (VIF). None of the final models included parameters with VIF > 5. For each number of training sites considered (25-285), we repeated the process of site selection and model development 100 times for each training set size (25-285) to generate a total of 900 “best” LUR models.

We evaluated the fitted LUR models using various approaches, including: 1) Leave-One-Out-Cross-Validation (LOOCV), an internal cross-validation within the selected training sites used to build the LUR models; and 2) Hold-Out evaluation comparing LUR-predicted concentrations with observed AQ model concentrations in test sites that were not used to develop the LUR models. For each iteration, sites that were not selected for model development were used for evaluation. The number of evaluation sites in each test dataset was equal to the difference between the total number of sites in the study area (318) and the number of sites included in the training datasets (25-285). Therefore, the number of test sites ranged from 33 tests sites (for 285 training sites) to 293 test sites (for 25 training sites).

The results of this paper reflect model characteristics and statistics for the 900 LUR models representing the 100 best models selected from each training site group (25-285). Percent contributions for each land-use variable was calculated as the product of the modeled regression coefficient and the mean for that predictor in the study area, divided by the mean concentration
for the pollutant of interest (NO$_x$, benzene, and PM$_{2.5}$) in the study area. All statistical analyses
were conducted in SAS 9.1.

3. Results

Descriptive statistics for pollutant concentrations and land-use characteristics in the study area
are shown in Table 1. There was a wide range in average summer pollutant concentrations. The
coefficients of variation for benzene, NO$_x$, and PM$_{2.5}$ were 92, 90, and 22%, respectively. The
highest levels were found in the city center, port areas, and near major roadways, and there was
also a high-degree of variation in the selected land-use variables. To build the LUR models, we
used 19 land use variables in seven different categories, and the number of independent
predictors in each LUR model ranged from 3 to 7. The LUR models for benzene, NO$_x$, and PM$_{2.5}$
were similar in terms of general model characteristics and performance. Benzene models
typically had the greatest number of independent predictors per model, followed by NO$_x$ and
PM$_{2.5}$.

One of the study objectives (under Aim 1) was to investigate the relative contribution of land-use
predictors to variability in air pollutant concentrations. Figure 5 shows the average percent
contribution of the six larger land use predictors. Percent contributions for each predictor land-
use variable, shown in Figure 5, were calculated as the product of the modeled regression
coefficient and the mean for that predictor in the study area, divided by the mean concentration
for the pollutant of interest (NO$_x$, benzene, and PM$_{2.5}$) in the study area, then standardized in
relation to other categories of predictors. For benzene and NO$_x$, the most important predictors
were traffic intensity and proximity to roadways, but for PM$_{2.5}$ the contribution of these two
variables was much smaller. This was expected, because PM$_{2.5}$ has a large component due to the
secondary formation, while the other two are driven by primary impact from mobile sources.

Looking across the chart in Figure 5, we compared the percent contribution by different type of
LUR models (e.g., using a limited number of observations to build the model versus using a
large number of observations We found that as the number of observations in training sites
increased, traffic intensity became more important compared to proximity to roadways. This
suggests that proximity variables explained the influence of traffic on receptor concentrations
when the number of observations was limited, whereas when the number of sites increased there
was more heterogeneity (e.g., hot-spots) and traffic intensity was much more robust predictor of concentrations than the simpler proximity measures, for these larger and more complex areas.

We then evaluated the fitted LUR models using various approaches, including LOOCV (Aim 2) and Hold-Out evaluation (Aim 3) as described in the Methods section. Finally, we examined whether the LUR models performed similarly for different pollutants. For this comparison, we used the following evaluation metrics: predicted versus observed correlation (adjusted model R²) and “Mean Residuals”, calculated as mean predicted minus observed concentrations. Both cross validation and hold-out evaluation results are provided in Table 2 (a-c) for benzene, NO₂, and PM₂.₅, respectively.

The results of model performance or explanatory power of LUR models (measured by adjusted model R² values) in training datasets are shown in Figure 6. Mean and variability in adjusted R² values (shown here as mean and the inter-quartile range or 25th – 75th percentiles of the mean adjusted R² values derived from 100 different iterations) for LUR models were inversely associated with the number of sites in the training set. Mean adjusted R² values for benzene ranged from 0.89 for training datasets with 25 sites to 0.67 for training datasets with 285 sites. For NO₂, mean adjusted R² values ranged from 0.79 for training datasets with 25 sites to 0.63 for those with 285 sites. Mean adjusted R² values for PM₂.₅ ranged from 0.83 for training datasets with 25 sites to 0.59 for training datasets with 285 sites. For LOOCV results (Table 4, a-c), model performance improved with increased number of training sites. Mean standardized prediction residuals approached zero and root mean square of standardized prediction residuals approached one as the number of training sites used to develop the models increased. These results were consistent for all three pollutants.

Figure 6 also shows the results from evaluation of LUR models based on the correlations between observed and predicted pollutant concentrations at test sites that were not used for model development (Aim 3). Training sites were selected randomly for each iteration of site selection, model development, and evaluation (e.g., N= 100 iterations for summer benzene models based on 25 training sites); and test sites included all sites that were not used in model development (e.g., 293 test sites for a training data set of 25 sites and 33 test sites for a training set of 285 sites). The individual test sites used to evaluate the LUR models also varied with each
LUR model set. Although adjusted $R^2$ values for LUR models in the training datasets decreased with number of training sites used to develop the models, model performance and robustness in the test datasets used for hold-out evaluation improved with increased number of training sites (Figure 6). Specifically, the correlation between observed pollutant concentrations and LUR estimates in sites that were not used to develop the models (adjusted $R^2$ predicted versus observed in the test dataset) was positively associated with number of training sites used to develop the models. The adjusted $R^2$ for LUR models in the training datasets and adjusted $R^2$ for predicted versus observed in test datasets began to converge at approximately 125 sites. These results were observed in LUR models for benzene, NO$_x$, and PM$_{2.5}$.

The mean difference between predicted and observed pollutant concentrations for benzene, NO$_x$, and PM$_{2.5}$ in the test datasets also decreased with number of training sites used to develop the LUR models. Mean difference was approximately 10-fold higher for LUR models based on 25 versus 285 training sites for all three pollutants. The mean difference between predicted and observed benzene concentrations was 0.32 µg/m$^3$ for LUR models developed using 25 sites versus 0.034 µg/m$^3$ for LUR models developed using 285 sites. Similarly, the mean difference for NO$_x$ was 18.7 µg/m$^3$ for LUR models based on 25 sites and 2.9 µg/m$^3$ for LUR models based on 285 sites. Finally, mean difference for PM$_{2.5}$ was 0.77 µg/m$^3$ for LUR models developed using 25 sites versus -0.049 µg/m$^3$ for LUR models developed using 285 sites.

We also evaluated models model performance for various ranges of air quality concentrations to investigate whether LUR model performance was different for either the low end or the high end of the distribution. The results indicate that LUR prediction error was not uniform across the entire range of air quality concentrations used as inputs to develop the models. Figure 7 displays the difference of model vs. observed concentrations for low (0 to 25$^{th}$ percentile), medium (25$^{th}$ to 75$^{th}$ percentile), and high (75$^{th}$ percentile to max) quartiles of the concentration distributions for NO$_x$. The prediction error was much higher for the upper tail of the concentration distribution. This behavior was observed for all range of numbers of training sites from 25 to 285 and for all three pollutants: benzene, NO$_x$, and PM$_{2.5}$. This phenomenon was largely expected under typical air pollution conditions. Pollutant concentrations typically exhibit less variability at the low end of the concentration distribution, which often represent background
conditions that are spatially and temporally more homogeneous, compared with conditions at the upper end of the concentration distributions which reflect the influence of pollution “hot spots” (Rao et al., 1985; Li et al., 2008). Thus, the confidence intervals for the upper tails of the prediction errors tend to be large due to the highly variable nature of the underlying high concentration data.

4. Discussion

In this study we linked land-use regression modeling approaches with combined regional-local scale air quality models in order to evaluate and improve LUR techniques. The air quality modeling results were first evaluated against available monitoring data to assure their reliability and later used to develop and evaluate LUR models in a hierarchical fashion using an iterative site selection approach. We evaluated the fitted LUR models in several different ways and examined the implications of varying the number of training sites used to develop the LUR models on LUR model performance for multiple pollutants.

Among the LUR models examined for benzene, PM$_{2.5}$ and NO$_x$, the benzene models had the best fit, with mean adjusted $R^2$ values ranging from 0.89 to 0.67 within the range of training sites considered (e.g., for models with N=25 to N=285 training sites, respectively) and from 100 different iterations (or alternative selection of receptors within each set of training sites) modeled. The LUR model fits for benzene were quite robust over the range of training sites considered (coefficient of variation for adjusted model $R^2$ values were within 10% to 20% of the mean adjusted $R^2$). Internal evaluation based on adjusted model $R^2$ values within the training sites used to fit the LUR models indicated a good fit to data. However, when these LUR models were tested against an independent hold-out data set ranging from N=293 to 33, their performance diminished considerably. For example, the average adjusted $R^2$ for predicted and observed values in the hold out data set dropped to 0.27 for benzene LUR models fitted to N=25 sites and slightly improved to 0.44 when fitted to N=100 sites (with associated inter-quartile range values for the adjusted $R^2$ estimates of 0.22 and 0.33, respectively). The model performance measured by average adjusted $R^2$ of the predicted vs. observed values using the independent test data asymptotically approached the average adjusted model $R^2$ values for LUR performance in the training data sets for the N=285 case as the coverage of the LUR models.
extended to the full study domain. These results suggest that model performance can improve as the number of training sites increases even though the adjusted model $R^2$ values within training sites may decrease.

Similar trends in the results for PM$_{2.5}$ and NO$_x$ LUR model results were also observed. Again, $R^2$ values determined from alternative realizations of LUR models fits to various size training sites were considerably inflated (e.g., usually 50% or greater than the $R^2$ values determined from those when models were tested against an independent data set). These findings have important implications for the design of LUR monitoring campaigns in support of epidemiology studies. Since sample size requirements for air pollution health effects studies are usually proportional to inverse $R^2$ of the predicted versus observed or true exposure estimates, some studies may be underpowered if the LUR model $R^2$ values are suspected to be biased low. When these situations may be of concern, a wider and more intensive monitoring network to support LUR model development may be advisable. In addition, preliminary analysis of mean predicted minus observed concentrations stratified by low (less than 25th percentile), medium (25th-75th percentile), and high (greater than 75th percentile) concentrations in the hold-out data sets suggested that LUR model prediction errors varied by pollutant level and as a function of the number of training sites used to fit the models. This issue may also have implications to epidemiological application of LUR models and will be explored in depth in future publications.

Our results confirm the challenges facing the LUR community when attempting to fit empirical response surfaces to spatially and temporally varying urban pollution levels. Even at the 2-month averaging periods considered in this initial research, it is clear that complex emissions and atmospheric processes due to meteorological, transport, diffusion and chemical mechanisms can substantially limit the predictive power of most straightforward LUR based models. Clearly, the greater the number of sites selected for building these models (e.g., above N=100) the fit of these models can improve up to a certain level. Unfortunately, the locations where these sites are selected within the airshed could introduce additional uncertainties and potential for exposure misclassification in community based air pollution epidemiology studies. Of course, the greatest concern is during the application of these empirical models for predicting ambient concentrations at hundreds of untested locations of health study subjects within the study community. We have
shown that the ambient concentration prediction errors greatly increase (nearly double) over the
wide range and type of LUR models we evaluated. The variations that we observed in LUR
model performance (e.g., across different models, pollutants and sample sizes) are most likely
reflected in the diverse range of LUR model fits reported in the published literature, especially
for NO$_x$ and PM$_{2.5}$.

Clearly, more work needs to be done in the future in order test how one might transfer LUR
model results across different geographical locations or even countries. Based on the results from
this work, we believe that future work should examine best ways to augment basic LUR models
with site-specific source-receptor information generated from air quality models. Despite their
known limitations (e.g., need for detailed emissions and meteorological information and
uncertainties due to model inputs, algorithms and outputs), air quality models have several
features that can be useful in improving LUR model applications. For example, air quality
models can reliably provide temporal (hourly) and spatial (at hundreds of locations) estimates,
and have a long history of use by regulatory agencies in multi pollutant mitigation strategies. Air
quality models are also based on physical and chemical principles, are widely used and tested in
permit applications, and are readily and freely available with extensive user support tools, such
as user's guides and manuals. Air quality models often undergo extensive peer review so model
improvements are continuously made to enhance its scientific credibility.

Finally, a model-based approach for characterizing complex urban concentrations in the future
will also allow for more effectively designing air monitoring campaigns required for developing
reliable multipollutant LUR models in support of exposure and health studies. In this study, we
examined the influence of site selection considerations on the performance of different LUR
models for each pollutant individually but not jointly. However, we expect that not only the
number of sites used but how they are oriented may influence the overall predictive power of
LUR models for multiple pollutants, using the information derived from a single monitoring
network. We plan to examine optimization strategies for fixed sampling networks that may better
support the development of LUR models for multiple pollutants. Finally, we also plan to
examine in future work various practical approaches to address temporal issues important to
most epidemiological studies, such as alternative exposure averaging periods, ranging from few
days to few weeks and across different seasons.

Acknowledgements

Markey Johnson performed most of the work while affiliated with the U.S. EPA. We thank
Luther Smith (Aliion Science and Technology, Inc.) for sharing his expertise in LUR
development and evaluation, and Ellen Kinnee (Computer Science Corporation) for generating
the land-use variables. We also thank the many contributors from New Haven for providing
support and assistance to EPA scientists in the development of coupled regional and local scale
air quality models. We thank Dr. ST Rao for his many helpful suggestions.

Disclaimer

The United States Environmental Protection Agency, through its Office of Research and
Development, partially funded and collaborated in the research described here under contract no.
Science and Technology, Inc.. It has been subjected to Agency review and approved for
publication. Approval does not signify that the contents reflect the views of the Agency nor does
mention of trade names or commercial products constitute endorsement or recommendation for
use.

References

quality monitoring: Factors affecting network design and interpretation of data. *Air Qual.

Brauer, M., Hoek, G., Van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys,
particulate air pollution concentrations by traffic indicators for epidemiological studies.
*Epidemiology* 14, 228–239.


List of Figures

Figure 1. Map of the study area

Figure 2. Spatial maps of modeled 2-month average concentrations in New Haven, CT for a) benzene, b) NO_{x}, and c) PM_{2.5}

Figure 3. Distributions of modeled and observed daily average PM_{2.5} concentrations in 2001 at three monitor locations

Figure 4. Trends of modeled and observed distributions of daily average NO_{x} concentrations at one monitor.

Figure 5. Percent contribution of land-use predictors by pollutant and number of training sites

Figure 6. Mean and inter-quartile range of adjusted R^{2} values for LUR models in both training (upper horizontal axis) and test datasets (lower horizontal axis) as a function of number of training sites for a) Benzene, b) NO_{x} and c) PM_{2.5}

Figure 7. LUR model prediction errors (defined as average +/- standard deviation) for NO_{x} concentrations as a function of three categories of concentrations (low, medium, high) and the number of training sites: a) 25 sites b) 100 sites and c) 285 sites.
Figure 1. Map of the study area
Figure 2. Spatial maps of modeled 2-month average concentrations in New Haven, CT for a) benzene, b) NO₃, and c) PM₂.₅
Figure 3. Distributions of modeled and observed daily average PM$_{2.5}$ concentrations in 2001 at three monitor locations.
Figure 4. Trends of modeled and observed distributions of daily average NO\textsubscript{x} concentrations at one monitor.
Figure 5. Percent contribution of land-use predictors by pollutant and number of observations
Figure 6. Mean and inter-quartile range of adjusted $R^2$ values for LUR models as a function of number of sites in both training (upper horizontal axis) and test datasets (lower horizontal axis) for a) Benzene, b) NOx and c) PM$_{2.5}$.
Figure 7. LUR model prediction errors (defined as average +/- standard deviation) for NO\textsubscript{x} concentrations as a function of three categories of concentrations (low, medium, high) and the number of training sites: a) 25 sites b) 100 sites and c) 285 sites.
Table 1. Summary statistics of pollutant concentrations and land-use variables

Table 2. Summary of LUR model evaluation for a) Benzene, b) NO\textsubscript{x}, and c) PM\textsubscript{2.5}
## Table 1. Summary statistics of pollutant concentrations and land-use variables

<table>
<thead>
<tr>
<th>Pollutants (µg/m³)</th>
<th>N</th>
<th>Mean value</th>
<th>Std</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average summer Benzene</td>
<td>318</td>
<td>1.1</td>
<td>1.0</td>
<td>92</td>
</tr>
<tr>
<td>Average summer NO₂</td>
<td>318</td>
<td>95.5</td>
<td>85.7</td>
<td>90</td>
</tr>
<tr>
<td>Average summer PM₂.₅</td>
<td>318</td>
<td>15.4</td>
<td>3.4</td>
<td>22</td>
</tr>
<tr>
<td><strong>Traffic Intensity (vehicles per day/km²)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within 125 meters</td>
<td>318</td>
<td>26200</td>
<td>91500</td>
<td>349</td>
</tr>
<tr>
<td>Within 250 meters</td>
<td>318</td>
<td>37800</td>
<td>101000</td>
<td>267</td>
</tr>
<tr>
<td>Within 500 meters</td>
<td>318</td>
<td>45000</td>
<td>81100</td>
<td>180</td>
</tr>
<tr>
<td>Within 1000 meters</td>
<td>318</td>
<td>49600</td>
<td>55200</td>
<td>111</td>
</tr>
<tr>
<td>Within 1500 meters</td>
<td>318</td>
<td>49400</td>
<td>46100</td>
<td>93</td>
</tr>
<tr>
<td>Within 2000 meters</td>
<td>318</td>
<td>49500</td>
<td>39900</td>
<td>81</td>
</tr>
<tr>
<td><strong>Population and Housing Density</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population density within census block group</td>
<td>318</td>
<td>2680</td>
<td>2780</td>
<td>104</td>
</tr>
<tr>
<td>Housing density within census block group</td>
<td>318</td>
<td>1200</td>
<td>1590</td>
<td>132</td>
</tr>
<tr>
<td><strong>Proximity to Roadways (1/km)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roads with ≥ 10,000 vehicles per day (vpd)</td>
<td>318</td>
<td>1.17</td>
<td>5.68</td>
<td>484</td>
</tr>
<tr>
<td>Roads with ≥ 20,000 vpd</td>
<td>318</td>
<td>0.39</td>
<td>1.99</td>
<td>509</td>
</tr>
<tr>
<td>Roads with ≥ 30,000 vpd</td>
<td>318</td>
<td>0.31</td>
<td>1.98</td>
<td>643</td>
</tr>
<tr>
<td>Roads with ≥ 40,000 vpd</td>
<td>318</td>
<td>0.20</td>
<td>0.75</td>
<td>376</td>
</tr>
<tr>
<td>Roads with ≥ 60,000 vpd</td>
<td>318</td>
<td>0.17</td>
<td>0.73</td>
<td>433</td>
</tr>
<tr>
<td>Roads with ≥ 70,000 vpd</td>
<td>318</td>
<td>0.16</td>
<td>0.73</td>
<td>464</td>
</tr>
<tr>
<td>Roads with ≥ 90,000 vpd</td>
<td>318</td>
<td>0.16</td>
<td>0.73</td>
<td>464</td>
</tr>
<tr>
<td><strong>Proximity to Ports and Harbors (1/km)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seaport</td>
<td>318</td>
<td>0.26</td>
<td>0.21</td>
<td>81</td>
</tr>
<tr>
<td>Harbor</td>
<td>318</td>
<td>0.63</td>
<td>1.15</td>
<td>185</td>
</tr>
<tr>
<td><strong>Proximity to Industrial Benzene Sources (1/km)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facilities emitting ≥ 100 lbs/year</td>
<td>318</td>
<td>0.34</td>
<td>0.70</td>
<td>208</td>
</tr>
<tr>
<td>Facilities emitting &lt; 100 lbs/year</td>
<td>318</td>
<td>0.92</td>
<td>1.33</td>
<td>145</td>
</tr>
<tr>
<td><strong>Proximity to Industrial NO₂ Sources (1/km)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facilities emitting ≥ 100 lbs/year</td>
<td>318</td>
<td>1.06</td>
<td>1.27</td>
<td>119</td>
</tr>
<tr>
<td>Facilities emitting &lt; 100 lbs/year</td>
<td>318</td>
<td>0.68</td>
<td>1.03</td>
<td>151</td>
</tr>
<tr>
<td><strong>Proximity to Industrial PM₂.₅ Sources (1/km)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facilities emitting ≥ 100 lbs/year</td>
<td>318</td>
<td>0.83</td>
<td>1.16</td>
<td>140</td>
</tr>
<tr>
<td>Facilities emitting &lt; 100 lbs/year</td>
<td>318</td>
<td>0.95</td>
<td>1.39</td>
<td>146</td>
</tr>
</tbody>
</table>
### Table 2-a. Summary of LUR model evaluation for benzene.

<table>
<thead>
<tr>
<th>Number of Training Sites</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
<th>125</th>
<th>150</th>
<th>175</th>
<th>200</th>
<th>285</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Benzene</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (Range)</td>
<td>0.89 (0.45-0.98)</td>
<td>0.76 (0.30-0.95)</td>
<td>0.72 (0.35-0.92)</td>
<td>0.71 (0.44-0.88)</td>
<td>0.72 (0.48-0.88)</td>
<td>0.69 (0.50-0.82)</td>
<td>0.67 (0.58-0.82)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LUR Model Performance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Model R² (training sites)</td>
<td>0.33 (-11.10-1.06)</td>
<td>0.10 (-2.43-0.24)</td>
<td>-0.02 (-1.81-0.11)</td>
<td>-0.03 (-0.92-0.08)</td>
<td>-0.06 (-0.66-0.06)</td>
<td>-0.02 (-0.54-0.06)</td>
<td>0.00 (-0.07-0.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean SPR</td>
<td>2.53 (1.09-1.63)</td>
<td>2.67 (1.08-1.17)</td>
<td>1.89 (1.07-0.99)</td>
<td>1.84 (1.05-0.89)</td>
<td>2.04 (1.07-1.22)</td>
<td>1.74 (1.07-1.07)</td>
<td>1.34 (1.15-1.90)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS of SPR</td>
<td>4.13 (1.09-5.58)</td>
<td>28.39 (1.09-28.39)</td>
<td>2.59 (1.09-1.82)</td>
<td>1.82 (1.05-1.21)</td>
<td>2.08 (1.04-1.88)</td>
<td>1.74 (1.04-1.77)</td>
<td>1.34 (1.15-1.90)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hold-Out Evaluation Results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Model R² (test sites)</td>
<td>0.27 (0.02-0.67)</td>
<td>0.39 (0.06-0.78)</td>
<td>0.44 (0.04-0.77)</td>
<td>0.52 (0.05-0.83)</td>
<td>0.60 (0.11-0.94)</td>
<td>0.71 (0.15-0.94)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Residuals</td>
<td>0.32 (-0.18-2.01)</td>
<td>0.27 (-0.15-1.80)</td>
<td>0.13 (-0.10-1.40)</td>
<td>0.09 (-0.12-0.64)</td>
<td>0.04 (-0.24-0.82)</td>
<td>0.03 (-0.47-0.25)</td>
<td>0.03 (-0.47-0.25)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2-b. Summary of LUR model evaluation for NO₂

<table>
<thead>
<tr>
<th>Number of Training Sites</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
<th>125</th>
<th>150</th>
<th>175</th>
<th>200</th>
<th>285</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NO₂</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (Range)</td>
<td>0.79 (0.08-0.99)</td>
<td>0.69 (0.32-0.89)</td>
<td>0.66 (0.41-0.90)</td>
<td>0.66 (0.40-0.85)</td>
<td>0.65 (0.43-0.83)</td>
<td>0.65 (0.45-0.83)</td>
<td>0.63 (0.54-0.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LUR Model Performance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Model R² (training sites)</td>
<td>0.00 (-6.01-1.22)</td>
<td>-0.09 (-1.14-0.11)</td>
<td>-0.06 (-0.86-0.08)</td>
<td>-0.03 (-0.73-0.07)</td>
<td>-0.02 (-0.47-0.06)</td>
<td>-0.01 (-0.22-0.04)</td>
<td>0.00 (-0.08-0.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean SPR</td>
<td>2.81 (1.10-3.58)</td>
<td>2.42 (1.11-1.64)</td>
<td>2.00 (1.10-0.93)</td>
<td>1.72 (1.08-0.93)</td>
<td>1.50 (1.08-0.93)</td>
<td>1.36 (1.07-0.93)</td>
<td>1.18 (1.07-0.93)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| RMS of SPR                | 18.68 (-19.33-111.17) | 9.36 (-97.01-69.62) | 6.78 (-14.65-56.40) | 8.92 (-99.38-66.21) | 5.55 (-15.12-62.64) | 1.91 (-90.80-78.29) | 1.03 (-16.83-43.96) | 2.87 (-40.78-46.85) | 30
<table>
<thead>
<tr>
<th>NO₂</th>
<th>Number of Training Sites</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
<th>125</th>
<th>150</th>
<th>175</th>
<th>200</th>
<th>285</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
<td>Mean (Range)</td>
</tr>
<tr>
<td></td>
<td>LUR Model Performance</td>
<td>Adjusted Model</td>
<td>0.83 (0.20-0.99)</td>
<td>0.72 (0.16-0.97)</td>
<td>0.72 (0.23-0.94)</td>
<td>0.65 (0.27-0.94)</td>
<td>0.66 (0.28-0.90)</td>
<td>0.66 (0.38-0.89)</td>
<td>0.63 (0.37-0.82)</td>
<td>0.61 (0.37-0.79)</td>
</tr>
<tr>
<td></td>
<td>R² (training sites)</td>
<td>Cross-Validation Results</td>
<td>-0.06 (-4.57-1.55)</td>
<td>0.00 (-1.58-0.27)</td>
<td>0.01 (-0.53-0.16)</td>
<td>0.02 (-0.27-0.11)</td>
<td>0.00 (-0.41-0.09)</td>
<td>0.01 (-0.25-0.07)</td>
<td>0.01 (-0.16-0.07)</td>
<td>0.01 (-0.05-0.03)</td>
</tr>
<tr>
<td></td>
<td>Mean SPR</td>
<td>RMS of SPR</td>
<td>3.11 (1.24-23.62)</td>
<td>2.42 (1.15-12.82)</td>
<td>1.89 (1.11-13.63)</td>
<td>1.91 (1.09-7.73)</td>
<td>1.49 (1.12-3.36)</td>
<td>1.64 (1.11-6.46)</td>
<td>1.52 (1.08-3.55)</td>
<td>1.52 (1.07-2.79)</td>
</tr>
<tr>
<td></td>
<td>Hold-Out Evaluation Results</td>
<td>Adjusted Model</td>
<td>0.29 (0.01-0.73)</td>
<td>0.32 (0.02-0.64)</td>
<td>0.33 (0.02-0.68)</td>
<td>0.38 (0.02-0.79)</td>
<td>0.42 (0.02-0.73)</td>
<td>0.44 (0.02-0.83)</td>
<td>0.47 (0.05-0.88)</td>
<td>0.48 (0.05-0.95)</td>
</tr>
<tr>
<td></td>
<td>R² (test sites)</td>
<td>Mean Residuals</td>
<td>0.77 (-0.78-5.61)</td>
<td>0.66 (-0.61-5.30)</td>
<td>0.48 (-0.53-5.83)</td>
<td>0.53 (-0.60-5.69)</td>
<td>0.29 (-0.57-2.35)</td>
<td>0.15 (-0.71-1.94)</td>
<td>0.16 (-0.71-2.16)</td>
<td>0.20 (-1.24-0.77)</td>
</tr>
</tbody>
</table>