6/15/2010

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. Özkaynak [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.

1. Introduction

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 dependant 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; Jerret 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.

89

90 The current study has the following four main aims:

91 1) Develop conventional LUR models for New Haven, CT using locally collected
92 demographic, housing, land-use, proximity to traffic and other main source information,
93 along with modeled air quality concentrations for benzene, NO_x and $PM_{2.5}$ (e.g., as
94 pseudo-monitoring data) at numerous receptors (318 total), instead of average pollutant
95 concentration measurements that are typically collected at limited number of monitoring
96 locations.
97 2) Internally evaluate the performance of fitted LUR models (ascertained by adjusted model
98 R^2 and other measures) as a function of the number and location of receptor (e.g., pseudo-
99 monitoring) sites selected, through an iterative stratified random sampling procedure.
100 3) Externally evaluate the predictive power of the fitted LUR models by comparing the
101 LUR model predictions to concentrations at receptors that were not used in fitting the
102 LUR models.
103 4) Examine and recommend strategies for optimum site selection in order to improve the
104 predictive power and transferability of LUR models.

105 ## 2. Methods

106 The air quality and LUR modeling analysis was done in two stages. First, we predicted air
107 quality concentrations of $PM_{2.5}$, NO_x, and benzene using hybrid modeling techniques based on
108 CMAQ and AERMOD model results. $PM_{2.5}$ was selected as a criteria pollutant that is mostly
109 regional and is largely formed in the atmosphere due to secondary reactions. NO_x was selected
110 because it is strongly influenced by local combustion sources. Benzene was selected as
111 representative of air toxics pollutants that are mostly emitted from mobile sources. We focused
112 on a 20 by 20 km area that includes the impacts from a majority of emission sources. The
113 locations of 318 census block group centroids and the modeling domain are shown in Figure 1.
114 Next, we used these modeled concentrations to develop and evaluate LUR models. The LUR
115 models were then evaluated to examine the implications of varying the number of training sites
116 (25, 50, 75, 100, 125, 150, 200, and 285 sites) used to build the models on LUR fit and
117 performance.
118

119 *2.1. Air Quality Modeling (Stage 1)*

120 Detailed, spatially-resolved hourly concentrations for multiple pollutants were estimated based
121 on a hybrid air quality model approach using the AERMOD dispersion model to provide
122 predictions of local concentration gradients (e.g., within a few meters to few kilometers from the
123 source) and the CMAQ regional model to provide concentrations due to photochemical reactions
124 and transport from sources outside the domain, e.g., background (Isakov et. al., 2009, Touma et.
125 al., 2006). The model results showed that there was a wide range of predicted ambient
126 concentrations across the study area. Spatial maps of modeled concentrations, shown in Figure 2
127 for all three pollutants, reveal spatial variability due to impact of mobile and stationary sources.
128 Predicted concentration levels were highest in the city center and near port areas and highways.
129
130 The air quality model estimates have been evaluated against available ambient measurements
131 (Cook et. al., 2008; Isakov, et. al. 2009). Figure 3 displays the observed and predicted $PM_{2.5}$
132 concentrations at the available three monitors in the area in 2001. For $PM_{2.5}$, there is a
133 noticeable difference in observed concentrations between the three monitors showing a degree of
134 spatial variability across the area. The model was able to reasonably simulate these spatial
135 differences; both the magnitude and distributions of modeled concentrations look similar to
136 observed concentrations at three monitor locations. For NO_x, there was insufficient amount of
137 observational data to perform a model to measurement comparison in a conventional way.
138 Therefore, we compared time trends between modeled concentrations and available monitor
139 data, and found a similar rate of decrease in NO_x concentrations in both modeled and monitor
140 data (Fig. 4). In this comparison, we used all observed concentrations from monitor 0027
141 (41.301111°N, -72.902778°W) for the four year period from 2005 to 2008, and daily average
142 concentrations at the closest receptor to the monitor location coordinates to create distributions
143 of NO_x concentrations for modeled years 2001 and 2010. The distributions show similar trends
144 for both mean and median concentrations. Model evaluation for benzene and CO is presented in
145 Cook et. al. (2008). For benzene, hourly model results were compared with existing monitor
146 data from monitor 9005 (latitude 41.34111 N, longitude 72.921389 W) at a suburban location.
147 Generally, modeled results agree with observed values to within a factor of two. Both modeled
148 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 geographic 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 ($R^2 \sim 1.0$) 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 Aikaike 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

240 for the pollutant of interest (NO_x, benzene, and $PM_{2.5}$) in the study area. All statistical analyses
241 were conducted in SAS 9.1.

242 **3. Results**

243 Descriptive statistics for pollutant concentrations and land-use characteristics in the study area
244 are shown in Table 1. There was a wide range in average summer pollutant concentrations. The
245 coefficients of variation for benzene, NO_x, and $PM_{2.5}$ were 92, 90, and 22%, respectively. The
246 highest levels were found in the city center, port areas, and near major roadways, and there was
247 also a high-degree of variation in the selected land-use variables. To build the LUR models, we
248 used 19 land use variables in seven different categories, and the number of independent
249 predictors in each LUR model ranged from 3 to 7. The LUR models for benzene, NO_x, and $PM_{2.5}$
250 were similar in terms of general model characteristics and performance. Benzene models
251 typically had the greatest number of independent predictors per model, followed by NO_x and
252 $PM_{2.5}$.
253
254 One of the study objectives (under Aim 1) was to investigate the relative contribution of land-use
255 predictors to variability in air pollutant concentrations. Figure 5 shows the average percent
256 contribution of the six larger land use predictors. Percent contributions for each predictor land-
257 use variable, shown in Figure 5, were calculated as the product of the modeled regression
258 coefficient and the mean for that predictor in the study area, divided by the mean concentration
259 for the pollutant of interest (NO_x, benzene, and $PM_{2.5}$) in the study area, then standardized in
260 relation to other categories of predictors. For benzene and NOx, the most important predictors
261 were traffic intensity and proximity to roadways, but for $PM_{2.5}$ the contribution of these two
262 variables was much smaller. This was expected, because $PM_{2.5}$ has a large component due to the
263 secondary formation, while the other two are driven by primary impact from mobile sources.
264 Looking across the chart in Figure 5, we compared the percent contribution by different type of
265 LUR models (e.g., using a limited number of observations to build the model versus using a
266 large number of observations We found that as the number of observations in training sites
267 increased, traffic intensity became more important compared to proximity to roadways. This
268 suggests that proximity variables explained the influence of traffic on receptor concentrations
269 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 R^2) 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, NO_x, and $PM_{2.5}$, respectively.

The results of model performance or explanatory power of LUR models (measured by adjusted model R^2 values) in training datasets are shown in Figure 6. Mean and variability in adjusted R^2 values (shown here as mean and the inter-quartile range or 25^{th} – 75^{th} percentiles of the mean adjusted R^2 values derived from 100 different iterations) for LUR models were inversely associated with the number of sites in the training set. Mean adjusted R^2 values for benzene ranged from 0.89 for training datasets with 25 sites to 0.67 for training datasets with 285 sites. For NO_x, mean adjusted R^2 values ranged from 0.79 for training datasets with 25 sites to 0.63 for those with 285 sites. Mean adjusted R^2 values for $PM_{2.5}$ 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 $\mu g/m^3$ for LUR models developed using 25 sites versus 0.034 $\mu g/m^3$ for LUR models developed using 285 sites. Similarly, the mean difference for NO_x was 18.7 $\mu g/m^3$ for LUR models based on 25 sites and 2.9 $\mu g/m^3$ for LUR models based on 285 sites. Finally, mean difference for $PM_{2.5}$ was 0.77 $\mu g/m^3$ for LUR models developed using 25 sites versus -0.049 $\mu g/m^3$ 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 25th percentile), medium (25th to 75th percentile), and high (75th 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

362 extended to the full study domain. These results suggest that model performance can improve as
363 the number of training sites increases even though the adjusted model R^2 values within training
364 sites may decrease.
365
366 Similar trends in the results for $PM_{2.5}$ and NO_x LUR model results were also observed. Again, R^2
367 values determined from alternative realizations of LUR models fits to various size training sites
368 were considerably inflated (e.g., usually 50% or greater than the R^2 values determined from
369 those when models were tested against an independent data set). These findings have important
370 implications for the design of LUR monitoring campaigns in support of epidemiology studies.
371 Since sample size requirements for air pollution health effects studies are usually proportional to
372 inverse R^2 of the predicted versus observed or true exposure estimates, some studies may be
373 underpowered if the LUR model R^2 values are suspected to be biased low. When these
374 situations may be of concern, a wider and more intensive monitoring network to support LUR
375 model development may be advisable. In addition, preliminary analysis of mean predicted minus
376 observed concentrations stratified by low (less than 25th percentile), medium (25th-75th
377 percentile), and high (greater than 75th percentile) concentrations in the hold-out data sets
378 suggested that LUR model prediction errors varied by pollutant level and as a function of the
379 number of training sites used to fit the models. This issue may also have implications to
380 epidemiological application of LUR models and will be explored in depth in future publications.
381
382 Our results confirm the challenges facing the LUR community when attempting to fit empirical
383 response surfaces to spatially and temporally varying urban pollution levels. Even at the 2-month
384 averaging periods considered in this initial research, it is clear that complex emissions and
385 atmospheric processes due to meteorological, transport, diffusion and chemical mechanisms can
386 substantially limit the predictive power of most straightforward LUR based models. Clearly, the
387 greater the number of sites selected for building these models (e.g., above N=100) the fit of these
388 models can improve up to a certain level. Unfortunately, the locations where these sites are
389 selected within the airshed could introduce additional uncertainties and potential for exposure
390 misclassification in community based air pollution epidemiology studies. Of course, the greatest
391 concern is during the application of these empirical models for predicting ambient concentrations
392 at hundreds of untested locations of health study subjects within the study community. We have

393 shown that the ambient concentration prediction errors greatly increase (nearly double) over the
394 wide range and type of LUR models we evaluated. The variations that we observed in LUR
395 model performance (e.g., across different models, pollutants and sample sizes) are most likely
396 reflected in the diverse range of LUR model fits reported in the published literature, especially
397 for NO_x and $PM_{2.5}$.
398
399 Clearly, more work needs to be done in the future in order test how one might transfer LUR
400 model results across different geographical locations or even countries. Based on the results from
401 this work, we believe that future work should examine best ways to augment basic LUR models
402 with site-specific source-receptor information generated from air quality models. Despite their
403 known limitations (e.g., need for detailed emissions and meteorological information and
404 uncertainties due to model inputs, algorithms and outputs), air quality models have several
405 features that can be useful in improving LUR model applications. For example, air quality
406 models can reliably provide temporal (hourly) and spatial (at hundreds of locations) estimates,
407 and have a long history of use by regulatory agencies in multi pollutant mitigation strategies. Air
408 quality models are also based on physical and chemical principles, are widely used and tested in
409 permit applications, and are readily and freely available with extensive user support tools, such
410 as user's guides and manuals. Air quality models often undergo extensive peer review so model
411 improvements are continuously made to enhance its scientific credibility.
412
413 Finally, a model-based approach for characterizing complex urban concentrations in the future
414 will also allow for more effectively designing air monitoring campaigns required for developing
415 reliable multipollutant LUR models in support of exposure and health studies. In this study, we
416 examined the influence of site selection considerations on the performance of different LUR
417 models for each pollutant individually but not jointly. However, we expect that not only the
418 number of sites used but how they are oriented may influence the overall predictive power of
419 LUR models for multiple pollutants, using the information derived from a single monitoring
420 network. We plan to examine optimization strategies for fixed sampling networks that may better
421 support the development of LUR models for multiple pollutants. Finally, we also plan to
422 examine in future work various practical approaches to address temporal issues important to

423 most epidemiological studies, such as alternative exposure averaging periods, ranging from few
424 days to few weeks and across different seasons.

425 **Acknowledgements**

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

432 **Disclaimer**

433 The United States Environmental Protection Agency, through its Office of Research and
434 Development, partially funded and collaborated in the research described here under contract no.
435 68-W-01-032 task 61 to Computer Sciences Corporation and Contract No. EP-D-05-065 to Alion
436 Science and Technology, Inc.. It has been subjected to Agency review and approved for
437 publication. Approval does not signify that the contents reflect the views of the Agency nor does
438 mention of trade names or commercial products constitute endorsement or recommendation for
439 use.

440 **References**

441 Baldauf, R., Watkins, N., Heist, D., Bailey, C., Rowley, P., Shores, R., 2009. Near-road air
442 quality monitoring: Factors affecting network design and interpretation of data. *Air Qual.*
443 *Atmos. Health.* 2: 1–9.

444 Brauer, M., Hoek, G., Van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys,
445 J., Bellander, T., Lewne, M., Brunekreef, B., 2003. Prediction of long term average
446 particulate air pollution concentrations by traffic indicators for epidemiological studies.
447 *Epidemiology* 14, 228–239.

448 Brunekreef, B., Holgate, S.T., 2002. Air pollution and health. *Lancet* 360, 1233–1242.

449 Cook, R.; Isakov, V.; Touma, J.; Benjey, W.; Thurman, J.; Kinnee, E.; Ensley, D. Resolving
450 Local Scale Emissions for Modeling Air Quality Near Roadways. *J. Air & Waste*
451 *Manage. Assoc.* 2008, 58, 451-461.

452 Gauderman, W.J., Avol, E., Lurmann, F., Kuenzli, N., Gilliland, F., Peters, J., McConnell, R.,
453 2005. Childhood asthma and exposure to traffic and nitrogen dioxide. *Epidemiology* 16
454 (6): 737–743.

455 Harrison, R.M., Tilling, R., Callen Romero, M.S., Harrad, S., Jarvis, K., 2003. A study of trace
456 metals and polycyclic aromatic hydrocarbons in the roadside environment. *Atmos*
457 *Environ* 37:2391–2402.

458 Health Effects Institute, 2010. Traffic-related air pollution: a critical review of the literature on
459 emissions, exposure, and health effects. Special Report #17, 2010-01-12. Available on-
460 line at: http://pubs.healtheffects.org/view.php?id=334 (accessed 02/11/2010).

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

464 Isakov, V., Touma, J., Khlystov, A. 2007. A Method of Assessing Air Toxics Concentrations
465 using Mobile Platform Measurements. *J. A&WMA.* **57**: 1286-1295.

466 Isakov, V., Touma, J., Burke, J., Lobdell, D., Palma, T., Rosenbaum, A., Özkaynak, H., 2009.
467 Combining Regional and Local Scale Air Quality Models with Exposure Models for Use
468 in Environmental Health Studies. *J. A&WMA* 59: 461-472.

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

472 Jerrett, M., Arain, M.A., Kanaroglou, P., Beckerman, B., Crouse, D., Gilbert, N.L., Brook, J.R.,
473 Finkelstein, N., Finkelstein, M.M., 2007. Modeling the intraurban variability of ambient
474 traffic pollution in Toronto, Canada. *Journal of Toxicology and Environmental Health* 70
475 (3&4), 200–212.

476 Kim, J.J., Smorodinsky, S., Lipsett, M., Singer, B.C., Hogdson, A.T., Ostro, B., 2004. Traffic-
477 related air pollution near busy roads: the East Bay Children's Respiratory Health Study.
478 *Am J Respir Crit Care Med* 170(5): 520–526.

479 Li, Y., Simmonds, D., Reeve, D. 2008. Quantifying uncertainty in extreme values of design
480 parameters with resampling techniques. *Ocean Engineering*. 35: 1029-1038.

481 Marshall, J. D., Nethery, E., Brauer M., 2008. Within-urban variability in ambient air pollution:
482 Comparison of estimation methods. *Atmospheric Environment,* 42, 1359–1369.

483 Mukerjee, S., Smith, L.A., Johnson, M.M., Neas, L.M., Stallings, C.A.. 2009. Spatial analysis
484 and land use regression of VOCs and NO2 from school-based urban air monitoring in
485 Detroit/Dearborn, USA. *Science of the Total Environment*, 407. 4642–4651

486 McConnell R., Berhane K., Yao L., Jerrett M., Lurmann F., Gilliland F., Kuenzli N., Gauderman
487 J., Avol E., Thomas D., Peters J., 2006. Traffic, susceptibility, and childhood asthma.
488 *Environ. Health Perspect.,* 114 (5): 766–772.

489 Rao, S.T., Sistla, G., Pagnotti, V., Petersen, W.B., Irwin, J.S., Turner, D.B., 1985. Resampling
490 and extreme value statistics in air quality model performance evaluation. *Atmos. Environ.*
491 19 (9): 1503-1518.

492 Ross, Z., English, P. B., Scalf, R., Gunier, R., Smorodinsky, S., Wall, S., and Jerrett, M. 2006.
493 Nitrogen dioxide prediction in Southern California using land use regression modeling:
494 Potential for environmental health analyses. *J. Expos. Sci. Environ. Epidemiol.* 16:106–
495 114.

496 Ryan, P.H. and LeMasters, G.K., 2007. A Review of Land-use Regression Models for
497 Characterizing Intraurban Air Pollution Exposure. *Inhalation Toxicology, 19 (Suppl.*
498 *1)*:127–133.

499 Samal M.T., Islam T., Gilliland F.D., 2008. Recent evidence for adverse effects of residential
500 proximity to traffic sources on asthma. *Curr Opin Pulm Med* 14 (1):3–8.

501 Samet J.M., 2007. Traffic, air pollution, and health. *Inhal Toxicol* 19:1021–1027.

Smith, L., Mukerjee, S., Gonzales, M., Stallings, C., Neas, L., Norris, G., Özkaynak, H. 2006. Use of GIS and ancillary variables to predict volatile organic compound and nitrogen dioxide levels at unmonitored locations. *Atmos. Environ.*, 40, 3773-3787.

Touma, J. S., V. Isakov, J. Ching, and C. Seigneur. Air quality modeling of hazardous pollutants: current status and future directions. *J. A&WMA.* **56**: 547-558 (2006).

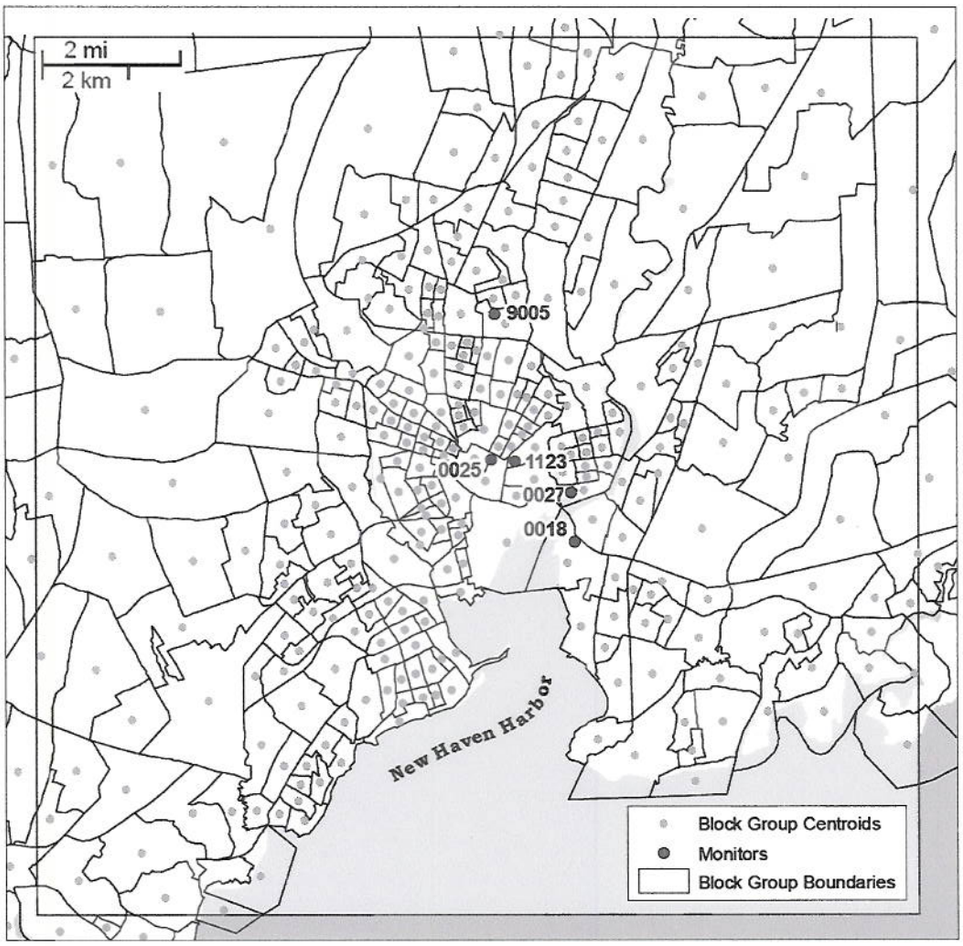
Vienneau, D., de Hoogh, K., Beelen, R., Fischer, P., Hoek, G., Briggs, D., 2010. Comparison of land-use regression models between Great Britain and the Netherlands. *Atmospheric Environment*, 44, 688-696.

Weijers, E.P.; Khlystov, A.Y.; Kos, G.P.A.; Erisman, J.W. Variability of Particulate Matter Concentrations along Roads and Motorways Determined by a Moving Measurement Unit; *Atmos. Environ.* **2004**, *38*,2993-3002.

Westerdahl, D.; Fruin, S.; Sax, T.; Fine, P.; Sioutas, C. Mobile Platform Measurements of Ultrafine Particles and Associated Pollutant Concentrations on Freeways and Residential Streets in Los Angeles; *Atmos. Environ.* **2005**, *39,* 3597-3610.

Zhu Y., Hinds W.C., Kim S., Sioutas C., 2002. Concentration and size distribution of ultrafine particles near a major highway. *J Air Waste Manage Assoc.* 52: 1032–1042.

518 **List of Figures**



533

534 Figure 1. Map of the study area

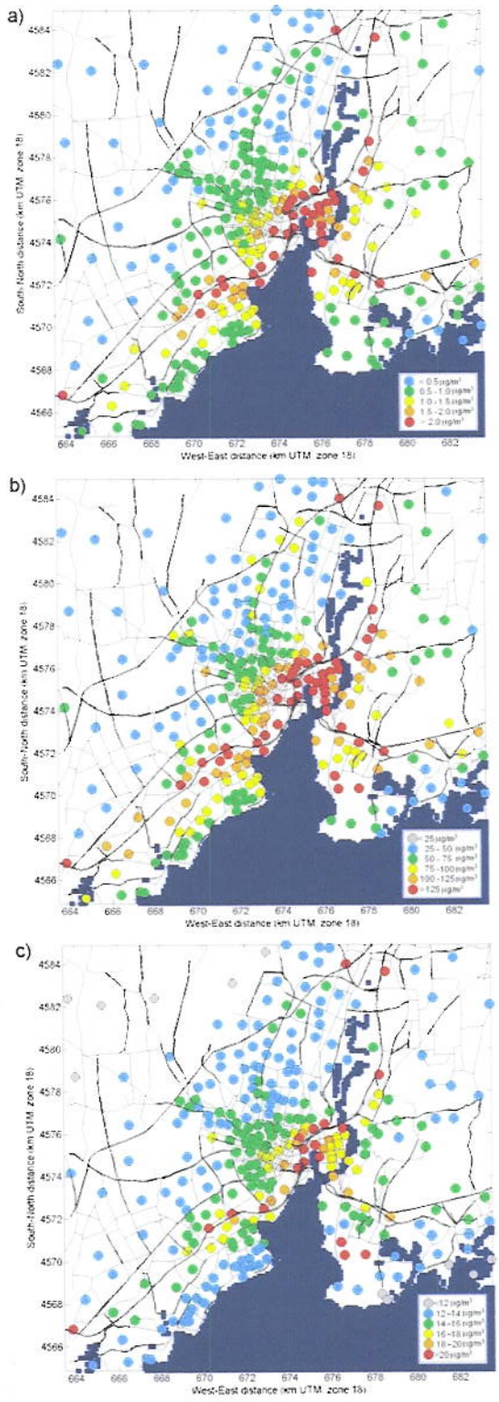
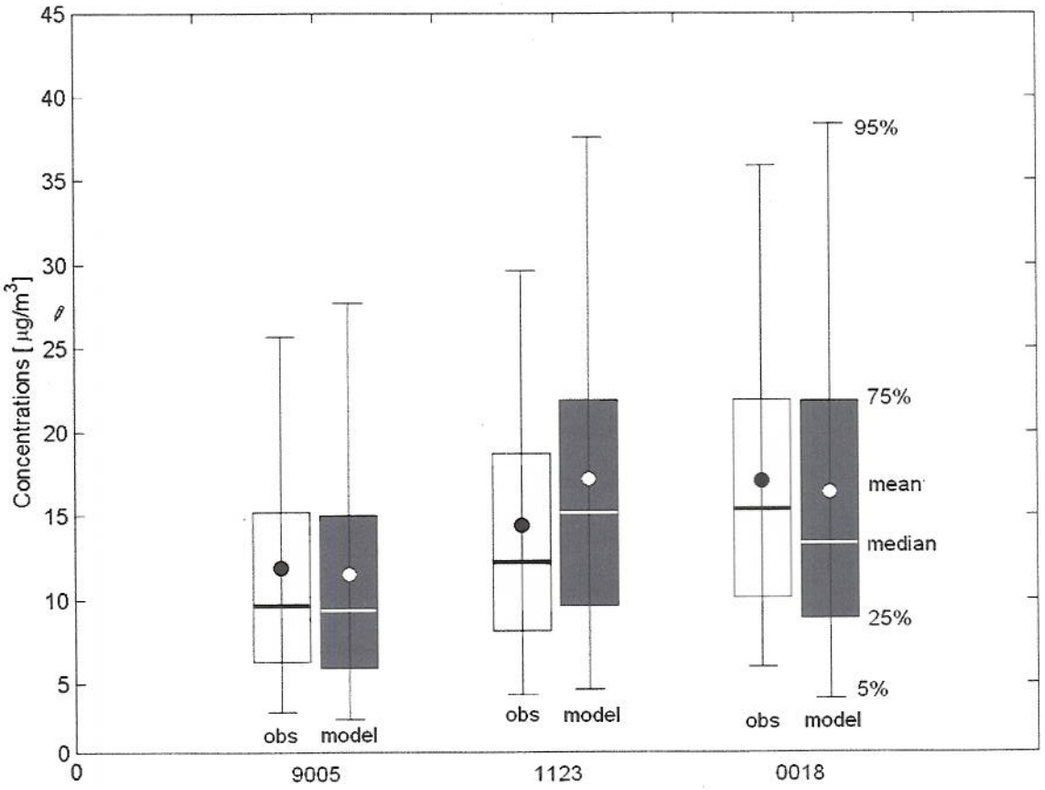
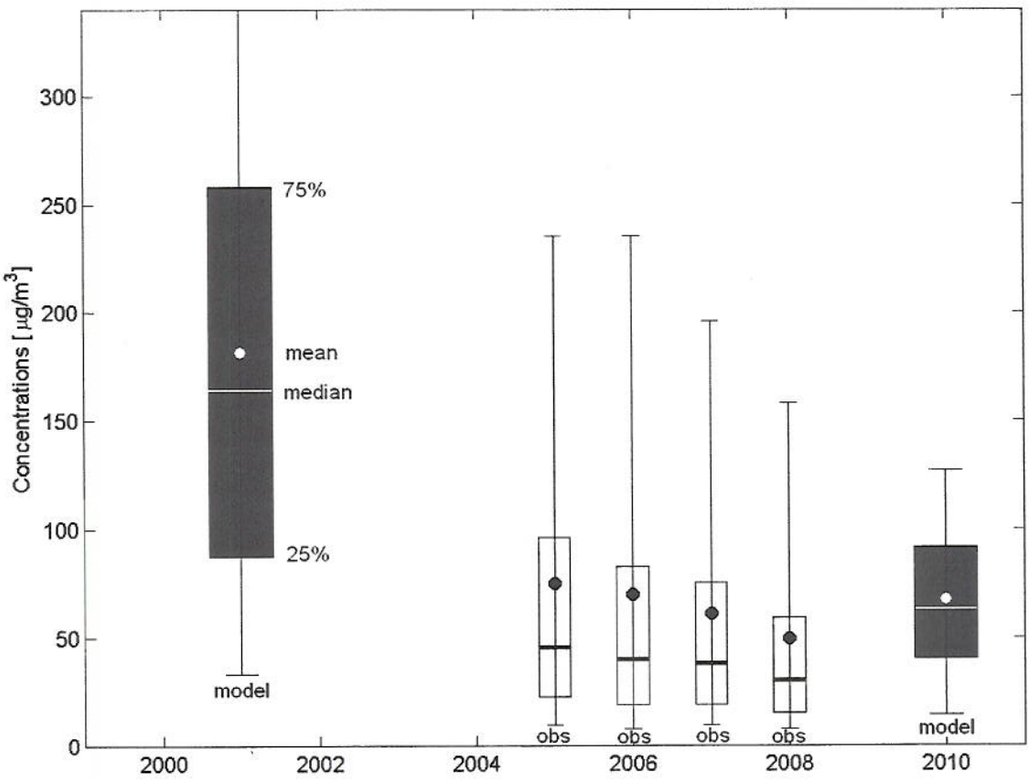


Figure 2. Spatial maps of modeled 2-month average concentrations in New Haven, CT for a) benzene, b) NO_x, and c) $PM_{2.5}$



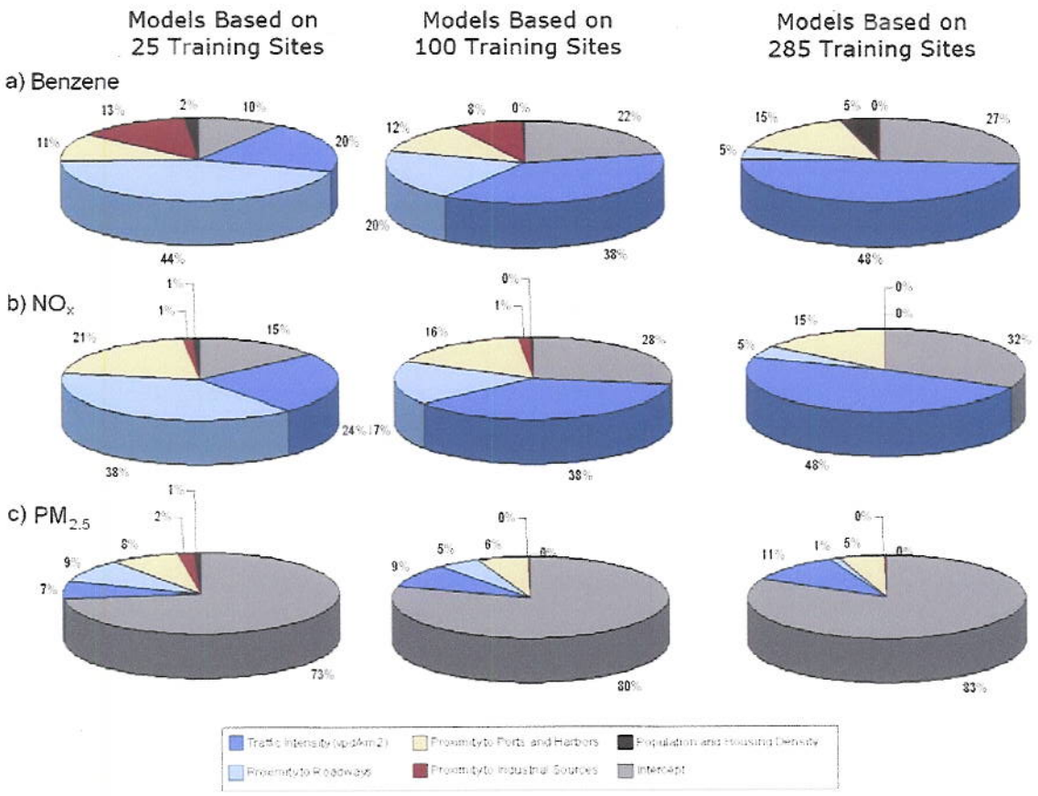
538

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



541

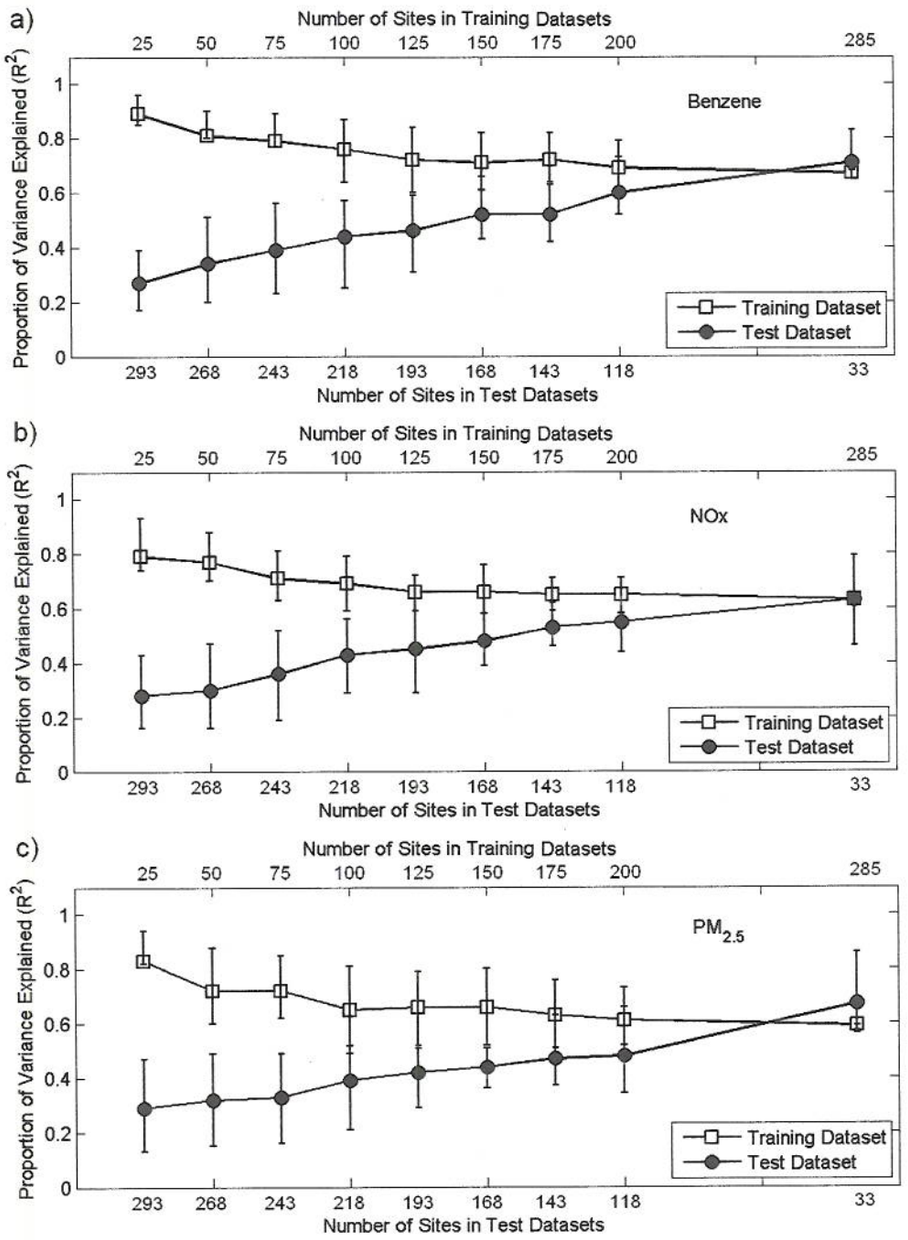
542 Figure 4. Trends of modeled and observed distributions of daily average NO_x concentrations at
543 one monitor.



544

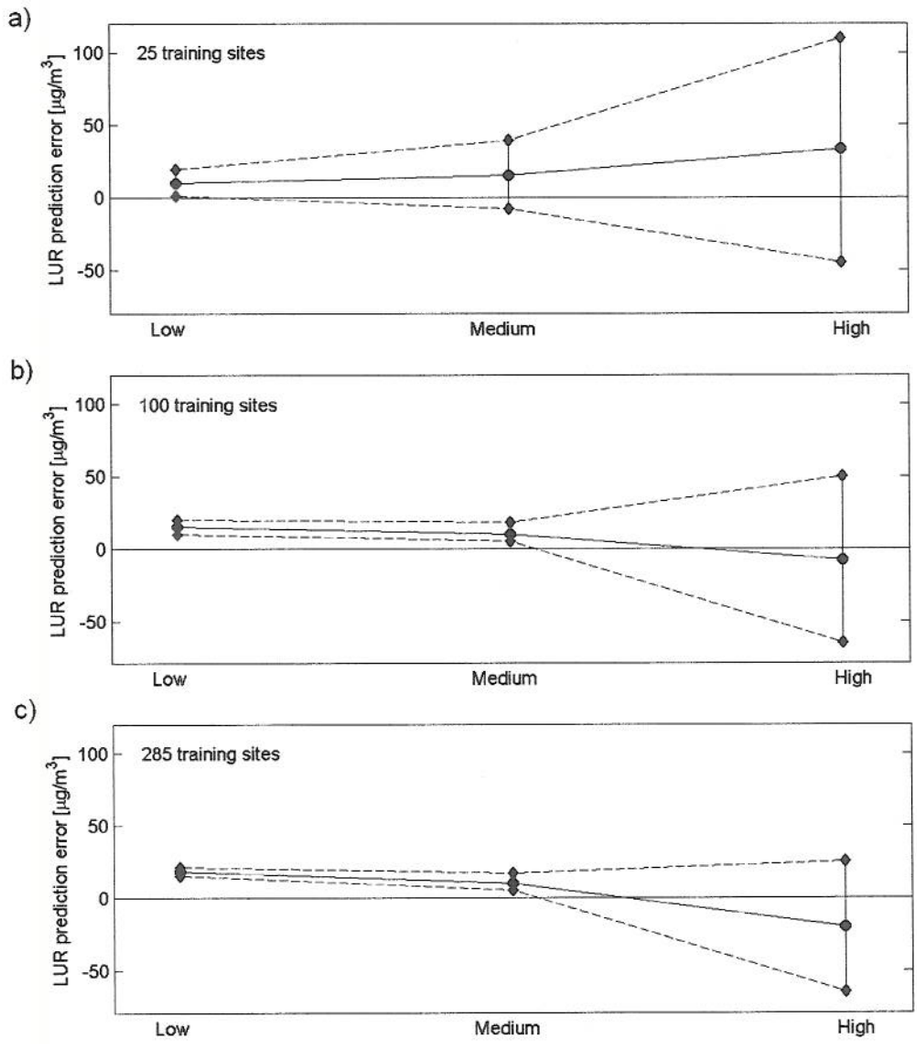
545

546 Figure 5. Percent contribution of land-use predictors by pollutant and number of observations



547

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



551

552 Figure 7. LUR model prediction errors (defined as average +/- standard deviation) for NO_x
553 concentrations as a function of three categories of concentrations (low, medium, high)
554 and the number of training sites: a) 25 sites b) 100 sites and c) 285 sites.

555

556

557 **List of Tables**

560 **Table 1. Summary statistics of pollutant concentrations and land-use variables**

	N	Mean value	Std	CV (%)
Pollutants (μg/m3)				
Average summer Benzene	318	1.1	1.0	92
Average summer NO_x	318	95.5	85.7	90
Average summer $PM_{2.5}$	318	15.4	3.4	22
Traffic Intensity (vehicles per day /km2)				
Within 125 meters	318	26200	91500	349
Within 250 meters	318	37800	101000	267
Within 500 meters	318	45000	81100	180
Within 1000 meters	318	49600	55200	111
Within 1500 meters	318	49400	46100	93
Within 2000 meters	318	49500	39900	81
Population and Housing Density				
Population density within census block group	318	2680	2780	104
Housing density within census block group	318	1200	1590	132
Proximity to Roadways (1/km)				
Roads with ≥ 10,000 vehicles per day (vpd)	318	1.17	5.68	484
Roads with ≥ 20,000 vpd	318	0.39	1.99	509
Roads with ≥ 30,000 vpd	318	0.31	1.98	643
Roads with ≥ 40,000 vpd	318	0.20	0.75	376
Roads with ≥ 60,000 vpd	318	0.17	0.73	433
Roads with ≥ 70,000 vpd	318	0.16	0.73	464
Roads with ≥ 90,000 vpd	318	0.16	0.73	464
Proximity to Ports and Harbors(1/km)				
Seaport	318	0.26	0.21	81
Harbor	318	0.63	1.15	185
Proximity to Industrial Benzene Sources (1/km)				
Facilities emitting ≥ 100 lbs/year	318	0.34	0.70	208
Facilities emitting < 100 lbs/year	318	0.92	1.33	145
Proximity to Industrial NO_x Sources (1/km)				
Facilities emitting ≥ 100 lbs/year	318	1.06	1.27	119
Facilities emitting < 100 lbs/year	318	0.68	1.03	151
Proximity to Industrial $PM_{2.5}$ Sources (1/km)				
Facilities emitting ≥ 100 lbs/year	318	0.83	1.16	140
Facilities emitting < 100 lbs/year	318	0.95	1.39	146

561

562 **Table 2-a. Summary of LUR model evaluation for benzene.**

	Number of Training Sites								
	25	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**285**
Benzene	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)
LUR Model Performance									
Adjusted Model R^2 (training sites)	0.89 (0.45-0.98)	0.81 (0.17-0.97)	0.79 (0.30-0.95)	0.76 (0.42-0.93)	0.72 (0.35-0.92)	0.71 (0.44-0.90)	0.72 (0.48-0.88)	0.69 (0.50-0.88)	0.67 (0.58-0.82)
Cross-Validation Results									
Mean SPR	-0.33 (-11.10-1.08)	-0.08 (-3.96-0.39)	-0.12 (-2.43-0.24)	-0.10 (-1.78-0.12)	-0.02 (-0.81-0.11)	-0.03 (-0.92-0.08)	-0.06 (-0.66-0.06)	-0.02 (-0.54-0.06)	0.00 (-0.07-0.04)
RMS of SPR	4.13 (1.09-55.08)	2.53 (1.09-28.39)	2.67 (1.07-21.40)	2.57 (1.08-17.93)	1.89 (1.05-9.99)	1.84 (1.04-12.25)	2.04 (1.07-8.13)	1.74 (1.07-7.66)	1.34 (1.15-1.90)
Hold-Out Evaluation Results									
Adjusted Model R^2 (test sites)	0.27 (0.02-0.67)	0.34 (0.02-0.74)	0.39 (0.06-0.78)	0.44 (0.02-0.77)	0.46 (0.04-0.79)	0.52 (0.05-0.78)	0.52 (0.11-0.83)	0.60 (0.15-0.85)	0.71 (0.01-0.94)
Mean Residuals	0.32 (-0.18-2.01)	0.27 (-0.15-1.80)	0.16 (-0.21-0.94)	0.13 (-0.20-1.36)	0.13 (-0.15-1.40)	0.09 (-0.12-0.64)	0.04 (-0.18-0.68)	0.04 (-0.24-0.82)	0.03 (-0.47-0.25)

563

564 **Table 2-b. Summary of LUR model evaluation for NO_x**

	Number of Training Sites								
	25	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**285**
NO_x	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)
LUR Model Performance									
Adjusted Model R^2 (training sites)	0.79 (0.08-0.99)	0.77 (0.25-0.97)	0.71 (0.30-0.92)	0.69 (0.32-0.89)	0.66 (0.41-0.90)	0.66 (0.40-0.85)	0.65 (0.43-0.83)	0.65 (0.45-0.83)	0.63 (0.54-0.74)
Cross-Validation Results									
Mean SPR	0.00 (-6.01-1.22)	-0.10 (-6.43-0.44)	-0.09 (-1.69-0.16)	-0.05 (-1.14-0.11)	-0.06 (-0.86-0.08)	-0.03 (-0.73-0.07)	-0.02 (-0.47-0.06)	-0.01 (-0.22-0.04)	0.00 (-0.08-0.02)
RMS of SPR	2.85 (1.10-36.58)	2.81 (1.11-45.53)	2.42 (1.11-16.54)	1.99 (1.10-10.20)	2.02 (1.10-9.93)	1.72 (1.08-9.17)	1.50 (1.08-6.66)	1.36 (1.07-3.58)	1.18 (1.07-1.92)
Hold-Out Evaluation Results									
Adjusted Model R^2 (test sites)	0.28 (0.02-0.59)	0.30 (0.02-0.63)	0.36 (0.05-0.67)	0.43 (0.08-0.69)	0.45 (0.03-0.71)	0.48 (0.06-0.78)	0.53 (0.14-0.78)	0.55 (0.16-0.83)	0.63 (0.18-0.93)
Mean Residuals	18.68 (-19.33-111.17)	14.40 (-18.39-104.00)	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)

565 **Table 2-c. Summary of LUR model evaluation for $PM_{2.5}$**

	Number of Training Sites								
	25	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**285**
NO_x	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)	Mean (Range)
LUR Model Performance									
Adjusted Model R^2 (training sites)	0.83 (0.20-0.99)	0.72 (0.16-0.97)	0.72 (0.23-0.94)	0.65 (0.27-0.94)	0.66 (0.28-0.90)	0.66 (0.41-0.89)	0.63 (0.38-0.88)	0.61 (0.37-0.82)	0.59 (0.47-0.79)
Cross-Validation Results									
Mean SPR	-0.06 (-4.57-1.55)	-0.06 (-4.07-0.39)	0.00 (-1.58-0.27)	0.01 (-0.53-0.16)	0.02 (-0.27-0.11)	0.00 (-0.41-0.09)	0.01 (-0.25-0.07)	0.01 (-0.16-0.07)	0.01 (-0.05-0.03)
RMS of SPR	3.11 (1.24-23.62)	2.42 (1.15-28.25)	1.89 (1.11-13.63)	1.91 (1.09-5.73)	1.49 (1.12-3.36)	1.64 (1.11-6.46)	1.52 (1.08-3.55)	1.52 (1.07-2.79)	1.34 (1.07-1.81)
Hold-Out Evaluation Results									
Adjusted Model R^2 (test sites)	0.29 (0.01-0.73)	0.32 (0.02-0.64)	0.33 (0.02-0.68)	0.38 (0.02-0.79)	0.42 (0.03-0.79)	0.44 (0.02-0.73)	0.47 (0.05-0.83)	0.48 (0.04-0.88)	0.67 (0.05-0.95)
Mean Residuals	0.77 (-0.78-5.61)	0.66 (-0.61-5.30)	0.48 (-0.53-5.83)	0.53 (-0.60-5.69)	0.29 (-0.57-2.35)	0.15 (-0.71-4.72)	0.16 (-0.54-1.94)	0.20 (-0.71-2.16)	-0.05 (-1.24-0.77)

566