Environmental Science & Technology

Article

pubs.acs.org/est

Evaluation of Land Use Regression Models for NO_2 and Particulate Matter in 20 European Study Areas: The ESCAPE Project

Meng Wang,[†,*] Rob Beelen,[†] Xavier Basagana,[‡,§,∥] Thomas Becker,[⊥] Giulia Cesaroni,[#] Kees de Hoogh,[▽] Audrius Dedele,[○] Christophe Declercq,[◆] Konstantina Dimakopoulou,[¶] Marloes Eeftens,[†] Francesco Forastiere,[#] Claudia Galassi,[⋈] Regina Gražulevičienė,[○] Barbara Hoffmann,[☼,⊕] Joachim Heinrich,[▰] Minas Iakovides,[☆] Nino Künzli,[△,1] Michal Korek,[Я] Sarah Lindley,[♦] Anna Mölter,[¥] Gioia Mosler,[▽] Christian Madsen,[ϕ] Mark Nieuwenhuijsen,[‡,§,∥] Harish Phuleria,[△,1] Xanthi Pedeli,[¶] Ole Raaschou-Nielsen,[∃] Andrea Ranzi,[⋈] Euripides Stephanou,[☆] Dorothee Sugiri,[☼] Morgane Stempfelet,[◆] Ming-Yi Tsai,[△,1,∞] Timo Lanki,[⇔] Orsolya Udvardy,[≫] Mihály J. Varró,[≫] Kathrin Wolf,[▰] Gudrun Weinmayr,[☼,√] Tarja Yli-Tuomi,[⇔] Gerard Hoek,[†] and Bert Brunekreef[†,∂]

[†]Institute for Risk Assessment Sciences, Utrecht University, P.O. Box 80178, 3508 TD Utrecht, The Netherlands

[‡]Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

[§]IMIM (Hospital del Mar Research Institute), Barcelona, Spain

[∥]CIBER Epidemiología y Salud Pública (CIBERESP), Spain

[⊥]Department of Environmental Science, Aarhus University, Denmark

[#]Epidemiology Department, Lazio Regional Health Service, Rome, Italy

[▽]MRC-HPA Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom

[○]Vytautas Magnus University, Kaunas, Lithuania

[◆]French Institute for Public Health Surveillance, Saint-Maurice, France

[¶]Department of Hygiene, Epidemiology & Medical Statistics, National and Kapodistrian University of Athens, Medical School, Athens, Greece

[⋈]Regional Reference Centre on Environment and Health, ARPA Emilia Romagna, Modena, Italy

[☼]IUF Leibniz Research Institute for Environmental Medicine, University of Düsseldorf, Düsseldorf, Germany

[⊕]Medical Faculty, Heinrich-Heine University of Düsseldorf, Düsseldorf, Germany

[▰]Helmholtz Zentrum München, German Research Center for Environmental Health, Institutes of Epidemiology I and II, Neuherberg, Germany

[☆]Environmental Chemical Processes Laboratory, Department of Chemistry, University of Crete, Heraklion, Greece

[△]Department of Epidemiology and Public Health, Swiss Tropical & Public Health Institute, Basel, Switzerland

[1]University of Basel, Basel, Switzerland

[Я]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

[♦]School of Environment and Development (Geography), The University of Manchester, Manchester, England, United Kingdom

[¥]Centre for Occupational and Environmental Health, The University of Manchester, Manchester, England, United Kingdom

[ϕ]Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway

[∃]Danish Cancer Society Research Center, Copenhagen, Denmark

[∞]Department of Environmental & Occupational Health Sciences, University of Washington, Seattle, Washington 98195, United States

[⇔]National Institute for Health and Welfare, Department of Environmental Health, Finland

[≫]Department of Air Hygiene, National Institute of Environmental Health, Budapest, Hungary

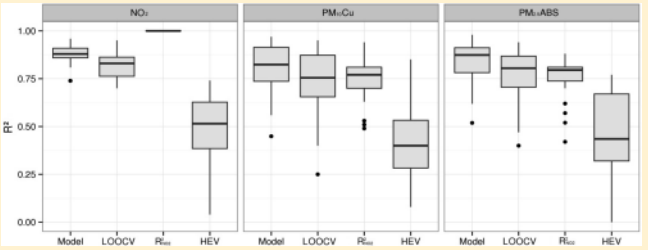
[√]Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany

[∂]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Received: December 14, 2012
Revised: March 21, 2013
Accepted: March 27, 2013
Published: March 27, 2013

Supporting Information

ABSTRACT: Land use regression models (LUR) frequently use leave-one-out-cross-validation (LOOCV) to assess model fit, but recent studies suggested that this may overestimate predictive ability in independent data sets. Our aim was to evaluate LUR models for nitrogen dioxide (NO_2) and particulate matter (PM) components exploiting the high correlation between concentrations of PM metrics and NO_2. LUR models have been developed for NO_2, $PM_{2.5}$ absorbance, and copper (Cu) in PM_{10} based on 20 sites in each of the 20 study areas of the ESCAPE project. Models were evaluated with LOOCV and "hold-out evaluation (HEV)" using the correlation of predicted NO_2 or PM concentrations with measured NO_2 concentrations at the 20 additional NO_2 sites in each area. For NO_2, $PM_{2.5}$ absorbance and PM_{10} Cu, the median LOOCV R^2s were 0.83, 0.81, and 0.76 whereas the median HEV R^2 were 0.52, 0.44, and 0.40. There was a positive association between the LOOCV R^2 and HEV R^2 for $PM_{2.5}$ absorbance and PM_{10} Cu. Our results confirm that the predictive ability of LUR models based on relatively small training sets is overestimated by the LOOCV R^2s. Nevertheless, in most areas LUR models still explained a substantial fraction of the variation of concentrations measured at independent sites.



INTRODUCTION

Epidemiological studies have suggested that long-term exposure to air pollution is associated with adverse health effects.[1−3] Some of these studies have relied on estimating air pollution concentrations at the home addresses of study participants using land use regression methods.[4,5] Within the ESCAPE (European Study of Cohort for Air Pollution Effects) project, a comprehensive measurement program was conducted in 36 European study areas between 2008 and 2011. Substantial spatial variability of nitrogen oxide (NO_2, NO_x) and particulate matter (PM) was identified within and between these areas.[6,7] To explain and predict within-area variability, land use regression (LUR) models were developed using a standardized approach.[8]

LUR modeling is a geographic information system (GIS) and statistics based method that exploits land use, geographic and traffic characteristics (e.g., traffic intensity, road length, population density) to explain spatial concentration variations at measured sites.[9] Within the ESCAPE project, PM and NO_2/NO_x models have been developed in 20 and 36 study areas, respectively, using a standardized method.[8,10] These models explained a large fraction of spatial variance in the measured pollution concentrations, as measured by R^2s ranging from 55−95% for NO_2 and for $PM_{2.5}$ absorbance.

Model evaluation is essential as the model R^2 may be artificially high.[11] Two common evaluation approaches are the internal "leave-one-out-cross-validation (LOOCV)" and the external "hold-out-evaluation (HEV)" against independent measurements set aside for model evaluation. The HEV is preferable as it likely better reflects the predictive power of the model at locations where no measurements were taken, such as addresses of subjects in an epidemiological study, assuming that validation sites are representative of the distribution of subject's addresses. In a study with 144 NO_2 monitoring sites, we previously reported that the model adjusted R^2 decreased slightly from 0.87 to 0.82 with the increasing size of the training sets used for model development. In contrast, the HEV R^2 increased from 0.60 to 0.74 with training set size from 24 to 120.[12] This is likely due to some overfitting.[11] Similar evaluations have been conducted in Girona, Spain and in Oslo, Norway with somewhat different results: in Girona, differences between LOOCV R^2 and HEV R^2 were larger than we found previously, in Oslo they were smaller.[13,14]

All these studies of LUR model performance evaluation were conducted for NO_2. Sampling of PM requires more effort and usually the number of sampling sites is not sufficient to allow for a separation into training and test data set (for validation purpose) of sufficient size. To the best of our knowledge, no evaluations have been conducted for particulate matter LUR models.

Within the ESCAPE study area specific PM models were developed based on 20 training sites per area in most of the study areas.[8] In view of the recent model evaluation studies which were restricted to single areas, the goal of this paper is to evaluate model performance in all 20 ESCAPE study areas for spatial variation of PM and NO_2.

MATERIALS AND METHODS

Study Design. ESCAPE study areas included 20 sites with simultaneous measurements of both PM and NO_2, and 20 sites where only NO_2 was measured in each area. As we did not have PM concentration data available for sampling sites other than the 20 PM sites in each area, we made use of the high correlation between the annual average concentration of traffic-related PM metrics such as $PM_{2.5}$ absorbance, copper in PM_{10} (PM_{10} Cu), and NO_2[7].We assessed the performances of LUR models developed using the PM/NO_2 sites to predict the NO_2 concentrations at the sites where only NO_2 was measured. We used this as a surrogate for the true hold-out validation. In the paper we will refer to the PM/NO_2 sites as training sites and the NO_2 only sites as test sites.

Study Areas and Air Pollution Measurements. Details of the ESCAPE study design and the measurement campaign have been described previously.[6,7] Briefly, an intensive monitoring campaign was conducted in 20 European study areas between October 2008 and May 2011. The abbreviations regarding to the study areas are shown in Supporting Information (SI) Table S1. In each area, we chose sampling sites at street, urban background, and regional background locations. These sites were selected to represent the spatial distribution of residential addresses of participants of cohort studies in these areas. Sampling of NO_2 was conducted at 40 sites, at half of which we also sampled PM. In

The Netherlands/Belgium and Cataluña measurements were performed at 40 PM sites and 80 NO_2 sites. At each of the PM sites, NO_2 was measured simultaneously. The site selection procedure (http://www.escapeproject.eu/manuals/index.php) specified that the 20 PM sites had to be a random selection of the 40 sites in each area. This was not always achieved as it is easier to find monitoring locations for the passive NO_2 sampler than for the active PM samplers. We compared the distributions of NO_2 concentrations measured at the sites where only nitrogen oxides were measured, to those at the sites where both nitrogen oxides and PM were measured. Each selected site was measured in three two-week sampling periods in the cold, warm and intermediate seasons. Due to limited amount of samplers, five sites and the reference site were measured simultaneously. The measured values were adjusted for temporal variation using continuous measurements at a background location which was not influenced by local pollution and annual average concentrations for each site were calculated and were used for model development.

NO_2 was measured using Ogawa badges and following the Ogawa analysis protocol (Ogawa & Co V 3.98, USA, Inc.). $PM_{2.5}$ and PM_{10} samples were collected on preweighted filters using Harvard Impactors. These filters were then used to measure absorbance and detect elemental composition (e.g., Cu) by energy dispersive X-ray fluorescence (ED-XRF) at Cooper Environmental Services (Portland, OR). More detail is provided in a separate paper (de Hoogh, in preparation). Briefly, 48 elements were measured. Quality assurance and control included analysis of NIST reference material(SRM 1128 and SRM987). All analysis batches passed quality criteria of the laboratory. In each study area, about 20 field blanks and field duplicates were taken. We calculated the mean field blank and the detection limit.

Predictor Variables for LUR Model. We extracted values for the GIS predictor variables at the coordinates of sampling sites using ArcGIS (ESRI, Redlands, California). Details of the predictor variables have been described in previous papers.[8,10] Briefly, the predictor variables were derived from both centrally available Europe-wide GIS database and locally collected GIS data from partners.

Central GIS predictor variables were comprised of road network, land use, population density and altitude data. High resolution digital road network was obtained from Eurostreets version 3.1(1:10 000 resolution) which were based on the TeleAtlas MultiNetTM data set for the year 2008. For all roads and major roads, the total lengths of roads were calculated within a buffer size of 25, 50, 100, 300, 500, 1000 m. Land use variables were derived from the CORINE (coordination and information on the environmental program) database for the year 2000 for the buffer sizes of 100, 300, 500, 1000, and 5000 m. Digital elevation data (SRTM 90m) were obtained through the Shuttle Radar Topographic Mission (http://srtm.csi.cgiar.org/).

Detailed road network with linked traffic intensity were available locally for most study areas. The accuracy should be at least 10m compared to the central road network. Data on traffic density were aggregated to annual means, as we were modeling annual mean concentrations. We did not obtain traffic counts for the exact monitoring hours as these traffic data were generally not available. Local land use, population density, altitude, and other local variables were also extracted for modeling.

LUR Model Development. Models for $PM_{2.5}$ absorbance were developed by local partners supervised centrally while models for PM_{10} Cu were built centrally at IRAS (Institute for Risk Assessment Sciences, Utrecht University). Separate models were built for each area, we did not attempt to build a universal model to cover all study areas in view of differences between areas not sufficiently characterized by the available GIS data. For this paper we further developed models for NO_2, using only the data from the training sites. Detailed procedures of model development and results have been published elsewhere.[8,10] LUR model results for elemental composition will be published later. A supervised stepwise regression was used to develop the LUR model. We first evaluated univariate regression of the corrected annual concentrations by entering all potential predictor variables. The variable producing the highest adjusted R^2 and having the a priori defined direction of effect (e.g., positive for traffic intensity) was selected as the first predictor. Second, the remaining variables were added separately and we assessed whether the variable with the highest increase in adjusted R^2 improved the model by at least 1%. This process continued until no more variable with the a priori specified sign could increase the model adjusted R^2 by at least 1%. In the final step, we excluded the variables which had a p value >0.1. We checked whether the variance inflation factor (VIF) was lower than 3 in order to avoid multicollinearity.

Model Evaluation. As previously described,[12] we performed two evaluation approaches:

1. Leave-one-out-cross-validation (LOOCV), which successively left out one site from the training data set and estimated models based on the remaining N-1 sites. In this procedure, the variables in the model were the same as identified using the full training data set; only the coefficients of the model changed.
2. Hold-out evaluation (HEV). For NO_2 this was straightforward as we compared NO_2 model predictions with measured NO_2 concentrations at test sites. True HEV for PM components was infeasible as training sets for PM were too small to split up for model building and validation. As an alternative, we evaluated PM models by investigating the correlations between the predicted values of PM metrics and the measured NO_2 at the test sets (HEV R^2).

A systematic check of the model evaluations was conducted in the following ways:

1. We restricted this analysis to PM components and areas with high correlations with measured NO_2 (squared Pearson correlation coefficient $R^2 > 0.5$).
2. We further evaluated whether the PM models could also fit NO_2 well by checking the correlations between predicted PM concentrations and measured NO_2 concentrations (R^2_{NO2}) at the training sets and included only areas where R^2_{NO2} was >0.5.
3. Finally, we compared the variability and tested the distributions of NO_2 in the training and the test sets of each area by simple boxplot and t tests to assess similarity of the two types of sites.

We compared the model performances of the PM metrics with the model performance of the NO_2 models, the latter reflecting true HEV. We evaluated the accuracy of the HEV only for the NO_2 model by calculating the root mean squared error (RMSE) and the mean difference between predictions and observations (MD) as the HEV for $PM_{2.5}$ absorbance and PM_{10} Cu was indirect. As a check of our approaching using correlation with NO_2 as surrogate for HEV, we made use of

two larger areas (The Netherlands & Belgium and Cataluña) with 40 PM sites. Ten data sets were randomly generated for model development (n = 20) and evaluation (n = 20) for $PM_{2.5}$ absorbance and PM_{10} Cu. We compared the indirect HEV R^2 (based on correlation with NO_2) with true HEV in these two areas.

We calculated the HEV R^2 by truncating the values of predictors in the test data sets that were outside the range of the values observed in the data set for model development. This is standard procedure within ESCAPE for exposure assignment and was done to prevent unrealistic predictions based on model extrapolations. Our previous study showed that with a small amount of locations for model building, the range of the variables for the model development may not cover the whole range when they were extended to larger numbers of independent test sites. Therefore, the predicted values may strongly deviate from the observations, especially when nonlinear functions are used such as 1/(distance to road).[12] We explored the impact of truncation on HEV R^2. Analyses were conducted with SAS 9.2.

RESULTS

Table 1 shows the squared Pearson correlation coefficients between NO_2 and selected PM components. Median correlations were high for both $PM_{2.5}$ absorbance and PM_{10} Cu. Substantial variability of correlations was found between study areas. For $PM_{2.5}$ absorbance, the R^2 with NO_2 in all the ESCAPE study areas were higher than 0.5. For PM_{10} Cu, Gyor was the only area with low correlation with NO_2. The highest correlations between NO_2 and PM components were frequently observed in big cities, for example, Munich (Germany), London/Oxford (United Kingdom), Barcelona (Spain) and Paris (France) with large spatial concentration contrasts compared with relatively small cities with smaller spatial contrast, for example, Gyor (Hungry) and Kaunas (Lithuania).[7]

Table 1. Squared Pearson Correlation Coefficients (R^2) between Measured NO_2 and $PM_{2.5}$ Absorbance and PM_{10} Cu in 20 European Study Areas

study areas	$PM_{2.5}$ absorbance	PM_{10} Cu
Oslo, Norway	0.75	0.73
Stockholm, Sweden	0.86	0.64
Helsinki, Turku, Finland	0.81	0.91
Copenhagen, Denmark	0.86	0.84
Kaunas, Lithuania	0.55	0.69
Manchester, UK	0.74	0.76
London, Oxford, UK	0.88	0.89
Netherlands & Belgium	0.86	0.83
Ruhr area, Germany	0.89	0.91
Munich, Germany	0.87	0.94
Vorarlberg, Austria	0.59	0.70
Paris, France	0.90	0.89
Györ, Hungary	0.65	0.25
Lugano, Switzerland	0.64	0.85
Turin, Italy	0.87	0.81
Rome, Italy	0.89	0.77
Barcelona, Spain	0.91	0.87
Catalunya, Spain	0.89	0.83
Athens, Greece	0.85	0.78
Herakion, Greece	0.63	0.66
Oslo, Norway	0.75	0.73
Median	0.86	0.82
Interquartile range	0.19	0.17

The variability of NO_2 concentrations was similar for the training sites and the test sites for most areas (Figure 1). The mean NO_2 concentration did not differ significantly between the training and the test sites, with the exception of the study areas of Paris, Heraklion, Turin, Ruhr area, Oslo, and Stockholm county ($p < 0.05$). Table 2 shows the distributions of model R^2 and LOOCV R^2 for NO_2, $PM_{2.5}$ absorbance and PM_{10}Cu and R^2 between predicted concentrations and measured NO_2 at the training sites (R^2_{NO2}) and test sites (HEV R^2).

Figure 2 and SI Tables S2−S4 show the model performance and structure for all individual study areas, including the predictor variables in the identified LUR models. Vorarlberg and Gyor were excluded from $PM_{2.5}$ absorbance and PM_{10} Cu, respectively, due to lower Correlation R^2 with measured NO_2 or R^2_{NO2} than 0.5. High median model R^2s were observed as 0.82 for PM_{10} Cu, 0.87 for $PM_{2.5}$ absorbance and 0.88 for NO_2. The median LOOCV R^2s were 5−6% lower than the model R^2s. The median correlations (R^2_{NO2}) of the PM model predictions with the measured NO_2 concentrations in the training data sets were as high as the squared correlations (Pearson R^2) between observations (Table 1), ranging from 0.77 for PM_{10} Cu to 0.80 for $PM_{2.5}$ absorbance. In contrast, the models explained substantially less variation in the independent test data sets. The NO_2 models developed on the 20 training sites had the best prediction ability (median HEV R^2 = 0.52). The RMSE and MD ranged from 3.18 to 18.57 $\mu g/m^3$ (median: 6.53 $\mu g/m^3$) and from −8.64 to 2.71 (median: −2.38 μg/m3), respectively. The $PM_{2.5}$ absorbance and PM_{10} Cu models explained only a slightly smaller fraction of the measured NO_2 concentration than the NO_2 models (median HEV R^2 = 0.44 and 0.40, respectively). The IQR of R^2s of each pollutant was higher for hold-out evaluations than for cross-validation and model development, indicating substantial variability of HEV R^2s across study areas.

In the sensitivity analysis with 10 sets of random selected 20 training and test PM sites in The Netherlands and Belgium and in Cataluña, the HEV R^2 validated by the same PM metric did not significantly deviate from the HEV R^2 validated by NO_2 for $PM_{2.5}$ absorbance and PM_{10} Cu (paired t test, $p > 0.1$). This supports our approach of using of NO_2 as proxy to evaluate the PM models (Table 3 and SI Figure S2). Similar differences were found between model R^2 and HEV R^2 for NO_2 in these two areas as in the analysis comprising all study areas.

The HEV was calculated with truncated predictors. We saw that by restricting the predictors in the test sets to the range of values that were obtained in the training sets, improved the median HEV R^2s by 8%, 5% and 8% for NO_2, $PM_{2.5}$ absorbance, and PM_{10} Cu, respectively (SI Table S5).

Figure 3 presents scatterplots of R^2 of LOOCV versus R^2 of HEV in individual areas. In general, there were positive associations between LOOCV R^2 and HEV R^2, indicating that better models as judged from LOOCV were on average better in HEV as well. The correlations were significant ($p < 0.1$) for PM_{10} Cu and $PM_{2.5}$ absorbance, but not for NO_2 ($p \geq 0.1$). There was however a wide scatter. In some areas, models that exhibited very stable performances in cross-validation reflected much lower HEV R^2s than the model R^2. For instance, the models of NO_2 and PM_{10} Cu in Turin have both high model R^2 (>0.87) and high LOOCV R^2 (>0.82), whereas the HEV R^2 dropped dramatically by over 66% from model R^2 (SI Tables S2,S4).

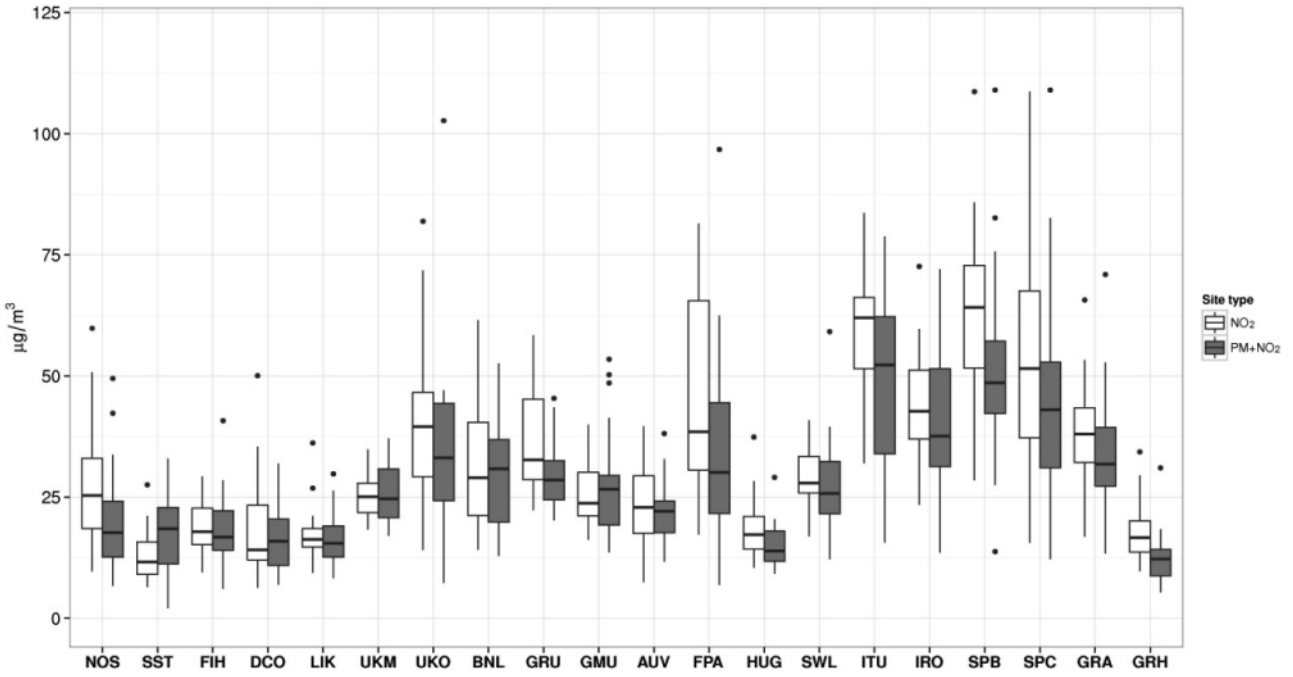


Figure 1. Boxplot of NO_2 concentrations at PM/NO_2 (training) sites and NO_2-only (test) sites in 20 ESCAPE study areas. The upper, middle and bottom layers of the box show the 75, 50, 25th percentiles of the data set. NOS: Oslo, Norway; SST: Stockholm, Sweden; FIH: Helsinki, Finland; DCO: Copenhagen, Denmark; LIK: Kaunas, Lithuania; UKM: Manchester, UK; UKO: London/Oxford, UK; BNL:Netherlands and Belgium; GRU: Ruhr area, Germany; GMU: Munich, Germany; AUV: Vorarlberg, Austria; FPA: Paris, France; HUG: Gyor, Hungry; SWL: Lugano, Switzerland; ITU: Turin, Italy; IRO: Rome, Italy; SPB: Barcelona, Spain; SPC: Cataluña, Spain; GRA: Athens, Greece; GRH: Heraklion, Greece.

Table 2. Comparison between model R^2 and LOOCV R^2 for NO_2 and PM Components (Training Sites), R^2 between Predicted Concentrations and Measured NO_2 at Training Sites (R^2_{NO2}) and R^2 between Predicted Concentrations and Measured NO_2 at Test Sites (HEV R^2) in 20 European Study Areas

	[a]model R^2		[b]LOOCV R^2		[c]R^2_{NO2}		[d]HEV R^2	
modeled pollutant	median	IQR	Median	IQR	Median	IQR	median	IQR
NO_2	0.88	0.05	0.83	0.10	1.00	0.00	0.52	0.24
$PM_{2.5}$ absorbance	0.87	0.13	0.81	0.16	0.80	0.07	0.44	0.35
PM_{10}Cu	0.82	0.18	0.76	0.22	0.77	0.11	0.40	0.25

[a]Model R^2: Model adjusted R^2. [b]LOOCV R^2: Leave-One-Out-Cross-validation R^2. [c]R^2_{NO2} shows the correlations between predicted NO_2 or PM components concentrations with measured NO_2 concentrations at the training sites, being the NO_2/PM sites. [d]HEV R^2 is hold-out evaluation R^2, approximated by the correlation of model predictions with measured NO_2 at test sites, which is NO_2-only sites.

This also applies to the models in a few other areas, for example, Paris, Kaunas, Heraklion, and Athens (SI Tables S2−S4). For NO_2, the five areas with the lowest HEV R^2 (<0.40) were predominantly in southern Europe (Turin, Paris, Athens, Heraklion and Rome). For absorbance, the lowest HEV R^2 (<0.30) were found more spread, specifically in Oslo, Helsinki, Kaunas, Athens, and Heraklion. For Cu, the lowest HEV R^2 (<0.30) were found more spread, specifically in Kaunas, Gyor, Turin, Athens, and Heraklion.

DISCUSSION

This study shows that for a wide range of study areas and pollutants including NO_2, $PM_{2.5}$ absorbance and PM_{10} Cu, model and LOOCV R^2 from land use regression models based on relatively small training sets overestimate predictive ability in independent test sets. Despite this overestimation, in most areas LUR models still explained a substantial fraction of the spatial variation measured at independent sites. The predictions were better for the areas, for example, Western Europe with more detailed predictor variables.

Evaluations of LUR predictive power and the effects of varying the number of sampling sites have been recently reported in four studies conducted in single areas for the pollutant NO_2.[12−15] The conclusions of these studies were variable, ranging from negligibly (LOOCV R^2: 0.67, HEV R^2: 0.64, N = 20)[14] to seriously inflated R^2 of model and LOOCV R^2s compared to HEV R^2 (LOOCV R^2: 0.72, HEV R^2: 0.22, N = 20).[13] Our results for NO_2 can be directly compared with these studies. Our models based on a large multicenter study showed similar patterns as observed in our recent work in The Netherlands only,[12] whereas the studies by Basagana et al. (2012)[13] and Johnson et al. (2010)[16] showed larger gaps between HEV R^2 and model or LOOCV R^2. In our current study the median HEV R^2 was still 52%, indicating that a substantial fraction of the measured variation was explained by the LUR models based upon 20 sites. In our previous work,[12] we found a HEV R^2 of 63% for models based upon 24 sites.

The differences between model R^2 and HEV R^2 for PM absorbance and Cu were evaluated with the NO_2 concentration at the test sites, because independent PM data were not available. The difference between model R^2 and HEV R^2 for

dx.doi.org/10.1021/es305129t | Environ. Sci. Technol. 2013, 47, 4357−4364

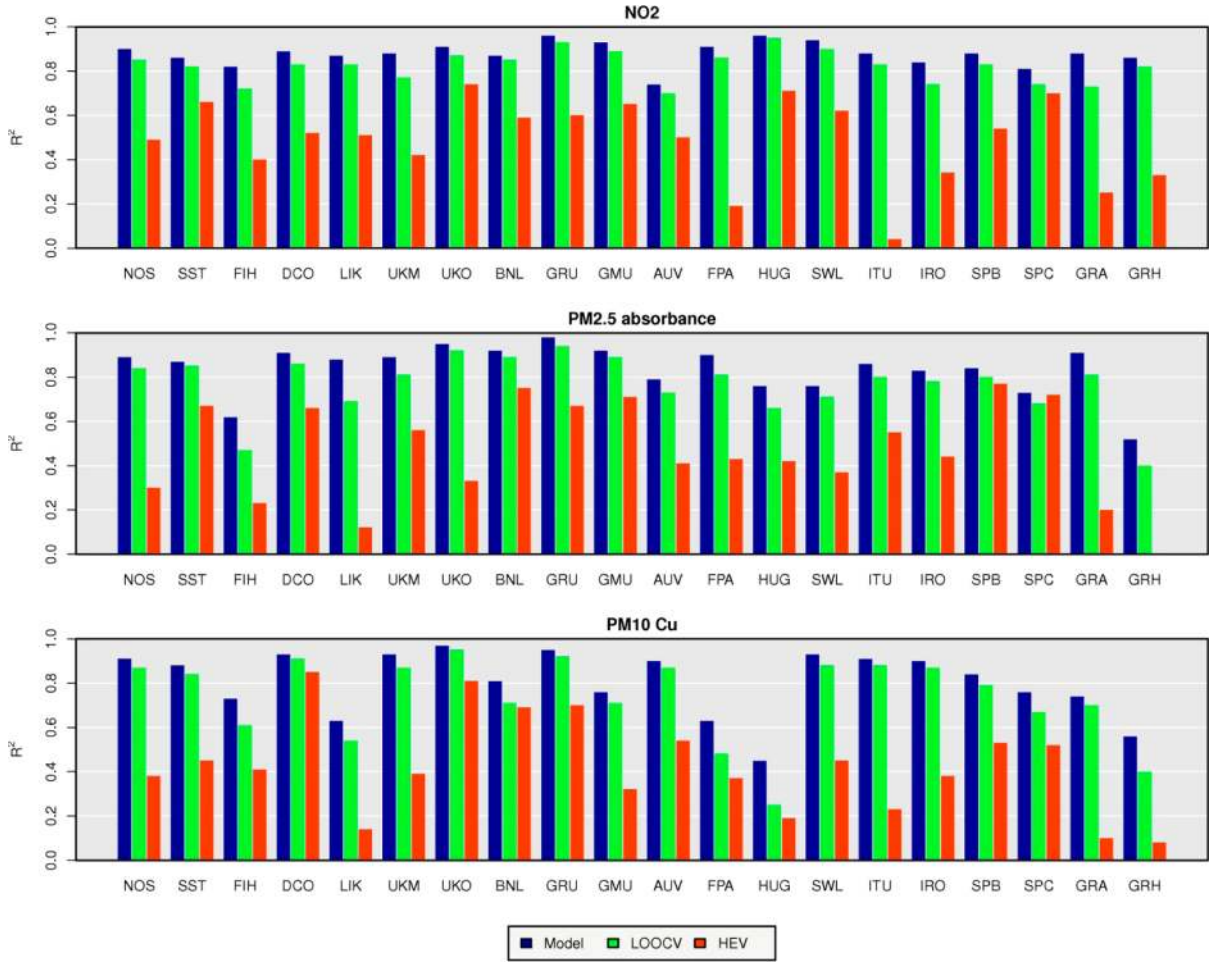


Figure 2. Model R^2; LOOCV R^2 (at NO_2/PM sites) for NO_2 & PM components; and R^2 of model predictions with NO_2 measurements (at NO_2 -only sites, HEV R^2) in 20 European countries. See Figure 1 for coding of the locations.

Table 3. Comparison between Model R^2 and LOOCV R^2 for PM Components and Indirect and Direct Hold out Evaluation in the Netherlands& Belgium and Cataluña[a]

		[b]model R^2		[c]LOOCV R^2		[d]R^2_{NO2}		[e]HEV R^2 (NO_2)		[f]HEV R^2(PM)	
pollutants	areas	median	IQR	median	IQR	median	IQR	median	IQR	median	IQR
$PM_{2.5}$ absorbance	BNL	0.90	0.06	0.87	0.08	0.83	0.04	0.68	0.11	0.76	0.13
	SPC	0.85	0.10	0.81	0.14	0.82	0.10	0.56	0.15	0.51	0.17
PM_{10}Cu	BNL	0.84	0.04	0.79	0.11	0.83	0.07	0.57	0.12	0.56	0.09
	SPC	0.82	0.08	0.77	0.09	0.71	0.10	0.45	0.32	0.45	0.36

[a]BNL: Belgium & the Netherlands; SPC: Cataluña, Spain. The 40 sites were randomly divided in test and training sets 10 times. [b]Model R^2: Model adjusted R^2. [c]LOOCV R^2: Leave-One-Out-Cross-validation R^2. [d]R^2_{NO2} shows the correlations between predicted NO_2 or PM components concentrations with measured NO_2 concentrations at the training sites, being the NO_2/PM sites. [e]Indirect HEV R^2(NO_2): correlations between predicted PM components and measured NO_2 at the 20 test sites. [f]Direct HEV R^2 (PM): correlations between predicted and measured PM components at the 20 test sites.

$PM_{2.5}$ absorbance and PM_{10} Cu was only slightly larger than for NO_2. For these PM metrics some of the gap is due to the use of NO_2 for the evaluation. To test this impact, we divided the HEV R^2 by the R^2_{NO2} in Table 3, which can be interpreted as the highest possible squared correlation for PM metrics. This resulted in median HEV R^2 of 62% and 52% for $PM_{2.5}$ absorbance and PM_{10} Cu, respectively. These adjusted HEVs are still much larger than the LOOCV. These PM metrics have strong relations to tailpipe and nontailpipe traffic emissions.[16,17] We restricted the evaluation to the areas with high correlation of the measured concentrations with NO_2 (Table 1) and high correlations of PM model predictions with NO_2 at the sites used for model development (SI Tables S2, S4) ($R^2 > 0.5$). Our sensitivity analysis indicated that use of NO_2 proxy for HEV showed no significant difference as compared to use of the same PM metrics for true HEV in The Netherlands & Belgium and Cataluña, suggesting that it was reasonable to use NO_2 to evaluate the prediction ability of $PM_{2.5}$ absorbance and PM_{10} Cu models in this study. A limitation of the use of NO_2 for PM metrics evaluation is that we can only evaluate the correlation and not the accuracy of the model. The evaluation of the NO_2

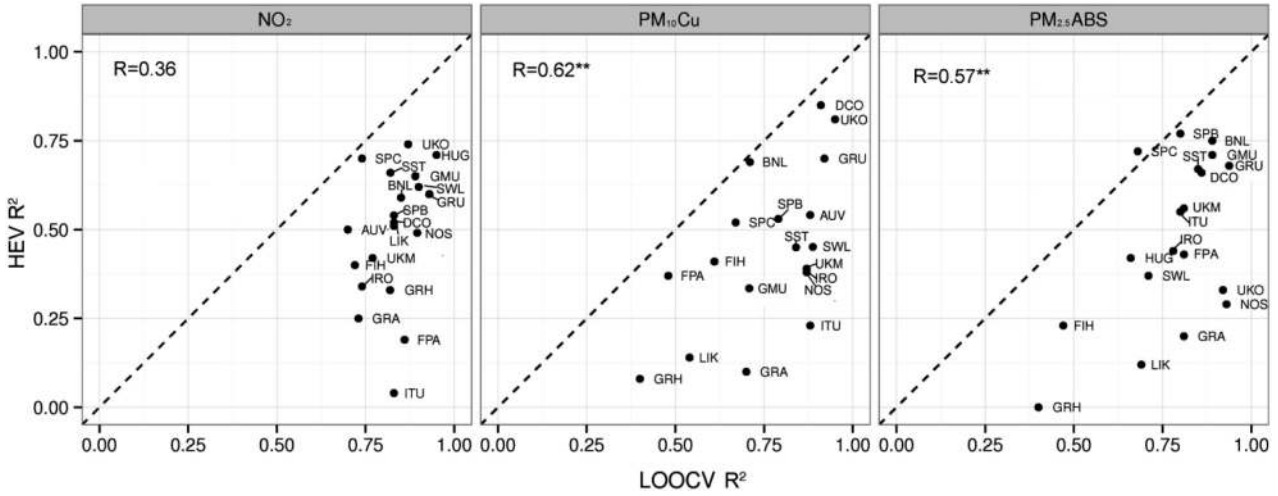


Figure 3. LOOCV R^2 (X-axis) versus HEV R^2 (Y-axis) in study areas. The codes corresponding to the areas are shown in Figure 1 and SI Table S1. R:*$p < 0.1$, **$p < 0.05$.

models suggested that the predictions may slightly underestimate concentrations in most of the study areas.

The differences between model R^2 and HEV R^2 were recognized as a phenomenon of some overfitting, in combination with incomplete representation of relevant area characteristics in small training sets.[11−13,18] The model R^2 and LOOCV R^2 may be inflated when models are based on small number of training sites and when many candidate predictors are available. In the ESCAPE study, we used a supervised approach with a priori defined directions of effects and restricted the potential predictors to limit the risk of overfitting. Our results showed that despite substantial variability of LOOCV R^2 and HEV R^2 in study areas, the areas with higher LOOCV R^2 tended to produce better predictions for the independent data, and therefore suggested more robust performances of models in predicting values at the cohort addresses in some areas.

We also noted that in a few areas, LOOCV R^2 was much lower than HEV R^2. This is likely explained at least in part by simple random variation (associations might have been different in these areas with other training and/or test sets in these same areas). However, the scatterplots in Figure 3 show that LOOCV R^2 and HEV R^2 were positively associated, suggesting that models in some areas were truly more predictive than in other areas. This is supported by SI Figure S1 which shows that the HEV R^2 is positively associated with the correlation between NO_2 and PM component measurements. The level of the HEV R^2 could be related to complexity of study areas and quality of measurements and predictor variables. With more detailed predictor variables, the models in the Western European centers generally performed better than the models in other areas. This suggests a sensitivity analysis in the epidemiological analysis using HEV R^2 rather than LOOCV R^2. Previous studies displayed a slight reduction of NO_2 model R^2s and LOOCV R^2s as a function of increasing number of training sites.[12,13,15] Our results supported this variation in model performances for a large number of areas using a standardized sampling and modeling method. We compared performances between NO_2 models which were centrally built for testing by IRAS based on 20 sites (40 for Netherlands & Belgium and Cataluña) and models which were optimized by local partners based on a full set of 40 sites (80 for Netherlands & Belgium and Cataluña). The median R^2s of model and LOOCV decreased from 0.88 to 0.81 and from 0.83 to 0.73, respectively (SI Figure.S3). The effect of restricting the out-of-range predictor values to the range of the training sets has been discussed elsewhere.[12,13] Our results support that the range truncation approach increases the HEV R^2 of our LUR models in most study areas. It is therefore important that the selected sites cover the variability of predictor variables and pollutant concentrations in the study area well.[8]

As the PM models will be applied to the epidemiological studies in all the ESCAPE study areas, the quality of estimated exposure of cohorts will largely depend on the prediction ability of models to the independent data set, that is, the HEV R^2. Although we cannot directly estimate absolute errors of PM metrics in the test sets, the HEV R^2 with measured NO_2 can still be informative to the health studies. We will, for instance, include model performance in meta-regressions of the cohort-specific effect estimates which are currently being developed. In summary, we found model R^2 and LOOCV R^2 to be substantially higher than HEV R^2 in LUR models developed for $PM_{2.5}$ absorbance and PM_{10} copper in 20 study areas across Europe. Despite this overestimation, in most areas LUR models still explained a substantial fraction of the variation measured at independent sites.

■ ASSOCIATED CONTENT

Supporting Information

Additional information as described in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +31 (0)30 2537517; fax: +31 (0)30 2539499; e-mail: M.Wang@uu.nl.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank all those who participated and were responsible for air pollution measurements, data management, model building and project supervision in all the ESCAPE study areas. The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/

2007-2011) under grant agreement number: 211250. The funding source had no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the manuscript for publication.

REFERENCES

(1) Brunekreef, B.; Holgate, S. T. Air pollution and health. *Lancet* **2002**, *360* (9341), 1233−1242.

(2) Pope, C. A.; Burnett, R. T.; Thun, M. J.; Calle, E. E.; Krewski, D.; Ito, K.; Thurston, G. D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA, J. Am. Med. Assoc.* **2002**, *287* (9), 1132−1141.

(3) Dockery, D. W.; Pope, C. A.; Xu, X. P.; Spengler, J. D.; Ware, J. H.; Fay, M. E.; Ferris, B. G.; Speizer, F. E. An association between air-pollution and mortality in 6 United-States cities. *N. Engl. J. Med.* **1993**, *329* (24), 1753−1759.

(4) Beelen, R.; Hoek, G.; van den Brandt, P. A.; Goldbohm, R. A.; Fischer, P.; Schouten, L. J.; Jerrett, M.; Hughes, E.; Armstrong, B.; Brunekreef, B. Long-term effects of traffic-related air pollution on mortality in a Dutch cohort (NLCS-AIR study). *Environ. Health Perspect.* **2008**, *116* (2), 196−202.

(5) Dijkema, M. B. A.; Mallant, S. F.; Gehring, U.; van den Hurk, K.; Alssema, M.; van Strien, R. T.; Fischer, P. H.; Nijpels, G.; Stehouwer, C. D. A.; Hoek, G.; Dekker, J. M.; Brunekreef, B., Long-term exposure to traffic-related air pollution and type 2 diabetes prevalence in a cross-sectional screening-study in the Netherlands. *Environ Health* **2011**, *10*.

(6) Cyrys, J.; Eeftens, M.; Heinrich, J.; Ampe, C.; Armengaud, A.; Beelen, R.; Bellander, T.; Beregszaszi, T.; Birk, M.; Cesaroni, G.; Cirach, M.; de Hoogh, K.; De Nazelle, A.; de Vocht, F.; Declercq, C.; Dedele, A.; Dimakopoulou, K.; Eriksen, K.; Galassi, C.; Grazuleviciene, R.; Grivas, G.; Gruzieva, O.; Hagenbjörk Gustafsson, A.; Hoffmann, B.; Iakovides, M.; Ineichen, A.; Krämer, U.; Lanki, T.; Lozano, P.; Madsen, C.; Meliefste, K.; Modig, L.; Mölter, A.; Mosler, G.; Nieuwenhuijsen, M. J.; Nonnemacher, M.; Oldenwening, M.; Peters, A.; Pontet, S.; Probst Hensch, N.; Quass, U.; Raaschou Nielsen, O.; Ranzi, A.; Sugiri, D.; Stephanou, E. G.; Taimisto, P.; Tsai, M.; Vaskövi, E.; Villani, S.; Wang, M.; Brunekreef, B.; Hoek, G. Variation of NO_2 and NO_x concentrations between and within 36 European study areas: Results from the ESCAPE study. *Atmos. Environ.* **2012**, *62*, 374−390.

(7) Eeftens, M.; Tsai, M.; Ampe, C.; Anwander, B.; Beelen, R.; Cesaroni, G.; Cirach, M.; Cyrys, J.; de Hoogh, K.; de Nazelle, A.; de Vocht, F.; Declercq, C.; dedele, A.; Eriksen, K. T.; Galassi, C.; Grazuleviciene, R.; Grivas, G.; Heinrich, J.; Hoffmann, B.; Iakovides, M.; Ineichen, A.; Katsouyanni, K.; Korek, M.; Krämer, U.; Kuhlbusch, T.; Lanki, T.; Madsen, C.; Meliefste, K.; Mölter, A.; Mosler, G.; Nieuwenhuijsen, M. J.; Oldenwening, W.; Pennanen, A.; Probst Hensch, N.; Quass, U.; Raaschou Nielsen, O.; Ranzi, A.; Stephanou, E. G.; Sugiri, D.; Udvardy, O.; Vaskövi, E.; Weinmayr, G.; Brunekreef, B.; Hoek, G. Spatial variation of $PM_{2.5}$, PM_{10}, $PM_{2.5}$ absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO_2 - Results of the ESCAPE project. *Atmos. Environ.* **2012**, *62*, 303−317.

(8) Eeftens, M.; Beelen, R.; de Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dedele, A.; Dons, E.; de Nazelle, A.; Dimakopoulou, K.; Eriksen, K. T.; Falq, G.; Fischer, P.; Galassi, C.; Grazuleviciene, R.; Heinrich, J.; Hoffmann, B.; Jerrett, M.; Keidel, D.; Korek, M.; Lankki, T.; Lindley, S.; Madsen, C.; Molter, A.; Nador, G.; Nieuwenhuijsen, M. J.; Nonnemacher, M.; Pedeli, X.; Raaschou Nielsen, O.; Patelarou, E.; Quass, U.; Ranzi, A.; Schindler, C.; Stempfelet, M.; Stephanou, E. G.; Sugiri, D.; Tsai, M.; Yli-Tuomi, T.; Varro, M. J.; Vienneau, D.; von Klot, S.; Wolf, K.; Brunekreef, B.; Hoek, G. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas; Results of the ESCAPE project. *Environ. Sci. Technol.* **2012**, 11195−11205.

(9) Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **2008**, *42* (33), 7561−7578.

(10) Beelen, R.; Hoek, G.; Vienneau, D.; Eeftens, M.; Dimakopoulou, K.; Pedeli, X.; Tsai, M.; Künzli, N.; Schikowski, T.; Marcon, A.; Eriksen, K.; Raaschou Nielsen, O.; Stephanou, E. G.; Patelarou, E.; Lanki, T.; Yli-Tuomi, T.; Declercq, C.; Falq, G.; Stempfelet, M.; Birk, M.; Cyrys, J.; von Klot, S.; Nádor, G.; János Varró, M.; Dedele, A.; Grazuleviciene, R.; Mölter, A.; Lindley, S.; Madsen, C.; Cesaroni, G.; Ranzi, A.; Badaloni, C.; Hoffmann, B.; Nonnemacher, M.; Krämer, U.; Kuhlbusch, T.; Cirach, M.; de Nazelle, A.; Nieuwenhuijsen, M. J.; Bellander, T.; Korek, M.; Olsson, D.; Strömgren, M.; Dons, E.; Jerrett, M.; Fischer, P.; Wang, M.; Brunekreef, B.; de Hoogh, K. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—The ESCAPE project. *Atmos. Environ.* **2013**, *72*, 10−23.

(11) Babyak, M. A. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosom. Med.* **2004**, *66*, 411−421.

(12) Wang, M.; Beelen, R.; Eeftens, M.; Meliefste, K.; Hoek, G.; Brunekreef, B. Systematic evaluation of land use regression models for NO_2. *Environ. Sci. Technol.* **2012**, *46* (8), 4481−4489.

(13) Basagana, X.; Rivera, M.; Aguilera, I.; Agis, D.; Bouso, L.; Elosua, R.; Foraster, M.; de Nazelle, A.; Nieuwenhuijsen, M.; Vila, J.; Kunzli, N. Effect of the number of measurement sites on land use regression models in estimating local air pollution. *Atmos. Environ.* **2012**, *54*, 634−642.

(14) Madsen, C.; Gehring, U.; Haberg, S. E.; Nafstad, P.; Meliefste, K.; Nystad, W.; Carlsen, K. C. L.; Brunekreef, B. Comparison of land-use regression models for predicting spatial NO_x contrasts over a three year period in Oslo, Norway. *Atmos. Environ.* **2011**, *45* (21), 3576−3583.

(15) Johnson, M.; Isakov, V.; Touma, J. S.; Mukerjee, S.; Ozkaynak, H. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmos. Environ.* **2010**, *44* (30), 3660−3668.

(16) He, L. Y.; Hu, M.; Zhang, Y. H.; Huang, X. F.; Yao, T. T. Fine particle emissions from on-road vehicles in the Zhujiang Tunnel, China. *Environ. Sci. Technol.* **2008**, *42* (12), 4461−4466.

(17) Bukowiecki, N.; Lienemann, P.; Hill, M.; Figi, R.; Richard, A.; Furger, M.; Rickers, K.; Falkenberg, G.; Zhao, Y. J.; Cliff, S. S.; Prevot, A. S. H.; Baltensperger, U.; Buchmann, B.; Gehrig, R. Real-world emission factors for antimony and other brake wear related trace elements: Size-segregated values for light and heavy duty vehicles. *Environ. Sci. Technol.* **2009**, *43* (21), 8072−8078.

(18) Craig, M. H.; Sharp, B. L.; Mabaso, M. L. H.; Kleinschmidt, I., Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. *Int. J. Health Geographics* **2007**, *6*.