Development and evaluation of land use regression models for black carbon based on bicycle and pedestrian measurements in the urban environment

Joris Van den Bossche[a,b,*], Bernard De Baets[b], Jan Verwaeren[b], Dick Botteldooren[c], Jan Theunis[a]

[a] *VITO - Flemish Institute for Technological Research, 2400 Mol, Belgium*
[b] *KERMIT, Dept. of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium*
[c] *INTEC, Dept. of Information Technology, Faculty of Engineering and Architecture, Ghent University, 9000 Ghent, Belgium*

Abstract

Land use regression (LUR) modelling is increasingly used in epidemiological studies to predict air pollution exposure. The use of stationary measurements at a limited number of locations to build a LUR model, however, can lead to an overestimation of its predictive abilities. We use opportunistic mobile monitoring to gather data at a high spatial resolution to build LUR models to predict annual average concentrations of black carbon (BC). The models explain a significant part of the variance in BC concentrations. However, the overall predictive performance remains low, due to input uncertainty and lack of predictive variables that can properly capture the complex characteristics of local concentrations. We stress the importance of using an appropriate cross-validation scheme to estimate the predictive performance of the model. By using independent data for the validation and excluding those data also during variable selection in the model building procedure, overly optimistic performance estimates are avoided.

Keywords: Land use regression, Spatial cross-validation, Mobile measurements, Opportunistic monitoring, Black carbon, Urban air quality

1. Introduction

The urban air quality shows a large spatial variability on a small scale, especially for traffic-related pollutants such as NOx, ultrafine particles (UFP) and black carbon (BC) (Vardoulakis et al., 2011; Peters et al., 2014; Wu et al., 2015). As the variation within a city may exceed the variation between cities (Jerrett et al., 2005; Cyrys et al., 2012), it is important to take this within-city variability into account for accurate exposure estimation in epidemiological studies (Hoek et al., 2008; Fruin et al., 2014). Land use regression (LUR) models intend to model this small-scale within-city variation by relating the air pollution concentration at certain locations with predictor variables, usually obtained through geographic information systems (GIS), holding information on surrounding land use and traffic characteristics (Jerrett et al., 2005; Hoek et al., 2008; Beelen et al., 2013). LUR models are increasingly used in epidemiological studies (Eeftens et al., 2012; Beelen et al., 2014; Dons et al., 2014; de Hoogh et al., 2014).

LUR modelling requires air quality measurements at multiple locations across the study area. Typically, stationary monitoring is used at 20 – 100 locations (Hoek et al., 2008). However, Basagaña et al. (2012) argue that LUR models for complex urban settings should be based on a large number of measurement sites (> 80 in their study). Mobile monitoring can provide an alternative way to gather data at a high spatial resolution (Van den Bossche et al., 2015). Some studies use a mobile platform to perform short-term measurements at many locations (e.g. Larson et al., 2009; Merbitz et al., 2012; Ghassoun et al., 2015; Montagne et al., 2015). Only few studies use mobile measurements as a basis for LUR modelling. For example, Hasenfratz et al. (2015) and Mueller et al. (2016) present a study on the modelling of particle number concentrations in Zurich using data from a tram-based mobile sensor network. Hankey and Marshall (2015) use bicycle-based, mobile measurements to build LUR models, and in studies of Kanaroglou et al. (2013), Patton et al. (2014) and Weichenthal et al. (2016b), van-based measurements are used. Mobile measurements can also be collected in participatory and community-based campaigns. Volunteers can systematically collect targeted data sets, or data are collected opportunistically during (repeated) daily activities or trips, to provide improved estimates of spatial variability (Snyder et al., 2013; Van den Bossche et al., 2016).

In this study, we will investigate the development of LUR models based on opportunistic mobile measurements to predict annual average concentrations at a high spatial resolution in the urban environment. This case study is based on measurements gathered by city wardens during their surveillance tasks, which were presented in Van den Bossche et al. (2016). The measurement campaign resulted in a higher spatial density of measurement locations compared to most LUR studies (sampling points at an approximate resolution of 50 m along the roads). Different techniques to build the LUR models, both linear and non-linear, and different methods to select the relevant predictor variables, will be evaluated. For the evaluation, a custom spatial cross-

*Corresponding author. KERMIT, Dept. of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, 9000 Ghent, Belgium.

Email address: jorisvandenbossche@gmail.com (Joris Van den Bossche)

validation scheme will be used to ensure a proper assessment of the predictive ability of the model.

2. Materials and methods

2.1. Study location and description

The study site is the city of Antwerp, Belgium, a medium-sized city of 480,000 inhabitants (51°12' N, 4°26' E, 985 inhabitants km^{-2}). The inner city (within the ring road) has an area of approximately 16 km^2. The study area where measurements were gathered comprises a quarter of this region (approximately 3.7 km^2), and is shown in Figure 1. This region consists mainly of residential and commercial areas, including main traffic roads and green areas. A highway (the ring road) is located at the border of the study area. There is no heavy industry located within the study area itself, but the port of Antwerp borders the city at the north. There are no significant differences in elevation throughout the study area.

2.2. Mobile air quality monitoring

The opportunistic mobile measurement campaign[1] was carried out with the collaboration of city wardens from July 2012 until June 2013. The Antwerp city wardens are city employees who are outdoors for a large part of the day carrying out surveillance tours by bicycle or on foot. These surveillance tours do not follow fixed routes or times. Black carbon was measured using the VITO airQmap platform[2]. The measurement unit consisted of a micro-aethalometer (MicroAeth Model AE51, AethLabs), a lightweight sensor that allows to measure BC at a high (1 s) frequency, and a GPS (Locosys Genie GT-31 GPS). The micro-aethalometer measures the concentration of optically absorbing aerosol particles (equivalent black carbon (EBC, in $\mu g\,m^{-3}$) using a mass-specific absorption cross-section (MAC) of 12.47 $m^2 g^{-1}$ at 880 nm (Petzold et al., 2013)).Three teams of two city wardens each were equipped with a measurement unit, and 393 hours of raw 1 second measurements were recorded for the three teams combined (459 hours of measurements before filtering for GPS quality), spread over 110 days. Most of the measurements were done between 10 am and 16 pm during working days and performed both on foot and by bike. The micro-aethalometers have been compared several times during the campaign to a reference monitoring station. More details on data collection, processing and quality control can be found in Van den Bossche et al. (2016).

2.3. Aggregated BC concentrations

As described in Van den Bossche et al. (2016), the data at 1 s resolution were aggregated over segments of approximately 50 m resolution along the roads (assigned to the midpoint of the corresponding segment). This resulted in different passages for each segment, where one passage is a continuous period of time during which measurements are performed in that segment. For each segment, an aggregated concentration level was calculated based on all passages using a trimmed mean and temporally adjusted to an annual average concentration. The temporal adjustment was performed through a combination of the additive and multiplicative method. More details can be found in Van den Bossche et al. (2016). The trimmed mean used in this study was calculated as the arithmetic mean after removing the 0.5 % largest and 0.5 % smallest values (Van den Bossche et al., 2015). The aggregated and adjusted values are the data points that will be used as the dependent variable in the LUR models. Because no fixed routes were followed, the number of passages was not identical for all segments. Only those segments with at least 5 passages were used for the models, resulting in 1457 sampling locations. Most segments were measured 9 to 27 times (interquartile range).

A few of the segments close to the ring road were removed from the target data set, in particular, the segments located at a bridge over the ring road. These data are not representative for the ring road itself and those high values for the traffic variables were not well represented within the dataset.

2.4. GIS data

Data were gathered for four categories of predictor variables: traffic variables (traffic intensity, road length, distance to roads), land use, population density and physical geography (urban morphology). The elevation was not considered as predictor variable. The different data sources were (i) OpenStreetMap (OSM), (ii) Urban Atlas, (iii) Central Reference Address Database (CRAB), (iv) a traffic model, (v) sky view factor data (open data Antwerp) and (vi) data on biking lanes from the Province of Antwerp. These sources are described in more detail in the Supplementary Material. Based on these data, predictor variables were calculated based on different buffer sizes around the measurement locations or as a point estimate. An overview of these variables is given in Table 1.

The relationship between the BC concentrations and a predictor variable, e.g. distance, is often not linear. Therefore, some transformations of the variables were included as additional variables (inverse distance and squared inverse distance for dist_near, dist_near_major and distance_to_traffic). Further, the following interactions between different predictor variables were included: traffic variables (trafload_50, trafnear, trafnear_heavy) multiplied with inverse (squared) distance, and sky view factor with traffic intensity. These interactions were included because both the distance and the local geometry (sky view factor) have an influence on the contribution of traffic to the local concentration levels. The additional predictor variables are listed in the Supplementary Material.

In the basic scenario, all described predictor variables were used for model building (described in more detail in the next section). To investigate the impact of the availability of certain predictor variables, additional models were built using different initial sets of predictor variables: all predictor variables, all predictor variables without manual transformations and interactions, all predictor variables without the sky view factor

[1] The dataset is available upon request.

[2] http://www.airqmap.com

Table 1: Overview of the predictor variables calculated from the GIS data. Additional predictor variables are transformations or combinations of these variables and are listed in the Supplementary Material.

Name	Description	Source	Buffer radii (m)
res_hd_xx	High-density residential area in a buffer with size XX m (Urban Atlas code 11100 and 11210) [m^2]	Urban Atlas	[100, 300, 500, 1000, 3000, 5000]
res_ld_xx	Low-density residential area in a buffer with size XX m (Urban Atlas code 11120, 11130 and 11240) [m^2]	Urban Atlas	[100, 300, 500, 1000, 3000, 5000]
industry_xx	Industrial or commercial area in a buffer with size XX m (Urban Atlas code 12100) [m^2]	Urban Atlas	[1000, 3000, 5000]
port_xx	Port area in a buffer with size XX m (Urban Atlas code 12300) [m^2]	Urban Atlas	[3000, 5000]
airport_xx	Airport area in a buffer with size XX m (Urban Atlas code 12400) [m^2]	Urban Atlas	[1000, 3000, 5000]
urban_green_xx	Urban green area in a buffer with size XX m (Urban Atlas code 14100) [m^2]	Urban Atlas	[300, 500, 1000, 3000, 5000]
nature_xx	Natural land in a buffer with size XX m (Urban Atlas code 30000 and 40000) [m^2]	Urban Atlas	[300, 500, 1000, 3000, 5000]
address_xx	Number of adresses in a buffer with size XX m	CRAB	[50, 100, 300, 500, 1000, 3000]
trafnear	Traffic intensity on the nearest road [Veh day^{-1}]	Traffic model	-
trafnear_heavy	Heavy traffic intensity on the nearest road [Veh day^{-1}]	Traffic model	-
trafload_xx	Sum of (traffic intensity * road length) in a buffer with size XX m [Veh day^{-1} m]	Traffic model	[50, 100, 300, 500, 1000]
trafload_heavy_xx	Sum of (traffic intensity (heavy traffic) * road length) in a buffer with size XX m [Veh day^{-1} m]	Traffic model	[50, 100, 300, 500, 1000]
trafloadhv_fraction_xx	Fraction of heavy traffic in a buffer with size XX m	Traffic model	[50, 100, 300, 500, 1000]
roadlength_xx	Total road length in a buffer with size XX m [m]	OpenStreetMap	[50, 100, 300, 500, 1000]
roadlength_major_xx	Total major road length in a buffer with size XX m [m]	OpenStreetMap	[50, 100, 300, 500, 1000]
dist_near	Distance to the nearest road [m]	OpenStreetMap	-
dist_near_major	Distance to the nearest major road [m] (OpenStreetMap primary and secondary)	OpenStreetMap	-
dist_highway	Distance to the nearest highway [m]	OpenStreetMap	-
distance_to_traffic	Distance between bike lane and traffic [cm]	Province of Antwerp	-
skyviewfactor	Fraction of visible sky	Open Data Antwerp	-

(skyviewfactor) and the distance between bike lane and traffic (distance_to_traffic) and all predictor variables without the variables derived from the traffic intensity data (from the traffic model).

2.5. Model building

Traditionally, most LUR studies use standard multiple linear regression techniques to relate the pollutant concentration with spatial predictor variables (Hoek et al., 2008). Some studies use non-linear techniques such as Generalized Additive Models (GAM) (e.g. Hasenfratz et al., 2015; Dekoninck et al., 2015). In this study, both multiple linear regression and a non-linear regression technique, support vector regression (SVR) using a radial basis function (RBF) kernel (Smola and Schölkopf, 2004), were used.

In many papers, the methodology for building LUR models as described in Henderson et al. (2007) and Eeftens et al. (2012) is used. We also used it in this study, and refer to it is as the 'classic' method. It is a supervised stepwise forward search of the best subset of predictor variables, based on an optimization of the adjusted R^2 and the significance of the coefficients in the linear regression. The supervised step checks whether the direction of the effect of each predictor variable (i.e. the sign of the coefficient in the linear model) corresponds to the predefined expected direction based on expert knowledge. The predictor variable that gives the highest adjusted R^2 in a univariate regression is used as a starting point. Subsequently, the predictor variable that yields the highest increase in adjusted R^2 is added in a stepwise manner, provided the following criteria are met: (i) the increase in adjusted R^2 is greater than 1 %, (ii) all variables have coefficients with a p-value < 0.05 and (iii) the sign of the coefficient agrees with the predefined effect and the sign of the other coefficients in the model does not change.

Next to the aforementioned 'classic' method, we also adopted different approaches of predictor variable selection based on cross-validation to ensure that those predictor variables that have the best generalization power are selected. The cross-validation approach is described in more detail in the next section. We used LASSO, a linear modelling approach that forces the estimated coefficient vector to be sparse using regularization (Tibshirani, 1996). In addition, both a forward search limiting the number of variables by requiring an increase in the cross-validation R^2 of 0.01 (forward CV), and a combined forward and backward search of a subset of variables that maximizes the cross-validation R^2 (optimal CV) were used. The latter approach is similar to the one taken in Beckerman et al. (2013). To summarize, the following model building procedures were used:

- **No selection**: using all available predictor variables in the model without selecting a subset.
- **Classic** (only for the linear models): using the 'classic' approach of a supervised stepwise forward variable selection.
- **LASSO** (only for the linear models): forcing the estimated coefficient vector to be sparse using regularization (and in this way selecting a subset of predictor variables). This method uses the cross-validation R^2 to optimize the regularization parameter.
- **Optimal CV**: using a combined forward and backward search for a subset of predictor variables that maximizes the cross-validation R^2.
- **Forward CV**: using a stepwise forward variable selection based on the cross-validation R^2, but with a stopping criterion of an increase in R^2 of 0.01.

In each of the model building procedures, we can distinguish two steps. Firstly, suitable predictor variables are selected and/or hyperparameters are optimized (hyperparameters are pa-

rameters whose values are set before fitting and not derived during fitting), which we will jointly refer to as **model selection**. Secondly, given a certain subset of variables or values for hyperparameters, the model is fitted on the data and the parameters are estimated (**model fitting**).

All but the classic and 'no selection' method used cross-validation in the model selection phase for optimizing hyperparameters (in case of LASSO and SVR models) or selecting predictor variables (in case of optimal and forward CV). In addition to model selection, cross-validation was also used for model evaluation. This will be elaborated in Section 2.6.3.

Further, all predictor variables were scaled by subtracting the mean and scaling to unit variance during model building (in case of cross-validation the mean and variance were determined based on the training dataset). For the SVR, a grid search was performed for the optimization of the hyperparameters. All regression analyses were performed using the Python packages scikit-learn (Pedregosa et al., 2011) and Statsmodels (Seabold and Perktold, 2010).

2.6. Model evaluation and spatial cross-validation

2.6.1. Performance metrics

The LUR models were evaluated using a set of different metrics. First, the R^2 (coefficient of determination) was used. It provides a measure of how well unseen samples will be predicted by the model. If $\hat{y}_i$ is the predicted value and y_i is the corresponding observed value of the i-th sample (with a total of n samples), then R^2 is defined as:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \quad (1)$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean observed value, RSS is the residual sum of squares and TSS is the total sum of squares. In addition to the R^2, also the explained variance (EV) was calculated:

$$\text{EV} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)} = 1 - \frac{\sum_{i=1}^{n}((y_i - \hat{y}_i) - \overline{(y_i - \hat{y}_i)})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \quad (2)$$

where Var denotes the variance. For a linear model with an intercept (fitted without regularization), the R^2 and explained variance are the same (and identical to the squared Pearson correlation), but once the model is applied to other data (e.g. during cross-validation), the model is fitted using regularization or when using non-linear models (e.g. SVR), these two metrics are not necessarily the same. There can be a high correlation between the measured and predicted values leading to a good explained variance, but still a poor prediction of the absolute values leading to a low or negative R^2. Further, also the root mean squared error (RMSE) between the model predictions and measurements was calculated.

The air quality measurements at short distances are known to be correlated. The spatial autocorrelation of the BC measurements and of the model residuals was evaluated using an empirical variogram and Moran's I statistic. To calculate the latter, appropriate spatial weights have to be defined. In this study, the inverse squared distance was used for the full matrix of all data points.

Figure 1: The different spatial zones for cross-validation. The zones are constructed as 1 x 1 km^2 areas based on the UTM coordinates. Some of the zones with fewer sampling locations are combined into one zone, resulting in six zones as indicated with numbers in the figure.

2.6.2. Spatial cross-validation scheme

To provide an unbiased evaluation of how well the models would predict the air quality for independent data (e.g. other locations within the same city), the metrics above were calculated using cross-validation. Given the high spatial autocorrelation in air quality data and given the high spatial density of sampling locations in this study, we have chosen to use a cross-validation scheme based on delineated areas instead of an n-fold cross-validation with random folds. Different zones were constructed within the study area (based on areas of 1 x 1 km^2, Figure 1). The samples within each zone were used as folds in the n-fold cross-validation ($n = 6$ in this case).

This spatial cross-validation scheme was used both during model building and for model evaluation. In each iteration of the cross-validation, the data of one fold (zone) is held out as test data to estimate the performance of the model that is built and/or fitted on the data of all other folds.

2.6.3. Cross-validation with and without full rebuilding of the model

When using the spatial cross-validation scheme for model evaluation, the model can either be refitted only or fully rebuilt (including model selection) for each fold. This distinction is important, and is not always clearly made in the LUR literature. To obtain an unbiased estimate of the predictive performance of the model, it is important that the validation data are fully independent, and thus not used at all in the full model building procedure. When a fold that is used during model selection (e.g. variable selection), but not for model fitting, is reused in the validation phase, the performance estimate will still be overly optimistic.

Therefore, we will report the performance metrics based on two different cross-validation schemes (illustrated in Figure 2).

- Cross-validation with refitting only of the models ($\mathbf{CV_F}$): in this scheme, the model is only refitted during cross-

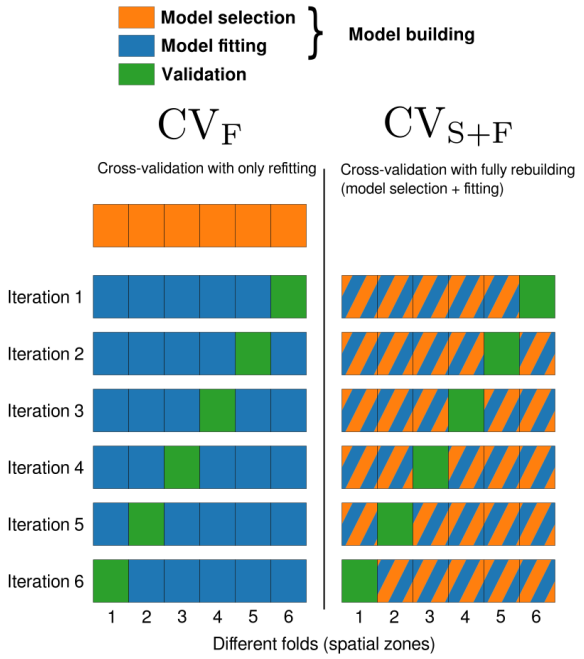

Figure 2: Illustration of the two cross-validation schemes CV_F and CV_{S+F}.

validation (parameter estimation). The model selection is only done once using all data.

- Cross-validation with full rebuilding of the models (**CV_{S+F}**): in this scheme, both model selection and model fitting are performed during cross-validation. This means that in each iteration of the cross-validation, a different model structure (in terms of selected variables and values for hyperparameters) is obtained.

In both cases, six different models are obtained during cross-validation (for the six iterations). When using CV_F, only the parameter values will differ between the models, whereas when using CV_{S+F}, also the model structure will vary, with each model having its own subset of predictor variables or values for hyperparameters. The variability between the models will also give an indication of the stability of the model selection. When reporting the performance metrics for the cross-validation, the metrics are both given as the range of the metrics obtained in all the iterations, as well as calculated using the pooled predictions (meaning the combination of the predictions obtained in all the iterations). For CV_F, also the R^2 of the model fitted and evaluated on all data is given (the model R^2).

In case the model selection itself also uses cross-validation, CV_{S+F} implies that a nested cross-validation will be performed. One fold is held out as test dataset, and the model will be built based on the other data using cross-validation with the remaining 5 folds. For the classic approach, the variable selection is not based on cross-validation, so for this method CV_{S+F} does not incur a nested cross-validation.

3. Results

3.1. Exploration of the target and predictor variables

A map of the measured concentrations is shown in Figure 3. The concentrations range from 0.4 $\mu g\,m^{-3}$ to 9.8 $\mu g\,m^{-3}$ with a mean of 3.3 $\mu g\,m^{-3}$ (median of 3.0 $\mu g\,m^{-3}$). The concentrations are spatially correlated (Figure 4). The largest increase in variance occurs up to 300 m, and after approximately 500 m the BC concentrations are no longer autocorrelated. Moran's I statistic is equal to 0.33 for the measured concentration levels.

The study area only includes urban area, no nature or low density residential areas. Therefore, predictor variables that were only present in those missing areas were left out. Most of the remaining predictor variables showed a large variability throughout the study area. The distribution of the BC concentrations is similar in most of the spatial zones (Figure S1).

3.2. Model results

The evaluation results of the different LUR models are summarized in Tables 2 and 3. The selected predictor variables for a selection of the individual models are given in the Supplementary Material. A large variability in selected variables can be noted. The measured and predicted concentration values are visualized in Figure 5 as scatter plots. A systematic underestimation of the high concentrations and an overestimation of the low concentrations is observed. Further, the predicted concentration values are also mapped in Figure 6.

3.2.1. Spatial cross-validation schemes

For all methods, the CV_{S+F} evaluation results in lower values for the metrics than the evaluation of the models based on all data (CV_F). For the classic linear model, for example, the model R^2 is 0.43 while the CV_F R^2 is 0.35. When fully rebuilding the model during cross-validation, the score decreases further: a CV_{S+F} R^2 of 0.24. A similar trend is observed for the other methods. This decrease in performance between CV_F and CV_{S+F} can be explained by overfitting and lack of validation on fully independent data in CV_F.

The importance of the spatially independent folds in the cross-validation is illustrated by testing other schemes. Using the custom spatial scheme, the CV_F R^2 is 0.30 for the SVR model without variable selection (Table 2). When using a random 10-fold cross-validation or leave-one-out cross-validation (LOOCV), the R^2 increases to 0.55 and 0.60, respectively.

3.2.2. Model-building methods

Based on the CV_{S+F} results, the differences between the models are not large. Most of the models have a similar predictive performance (R^2 ranging between 0.24 and 0.26), with slightly lower values for the forward and optimal CV linear models and the optimal CV SVR model. For the CV_F R^2 values, a larger difference between the different models is found.

The optimal CV models score better on CV_F R^2. For example, the linear optimal CV model has a CV_F R^2 of 0.46, but an R^2 of only 0.22 for CV_{S+F}. This large difference between CV_F and CV_{S+F} performance indicates that there is overfitting during the variable selection phase. The optimal CV models also

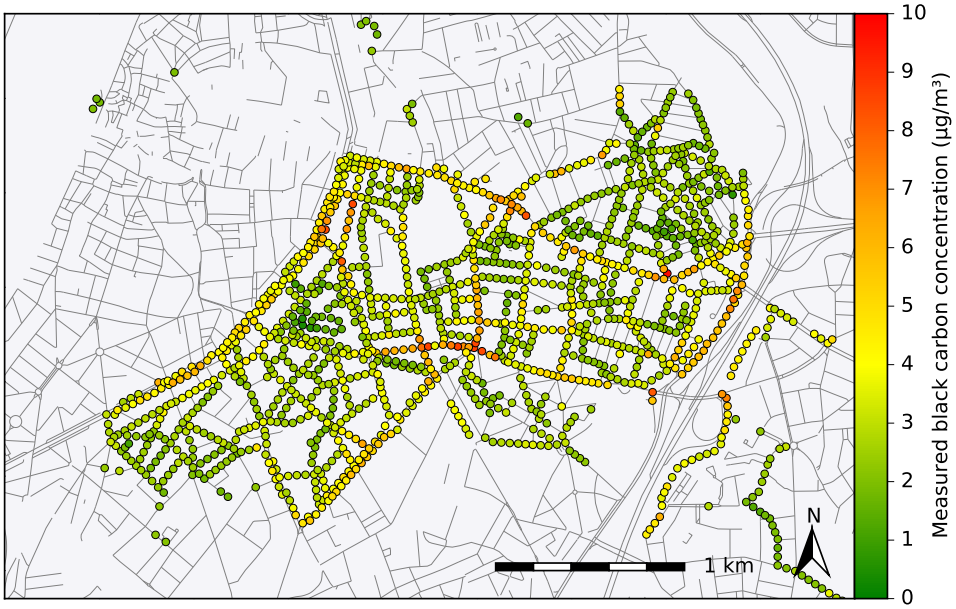

Figure 3: Map of the measured BC concentrations.

Table 2: Overview of the evaluation results of the different LUR models. For the CV_F scheme (model selection based on all data), the model R^2 and the cross-validation R^2, explained variance (EV) and RMSE are given. Moran's I statistic is calculated for the residuals of the model fitted on all data. For the CV_{S+F} scheme (cross-validation with full rebuilding), the R^2, EV and RMSE of the different models are given. The CV metrics are reported for the pooled predictions and as the minimum and maximum of the individual models as [min, max].

	CV_F (only refitting)					CV_{S+F} (fully rebuilding)		
	Model R^2	R^2	EV	RMSE	Moran's I	R^2	EV	RMSE
Linear regression								
No selection	0.56	-0.08 [-1.24, 0.32]	-0.07 [-1.07, 0.36]	1.4 [1.1, 2.2]	0.16	-	-	-
Classic	0.43	0.35 [-0.39, 0.54]	0.35 [0.26, 0.55]	1.0 [0.8, 1.2]	0.26	0.24 [-0.65, 0.52]	0.25 [0.28, 0.53]	1.1 [1.0, 1.3]
LASSO	0.35	0.28 [0.01, 0.48]	0.28 [0.24, 0.48]	1.1 [1.0, 1.3]	0.31	0.26 [-0.26, 0.44]	0.26 [0.26, 0.45]	1.1 [1.0, 1.3]
Optimal CV	0.52	0.46 [0.31, 0.57]	0.46 [0.36, 0.58]	1.0 [0.8, 1.1]	0.19	0.22 [-0.70, 0.50]	0.23 [-0.02, 0.52]	1.1 [1.0, 1.3]
Forward CV	0.38	0.34 [0.08, 0.49]	0.34 [0.29, 0.50]	1.1 [0.9, 1.3]	0.29	0.20 [-0.75, 0.36]	0.22 [0.22, 0.41]	1.2 [0.9, 1.3]
SVR								
No selection	0.68	0.30 [0.07, 0.40]	0.30 [0.16, 0.42]	1.1 [1.0, 1.3]	0.13	0.26 [0.06, 0.36]	0.26 [0.16, 0.38]	1.1 [0.9, 1.3]
Optimal CV	0.60	0.46 [0.25, 0.57]	0.46 [0.30, 0.58]	1.0 [0.8, 1.1]	0.16	0.19 [-0.60, 0.40]	0.20 [0.04, 0.43]	1.2 [1.0, 1.3]
Forward CV	0.48	0.40 [0.24, 0.58]	0.41 [0.35, 0.58]	1.0 [0.8, 1.2]	0.23	0.24 [-0.03, 0.37]	0.26 [0.17, 0.41]	1.1 [0.9, 1.3]

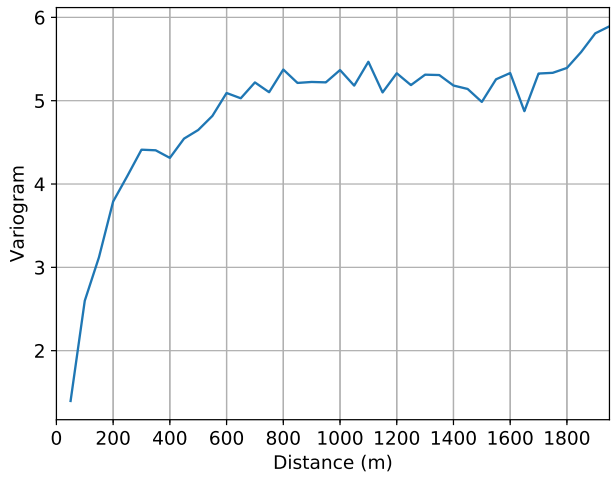

Figure 4: Variogram of BC measurements.

include a larger number of variables: for the linear models, the optimal CV method selects on average 14 variables in CV_{S+F}, compared to six variables for the classic supervised method.

Using the forward CV method for the linear model results in a lower CV_{S+F} R^2 compared to the classic method, although both methods use the stopping criterion of a minimum of 1 % increase in R^2. In the case of SVR, the forward CV selection method gives slightly worse results based on CV_{S+F} compared to using all variables (0.24 vs 0.26 for R^2, respectively, and an EV of 0.26 for both). However, the resulting forward CV model only uses five to six predictor variables. Therefore, from a practical point of view (fewer variables are preferred when applying the model), the forward CV method is used for the next paragraph.

3.2.3. Available predictor variables

To investigate the impact of the available predictor variables, both linear and SVR models are built using the different ini-

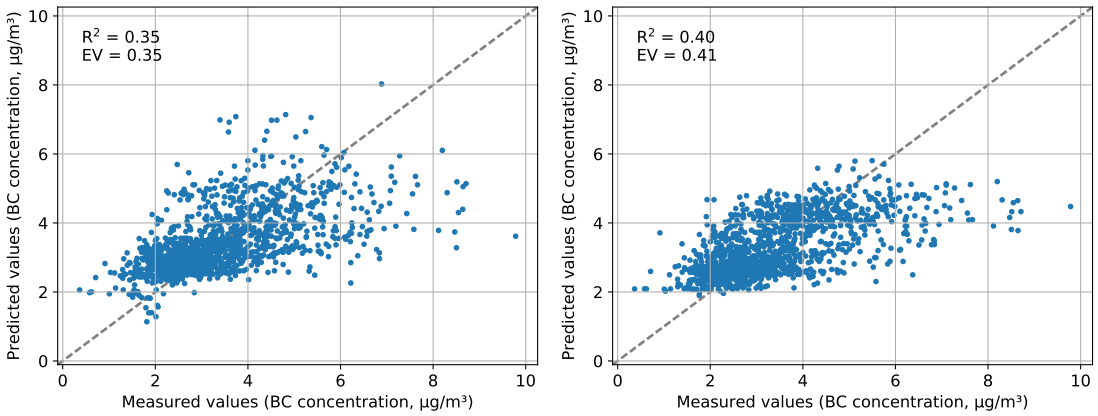

Figure 5: Scatter plots of the measured and predicted BC concentrations for the optimal (a) linear model (classic) and (b) SVR model (forward CV). The predicted BC concentrations are the pooled predictions of the CV_F models.

tial sets. The classic approach is used for the linear models and forward CV for SVR. An overview of the results is given in Table 3. The best results are obtained when using all predictor variables including the transformations and interactions. In each of the models at least one interaction of a traffic variable with the distance between bike lane and traffic (trafnear/distance_to_traffic or trafload_50/distance_to_traffic) is included (Table S1). Leaving out all variables using the sky view factor (skyviewfactor) or the distance between bike lane and traffic (distance_to_traffic) also leads to a lower R^2. Leaving out traffic intensity variables leads to a much lower performance. The traffic variables are then mainly replaced by the total length of major roads (roadlength_major) in buffer zones of 50 and 100 m. For the SVR models, similar trends can be noted.

3.2.4. Model residuals

The model residuals are analysed by calculating Moran's I statistic for each of the models (based on the model fitted on all data). For most models there is a small decrease in Moran's I statistic compared to the value of 0.33 for the measured concentrations. But, the values did not fall to zero, indicating that there still is a considerable spatial autocorrelation in the model residuals. For the classic linear model, Moran's I statistic decreased to 0.26. The variogram for this model (not shown) shows that the influential distance has become smaller, confirming the decrease in Moran's I statistic. The residuals for this model are also visualized on a map (Figure 7).

4. Discussion

4.1. Different evaluation approaches

The performance of the LUR models was evaluated using a custom spatial 6-fold cross-validation scheme to ensure spatial independence between the training and the test set. Further, the performance was assessed both by cross-validation of the model based on all data (CV_F) and by excluding one fold in model building for evaluation during cross-validation (CV_{S+F}). Generally, there is a clear difference between the evaluation approaches resulting in lower R^2 and EV values for CV_{S+F} compared to CV_F. The CV_{S+F} approach ensures an unbiased estimate of the predictive performance by excluding the validation data from the model building phase. This indicates that not only the model R^2, but also the CV_F cross-validation R^2, does not necessarily reflect the predictive ability of the model. The results stress the importance of a proper evaluation method when assessing the predictive performance of LUR models.

In literature, LUR models often use leave-one-out cross-validation (LOOCV) to assess the model performance (e.g. Hoek et al., 2008; Eeftens et al., 2012; Beelen et al., 2013), but it is known that this may overestimate the predictive ability on independent data sets (Wang et al., 2012; Basagaña et al., 2012; Wang et al., 2013). Some other studies use hold-out validation (HV) or random n-fold cross-validation to get a more reliable estimate of the performance (e.g. Hasenfratz et al., 2015; Kanaroglou et al., 2013; Montagne et al., 2015). For these approaches, sufficient sampling locations are required to be able to split up the dataset. When using a random 10-fold cross-validation or LOOCV instead of the spatial cross-validation scheme in our case study, the R^2 increases from 0.30 to 0.55 and 0.60, respectively (using the SVR model without variable selection). This stresses the importance of the spatial independence of the validation data. The spatial cross-validation scheme tries to ensure this independence. At the borders of the zones, there will still be a spatial autocorrelation between samples in the training and validation dataset, but in view of the spatial autocorrelation this method is more appropriate than LOOCV or random n-fold cross-validation. A possible improvement would be not to work with fixed spatial zones, but to use a buffer of certain size around each individual sampling location.

Recent studies show that LUR models can perform well in the LOOCV setting, but perform worse in evaluation on an independent dataset (Wang et al., 2012; Basagaña et al., 2012; Wang et al., 2013). However, we have shown here that also cross-validation or hold-out approaches can still give an over-

Table 3: Overview of the evaluation results of the LUR models starting from different initial subsets of predictor variables. The metrics are the same as in Table 2.

	CV_F (only refitting)					CV_{S+F} (fully rebuilding)		
	Model R^2	R^2	EV	RMSE	Moran's I	R^2	EV	RMSE
Linear regression								
All variables	0.42	0.30 [-0.23, 0.43]	0.31 [0.18, 0.46]	1.1 [0.9, 1.2]	0.25	0.19 [-0.78, 0.33]	0.20 [0.20, 0.44]	1.2 [0.9, 1.3]
All variables + interactions	0.43	0.35 [-0.39, 0.54]	0.35 [0.26, 0.55]	1.0 [0.8, 1.2]	0.26	0.24 [-0.65, 0.52]	0.25 [0.28, 0.53]	1.1 [1.0, 1.3]
Without skyview/distance_to_traffic	0.42	0.29 [-0.30, 0.41]	0.30 [0.23, 0.47]	1.1 [0.9, 1.3]	0.24	0.18 [-0.80, 0.33]	0.19 [0.22, 0.42]	1.2 [0.9, 1.3]
Without traffic	0.31	0.14 [-0.82, 0.44]	0.14 [-0.04, 0.46]	1.2 [0.9, 1.3]	0.28	0.04 [-1.15, 0.31]	0.05 [-0.08, 0.34]	1.3 [0.9, 1.4]
SVR								
All variables	0.46	0.39 [0.18, 0.57]	0.39 [0.32, 0.58]	1.0 [0.8, 1.2]	0.24	0.20 [-0.53, 0.41]	0.20 [0.07, 0.44]	1.2 [1.0, 1.4]
All variables + interactions	0.48	0.40 [0.24, 0.58]	0.41 [0.35, 0.58]	1.0 [0.8, 1.2]	0.23	0.24 [-0.03, 0.37]	0.26 [0.17, 0.41]	1.1 [0.9, 1.3]
Without skyview/distance_to_traffic	0.56	0.41 [0.33, 0.52]	0.41 [0.37, 0.55]	1.0 [0.8, 1.2]	0.19	0.17 [-0.47, 0.37]	0.18 [-0.03, 0.39]	1.2 [1.0, 1.4]
Without traffic	0.50	0.30 [0.12, 0.41]	0.31 [0.20, 0.41]	1.1 [0.8, 1.2]	0.17	0.05 [-0.48, 0.20]	0.06 [-0.37, 0.21]	1.3 [1.1, 1.4]

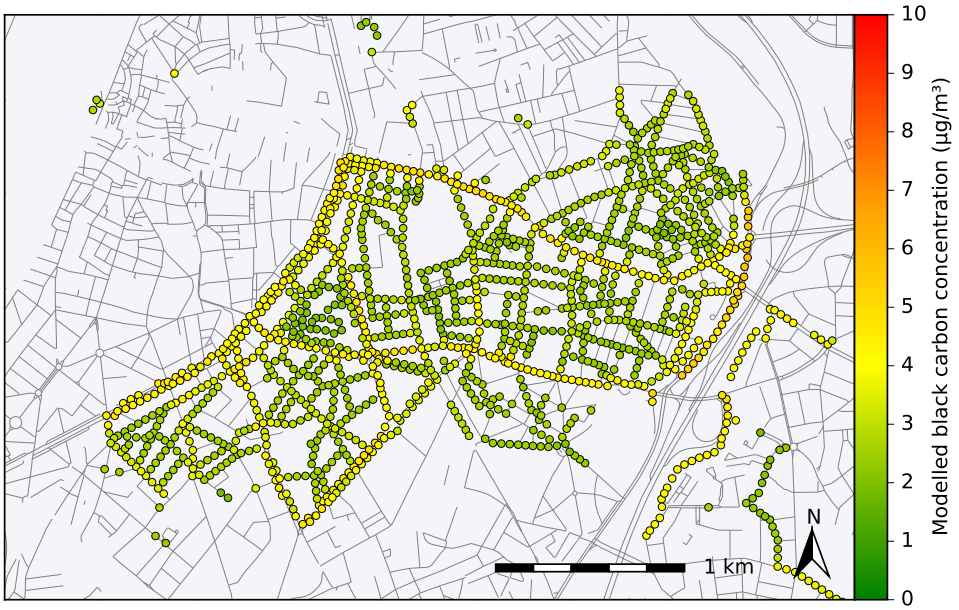

Figure 6: Map of the predicted BC concentrations for the classic linear model (at the same locations where measurements have been performed). The predicted values are the pooled predictions of the CV_F models.

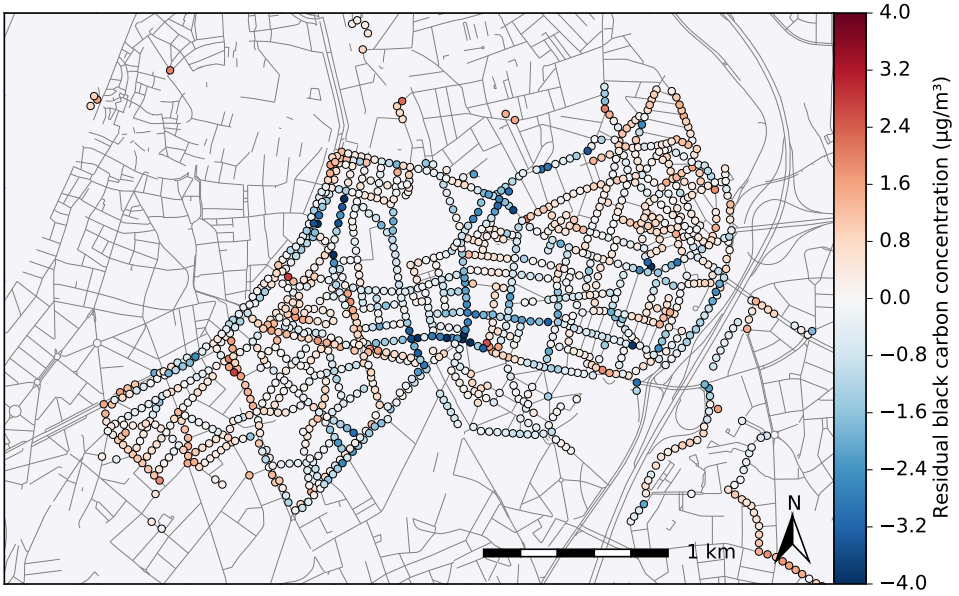

Figure 7: Map of the residuals (predicted concentrations minus measured concentrations) for the classic linear model. Negative values indicate an underestimation of the LUR model. The predicted values are the pooled predictions of the CV_F models.

estimation if the model is not fully rebuilt. Often, the model is not rebuilt but only refitted (parameter estimation), in both LOOCV and HV settings (e.g. Eeftens et al., 2012; Beelen et al., 2013; Kanaroglou et al., 2013). Our results show the importance of doing a full rebuild (i.e. redoing the variable selection, the CV_{S+F} approach) for each fold of the cross-validation to get an estimate of the predictive ability of the model. The CV_{S+F} approach yields the best possible estimate of the predictive ability of the model when applied to actual unseen data.

4.2. Different model building techniques

Multiple models and model building techniques were tested in this study. When using the CV_{S+F} cross-validation scheme, there was not much difference between the obtained models in terms of predictive ability. For the linear models, it is clear that variable selection is important given the negative cross-validation R^2 when using all variables.

The supervised stepwise regression is widely used (Hoek et al., 2008; Eeftens et al., 2012). This approach focuses on selecting models with a limited number of predictor variables that have plausible (predefined) effects. The motivation of this supervised approach is that it results in a more interpretable model, that the model could be applied more easily in other study areas and that it limits the risk of overfitting (Beelen et al., 2013). In this approach, the variable selection in the model building process is based on all data. Alternatively, the variable selection can also be based on the CV performance, e.g. the cross-validation R^2 instead of the adjusted model R^2, to select those variables that give the best generalization and to minimize the risk of overfitting. Basagaña et al. (2012) compared different techniques: the classic approach with a forward selection based on adjusted R^2, the same algorithm but forward selection based on LOOCV R^2 and the deletion/substitution/addition (DSA, Su et al. 2015) algorithm that searches through the variable space in order to minimize the squared prediction error during cross-validation. They concluded that the techniques performed similarly in terms of predictive ability on the validation dataset. Our results also do not show much difference between the different techniques, and an even worse predictive performance for the models based on an optimization of the cross-validation R^2 that do not limit the number of predictor variables. Those models have a better cross-validation CV_F R^2 compared to the classic approach, but the CV_{S+F} R^2 is lower. This means that, despite the cross-validation during variable selection, there is some overfitting. To have generalizable models, it seems important to limit the number of predictor variables by early stopping. The methods that do not use a custom variable selection method but have regularization built in, i.e. LASSO and SVR without manual variable selection, show less difference between the CV_F and CV_{S+F} performance metrics. For these methods, less overfitting occurs

The regularized linear model, LASSO, performed similarly to the classic linear model (CV_{S+F} R^2 of 0.26 vs 0.24), even though LASSO does not make use of the additional information on predefined effects as used in the classic method. The non-linear SVR models (based on all variables and forward CV) also performed similarly. Introducing non-linearities in the model by transforming predictor variables seems to be enough for the linear model to perform as well as the non-linear model. The differences between the models are, however, very small and therefore it is not possible to draw clear conclusions. Weichenthal et al. (2016a) also compared multiple linear regression and a non-linear approach. The non-linear model had a higher model R^2, but when evaluating the model with cross-validation the difference was no longer significant. To conclude, using SVR as a non-linear technique did not yield compelling improvements over linear regression for the data set in this study.

4.3. Mobile monitoring as the basis for LUR models

A limited number of studies have used mobile monitoring to build LUR models, and more research is needed to determine best practices (Hankey and Marshall, 2015). When using mobile monitoring as the basis for LUR models, an important aspect is how to aggregate the mobile measurements into a suitable form that can be fed into the model. Based on the analysis in Van den Bossche et al. (2015), we adopted a 50 m resolution and a trimmed mean as the aggregation statistic. The measurements were aggregated over segments of an approximately equal length of 50 m along the road network. Other approaches include assigning the collected data to the midpoints of the corresponding road segment (e.g. Weichenthal et al. (e.g. 2016b) or regular grids (e.g. Hasenfratz et al., 2015).

Hankey and Marshall (2015) tested different metrics to aggregate their data over different passages. They chose the median concentration in their best-case models because this gave a lower error than for the mean concentration. In previous work we showed that the trimmed mean used in this paper reduces the impact of extreme values on the average concentration of a segment and gives a better estimate of the true mean than the arithmetic mean or the median (Van den Bossche et al., 2015).

The advantage of mobile measurements is the ability to monitor many locations, leading to a high spatial density of sampling locations in the study area. This yields a higher spatial resolution than can be achieved with fixed site monitoring. In this study, 1457 sampling locations were included, and the study of Hankey and Marshall (2015) used 1101 locations. Basagaña et al. (2012) showed that building LUR models to characterize local air pollution levels in complex urban settings should be based on a larger number of measurement sites than the 20–100 sites that are typically used in LUR studies, and thus the large number of sampling locations obtained by mobile monitoring can yield a more adequate dataset for such LUR models. But, mobile monitoring also holds some drawbacks compared to stationary monitoring. Due to measuring at many locations for only a short time period, an increased variation in the measured concentrations can be expected (Van den Bossche et al., 2015). This may result in a lower model fit, which is further discussed in the next section. To cope with this, a sufficient number of repeated measurements at the same location is needed to get a reliable estimate of the average concentration at each location (Van den Bossche et al., 2015). Further, mobile measurements will more likely be limited to the day-time

hours when people are active and to publicly accessible locations (streets, squares, urban green, etc.). Care should be taken to avoid a sampling bias towards traffic environments. In the present case study, the spatial spread was appropriate and also non-traffic environments were included due to the nature of the city wardens' work, but the temporal coverage was limited to working hours.

4.4. Evaluation of the performance

4.4.1. Comparison to other studies

It is difficult to compare the obtained performance with that of other studies. Firstly, often different evaluation methods are used (different metrics or cross-validation approaches) and it is not always clear which method is exactly used. As explained above, the choice of the method to evaluate the performance can have a large impact on the obtained values. Secondly, the number of sampling locations can vary greatly. Based on a resampling exercise, Basagaña et al. (2012) concluded that higher R^2 values are found for LUR models based on a smaller number of measurement sites, but that this does not necessarily reflect the true predictive ability. Finally, depending on the data quality, part of the unexplained variance can be attributed to uncertainty in the data itself, which can result in a lower model fit.

In this study, a CV_F R^2 of 0.35 and CV_{S+F} R^2 of 0.24 were obtained for the classic model. In other studies using mobile monitoring, similar performances were found. Hasenfratz et al. (2015) obtained an R^2 of 0.38 for a yearly map with 10-fold cross-validation based on mobile tram measurements. In the study of Kanaroglou et al. (2013), a LUR model for SO_2 with a resolution of 50 m was built using mobile van-based measurements and they obtained an R^2 of 0.30 for a 50 % hold-out cross-validation data set (with a fixed set of selected variables during cross-validation, thus without fully rebuilding the model). Hankey and Marshall (2015) also used on-road bike measurements, and the LUR model for BC showed an R^2 of 0.20 to 0.35 using a random 1/3 hold-out cross-validation. A higher R^2 of 0.50 based on an independent validation dataset was found by Weichenthal et al. (2016b) for a UFP model. However, the validation data were scattered throughout the study area and therefore possibly not spatially independent.

The performance on spatially independent data during cross-validation varies considerably between folds: a pooled CV_{S+F} R^2 of 0.24 (and EV of 0.25) is found for the classic linear model while the R^2 values for the individual folds range between -0.65 and 0.52 (EV ranges between 0.28 and 0.53, Table 2). The negative R^2 value means that for this specific zone (zone 1 in Figure 1) the absolute values are not predicted well. The variability in the selected predictor variables (Tables S1 and S2 in the Supplementary Material) and the variation in performance between the different folds during CV_{S+F} can be caused by differences in land use between the different zones (Figure S1). Similar trends of high variability between the folds are observed for the other methods as well. This lower performance during CV_{S+F} cross-validation indicates that the model still has difficulty to generalize to the full city and that the performance in predicting the concentrations, especially absolute values, for areas where no measurements took place (outside of the study area) is limited.

Hankey and Marshall (2015) also performed a systematic validation using data from two routes to predict the third route, leading to low R^2 values (0.01 to 0.20 for BC). A possible reason they gave was that the range spanning the predictor variables within each of the routes was not fully balanced. In the study of Patton et al. (2015), measurements were performed in four different neighbourhoods. When models built for one of the neighbourhoods were transferred to the other neighbourhoods, the models performed poorly ($R^2 < 0.17$, compared to R^2 of 0.23 to 0.42 for the neighbourhood-specific models). These two studies also have difficulties to generalize the model to other parts of the same city. This evaluation of the transferability of the models is similar to the spatial cross-validation in the present study, although the neighbourhoods in Patton et al. (2015) are not contiguous but around 3 to 12 km apart, and the routes in Hankey and Marshall (2015) also cover a larger area (about 8 x 12 km^2).

4.4.2. Selected predictor variables

The traffic intensity variables are the primary selected variables in all models. For example, the SVR model has a CV_F R^2 of 0.29 when only the traffic load in a buffer of 50 m (trafload_50) is used as predictor variable, and 0.39 when also including a variable on heavy traffic and the sky view factor. The problem with the traffic intensity variables is that they are not always easily accessible. Hoek et al. (2008) reported that several LUR studies have successfully explored the use of the length of specific road types without traffic intensity data (e.g. Henderson et al., 2007). In the present study, however, the models that were forced not to include traffic variables had a low performance (Table 3) and the different road length and distance to road variables did not prove to be decent substitutes. This could be attributed to the generally high road density in the city centre. These results emphasize the importance of the availability of traffic data when building LUR models.

We have also built models for a scenario without the availability of the sky view factor and the distance between bike lane and traffic (distance_to_traffic) (Table 3) because these data are rather specific and not generally available. However, given the lower performance of the models, those predictor variables have a clear added value in explaining the variability in BC concentrations in the urban environment. Eeftens et al. (2013) also concluded that street canyon indicators such as the sky view factor could be valuable to consider in air pollution models. The results correspond with the findings in Peters et al. (2014) that traffic intensity, distance to the traffic and street topology are determinant for cyclist exposure.

4.4.3. Explaining low performance: quality of data and predictor variables

Generally low R^2 and EV values are obtained for the final models. In the previous paragraphs we discussed the impact on the performance of the validation approach and the number of sampling locations. Here, we will further discuss the low performance.

The opportunistic mobile monitoring methodology had an impact on the quality of the data, as discussed in Van den Bossche et al. (2016). There is still a rather large uncertainty on the average concentration levels at a spatial resolution of 50 m due to a limited number of measurements for many of the locations, limited temporal coverage and sampling bias. Insufficient data were gathered to get an accurate estimate of the BC concentration at all locations. Part of the unexplained variance can therefore be attributed to uncertainty in the data itself. The data quality could be improved by using more mobile platforms for the data collection, a stronger follow-up of the participants, etc. (see Van den Bossche et al. (2016) for more detailed discussion).

However, there are also systematic errors indicating that part of the unexplained variance is also due to missing to explanatory power of the model and predictor variables. The model residuals show a clear relation with the BC concentrations: there is a systematic underestimation of the high concentrations and an overestimation of the low concentrations. The model residuals also exhibit a considerable spatial autocorrelation. This analysis of the model residuals indicates that not yet all explanatory factors are captured in the predictor variables.

The unexplained variance can be explained by the insufficient quality of certain predictor variables. For example, the traffic intensity is based on a traffic model and is known to be rather coarse. The influence of congestion and frequent start-stop behaviour of traffic in the urban environment is also not included. The modelled traffic intensity may therefore be a poor predictor of actual traffic emissions. Further, the complex interplay between local emissions, local street geometry and local meteorology may not be adequately captured in the predictor variables.

4.5. Limitations of this study

The applicability of the LUR models obtained in this study is restricted by the characteristics of the input (pollution) data. The measured concentrations are representing street-level (cyclist or pedestrian) daytime exposure values. Further, the LUR models are applied in a relatively small study area. The question how well the model would perform at a larger scale (e.g. including the peripheries and not only the city centre of Antwerp) remains unanswered. The goal of the present study was to obtain an annual average map and we did not distinguish between different seasons. Further, as discussed above and in Van den Bossche et al. (2016), the dataset of opportunistic measurements used in this study has limitations with regard to the data quality. Despite the uncertainty on the concentration levels, large spatial patterns within the city are clearly captured with the mobile campaign.

5. Conclusion

The goal of this paper was to develop and evaluate LUR models to predict annual average BC concentrations in an urban environment based on opportunistic mobile measurements. We can conclude that mobile monitoring is suited for building LUR models at a high spatial resolution. Mobile monitoring can provide the high spatial resolution data needed to characterize the spatial variability in the complex urban environment.

We illustrated the importance of a careful evaluation approach for estimating the predictive performance of the model using an appropriate cross-validation scheme. It is crucial to use spatially independent data for the validation. These test data should be excluded during variable selection in the model building procedure. Many papers in literature do not use such a rigorous evaluation approach, often because the limited data available do not allow them to do so, and find overly optimistic performance estimates. Different model building techniques were tested. LASSO, a regularized linear model, performed slightly better than the classic supervised approach, and the non-linear SVR technique did not show much improvement over a linear model. But, due to the generally low R^2 and the small differences, it is not possible to draw clear conclusions on which model building technique is preferred.

The LUR models obtained in this study explain a significant part of the variance in the BC concentrations. The relatively low R^2 values can be attributed partly to uncertainty in the data related to the set-up of the opportunistic mobile monitoring campaign. However, there is a systematic underestimation of the high concentrations and an overestimation of the low concentrations, indicating that not all explanatory factors are captured in the predictor variables. Further, the generalization of the LUR model to areas where no measurements were made is limited, especially in predicting absolute concentrations.

Acknowledgement

We would like to thank the city of Antwerp, Environmental Services and especially the city wardens (Dienst Samen Leven - Stadstoezicht) for the collaboration in the measurement campaign. We would also like to thank the Province of Antwerp and the Fietsersbond (cyclists association) for the biking lane data.

References

Basagaña, X., Rivera, M., Aguilera, I., Agis, D., Bouso, L., Elosua, R., Foraster, M., de Nazelle, A., Nieuwenhuijsen, M., Vila, J., Künzli, N., 2012. Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmospheric Environment 54, 634–642.

Beckerman, B.S., Jerrett, M., Martin, R.V., van Donkelaar, A., Ross, Z., Burnett, R.T., 2013. Application of the Deletion/Substitution/Addition algorithm to selecting Land Use Regression models for interpolating air pollution measurements in California. Atmospheric Environment 77, 172–177.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.Y., Künzli, N., Schikowski, T., Marcon, A., Eriksen, K., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., Birk, M., Cyrys, J., von Klot, S., Nádor, G., Varró, M.J., Dėdelė, A., Gražulevičienė, 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., Bellander, T., Korek, M., Olsson, D., Strömgren, M., Dons, E., Jerrett, M., Fischer, P., Wang, M., Brunekreef, B., de Hoogh, K., 2013. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – the ESCAPE project. Atmospheric Environment 72, 10–23.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., Vineis, P., Xun, W.W., Katsouyanni, K., Dimakopoulou, K., Oudin, A., Forsberg, B., Modig, L., Havulinna, A.S., Lanki, T., Turunen, A., Oftedal, B., Nystad, W., Nafstad, P., De Faire, U., Pedersen, N.L., Östenson, C.G., Fratiglioni, L., Penell, J., Korek, M., Pershagen, G., Eriksen, K.T., Overvad, K., Ellermann, T., Eeftens, M., Peeters, P.H., Meliefste, K., Wang, M., Bueno-de Mesquita, B., Sugiri, D., Krämer, U., Heinrich, J., de Hoogh, K., Key, T., Peters, A., Hampel, R., Concin, H., Nagel, G., Ineichen, A., Schaffner, E., Probst-Hensch, N., Künzli, N., Schindler, C., Schikowski, T., Adam, M., Phuleria, H., Vilier, A., Clavel-Chapelon, F., Declercq, C., Grioni, S., Krogh, V., Tsai, M.Y., Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Katsoulis, M., Trichopoulou, A., Brunekreef, B., Hoek, G., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. The Lancet 383, 785–795.

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., Dėdelė, A., Dimakopoulou, K., Eriksen, K., Galassi, C., Gražulevičienė, R., Grivas, G., Gruzieva, O., Gustafsson, A.H., Hoffmann, B., Iakovides, M., Ineichen, A., Krämer, U., Lanki, T., Lozano, P., Madsen, C., Meliefste, K., Modig, L., Mölter, A., Mosler, G., Nieuwenhuijsen, M., 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.Y., Vaskövi, É., Villani, S., Wang, M., Brunekreef, B., Hoek, G., 2012. Variation of NO2 and NOx concentrations between and within 36 European study areas: Results from the ESCAPE study. Atmospheric Environment 62, 374–390.

Dekoninck, L., Botteldooren, D., Int Panis, L., 2015. Using city-wide mobile noise assessments to estimate bicycle trip annual exposure to Black Carbon. Environment International 83, 192–201.

Dons, E., Van Poppel, M., Int Panis, L., De Prins, S., Berghmans, P., Koppen, G., Matheeussen, C., 2014. Land use regression models as a tool for short, medium and long term exposure to traffic related air pollution. Science of the Total Environment 476-477, 378–386.

Eeftens, M., Beekhuizen, J., Beelen, R., Wang, M., Vermeulen, R., Brunekreef, B., Huss, A., Hoek, G., 2013. Quantifying urban street configuration for improvements in air pollution models. Atmospheric Environment 72, 1–9.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dėdelė, A., Dons, E., de Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Gražulevičienė, R., Heinrich, J., Hoffmann, B., Jerrett, M., Keidel, D., Korek, M., Lanki, T., Lindley, S., Madsen, C., Mölter, A., Nádor, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-Nielsen, O., Patelarou, E., Quass, U., Ranzi, A., Schindler, C., Stempfelet, M., Stephanou, E., Sugiri, D., Tsai, M.Y., Yli-Tuomi, T., Varró, M.J., Vienneau, D., von Klot, S., Wolf, K., Brunekreef, B., Hoek, G., 2012. 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. Environmental Science & Technology 46, 11195–205.

Fruin, S.A., Urman, R., Lurmann, F., McConnell, R., Gauderman, J., Rappaport, E., Franklin, M., Gilliland, F.D., Shafer, M., Gorski, P., Avol, E., 2014. Spatial variation in particulate matter components over a large urban area. Atmospheric Environment 83, 211–219.

Ghassoun, Y., Ruths, M., Löwner, M.O., Weber, S., 2015. Intra-urban variation of ultrafine particles as evaluated by process related land use and pollutant driven regression modelling. Science of the Total Environment 536, 150–160.

Hankey, S., Marshall, J.D., 2015. Land Use Regression models of on-road particulate air pollution (Particle Number, Black Carbon, PM2.5, Particle Size) using mobile monitoring. Environmental Science & Technology 49, 9194–9202.

Hasenfratz, D., Saukh, O., Walser, C., Hueglin, C., Fierz, M., Arn, T., Beutel, J., Thiele, L., 2015. Deriving high-resolution urban air pollution maps using mobile sensor nodes. Pervasive and Mobile Computing 16, 268–285.

Henderson, S.B., Beckerman, B., Jerrett, M., Brauer, M., 2007. Application of Land Use Regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. Environmental Science & Technology 41, 2422–2428.

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

de Hoogh, K., Korek, M., Vienneau, D., Keuken, M., Kukkonen, J., Nieuwenhuijsen, M.J., Badaloni, C., Beelen, R., Bolignano, A., Cesaroni, G., Pradas, M.C., Cyrys, J., Douros, J., Eeftens, M., Forastiere, F., Forsberg, B., Fuks, K., Gehring, U., Gryparis, A., Gulliver, J., Hansell, A.L., Hoffmann, B., Johansson, C., Jonkers, S., Kangas, L., Katsouyanni, K., Künzli, N., Lanki, T., Memmesheimer, M., Moussiopoulos, N., Modig, L., Pershagen, G., Probst-Hensch, N., Schindler, C., Schikowski, T., Sugiri, D., Teixidó, O., Tsai, M.Y., Yli-Tuomi, T., Brunekreef, B., Hoek, G., Bellander, T., 2014. Comparing land use regression and dispersion modelling to assess residential exposure to ambient air pollution for epidemiological studies. Environment International 73, 382–92.

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

Kanaroglou, P.S., Adams, M.D., De Luca, P.F., Corr, D., Sohel, N., 2013. Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model. Atmospheric Environment 79, 421–427.

Larson, T., Henderson, S.B., Brauer, M., 2009. Mobile monitoring of particle light absorption coefficient in an urban area as a basis for land use regression. Environmental Science & Technology 43, 4672–4678.

Lefebvre, W., Van Poppel, M., Maiheu, B., Janssen, S., Dons, E., 2013. Evaluation of the RIO-IFDM-street canyon model chain. Atmospheric Environment 77, 325–337.

Merbitz, H., Fritz, S., Schneider, C., 2012. Mobile measurements and regression modeling of the spatial particulate matter variability in an urban area. Science of the Total Environment 438, 389–403.

Montagne, D.R., Hoek, G., Klompmaker, J.O., Wang, M., Meliefste, K., Brunekreef, B., 2015. Land use regression models for ultrafine particles and black carbon based on short-term monitoring predict past spatial variation. Environmental Science & Technology 49, 8712–8720.

Mueller, M., Hasenfratz, D., Saukh, O., Fierz, M., Hueglin, C., 2016. Statistical modelling of particle number concentration in Zurich at high spatio-temporal resolution utilizing data from a mobile sensor network. Atmospheric Environment 126, 171–181.

Patton, A.P., Collins, C., Naumova, E.N., Zamore, W., Brugge, D., Durant, J.L., 2014. An hourly regression model for ultrafine particles in a near-highway urban area. Environmental Science & Technology , 3272–3280.

Patton, A.P., Zamore, W., Naumova, E.N., Levy, J.I., Brugge, D., Durant, J.L., 2015. Transferability and generalizability of regression models of ultrafine particles in urban neighborhoods in the Boston Aarea. Environmental Science & Technology 49, 6051–6060.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peters, J., Van den Bossche, J., Reggente, M., Van Poppel, M., De Baets, B., Theunis, J., 2014. Cyclist exposure to UFP and BC on urban routes in Antwerp, Belgium. Atmospheric Environment 92, 31–43.

Petzold, A., Ogren, J.A., Fiebig, M., Laj, P., Li, S.M., Baltensperger, U., Holzer-Popp, T., Kinne, S., Pappalardo, G., Sugimoto, N., Wehrli, C., Wiedensohler, A., Zhang, X.Y., 2013. Recommendations for reporting "black carbon" measurements. Atmospheric Chemistry and Physics 13, 8365–8379.

Seabold, J., Perktold, J., 2010. Statsmodels: Econometric and Statistical Modeling with Python, in: van der Walt, S., Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference, pp. 57–61.

SGS, 2010. Strategische geluidsbelastingskaarten agglomeratie Antwerpen. Rapport in opdracht van Stad Antwerpen, 090357-2-v1, SGS Belgium NV. Technical Report.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14, 199–222.

Snyder, E.G., Watkins, T.H., Solomon, P.A., Thoma, E.D., Williams, R.W., Hagler, G.S.W., Shelow, D., Hindin, D.A., Kilaru, V.J., Preuss, P.W., 2013. The changing paradigm of air pollution monitoring. Environmental Science & Technology 47, 11369–11377.

Su, J.G., Hopke, P.K., Tian, Y., Baldwin, N., Thurston, S.W., Evans, K.,

Rich, D.Q., 2015. Modeling particulate matter concentrations measured through mobile monitoring in a deletion/substitution/addition approach. Atmospheric Environment 122, 477–483.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.

Van den Bossche, J., Peters, J., Verwaeren, J., Botteldooren, D., Theunis, J., De Baets, B., 2015. Mobile monitoring for mapping spatial variation in urban air quality: development and validation of a methodology based on an extensive dataset. Atmospheric Environment 105, 148–161.

Van den Bossche, J., Theunis, J., Elen, B., Peters, J., Botteldooren, D., De Baets, B., 2016. Opportunistic mobile air pollution monitoring: A case study with city wardens in Antwerp. Atmospheric Environment 141, 408–421.

Vardoulakis, S., Solazzo, E., Lumbreras, J., 2011. Intra-urban and street scale variability of BTEX, NO2 and O3 in Birmingham, UK: Implications for exposure assessment. Atmospheric Environment 45, 5069–5078.

Wang, M., Beelen, R., Basagana, X., Becker, T., Cesaroni, G., de Hoogh, K., Dedele, A., Declercq, C., Dimakopoulou, K., Eeftens, M., Forastiere, F., Galassi, C., Gražulevičienė, R., Hoffmann, B., Heinrich, J., Iakovides, M., Künzli, N., Korek, M., Lindley, S., Mölter, A., Mosler, G., Madsen, C., Nieuwenhuijsen, M., Phuleria, H., Pedeli, X., Raaschou-Nielsen, O., Ranzi, A., Stephanou, E., Sugiri, D., Stempfelet, M., Tsai, M.Y., Lanki, T., Udvardy, O., Varró, M.J., Wolf, K., Weinmayr, G., Yli-Tuomi, T., Hoek, G., Brunekreef, B., 2013. Evaluation of land use regression models for NO2 and particulate matter in 20 European study areas: the ESCAPE project. Environmental Science & Technology 47, 4357–64.

Wang, M., Beelen, R., Eeftens, M., Meliefste, K., Hoek, G., Brunekreef, B., 2012. Systematic evaluation of land use regression models for NO2. Environmental Science & Technology 46, 4481–9.

Weichenthal, S., Ryswyk, K.V., Goldstein, A., Bagg, S., Shekkarizfard, M., Hatzopoulou, M., 2016a. A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. Environmental Research 146, 65–72.

Weichenthal, S., Van Ryswyk, K., Goldstein, A., Shekarrizfard, M., Hatzopoulou, M., 2016b. Characterizing the spatial distribution of ambient ultrafine particles in Toronto, Canada: A land use regression model. Environmental Pollution 208, 241–248.

Wu, H., Reis, S., Lin, C., Beverland, I.J., Heal, M.R., 2015. Identifying drivers for the intra-urban spatial variability of airborne particulate matter components and their interrelationships. Atmospheric Environment 112, 306–316.

Appendix A. Supplementary material

Appendix A.1. GIS data sources

The different data sources for the predictor variables are described below (the derived variables are given in Table 1).

OpenStreetMap. OpenStreetMap is a collaborative and openly licensed map[3]. In particular, we used the information on roads (the *highway*-key in the database), including the type of road (primary, secondary, motorway, etc.). The distance from the measurement location to the closest road was calculated, as well as the total road length in surrounding buffers.

Urban Atlas. The Urban Atlas[4] is providing pan-European comparable land use and land cover data for large urban zones based on satellite images. Land use classes including residential, industrial, port, airport and green urban areas were used.

Address locations. The Central Reference Address Database (CRAB) is a freely available[5] database containing street names, house numbers and information about the geographical positioning of addresses for Flanders. The address positions were used to calculate the number of addresses in the surrounding area, which could possibly link the air quality to domestic emissions (e.g. heating).

Traffic data. Traffic intensity data was obtained from a traffic model specifically developed for the city of Antwerp (SGS, 2010) and previously applied in a study of Lefebvre et al. (2013). The model provides average daytime traffic intensity for light, medium and heavy traffic, and for all streets in the study area. Different variables were calculated: the total traffic load (sum of light, medium and heavy traffic), the heavy traffic load and the fraction of heavy to total traffic load.

Sky view factor. The sky view factor (SVF) is a measure of the total fraction of visible sky from the position of an observer on the ground. This measure can be used as a street canyon indicator (Eeftens et al., 2013). The SVF can be calculated using 3-dimensional building data or a digital surface model, but for the city of Antwerp the SVF data were made available as open data[6].

Cycling-specific data. Thanks to the Province of Antwerp and the Fietsersbond (cyclists association), there is an extensive dataset available describing several aspects of biking lanes, including the distance to the traffic lane. This information was used to construct an extra variable (distance_to_traffic), which describes the estimated distance of the measurement location (where the cyclist was positioned) to the traffic. For those locations where no biking lane or no data were present, the distance to the nearest road (dist_near, calculated based on the OpenStreetMap data) was used to fill the missing locations.

All the derived variables are given in Table 1. Additionally, some transformations and combinations of the predictor variables were included as well: 1/distance_to_traffic, 1/distance_to_traffic^2, 1/dist_near, 1/dist_near^2, 1/dist_near_major, 1/dist_near_major^2, log(distance_to_traffic), log(dist_near), trafload_50/distance_to_traffic, trafload_50/distance_to_traffic^2, trafload_50/dist_near, trafload_50/dist_near^2, trafnear/distance_to_traffic, trafnear/dist_near, trafnear/distance_to_traffic^2, trafnear/dist_near^2, trafnear_heavy/distance_to_traffic, trafnear_heavy/dist_near, trafnear_heavy/distance_to_traffic^2, trafnear_heavy/dist_near^2, skyview*trafload_50, trafload_50/skyview.

Appendix A.2. Variation in variables between spatial cross-validation zones

Appendix A.3. Selected predictor variables

[3] https://www.openstreetmap.org/about
[4] http://www.eea.europa.eu/data-and-maps/data/urban-atlas
[5] https://www.agiv.be/international/en/products/crab-en
[6] http://opendata.antwerpen.be/datasets/skyviewfactor-hittekaart

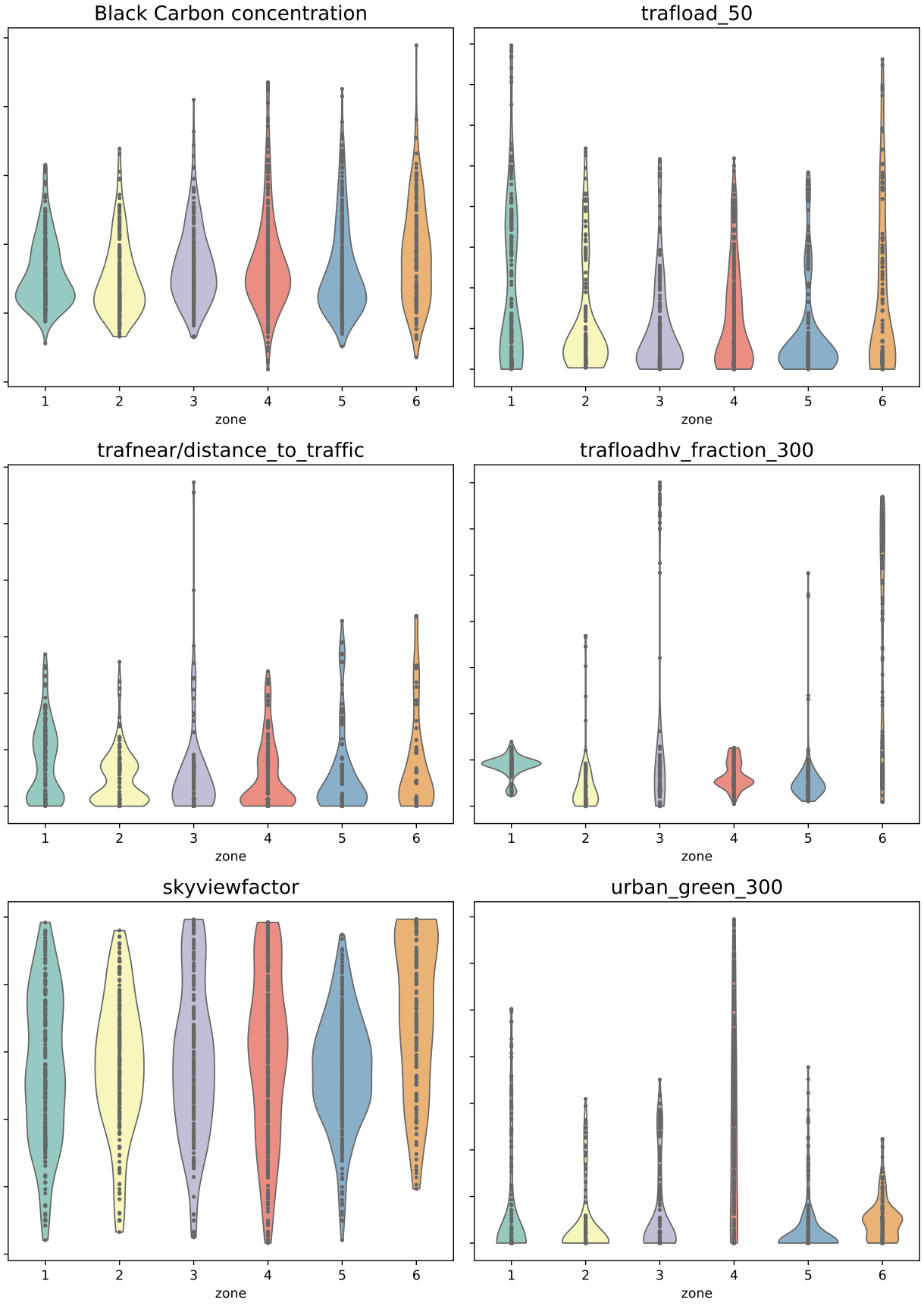

Figure S1: Violin plots to illustrate the variation between the different spatial cross-validation zones (see Figure 1 for the zones) for the measured BC concentrations as well as a selection of predictor variables. Labels on the y-axis have been left out since the predictor variables are scaled when building the models.

Table S1: Overview of the selected predictor variables and performance metrics for the classic linear model. Results are shown for the CV_F model based on all data and for the different models obtained in each iteration of the spatial CV_{S+F} cross-validation.

	CV_F (model built using all data)	CV_{S+F} (cross-validation with rebuilding) 1	2	3	4	5	6	pooled
Performance								
R^2	0.35	-0.65	0.06	0.30	0.25	0.40	0.52	0.24
EV	0.35	0.28	0.36	0.37	0.28	0.40	0.53	0.25
RMSE	1.05	1.26	1.07	0.95	1.28	1.07	1.03	1.13
Selected features								
address_1000	-	-	-	x	-	-	-	
airport_5000	x	-	x	x	-	x	x	
industry_5000	-	-	-	-	x	-	-	
port_3000	-	-	-	-	-	x	-	
port_5000	-	-	x	-	-	-	x	
res_hd_100	-	x	-	-	-	-	-	
res_hd_1000	x	x	-	-	x	x	x	
trafload_100	x	-	x	x	x	-	x	
trafload_50	x	x	x	x	-	x	x	
trafload_50/distance_to_traffic	-	-	-	-	x	-	-	
trafload_heavy_300	-	x	-	-	-	x	-	
trafnear	-	-	-	-	x	-	-	
trafnear/distance_to_traffic	x	x	x	x	-	x	x	
urban_green_300	x	-	x	-	-	x	x	

Table S2: Overview of the selected features and performance metrics for the forward CV SVR model. Results are shown for the CV_F model based on all data and for the different models obtained in each iteration of the spatial CV_{S+F} cross-validation.

	CV_F (model built using all data)	CV_{S+F} (cross-validation with rebuilding) 1	2	3	4	5	6	pooled
Performance								
R^2	0.40	0.19	0.24	-0.03	0.18	0.37	0.24	0.24
EV	0.41	0.26	0.31	0.17	0.23	0.41	0.35	0.26
RMSE	1.00	0.88	0.96	1.15	1.33	1.09	1.29	1.14
Selected features								
address_100	-	-	-	-	-	x	-	
airport_1000	-	-	x	-	-	x	-	
airport_3000	-	-	-	-	-	-	x	
industry_1000	-	-	-	-	x	-	-	
port_3000	-	-	-	x	-	-	-	
roadlength_300	-	-	-	x	-	-	-	
skyviewfactor	x	x	x	-	x	x	-	
trafload_50	x	-	x	x	-	x	-	
trafload_50/skyview	-	x	-	-	x	-	x	
trafload_heavy_300	-	x	-	-	-	-	-	
trafloadhv_fraction_1000	-	-	-	-	-	-	x	
trafloadhv_fraction_300	x	-	x	-	x	x	-	
trafnear	-	-	-	-	x	-	-	
trafnear/dist_near	-	x	-	-	-	-	-	
trafnear/distance_to_traffic	-	-	-	-	-	-	x	
trafnear_heavy/dist_near	x	x	x	-	-	x	-	
trafnear_heavy/distance_to_traffic	-	-	-	x	-	-	x	
urban_green_500	-	-	x	x	-	-	-	