Environmental Modelling & Software 99 (2018) 58–69

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

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

ARTICLE INFO

Article history:
Received 31 March 2017
Received in revised form
31 August 2017
Accepted 29 September 2017

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

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.

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.

* Corresponding author. KERMIT, Dept. of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, 9000 Ghent, Belgium.
E-mail address: jorisvandenbossche@gmail.com (J. Van den Bossche).

https://doi.org/10.1016/j.envsoft.2017.09.019
1364-8152/

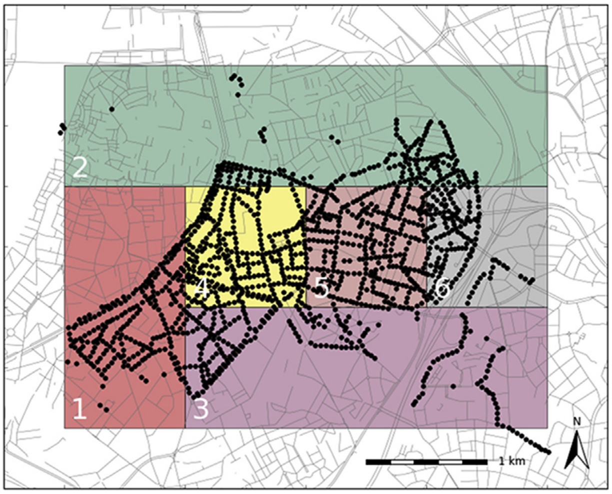

Fig. 1. The different spatial zones for cross-validation. The zones are constructed as 1×1 km^2 areas based on the UTM coordinates. Some of the zones with fewer sampling locations are combined into one zone, resulting in six zones as indicated with numbers in the figure.

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-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 Fig. 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 μg 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 h of raw 1 s measurements were recorded for the three teams combined (459 h of measurements before filtering for GPS quality), spread over 110 days. Most of the measurements were done between 10 a.m. and 16 p.m. 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

[1] The dataset is available upon request.

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