Environmental Research 146 (2016) 65–72

Contents lists available at ScienceDirect

Environmental Research

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

A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach

Scott Weichenthal [a,d,*], Keith Van Ryswyk [a], Alon Goldstein [b], Scott Bagg [b], Maryam Shekkarizfard [c], Marianne Hatzopoulou [e]

[a] Air Health Science Division, Health Canada, Ottawa, Canada
[b] School of Urban Planning, McGill University, Montreal, Canada
[c] Department of Civil Engineering, McGill University, Montreal, Canada
[d] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada
[e] Department of Civil Engineering, University of Toronto, Toronto, Canada

ARTICLE INFO

Article history:
Received 9 October 2015
Received in revised form
9 December 2015
Accepted 14 December 2015
Available online 22 December 2015

Keywords:
Ultrafine particles
Land use regression
Traffic
Built environment

ABSTRACT

Existing evidence suggests that ambient ultrafine particles (UFPs) (< 0.1 μm) may contribute to acute cardiorespiratory morbidity. However, few studies have examined the long-term health effects of these pollutants owing in part to a need for exposure surfaces that can be applied in large population-based studies. To address this need, we developed a land use regression model for UFPs in Montreal, Canada using mobile monitoring data collected from 414 road segments during the summer and winter months between 2011 and 2012. Two different approaches were examined for model development including standard multivariable linear regression and a machine learning approach (kernel-based regularized least squares (KRLS)) that learns the functional form of covariate impacts on ambient UFP concentrations from the data. The final models included parameters for population density, ambient temperature and wind speed, land use parameters (park space and open space), length of local roads and rail, and estimated annual average NO_x emissions from traffic. The final multivariable linear regression model explained 62% of the spatial variation in ambient UFP concentrations whereas the KRLS model explained 79% of the variance. The KRLS model performed slightly better than the linear regression model when evaluated using an external dataset (R^2=0.58 vs. 0.55) or a cross-validation procedure (R^2=0.67 vs. 0.60). In general, our findings suggest that the KRLS approach may offer modest improvements in predictive performance compared to standard multivariable linear regression models used to estimate spatial variations in ambient UFPs. However, differences in predictive performance were not statistically significant when evaluated using the cross-validation procedure.

1. Introduction

Ambient ultrafine particles (UFPs) (< 0.1 μm) may contribute to acute cardiovascular morbidity including changes in heart rate variability and endothelial function (Weichenthal, 2012). However, little is known about the long-term health effects of these traffic pollutants owing in part to a need for exposure surfaces suitable for use in large population-based studies. Recently, Ostro et al. (2015) used a chemical transport model to examine the relationship between UFP and cardiovascular mortality and reported an increased risk of ischemic heart disease mortality among participants in the California Teachers Study Cohort. Other studies of the long-term health effects of UFPs have not been conducted to date but land use regression models have been developed for several cities including Vancouver (Abernethy et al., 2013) and Toronto, Canada (Sabaliauskas et al., 2015; Weichenthal et al., 2016), Barcelona, Spain (Rivera et al., 2012), and Amsterdam, Netherlands (Hoek et al., 2011). In general, these models suggest that within-city spatial variations in ambient UFPs can be predicted using various land use, traffic, and meteorological parameters with R^2 values generally exceeding 50%. Moreover, Klompmaker et al. (2015) demonstrated that short-term monitoring campaigns may be an efficient means of developing land use regression models for ambient UFPs and that these models may provide reasonable estimates of historical spatial contrasts. In developing such models,

* Corresponding author at: Air Health Science Division, Health Canada, 269 Laurier Ave West, Ottawa, Ontario, Canada K1A 0K9.
E-mail address: scott.weichenthal@hc-sc.gc.ca (S. Weichenthal).

http://dx.doi.org/10.1016/j.envres.2015.12.016
0013-9351/

mobile monitoring offers an efficient means of data collection as recently highlighted by studies in Toronto, Canada (Weichenthal et al., 2016) and Minneapolis, USA (Hankey and Marshall, 2015a). In this study, we developed a land use regression model for ambient UFPs in Montreal, Canada using data collected with both bicycle and vehicle-based mobile platforms. In doing so, we examined two different approaches including standard multivariable linear regression and a machine learning method (kernel-based regularized least squares (KRLS)) that does not impose strong function form assumptions on covariate impact on ambient UFP concentrations.

2. Methods

2.1. Mobile monitoring of ultrafine particles

Mobile monitoring data for ambient UFPs were collected at 1-s resolution using portable condensation particle counters (TSI CPC Model 3007) mounted on bicycles (for summer monitoring) and vehicle roof –racks (for winter monitoring). Details of the two monitoring campaign are described in detail elsewhere (Weichenthal et al., 2015; Farrell et al., 2015). Briefly, winter UFP data were collected using three separate vehicles (Chevrolet Grand Caravans) driving for six hours a day (between 7:00–10:00 and 15:00–18:00) for 5 consecutive weekdays in March 2011. Time periods were selected to capture peak ambient concentrations and also to allow for time to download and process the data between trips each day. Each vehicle focused on covering a different area of the city including downtown areas, major highways, and suburban areas; the spatial coverage of the winter monitoring campaign is illustrated in Supplemental Fig. 1.

The bicycle monitoring campaign took place on 23 weekdays during the months of June and July, 2012. All cycling took place between 8:00–10:00 and 15:00–17:00. Two pairs of research assistants used condensation particle counters (TSI CPC Model 3007) affixed to bicycles to measure UFP concentrations along 25 routes charted around the Island of Montreal. The routes were designed to cover both downtown and suburban locations, urban canyons and low built-up areas (i.e. areas with 2–3 storey buildings). Each route was a circuit of approximately 25 km in circumference. The extent of the network is presented in Supplemental Fig. 2. In total, over 475 km of unique roadways were covered.

Ambient temperature and relative humidity data were collected on mobile platforms at the same time as UFP data at 1-s resolution. Mean wind speed data were collected from the nearest Environment Canada site and matched to the time of data collection. All UFP and meteorological data were pooled and averaged for each individual road segment.

2.2. Assigning ultrafine particle concentrations to road segments

All air quality data were matched with their respective GPS coordinates based on the time-stamp of the recording (at a frequency of 1 Hz). Every GPS reading coupled with a UFP level was then associated with the road segment where the monitoring was designed to occur based on the initial identification of daily trajectories. A road segment is defined as a link between two successive intersections; road segments had a mean length of 377 m (interquartile range: 159–415 m). In the case of the cycling data, points were also related with a non-motorized trail if it was ridden on or alongside, as is the case when riding within parks. All UFP data (i.e. from monitoring campaigns over both seasons) associated with each road segment were averaged (i.e. by pooling data from both surveys) and the number of GPS points or seconds associated with the mean UFP per segment was recorded. All analyses are based on mean UFP data assigned to road segments over the entire monitoring campaign Moreover, both monitoring campaigns were designed so that the distributions of visits across days and time periods would remain relatively stable across locations.

Table 1
Descriptive statistics for UFP concentrations (count/cm^3).

Statistic	UFP
Minimum	5689
10th percentile	14,165
First quartile	18,765
Mean (SD)	39,199 (34,582)
Median	26,497
Third quartile	48,236
90th percentile	83,762
Maximum	234,976

Data reflects a total of 414 road segments with at least 200 points/segment.

The number of data points available for each road segment varied depending on the number of times it was traversed during the mobile monitoring campaigns. All statistical analyses are based on road segments with at least 200 points/segment (mean: 405 points/segment; interquartile range: 235–449) as this cut-off provided the best balance of spatial coverage and points/segment. As sensitivity analysis, a multivariable linear regression model was also examined using road segments with at least 250 points/segment and this did not change the results (Supplemental Table 3 and 4); therefore, we selected the lower cut-point to increase spatial coverage (Tables 1 and 2).

2.3. Derivation of land use and built environment data for model development

Each road segment was associated with a number of land-use and built environment characteristics. These included variables computed as distances between the mid-point of the road segment and potential sources of UFP (e.g. nearest highway, nearest major road, nearest bus route, and Trudeau International Airport). In addition, a number of land-use variables were computed within buffers of sizes ranging from 100 to 300 m. These include: number of bus stops, length of bus routes (in meters), length of rail lines, number of restaurants, number of trees, length of expressways (in meters), length of primary highways (in meters), length of secondary highways (in meters), length of major roads (in meters), length of local roads (in meters), population density (number of individuals/km^2), number of trees, and proportion land occupied by different land-use types (e.g. commercial, governmental/institutional, open areas, parks/recreational, residential, resource/industrial, water body). The decision to limit buffers to a maximum of 300 m was based on the fact that UFPs are highly dominated by local emissions occurring in the direct vicinity of each sampling location. In addition, the magnitude of covariate impacts on ambient UFPs tended to decrease with increasing buffer sizes (Table 3).

In addition, we made use of prior research into a mesoscopic traffic simulation model that was developed for the Greater Montreal Area (Sider et al., 2013). The model generated outputs at the level of the road segment for vehicular composition, volume, and speed. In order to refine our measure of road traffic, we used the output of the same traffic assignment model and transformed traffic volumes, compositions, and speeds into a measure of daily nitrogen oxide (NO_x) emissions per road segment (in grams). In order to calculate the NO_x emissions potentially affecting each