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

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

Table 2
Descriptive statistics for candidate predictor variables.

Independent variable	Buffer size (m)	Mean (SD)	Minimum	Maximum
Distance to airport (m)		13,600 (4360)	2741	31,483
Land use proportions				
Residential	100	0.41 (0.33)	0	1
	200	0.42 (0.27)	0	1
	300	0.42 (0.25)	0	0.96
Commercial	100	0.052 (0.12)	0	0.67
	200	0.054 (0.10)	0	0.57
	300	0.051 (0.084)	0	0.60
Industrial	100	0.19 (0.26)	0	1
	200	0.20 (0.24)	0	1
	300	0.19 (0.22)	0	0.96
Park space	100	0.10 (0.23)	0	1
	200	0.10 (0.21)	0	1
	300	0.10 (0.19)	0	1
Open space	100	0.14 (0.22)	0	1
	200	0.11 (0.18)	0	1
	300	0.095 (0.16)	0	1
Water	100	0.032 (0.14)	0	1
	200	0.038 (0.14)	0	1
	300	0.042 (0.14)	0	1
Length of roadways(m)				
Expressways	100	109 (235)	0	1171
	200	251 (525)	0	3317
	300	421 (831)	0	4959
Primary highways	100	15.6 (77)	0	806
	200	52.5 (217)	0	2821
	300	107 (338)	0	3556
Secondary highways	100	1.1 (16.4)	0	273
	200	5.9 (46.8)	0	562
	300	11.6 (80.7)	0	772
Major roads	100	176 (218)	0	1726
	200	540 (525)	0	3009
	300	1031 (859)	0	3826
Local roads	100	267 (193)	0	1051
	200	1140 (648)	0	3057
	300	2614 (1332)	0	5771
Number of Restaurants	100	2.0 (4.9)	0	44
	200	7.9 (16)	0	97
	300	17 (31)	0	206
Number of Bus stops	100	0.85 (1.4)	0	7
	200	3.8 (3.4)	0	15
	300	7.9 (5.9)	0	27
Length of Bus routes (m)	100	573 (655)	0	3536
	200	1871 (1600)	0	9883
	300	3693 (2698)	0	15,239
Length of Rail (m)	100	78 (160)	0	1214
	200	312 (605)	0	4320
	300	631 (1141)	0	8516
Number of Trees	100	25.8 (30)	0	247
	200	101 (108)	0	721
	300	227 (227)	0	1614
Annual average NO_x emissions (g)	100	2897 (4725)	0	18,929
	200	7301 (9873)	0	53,363
	300	12,745 (15,207)	0	87,480
Population density (km^2)	100	6643 (6645)	0	45,297
	200	6574 (6063)	0	41,023
	300	6445 (5338)	0	26,811
Ambient temperature (°C)		16.4 (9.9)	−4.3	28.5
Relative humidity (%)		49.8 (10.4)	0.0	84.0
Wind speed (m/s)		14.3 (5.6)	0.0	51.3

Table 3
Single-predictor linear regression models for mean UFP concentrations in Montreal, Canada.

Independent variable	Buffer (m)	Linear regression models β (95% CI)	R^2	RMSE
Distance to airport		−2.07 (−2.81, −1.33)	0.068	33,420
ln(distance to airport)		−30,482 (−39,541, −21,424)	0.096	32,920
Land use proportions				
Residential	100	−13,887 (−24,057, −3717)	0.017	34,325
	200	−7981 (−20,215, 4252)	0.004	34,555
	300	−3463 (−17,099, 10,172)	0.00	34,614
Commercial	100	3755 (−24,650, 32,160)	0.00	34,621
	200	548 (−33,222, 34,319)	0.00	34,624
	300	6654 (−32,996, 46,306)	0.00	34,620
Industrial	100	3258 (−9451, 15,969)	0.00	34,614
	200	10,061 (−4057, 24, 179)	0.00	34,542
	300	9913 (−5350, 25, 178)	0.00	34,556
Park space	100	−26,033 (−40,182, −11,884)	0.03	34,087
	200	−30,461 (−46,305, −14,618)	0.03	34,039
	300	−32,710 (−50,182, −15,238)	0.03	34,069
Open space	100	50,118 (35,723, 64,514)	0.10	32,810
	200	42,109 (23,886, 60,332)	0.05	33,789
	300	37,360 (16,876, 57,844)	0.03	34,097
Water	100	12,949 (−11,088, 36,988)	0.00	34,577
	200	7787 (−15,307, 30,881)	0.00	34,606
	300	7372 (−16,473, 31,219)	0.00	34,609
Length of roadways				
Expressways	100	78 (66, 90)	0.28	29,361
	200	32 (27, 38)	0.24	30,153
	300	19 (16, 23)	0.22	30,579
Primary highways	100	−14 (−57, 30)	0.00	34,608
	200	−1.8 (−17, 14)	0.00	34,622
	300	−2.4 (−12, 7.5)	0.00	34,615
Secondary highways	100	−45.5 (−250, 159)	0.00	34,616
	200	−44 (−115, 28)	0.00	34,563
	300	−32 (−73, 10)	0.00	34,530
Major roads	100	30 (15, 45)	0.04	33,994
	200	10 (4.0, 17)	0.02	34,190
	300	5.5 (1.6, 9.3)	0.02	34,303
Local roads	100	−49 (−66, −33)	0.08	33,288
	200	−9.4 (−14, −4.3)	0.03	34,087
	300	−3.5 (−5.9, −0.97)	0.02	34,316
Bus stops	100	−1063 (−3433, 1307)	0.00	34,592
	200	416 (−561, 1392)	0.00	34,595
	300	−37 (−605, 531)	0.00	34,624
Length of bus routes	100	10 (5.4, 15)	0.04	33,949
	200	3.7 (1.6, 5.8)	0.03	34,114
	300	1.8 (0.61,3.1)	0.02	34,265
Length of rail	100	35 (14,56)	0.03	34,165
	200	7.4 (1.9, 13)	0.02	34,336
	300	4.0 (1.1, 6.9)	0.02	34,318
Number of Restaurants	100	28 (−655, 711)	0.00	34,624
	200	−57 (−270, 157)	0.00	34,613
	300	−64 (−171, 42)	0.00	34,566
Number of Trees	100	−163 (−271, −55)	0.02	34,261
	200	−41 (−72, −10)	0.02	34,342
	300	−16 (−30, −1.1)	0.01	34,437
Annual Average NO_x	100	4.7 (4.2, 5.3)	0.42	26,482
	200	2.2 (1.9, 2.4)	0.38	27,202
	300	1.3 (1.2, 1.5)	0.35	27,934
Population density	100	−0.65 (−1.12, −0.15)	0.02	34,353
	200	−0.76 (−1.3, −0.21)	0.02	34,315
	300	−0.92 (−1.5, −0.29)	0.02	34,276
Ambient temperature		−2287 (−2543, −2030)	0.43	26,208
Relative humidity		−777 (−1088, −465)	0.06	33,657
Wind speed		−1772 (−2342, −1202)	0.08	33,154

Variables included in the final model are in bold.

road segment monitored, we generated buffers of different sizes (50–300 m) around the midpoint of every segment and intersected with a map of NO_x emissions (in grams) on the road network. The sum of NO_x emissions occurring within each buffer was then extracted. Buffering and intersections were done using ArcMap 10.2. The final exposure surface for UFPs was generated first by superimposing a raster of 100×100 m^2 grid cells on the city of Montreal. Buffers were drawn around the centroids of each

grid cell in order to compile the set predictors for each cell; final model coefficients were then applied to each cell.

2.4. Statistical analysis

Land use regression models were constructed for mean UFP concentrations (and the natural logarithm of mean UFP concentrations) according to the following approach. First, single predictor linear regression models were examined for each candidate predictor to identify parameters that were significantly associated (95% CI excluded the null) with ambient UFPs. The list of candidate predictors evaluated is provided in Table 2 and included land use categories (e.g. residential, commercial, industrial), length of rail, roadways, and bus routes, number of restaurants, number of bus stops, number of trees, population density, estimate annual NO_x emissions (as a surrogate measure of traffic counts), and mean ambient temperature, relative humidity, and wind speed. Candidate predictor variables that were associated with mean UFP concentrations in single predictor models were considered in final multivariable models. Spearman's correlations were examined between candidate predictors and highly correlated variables ($r > 0.80$) were eliminated from the analysis; we retained the variable that was most strongly associated with ambient UFPs (i.e. the largest R^2 and lowest root mean square error (RMSE)). Since all road segments were not monitored simultaneously, mean ambient temperature was included in all final models in order to control for temporal differences in ambient UFP concentrations (even after averaging concentrations across monitoring campaigns) which are known to be inversely correlated with ambient temperature (Alm et al., 1999; Kaur and Nieuwenhuijsen, 2009; Weichenthal et al., 2008; 2014a, 2015, 2016). We did not place a priori restrictions on the directions of observed relationships between candidate predictors and UFPs as the primary purpose of the model was prediction. Both linear and quadratic terms were evaluated for ambient temperature to account for potential non-linearity in the relationship between ambient temperature and UFPs.

The final multivariable linear regression model was selected by first including all candidate predictors that were associated with UFPs in single predictor models. Candidate predictors that were not significantly associated with UFPs in the multivariable model were only removed if doing so decreased or did not substantially change ($< 1\%$) the RMSE of the model. In addition to the multivariable linear regression model outlined above, a second multivariable model was selected using 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 (Ferwerda et al., 2013; Hainmueller and Hazlett, 2013). The KRLS program in Stata uses Tikhonov regularization to preferentially select smoother, less complicated functions and to limit over-fitting (Ferwerda et al., 2013; Hainmueller and Hazlett, 2013). In this study, the KRLS model included the same covariates as the multivariable linear regression model above but allowed covariate impacts on ambient UFP concentrations to vary across the parameter space. This analysis was implemented using the *krls* command in Stata (version 13) (Statacorp, College Station, Texas), and provides an estimate of the pointwise marginal effect of a given parameter on UFP concentrations (like a β coefficient from linear regression) along with heterogeneity in the marginal effect expressed as an interquartile range (25th–75th) across the parameter space.

Bias and precision of model estimates were evaluated using linear regression models relating measured and predicted values for ambient UFPs; separate models were examined for multivariable linear regression and KRLS models. First, models developed using data from road segments with at least 200 points/segment were evaluated using data from road segments with 100–199 points/segment. The slopes of these models provided an estimate of the strength of the linear relationship between measured and predicted values. In addition, a 10-fold cross validation procedure was conducted whereby models were developed using 90% of the data and tested on the remaining 10%. This procedure was repeated 10 times so that all of the data were used at least once for both model development and evaluation. Mean alpha, beta, R^2 and RMSE values (and 95% confidence intervals) were calculated for the multivariable linear regression model and the KRLS model using the cross validation data. Mover, mean differences (and 95% confidence intervals) in these parameters were evaluated between the two models.

3. Results

In total, 414 road segments had at least 200 data points/segment and were used for model development (Supplemental Fig. 3). Ambient temperatures during winter monitoring ranged from 3.7–15 °C (mean = 9.4 °C) whereas ambient temperatures during summer monitoring ranged from 20–28 °C (mean = 24.8 °C). Ambient UFP concentrations and candidate predictors varied substantially across road segments (Tables 1–2) and a number of parameters were identified as potential predictors of ambient UFP. In particular, the following parameters were associated with ambient UFPs in single predictor models: distance to Trudeau International Airport, residential land use (100 m buffer), park space (200 m buffer), open space (100 m buffer), length of expressways, major roads, local roads, rail, and bus routes (100 m buffer), number of trees (100 m buffer), estimated annual average NO_x (100 m buffer), population density (300 m buffer), ambient temperature, relative humidity, and wind speed. None of these parameters were strongly correlated ($r \leq 0.69$) and thus none were eliminated from final multivariable models because of collinearity. In general, length of expressways, estimated annual average NO_x emissions, and ambient temperature were most strongly associated with ambient UFPs in single predictor models with other variables explaining less than 10% of the variation in ambient concentrations.

The final multivariable model for mean UFP concentrations is shown in Table 4 and included terms for ambient temperature and wind speed, park space (200 m buffer), open space (100 m buffer), length of rail and local roads (100 m buffers), estimated annual average NO*x* emissions (100 m buffer), and population density (300 m buffer). The model for the natural logarithm of ambient UFPs is shown in Supplemental Table 1 and explained a similar proportion of the variance in ambient UFP concentrations. Coefficients for park space and open space were not statistically significant in the mean UFP model; however, removing these parameters decreased the R^2 value by approximately 10% and increased the RMSE value by 12% and thus these parameters were retained in the model. Including a quadratic term for temperature or an interaction term between temperature and wind speed did not improve model fit and thus these predictors were not included in the final model (data not shown). Variance inflation factors ranged from 1.05–1.81 (mean = 1.43) suggesting limited collinearity among variables in the model.

The predicted spatial distribution of ambient UFPs in Montreal, Canada is shown in Fig. 1 (temperature and wind speed parameters were set to mean values shown in Table 2).

The final KRLS model is also shown in Table 4 and explained a greater proportion of the variance in ambient UFP concentrations than the final multivariable linear regression model (79% vs. 62%). Moreover, evidence of nonlinearity was apparent for a number of the parameters in the model as shown by the interquartile ranges

Table 4
Final models for mean UFP concentrations in Montreal, Canada ($n=414$).

Model	Alpha	Independent variables	Effect estimate (95% CI)	Adjusted R^2	RMSE	KRLS[a] 25th–75th
Linear regression	72,882	Temperature	−1322 (−1593, −1052)	0.62	21,411	–
		Wind speed	−1534 (−1912, −1156)			
		Park space[b]	−5176 (−16,248, 5895)			
		Open space[c]	−9705 (−22,040, 2630)			
		Local roads[c]	−14.8 (−27, −2.1)			
		Length of rail[c]	23 (9.4, 36)			
		Annual NO_x[c]	3.4 (2.7, 3.9)			
		Population density[d]	0.66 (0.19, 1.1)			
KRLS	–	Temperature	−1049 (−1262, −835)	0.79	–	−1322, −714
		Wind speed	−885 (−1367, −403)			−1348, −391
		Park space[b]	1635(−10,682, 13,952)			−12,976, 12,667
		Open space[c]	7825 (−5583, 21,233)			−10,302, 23, 354
		Local roads[c]	−13 (−22, −4.0)			−21.5, −1.0
		Length of rail[c]	25 (9.2, 41)			5.5, 37
		Annual NO_x[c]	0.82 (0.056, 1.6)			0.23, 1.3
		Population density[d]	0.70 (0.35, 1.05)			0.0072, 1.1

KRLS, kernel-based regularized least squares.

[a] Interquartile range of pointwise marginal effects in the KRLS model.
[b] 200 m buffer.
[c] 100 m buffer.
[d] 300 m buffer.

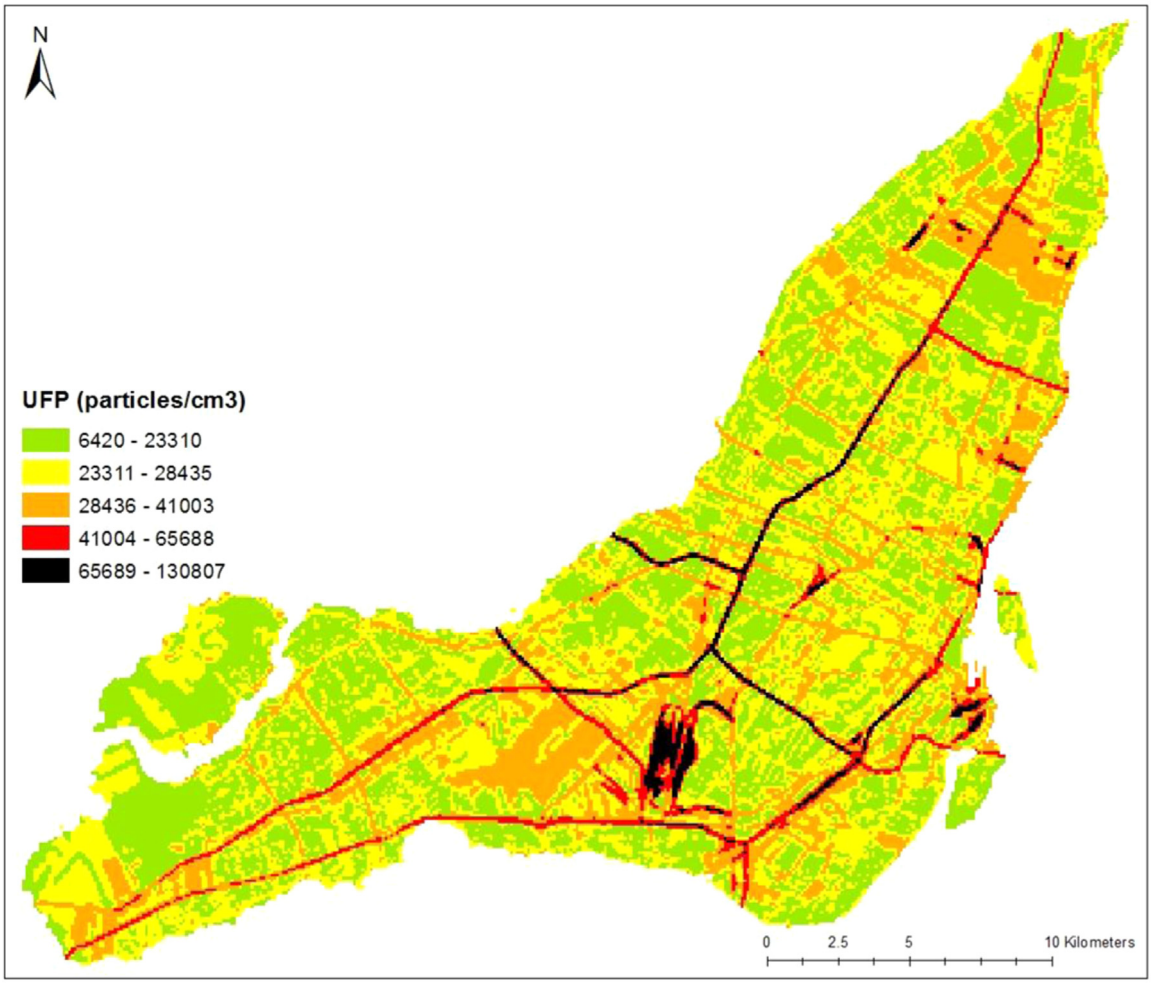

Fig. 1. Spatial distribution of ambient UFPs in Montreal, Canada.

of pointwise marginal effects in the last column of Table 4 (histograms of marginal effects are available in Supplemental Fig. 4). In general, the ability of the KRLS model to capture non-linear associations likely explains the improved performance of this modeling approach. The KRLS model for the natural logarithm of ambient UFPs is shown in Supplemental Table 1 and performed slightly better than the model for mean UFPs ($R^2=0.84$).

In total, model evaluation was conducted using data from 893

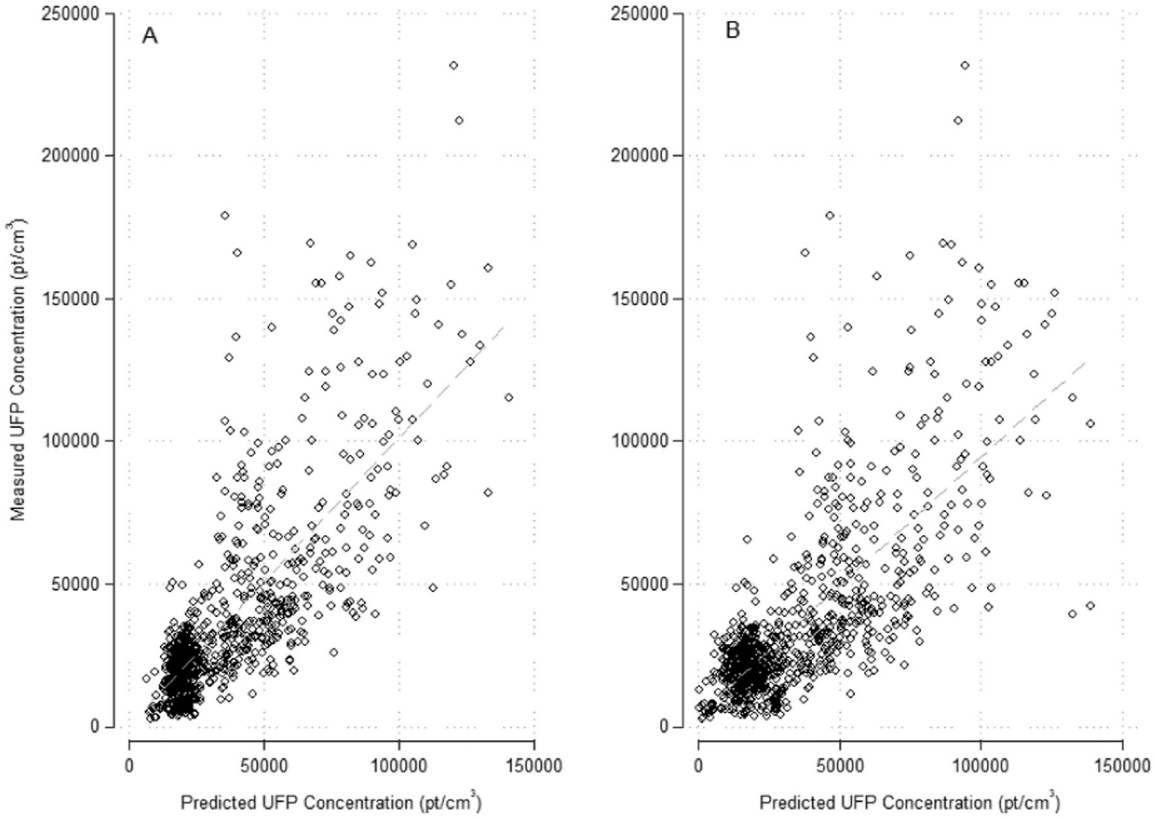

Fig. 2. Relationship between measured and predicted mean UFP concentrations using the KRLS (A) and multivariable linear regression (B). Model evaluation was conducted on an external dataset of 893 road segments not used for model development.

Table 5
Model evaluation using an external dataset (n=893).

Model	α (95% CI)	β[a] (95% CI)	R^2[a]	RMSE[a]
MVLR	4050 (1620, 6480)	0.90 (0.85, 0.96)	0.55	21,736
KRLS	404 (−2043, 2852)	1.00 (0.95, 1.06)	0.58	20,851

MVLR, multivariable linear regression; KRLS, kernel-based regularized least squares. Model evaluation is based on 893 road segments with at least 100–199 points/segment.

[a] For linear regression model relating measured and predicted values.

Table 6
Model evaluation using a 10-fold cross validation procedure.

Model	Mean α (95% CI)	Mean β[a] (95% CI)	Mean R^2 (95% CI)[a]	Mean RMSE (95% CI)[a]
MVLR	1741 (−3467, 6950)	0.94 (0.75, 1.1)	0.60 (0.54, 0.67)	20,264 (16,327, 24,200)
KRLS	−876 (−5693, 3940)	1.02 (0.89, 1.2)	0.67 (0.59, 0.77)	18,594 (14,155, 23,034)
Mean Difference (95% CI)	2618 (−3974, 9209)	−0.085 (−0.39, 0.13)	−0.063 (−0.16, 0.037)	1669 (−3846, 7185)

MVLR, multivariable linear regression; KRLS, kernel-based regularized least squares. Model evaluation is based on a cross validation procedure using 90% of the data to build the model and a 10% test sample, repeated 10 times.

[a] For linear regression model relating measured and predicted values.

road segments with between 100 and 199 points/segment (Fig. 2). On average, the multivariable linear regression model tended to underestimate mean UFPs by 4050 particles/cm^3 (95% CI: 1620, 6480) whereas no systematic bias was apparent for the KRLS model (Table 5). However, the R^2 value for the KRLS model decreased more than the linear regression model when evaluated in the external dataset suggesting that the KRLS method may have overfit the data. Nevertheless, model predictions from the KRLS model were more precise and the slope relating measured and predicted values was 1.00 (95%CI: 0.95, 1.06) compared to 0.90 (95% CI: 0.85, 0.96) for the multivariable linear regression model (Table 5). Similarly, the KRLS model performed slightly better than the multivariable linear regression model when evaluated using a 10-fold cross-validation procedure (Table 6). Specifically, the mean R^2 value for the KRLS model was 67% (95% CI: 59–77) whereas a mean value of 60% (95% CI: 54–67) was observed for the multivariable linear regression model. The RMSE value for the KRLS model was also slightly lower, although differences between model intercepts, slopes, R^2, and RMSE values were not statistically significant (Table 6).

4. Discussion

Short-term exposures to ambient UFPs have been associated with acute changes in physiological measures of cardiovascular health including endothelial function and heart rate variability (Weichenthal, 2012, 2014b); however, little is known about the long-term health effects of these pollutants. In this study we developed a land use regression model for ambient UFPs in Montreal, Canada using mobile monitoring data collected using both bicycle and vehicle platforms. This model will be used in future cohort studies to evaluate the chronic health risks of UFPs.

In general, our model explained the majority of the spatial variation in ambient UFPs on the island of Montreal and performed reasonably well when evaluated in an external dataset and

using a cross-validation procedure. In particular, the model developed using the KRLS method slightly outperformed the multivariable linear regression model although differences in predictive performance were not statistically significant when the results of the cross-validation procedure were compared. However, differences in model coefficients were apparent between the two models, particularly for annual average NO_x (which had a smaller coefficient in the KRLS mode) and open space and park space which had opposite directions between the two models (but were not statistically significant). These differences are likely explained by the fact that the KRLS method captures potential non-linear associations between candidate predictors and ambient UFPs and thus the estimated marginal effect of each parameter reflects heterogeneity across the parameter space (Ferwerda et al., 2013; Hainmueller and Hazlett, 2013). To our knowledge, this is the first study to compare machine learning and standard linear regression models in the development of predictive models for spatial differences in ambient UFPs and our overall findings suggest that the KRLS method may offer modest improvements over the standard approach.

The parameters included in our final model are generally consistent with those from other cities in Canada (Abernethy et al., 2013; Sabaliauskas et al., 2015; Weichenthal et al., 2016) and elsewhere (Hoek et al., 2011; Rivera et al., 2012) and largely reflect traffic sources and meteorology. Recent studies have identified airports as potentially important sources of ambient UFPs in both Los Angeles (Hudda et al., 2014) and Toronto (Weichenthal et al., 2016) and while airport proximity was associated with ambient UFPs in Montreal in single pollutant models it was not retained in the final model. However, the length of rail within a 100 m buffer was an important predictor of UFPs in Montreal and remained a significant predictor in final models. This is not surprising as diesel vehicles are known to be important sources of ambient UFPs (Hankey and Marshall, 2015b; Hatzopoulou et al., 2013; Weichenthal et al., 2015); however, to our knowledge this is the first land use regression model for UFPs to incorporate a parameter for the length of rail within a given buffer. To date, little (if any) research has focused on the impact of railway emissions on ambient UFP exposures and future studies should aim to characterize the impacts of non-road transportation sources on ambient UFPs.

A second novel variable that is present in our model is the estimate of annual average NO_x emissions. Since it is challenging to collect detailed traffic count data across large geographical areas, we used an estimate of simulated NO_x emissions from traffic extracted from a transportation-emissions model. Our measure of NO_x emissions reflected traffic volume as well as traffic speed (since speed is highly associated with emissions). In order to estimate the model, we used average annual NO_x emissions as long-term exposures are of particular interest given the current lack of information related to the chronic health effects of UFPs. However, it is important to note that NO_x emissions are available for each hour of the day; therefore, this data may be used to develop time-varying exposure surfaces on a finer temporal scale in future studies.

While this study had a number of important advantages including broad spatial coverage over multiple seasons it is important to note several limitations. In particular, while winter monitoring covered a broad geographic area it was limited to a single week in March and thus may provide an imprecise estimate of UFP concentrations over all winter months (December–March). Specifically, ambient temperatures during winter monitoring were mild (by Montreal standards) and do not reflect the extremely cold temperatures that often occur in Montreal. As a result, winter data included in our model likely underestimate exposures during colder periods given the inverse relationship between ambient temperature and UFPs. In addition, our model may overestimate long-term ambient concentrations as monitoring was not conducted on evenings or weekends when ambient UFP levels may be lower. However, with respect to health analyses, so long as spatial differences are adequately represented this should not bias risk estimates for incremental changes in ambient concentrations between regions. In addition, our campaign included a large number of sites in residential areas that were not heavily impacted by traffic sources and thus we feel that our model does capture overall population exposures not just near road concentrations. More generally, the relatively short monitoring periods used to assign ambient UFP concentrations to road segments may also be viewed as a limitation; however other studies have also used short-term monitoring campaigns to build land use regression models for UFPs (Rivera et al., 2012; Weichenthal et al., 2016) and the performance of our model was comparable to models presented in these studies. Moreover, Montagne et al. (2015) recently reported that short-term monitoring is an efficient means of developing land use regression models for ambient UFPs and that such models provide reasonable estimates of historical spatial contrasts in. Futures studies may address this limitation by including longer monitoring periods for each road segment but this will likely come at a cost of decreased spatial coverage.

The use of ambient temperature and wind speed data to adjust for temporal variations in ambient UFPs may also be viewed as a limitation as all we did not have fixed-site regional UFP data for the Montreal area. However, ambient temperature and wind speed are known to be important predictors of temporal variations in ambient UFP concentrations (Alm et al., 1999; Kaur and Nieuwenhuijsen, 2009; Weichenthal et al., 2008, 2014a, 2015, 2016) and both were strong predictors of ambient UFPs in our models. Nevertheless, we cannot rule out some residual impact of temporal variations on the spatial gradients presented in this study. On the other hand, the inclusion of temperature and wind speed in the model also has advantages as it adds a temporal component directly to the model which allows predictions to be made throughout the year as opposed to one static long-term average estimate.

5. Conclusions

A land use regression model is now available for ambient UFPs in Montreal, Canada. This model explains the majority of the spatial variation in ambient UFPs and performed well when evaluated using an external dataset of road segments not used for model development. This model will be applied to estimate the chronic health effects of long-term UFP exposures. However, it may also be used to assign exposures at a finer temporal scale as the terms for ambient temperature and wind speed facilitate predictions throughout the year.

Financial interests' declaration

None declared.

Acknowledgments

This work was supported by CHRP (Canadian Collaborative Health Research Projects) and Health Canada.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.envres.2015.12.016.

References

Abernethy, R.C., Allen, R.W., McKendry, I.G., Brauer, M., 2013. A land use regression model for ultrafine particles in Vancouver. Can. Environ. Sci. Technol. 47, 5217–5225.

Alm, S., Jantunen, M.J., Vartiainen, M., 1999. Urban commuter exposure to particle matter and carbon monoxide inside and automobile. J. Expo. Anal. Environ. Epidemiol. 9, 237–244.

Farrell, W., Deville-Cavellin, L., Weichenthal, S., Goldberg, M., Hatzopoulou, M., 2015. Capturing the urban canyon effect on particle number concentrations across a large road network using spatial analysis tools. Build. Environ. 92, 328–334.

Ferwerda, J., Hainmueller, J., Hazlett, C.J., 2013. KRLS: a Stata package for kernel-based regularized least squares. J. Stat. Softw. 55, 1–24.

Hainmueller, J., Hazlett, C., 2013. Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. Polit. Anal. 22, 143–168.

Hankey, S., Marshall, J., 2015a. Land use regression models of on-road particulate air pollution (particle number, black carbon, $PM_{2.5}$, particle size) using mobile monitoring. Environ. Sci. Technol. 49, 9194–9202.

Hankey, S., Marshall, J., 2015b. On-bicycle exposure to particulate air pollution: particle number, black carbon, $PM_{2.5}$, and particle size. Atmos. Environ. 122, 65–73.

Hatzopoulou, M., Weichenthal, S., Dugum, H., Pickett, G., Miranda-Moreno, L., Kulka, R., Anderson, R., Goldberg, M., 2013. The impact of traffic volume, composition, and road geometry on personal air pollution exposures among cyclists in Montreal. Can. J. Expo. Sci. Environ. Epidemiol. 23, 46–51.

Hoek, G., Beelen, R., Kos, G., Dijkema, M., Van Der Zee, S.C., Fischer, P.H., Brunekreef, B., 2011. Land use regression model for ultrafine particles in Amsterdam. Environ. Sci. Technol. 45, 622–628.

Hudda, N., Gould, T., Hartin, K., Larson, T.V., Fruin, S.A., 2014. Emissions from an international airport increase particle number concentrations 4-fold at 10 km downwind. Environ. Sci. Technol. 48, 6628–6635.

Kaur, S., Nieuwenhuijsen, M., 2009. Determinants of personal exposure to $PM_{2.5}$, ultrafine particle counts, and CO in a transport microenvironment. Environ. Sci. Technol. 43, 4737.

Klompmaker, J.O., Montagne, D.R., Meliefste, K., Hoek, G., Brunekreef, B., 2015. Spatial variation of ultrafine particles and black carbon in two cities: results from a short-term measurement campaign. Sci. Total. Environ. 508, 266–275.

Ostro, B., Hu, J., Goldberg, D., Reynolds, P., Hertz, A., Bernstein, L., Kleeman, M.J., 2015. Associations of mortality with long-term exposures to fine and ultrafine particles, species and sources: results from the California Teachers Study Cohort. Environ. Health Perspect. 123, 549–556.

Rivera, M., Basagana, X., Aguilera, I., Agis, D., Bouso, L., Foraster, M., Medina-Ramon, M., Pey, J., Kunzli, N., Hoek, G., 2012. Spatial distribution of ultrafine particles in urban settings: a land use regression model. Atmos. Environ. 54, 657–666.

Sabaliauskas, K., Jeong, C.H., Yao, X., Reali, C., Sun, T., Evans, G., 2015. Development of a land-use regression model for ultrafine particles in Toronto. Can. Atmos. Environ. 110, 84–92.

Sider, T., Alam, A., Zukari, M., Dugum, H., Goldstein, N., Eluru, N., Hatzopoulou, M., 2013. Land-use and socio-economics as determinants of traffic emissions and individual exposure to air pollution. J. Transp. Geogr. 33, 230–239.

Weichenthal, S., Dufresne, A., Infante-Rivard, C., Joseph, L., 2008. Determinants of ultrafine particle exposures in transportation environments: findings of an 8-month survey conducted in Montreal. Can. J. Expo. Sci. Environ. Epidemiol. 18, 551–563.

Weichenthal, S., 2012. Selected physiological effects of ultrafine particle in acute cardiovascular morbidity. Environ. Res. 115, 26–36.

Weichenthal, S., Farrell, W., Goldberg, M., Joseph, L., Hatzopoulou, M., 2014a. Characterizing the impact of traffic and the built environment on near-road ultrafine particle and black carbon concentrations. Environ. Res. 132, 305–310.

Weichenthal, S., Hatzopoulou, H., Goldberg, M., 2014b. Exposure to traffic-related air pollution during physical activity and acute changes in blood pressure, autonomic and micro-vascular function in women: a cross-over study. Part Fibre Toxicol. 11, 70.

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

Weichenthal, S., Van Ryswyk, K., Kulka, R., Sun, L., Wallace, L., Joseph, L., 2015. In-vehicle exposures to particulate air pollution in Canadian metropolitan areas: the urban transportation exposure study. Environ. Sci. Technol. 49, 597–605.