Transportation Research Part D 88 (2020) 102599

Contents lists available at ScienceDirect

Transportation Research Part D

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

Potential of machine learning for prediction of traffic related air pollution

An Wang, Junshi Xu, Ran Tu, Marc Saleh, Marianne Hatzopoulou *

Department of Civil and Mineral Engineering, University of Toronto, 35 St George St., Toronto, ON M5S 1A4, Canada

ARTICLE INFO

Keywords:
Mobile sampling
Land use regression
Machine learning
Traffic-related air pollution
Fine particulate matter
Black carbon
Traffic pattern recognition

ABSTRACT

Land use regression (LUR) has been extensively used to capture the spatial distribution of air pollution. However, regional background and non-linear relationships can be challenging to capture using linear approaches. Machine learning approaches have recently been used in air quality prediction. Using data from a mobile campaign of fine particulate matter and black carbon in Toronto, Canada, this study investigates the boundaries of LUR approaches and the potential of two different machine learning models: Artificial Neural Networks (ANN) and gradient boost. In addition, a moving camera was used to collect real-time traffic. Models developed for fine particulate matter performed better than those for black carbon. For the same pollutants, machine learning exhibited superior performance over LUR, demonstrating that LUR performance could benefit from understanding how explanatory variables were expressed in machine learning models. This study unveils the black-box nature of machine learning algorithms by investigating the performance of different models in the context of how they capture the relationship between air quality and various predictors.

1. Introduction

Fine particulate matter ($PM_{2.5}$) consists of a mixture of particles that form a major component of ambient air pollution, and it has been associated with adverse health effects at various levels of exposure (Pinault et al., 2017). Black carbon (BC), a component of $PM_{2.5}$, has been associated with both human health effects (World Health Organization, 2013) and climate forcing (Brewer, 2019). BC has been associated with heavy-duty vehicles, which are typically diesel-fuelled (de Miranda et al., 2019; Xu et al., 2018).

Land use regression (LUR) models have been used to capture the spatial variability of traffic-related air pollution in urban areas and can handle almost all air pollutants of interest (Hoek et al., 2008). The data inputs for LUR models are generally available from municipal databases or open maps. Satellite remote sensing (SRS) provides another important data source for LUR models. Abundant studies have highlighted the usefulness of LUR models developed from SRS data in capturing air quality hotspots (Lee et al., 2016; Shen et al., 2002; Yang et al., 2018; Zheng et al., 2018). Traditionally, multiple linear regression models estimated at the urban level were based on measurements from fixed monitoring stations (Liu et al., 2016; Wolf et al., 2017). With low-cost sensors and data mining techniques improving in performance, an increasing number of studies have complemented traditional fixed-site LUR models with short-term stationary measurements (Basu et al., 2019; Kerckhoffs et al., 2016; Minet et al., 2018), and mobile measurements (Hankey & Marshall, 2015; Messier et al., 2018). LUR model results are generally intuitive, interpretable and can be extrapolated and validated.

* Corresponding author at: Department of Civil and Mineral Engineering, University of Toronto, 35 St George St., Toronto, ON M5S 1A4, Canada.
E-mail address: marianne.hatzopoulou@utoronto.ca (M. Hatzopoulou).

https://doi.org/10.1016/j.trd.2020.102599

Available online 28 October 2020

However, the simple statistical nature of LUR limits its performance when exposed to complex air quality data.

More recently, various modelling techniques have been tested to overcome the limitations of LUR at capturing the non-linear relationships between pollutants and predictors. Several studies have compared the performance of machine learning methods and LUR using fixed data at various spatial scales (J. Chen et al., 2019; Weichenthal et al., 2016; Zhan et al., 2017). Kerckhoffs et al. (2019) compared different modelling techniques, including linear regression, non-linear regression, tree-based, and neural network-based machine learning based on mobile and fixed ultrafine particle data collected in the Netherlands. Their results indicate a worsening in the performance of machine learning methods when the models are tested using external data. Other studies demonstrated improvements in model performance after applying machine learning methods to predict $PM_{2.5}$ (Brokamp et al., 2017; Lim et al., 2019). It is worth noting that these studies have varying data quality and quantity, air pollutants of interest, and model hyper-parameter tuning, which are critical factors that influence model performance.

Additional research is needed to understand the behaviour and performance of different empirical techniques and explore the boundaries around their appropriate uses. In this study, $PM_{2.5}$ and BC predictive models were developed for Toronto, Canada, while comparing the performance of LUR models and machine learning techniques, namely artificial neural network (ANN) and gradient boosting such as XGBoost. The impacts of built environment variables and real-time traffic information were evaluated using data collected during a mobile sampling campaign. It is expected that different road segmentation approaches, sample sizes, and cross-validation schemes would have varying impacts on the empirical models. This study capitalizes on mobile air quality data and identifies appropriate setups for different modelling techniques. Our study unveils the behaviour of different machine learning models and explains discrepancies in model performance. By investigating the response of machine learning models to the data, this study provides information that can even be useful for traditional LUR studies

2. Materials and methods

2.1. Air quality sampling campaign

Air quality data were collected over the course of 4 weeks between March and June 2019 in downtown Toronto, Canada. Second-by-second $PM_{2.5}$ particle number was collected using a TSI model 3330 Optical Particle Sizer (OPS), and 10-sec BC mass concentration (in ng/m^3) was collected using a microaethalometre (MicroAeth model AE51) both sampling at the rooftop of a moving vehicle. $PM_{2.5}$ counts were collected in 14 different size bins ranging from 0.3 to 2.5 μm. A dashboard video camera was installed to capture local traffic while the vehicle is moving. Besides, a Global Positioning System (GPS) device (Qstarz BT-Q1000X Travel Recorder) was installed to capture real-time vehicle location and speed. The accuracy of this GPS device has been investigated in a previous study (Xu et al., 2019), and it was observed that the median error for trips in urban canyons was 1.5 m (Schipperijn et al., 2014). All instrument clocks were synchronized daily.

All sampling activities were conducted on non-rainy weekdays between 7:00 AM and noon. Our sampling campaign captured data spanning from morning peak hours to midday off-peak. In terms of spatial coverage, 19 unique corridors (about 120 unique kilometres and 80 natural segments) in a 4 km-by-6 km area in downtown Toronto were covered at least 4 times in each direction. Specifically, a natural segment (which will be referred to as a segment in the rest of the document) represents a stretch of road extending from one intersection to another. A corridor includes several consecutive segments in the same direction (usually with the same street name), spanning between two boundaries of the study area. Each corridor was sampled in two directions within the same hour. The measuring sequence and direction were randomly picked. Messier et al. (2018) demonstrated that with at least 4 repeated sampling visits on the same road segment with 30% road coverage, a robust LUR model could be developed.

2.2. Data processing

2.2.1. Video camera and analysis of local traffic

Traffic video recordings were processed using a convolutional neural network (CNN)-based object detection system. Vehicles were counted and classified frame-by-frame in two phases. Videos were captured by the camera at a rate of 30 frames per second (FPS). First, each frame was treated as an image. All objects in this image were detected and put into bounding boxes using a real-time object detection YOLOv3 (Redmon & Farhadi, 2018). A default classifier in YOLOv3 that was pre-trained on the COCO benchmark (Lin et al., 2014) was used to classify vehicle types, including car, truck, and bus. During the process, each classified object was associated with a confidence level of the prediction. The minimum confidence level to detect vehicles was set as 0.6. Then, the Deep SORT method, which combines Kalman filtering and the Hungarian Algorithm was applied to track multiple vehicles detected by bounding boxes (Wojke et al., 2018). Previous studies have combined these two methods to solve multi-object detection and tracking problems (Munich et al., 2018; Nakashima, Arai, & Fujikawa, 2019). As such, vehicles were counted in videos as those overtaking us and those overtaken by us in the same direction or those in the opposite direction if they pass reference lines in video frames. It is worth mentioning that since curb-parked vehicles are hard to differentiate from those overtaken by us in the same direction, parked vehicles were manually counted and used to adjust our overall counts accordingly. Idling vehicles were not counted as parked vehicles.

The counting system was validated and tested by comparing manual vehicle counts with the system's counting and classification results. In total, 60 min video clips were used for testing, and every minute was selected randomly from our database with over 27 h of videos. The validation results indicate that the Pearson correlation coefficients between counts derived from the computer vision method and manual counts were 0.96, 0.83, and 0.77 for car, truck, and bus, respectively. Vehicle counts in videos were then converted to traffic volumes on-road segments using the moving observer method, which was proposed by Wardrop and Charlesworth

(1954). The detailed algorithm is presented in the Supporting Information, Section 1.

2.2.2. Land Use, meteorology and road network characteristics

To extract accurate land use and built environment variables for every location, all GPS points were first matched and associated with the road network. Land use and road network variables were derived from shapefiles provided by the City of Toronto open data portal and DMTI spatial Inc (City of Toronto, 2019; DMIT Spatial Inc., 2014) using ArcMap 10.4.1. For each GPS point, the distance was calculated between its location and the nearest major arterial, highway, bus route, railway, and shore. We then summarized the length of major arterials, highways, bus routes, railways, and all types of roads in buffers of 25, 50, 100, 200, 500, and 1000 m. Within the same buffer sizes, the areas of different land use types were extracted. In addition, annual average daily traffic (AADT) was derived from traffic count data provided by the City of Toronto. Hourly meteorology was obtained from an Environment Canada weather station close downtown Toronto (Toronto City Centre Airport Station). Each air quality reading was matched with hourly temperature, relative humidity, and wind speed. Real-time traffic derived in Section 2.2.1 was also matched with GPS points based on location.

2.2.3. Air quality data processing

$PM_{2.5}$ counts were collected second-by-second in 14 size bins. For each size bin, $PM_{2.5}$ was first converted to a number concentration, using Equation (1) (TSI Inc., 2016):

$$C_i = \frac{N_i}{Q \times (t_s - DTC \times t_d)} \tag{1}$$

Where C_i is the number concentration in size bin i, N_i is the number count in size bin i, Q is the sample flow rate,t_s is sample time in seconds,t_d is deadtime in seconds, DTC is a deadtime correction factor. Specifically, deadtime is the period that the particle counter is not able to detect an event after another detection event, which is due to the counter's reaction time (capturing a particle and creating an electronic pulse). Therefore, the deadtime correction factor was introduced to account for the missing particle counts. In our case, the manufacturer's suggested deadtime correction factor was 1 (TSI Inc., 2016).

Number concentration was then converted to mass concentration by assuming all $PM_{2.5}$ particles are spherical, with diameters equal to the median of size bins they belong to. Particle effective density was assumed to be 1.6 g/cm^3. We understand that the composition and source of fine particulate matter can largely impact its effective density (Buonanno et al., 2009; Hand et al., 2010; Shen et al., 2002), but the selection of a single estimate of effective density should not affect our models' performance and conclusions. $PM_{2.5}$ mass concentration was calculated as the sum of all mass concentrations in size bins with diameters smaller than 2.5 μm.

BC concentration was collected by microaethalometres at a 10 s time interval. The Optimized Noise-Reduction Algorithm (ONA) developed by the U.S. EPA was employed to reduce the occurrence of negative values (approximately 3%) to virtually zero while preserving the significant dynamic trends in time series (Hagler et al., 2011).

The $PM_{2.5}$ data were associated with second-by-second GPS points based on their timestamp. For BC concentrations, every 10 GPS points were assigned the same BC concentration.

2.2.4. Data segmentation

The GPS data were organized into five different segmentation schemes: (1) natural segmentation, which refers to the distance between the end of one intersection and the end of the next intersection (on average 90 m long, 9 s drive); (2) 30-metre segments (about 3 s drive); (3) 50-metre segments (about 5 s drive); (4) 100-metre segments (about 10 s drive); (5) 200-metre segments (about 15–20 s drive). Air quality, meteorology, traffic flow, and land use variables associated with each GPS point were averaged and summarized over the corresponding segment (in the five segmentation schemes). The different segmentation schemes were adopted as an additional dimension to model performance testing. It is hypothesized that while a small segment length provides a larger sample size, which could be appropriate for the training of a machine learning model, it also introduces variability into the sample that a LUR model may not explain.

2.3. Empirical models

2.3.1. Land use regression models

LUR models were developed using a forward selection procedure following Messier et al. (2018). An ordinary least-squares LUR was developed by first fitting an intercept-only model, and one at a time, explanatory variables were added to the model based on the ranking of their correlations with the response variable (log-transformed $PM_{2.5}$). A new variable was added if: it statistically significantly increases the model coefficient of determination (pseudo-R-squared); its variance inflation factor (VIF) was less than 3; it had a sign that was in line with prior knowledge of the direction of its effect. Since the probability distributions of $PM_{2.5}$ and BC mass concentrations were log-normal, the natural logarithm of the concentrations was used as response variables. All potential explanatory variables are summarized as follows:

- Areas of various land uses, including residential, commercial, government and institutional, open area, park and recreational, resource and industrial, waterbody within buffer sizes of 25, 50, 100, 200, 500, and 1000 m;
- Length of different road types, including major arterials, highways, bus routes, railroads; and the sum of all road types within buffer sizes of 25, 50, 100, 200, 500, and 1000 m;
- Distances to nearest major arterials, highways, bus routes, railroads, and shore;

- Traffic information, including annual average daily traffic, measured average speed, real-time traffic density, and truck count;
- Meteorology, including temperature, relative humidity, and wind speed.

2.3.2. *Artificial neural Networks and One-at-a-Time analysis*

With the same input database for LUR models, an artificial neural network (ANN) model was also developed. Input data were standardized to improve model performance. Different hyper-parameters were tested, including network layers, neurons in layers, learning rate (and leaning rate decay), training algorithm, performance function, transfer functions, and regularization methods. Finally, a fully connected feedforward ANN with 2 hidden layers and 4 neurons in each layer was selected based on test data performance. The activation function for hidden layers 1 and 2 is tan-sig function with regulated output between -1 and 1. The ANN was trained using a Bayesian regularization backpropagation algorithm. This algorithm determines the regularization parameters in an automated fashion so that it can help improve model generalization and avoid overfitting. The loss function is the widely used mean squared error (MSE), defined as $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$, where n is the sample size, $Y_i$and $\widehat{Y}_i$ are observed and predicted values.

ANN's black box nature has limited its usage in understanding the sources and dispersion patterns of near-road quality. In light of the simple structure of the ANN trained, a One-At-A-Time (OAAT) analysis was employed to understand the behaviour of each predictor. The OAAT analysis was applied to trained ANN models with the best cross-validation performance. Recall that ANN input vectors were standardized by mapping them to the distribution with a mean of 0 and standard deviation (σ) of 1, before inputting them into the model. Artificial input vectors were created with standardized input variables ranging from -2σ to 2σ while all other variables were kept as 0, as shown in Fig. 1. Then, these input vectors were fed through the trained ANN models, and this process was repeated for all input variables.

According to the universal approximation theorem, if the ANN model is considered as a continuous function describing $PM_{2.5}$ or BC concentrations, this OAAT analysis develops a projection of the function on the plane defined by the response variable ($PM_{2.5}$ or BC concentrations) and every explanatory variable. In this fashion, the effect of the ANN model on every single variable can be captured. It is worth noting that the OAAT approach isolates each variable and cannot account for the relative importance of different input variables. Nevertheless, it clearly illustrates how ANN models treat each variable by projecting the complex ANN-learnt function to a 2-dimensional plane.

2.3.3. *Xgboost and feature attribution method*

The XGBoost approach is a tree-based gradient boosting framework which can naturally deal with sparse features, continuous features, and categorical features as well as high-order interactions between features (T. Chen & He, 2015). Besides, it can overcome the overfitting issue by penalizing more complex models through both LASSO and Ridge regularizations. Most importantly, feature importance measures can be employed with tree-based models to qualitatively and quantitatively investigate each feature's impact on the model output. In this study, the scikit-learn and XGBoost Python libraries were used (Pedregosa et al., 2011). A set of hyper-parameters including learning rate, maximum depth of the tree, number of trees to grow, the minimum number of samples on a leaf, and minimum loss reduction were tuned using a combination of random and grid search. Root-mean-square-error showed in Equation (2) was employed as the loss function, where n is the sample size, $\widehat{y}_i$ is the predicted value, and y_i is the measured value:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}{n}} \tag{2}$$

Once the best model was selected, the SHapley Additive exPlanations (SHAP) method was applied to capture the order of feature importance and the impact of features on the model output. This method evaluates the importance of a feature by comparing the performance of the model with and without the feature (Lundberg & Lee, 2017). Lundberg and Lee (2018) developed fast and exact tree solutions for SHAP values by averaging differences in predictions over all possible orderings of the features. Results are represented in a SHAP summary plot. It sorts the features by their global impact, capturing the impact of each feature on the model output for the individual record. In this context, response variables were $PM_{2.5}$ and BC concentrations.

2.3.4. *Model performance and evaluation*

All models (LUR, ANN, and XGBoost) were developed for 5 different data segmentation schemes: natural segments, artificial segments of 30 m, 50 m, 100 m, and 200 m. For LUR models, a Monte-Carlo approach was implemented with 5-fold cross-validation (CV) to assess LUR model performance. Each fold was used as a test set once, while the other 4 folds were used for estimating the LUR model parameters. The mean pseudo-R-squared for the 5 models was used as a performance measure. This 5-fold CV process was repeated 100 times as a Monte-Carlo simulation to assess LUR model performance. Each time, data were shuffled to generate a new 5-fold dataset. Similarly, for the ANN and XGBoost methods, they were trained and tested by subsampling 90% of observations without replacement (no observation can be selected twice) as the training set, with the remaining as the test set. This process was also repeated

$$\begin{pmatrix} -2\sigma & -1.5\sigma & \dots & 1.5\sigma & 2\sigma \\ 0 & 0 & \vdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} \gg\gg \textit{into trained ANN in this direction}$$

Fig. 1. A conceptual illustration of OAAT input vectors for one explanatory variable.

100 times. Detailed model development is captured in Fig. 2.

When subsampling the training and testing sets for machine learning methods, a grouped non-stratified partition scheme was used to minimize overfitting. Road segments were first grouped according to the roadway corridor they belong to and then used a non-stratified partition based on the groups. This technique sets a more stringent validation scheme on the machine learning models because an entire corridor could be withheld from the training test but then included for prediction. In contrast, random cross-validation (CV) was applied in the case of the LUR models.

Two statistics were used to evaluate model performance: pseudo-R-squared (squared Pearson correlation between log-transformed predicted and observed values only in the test dataset) and normalized root mean squared error (NRMSE). NRMSE is defined in Equation (3):

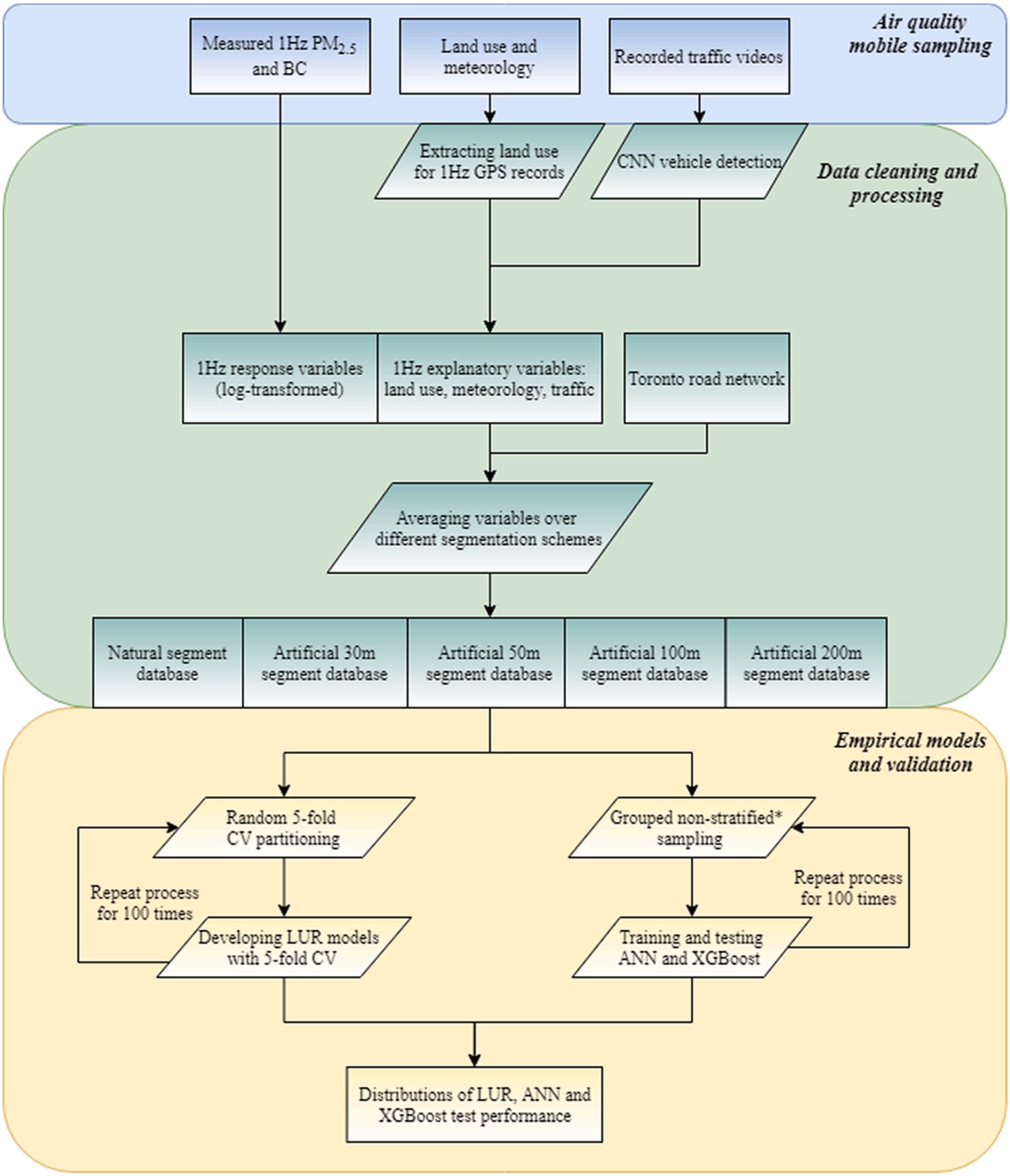
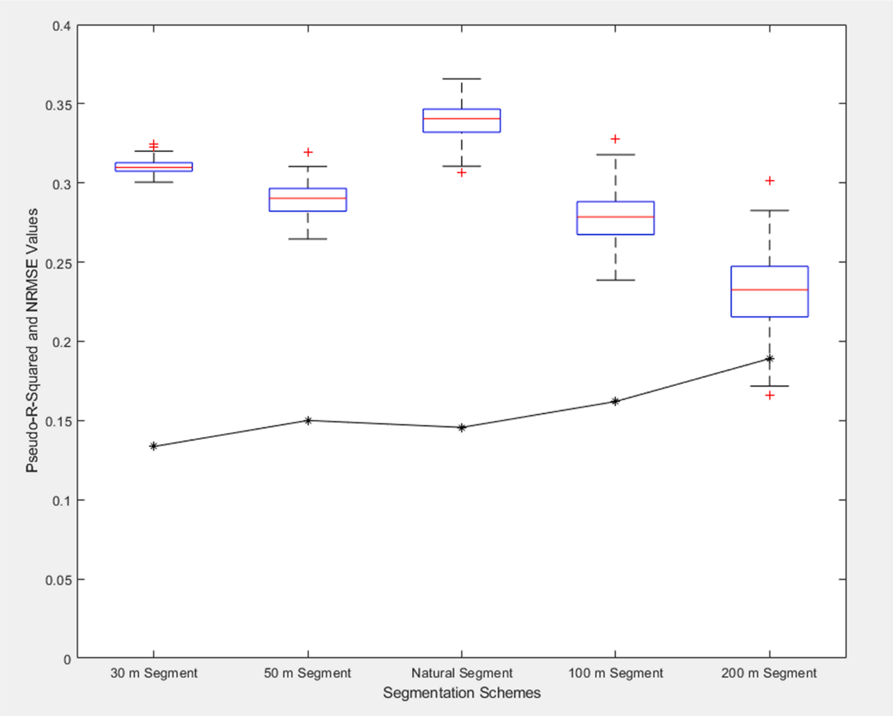
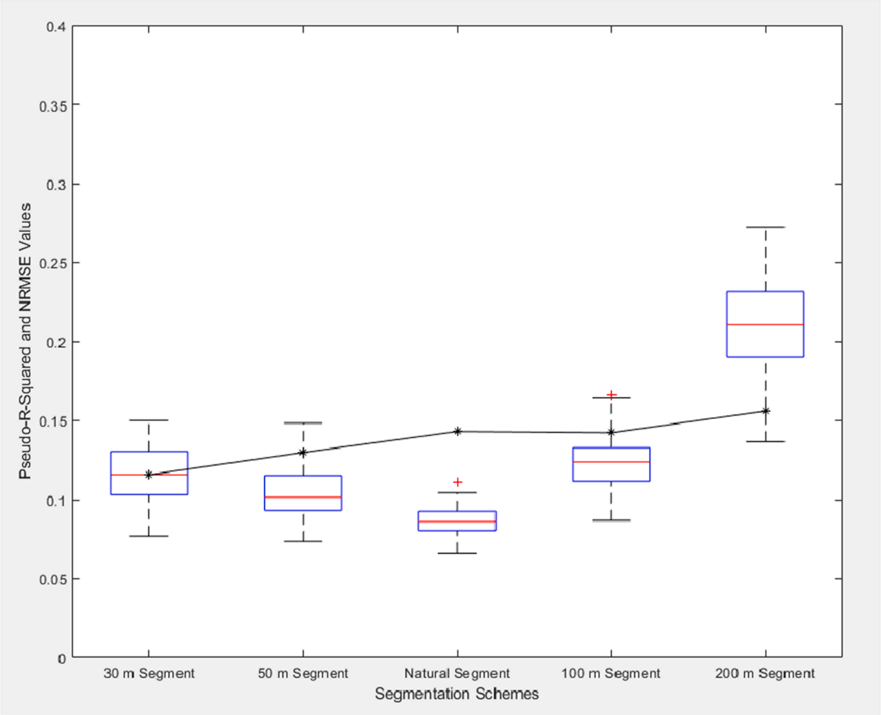

Air quality mobile sampling
Measured 1Hz $PM_{2.5}$ and BC
Land use and meteorology
Recorded traffic videos
Data cleaning and processing
Extracting land use for 1Hz GPS records
CNN vehicle detection
1Hz response variables (log-transformed)
1Hz explanatory variables: land use, meteorology, traffic
Toronto road network
Averaging variables over different segmentation schemes
Natural segment database
Artificial 30m segment database
Artificial 50m segment database
Artificial 100m segment database
Artificial 200m segment database
Empirical models and validation
Random 5-fold CV partitioning
Grouped non-stratified* sampling
Repeat process for 100 times
Developing LUR models with 5-fold CV
Repeat process for 100 times
Training and testing ANN and XGBoost
Distributions of LUR, ANN and XGBoost test performance

Fig. 2. Data collection, processing, and model development. * Grouped non-stratified sampling: grouping segments by corridors first, then partitioning based on groups.

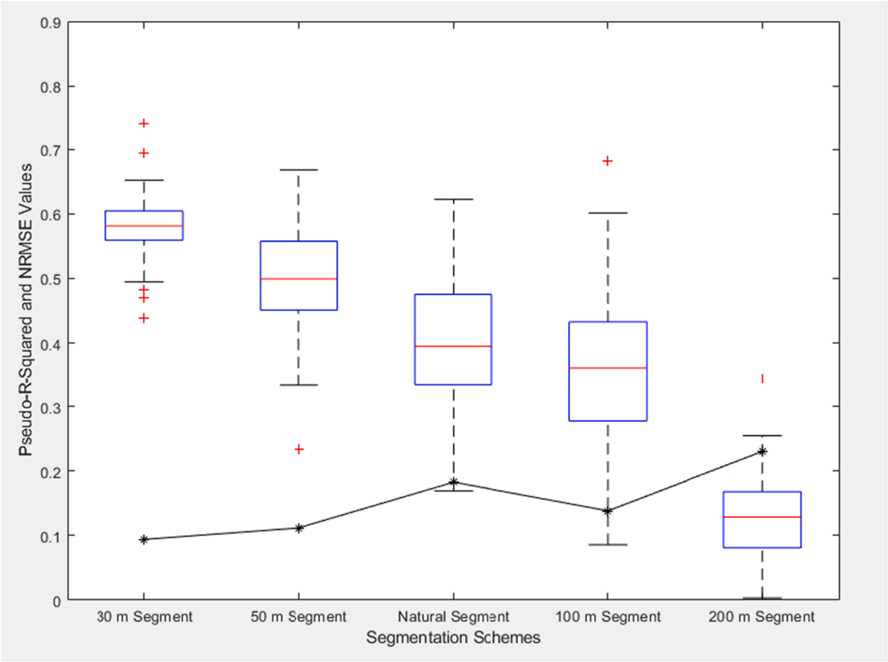


(a) LUR model performance for $PM_{2.5}$

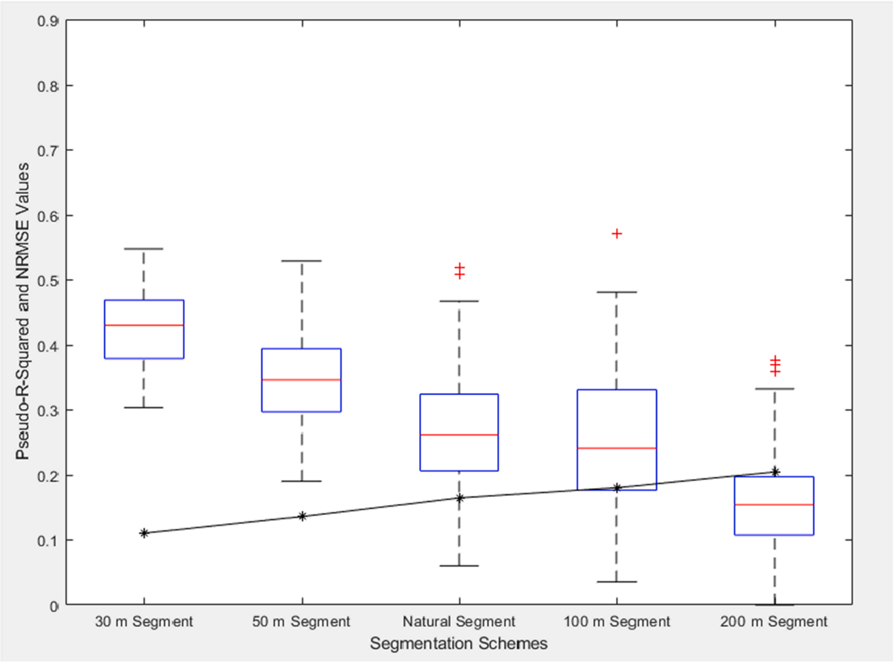


(b) LUR model performance for BC

Fig. 3. LUR model performance using different segmentations; boxes represent pseudo-R-squared and black asterisks represent the average NRMSE.

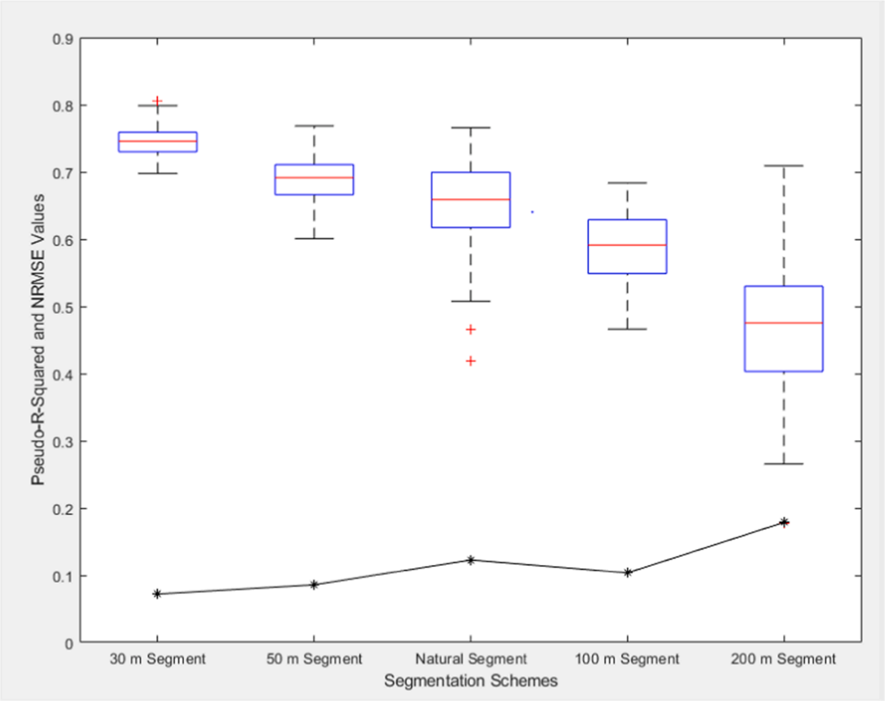


(a) ANN model performance for $PM_{2.5}$

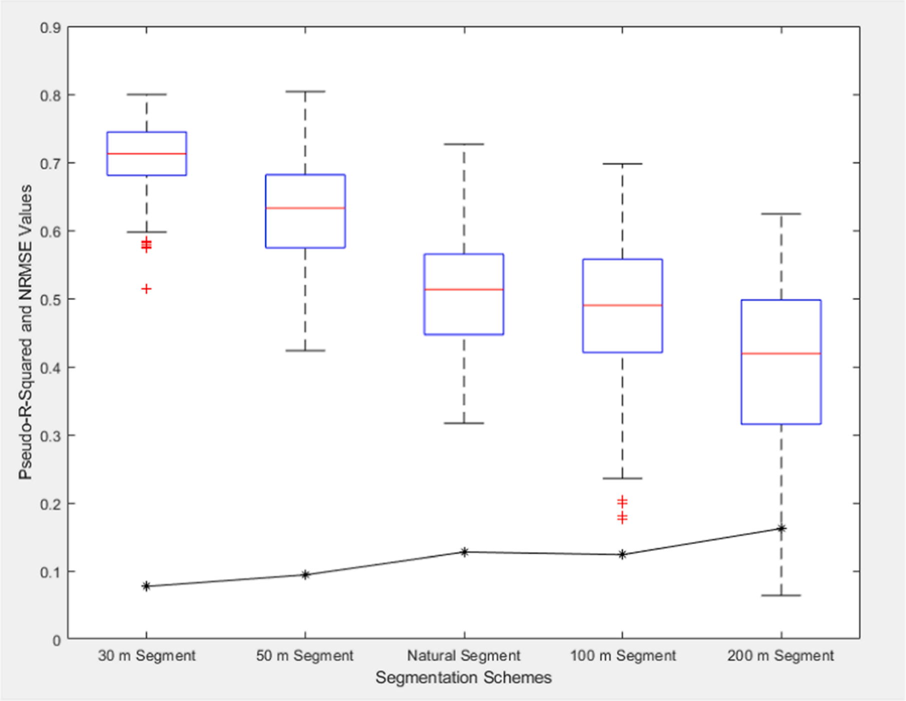


(b) ANN model performance for BC

Fig. 4. ANN and XGBoost model performance using different segmentation schemes; boxes represent the pseudo-R-squared, and black asterisks represent average NRMSE.



(c) XGBoost model performance for $PM_{2.5}$



(d) XGBoost model performance for BC

Fig. 4. (*continued*).

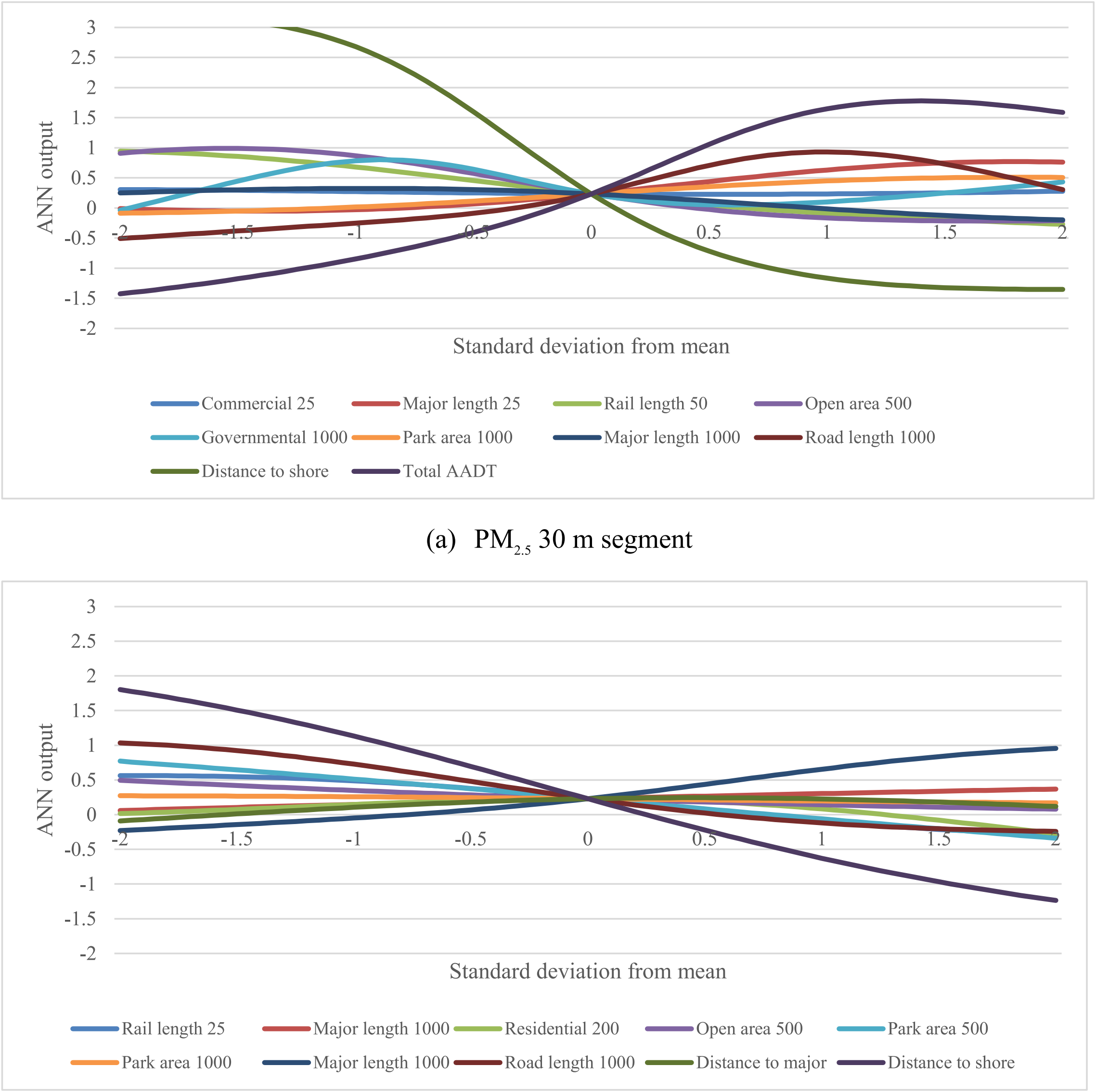
$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n} \frac{\left(y_i - \widehat{y}_i\right)^2}{n}}}{\max(Y) - \min(Y)} \tag{3}$$

wheren is the number of samples in each fold, y_i and $\widehat{y}_i$ are observed and predicted response variables, and Y is the observed response variable vector. It is worth noting that both performance statistics are calculated with test data that are never used in model training.

3. Results

3.1. Descriptive analysis

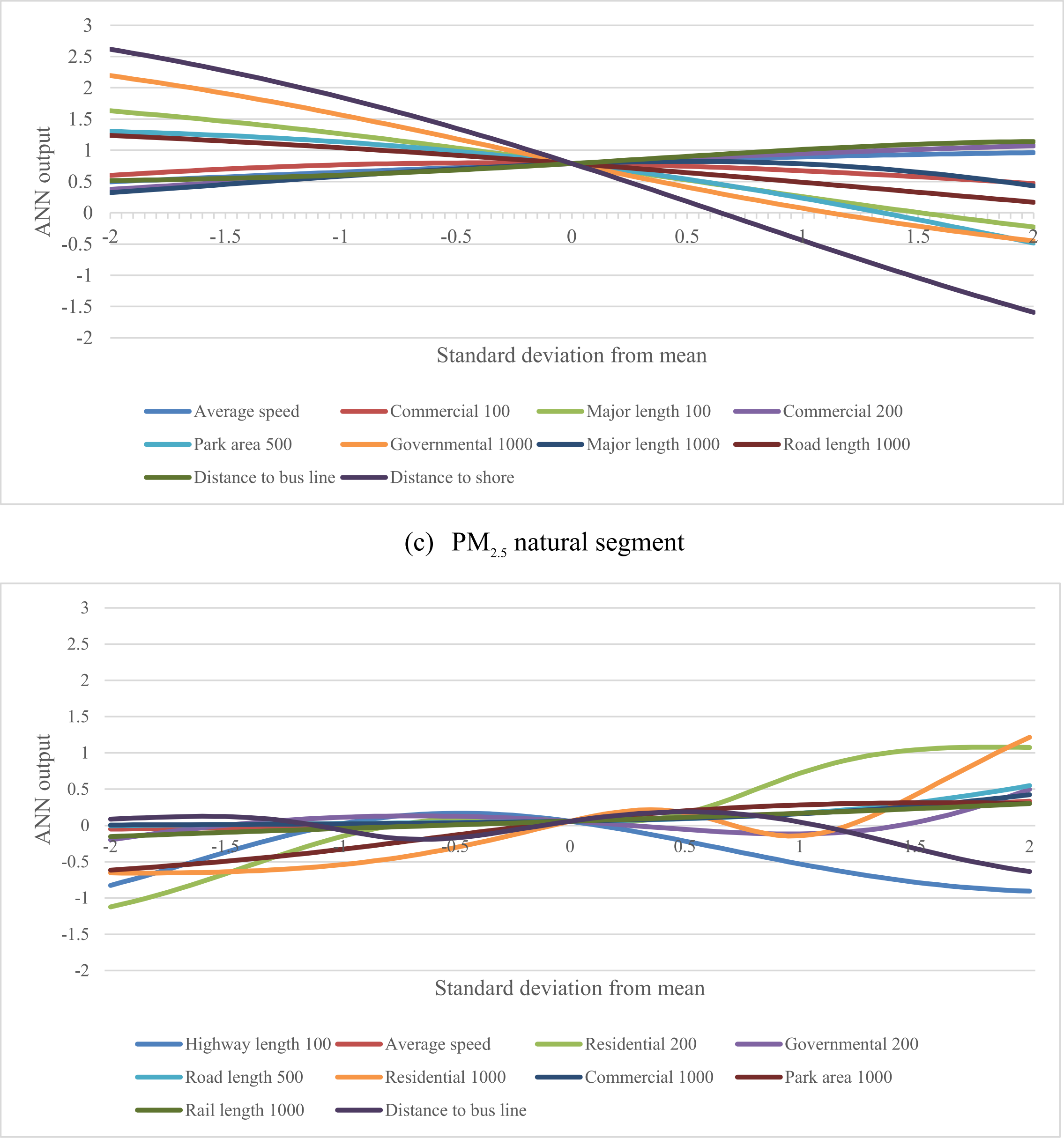
After data processing, a total of 150,000 effective records (in seconds) were obtained along 19 unique roadway corridors. Each record is associated with time, coordinates, $PM_{2.5}$ and BC concentrations, traffic, weather, and land use variables. By matching each



(a) $PM_{2.5}$ 30 m segment

(b) $PM_{2.5}$ 50 m segment

Fig. 5. One-at-a-time analysis of best ANN models using different segmentations.



(c) $PM_{2.5}$ natural segment

(d) BC 30 m segment

Fig. 5. (*continued*).

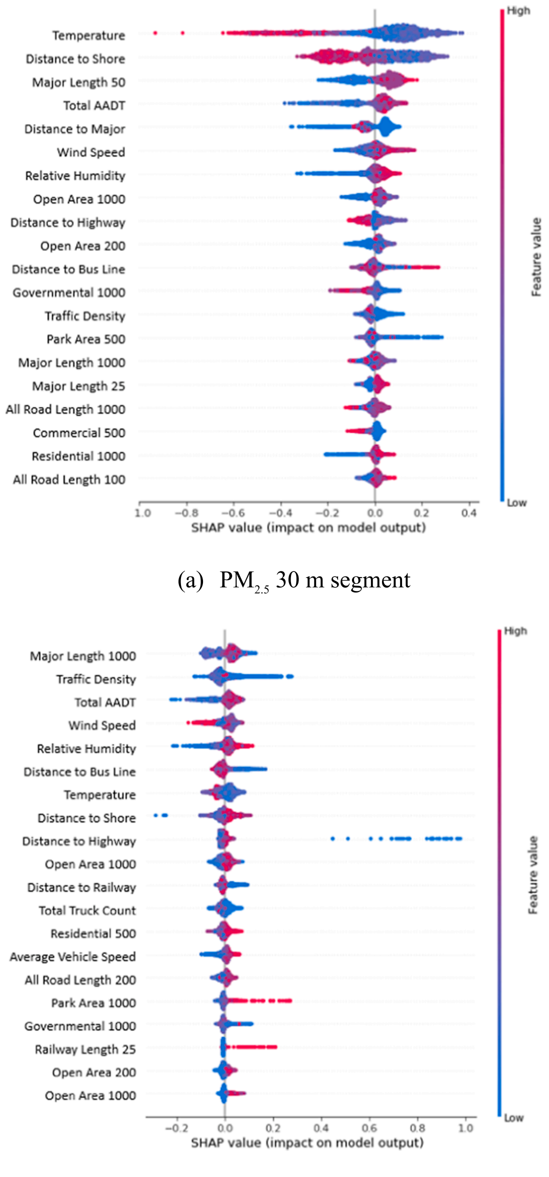
record to the corresponding road segment and averaging, 810, 2979, 1519, 787, and 349 independent segment-level observations were obtained, based on natural segmentation, and segments of lengths 30 m, 50 m, 100 m, and 200 m. Note that the lengths of natural segments are around 100 m, which is captured by a similar number of segments obtained in the two schemes. In downtown Toronto, across the entire network covered, the average $PM_{2.5}$ and BC concentrations were 3.94 ug/m^3 and 1.75 ug/m^3, respectively. The correlations were examined between each explanatory variable and response variables ($PM_{2.5}$ and BC), and the results are presented in Section 2 of the Supporting Information.

3.2. Land use regression models

LUR models developed for different segmentation schemes are summarized in Section 3 of the Supporting information. Their performance is shown in Fig. 3. For $PM_{2.5}$, the models with the highest predictive power were those with GPS data segmented according to natural segments with an average pseudo-R-squared of 0.35. In contrast, the 200 m segment scheme led to the BC LUR models with the highest predictive power, and natural segmentation led to the BC models with the lowest predictive power. $PM_{2.5}$ LUR models achieved higher predictive power (similar average NRMSE but higher pseudo-R-Squared) than BC models.

3.3. Machine learning models

The impact of the segmentation scheme on model performance was more evident with the machine learning approaches. ANN performance is summarized in Fig. 4 (a) and (b). When segments are longer, model performance drops, and average NRMSE values increase. It is worth noting that as segmentation becomes more aggregated, the training sample size also decreases, as stated in Section 3.1. We explored further to determine whether model performance deterioration was mostly caused by changes in the aggregation of explanatory variables or due to a decrease in sample size. Three hundred observations were randomly subsampled from the entire training dataset for different segmentations and reproduced all LUR and ANN models accordingly. LUR models maintained similar average Pseudo-R-Squared as the model trained with all observations, but with larger variability in model performance. Yet, ANN model performance declined drastically with increasing segment lengths. These results can be found in the Supporting Information, Section 6.



(a) $PM_{2.5}$ 30 m segment

(b) BC 30 m segment

Fig. 6. SHAP plots of best XGBoost models using different segmentation schemes.

The XGBoost model performance is summarized in Fig. 4 (c) and (d). With the segmentation becoming more aggregated, the model's pseudo-R-square drops, whereas the average NRMSE value, as well as the variability in model performance, increases. Similar to ANN, the XGBoost model performance highly depends on the sample size of the training data (segmentation scheme). Both machine learning methods exhibited a higher pseudo-R squared than the LUR model, but only when the sample size is large. Smaller sample sizes obtained with the larger segmentation schemes were observed to compromise model performance. Prediction maps using best-performing LUR, ANN, and XGBoost models are presented in the Supporting Information, Section 7.

3.4. *Understanding land use regression and Machine learning models*

3.4.1. *One-At-A-Time analysis for artificial neural network*

The OAAT approach captures the effect of individual variables on the ANN model output. For cases where ANN performs better than LUR, ANN's explanatory power was exploited to understand LUR models' behaviours. For this purpose, the most frequently selected explanatory variables in the best set of 5-fold LUR models for each segmentation scheme were identified for OAAT analysis. Fig. 5 presents the OAAT relationship between LUR selected variables and the ANN model output. Fig. 5 (a) captures the effect of variables picked by LUR on $PM_{2.5}$ under the 30 m segmentation scheme. In general, the curves shown in Fig. 5 (a) have apparent slopes, which are mostly monotonic, except for governmental area and all road lengths. In Fig. 5 (b), the OATT analysis indicates that LUR models could potentially misunderstand the trend due to smaller gradients, and the advantage of having a nearly linear relationship is also diminished. In the previous section, it was observed that the $PM_{2.5}$ natural segmentation scheme has the best average performance among all LUR models; this can be seen in Fig. 5 (c) as the variables generally have steep slopes and monotonicity. Fig. 5 (d) illustrates the BC 30 m scheme, in which it is observed that the small gradient and strong non-monotonicity can lead to poor average performance and higher variability for this set of LUR models.

More generally, three observations were concluded related to LUR model performance based on OAAT analysis: 1) The LUR model will have less variability in its performance if the relationship between each explanatory variable and the outcome variable has a large gradient which enables the LUR model to capture the pattern; 2) The LUR model performs better if the curve of each explanatory variable is monotonic and ideally linear; 3) A curve's gradient/slope can amplify or diminish the impacts of monotonicity. Further analysis for other schemes is presented in the Supporting Information, Section 4.

3.4.2. *Feature attribution analysis for XGBoost*

The relative importance of explanatory variables on response variables ($PM_{2.5}$ and BC) was ranked using SHAP values. In Fig. 6, the dot is coloured based on the value of the feature: low (blue) to high (red), with changes in colour. The dots will be stacked vertically if this feature has a similar impact on the model's output for different observations (Lundberg et al., 2018).

Similar results are observed as those from the LUR and ANN models. For BC, the most influential variables are major road length, traffic density, total AADT, wind speed, and relative humidity. Also, in SHAP plots, if the impact of an explanatory variable on the response variable is monotonic, a smooth change in colouring should be observed (red–purple-blue or vice versa). Furthermore, if the impact is linear, the colour will change smoothly, and vertical stacking at each SHAP value would have a similar number of dots. Good examples of monotonicity are distance to shore and major road length in 50 m segments in Fig. 6 (a). The XGBoost model for BC 30 m segmentation (Fig. 6 (b)) has shown not only mixed colouring indicating non-linearity, but also small span in SHAP value for each explanatory variable indicating a small gradient as discussed in Section 3.4.1. This finding agrees with the three observations derived from ANN OAAT analysis and explains why the BC 30 m segmentation LUR model has poorer performance than that of $PM_{2.5}$.

4. Discussion

In this study, $PM_{2.5}$ and BC models were developed with mobile data using LUR and machine learning methods for 5 data segmentation schemes. Multiple techniques were used to contrast their performance and, more importantly, understand their behaviours. Since no significant difference in performance was found among linear forward and backward stepwise regression models (J. Chen et al., 2019; Kerckhoffs et al., 2019; Minet et al., 2018), LUR models were only developed and analyzed following the same forward selection procedure as Messier et al. (2018). In this process, LUR models were found to be highly dependent on researchers' a priori knowledge and subjective judgment, especially when variable interactions were involved. In general, the $PM_{2.5}$ data was observed to fit better with the LUR approach than BC (best average 5-fold CV pseudo-R-squared at 0.35 vs. 0.22) irrespective of the segmentation scheme used.

The same dataset and explanatory variables used in LUR were applied to develop the ANN and XGBoost models with better model performance, except for the 200 m segmentation scheme. With the segmentation getting more aggregated, ANN and XGBoost model performance decreases. This is mainly due to a decrease in sample size when data is more aggregated. In general, LUR maintains its explanatory power when the size of the dataset decreases, while the machine learning models are quite data-hungry. According to Messier et al. (2018), the adequate sample size should be derived from the proper aggregation/segmentation scheme of explanatory variables with at least 4–8 independent repetitions on each site/road segment. In this study, about 1500 sites (with 50 m segments) were achieved, each repeated 8 times on different days; this segmentation seemed to lead to the best performing machine learning models.

ANN's superior performance to linear regression models is guaranteed by the universal approximation theorem of functions (Hornik, 1991). It states that a fully connected multilayer (layer number > 1) feedforward neural network with continuous, bounded, and nonconstant activation function can act as a universal approximator for any smooth mapping to any accuracy. Meanwhile, linear regression-based models without variable interaction can be interpreted as first-order Taylor Series approximation of functions. Nevertheless, ANN models also have much more variability in performance than LUR due to random initialization of weights and biases. Furthermore, when comparing linear and non-linear models, the coefficient of determination (R-squared) should be interpreted with extra caution (Kerckhoffs et al., 2019). For linear models, R-squared represents the fraction of explained variance as well as the squared correlation coefficient. As in the ordinary least square linear models, the explained sum of squares and residual sum of squares add up to the exact total sum of squares. However, for many non-linear models, this condition does not hold, so that R-squared can only be calculated as the squared correlation coefficient between predicted and observed values. It cannot be interpreted more than a measure of the correlation between predicted and observed values. Other measures, such as bias or NRMSE, should be used as a complement to evaluate prediction errors, especially for systematic performance comparison involving non-linear models. Proper cross-validation and external testing techniques should be applied to avoid overfitting and improve model robustness.

Efforts were made in the OAAT analysis as well as SHAP plots to unveil the black-box nature of ANN and XGBoost methods. By contrasting the expression of explanatory variables in LUR and machine learning models, there are three observations that relate to situations where the performance of LUR is deemed adequate. Based on these observations, it is acknowledged that future studies on traffic-related air pollution using methods that can recognize non-linear relationships can promote better LUR application. Furthermore, the development of machine learning methods, provide opportunities to understand complex interactions and non-linear relationships between response and explanatory variables. Future studies should not only compare machine learning model performance but also understand how machine learning models interpret data. With growing studies conducted in this way, a more comprehensive and accurate a priori knowledge pool can be established, which can better inform LUR developments.

Our analysis was associated with a set of limitations. First, the same neural network architecture was applied with similar hyperparameters to all datasets, which can be further tuned for better performance. While we concluded that a properly tuned ANN should fit better the observations than a linear model, one primary assumption here is that $PM_{2.5}$ and BC concentrations can be expressed as continuous functions of the explanatory variables. This is generally true in the context of our study, but will not always hold, especially when discrete or categorical explanatory variables are introduced within the ANN.

CRediT authorship contribution statement

An Wang: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing - original draft. **Junshi Xu:** Conceptualization, Data curation, Formal analysis. **Ran Tu:** Data curation. **Marc Saleh:** . **Marianne Hatzopoulou:** Conceptualization, Funding acquisition, Methodology, Resources, Supervision.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.trd.2020.102599.

References

Basu, B., Alam, M.S., Ghosh, B., Gill, L., McNabola, A., 2019. Augmenting limited background monitoring data for improved performance in land use regression modelling: Using support vector regression and mobile monitoring. Atmos. Environ. 201, 310–322. https://doi.org/10.1016/J.ATMOSENV.2018.12.048.

Brewer, T.L., 2019. Black carbon emissions and regulatory policies in transportation. Energy Policy 129 (March), 1047–1055. https://doi.org/10.1016/j.enpol.2019.02.073.

Brokamp, C., Jandarov, R., Rao, M.B., LeMasters, G., Ryan, P., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmos. Environ. 151, 1–11. https://doi.org/10.1016/j.atmosenv.2016.11.066.

Buonanno, G., Isola, M.D., Stabile, L., Viola, A., Isola, M.D., Stabile, L., Viola, A., 2009. Uncertainty Budget of the SMPS – APS System in the Measurement of PM 1, PM 2.5, and PM 10. Aerosol Sci. Technol. 6826 https://doi.org/10.1080/02786820903204078.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Hoek, G., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934. https://doi.org/10.1016/J.ENVINT.2019.104934.

Chen, T., He, T., 2015. xgboost : eXtreme Gradient Boosting. R Package Version (4-2), 1–4.

City of Toronto. (2019). Toronto Centreline (TCL) - Data catalogue - Open Data | City of Toronto. Retrieved July 30, 2019, from http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=9acb5f9cd70bb210VgnVCM1000003dd60f89RCRD&vgnextchannel=1a66e03bb8d1e310VgnVCM10000071d60f89RCRD.

de Miranda, R. M., Perez-Martinez, P. J., de Fatima Andrade, M., & Ribeiro, F. N. D. (2019). Relationship between black carbon (BC) and heavy traffic in São Paulo, Brazil. Transportation Research Part D: Transport and Environment, 68(September 2017), 84–98. https://doi.org/10.1016/j.trd.2017.09.002.

DMIT Spatial Inc. (2014). CanMap ® RouteLogistics.

Hagler, G.S.W., Yelverton, T.L.B., Vedantham, R., Hansen, A.D.A., Turner, J.R., 2011. Post-processing method to reduce noise while preserving high time resolution in aethalometre real-time black carbon data. Aerosol Air Qual. Res. 11 (5), 539–546. https://doi.org/10.4209/aaqr.2011.05.0055.

Hand, J. L., Kreidenweis, S. M., Hand, J. L., & Kreidenweis, S. M. (2010). A New Method for Retrieving Particle Refractive Index and Effective Density from Aerosol Size Distribution Data A New Method for Retrieving Particle Refractive Index and Effective Density from Aerosol Size Distribution Data, 6826. https://doi.org/10.1080/0278682029009227.

Hankey, S., Marshall, J.D., 2015. 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 (15), 9194–9202. https://doi.org/10.1021/acs.est.5b01209.

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. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2008.05.057.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4 (2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-T.

Kerckhoffs, J., Hoek, G., Messier, K.P., Brunekreef, B., Meliefste, K., Klompmaker, J.O., Vermeulen, R., 2016. Comparison of ultrafine particle and black carbon concentration predictions from a mobile and short-term stationary land-use regression model. Environ. Sci. Technol. 50 (23), 12894–12902. https://doi.org/10.1021/acs.est.6b03476.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., Vermeulen, R.C.H., 2019. Performance of Prediction Algorithms for Modelling Outdoor Air Pollution Spatial Surfaces. Environ. Sci. Technol. 53 (3), 1413–1421. https://doi.org/10.1021/acs.est.8b06038.

Lee, H.J., Chatfield, R.B., Strawa, A.W., 2016. Enhancing the applicability of satellite remote sensing for $PM_{2.5}$ estimation using MODIS deep blue AOD and land use regression in California, United States. Environ. Sci. Technol. 50 (12), 6546–6555. https://doi.org/10.1021/acs.est.6b01438.

Lim, C.C., Kim, H., Vilcassim, M.J.R., Thurston, G.D., Gordon, T., Chen, L.C., Kim, S.Y., 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul. South Korea. *Environment International* 131 (March), 105022. https://doi.org/10.1016/j.envint.2019.105022.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 8693 *LNCS*(PART 5) 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, C., Henderson, B.H., Wang, D., Yang, X., Peng, Z., 2016. A land use regression application into assessing spatial variation of intra-urban fine particulate matter (PM2.5) and nitrogen dioxide (NO2) concentrations in City of Shanghai. China. *Science of The Total Environment* 565, 607–615. https://doi.org/10.1016/J.SCITOTENV.2016.03.189.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles.

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. (Section 2), 1–10.

Messier, K.P., Chambliss, S.E., Gani, S., Alvarez, R., Brauer, M., Choi, J.J., Apte, J.S., 2018. Mapping Air Pollution with Google Street View Cars : Efficient Approaches with Mobile Monitoring and Land Use Regression. Environ. Sci. Technol. 52 (21), 12563–12572. https://doi.org/10.1021/acs.est.8b03395.

Minet, L., Liu, R., Valois, M.-F.-M.-F.-M.-F., Xu, J., Weichenthal, S., Hatzopoulou, M., 2018. Development and Comparison of Air Pollution Exposure Surfaces Derived from On-Road Mobile Monitoring and Short-Term Stationary Sidewalk Measurements. Environ. Sci. Technol. 52 (6), 3512–3519. https://doi.org/10.1021/acs.est.7b05059.

Pedregosa, F., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Weiss, R., 2011. Scikit-learn : Machine Learning in Python To cite this version : Scikit-learn : Machine Learning in Python. Journal of Machine Learning Research 12.

Pinault, L.L., Weichenthal, S., Crouse, D.L., Brauer, M., Erickson, A., van Donkelaar, A., Burnett, R.T., 2017. Associations between fine particulate matter and mortality in the 2001 Canadian Census Health and Environment Cohort. Environ. Res. 159, 406–415. https://doi.org/10.1016/J.ENVRES.2017.08.037.

Redmon, J., Farhadi, A., 2018. YOLO vol 3. Tech Report 1–6.

Schipperijn, J., Kerr, J., Duncan, S., Madsen, T., Klinker, C.D., Troelsen, J., 2014. Dynamic Accuracy of GPS Receivers for Use in Health Research: A Novel Method to Assess GPS Accuracy in Real-World Settings. Front. Public Health 2 (March), 1–8. https://doi.org/10.3389/fpubh.2014.00021.

Shen, S., Jaques, P.A., Zhu, Y., Geller, M.D., Sioutas, C., 2002. Evaluation of the SMPS-APS system as a continuous monitor for measuring PM2.5, PM10 and coarse (PM2.5-10) concentrations. Atmos. Environ. 36 (24), 3939–3950. https://doi.org/10.1016/S1352-2310(02)00330-8.

TSI Inc. (2016). Optical particle sizer spectrometre model 3330. Retrieved from https://www.tsi.com/optical-particle-sizer-3330/.

Wardrop, J.G., Charlesworth, G., 1954. a Method of Estimating Speed and Flow of Traffic From a Moving Vehicle. Proc. Inst. Civ. Eng. 3 (1), 158–171. https://doi.org/10.1680/ipeds.1954.11628.

Weichenthal, S., Ryswyk, K.V., Goldstein, A., Bagg, S., Shekkarizfard, M., Hatzopoulou, M., 2016. A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. Environ. Res. 146, 65–72. https://doi.org/10.1016/J.ENVRES.2015.12.016.

Wojke, N., Bewley, A., & Paulus, D. (2018). Simple online and real-time tracking with a deep association metric. Proceedings - International Conference on Image Processing, ICIP, 2017-Septe, 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962.

Wolf, K., Cyrys, J., Harciníková, T., Gu, J., Kusch, T., Hampel, R., Peters, A., 2017. Land use regression modelling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. Sci. Total Environ. 579, 1531–1540. https://doi.org/10.1016/J.SCITOTENV.2016.11.160.

World Health Organization. (2013). Review of evidence on health aspects of air pollution – REVIHAAP Project, 309. Retrieved from http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/2013/review-of-evidence-on-health-aspects-of-air-pollution-revihaap-project-final-technical-report.

Xu, J., Saleh, M., Wang, A.A., Tu, R., Hatzopoulou, M., 2019. Embedding local driving behaviour in regional emission models to increase the robustness of on-road emission inventories. Transportation Research Part D: Transport and Environment 73 (June), 1–14. https://doi.org/10.1016/j.trd.2019.05.011.

Xu, J., Wang, J., Hilker, N., Fallah-Shorshani, M., Saleh, M., Tu, R., Hatzopoulou, M., 2018. Comparing emission rates derived from a model with those estimated using a plume-based approach and quantifying the contribution of vehicle classes to on-road emissions and air quality. Journal of the Air and Waste Management Association 68 (11), 1159–1174. https://doi.org/10.1080/10962247.2018.1484395.

Yang, X., Zheng, Y., Geng, G., Liu, H., Man, H., Lv, Z., He, K., de Hoogh, K., 2018. Development of $PM_{2.5}$ and NO_2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. Environ. Pollut. 226 (2), 143–153. https://doi.org/10.1016/j.envpol.2017.03.079.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M.L., Shen, X., Zhang, M., 2017. Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139. https://doi.org/10.1016/J.ATMOSENV.2017.02.023.

Zheng, C., Zhao, C., Li, Y., Wu, X., Zhang, K., Gao, J., Qiao, Q., Ren, Y., Zhang, X., Chai, F., 2018. Spatial and temporal distribution of NO_2 and SO_2 in Inner Mongolia urban agglomeration obtained from satellite remote sensing and ground observations. Atmos. Environ. 188, 50–59. https://doi.org/10.1016/j.atmosenv.2018.06.029.