Environment International 178 (2023) 108106

Contents lists available at ScienceDirect

Environment International

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

Full length article

Predicting spatial variations in annual average outdoor ultrafine particle concentrations in Montreal and Toronto, Canada: Integrating land use regression and deep learning models

Marshall Lloyd [a], Arman Ganji [b], Junshi Xu [b], Alessya Venuta [a], Leora Simon [a], Mingqian Zhang [b], Milad Saeedi [b], Shoma Yamanouchi [b], Joshua Apte [c,d], Kris Hong [a], Marianne Hatzopoulou [b], Scott Weichenthal [a,*]

[a] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Québec H3A 1G1, Canada
[b] Department of Civil and Mineral Engineering, University of Toronto, Toronto, Ontario M5S 1A4, Canada
[c] Department of Civil and Environmental Engineering, University of California at Berkeley, Berkeley, CA 94720, United States
[d] School of Public Health, University of California, Berkeley, CA 94720, United States

ARTICLE INFO

Handling Editor: Xavier Querol

Keywords:
Ultrafine particles
Black Carbon
Deep learning
Images
Land use regression

ABSTRACT

Background: Concentrations of outdoor ultrafine particles (UFP; <0.1 μm) and black carbon (BC) can vary greatly within cities and long-term exposures to these pollutants have been associated with a variety of adverse health outcomes.

Objective: This study integrated multiple approaches to develop new models to estimate within-city spatial variations in annual median (i.e. average) outdoor UFP and BC concentrations as well as mean UFP size in Canada's two largest cities, Montreal and Toronto.

Methods: We conducted year-long mobile monitoring campaigns in each city that included evenings and weekends. We developed generalized additive models trained on land use parameters and deep Convolutional Neural Network (CNN) models trained on satellite-view images. Using predictions from these models, we developed final combined models.

Results: In Toronto, the median observed UFP concentration, UFP size, and BC concentration values were 16,172pt/cm^3, 33.7 nm, and 1225 ng/m^3, respectively. In Montreal, the median observed UFP concentration, UFP size, and BC concentration values were 14,702pt/cm^3, 29.7 nm, and 1060 ng/m^3, respectively. For all pollutants in both cities, the proportion of spatial variation explained (i.e., R^2) was slightly greater (1–2 percentage points) for the combined models than the generalized additive models and a greater (approximately 10 percentage points) than the deep CNN models. The Toronto combined model R^2 values in the test set were 0.73, 0.55, and 0.61 for UFP concentrations, UFP size, and BC concentration, respectively. The Montreal combined model R^2 values were 0.60, 0.49, and 0.60 for UFP concentration, UFP size, and BC concentration models respectively. For each pollutant, predictions from the combined, deep CNN, and generalized additive models were highly correlated with each other and differences between models were explored in sensitivity analyses.

Conclusion: Predictions from these models are available to support future epidemiological research examining long-term health impacts of outdoor UFPs and BC.

Abbreviations: UFP, ultrafine particles; BC, black carbon; LUR, land use regression; CNN, convolutional neural networks; GAM, generalized additive models; MSE, mean squared error; RMSE, root mean squared error.

* Corresponding author at: Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, 2001 McGill College Avenue, Room 1277, Montreal, QC H3A 1G1, Canada.

E-mail addresses: marshall.lloyd@mail.mcgill.ca (M. Lloyd), arman.ganji@utoronto.ca (A. Ganji), junshi.xu@mail.utoronto.ca (J. Xu), alessya.venuta@mail.mcgill.ca (A. Venuta), leora.simon@mail.mcgill.ca (L. Simon), mingqianz.zhang@utoronto.ca (M. Zhang), milad.saeedi@mail.utoronto.ca (M. Saeedi), shoma.yamanouchi@mail.utoronto.ca (S. Yamanouchi), apte@berkeley.edu (J. Apte), kris.hong@alumni.ubc.ca (K. Hong), marianne.hatzopoulou@utoronto.ca (M. Hatzopoulou), scottandrew.weichenthal@mcgill.ca (S. Weichenthal).

https://doi.org/10.1016/j.envint.2023.108106
Received 8 February 2023; Received in revised form 28 June 2023; Accepted 19 July 2023
Available online 22 July 2023

1. Introduction

Ambient air pollution is a heterogenous mix of particles and gases that varies over space and time (de Hoogh et al., 2018; Hoek, 2017; Maciejczyk et al., 2021; HEI, 2022). Associations between exposure to fine particulate matter ($PM_{2.5}$; diameter less than 2.5 μm) and a variety of adverse health outcomes have been well documented and have led to widespread regulations limiting ambient mass concentrations of $PM_{2.5}$ (Boogaard et al., 2019; Boogaard et al., 2022; U.S. EPA, 2019). There has been less research and regulatory action with respect to other forms of particulate matter such as ultrafine particles (UFP, <0.1 μm) and black carbon (BC). Emerging evidence suggests that long-term exposures to UFPs and BC are associated with adverse health outcomes such as cardiovascular mortality and brain tumour incidence; however, exposure assessment in these studies is often a challenge owing to the high spatial variability of UFPs and BC compared to $PM_{2.5}$ (Bouma et al., 2023; Magalhaes et al., 2018; Ohlwein et al., 2019; Weichenthal et al., 2020; HEI, 2022).

A common approach to exposure modelling for UFPs and BC relies on mobile monitoring (Abernethy et al., 2013; Apte et al., 2017; Hankey and Marshall, 2015; Kerckhoffs et al., 2017; Messier et al., 2018; Weichenthal et al., 2014; HEI, 2022; Lloyd et al., 2021). Due to resource constraints, past mobile monitoring campaigns were often less than a month long and restricted to weekdays, which can result in biased estimates of annual averages (Apte et al., 2017; Presto et al., 2021; Saha et al., 2019; HEI, 2022). Recent studies address this by collecting mobile monitoring data over extended periods of time (Blanco et al., 2022). Monitoring data are generally used to develop land use regression (LUR) models which use land use and traffic data from geographical information systems (GIS) to predict spatial variations in pollutant levels (Hoek, 2017, Apte et al., 2017, Kerckhoffs et al., 2017, Hoek et al., 2008). There are various approaches to LUR model development including linear regression and machine learning algorithms. The latter use the same GIS inputs as linear regression models, but may be able to find more complex relationships in the data (Bellinger et al., 2017). Directly comparing LUR model performance statistics across different studies is a challenge due to differences in monitoring and validation approaches, but comparisons of UFP LUR model development approaches within studies have found only modest differences in performance (Kerckhoffs et al., 2019, Weichenthal et al., 2016).

Instead of using GIS data, an emerging approach trains models on images of the urban environment. Street-level and satellite view (i.e. orthogonal) images can be a useful data source since they may contain information not found in traditional GIS databases and models may be able to learn complex associations between the urban environment and ambient air pollution levels (Weichenthal et al., 2019). In one approach, researchers have used various algorithms to extract specific types of features (e.g., greenspace or types of vehicles) from images and were able to estimate associations between these extracted features and ambient air pollution (Qi et al., 2022; Ganji et al., 2020; Xu et al., 2022; Liu et al., 2021). Alternatively, deep Convolutional Neural Networks (CNN) can be trained to directly predict air pollution levels by iteratively learning what combinations of features in the images are associated with the outcome of interest (Albawi et al., 2017, LeCun and Bengio, 1995). Once a such a CNN model is trained, it provides an estimate of ambient air pollution levels at a given location based only on digital images (Lloyd et al., 2021, Xu et al., 2022, Hong et al., 2019, Hong et al., 2020). CNN models have been trained on street-level or satellite view images to predict within-city spatial variation of various air pollutants in London, UK; Vancouver, Toronto, and Montreal, Canada; Los Angeles, USA; and Bucaramanga, Colombia (Lloyd et al., 2021, Hong et al., 2019; Hong et al., 2020; Sorek-Hamer et al., 2022). While LURs are more widely used and accepted, the CNN approach is an emerging tool that uses a separate data stream (i.e., images instead of GIS data) to complement existing methods and our recent research suggested that combined use of LUR and CNN models may outperform either approach on their own (Lloyd et al., 2021). In particular, we observed greatly reduced spatial heterogeneity in model performance in 10-fold cross-validation when LUR and CNN model predictions were combined compared to using the LUR model on its own (Lloyd et al., 2021).

In this study, we conducted year-long mobile monitoring campaigns across Montreal and Toronto, Canada for ambient UFP number concentrations, mean UFP size, and BC mass concentration. These monitoring data were aggregated to 100 m road segments in order to represent annual medians (i.e. averages) at those road segments and the data were split into training, validation, and test sets for model development. Using the aggregated monitoring data, we developed LUR models based on land use and traffic characteristics, and we developed CNN models using aerial images. Predictions from the LUR and CNN models were combined to generate high-resolution models of annual median ambient UFP number concentrations, mean UFP size, and BC mass concentrations. Predictions from these models are available for use in future research, including application in population-based cohorts to investigate the health impacts of long-term exposure to UFPs and possible effect modification by mean UFP size.

2. Materials and methods

2.1. Study setting

The Montreal study area included all municipalities on the island of Montreal (population 1.9 million) and the Toronto study area was within the post-amalgamation political border of the city of Toronto (population 2.9 million). Both cities border major bodies of water, are surrounded by large suburban communities, and have similar climates (Table S1).

2.2. Mobile monitoring study

We conducted a year-long (September 2020 to August 2021) mobile monitoring campaign of outdoor UFP number concentrations, mean UFP size, and BC concentrations. Monitoring routes (Fig. S1) were designed to capture a wide variety of land use and road types. This was done by dividing the study areas into 100 m by 100 m grids, extracting land use parameters for each grid square, and conducting a principal component analysis to identify components that explain the greatest amount of variance in the data. These components were used in Silhouette (Rousseeuw, 1987) and Davies-Bouldin (Davies and Bouldin, 1979) method cluster analyses on the study area grid squares to identify sets of grid squares that, when grouped together explained the greatest amount of variance in land use. Routes were then selected along multiple road types within each of the clusters. To obtain measurements representative of annual medians, these pre-specified routes were repeatedly monitored at various times of day between 7 am and 11 pm, on all days of the week including weekends, and in all four seasons. We randomly selected the time of day, day of week, and the order in which routes were monitored each week. Each monitoring route took approximately 1 h and 3–15 routes were completed each week. Data collected while driving between routes was retained in order to maximize spatial coverage.

2.3. Air pollution measurements

UFP and BC monitors were time-synched with GPS monitors and sampled data at 1 s intervals. The BC monitor was a microAeth MA350 (Aethlabs). The UFP monitors were the Naneos Partector 2 (Naneos) and Testo DiscMini (Testo). We conducted limited collocated measurements, but based on advice from Naneos we elected to not adjust for possible instrument differences. We treated data from both instruments as equivalent since both monitors measure UFPs using the same operating principles and were factory calibrated for the monitoring campaign. For each run, a BC monitor was mounted on the roof of the vehicle and a UFP

monitor was mounted inside the vehicle (Nissan Micra for Toronto and Nissan Rogue for Montreal). The UFP monitor mounted inside the vehicle had its sampling tube inlet extended out the rear passenger side window and pointed down to prevent water from entering the instrument. The tube inlets were roughly 2 m above and in front of the vehicle tailpipes, but the vehicle emissions may have increased air pollution concentrations at the tube inlet. However, it is important to note that we observed UFP concentrations below 1,000 pt/cm^3 and BC concentrations below 50 ng/m^3 suggesting that this contribution was not substantial. After each monitoring run, instruments were inspected and the data offloaded to local and cloud data storage. Files were inspected and data points associated with instrument error codes or implausible temporal or spatial patterns (e.g. constant concentrations) were removed from the analysis. Values above and below the manufacturer's reported limits of detection (listed in Table S2) were imputed with the upper and half of the lower limit of detection, respectively.

2.4. Data Pre-Processing

Air pollution and geospatial position data were joined by matching time stamps. Hourly median ambient weather conditions during monitoring were recorded at airport Automated Surface Observing Systems located within the study area (Table S3) were downloaded using the riem (Salmon and Anderson, 2016) package in version 4.1.2 of the R statistical computing environment (R Core Team, Vienna, Austria) and were matched to the monitoring data by time. Road networks in each city were divided into 100 m road segments. The median of all 1-second air pollution measurements along a given road segment (i.e. the grand median) and the median ambient weather conditions during monitoring were assigned to the centroid of the road segment. The median air pollution levels were our estimates of annual median ambient pollutant levels and are referred to as the "observed" values for the remainder of this text. The unit of analysis for this study was annual median ambient pollutant levels at the 100 m road segments (i.e., the temporal median of all monitoring data along each 100 m road segment; illustrated in Figs. S2 and S3). Road segments monitored on fewer than 6 separate days throughout the campaign were excluded from the analysis. This cut-off was a trade-off between improving temporal stability and retaining good spatial coverage. In total, mobile monitoring data were aggregated to 7051 and 5819 road segments in Montreal and Toronto, respectively. Six-digit geohash codes were assigned to each road segment based on their location and the road segments were randomly split by geohash code into training (70%), validation (15%) and test (15%) sets. The geohash geocoding system spatially splits the globe into cells, each with its own alphanumeric code. A cell with six-digit geohash code is approximately 1.2 km by 0.6 km. Stratifying the random split by geohash code (i.e., geospatial position) increases the independence of the test set and reduces the overlap of images from the training, validation, and test sets (split visualized in Figs. S4 and S5). The distributions of observed UFP number concentrations and BC concentrations were left-skewed thus log-transformed for model development. The distribution of observed mean UFP size was not skewed and not log-transformed.

2.5. Image data for training convolutional neural networks

For each road segment centroid, two aerial (i.e., satellite-view) images were downloaded from Google maps at different zoom levels (18 and 19) using the ggmap (Kahle and Wickam, 2013) package in R. Zoom 18 and 19 images covered areas of approximately 280 m × 280 m and 140 m × 140 m respectively. We compiled our database using images from Google maps because it was an efficient approach to access standardized digital images with good quality control and it is an approach that any researcher can use for any study area in the world. Images were downloaded in 2021 in order to be temporally aligned with the monitoring campaign, though we did not have access to the specific date the images were captured. Images were downloaded with a resolution of 604 × 640 × 3 pixels (three color channels) that was resized to 256 × 256 × 3 to allow for larger training batch sizes and potentially faster CNN model training. Lastly the images were linked to the aggregated monitoring data.

2.6. Land use and traffic

Using an approach similar to previous studies, (Weichenthal et al., 2016, Zalzal et al., 2019; Ripley et al., 2022; Hatzopoulou et al., 2017) land use and traffic parameters plausibly associated with ambient air pollution were extracted using ArcMap 10.8.1 (ESRI, Redlands, USA) from the following data sources: DMTI Spatial (Richmond Hill, CA), Emme (INRO, Montreal, CA), City of Montreal, City of Toronto, Canadian National Pollution Release Inventory, Statistics Canada, Toronto Transit Commission, and Société de Transport de Montreal. Types of land use parameters were: land cover (e.g. industrial area), type of transportation infrastructure (e.g. length or highway or railroad), and points associated with emissions (e.g. distance to airport or number of restaurants). Table S4 describes all 78 spatial predictor variables examined in this study. Buffer sizes were 100 m, 200 m, and 300 m. Larger buffers were not considered because road segment centroids were 100 m apart and UFP and BC concentrations can vary over very short distances.

2.7. Land use regression model development

Generalized additive models (GAM) (Hastie and Tibshirani, 1986, Wood, 2020) for each city were developed to predict spatial variations in annual median outdoor UFP number concentrations, mean UFP size, and BC concentrations in each city (six GAM LUR models). LUR variable selection and model training was done using training set data and following the same method for all three measures of air pollution. Median ambient temperature, relative humidity, and windspeed during monitoring along each road segment were included in all models to account for weather-related temporal variations in air pollution during the monitoring campaign. This "temporal adjustment" is a common approach when developing spatial models using mobile monitoring data (Hankey and Marshall, 2015, Jones et al., 2020, Montagne et al., 2015) and assumes a spatially constant temporal structure for each pollutant across each study area (i.e. the relationships between each meteorological condition and each pollutant are constant throughout each city). Our monitoring campaign was designed to be temporally-balanced, but due to the relatively low number of visits at certain sites (i.e., as few as 6 visits), there were some temporal imbalances between sites and we used temporal adjustment to account for chance weather-related temporal variations in the monitoring data while developing purely spatial models. To select variables for inclusion in the LUR, air pollution levels were first regressed onto each land use variable (listed in Table S4) in univariable regressions. Using those results, variables associated with the air pollutant (95% confidence interval excluding the null) without being driven by outliers (predictor variable values greater than 2 standard deviations from the mean) became candidates land use variables for inclusion in the LUR. To reduce possible collinearity, pairs of candidate land use variables that were correlated with other candidate land use variables were identified (Spearman's correlation >0.7) and in each pair, the land use variable with a higher MSE in the univariable regressions was excluded. The remaining candidate land use variables were used to train the multi-variable LUR model. The multi-variable LUR was trained as a GAM estimated using restricted maximum likelihood. Thin plate splines were used on land use parameters and temporal adjustments to allow for non-linear slopes (limited to 3 basis functions to avoid overfitting). Additional spatial dependencies not captured by land use and traffic parameters were modelled by including road segment latitude and longitude in a tensor product smooth (Wood, 2020, Wood, 2006). Models without latitude and longitude were developed as a

sensitivity analysis. All road segments were treated equally in the analysis and to explore the possible impact of the imbalance in monitoring time across road segments, LUR models were developed using road segments visited on 6 to 12 different days as well as only the road segments along the pre-planned monitoring routes (i.e., excluding data recorded when travelling between monitoring routes).

2.8. Convolutional neural network model development

CNN models for each measure of air pollution in each city were developed using satellite-view images at two different zoom levels (six CNN models). The keras package (Chollet, 2015) in Python was used to train models on the training set data (70%; 9300 images for Montreal and 7330 images for Toronto) and hyperparameter tuning was based on MSE in the validation set (15%). The Xception model architecture (Chollet, 2017) was used because it is a relatively small model that can accept a wide range of image sizes and has demonstrated high accuracy (Chollet, 2017). To reduce training time, we started with pre-trained models using ImageNet initial weights (Deng et al., 2009). The Nadam learning rate optimizer was used for updating weight parameters and minimizing the loss function because it has demonstrated relatively fast convergence when compared to other optimizers (Choi et al., 2020, Dozat, 2016). The initial learning rate was 0.0001 and was reduced if MSE in the validation set plateaued for 5 epochs. Batch size was 128 (32 images per GPU) and models were trained for up to 100 epochs, though training stopped early if MSE in the validation set plateaued for 10 epochs. Model weights that resulted in the lowest MSE in the validation were selected. The CNN models were trained using the aggregated mobile monitoring data (i.e., estimate annual medians) linked to digital satellite images downloaded in 2021. CNN predictions did not account for weather-related temporal variations in the monitoring data and to address this we used the same approach as a previous study (Lloyd et al., 2021) whereby observed air pollution levels are regressed onto the CNN predictions in the validation set with median ambient weather conditions during monitoring included in the model to account for weather-related temporal variations in air pollution (steps illustrated in Fig. S6). The coefficients from this regression were used to adjust CNN model predictions for temporal variations in the monitoring data. As sensitivity analyses, the LUR models were trained using the same approach for temporal adjustment (i.e., training the model in the training set without weather parameters and then adjusting for weather-related temporal variations in the validation set) and both LUR and CNN models were trained without any temporal adjustment.

2.9. Combined model development

For the primary analysis of this study, final combined models for each measure of air pollution in each city were developed by combining LUR predictions with the temporally adjusted CNN model predictions (six combined models). This was done using validation set data in a linear regression:

$$y_i = \beta_0 + \beta_1 x_{LURi} + \beta_2 x_{CNNi} + \epsilon_i$$

where y is the annual median outdoor pollution level and x_{LUR} and x_{CNN} are predictions from the LUR and CNN models respectively for the i^{th} road segment (non-linear regression did not improve model performance). These combined models captured all available information from both the LUR and CNN models.

2.10. Model evaluation

We developed city-specific LUR, CNN, and combined models to predict spatial variations of each pollutant within in each city. Models were trained on the training data (70%), the validation data (15%) were used for CNN hyperparameter tuning, temporal adjustment and combining the LUR and CNN models, and all models were evaluated using the test data (15%). We chose hold-out instead of k-fold cross validation because we had a relatively large sample size (i.e., less variance in sets) and we had a large number of models to develop. Though easier to implement, using a single hold-out set for model evaluation can result in unstable estimates of model performance. The unit of analysis for all models was annual median ambient pollutant levels at the 100 m road segments (i.e., the monitoring data aggregated to road segments). As a sensitivity analysis, multi-city models were developed by pooling data from both cities. Predictions (i.e., estimates) from each model were generated in the test set data and compared to observed values. RMSE and R^2 were used to described model performance. Model residuals for all data were plotted and inspected for spatial clustering. To quantify the spatial clustering of residuals, Moran's Index was calculated on the model residuals in the test set using inverse distance between road segments as the weights.

2.11. Prediction surfaces

The study areas were divided into 100 m $\times$ 100 m cells for the prediction surfaces. For each cell, land use and traffic parameters were extracted and used to generate LUR model predictions and satellite-view images at both zoom levels were downloaded to generate CNN models predictions. spatially invariant (i.e., constant) annual median temperature, humidity, and wind speed at local airports (i.e., the same data source used for weather conditions during monitoring) were used for each city when generating predictions. This assumes a spatially constant temporal structure across the study areas and also assumes that predictions of pollutant levels under average regional meteorological conditions represent annual median outdoor pollutant levels. Surfaces from the combined models provided estimates of within-city spatial variation of annual ambient pollution.

2.12. CNN model behaviour

CNN models lack the easily interpretable coefficients of regression models; thus, we explored CNN predictions to investigate model behaviour. CNN prediction surfaces were inspected and compared LUR surfaces. Following an approach described by Sorek-Hamer et al. (2022), digital images were modified and CNN model predictions were compared to expected values. For example, pasting the image of a highway into the image of a residential area was expected to increase the predicted concentration with respect to the unmodified image of the residential area. The resulting prediction was compared to expectations to determine if the model behaved in a manner consistent with expectations. Google periodically updates satellite-view images, and we generated CNN predictions using images from different time periods to explore the sensitivity of CNN models to time of year an image was captured (e.g., comparing predictions based on images captured during the summer when trees are full of green leaves to images captured during early spring when there are no green leaves).

3. Results

3.1. Monitoring

We conducted over 700 h of mobile monitoring, of which data from over 500 h were retained for model development. Monitoring data were retained for 12,870 road segments (7051 in Montreal and 5819 in Toronto) that were visited on at least 6 different days during the campaign. Median observed pollutant levels on road segments monitored on fewer than 6 different days were considered temporally unstable and not representative of annual outdoor medians, thus they were excluded from the analysis (Tables S5–S6 and Figs. S7–S8 compare retained and discarded monitoring data). Approximately 0.5% of the recorded data was outside of instrument detection ranges and was thus

imputed. Road segments (100 m) were visited on a median of 10 different days (sd = 8) and monitored for a median total duration of 63 s (sd = 640). As shown in Table 1, Toronto had slightly higher median observed concentrations of UFP and BC and larger median observed UFP size. Table S7 shows descriptive statistics for the training, validation, and test sets. Tables S8 and S9 show the months and days of the week that monitoring occurred. Observed UFP and BC concentrations during monitoring were slightly lower than observed values during previous campaigns in Montreal and Toronto (Table S10; Weichenthal et al., 2016, Weichenthal et al., 2016, Minet et al., 2018), likely due to this campaign including evening and weekends whereas the previous campaign focused on rush-hour periods during weekdays.

3.2. Model performance

Variables selected for each LUR model are listed in Table S11. Nineteen, 22, and 16 land use and traffic variables were selected into the Montreal UFP concentration, UFP size and BC concentrations LUR models, respectively. Fourteen, 18, and 13 land use and traffic variables were selected into the Toronto UFP concentration, UFP size and BC concentrations LUR models, respectively. Correlations between predictor variables are shown in Figs. S9–S14. All city-specific model R^2 and combined models coefficients are in Table 2 (RMSE in Table S12). All models had generally similar performance, with combined models having the highest R^2 though it was only slightly higher than the LUR R^2. LUR and CNN model predictions were highly correlated (Table S13 shows Pearson r and for all models it was near 0.8), but the LUR models had higher R^2. The LUR predictions also had slightly larger coefficients in the combined models, thus made somewhat greater contributions to the combined models than the CNN predictions. All city-specific models had higher R^2 than the multi-city models trained on pooled data (Table S14). LUR temporal adjustment in the validation set instead of the training set and omitting the temporal adjustment of LUR and CNN models had very little impact on model R^2 (Tables S15 and S16). Maps of median meteorological conditions during monitoring show some spatial variation (Fig. S15) and response curves for meteorological terms in the models (Figs. S16 and S17) show relatively modest associations with pollutant concentrations across the monitoring sites. Restricting the training data to road segments visited 6 to 12 times and restricting it to only the pre-planned monitoring routes both had very little impact on model R^2 (Table S17 and Fig. S18).

3.3. Model bias

Table 3 shows median differences between observed values (i.e., aggregated monitoring data) and model predictions in the test set and compared to the range of observed values of each pollutant, the mean differences were relatively small. The median differences for all the Montreal UFP concentration models were similar, but the Toronto LUR on average slightly underpredicted compared to observed values, whereas the Toronto CNN and combined models slightly overpredicted. For BC concentrations, all models slightly underpredicted on average. This contrast in LUR and CNN UFP concentration model behaviour between cities was explored further in scatter plots of observed and predicted UFP number concentrations (Fig. 1) and plots of LUR and CNN predictions (Fig. 2, Figs. S19–S21, Table S18, and Table S19). The Toronto CNN and combined UFP concentration models generated more predictions above 45,000 pt/cm^3 than the Toronto LUR model. Conversely, the Montreal LUR model generated a greater number of elevated UFP concentration predictions than the CNN or combined model. This diverging pattern in UFP concentration predictions was much less pronounced when using multi-city CNN and LUR models that were trained on both Montreal and Toronto data (Figs. S22–S25). This suggests that the CNN UFP concentration model trained on Toronto data alone may have learned features specific to Toronto that were associated with very elevated UFP concentrations (Fig. S26). BC mass concentration and UFP size predictions did not exhibit the diverging patterns observed for UFP concentrations in Fig. 1 (Figs. S27–S29). The Toronto LUR UFP concentration predictions exhibited unexpected clustering, but the Toronto CNN model predictions did not (Fig. S30). This was in part due to some of the test set data being clustered along major highways (Figs. S31 and S32) that had distinctly elevated values of vehicle traffic (Fig. S33) which was an important variable the LUR model (Fig. S34). The same clustered test set did not lead to clustered CNN model predictions because CNNs are not trained on distinct categories of parameters, but instead learn complex features in digital images. Slopes and intercepts of the scatter plots are listed in Table S20. Model errors were mapped and although they did not appear to be spatially clustered (Figs. S35–S38), the test of Moran's Index (Table S21) for all models rejected the null hypothesis that there was no spatial clustering of residuals.

3.4. Prediction surfaces

Fig. 3 shows the UFP number concentration LUR, CNN, and combined model prediction surfaces for Toronto (A) and Montreal (B). Fig. 4 shows the combined model surfaces for UFP size and areas with small particle sizes were areas with generally higher UFP concentrations. Combined model BC concentration prediction surfaces shown in Fig. 5 and other models shown in Figs. S39–S44. Differences between LUR and CNN prediction surfaces are in Figs. S45–S47 and show that for major highways, the Montreal CNN model consistently generated higher UFP number concentration predictions than the LUR model, whereas for Toronto it was the opposite. This contrast in prediction surfaces was consistent with the model behaviour observed in the test set and discussed in the Model Bias section (Figs. S19–S25). LUR surfaces appeared more spatially smooth than CNN surfaces, but removing latitude and longitude from LUR models resulted in less spatial smoothing (Figs. S48–S50) and reduced the very elevated UFP number concentration predictions in North-Western Toronto. CNN model predictions using modified images were generally consistent with our expectations (e.g., the CNN UFP concentration prediction using an unmodified image of a Montreal residential area was 7935 pt/cm^3, but when we modified the image by inserting the image of a major highway, the CNN prediction increased to 25,259 pt/cm^3; Figs. S51–S54). However, there were some predictions on modified images that did not meet our expectations such as when a BC concentration prediction using the image of a Montreal residential area increased by 99 ng/m^3 after the image of a forest was inserted (Fig. S52). The handful of predictions that did not meet our expectations may indicate a source of error in model predictions, but this

Table 1
Mobile Monitoring Campaign Descriptive Statistics by City.

Pollutant	City	Median (IQR)	5th – 95th Percentile	Number of 100 m Road Segments
UFP number concentration (pt/cm^3)	Toronto	16,172 (14,991)	6895 – 53,710	5819
	Montreal	14,702 (13,549)	4868 – 46,063	7051
Mean UFP Size (nm)	Toronto	33.7 (7.7)	23 – 44	5819
	Montreal	29.7 (10.1)	19 – 46	7051
BC mass concentration (ng/m^3)	Toronto	1225 (1151)	447 – 3197	5348
	Montreal	1060 (1006)	277 – 2789	7112

Table 2
Model performance in test set and combined model coefficients. *UFP number concentration and BC concentration log-transformed for model development.

City	Pollutant	R^2 in Test Set			Combined Model Coefficients		
		LUR	CNN	Combined	Intercept	LUR	CNN
Montreal	UFP PNC* (pt/cm^3)	0.59	0.49	0.60	−0.25	0.57	0.46
	UFP Size (nm)	0.48	0.41	0.49	1.49	0.52	0.44
	BC Conc.* (ng/m^3)	0.58	0.50	0.60	−0.13	0.52	0.50
Toronto	UFP PNC* (pt/cm^3)	0.71	0.66	0.73	−1.34	0.65	0.49
	UFP Size (nm)	0.56	0.43	0.55	−2.34	0.56	0.51
	BC Conc.* (ng/m^3)	0.60	0.53	0.61	−0.69	0.72	0.38

Table 3
Mean difference between observed and predicted values in the test set and the 5th to 95th percentile range of observed values used for model development (i.e., the aggregated monitoring data).

Pollutant	Model	Median Difference (5th, 95th percentile)		Observed 5th-95th Percentile Range	
		Montreal	Toronto	Montreal	Toronto
UFP PNC (pt/cm3)	LUR	−338 (−11332, 17177)	533 (−12285, 22961)	41,195	46,815
	CNN	−232 (−8948, 24100)	−1530 (−18211, 16988)		
	Combined	−125 (−8924, 21220)	−770 (−20864, 16063)		
UFP Size (nm)	LUR	−0.35 (−9.18, 10.14)	−0.06 (−6.37, 6.8)	27	21
	CNN	−0.79 (−10.17, 11.83)	0.72 (−6.64, 8.75)		
	Combined	−0.8 (−9.64, 10.2)	0.54 (−5.96, 7.37)		
BC Concentration (ng/m3)	LUR	51 (−617, 1087)	71 (−812, 1282)	2512	2750
	CNN	91 (−511, 1331)	44 (−818, 1312)		
	Combined	75 (−462, 1156)	40 (−1031, 1202)		

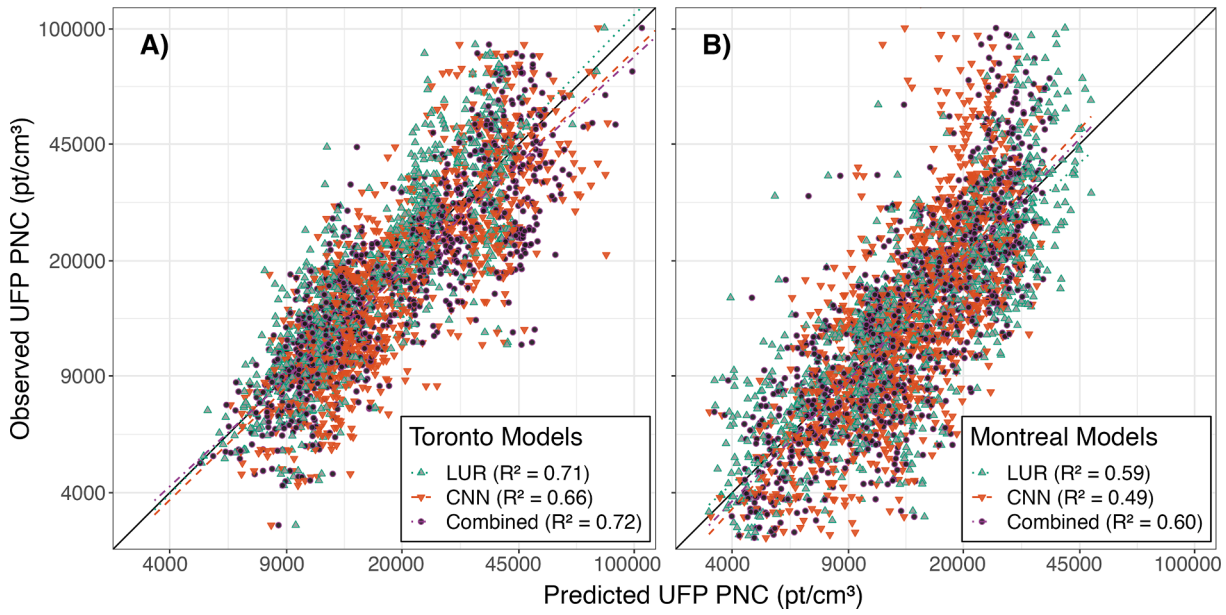

Fig. 1. Comparing observed to predicted UFP particle number concentrations (PNC) in Toronto (A) and Montreal (B) with legends showing model performance in the test set. Figs. S27–S29 show comparisons for each pollutant in 6-panel plots instead of the 2-panel plots shown here.

sensitivity analysis using modified images is likely an inadequate test of the complex relationships learned by the CNN models. Some CNN predictions appeared to be somewhat sensitive to the season in which images were captured, but the impact of season on estimated air pollution levels was approximately normally distributed around zero (Figs. S55–S57), suggesting a reduction in prediction precision rather than biased predictions. Figs. S58–S81 show several explorations of CNN model behaviour including comparisons with LUR predictions.

4. Discussion

In this study, we developed new high-resolution exposure models to predict within-city spatial variations in outdoor UFP number concentrations, mean UFP size, and BC mass concentrations for Canada's two largest cities. This analysis improves on our earlier models (Weichenthal et al., 2016, Weichenthal et al., 2016) by increasing the spatial coverage of the monitoring campaign, increasing the total monitoring time, extending the monitoring period over an entire year, randomly sampling all days of the week and most times of day, and incorporating information from digital satellite-view images into model predictions using convolutional neural networks. The increased spatial and temporal coverage of this monitoring campaign compared to our previous effort likely resulted in a more representative sample of the within-city spatial variations of annual median ambient air pollution. Model R^2 values cannot be directly compared across studies, but the R^2 values of our LUR, CNN, and combined models fell within the range of

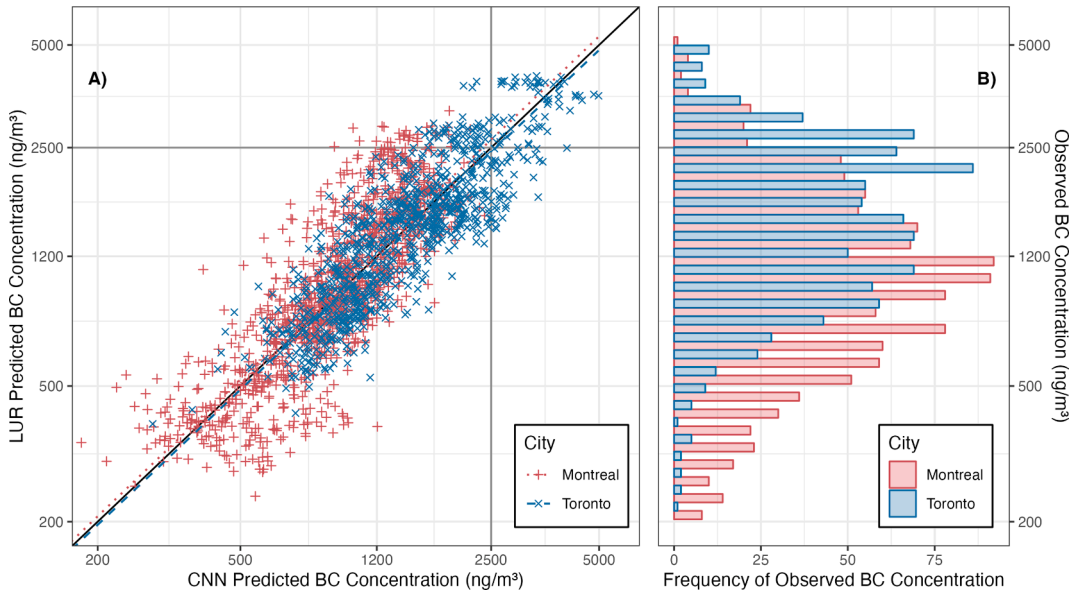

Fig. 2. Comparing test set LUR UFP number concentration predictions to CNN predictions in Montreal and Toronto (A) with histogram of observed values for reference (B). For high UFP number concentration (e.g., near 45,000 pt/cm^3), the Montreal CNN predictions appeared to the systematically lower than the Montreal LUR predictions, whereas it was the opposite in Toronto. Predictions for road segments with observed UFP number concentration values greater than 45,000 pt/cm^3 are indicated with a grey circle in A) and were found on major highways (Fig. S32). The mean difference in predicted UFP number concentration between the Montreal LUR and Montreal CNN was 271 pt/cm^3 (5th, 95th percentile: −5469, 12,175). For Toronto it was −2232 pt/cm^3 (5th, 95th percentile: −17,070, 6114). Pearson correlation coefficient of the LUR and CNN model predictions was 0.80 for Montreal and 0.86 for Toronto.

published R^2 values from other studies that developed linear regression LUR models, machine learning LUR models, or CNN models trained on images of the urban environment (Lloyd et al., 2021, Kerckhoffs et al., 2019, Hong et al., 2020, Weichenthal et al., 2016, van Nunen et al., 2017). The magnitude of the bias observed in our models was similar to the bias reported by other studies as well (Hankey and Marshall, 2015, Lloyd et al., 2021, Kerckhoffs et al., 2019, Weichenthal et al., 2016, Montagne et al., 2015, van Nunen et al., 2017). This consistency of model performance and bias suggests that each approach in this study could be useful for estimating within-city spatial variations in air pollution. A further improvement over our earlier models was the development of mean UFP size models that generally predicted small mean UFP sizes in areas of elevated UFP concentrations, which is consistent with our understanding of particle growth (i.e. fresh emissions consisting of high concentrations of very small particles) (Kittelson et al., 2022, Kwon et al., 2020). The development of mean UFP size models is important because particle size may play a role in UFP toxicity (Huang et al., 2021; Moreno-Ríos et al., 2022; Shang et al., 2021) and has the potential to confound the relationship between UFP concentrations and adverse health outcomes (i.e., smaller particles may be more harmful and are typically present in higher concentrations) (Weichenthal et al., 2022). Collectively, this investigation produced a number of interesting results.

First, we observed high within-city spatial variations in outdoor UFP concentrations, mean UFP size, and BC concentrations during monitoring. This is consistent with findings from other studies that report outdoor UFP and BC concentrations having much greater within-city spatial variation than outdoor $PM_{2.5}$ concentrations (Alonso-Blanco et al., 2018; Apte et al., 2017; Chambliss et al., 2020; Presto et al., 2021; Evans et al., 2019; HEI, 2022). The high-resolution models we developed explained more than half of the observed spatial variation in UFP and BC concentrations in the test sets. Due to their relatively low R^2, applying the UFP size models in a health study may introduce measurement error and reduce the precision of estimated health effects. City-specific models performed better than multi-city models trained on pooled data, which is consistent with the documented difficulty of transferring models between study areas (Hoek, 2017, Zalzal et al., 2019, Allen et al., 2011). The CNN models performed somewhat worse than the LUR models, but they took advantage of an alternative data stream (i.e., images instead of GIS data) and may be learning complex associations that are not present in GIS data alone (Fig. S26). Combined models performed better than any LUR or CNN models on their own, though with only a modest increase in R^2 compared to the LUR models. This is consistent with the results from a similar study (Lloyd et al., 2021) and suggests that CNNs trained on images can be useful for predicting within-city spatial variations in outdoor air pollution, especially when combined with LUR models. Nonetheless, CNNs can learn unintended associations between image features and underlying structures in the data which can affect generalizability (Bowyer et al., 2020; Zech et al., 2018; Ribeiro et al., 2016). An example of an underlying structure is images in a database that are captured during different seasons. Our CNNs appeared to be somewhat sensitive to the time of year images were captured (Figs. S55–S57) which likely introduced some random error into our estimates. Nevertheless, CNN models are likely most useful in places lacking large curated databases of land use and traffic information as recently demonstrated in Bucaramanga, Colombia (Lloyd et al., 2021).

Secondly, for each pollutant the LUR and CNN models generated similar prediction surfaces, yet there were several interesting differences (Figs. 3 - 4 and Figs. S45–S47). For instance, the LUR model prediction surfaces were generally smoother than those from CNN models. This was due in part to the inclusion of latitude and longitude in the LUR models (Figs. S48–S50). Latitude and longitude vary incrementally throughout the study area and smoothed the LUR predictions, whereas each CNN-generated prediction was based solely on digital images that covered up to 280 m × 280 m of the earth's surface (i.e., the CNN prediction for a given point was naïve of any information beyond the edge of the image centered on that point). UFP and BC concentrations can vary greatly over very short distances (Apte et al., 2017, Presto et al., 2021, Evans et al., 2019; Alonso-Blanco et al., 2018; Chambliss et al., 2020) and it is possible LURs may have over-smoothed the spatial variations in certain areas (Figs. S67 and S77). In other areas however, the CNN being naïve of information beyond the limits of the images may have resulted in under-smoothing (Figs. S68 and S74). Combined model prediction

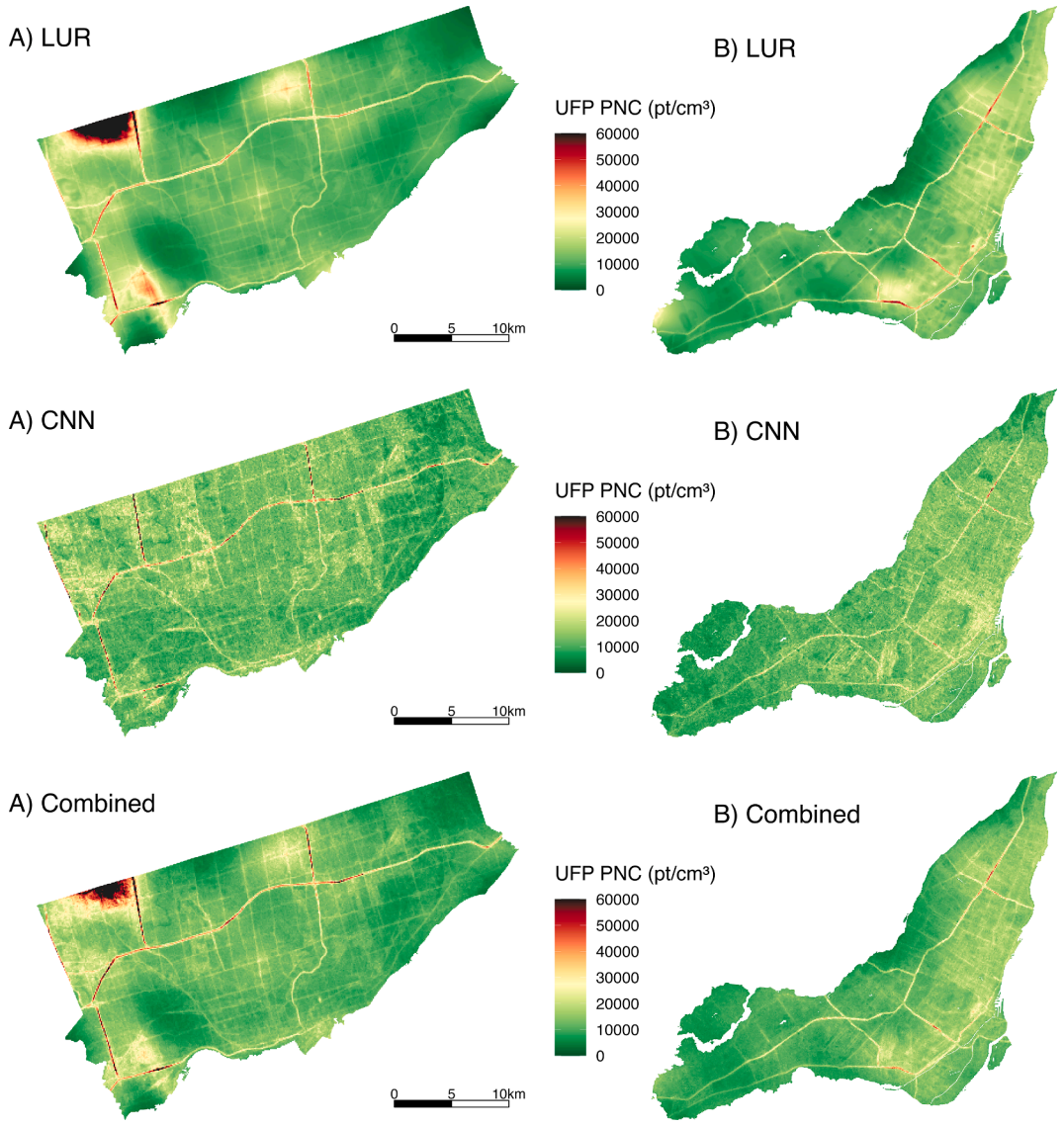

Fig. 3. Toronto (A) and Montreal (B) UFP particle number concentration (PNC) prediction surfaces for the LUR, CNN, and combined models (100 m × 100 m resolution). Figs. S58–S81 show an exploration of the Toronto CNN surface dark green bands and other areas of interest. Figs. S48–S50 show LUR model surfaces without latitude and longitude, which were generally more similar to the CNN surfaces. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

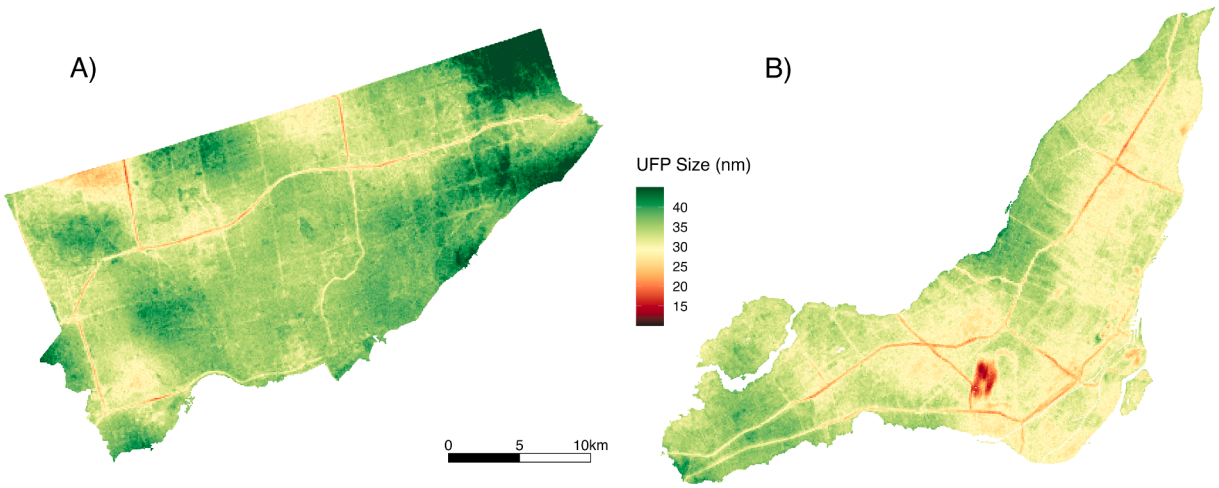

Fig. 4. Toronto (A) and Montreal (B) combined model UFP size prediction surfaces (100 m × 100 m resolution). LUR and CNN model surfaces are in Figs. S39–S41.

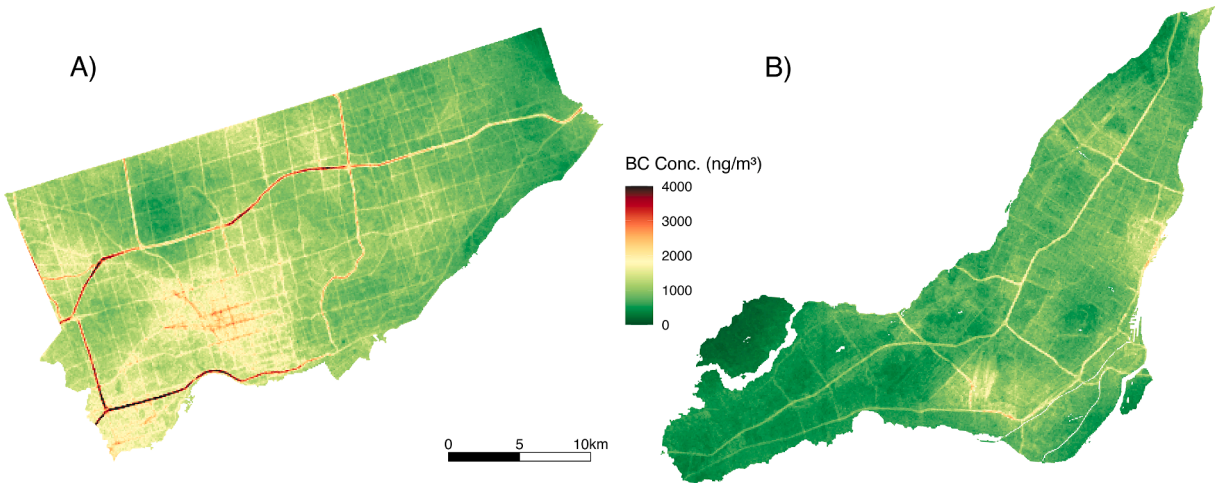

Fig. 5. Toronto (A) and Montreal (B) combined model BC concentration prediction surfaces (100 m × 100 m resolution). LUR and CNN model surfaces are in Figs. S42–S44.

surfaces appeared to integrate the smoothness of the LUR surfaces with the sharp gradients of the CNN surfaces (Fig. 3), which may be a useful compromise between the two approaches. Furthermore, mapping the difference between LUR and CNN model predictions highlighted interesting contrasts. For example, on major highways the Montreal LUR model consistently predicted higher UFP number concentrations than the Montreal CNN model whereas in Toronto it was the opposite (Figs. S19–S25 and S45–S47). In general, combining LUR and CNN predictions resulted in only a modest increase in overall model performance compared to the LUR models alone, but may help generate more robust predictions throughout the modelling areas by taking advantage of information from both land use data and digital images. Indeed, in our previous study conducted in Colombia, spatial variations in model errors were lower for CNN models than for LUR models (Lloyd et al., 2021).

A strength of this study was the large scope of the monitoring campaign and mobile monitoring was an efficient approach to maximize spatial coverage, though a limitation was relatively low monitoring time per road segment compared to stationary monitoring (Kerckhoffs et al., 2017, van Nunen et al., 2017). On average, road segments were visited on 10 different days for a total of roughly 60 s of monitoring. Although researchers have successfully developed models based on similar levels of monitoring, (Messier et al., 2018, Kerckhoffs et al., 2017) longer monitoring times would likely provide more stable estimates of annual median ambient UFP and BC levels and reduced temporal imbalances between sites which may have eliminated the need for temporal adjustment of the models and the assumption of spatially constant temporal structures across the study areas. Additionally, mobile monitoring was conducted using internal combustion engine vehicles which may have contributed to the monitoring vehicle emissions being measured. Furthermore, monitoring was conducted on roads and major highways, which likely resulted in our measured values of air pollution being higher than the air pollution values immediately outside residences. Though we monitored on-road, we still observed low UFP levels at many monitoring sites (e.g. over 400 sites with median values less than 5000 pt/cm^3) and our approach did not preclude the identification of such locations. Nonetheless, future monitoring campaigns could address these limitations by following the approach used by Blanco et al. (2022), which involved monitoring from stationary vehicles parked at pre-specified sites on the sides of non-highway roads (Blanco et al., 2022). If applied to a health study, our models would likely overestimate the absolute value of residential outdoor concentrations and thus investigating absolute ambient air pollution thresholds using our estimates could be challenging. Nonetheless, the spatial contrasts between residential exposures would still be informative and useful in epidemiological analyses examining the long-term health impacts of these exposures. Another strength of this study was the incorporation of information from digital images to improve predictions, however the application of CNNs can be challenging. Firstly, CNNs require a large amount of training data and may not be applicable for smaller monitoring campaigns. CNNs also do not have the easily interpretable coefficients of linear regression models, thus we explored CNN model behaviour using several approaches (Figs. S51–S81). Lastly, quality control of digital images is an extremely important and potentially resource intensive step (Santos et al., 2021, Pelletier et al., 2017) when training CNN models. For example, past applications of CNNs have erroneously learned structural flaws in the data when training models on images from multiple databases (Noseworthy et al., 2020, Heaven, 2019). Using R to download Google Maps satellite images was an efficient approach to compile a high-quality database of digital images, but we could not control the exact timing of image capture. This led to some images being from different seasons during the year-long campaign and likely had a small impact on CNN model predictions. Future studies should consider allocating resources to establishing high-quality databases of digital images for CNN model training and possibly developing methods to take advantage of seasonal differences in digital images to generate robust estimates of spatial variations in air pollution.

5. Conclusion

We conducted a year-long monitoring campaign and developed new high-resolution models of within-city spatial variation in annual median (i.e. average) outdoor UFP number concentrations, mean UFP size, BC concentrations for Canada's two largest cities. The best Toronto models had R^2 values in the test set of 0.73, 0.55, and 0.61 for UFP concentrations, UFP size, and BC concentration, respectively. The best Montreal models had R^2 values in the test set of 0.60, 0.49, and 0.60 for UFP concentration, UFP size, and BC concentration models respectively. The CNN models had somewhat lower R^2 values than the LUR models, but still showed good performance and had the advantage that they did not need an extensive database of land use information. These models are available for use in future research, including application in population-based cohorts to investigate the health impacts of long-term exposure to these pollutants.

Funding Sources.

The research described in this article was conducted under contract to the Health Effects Institute (HEI) (Grant Number: 4976-RFA19-1/20–10), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award CR 83998101) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI or its sponsors, nor do

they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers.

CRediT authorship contribution statement

Marshall Lloyd: Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Supervision. **Arman Ganji:** Software, Formal analysis, Investigation, Data curation. **Junshi Xu:** Software, Formal analysis, Investigation, Data curation. **Alessya Venuta:** Investigation, Data curation. **Leora Simon:** Investigation, Data curation, Project administration. **Mingqian Zhang:** Investigation, Data curation. **Milad Saeedi:** Software, Formal analysis, Investigation, Data curation. **Shoma Yamanouchi:** Investigation, Data curation. **Joshua Apte:** Conceptualization, Methodology. **Kris Hong:** Software, Methodology. **Marianne Hatzopoulou:** Conceptualization, Methodology, Investigation, Resources, Writing – review & editing, Supervision. **Scott Weichenthal:** Conceptualization, Methodology, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Scott Weichenthal reports financial support was provided by Health Effects Institute.

Data availability

Data will be made available on request.

Acknowledgements

The research described in this article was conducted under contract to the Health Effects Institute (HEI) (Grant Number: 4976-RFA19-1/20-10), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award CR 83998101) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2023.108106.

References

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

Aethlabs. microAeth® / MA350. Retrieved on 20 April, 2022. https://aethlabs.com/microaeth/ma350/tech-specs.

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET) 1–6 (2017). doi:10.1109/ICEngTechnol.2017.8308186.

Allen, R.W., Amram, O., Wheeler, A.J., Brauer, M., 2011. The transferability of NO and NO2 land use regression models between cities and pollutants. Atmos. Environ. 45, 369–378.

Alonso-Blanco, E., Gómez-Moreno, F.J., Artíñano, B., Iglesias-Samitier, S., Juncal-Bello, V., Piñeiro-Iglesias, M., López-Mahía, P., Pérez, N., Brines, M., Alastuey, A., García, M.I., Rodríguez, S., Sorribas, M., del Águila, A., Titos, G., Lyamani, H., Alados-Arboledas, L., 2018. Temporal and spatial variability of atmospheric particle number size distributions across Spain. Atmos. Environ. 190, 146–160.

Apte, J.S., Messier, K.P., Gani, S., Brauer, M., Kirchstetter, T.W., Lunden, M.M., Marshall, J.D., Portier, C.J., Vermeulen, R.C.H., Hamburg, S.P., 2017. High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. Environ. Sci. Technol. 51, 6999–7008.

Bellinger, C., Mohomed Jabbar, M.S., Zaïane, O., Osornio-Vargas, A., 2017. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health 17, 907.

Blanco, M.N., Gassett, A., Gould, T., Doubleday, A., Slager, D.L., Austin, E., Seto, E., Larson, T.V., Marshall, J.D., Sheppard, L., 2022. Characterization of Annual Average Traffic-Related Air Pollution Concentrations in the Greater Seattle Area from a Year-Long Mobile Monitoring Campaign. Environ. Sci. Technol. 56, 11460–11472.

Boogaard, H., Walker, K., Cohen, A.J., 2019. Air pollution: the emergence of a major global health risk factor. Int. Health 11, 417–421.

Boogaard, H., Patton, A.P., Atkinson, R.W., Brook, J.R., Chang, H.H., Crouse, D.L., Fussell, J.C., Hoek, G., Hoffmann, B., Kappeler, R., Kutlar Joss, M., Ondras, M., Sagiv, S.K., Samoli, E., Shaikh, R., Smargiassi, A., Szpiro, A.A., Van Vliet, E.D.S., Vienneau, D., Weuve, J., Lurmann, F.W., Forastiere, F., 2022. Long-term exposure to traffic-related air pollution and selected health outcomes: A systematic review and meta-analysis. Environ. Int. 164, 107262.

Bouma, F., Janssen, N.A., Wesseling, J., van Ratingen, S., Strak, M., Kerckhoffs, J., Gehring, U., Hendricx, W., de Hoogh, K., Vermeulen, R., Hoek, G., 2023. Long-term exposure to ultrafine particles and natural and cause-specific mortality. Environ. Int. 175, 107960.

Bowyer, K.W., King, M.C., Scheirer, W.J., Vangara, K., 2020. The, "Criminality From Face" Illusion. IEEE Trans. Technol. Soc. 1, 175–183.

Chambliss, S.E., Preble, C.V., Caubel, J.J., Cados, T., Messier, K.P., Alvarez, R.A., LaFranchi, B., Lunden, M., Marshall, J.D., Szpiro, A.A., Kirchstetter, T.W., Apte, J.S., 2020. Comparison of Mobile and Fixed-Site Black Carbon Measurements for High-Resolution Urban Pollution Mapping. Environ. Sci. Technol. 54, 7848–7857.

Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C.J., Dahl, G.E., 2020. On Empirical Comparisons of Optimizers for Deep Learning. Preprint at https://doi.org/10.48550/arXiv.1910.05446 (2020).

Chollet, F., 2015. Keras: Deep learning library for theano and tensorflow. URL: https://keras.io/k **7**, T1 (2015).

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 1251–1258.

Davies, D.L., Bouldin, D.W., 1979. A Cluster Separation Measure. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-1**, 224–227 (1979).

de Hoogh, K., Chen, J., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Klompmaker, J., Martin, R.V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Brunekreef, B., Hoek, G., 2018. Spatial PM2.5, NO2, O3 and BC models for Western Europe – Evaluation of spatiotemporal stability. Environ. Int. 120, 81–92.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition 248–255 (2009).

Dozat, T. Incorporating Nesterov Momentum into Adam. ICLR 2016 workshop paper 107 review 10 (2016).

Evans, G. J., Audette, C., Badali, K., Celo, V., Dabek-Zlotorszynka, E., Debosz, J., Ding, L., Doerksen, G. N., Healy, R. M., Henderson, D., Herod, D., Hilker, N., Jeong, C.-H., Johnson, D., Jones, K., Munoz, A., Noble, M., Reid, K., Schiller, C., Sofowote, U., Su, Y., Wang, J. & White, L. Near-Road Air Pollution Pilot Study Final Report. (2019).

Ganji, A., Minet, L., Weichenthal, S., Hatzopoulou, M., 2020. Predicting Traffic-Related Air Pollution Using Feature Extraction from Built Environment Images. Environ. Sci. Technol. 54, 10688–10699.

Hankey, S., Marshall, J.D., 2015. Land Use Regression Models of On-Road Particulate Air Pollution (Particle Number, Black Carbon, PM2.5, Particle Size) Using Mobile Monitoring. Environ. Sci. Technol. 49, 9194–9202.

Hastie, T., Tibshirani, R., 1986. Generalized Additive Models. Stat. Sci. 1, 297–310.

Hatzopoulou, M., Valois, M.F., Levy, I., Mihele, C., Lu, G., Bagg, S., Minet, L., Brook, J., 2017. Robustness of Land-Use Regression Models Developed from Mobile Air Pollutant Measurements. Environ. Sci. Technol. 51, 3938–3947.

Heaven, D., 2019. Why deep-learning AIs are so easy to fool. Nature 574, 163–166.

HEI, 2022. HEI Panel on the Health Effects of Long-Term Exposure to Traffic-Related Air Pollution - Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. Special Report 23. Boston, MA: Health Effects Institute (2022).

Hoek, G., 2017. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. Curr. Environ. Health Reports 4, 450–462.

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. 42, 7561–7578.

Hong, K.Y., Pinheiro, P.O., Minet, L., Hatzopoulou, M., Weichenthal, S., 2019. Extending the spatial scale of land use regression models for ambient ultrafine particles using satellite images and deep convolutional neural networks. Environ. Res. 176, 108513.

Hong, K.Y., Pinheiro, P.O., Weichenthal, S., 2020. Predicting outdoor ultrafine particle number concentrations, particle size, and noise using street-level images and audio data. Environ. Int. 144, 106044.

Huang, C., Tang, M., Li, H., Wen, J., Wang, C., Gao, Y., Hu, J., Lin, J., Chen, R., 2021. Particulate matter air pollution and reduced heart rate variability: How the associations vary by particle size in Shanghai, China. Ecotoxicol. Environ. Saf. 208, 111726.

Jones, R.R., Hoek, G., Fisher, J.A., Hasheminassab, S., Wang, D., Ward, M.H., Sioutas, C., Vermeulen, R., Silverman, D.T., 2020. Land use regression models for ultrafine particles, fine particles, and black carbon in Southern California. Sci. Total Environ. 699, 134234.

Kahle, D., Wickam, H., 2013. ggmap: Spatial Visualization with ggplot2. The R Journal. 5 (1), 144-161. https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

Kerckhoffs, J., Hoek, G., Vlaanderen, J., van Nunen, E., Messier, K., Brunekreef, B., Gulliver, J., Vermeulen, R., 2017. Robustness of intra urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring. Environ. Res. 159, 500–508.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., Vermeulen, R.C.H., 2019. Performance of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces. Environ. Sci. Technol. 53, 1413–1421.

Kittelson, D., Khalek, I., McDonald, J., Stevens, J., Giannelli, R., 2022. Particle emissions from mobile sources: Discussion of ultrafine particle emissions and definition. J. Aerosol Sci 159, 105881.

Kwon, H.-S., Ryu, M.H., Carlsten, C., 2020. Ultrafine particles: unique physicochemical properties relevant to health and disease. Exp. Mol. Med. 52, 318–328.

LeCun, Y., Bengio, Y., 1995. Convolutional networks for images, speech, and time series. Handbook Brain Theory Neural Networks 3361, 10.

Liu, X., Zhang, X., Schnelle-Kreis, J., Jakobi, G., Cao, X., Cyrys, J., Yang, L., Schloter-Hai, B., Abbaszade, G., Orasche, J., Khedr, M., Kowalski, M., Hank, M., Zimmermann, R., 2021. Spatiotemporal Characteristics and Driving Factors of Black Carbon in Augsburg, Germany: Combination of Mobile Monitoring and Street View Images. Environ. Sci. Technol. 55, 160–168.

Lloyd, M., Carter, E., Diaz, F. G., Magara-Gomez, K. T., Hong, K. Y., Baumgartner, J., Herrera G, V. M. & Weichenthal, S., 2021. Predicting Within-City Spatial Variations in Outdoor Ultrafine Particle and Black Carbon Concentrations in Bucaramanga, Colombia: A Hybrid Approach Using Open-Source Geographic Data and Digital Images. *Environ. Sci. Technol.* **55**, 12483–12492.

Maciejczyk, P., Chen, L.-C., Thurston, G., 2021. The Role of Fossil Fuel Combustion Metals in PM2.5 Air Pollution Health Associations. Atmos. 12, 1086.

Magalhaes, S., Baumgartner, J., Weichenthal, S., 2018. Impacts of exposure to black carbon, elemental carbon, and ultrafine particles from indoor and outdoor sources on blood pressure in adults: A review of epidemiological evidence. Environ. Res. 161, 345–353.

Messier, K.P., Chambliss, S.E., Gani, S., Alvarez, R., Brauer, M., Choi, J.J., Hamburg, S.P., Kerckhoffs, J., LaFranchi, B., Lunden, M.M., Marshall, J.D., Portier, C.J., Roy, A., Szpiro, A.A., Vermeulen, R.C.H., 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, 12563–12572.

Minet, L., Liu, R., Valois, 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, 3512–3519.

Montagne, D.R., Hoek, G., Klompmaker, J.O., Wang, M., Meliefste, K., Brunekreef, B., 2015. Land Use Regression Models for Ultrafine Particles and Black Carbon Based on Short-Term Monitoring Predict Past Spatial Variation. Environ. Sci. Technol. 49, 8712–8720.

Moreno-Ríos, A.L., Tejeda-Benítez, L.P., Bustillo-Lecompte, C.F., 2022. Sources, characteristics, toxicity, and control of ultrafine particles: An overview. Geosci. Front. 13, 101147.

Naneos. Partector 2 - the world's smallest multimetric nanoparticle detector. Retrieved April 28, 2022. https://www.naneos.ch/partector2.html.

Noseworthy, P.A., Attia, Z.I., Brewer, L.C., Hayes, S.N., Yao, X., Kapa, S., Friedman, P.A., Lopez-Jimenez, F., 2020. Assessing and Mitigating Bias in Medical Artificial Intelligence. Circ. Arrhythm. Electrophysiol. https://doi.org/10.1161/CIRCEP.119.007988.

Ohlwein, S., Kappeler, R., Kutlar Joss, M., Künzli, N., Hoffmann, B., 2019. Health effects of ultrafine particles: a systematic literature review update of epidemiological evidence. Int J Public Health 64, 547–559.

Pelletier, C., Valero, S., Inglada, J., Champion, N., Marais Sicre, C., Dedieu, G., 2017. Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series. Remote Sens. (Basel) 9, 173.

Presto, A.A., Saha, P.K., Robinson, A.L., 2021. Past, present, and future of ultrafine particle exposures in North America. Atmos. Environ.: X 10, 100109.

Qi, M., Dixit, K., Marshall, J.D., Zhang, W., Hankey, S., 2022. National Land Use Regression Model for NO2 Using Street View Imagery and Satellite Observations. Environ. Sci. Technol. 56, 13499–13509.

Ribeiro, M. T., Singh, S., Guestrin, C., 2016. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. Preprint at https://doi.org/10.48550/arXiv.1602.04938.

Ripley, S., Minet, L., Zalzal, J., Godri Pollitt, K., Gao, D., Lakey, P.S.J., Shiraiwa, M., Maher, B.A., Hatzopoulou, M., Weichenthal, S., 2022. Predicting Spatial Variations in Multiple Measures of PM2.5 Oxidative Potential and Magnetite Nanoparticles in Toronto and Montreal, Canada. Environ. Sci. Technol. 56, 7256–7265.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Saha, P.K., Li, H.Z., Apte, J.S., Robinson, A.L., Presto, A.A., 2019. Urban Ultrafine Particle Exposure Assessment with Land-Use Regression: Influence of Sampling Strategy. Environ. Sci. Technol. 53, 7326–7336.

Salmon, M., Anderson, B., 2016. riem: Accesses Weather Data from the Iowa Environment Mesonet.

Santos, L.A., Ferreira, K.R., Camara, G., Picoli, M.C.A., Simoes, R.E., 2021. Quality control and class noise reduction of satellite image time series. ISPRS J. Photogramm. Remote Sens. 177, 75–88.

Shang, Y., Chen, R., Bai, R., Tu, J., Tian, L., 2021. Quantification of long-term accumulation of inhaled ultrafine particles via human olfactory-brain pathway due to environmental emissions – a pilot study. NanoImpact 22, 100322.

Sorek-Hamer, M., Von Pohle, M., Sahasrabhojanee, A., Akbari Asanjan, A., Deardorff, E., Suel, E., Lingenfelter, V., Das, K., Oza, N.C., Ezzati, M., Brauer, M., 2022. A Deep Learning Approach for Meter-Scale Air Quality Estimation in Urban Environments Using Very High-Spatial-Resolution Satellite Imagery. Atmos. 13, 696.

Testo. DiSCmini Handheld Nanoparticle Counter. Retrieved on 1 October, 2022. https://www.testo.com/en-US/testo-discmini/p/133.

U.S. EPA, 2019. Integrated Science Assessment (ISA) for Particulate Matter. (Final Report, Dec 2019). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-19/188 (2019).

van Nunen, E., Vermeulen, R., Tsai, M.-Y., Probst-Hensch, N., Ineichen, A., Davey, M., Imboden, M., Ducret-Stich, R., Naccarati, A., Raffaele, D., Ranzi, A., Ivaldi, C., Galassi, C., Nieuwenhuijsen, M., Curto, A., Donaire-Gonzalez, D., Cirach, M., Chatzi, L., Kampouri, M., Vlaanderen, J., Meliefste, K., Buijtenhuijs, D., Brunekreef, B., Morley, D., Vineis, P., Gulliver, J., Hoek, G., 2017. Land Use Regression Models for Ultrafine Particles in Six European Areas. Environ. Sci. Technol. 51, 3336–3345.

Weichenthal, S., Farrell, W., Goldberg, M., Joseph, L., Hatzopoulou, M., 2014. 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., 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.

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., Hatzopoulou, M., Brauer, M., 2019. A picture tells a thousand… exposures: Opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. Environ. Int. 122, 3–10.

Weichenthal, S., Olaniyan, T., Christidis, T., Lavigne, E., Hatzopoulou, M., Van Ryswyk, K., Tjepkema, M., Burnett, R., 2020. Within-city Spatial Variations in Ambient Ultrafine Particle Concentrations and Incident Brain Tumors in Adults. Epidemiology 31, 177–183.

Weichenthal, S., Ripley, S., Korsiak, J., 2022. Fine Particulate Air Pollution and the 'No-Multiple-Versions-of-Treatment' Assumption: Does Particle Composition Matter for Causal Inference? Am. J. Epidemiol. kwac191 https://doi.org/10.1093/aje/kwac191.

Wood, S.N., 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62, 1025–1036.

Wood, S., 2020. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.

Xu, J., Zhang, M., Ganji, A., Mallinen, K., Wang, A., Lloyd, M., Venuta, A., Simon, L., Kang, J., Gong, J., Zamel, Y., Weichenthal, S., Hatzopoulou, M., 2022. Prediction of Short-Term Ultrafine Particle Exposures Using Real-Time Street-Level Images Paired with Air Quality Measurements. Environ. Sci. Technol. 56, 12886–12897.

Zalzal, J., Alameddine, I., El Khoury, C., Minet, L., Shekarrizfard, M., Weichenthal, S., Hatzopoulou, M., 2019. Assessing the transferability of landuse regression models for ultrafine particles across two Canadian cities. Sci. Total Environ. 662, 722–734.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 15, e1002683.