Environmental Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Environmental Research

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

Predicting intraurban $PM_{2.5}$ concentrations using enhanced machine learning approaches and incorporating human activity patterns

Mehdi Ashayeri [a,1], Narjes Abbasabadi [a,2], Mohammad Heidarinejad [b], Brent Stephens [b,*]

[a] College of Architecture, Illinois Institute of Technology, Chicago, IL, USA
[b] Department of Civil, Architectural, and Environmental Engineering, Illinois Institute of Technology, Chicago, IL, USA

ARTICLE INFO

Keywords:
Outdoor air quality
Artificial intelligence
Human activity
Air pollution modeling
Statistical prediction model

ABSTRACT

Urban areas contribute substantially to human exposure to ambient air pollution. Numerous statistical prediction models have been used to estimate ambient concentrations of fine particulate matter ($PM_{2.5}$) and other pollutants in urban environments, with some incorporating machine learning (ML) algorithms to improve predictive power. However, many ML approaches for predicting ambient pollutant concentrations to date have used principal component analysis (PCA) with traditional regression algorithms to explore linear correlations between variables and to reduce the dimensionality of the data. Moreover, while most urban air quality prediction models have traditionally incorporated explanatory variables such as meteorological, land use, transportation/mobility, and/or co-pollutant factors, recent research has shown that local emissions from building infrastructure may also be useful factors to consider in estimating urban pollutant concentrations. Here we propose an enhanced ML approach for predicting urban ambient $PM_{2.5}$ concentrations that hybridizes cascade and PCA methods to reduce the dimensionality of the data-space and explore nonlinear effects between variables. We test the approach using different durations of time series air quality datasets of hourly $PM_{2.5}$ concentrations from three air quality monitoring sites in different urban neighborhoods in Chicago, IL to explore the influence of dynamic human-related factors, including mobility (i.e., traffic) and building occupancy patterns, on model performance. We test 9 state-of-the-art ML algorithms to find the most effective algorithm for modeling intraurban $PM_{2.5}$ variations and we explore the relative importance of all sets of factors on intraurban air quality model performance. Results demonstrate that Gaussian-kernel support vector regression (SVR) was the most effective ML algorithm tested, improving accuracy by 118% compared to a traditional multiple linear regression (MLR) approach. Incorporating the enhanced approach with SVR algorithm increased model performance up to 18.4% for year-long and 98.7% for month-long hourly datasets, respectively. Incorporating assumptions for human occupancy patterns in dominant building typologies resulted in improvements in model performance by between 4% and 37%. Combined, these innovations can be used to improve the performance and accuracy of urban air quality prediction models compared to conventional approaches.

1. Introduction

Exposure to ambient air pollution contributes substantially to the global burden of death and disease (WHO | Air pollution; Cohen et al., 2017). In particular, exposure to $PM_{2.5}$, or the mass concentration of particles smaller than 2.5 μm in aerodynamic diameter, is a major contributor to the adverse health effects associated with exposure ambient air pollution, leading to an estimated 8.9 million deaths globally in 2015 (Burnett et al., 2018). Urban areas are important contributors to human exposure to outdoor air pollution, as concentrations of many airborne pollutants (including $PM_{2.5}$) are present at elevated levels in cities worldwide (Mage et al., 1996; Mayer, 1999), and over half of the global human population lives in urban areas (Overview. World Bank n.d., 2019). Monitoring and predicting urban air pollutant concentrations are key components in any urban air quality management plan (Gulia et al., 2015).

* Corresponding author. Department of Civil, Architectural, and Environmental Engineering, Illinois Institute of Technology, Alumni Memorial Hall Room 228E, 3201 South Dearborn Street Chicago, IL, 60616, USA.
E-mail address: brent@iit.edu (B. Stephens).
[1] Present Address: College of Health and Human Sciences, School of Architecture, Southern Illinois University Carbondale, IL USA.
[2] Present Address: College of Architecture, Planning and Public Affairs, University of Texas at Arlington, TX USA.

https://doi.org/10.1016/j.envres.2020.110423
Received 25 March 2020; Received in revised form 14 August 2020; Accepted 31 October 2020
Available online 4 November 2020
0013-9351/

Governments and researchers around the world monitor ambient concentrations of $PM_{2.5}$ and other airborne pollutants for regulatory and monitoring purposes (e.g., the U.S. Environmental Protection Agency (EPA) (ntegrated Scienc, 2009; Solomon and Sioutas, 2008; Solomon et al., 2014)). However, ambient monitoring networks require significant investments in infrastructure (Yuan et al., 2012) and still leave spatiotemporal gaps in air pollution data – including in both rural and intraurban locations – that need to be filled for accurate human exposure assessments. Approaches to fill these gaps commonly include satellite/remote sensing (van Donkelaar et al., 2010, 2014; Ma et al., 2016; Lin et al., 2018), atmospheric physics/chemistry simulations (Zhang et al., 2018; Fann et al., 2012, 2018; Liu et al., 2010; Chemel et al., 2010), networks of low-cost monitors (Gao et al., 2015; Moltchanov et al., 2015), and statistical models (Singh et al., 2012, 2013; Karppinen et al., 2000a, 2000b; Elbir, 2003), or combinations of one or more of these methods (van Donkelaar et al., 2016).

In particular, numerous statistical prediction models with a variety of approaches and explanatory variables have been used to estimate ambient concentrations of $PM_{2.5}$ and other pollutants, including within urban areas. Statistical prediction models for urban air quality commonly have two major applications. The first is to explore predictors of time-series data from actual monitoring stations (Zhai and Chen, 2018; Karimian et al., 2019; Liu et al., 2019; Biancofiore et al., 2017; Russo et al., 2015; He et al., 2014; Yang et al., 2018); the second is to use predictive models such as land use regression (LUR) to estimate air quality in locations for which there is no air quality information available (Dons et al., 2013; Ryan and LeMasters, 2007; Hoek et al., 2008; Jerrett et al., 2005; Eeftens et al., 2012). The focus of this work is on the first application.

Statistical prediction models for urban air quality have generally considered meteorological factors as exploratory variables (e.g., (Chen et al., 2017; Hou and Wu, 2016; Yousefian et al., 2020)), in addition to other factors such as auxiliary pollutants (i.e., co-pollutants occurring at the same time and space) (Biancofiore et al., 2017; Russo et al., 2015), mobility-related factors (Nyhan et al., 2019; Cyrys et al., 2003; Fan et al., 2009), and demographic factors (Burke et al., 2001; Chowdhury et al., 2018; Isukapalli et al., 2013). A limited number of studies have also included factors that attempt to characterize local emissions from building infrastructure as predictor variables to estimate local $PM_{2.5}$ concentrations. For example, the number of bedrooms, fireplaces, and kitchens in residences (Masiol et al., 2018) or the density of oil-burning boilers (Clougherty et al., 2013), both of which could conceivably account for local combustion emissions from buildings to local ambient air via exfiltration and/or exhaust ventilation. However, to date, building-related factors have typically been considered as static spatial values rather than temporal values, although they likely vary with human activity patterns in and around buildings. One of the aims of this study is to test the effectiveness of incorporating dynamic building-related factors such as human activity patterns for improving urban air quality prediction models.

Dynamic building-related factors that account for indoor human activity patterns and their impacts on local ambient air quality may be important to consider since people spend nearly 90% of their time indoors (Klepeis et al., 2001) and they generate numerous indoor pollutants through both combustion and non-combustion sources that could conceivably contribute to local intraurban ambient $PM_{2.5}$ concentrations. For example, one recent study demonstrated that indoor emissions of volatile organic compounds (VOCs) from consumer products contribute substantially to local outdoor air quality, as exfiltrated/exhausted VOCs migrate from buildings to outdoors and act as precursors to $PM_{2.5}$ formation (McDonald et al., 2018). Other studies have similarly demonstrated that pollutants such as semi-volatile organic compounds (SVOCs) (e.g., polybrominated diphenyl ethers, or PBDEs) migrate from building ventilation systems (Björklund et al., 2012) and contribute to ambient $PM_{2.5}$ concentrations (Li et al., 2015). Given that one recent study estimated that 46% of the human-origin $PM_{2.5}$ in U.S. cities remains unaccounted for even after considering traffic, industry, domestic fuel combustion, and natural sources (Karagulian et al., 2015), exploring the role of dynamic building-related factors such as indoor human activity patterns may provide further insights into the contributors to urban air pollution.

Additionally, until recently, statistical prediction models for urban air quality have primarily involved linear models and large-dimensional data. More recently, various scientific communities have demonstrated that artificial intelligence (AI) approaches, and more specifically machine learning (ML) approaches, can help improve model accuracy and precision in a wide range of applications (e.g., (Wang et al., 2019; Philibert et al., 2013)). However, in large-dimensional data, correlations between explanatory variables frequently occur. Further, large-dimensional data reduces the predictive model-performance for ML approaches, particularly those of non-linear algorithms (e.g., Artificial Neural Networks (ANNs) and Support Vector Regressions (SVRs)), as these algorithms already demand high-computational capacity to be executed. Thus, finding meaningful low-dimensional data-structures that are embedded in their high-dimensional observations can help reduce correlation effects and increase the accuracy and performance of these models.

Traditionally, principal component analysis (PCA) is used to reduce the dimensionality of the data (Björklund et al., 2012; Li et al., 2015) and to embed the data in a low-dimensional input space that best preserves their variance as captured in the original high-dimensional data. PCA explores correlations between variables and replaces them with newly created variables, which minimizes correlation effects. However, nonlinear patterns are invisible to PCA (Tenenbaum, 2000). Incorporating ML algorithms into PCA and cascading are both approaches to reduce the dimensionality of the data-space, particularly for ML-based algorithms (Tenenbaum, 2000). To our knowledge, the cascading approach has not been used for detecting correlations between co-variates yet. Cascade approaches are primarily implemented for two purposes: (i) performance (low-execution time) and (ii) accuracy (precision). To expedite performance, the cascade approach partitions the data-space into smaller portions based on either reducing data-dimensionality (Kramer, 2015) or reducing the number of observations (Graf et al, ; You et al., 2015; Lujan, 2012). To improve accuracy, the data-space is partitioned into smaller sub-categories by which multiple response variables are predicted in parallel and the obtained predicted values for each sub-category are used as inputs for the core algorithm to layout the overall response (Garg and Gupta, 2008). In addition, the latter approach can also be applied to predict and add new variables into the n-dimensional data-space in which the n+1 dimensional data-space predicts the overall response. This approach was proposed by (Franceschini et al., 2018) in which an ANN algorithm was used with a newly created binary variable. However, to the best of our knowledge, a method that integrates the power of ML, PCA, and cascade methods together by which the model performance and accuracy are both improved, also considering correlation issues, has not been investigated.

It is believed that non-rule-based ML algorithms such as ANNs, SVRs, and k-nearest neighbors (k-NNs) are 'black-box' (Olden et al., 2004) approaches, such that explaining their results is challenging (Greenwell, 2017). There exist, however, several methods to 'illuminate' (Beck, 2018) the 'black-box', in which model-agnostic methods are used to implement such a task, regardless of what algorithm is applied. The Partial Dependence Plot (PDP) approach introduced by (Friedman, 2001) is one of those approaches that provides an opportunity to evaluate interactions between predictor variables. PDPs display how the average effect of a predictor changes when a predictor is changed over the range of variation by considering other predictors at their constant values. PDP is applied on the user's fitted data, regardless of what algorithm fitted that data. The Individual Conditional Expectation (ICE) (Goldstein et al., 2013) is a new model-agnostic approach that is based upon the PDP method but that displays the impacts of each instance of a

predictor separately.

With this background on urban air quality and statistical prediction models, the objectives of this research are to (1) bring the power of AI into urban air quality models, with a focus on hybridizing cascade ML algorithms and PCA approaches for predicting intraurban $PM_{2.5}$ concentrations, and (2) apply these approaches to explore the influence of dynamic human-related factors, including mobility (i.e., traffic) and building occupancy patterns, on model performance. The workflow entails four main steps: (i) testing multiple state-of-the-art ML algorithms to find the most effective algorithm for modeling intraurban $PM_{2.5}$ variations; (ii) incorporating PCA and cascade ML approaches together into a single model to improve accuracy of predictions; (iii) incorporating dynamic building-related factors that could conceivably account for emissions from buildings and occupants as indoor sources to the local ambient environment (i.e., improving explanatory variables); and (iv) explaining the relative importance of all sets of factors on intraurban air quality model performance. We hypothesize that these innovations will improve the performance and accuracy of urban air quality prediction models compared to conventional approaches. We test this approach using time series air quality datasets from three air quality monitoring sites in different urban neighborhoods in Chicago, IL.

2. Materials and methods

2.1. Data sources, sampling site, and dataset processing

Three air pollution monitoring locations in Chicago, IL, were selected in this study: one located in Logan Square, one in the Loop (Downtown), and one in the Ashburn neighborhood. The dominant morphology of these neighborhoods is residential, commercial, and residential-industrial properties, respectively (Fig. 1). This study uses two different data sources due to the lack of reliable high resolution outdoor pollutant data in all locations. First, air quality data for the Loop location were obtained from a weeklong study of several ambient pollutants in the outdoor ventilation air intakes along the height of a tall building conducted in June 2017, including size-resolved particulate matter, which were used to estimate $PM_{2.5}$ mass concentrations (Azimi et al., 2018). Only the 2nd floor (closest to ground-level) data were used from this location. Second, a full year of hourly air quality data in 2017 for the Logan Square and Ashburn neighborhoods were obtained from the U.S. EPA's AirData database (AirData website File Down, 2019).

We selected four categories of explanatory variables for predicting intraurban $PM_{2.5}$ concentrations at each site, including (i) meteorological, (ii) human activity patterns, (iii) daytime/calendar, and (iv) auxiliary air pollutants including ozone (O_3) and nitrogen oxides (NO_x), each of which is described below. Table 1 lists all variables that were used in this research. Meteorological variables comprise local wind speed, wind gust, temperature, relative humidity, atmospheric pressure, and solar radiation obtained from local Personal Weather Stations (PWS) on the Weather Underground portal (Local Weather Forecast and N, 2019). The closest PWS to each air pollution monitoring site for each neighborhood was selected (including KILCHICA114 for Logan Square and KILOAKLA6 for Ashburn). Human activity pattern factors comprise those that influence mobility and building occupancy. Wind "gust" was used to incorporate a measure of the transient (instantaneous) behavior of the wind (i.e., duration $\leq$ 20 s and speed $\geq$ 8.2 m/s), in addition to its continuous speed (i.e., "wind speed"), in part because of recent literature demonstrating that wind gust is associated with an increase in $PM_{2.5}$ concentrations a decrease in visibility (Kelley et al., 2020). Wind direction was not included because of considerable missing values for this factor in the databases that we retrieved (and thus remains a limitation of this work). Mobility factors include hourly-averaged traffic congestion and speed of traffic obtained from the *Illinois Department of Transportation* (Traffic Count Database Sy, 2019) and *Chicago Traffic*

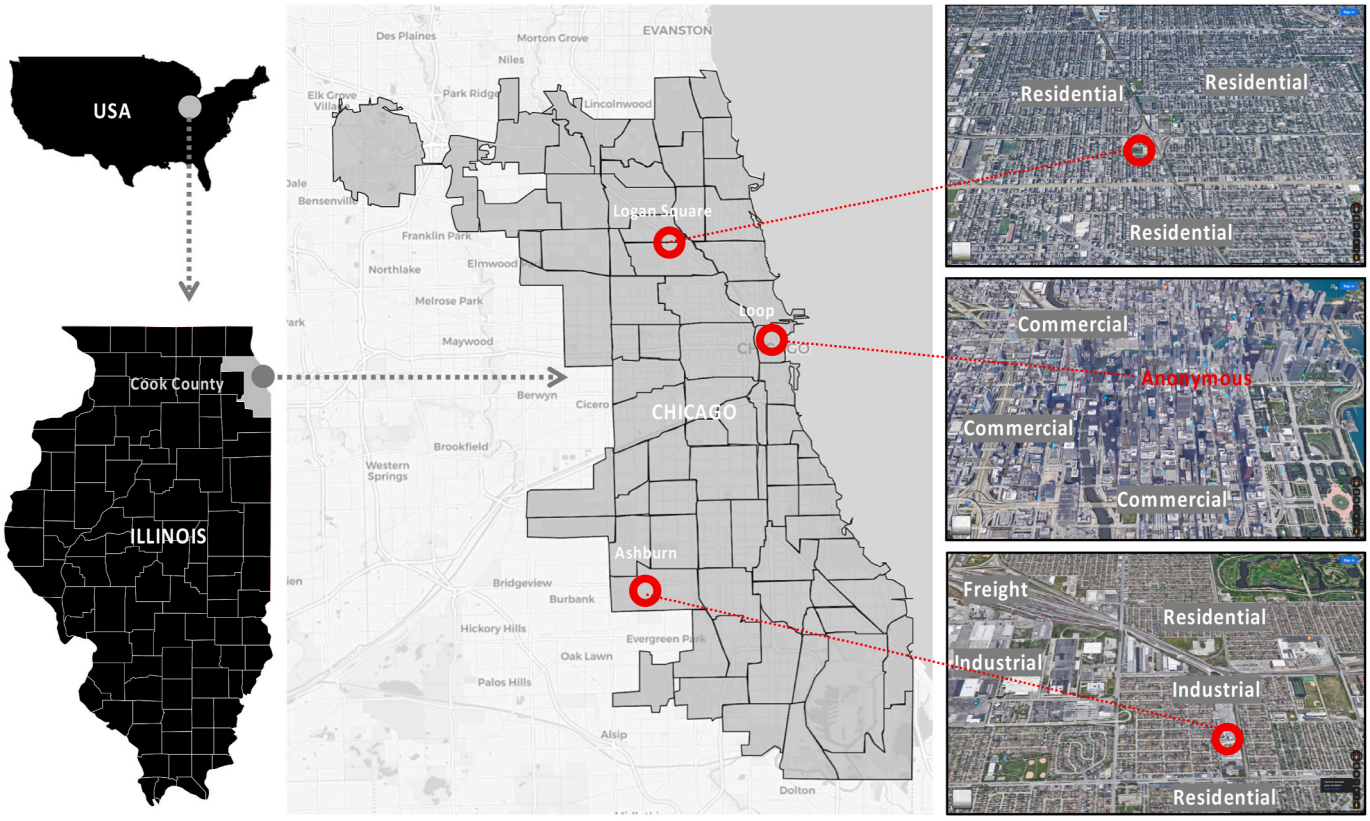

Fig. 1. Maps of dominant urban morphology across three neighborhoods in Chicago, including Logan Square, Loop, and Ashburn. RStudio was used for mapping out geo data. Image source: Google Street View (GSV), 2019. Red circles show the location of $PM_{2.5}$ monitoring locations. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Factors used for predicting intraurban $PM_{2.5}$ concentrations.

Category	Variable	Abbreviation	Data type	Unit
X1: Meteorological	Wind speed	Ws	Continuous	m s^{-1}
	Wind gust	Wg	Continuous	–
	Temperature	T	Continuous	°C
	Relative humidity	RH	Continuous	–
	Atmospheric pressure	P	Continuous	Pa
	Solar radiation	Sol	Continuous	W m^{-2}
X2: Human activity patterns				
X2-A: Mobility	Hourly traffic speed profile	Tspeed	Continuous	km h^{-1}
	Hourly traffic count profile	Tcount	Continuous	–
X2-B: Building occupancy	Hourly residential occupancy profile	Res	Discrete	–
	Hourly commercial occupancy profile	Comm	Discrete	–
X3: Daytime/ Calendar	Hour of Day	Hour	Discrete	–
	Weekdays	Weekday	Discrete	–
	Day of moth	Day	Discrete	–
	Month of Year	Month	Discrete	–
X4: Auxiliary pollutants	Nitrogen oxides	NO_X	Continuous	ppm
	Ozone	O_3	Continuous	ppm
Y: Response variable	$PM_{2.5}$ concentrations	$PM_{2.5}$	Continuous	μg m^{-3}

Tracker-Congestion Estimates by Segments-2011-2018 (Chicago Traffic Tracker), respectively.

Prototypical hourly occupancy patterns for commercial and residential apartment buildings were extracted from the U.S. Department of Energy (DOE) Commercial Prototype Building Models (Commercial Prototype Buil, 2019; Deru et al., 2011). Prototypical hourly occupancy patterns for single- and low-rise multi-family residential buildings were extracted from the DOE Residential Prototype Building Models (Residential Prototype Bui, 2019). These datasets have been used in many previous studies to support energy consumption analysis across building types (Sohn and Dunn, 2019) and archetypes used in urban scale energy modeling (e.g., (Heidarinejad et al., 2017; Heiple and Sailor, 2008)). These patterns are considered nationally representative and do not necessarily represent actual data from our specific sites; however, they represent a reasonable first approach to incorporate dynamic building occupancy factors in urban air quality modeling. The occupancy profiles used herein comprise discrete numbers between 0 and 1, while traffic profiles are based on continuous values.

Similar to spatiotemporal models such as land use regression (LUR) models, we defined buffer zones for building and traffic-related factors separately to investigate their local impacts on outdoor $PM_{2.5}$ concentrations across those zones. Buffer radii of 1000 m and 100 m were defined for local building occupancy profiles and traffic profiles, respectively. The selected buffer radii for occupancy profiles were informed by a recently developed LUR model for estimating hourly $PM_{2.5}$ concentrations in Monroe County, New York, in which the radii were calculated ranging from 50 to about 2000 m for multiple (static) building-related factors (Masiol et al., 2018). We assumed a 1000 m radius, which is in the middle of that range. The predominant building typology across the defined buffer zone was then captured from the Chicago Building Footprint dataset (Building Footprints (curr, 2018) in order to define the predominant occupancy profile for that zone. Meanwhile, traffic counts and vehicle speed data were averaged based on hourly traffic congestion data for all street nodes within the 100 m buffer (Masiol et al., 2018; Weichenthal et al., 2016) to be used as hourly local traffic profiles for the zone.

Daytime/calendar variables comprised hour-of-day (24 h), day-of-week (seven days), month-of-year, and day-of-month. Each was latently available at all-time series. Holidays (e.g., New Year's and the Fourth of July) were excluded from datasets because of anomalous firework-related pollutant sources. For weekly and monthly analysis, day-of-month and month-of-year were excluded, as there are no repetitions of these factors during the time duration. For seasonal analysis, January, February, and December were assumed as winter; March, April, and May as spring; June, July, and August as summer; and September, October, and November as fall seasons (Zhang et al., 2015; Xu et al., 2017; Di et al., 2016). The *lubridate* package from CRAN library (Grolemund and Wickham, 2011) in R software was applied to generate values for this category of predictors.

We also used ambient concentrations of two auxiliary pollutants (i.e., co-pollutants) – O_3 and NO_x – in some models to investigate two potentially important mechanisms: (a) the impacts of these pollutants parallel to other explanatory variables, and (b) the relationship between these pollutants and human activity patterns. NO_x and O_3 were chosen as auxiliary explanatory variables because (1) they are both known precursor gaseous components for the formation of $PM_{2.5}$ (Anenberg et al., 2012) (albeit not the only precursor components) and (2) they are the only co-pollutants for which we had access to data from local regulatory monitoring networks. These two co-pollutants were tested specifically on the July and December datasets in detail. Additionally, regarding relationships between these pollutants and human activity patterns, a recent study indicated that a portion of methane (CH_4), which is a leading ozone precursor (von Schneidemesser et al., 2015), in cities is emitted from intra-building natural gas distribution and end-use facilities (Fischer et al., 2018). These emissions occur more from residential buildings than commercial buildings (Saint-Vincent and Pekney, 2019). Another recent study demonstrate that methane leakage in buildings in the U.S. cities is likely twice as what has been previously assumed (Perkins, 2019). Therefore, we hypothesize that inclusion of these two co-pollutants could conceivably improve predictions of local ambient $PM_{2.5}$ concentrations, both directly and indirectly.

2.2. Testing various ML algorithms

Using these datasets, we first tested 9 state-of-the-art ML algorithms to find the most effective algorithm for predicting intraurban $PM_{2.5}$ concentrations using *regular* models that do not reduce correlation impacts between variables (i.e., prior to including *enhanced* approaches, which is explored later). The logic for selecting these 9 ML algorithms is as follows.

ANN has received significant attention in the literature to be applied for developing predictive models. We tested three different ANN algorithms: Multi-layer perception (MLP) with a single hidden layer optimized by the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Venables and Ripley, 2002); MLP with multiple hidden layers optimized by resilient backpropagation algorithm (Fritsch et al., 2019); and radial-basis (Gaussian) kernel (RBF) ANN. The support vector machine (SVM) with polynomial and Gaussian kernels (Meyer et al., 1071), k-nearest neighbor (kNN) with Gaussian kernel (Grolemund and Wickham, 2011), and rule-based techniques including regression tree (CART) (Terry and Atkinson, 2018), random forest (RF) (von Schneidemesser et al., 2015), and gradient boosting machine (GBM) (Friedman, 2001; GreenwellBradley et al., 2019) were also implemented. A brief definition and differences between these algorithms can be found in (Abbasabadi and Ashayeri, 2019). We used grid search optimization method to tune the hyper-parameters for each of these ML algorithms in order to avoid a pre-deterministic parameter-selection approach. To this end, we defined a wide range of variance for each of parameters, as described later.

SVM, ANN, and kNN are known as non-linear and black-box approaches because of their sophisticated intra-processing tasks and resulted weights, while the rule-based techniques provide the simulation process due to their simple if/else-based functions. RF and GBM are based on the ensemble of many trees, which enable them to provide less biased results than a single tree. The multiple linear regression (MLR)

algorithm was also used as a traditional predictive approach to compare the effectiveness of this algorithm with those of ML-based approaches.

To first select the most effective algorithm out of this pool of 9 algorithms, we tested each algorithm for predicting hourly $PM_{2.5}$ concentrations across multiple time horizons (including month-long, season-long, and year-long hourly data) in our two larger datasets: Logan Square and Ashburn. This yielded a total of 34 datasets for spatiotemporal application and comparision (i.e., 24 month-long, 8 seasonal, and two yearlong datasets for each of the two locations, Ashburn and Logan Square). The model structure utilized in this section included all possible explanatory factors shown in Table 1, with the exception of auxiliary pollutants (i.e., all variables in the meteorological, daytime/calendar, and human activity patterns categories). This model structure is later referred to as "model-7" in Table 3. We averaged all R^2 results obtained from these approaches applied across all time horizons. We also compared model performance between MLR and the best ML algorithm to highlight the potential improvements achievable by ML approaches.

2.3. Development of an enhanced predictive approach with ML algorithms

Next, we developed a enhanced approach for predicting hourly $PM_{2.5}$ concentrations across neighborhoods with ML algorithms. In the enhanced approach, the variance inflation factor (VIF) test (Allison, 1999) was implemented to detect multicollinearity between each pair of explanatory variables. Multi-collinearity is an issue in data simulation by which one explanatory variable is predicted linearly with a significant degree of accuracy by another explanatory variable. The obtained results under multi-collinearity can be biased in spite of likely improvements in the model. The VIF test is implemented before simulating predictive models either by statistical or ML-based regression algorithms (Karimian et al., 2019). VIF lesser or equal to 5 (Masiol et al., 2018) was used to exclude collinear variables prior to subsampling datasets.

In addition to the VIF test, a quasi PCA approach with ML algorithms was developed to create a new variable instead of a pair of correlated variables. This approach was applied when the maximum coefficient of correlation (R) between the same sub-category of predictors and between the different categories of predictors exceeded 0.6 and 0.95, respectively, with a p-value lower than 0.05 (Dons et al., 2013). This approach was used to replace correlated variables with a new variable rather than omitting them. In doing so, a pair of correlated variables were placed as the predictor and response with shuffling roles; the option with the highest value for R was selected to define the ultimate combination for the selected sub-category. This enables keeping the influence of all predictors in the model while removing correlation effects. Since the datasets used herein are time series, the daytime/calendar predictors were added in each sub-category. An automation code was scripted in R software to implement such a process in a single execution.

Fig. 2 illustrates the workflow of the enhanced model development for predicting intraurban $PM_{2.5}$ concentrations. In this figure, X_{pre} denotes non-processed predictors; X_{prep} denotes newly created variables; Final Dataset denotes the final structure of data adding newly created variables and excluding correlated and collinear variables; Y denotes the final output (i.e., intraurban $PM_{2.5}$ concentrations); C-variable denotes the new variables that are created instead of those variables that are correlated for each category; D_1 to D_n denotes the final datasets for each category that include combinations of X and C variables that remove collinearities and correlations in all sub-categories; and D is the final dataset merging D_1 to D_n that removes collinearities and correlations between different categories of both X and C variables.

For comparing the performance of enhanced vs. regular ML approaches, we used multiple time horizons of hourly datasets, including month-long, season-long, and year-long datasets separately for Ashburn and Logan Square neighborhoods (longer datasets were not available in the Loop location), both with and without auxiliary pollutants for Ashburn (auxiliary variables were only available for Ashburn datasets). Table 2 shows the pseudo-code used for executing the proposed workflow.

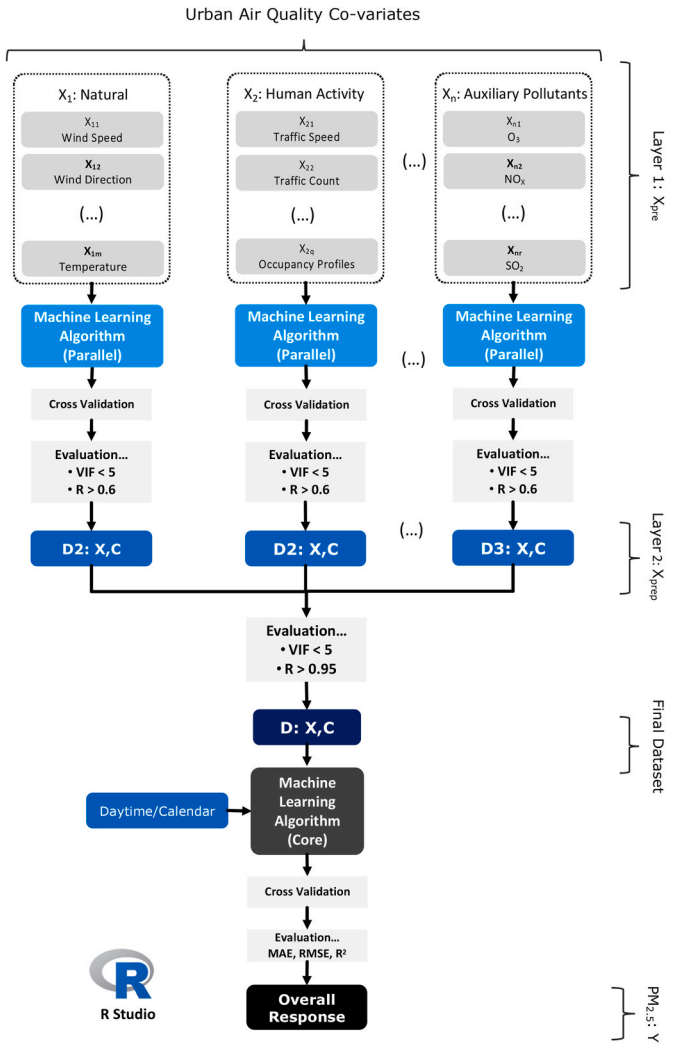

Fig. 2. The workflow for developing enhanced models for predicting intra-urban $PM_{2.5}$ concentrations.

2.4. Model sub-sampling, validation, and evaluation

In order to avoid biasing results and over-fitting issues in the application of the ML techniques, a 10-repeated 5-fold cross-validation (10 × 5CV) method was applied (Li et al., 2006). The datasets for all ML approaches were divided into two parts, training sets (80%) and test sets (20%), using a random subsampling approach. The process was repeated until the histogram of both training and test sets had approximately the same ratio per bins, in which the number of bins was set to be selected automatically. The *hist()* function was used in RStudio to implement this process. The VIF test and the subsampling approach was used for all models.

We applied the mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) as the most common-use metrics to compare the effectiveness of different models for urban air quality prediction (Venables and Ripley, 2002). The MAE shows the degree of difference between the predicted and the actual values. The RMSE is a relative error metric that focuses on the impact of extreme

Table 2
Pseudo-code used for automating the cascading ML algorithm and replacing correlated variables with newly created variable.

for (i in vector X (i = 1:n, n = number of categories without daytime/calendar category) {. for (j in vector X (j = 1:m, m = number variables for each category)
Step A: intra-category
do train [X_j] over [X_{j+1} & $A_{Daytime/calendar}$] factors by a 10 repeated 5-fold cross validation,
do test [X_j] over [X_{j+1} & $A_{Daytime/calendar}$],
do paste (R) for both train and test sets.
if $R_{test} \geq 0.60$ and p-value<0.05 (in the same category of variables)
do replace both [X_{ij}] and [$X_{i(j+1)}$ & $X_{Daytime/calendar}$] togather $_{by}$ predicted value for test set,
do name the trained new variable as $C_{s\text{-}tr\text{-}p}$ (p = 1:p, p ∈ *N*, p ≥ 1)
do replace both [X_{ij}] and [$X_{i(j+1)}$ & $X_{Daytime/calendar}$] togather with trained value for train set,
do name the predicted new variable as $C_{s\text{-}ts\text{-}p}$ (p = 1:k, p ∈ *N*, p ≥ 1)
do merge $C_{s\text{-}tr\text{-}p}$ and $C_{s\text{-}ts\text{-}p}$
do name it C_{sp}
Step B: inter-category
do train [X_{ij}] over [$X_{i(j+1)}$ & $A_{Daytime/calendar}$] factors by a 10 repeated 5-fold cross validation,
do test [X_{ij}] over [$X_{i(j+1)}$ & $A_{Daytime/calendar}$],
do paste (R) for both train and test sets.
if R ≥ 0.95 and p-value<0.05 (between different category of variables),
do replace both [X_{ij}] and [$X_{(i+1)j}$ & $X_{Daytime/calendar}$] togather by predicted value for test set,
do name the trained new variable as $C_{d\text{-}tr\text{-}p}$ (q = 1:q, q ∈ *N*, q ≥ 1)
do replace both [X_{ij}] plus [$X_{(i+1)j}$ & $X_{Daytime/calendar}$] togather with trained value for train set,
do name the predicted new variable as $C_{d\text{-}ts\text{-}p}$ (q = 1:k, q ∈ *N*, q ≥ 1)
do merge $C_{d\text{-}tr\text{-}q}$ and $C_{d\text{-}ts\text{-}q}$
do name it C_{dq}
do merge C_{sp} and C_{dq} with remaining factors and daytime/calendar variables as final dataset. do name final dataset as D
do train D by a 10 repeated 5-fold cross validation,
do test D,
do paste (RMSE, MAE, R^2, and adjusted R^2 all in a single matrix) for both train and test sets
stop
}

Table 3
Developed models for inter-model sensitivity analysis based on the regular ML approach.

Model	Explanatory variables without auxiliary pollutants	Explanatory variables with auxiliary pollutants
0	Met	Met, Pollutant[a]
p	n/a	Pollutant
1	Hour	Hour, Pollutant
2	Hour, Met	Hour, Met, Pollutant
3	Hour, Weekday, Met (Base Model)	Hour, Weekday, Met, Pollutant (Base Model)
4	Hour, Weekday, Met, Tcount, Tspeed	Hour, Weekday, Met, Pollutant, Tcount, Tspeed
5	Hour, Weekday, Met, Comm	Hour, Weekday, Met, Pollutant, Comm
6	Hour, Weekday, Met, Tcount, Tspeed, Comm	Hour, Weekday, Met, Pollutant, Tcount, Tspeed, Comm
7	Hour, Weekday, Met, Tcount, Tspeed, Comm, Res	Hour, Weekday, Met, Pollutant, Tcount, Tspeed, Comm, Res
8	Hour, Weekday, Met, Tcount, Tspeed, Res	Hour, Weekday, Met, Pollutant, Tcount, Tspeed, Res
9	Hour, Weekday, Met, Res	Hour, Weekday, Met, Pollutant, Res

[a] Pollutant = O_3 or NO_x separately.

values and represents the performance of a linear model based on the model-fitting approach. Theoretically, RMSE provides a larger value than MAE for the same problem (Chai and Draxler, 2014). The metrics are formulated as the following Equations (1-3):

$$MAE = \frac{1}{k}\sum_{i=1}^{k}\left|y_{pred,\,i} - y_{act,\,i}\right| \quad (1)$$

$$RMSE = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\frac{\sum_{i=1}^{k}\left(y_{predt,\,i} - y_{act,\,i}\right)^2}{k}} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{k}\left(y_{pred,\,i} - y_{act,\,i}\right)^2}{\sum_{i=1}^{k}\left(y_{act,\,i} - \bar{y}_{act}\right)^2} \quad (3)$$

Where $y_{act,\,i}$ and $y_{pred,\,i}$ denote the actual and predicted $PM_{2.5}$ concentration for i^{th} observation, $\bar{y}_{act}$ denotes the mean $PM_{2.5}$ concentration, and *k* denotes the total number of observations.

2.5. Inter-model sensitivity analysis

To address the second objective of this work (i.e., exploring the impacts of including human activity patterns in urban air quality models), we conducted an inter-model sensitivity analysis in which the best performing ML algorithms from Section 2.2 were applied to various time horizons of our three air quality datasets using both regular and enhanced approaches (Section 2.3). For both approaches, we implemented a variable selection process that allowed for exploring the influence of including or excluding individual explanatory variables and/or sets of explanatory variables. We implemented this analysis for understanding the performance of models based on (i) complexity of inclusion/exclusion of factors and the correlation effects between covariates, and (ii) exploring the impacts of human activity patterns based on the dominant typology of buildings and local traffic patterns in each neighborhood.

Table 3 summarizes the developed models, including a total of 10 regular models that exclude auxiliary pollutants and 11 regular models that include auxiliary pollutants (Pollutant). Each model is based on different combinations of predictors and is defined by adding and subtracting predictor variables. A study by (Levy et al., 2010) implemented a similar approach for inter-model sensitivity analysis for predicting intraurban $PM_{2.5}$ concentrations by adding and subtracting human activity patterns (traffic profiles). For the models that excluded auxiliary pollutants, we first developed model-0, which includes only natural (Meteorological, or Met) factors. Model-1 includes only the hour-of-day (Hour) factor and Model-2 includes both meteorological and hour factors. Model-3 adds the weekday factor into model-2. We refer to model-3 as the "base model," without human activity factors, because this is the simplest model that is used for predicting hourly $PM_{2.5}$ concentrations for datasets no longer than a month horizon. In model-4, mobility factors, including traffic count and traffic speed data, are added to the base model (model-3). Model-5 adds commercial building occupancy (Comm) into the base model, while model-6 adds Comm into model-4. Model-7 includes all available variables in the model, adding residential building occupancy patterns (Res) to model-6. Model-8 subtracts Comm from the model and keeps only residential building occupancy pattern (Res) as the occupancy factor, and model-9 subtracts mobility from model-8 and keeps Res as the sole human activity pattern in the model. For the models that included auxiliary pollutants (O_3 and NO_x), we simply added co-pollutant concentrations (where available; i.e., Ashburn only) as an additional explanatory variable to each of the 10 models constructed without auxiliary pollutants, and also added an 11th model that included only co-pollutants as predictors. We added O_3 and NO_x separately to explore the significance of these pollutants in explaining local $PM_{2.5}$ concentrations in the predictive models.

Next, we developed 6 additional enhanced models that exclude auxiliary pollutants and 8 additional enhanced models that include auxiliary pollutants (Table 4). In the enhanced models without auxiliary pollutants, we created three new variables in the time series data that

Table 4
Developed models for inter-model sensitivity analysis based on the enhanced ML approach.

Model	Explanatory variables without auxiliary pollutants	Explanatory variables with auxiliary pollutants
10	Hour, Weekday, Met, C1	Hour, Weekday, Met, Pollutant, C1
11	Hour, Weekday, Met, Comm, C1	Hour, Weekday, Met, Comm, Pollutant, C1
12	Hour, Weekday, Met, Res, C1	Hour, Weekday, Met, Res, Pollutant, C1
13	Hour, Weekday, Met, C2	Hour, Weekday, Met, Pollutant, C2
14	Hour, Weekday, Met, Res, C2	Hour, Weekday, Met, Res, Pollutant, C2
15	Hour, Weekday, Met, C3	Hour, Weekday, Met, Pollutant, C3
16	n/a	Hour, Weekday, Met, C3, C4
17	n/a	Hour, Weekday, Met, NO_x, C3, C4

C1 = Tcount & Tspeed; C2 = Comm & Tcount; C3 = Res & Tspeed; Pollutant = O_3 or NO_x; C4 = (O_3 & Sol) & Tcount.

combined the variables related to human activity patterns that had the highest correlations when treated individually: C1 was created as a combination of Tcount and Tspeed (i.e., traffic variables); C2 was created as a combination of Comm and Tspeed (i.e., traffic speed and commercial building occupancy patterns); and C3 was created as a combination of Res and Tspeed (i.e., traffic and residential building occupancy patterns). Daytime/calendar and meteorological data were kept as control variables (similar to the base model), and these additional human activity pattern variables were added one by one to create additional models. In the enhanced models with auxiliary pollutants, we took a similar approach albeit with inclusion of co-pollutants for each model increment and with the addition of two additional models with one new variable (C4) that was created to remove correlations between O_3, Sol, and Tcount variables.

For the inter-model sensitivity analysis, we used the following datasets: (1) weeklong hourly datasets in June for all three neighborhoods (this was the only concurrent period for which data were avaiable at all locations) and (2) a full month of hourly data for June, July, and December each for Ashburn and Logan Square. These months were selected to be representative of summer and winter seasons, which are known to have very different contributors to ambient particulate matter concentrations (e.g., (Dons et al., 2013; Weichenthal et al., 2016; Tripathy et al., 2019; Tunno et al., 2016a)). In this analysis, we compared only among ML algorithms (i.e., we did not compare MLR vs. ML algorithms). Table 5 shows a matrix of datasets with their time horizons used for each neighborhood and the number of developed models used in the inter-model sensitivity analysis for each dataset.

Table 5
Datasets with their available variables and time horizons for each neighborhood and the number of models developed per dataset for inter-model sensitivity analysis based on the regular ML approach.

Duration	Neighborhood	Dataset with and without auxiliary pollutants	Number of models
Weeklong (in June 2017)	Loop	without	10
	Ashburn	without	10
	Logan Square	without	10
Month-long (June, July, December 2017)	Ashburn	without & with	10,11
	Logan Square	without	10

2.6. Intra-model sensitivity analysis

To better understand each of the individual model approaches from Section 2.6, we also conducted an intra-model sensitivity analysis using a set of techniques to capture the relative influential factors for predicting interurban air quality. The PDPs (Friedman, 2001) and ICE Plots (Goldstein et al., 2013) were used to handle such applications as they represent model-agnostic methods. This approach helps explain outputs (i.e., fitted response) from 'black-box' techniques better than within-model parameters and internal actions. ICE plots depict the heterogeneity of instances of a predictor over the fitted response variable and are more helpful when interactions between the predictors for which the PDP is calculated and the other predictors are not weak. By applying ICE plots and PDPs, we quantify variable importance (VI), which is based on a model-free approach. VI plots depict the relative magnitude of each predictor for the fitted response. We applied ICE plots averaged by PDP curves based on a method proposed by (Greenwell et al., 2018). This method identifies VI as the flatness of PDP curves in which the more flatter curves represent the lower relative VI for the interest predictor. In other words, the less interaction that a target predictor has with other predictors, the less influential role that variable plays in that model. The flatness of PDP curves was calculated using Equation (4):

$$i(x_1) = \begin{cases} \sqrt[2]{\dfrac{1}{k-1}\sum_{i=1}^{k}\left[\overline{f_1}(x_{1i}) - \dfrac{1}{k}\sum_{i=1}^{k}\overline{f_1}(x_{1i})\right]^2} & \text{if } x_1 \text{ is continous} \\ \left[\max_i\left(\overline{f}_1(x_{1i})\right) - \min_i\left(\overline{f}_1(x_{1i})\right)\right] \Big/ 4 & \text{if } x_1 \text{ is categorical} \end{cases} \quad (4)$$

Where, *i(x)* denotes the flatness of the PDP curve per level of the standard deviation of a target variable, $\overline{f_1}$ displays little variability, x_{1i} denotes the first observation for each level, and *k* denotes the total number of data points per level. We used the *ICEbox* package in CRAN library to plot the ICE and PDP curves, and then, manually quantified the "flatness" of PDP curves to quantify VI. This study used the ggplot2 package

Table 6
Average of R^2 for the 34 hourly datasets without auxiliary variables.

Model	Technique	Algorithm	R-Package (CRAN Library)	R^2 (test set)
a	SVR	Gaussian kernel	e1071	**0.577**
b	GBM	Gaussian distribution	Gbm	0.545
c	kNN	Gaussian kernel	KernelKnn	0.505
d	ANN	deep MLP: resilient BP with 3-hidden layers	Neuralnet	0.487
e	RF	General	randomForest	0.479
f	ANN	Gaussian kernel	Kernlab	0.443
g	SVR	Polynomial kernel (deg. = 4)	e1071	0.434
h	ANN	MLP: BFGS with single hidden layer	Nnet	0.391
i	TREE	General	Rpart	0.322
j	MLR	General	Base	0.208

Model	**Hyper Parameter per ML Algorithm**
a	$C = 2^{\wedge}$ [-15: 10], $\gamma = 2^{\wedge}$ [-15:4], ε = [0, 1, seq = 0.1]
b	Number of Trees = 5000, Interaction Depth = [2, 10, seq = 1], Shrinkage = {0.1, 0.01, 0.001}
c	K = [2, |$\sqrt{}$number of Variables|, seq = 2]
d	Neurons = {L1 = [2, |2/3 * number of Variables|, seq = 1], L2 = 3, L3 = L1/}, Decay = {0.1,0.5}
e	Number of Trees = 5000
f	$\sigma = = 2^{\wedge}$ [-8: 8], Units = [2, |2/3 * number of Variables|, seq = 1], Decay = {0.1,0.5}
g	Degree = {3,4}, C = $2^{\wedge}$ [-8, −4, −2, 0, 1, 2, 4, 8], $\gamma = 2^{\wedge}$ [-15:4]
h	Neurons = [2, |2/3 x number of Variables|, seq = 1]
i	Min Split = {5,10}, Min Depth = [2,10, seq = 1], cp = {0.01, 0.001. 0.0001}

in R library to visualize bar plots of the relative VIs (Wickham, 2009).

3. Results and discussion

3.1. The most effective ML algorithms

Table 6 (top) shows model performance (R^2) from the 9 initially tested ML algorithms (labeled as model-a through i) as well as the MLR model (model-j) applied to the 34 datasets of hourly data (i.e., 24 month-long, 8 seasonal, and two yearlong) for Logan Square and Ashburn, as described in section 2.2. The results suggest that Gaussian kernel SVR (model-a), Gaussian distribution GBM (model-b), and Gaussian-kernel kNN (model-c) were the most effective models for predicting $PM_{2.5}$ concentrations. These three Gaussian kernels based algorithms explained $PM_{2.5}$ concentrations with 177.4%, 162.0%, and 142.8% higher accuracy in terms of R^2, respectively, compared to the MLR model as the conventional regression approach. Further, the results suggest that these three models, which are based on Gaussian kernels, can provide higher accuracy than the models that are built upon polynomial and linear functions for handling urban air quality problems. The Gaussian-kernel SVM is an approach that could reasonably be a proper choice in selecting an ML technique for solving problems and can outperform ANNs if its hyper-parameters (C, γ, ε) are appropriately selected (Hassan et al., 2010; Hsu et al., 2003). Additionally, it has been reported that the performance of SVRs in predicting air pollution in urban areas outperforms neural networks (Lu and Wang, 2005). Table 6 (bottom) shows the variance of hyper-parameters per algorithm in which values were automatically tuned through the grid-search optimization approach to obtain results with no predeterministic selection approach.

Moreover, the Gaussian kNN algorithm (model-c) has shown its effectiveness for predicting urban scale data, which is superior to the Tree, RF and MLP algorithms (Abbasabadi et al., 2019). Further, the results suggest that the ANN model with a deep MLP approach (model-d), with three hidden layers, is more effective than the RF algorithm. This model, which was optimized by the Resilient Back Propagation algorithm, showed higher performance relative to the ANN model based on the Gaussian kernel algorithm (model-f) and the MLP-based ANN with BFGS optimization (model-h) (about 9.9% and 24.5% higher, respectively, based on the R^2 metric). The SVR model with the polynomial kernel (model-g) was found to be more effective than BFGS (model-h). Thus, based on these results, we kept only the kernel SVM algorithm for developing regular and enhanced predictive models moving forward.

3.2. Evaluating enhanced vs. regular approaches

Table 7 shows results of model performance (R^2) obtained from regular and enhanced SVR approaches for hourly $PM_{2.5}$ prediction applied to the month-long, season-long, and year-long datasets in both Ashburn and Logan Square, both with and without auxiliary explanatory variables, including NO_x and O_3 as co-pollutants where available (i.e., Ashburn). Model-7 in Table 3 is used for each comparison in Table 7. Several key results are illustrated.

First, across all durations of datasets in both locations, model performance for predicting hourly $PM_{2.5}$ concentrations improved with increasing model complexity, with R^2 values ranging from 0.03 to 0.40 for MLR applied across all time horizons, increasing to 0.49–0.67 for

Table 7
Model performance (R^2) obtained using MLR, regular SVR, and enahnced SVR approaches predicting hourly $PM_{2.5}$ concentrations across multiple time horizons, including month-long, season-long, and year-long data, in Ashburn and Logan Square in 2017. Model-7 in Table 3 is used for each comparison.

R^2 for hourly $PM_{2.5}$ prediction														
	Logan Square, 2017					Ashburn, 2017								
Time span	N	Without auxiliary variables				N	Without auxiliary variables				With auxiliary variables (NO_x and O_3)			
		Daytime + Met[1]+Mobility + Occupancy					Daytime + Met + Mobility + Occupancy				Daytime + Met + Mobility + Occupancy + Pollutant			
		Regression Algorithm					Regression Algorithm				Regression Algorithm			
		MLR	SVR	Enhanced SVR	Change %[2]		MLR	SVR	Enhanced SVR	Change %[2]	MLR	SVR	Enhanced SVR	Change %[2]
Month-long														
January	719	0.195	0.678	0.828	+22.14	707	0.607	0.846	0.906	+7.09	0.663	0.857	0.938	+9.45
February	667	0.275	0.725	0.791	+9.16	544	0.238	0.638	0.688	+7.84	0.312	0.692	0.718	+3.76
March	681	0.137	0.503	0.557	+10.74	642	0.130	0.492	0.532	+8.13	0.285	0.544	0.582	+6.99
April	706	0.170	0.377	0.526	+39.56	689	0.371	0.493	0.533	+8.11	0.392	0.525	0.563	+7.24
May	638	0.167	0.627	0.711	+13.40	543	0.450	0.651	0.720	+10.77	0.541	0.777	0.827	+6.44
June	715	0.276	0.407	0.622	+52.83	682	0.075	0.419	0.642	+53.22	0.225	0.545	0.673	+23.49
July	611	0.100	0.231	0.459	+98.70	684	0.027	0.357	0.518	+45.10	0.050	0.540	0.582	+7.78
August	628	0.167	0.507	0.576	+13.70	630	0.216	0.465	0.535	+15.05	0.415	0.529	0.607	+14.74
September	690	0.507	0.573	0.683	+19.11	668	0.456	0.675	0.725	+7.41	0.532	0.664	0.755	+13.70
October	480	0.278	0.528	0.598	+13.28	718	0.282	0.543	0.603	+11.05	0.372	0.504	0.633	+25.60
November	707	0.256	0.573	0.674	+17.59	573	0.466	0.703	0.753	+7.11	0.474	0.742	0.793	+6.87
December	725	0.284	0.656	0.834	+27.13	375	0.550	0.654	0.822	+25.69	0.563	0.729	0.891	+22.22
Average		**0.234**	**0.532**	**0.655**	**+28.11**		**0.322**	**0.578**	**0.665**	**+17.21**	**0.402**	**0.637**	**0.714**	**+12.36**
Season-long														
Spring	2025	0.143	0.589	0.626	+6.28	1874	0.347	0.643	0.683	+6.22	0.384	0.721	0.75	+4.17
Summer	1954	0.131	0.521	0.573	+9.98	1996	0.293	0.539	0.599	+11.13	0.335	0.619	0.709	+5.71
Fall	1877	0.226	0.696	0.734	+5.46	1959	0.249	0.736	0.776	+5.43	0.282	0.754	0.79	+3.95
Winter	2111	0.186	0.733	0.812	+10.78	1626	0.48	0.792	0.842	+6.31	0.531	0.802	0.887	+8.54
Average		**0.172**	**0.635**	**0.686**	**+8.11**		**0.335**	**0.643**	**0.696**	**+8.82**	**0.383**	**0.666**	**0.743**	**+11.67**
Year-long														
Annual	7967	**0.0326**	**0.5517**	**0.614**	**+11.29**	7455	**0.255**	**0.493**	**0.563**	**+14.20**	**0.245**	**0.63**	**0.746**	**+18.41**

1: Met: Meteorological factors.
2: Change of enhanced SVR relative to SVR.

SVR and 0.56–0.75 for enhanced SVR applied across all time horizons. For the two model applications without auxiliary co-pollutants, using the enhanced approach improved model performance by between +5.5% and +98.7% compared with the regular SVR, depending on the month of data utilized.

Second, model performance varied by time horizon, especially for MLR, with lower accuracy when applied to longer time horizons (i.e., annual vs. seasonal vs. monthly). However, differences in model performance by time horizon were smaller for SVR and enhanced SVR, suggesting these approaches are more flexible for predicting hourly concentrations across varied time horizons ranging from month-long to year-long.

Third, model performance for monthly (and seasonal) time horizons varied by month (and season) for all model approaches. Among the month-long dataset applications, hourly $PM_{2.5}$ concentrations were predicted by enhanced SVR (the best performing approach) with the highest accuracy in January (R^2 ranging 0.83–0.94), followed by December (R^2 ranging 0.82–0.89), February (R^2 ranging 0.69–0.79), and November (R^2 ranging 0.67–0.79) across both locations. Conversely, model performance was lowest for the July datasets for all approaches, with R^2 ranging 0.46–0.58 for enhanced SVR, 0.23–0.54 for SVR, and 0.03–0.10 for MLR. Aggregating on a seasonal basis, model performance was highest for winter followed by fall, spring, and summer in descending order of accuracy. Additionally, moving from regular SVR to enhanced SVR in both locations (ignoring co-pollutants) generally had the largest impact on improving model performance in summer months, with improvements as high as +98.7% in July in Logan square (i.e., R^2 increasing from 0.23 to 0.46).

Fourth, the addition of auxiliary co-pollutants to the Ashburn datasets increased model performance in all months. The largest increase in model performance resulting from introducing co-pollutants was for the annual dataset, increasing R^2 by ~33% (from 0.56 to 0.75). Additionally, the magnitude of improvement varied by month and season, with higher improvements observed in spring and summer months than fall and winter months. For example, comparing enhanced SVR approaches in Ashburn with and without co-pollutants, adding co-pollutants increased R^2 on an absolute basis by an average of 0.06 (~10% relative basis) in spring and summer months and by an average of 0.04 (~5% relative basis) in fall and winter months. The highest individual months of relative improvement in model performance achieved by adding co-pollutants to the enhanced SVR approach were May (15%), August (13%), and July (12%), respectively. Model accuracy was never reduced by including co-pollutants in the Ashburn datasets. Because of these significant improvements, we subsequently explore in more detail the impacts of including auxiliary pollutants in the inter-model sensitivity analysis in section 3.3

Table 8 provides results from the intra-model sensitivity analysis of the regular SVR approaches (without auxiliary pollutants) applied to the Ashburn and Logan Square datasets using the variable importance (VI) technique. Seasonal and yearlong hourly horizons were applied and the VIs were aggregated into the variable-category level, including meteorological, daytime/calendar, and human activity patterns, and normalized between 0% and 100% (Type A). To further explore the influence of including human activity patterns in these models, we also divided this category into mobility and occupancy patterns. We also normalized the VIs between 0% and 100% for these two sub-categories (Type B).

For Type A analysis, results suggest that the meteorological category of variables is more influential than the other two categories of variables for predicting both seasonal and yearlong hourly datasets. Further, including human activity patterns is more influential in summer than winter seasons (i.e., 32.9% and 26.2% vs. 18.2% and 20.0% for Ashburn and Logan Square, respectively). Second, for Type B analysis, it was found that for all durations in Ashburn and Logan Square, mobility factors are more influential than occupancy profiles in the model. For example, the magnitude of occupancy importance in winter was found to be 29% and 27% for Ashburn and Logan Square, respectively,

Table 8
Percentage of relative VIs for intra-model sensitivity analysis based on flatness of PDP curve over ICE plots. Seasonal and annual hourly $PM_{2.5}$ concentrations based on datasets without auxiliary pollutants were predicted using the regular SVR approach for Ashburn and Logan Square.

Type	Category of variables	Spring	Summer	Fall	Winter	Annual
Ashburn						
A	Meteorological	61.1%	53.1%	56.6%	62.3%	60.9%
	Human activity patterns	22.2%	32.9%	19.1%	18.2%	21.3%
	Daytime/calendar	16.7%	14.1%	24.3%	19.5%	17.8%
		Percentage of occupancy vs mobility				
B	Occupancy profiles	17.1%	16.0%	16.2%	29.0%	24.8%
	Mobility profiles	82.9%	84.0%	83.8%	71.0%	75.2%
Logan Square						
A	Meteorological	63.0%	48.7%	65.3%	63.0%	59.5%
	Human activity patterns	23.6%	26.2%	21.0%	20.0%	18.4%
	Daytime/calendar	13.4%	25.2%	13.7%	17.0%	22.1%
		Percentage of occupancy vs mobility				
B	Occupancy profiles	14.8%	12.3%	18.6%	27.0%	12.5%
	Mobility profiles	85.2%	87.7%	81.4%	73.0%	87.5%

compared with only 16.2% and 12.3%, respectively, in summer. This may be attributable to higher combustion emissions by the buildings sector during winter compared to summer in Chicago (Energy Usage 2010 | City, 2010). Research conducted by (Clougherty et al., 2013) confirms such impacts of buildings in wintertime for explaining variations in intraurban $PM_{2.5}$ concentrations for New York City.

3.3. Understanding impacts of human activity patterns based on enhanced vs. regular approaches using inter-model sensitivity analysis

3.3.1. Models without auxiliary pollutants

Table 9 shows results obtained from the application of both the regular and enhanced kernel SVR models to weeklong hourly data with various combinations of variables for predicting $PM_{2.5}$ concentrations across the three urban locations in Chicago. First, limiting only to the regular models, in Logan Square, model-9, which includes residential building occupancy profiles as a way to describe human activity patterns, provided the most accurate results ($R^2 = 0.81$), even more effectively than the models that employed mobility (traffic) factors (i.e., models 4 through 7 with R^2 ranging 0.73–0.78). This suggests that residential building occupancy patterns may impact local ambient $PM_{2.5}$ concentrations in this neighborhood with its primarily residential morphology more than commercial building occupancy patterns or even traffic patterns. Similarly, hourly concentrations of $PM_{2.5}$ in the Loop were best explained by the regular models that included commercial building occupancy profiles (i.e., model-6 with $R^2 = 0.82$), which further suggests that building occupancy patterns may impact local $PM_{2.5}$ concentrations in this neighborhood with its primarily commercial building morphology. Conversely, concentrations of $PM_{2.5}$ in Ashburn, which contains a combination of low-rise residential and industrial buildings, were best explained with model-3, which did not include human activity or traffic patterns. Additionally, air pollution and most of the weather data in the Loop were collected from the same location; thus, the level accuracy of the best model results for this location is higher than the other neighborhoods (i.e., $R^2 = 0.91$ vs. 0.81 and 0.82).

Compared to the regular models in Table 9, the enhanced models explained variations in $PM_{2.5}$ concentrations with greater accuracy. For example, Model-15 is based on residential occupancy and traffic profiles combined, which includes the newly created variable (C3). This model explained variations in $PM_{2.5}$ concentrations in the Logan Square

Table 9
Results of inter-model sensitivity analysis based on R^2 metric obtained from regular and enhanced SVR models for weeklong hourly data in June 2017 with various combinations of variables for predicting intraurban $PM_{2.5}$ concentrations in Logan Square, Ashburn, and Loop.

Model	Explanatory Variables	Logan Square				Ashburn				Loop (Downtown)			
		June 23–30, 2017				June 23–30, 2017				June 22–29, 2017			
		n = 172				n = 164				n = 165			
		R^2	adj. R^2	RMSE	MAE	R^2	adj. R^2	RMSE	MAE	R^2	adj. R^2	RMSE	MAE
Regular													
0	Met	0.714	0.705	0.122	0.094	0.682	0.672	0.127	0.102	0.768	0.761	0.140	0.097
1	Hour	0.003	−0.028	0.209	0.166	0.000	−0.032	0.230	0.174	0.052	0.022	0.243	0.197
2	Hour, Met	0.767	0.760	0.114	0.091	0.718	0.708	0.126	0.100	0.852	0.847	0.120	0.082
3	Hour, Weekday, Met (Base Model)	0.807	0.801	0.106	0.086	**0.821**	**0.815**	**0.109**	**0.087**	0.892	0.888	0.104	0.075
4	Hour, Weekday, Met, Tcount, Tspeed	0.777	0.770	0.112	0.089	0.709	0.700	0.132	0.104	0.862	0.858	0.102	0.075
5	Hour, Weekday, Met, Comm	0.736	0.728	0.126	0.097	0.759	0.751	0.124	0.095	0.898	0.895	0.098	0.069
6	Hour, Weekday, Met, Tcount, Tspeed, Comm	0.770	0.763	0.113	0.095	0.625	0.613	0.148	0.117	**0.910**	**0.908**	**0.091**	**0.067**
7	Hour, Weekday, Met, Tcount, Tspeed, Comm, Res	0.778	0.771	0.111	0.091	0.587	0.574	0.154	0.121	0.900	0.897	0.089	0.067
8	Hour, Weekday, Met, Tcount, Tspeed, Res	0.811	0.805	0.106	0.083	0.665	0.654	0.138	0.111	0.865	0.860	0.101	0.075
9	Hour, Weekday, Met, Res	**0.811**	**0.805**	**0.106**	**0.085**	0.805	0.799	0.114	0.091	0.866	0.862	0.131	0.086
Enhanced													
10	Hour, Weekday, Met, C1[1]	0.818	0.812	0.110	0.086	**0.843**	**0.838**	**0.096**	**0.080**	0.873	0.869	0.108	0.077
11	Hour, Weekday, Met, Comm, C1	0.804	0.798	0.118	0.093	0.778	0.771	0.106	0.083	**0.922**	**0.919**	**0.088**	**0.059**
12	Hour, Weekday, Met, Res, C1	0.806	0.800	0.113	0.088	0.809	0.803	0.111	0.094	0.872	0.868	0.100	0.065
13	Hour, Weekday, Met, C2[2]	0.824	0.818	0.102	0.083	0.791	0.784	0.115	0.095	0.848	0.843	0.141	0.105
14	Hour, Weekday, Met, Res, C2	0.817	0.812	0.106	0.085	0.803	0.797	0.107	0.088	0.856	0.852	0.129	0.091
15	Hour, Weekday, Met, C3[3]	**0.842**	**0.837**	**0.101**	**0.081**	0.822	0.817	0.110	0.090	0.861	0.857	0.124	0.095

C1 = Tcount & Tspeed; C2 = Comm & Tcount; C3 = Res & Tspeed.
Cells highlighted in bold lettering highlight the best performing model scenario in the table of model comparisons.

neighborhood slightly more accurately than the best performing regular model (R^2 = 0.84 vs. 0.81). Similarly, the enhanced model-10, which includes traffic speed variables but not building occupancy variables, best explained variations in $PM_{2.5}$ concentrations in the Ashburn neighborhood (R^2 = 0.84 vs. 0.82). The enhanced model-11, which includes both commercial building occupancy and mobility (traffic) profiles (via C1), explained variations in $PM_{2.5}$ concentrations in the Loop neighborhood with the highest accuracy (R^2 = 0.92). It should be noted that the accuracy of enhanced models in the inter-model sensitivity analysis is lower than the models developed in Table 7 because we only cascaded one category of variables (i.e., human activity patterns) for this section to capture how cascading this category of variables improves model accuracy. This is a necessary limitation to answer this specific question. The enhanced approach enabled explanation of local $PM_{2.5}$

Table 10
Results of inter-model sensitivity analysis based on R^2 metric obtained from regular and enhanced SVR models for month-long hourly data in June 2017 with various combinations of variables for predicting intraurban $PM_{2.5}$ concentrations in Logan Square and Ashburn.

Model	Explanatory Variables	Logan Square				Ashburn			
		June 1–31, 2017				June 1–31, 2017			
		n = 717				n = 682			
		R^2	adj.R^2	RMSE	MAE	R^2	adj.R^2	RMSE	MAE
Regular									
0	Met	0.323	0.319	0.116	0.082	0.329	0.324	0.127	0.093
1	Hour	0.027	0.022	0.140	0.106	0.002	−0.005	0.155	0.115
2	Hour, Met	0.364	0.361	0.112	0.079	0.397	0.392	0.121	0.087
3	Hour, Weekday, Met (Base Model)	**0.422**	**0.419**	**0.108**	**0.076**	**0.603**	**0.600**	**0.099**	**0.073**
4	Hour, Weekday, Met, Tcount, Tspeed	0.396	0.394	0.110	0.078	0.554	0.550	0.106	0.074
5	Hour, Weekday, Met, Comm	0.335	0.332	0.118	0.083	0.527	0.524	0.110	0.083
6	Hour, Weekday, Met, Tcount, Tspeed, Comm	0.323	0.320	0.118	0.084	0.436	0.432	0.126	0.096
7	Hour, Weekday, Met, Tcount, Tspeed, Comm, Res	0.332	0.329	0.117	0.084	0.419	0.414	0.129	0.097
8	Hour, Weekday, Met, Tcount, Tspeed, Res	0.390	0.388	0.111	0.079	0.478	0.474	0.119	0.086
9	Hour, Weekday, Met, Res	0.420	0.418	0.108	0.076	0.596	0.593	0.099	0.076
Enhanced									
10	Hour, Weekday, Met, C1	0.431	0.429	0.106	0.076	**0.625**	**0.623**	**0.089**	**0.067**
11	Hour, Weekday, Met, Comm, C1	0.379	0.376	0.110	0.080	0.463	0.459	0.117	0.088
12	Hour, Weekday, Met, Res, C1	0.418	0.415	0.107	0.078	0.496	0.492	0.112	0.083
13	Hour, Weekday, Met, C2	0.356	0.353	0.114	0.082	0.565	0.562	0.102	0.078
14	Hour, Weekday, Met, Res, C2	0.353	0.350	0.114	0.081	0.529	0.526	0.106	0.082
15	Hour, Weekday, Met, C3	**0.440**	**0.437**	**0.106**	**0.074**	0.562	0.559	0.103	0.078

C1 = Tcount & Tspeed; C2 = Comm & Tcount; C3 = Res & Tspeed.
Cells highlighted in bold lettering highlight the best performing model scenario in the table of model comparisons.

variations in these datasets and models with slightly higher accuracy by reducing Pearson correlations between factors in which both mobility and occupancy were incorporated together rather than using human-related factors independently.

Table 10 and Table 11 present performance metrics for predictions of $PM_{2.5}$ concentrations for month-long periods of hourly data for Ashburn and Logan Square in June 2017 and July 2017, respectively, using both regular and enhanced approaches. Similarly, Table 12 shows model performance in both Ashburn and Logan Square in December for a month-long duration. Unlike the results obtained from the week-long summertime data, the model effectiveness for the month-long hourly data was much lower, indicating that weekly patterns cannot be easily extracted out of a month-long hourly duration in the summer season. Thus, the generalization power of the month-long summer data in the model is lower than those of weeklong durations. This may be because of the lowest influence of the weekday factor, which explains less of the weekly $PM_{2.5}$ variations in the summer season (VI = 3.9%, 4.2%) compared to the winter season (VI = 9.9%, 9.6%) for Ashburn and Logan Square, respectively, based on the magnitude of flatness of the PDP curves over ICE Plots (Table 13). Moreover, among the regular models, model-3, which considered only daytime and meteorological factors without any human-related factors, provided the highest accuracy for the Logan Square and Ashburn locations in both June and July. This indicates that in the summertime the regular approach handles data with the base model including no human activity factors more effectively. Conversely, the enhanced models achieved slightly higher accuracy for both locations, with model-15 and model-10 performing best for Logan Square and Ashburn, respectively. These models included newly created variables (C1 and C3), indicating that human activity patterns can moderately improve model performance in these datasets if correlations between these patterns are removed through the proposed enhanced approach.

The best performing regular and enhanced models ifor December 2017 in both Logan Square (model-9 and model-15) and Ashburn (model-9 and model-14) included residential building occupancy patterns, which as mentioned earlier, may be because of higher combustion emissions by residential buildings during the winter in Chicago (Energy Usage 2010 | City, 2010). These results indicate that the variation of ambient $PM_{2.5}$ concentrations in Logan Square, which serves as a representative of a primarily residential neighborhood, is defined well by residential building occupancy and traffic profiles, while in Ashburn, which serves as a representative of a primarily residential and industrial neighborhood, the variation is explained most effectively by residential and commercial occupancies and traffic profiles together.

Additionally, contrary to the month-long summer data, the accuracy of models for the month-long winter data increased (i.e., from R^2 ranging 0.38–0.60 to R^2 ranging 0.76–0.80). Potential reasons for this better fit to month-long wintertime data may include the lack of inclusion of biogenic factors in the model that are known to associate with the formation of fine particulate matter in the summertime (Hallquist et al., 2009) and the accumulation of pollution in the summertime because of stagnant weather conditions in Chicago (Jing et al., 2016). Stagnation is a meteorological phenomenon that traps air pollutants, limiting their removal from an airshed (Leung, 2005). To explore the potential for stagnation effects, Fig. 3 illustrates wind speed vs. $PM_{2.5}$ variations against each other in the Logan Square and Ashburn neighborhoods in both July and December 2017 using centered-ICE plots averaged with PDP curves. The average of interaction between wind speed and $PM_{2.5}$ for all instances in July is almost negligible based on the flatter PDP curves in Fig. 3a (Ashburn) and 3c (Logan Square). Conversely, the magnitude of changes on y-axis is more variable in December (Fig. 3b and d) than in July for both neighborhoods. These data suggest that wind speed in wintertime contributes more to the model prediction of local $PM_{2.5}$ concentrations than in summertime, which is consistent with the literature (e.g., (Tunno et al., 2016b)).

Fig. 4 illustrates relative VI comparing regular and enhanced SVR approaches for predicting month-long hourly $PM_{2.5}$ concentrations in Ashburn in December (i.e., data from Table 12). This VI plots show how the cascading approach helps improve model performance based on reducing model complexity. For example, in model-7 (Fig. 4-c), the summation of relative VI of traffic speed and residential occupancy patterns is calculated to be 16%, while in the enhanced model (Fig. 4-d), the newly created variable C3 has a higher percentage of relative VI (17.9%) along with higher model performance (R^2 = 0.79 vs. 0.66),

Table 11
Results of inter-model sensitivity analysis based on R^2 metric obtained from regular and enhanced SVR models for month-long hourly data in July 2017 with various combinations of variables for predicting intraurban $PM_{2.5}$ concentrations in Logan Square and Ashburn.

Model	Explanatory Variables	Logan Square				Ashburn			
		July 1–31, 2017				July 1–31, 2017			
		n = 714				n = 690			
		R^2	adj.R^2	RMSE	MAE	R^2	adj.R^2	RMSE	MAE
Regular									
0	Met	0.223	0.219	0.105	0.081	0.319	0.315	0.089	0.070
1	Hour	0.119	0.114	0.111	0.088	0.002	0.004	0.108	0.086
2	Hour, Met	0.332	0.328	0.098	0.075	0.373	0.369	0.086	0.069
3	Hour, Weekday, Met (Base Model)	**0.378**	**0.375**	**0.093**	**0.071**	**0.446**	**0.443**	**0.080**	**0.062**
4	Hour, Weekday, Met, Tcount, Tspeed	0.303	0.299	0.100	0.075	0.427	0.423	0.081	0.064
5	Hour, Weekday, Met, Comm	0.332	0.328	0.098	0.075	0.347	0.343	0.088	0.068
6	Hour, Weekday, Met, Tcount, Tspeed, Comm	0.224	0.220	0.110	0.084	0.359	0.355	0.086	0.069
7	Hour, Weekday, Met, Tcount, Tspeed, Comm, Res	0.231	0.227	0.109	0.084	0.357	0.353	0.086	0.069
8	Hour, Weekday, Met, Tcount, Tspeed, Res	0.312	0.308	0.100	0.075	0.414	0.410	0.082	0.066
9	Hour, Weekday, Met, Res	0.359	0.355	0.095	0.073	0.445	0.442	0.080	0.063
Enhanced									
10	Hour, Weekday, Met, C1	0.345	0.342	0.095	0.073	**0.455**	**0.451**	**0.079**	**0.062**
11	Hour, Weekday, Met, Comm, C1	0.323	0.320	0.096	0.073	0.380	0.376	0.085	0.067
12	Hour, Weekday, Met, Res, C1	0.352	0.349	0.094	0.072	0.436	0.432	0.079	0.063
13	Hour, Weekday, Met, C2	0.335	0.331	0.095	0.073	0.398	0.394	0.085	0.066
14	Hour, Weekday, Met, Res, C2	0.346	0.343	0.094	0.072	0.406	0.402	0.083	0.066
15	Hour, Weekday, Met, C3	**0.412**	**0.409**	**0.090**	**0.069**	0.450	0.446	0.083	0.065

C1 = Tcount & Tspeed; C2 = Comm & Tcount; C3 = Res & Tspeed.
Cells highlighted in bold lettering highlight the best performing model scenario in the table of model comparisons.

Table 12
Results of inter-model sensitivity analysis based on R^2 metric obtained from regular and enhanced SVR models for month-long hourly data in December 2017 with various combinations of variables for predicting intraurban $PM_{2.5}$ concentrations in Logan Square and Ashburn.

Model	Explanatory Variables	Logan Square				Ashburn			
		December 1–31, 2017				December 1–31, 2017			
		n = 714				n = 690			
		R^2	adj.R^2	RMSE	MAE	R^2	adj.R^2	RMSE	MAE
Regular									
0	Met	0.333	0.330	0.148	0.104	0.443	0.440	0.128	0.086
1	Hour	0.006	0.002	0.186	0.133	0.003	−0.002	0.169	0.121
2	Hour, Met	0.407	0.404	0.140	0.096	0.530	0.528	0.117	0.075
3	Hour, Weekday, Met (Base Model)	0.662	0.661	0.111	0.08	0.695	0.693	0.093	0.062
4	Hour, Weekday, Met, Tcount, Tspeed	0.627	0.625	0.113	0.074	0.702	0.701	0.094	0.067
5	Hour, Weekday, Met, Comm	0.664	0.663	0.113	0.082	0.702	0.701	0.092	0.061
6	Hour, Weekday, Met, Tcount, Tspeed, Comm	0.643	0.641	0.118	0.089	0.695	0.693	0.097	0.069
7	Hour, Weekday, Met, Tcount, Tspeed, Comm, Res	0.656	0.654	0.118	0.090	0.684	0.682	0.099	0.068
8	Hour, Weekday, Met, Tcount, Tspeed, Res	0.678	0.677	0.114	0.086	0.709	0.708	0.095	0.069
9	Hour, Weekday, Met, Res	**0.755**	**0.754**	**0.098**	**0.077**	**0.761**	**0.760**	**0.084**	**0.061**
Enhanced									
10	Hour, Weekday, Met, C1	0.748	0.747	0.094	0.069	0.750	0.749	0.085	0.064
11	Hour, Weekday, Met, Comm, C1	0.735	0.734	0.097	0.071	0.746	0.745	0.086	0.064
12	Hour, Weekday, Met, Res, C1	0.763	0.761	0.091	0.068	0.757	0.756	0.084	0.062
13	Hour, Weekday, Met, C2	0.735	0.733	0.101	0.076	0.743	0.741	0.084	0.063
14	Hour, Weekday, Met, Res, C2	0.756	0.755	0.092	0.069	**0.794**	**0.793**	**0.076**	**0.059**
15	Hour, Weekday, Met, C3	**0.798**	**0.797**	**0.084**	**0.063**	0.788	0.787	0.078	0.060

C1 = Tcount & Tspeed; C2 = Comm & Tcount; C3 = Res & Tspeed.
Cells highlighted in bold lettering highlight the best performing model scenario in the table of model comparisons.

Table 13
Percentage of VIs for weekday factor for seasonal and annual hourly $PM_{2.5}$ predictions for Ashburn and Logan Square using regular SVR models in table based on calculating flatness of PDP curves over ICE plots.

Percentage of relative VIs for weekday factors for seasonal and annual hourly prediction					
	Spring	Summer	Fall	Winter	Annual
Ashburn					
Weekday	7.1%	**3.9%**	7.2%	9.9%	3.8%
Logan Square					
Weekday	6.8%	**4.2%**	6.7%	9.6%	6.8%

Cells highlighted in bold lettering highlight the best performing model scenario in the table of model comparisons.

indicating that the enhanced approach can effectively predict local $PM_{2.5}$ while data dimensionality is reduced. The Pearson correlation for creating the C3 variable was R = 0.99. We plotted these plots for July as well as December and July for Logan Square and found the same model improving process (although results are not shown for brevity). It was also found that wind speed and wind gust had marginal impacts on $PM_{2.5}$ concentrations. Research by (Ito et al., 2007) confirms lower correlations between $PM_{2.5}$ and wind speed for temporal trends. On the other hand, relative humidity was found to be an influencial factor in the model during the wintertime. Others have similarly found RH to be a significant predictor of urban $PM_{2.5}$ concentrations (Chu et al., 2010; Su et al., 2016; Tai et al., 2010; Cheng et al., 2015). Further, another study indicated that 70% of $PM_{2.5}$ concentration reductions in December was associated with variations in natural (meteorological) conditions (Zhang et al., 2019). Therefore, the findings from this VI analysis for both locations across seasons as well as the annual time horizon in Table 8 indicate the importance of meteorological over anthropogenic factors for the selected locations in Chicago.

Fig. 5 illustrates predicted out of sample (test set) $PM_{2.5}$ concentrations based on month-long data for Ashburn and Logan Square during July and December, using models obtained from Tables 11 and 12. The graph shows that model-0 (orange line) cannot handle extreme values in the model and moved across the mean values of the actual $PM_{2.5}$. In July,

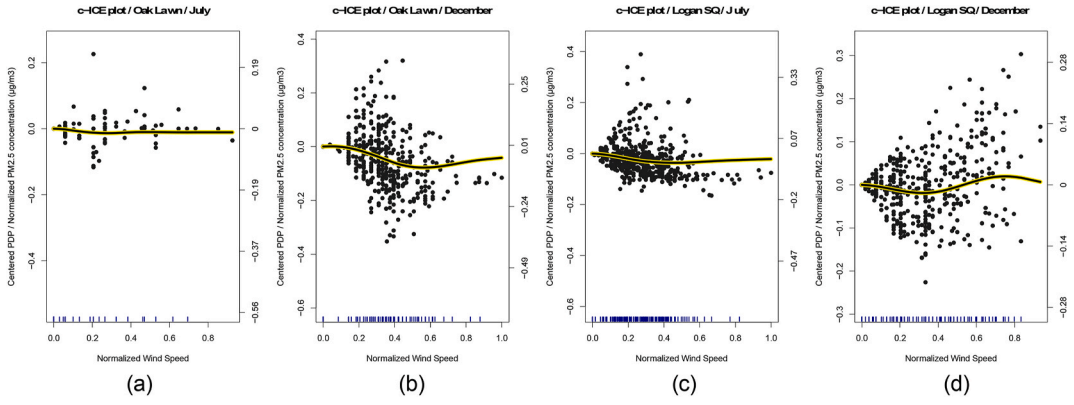

Fig. 3. Centered ICE plots averaged by PDP curves for visualization of interactions between wind speed and $PM_{2.5}$ variations in Ashburn (a, b) and Logan Square (c, d) in July (a, c) and December (b, d) 2017.

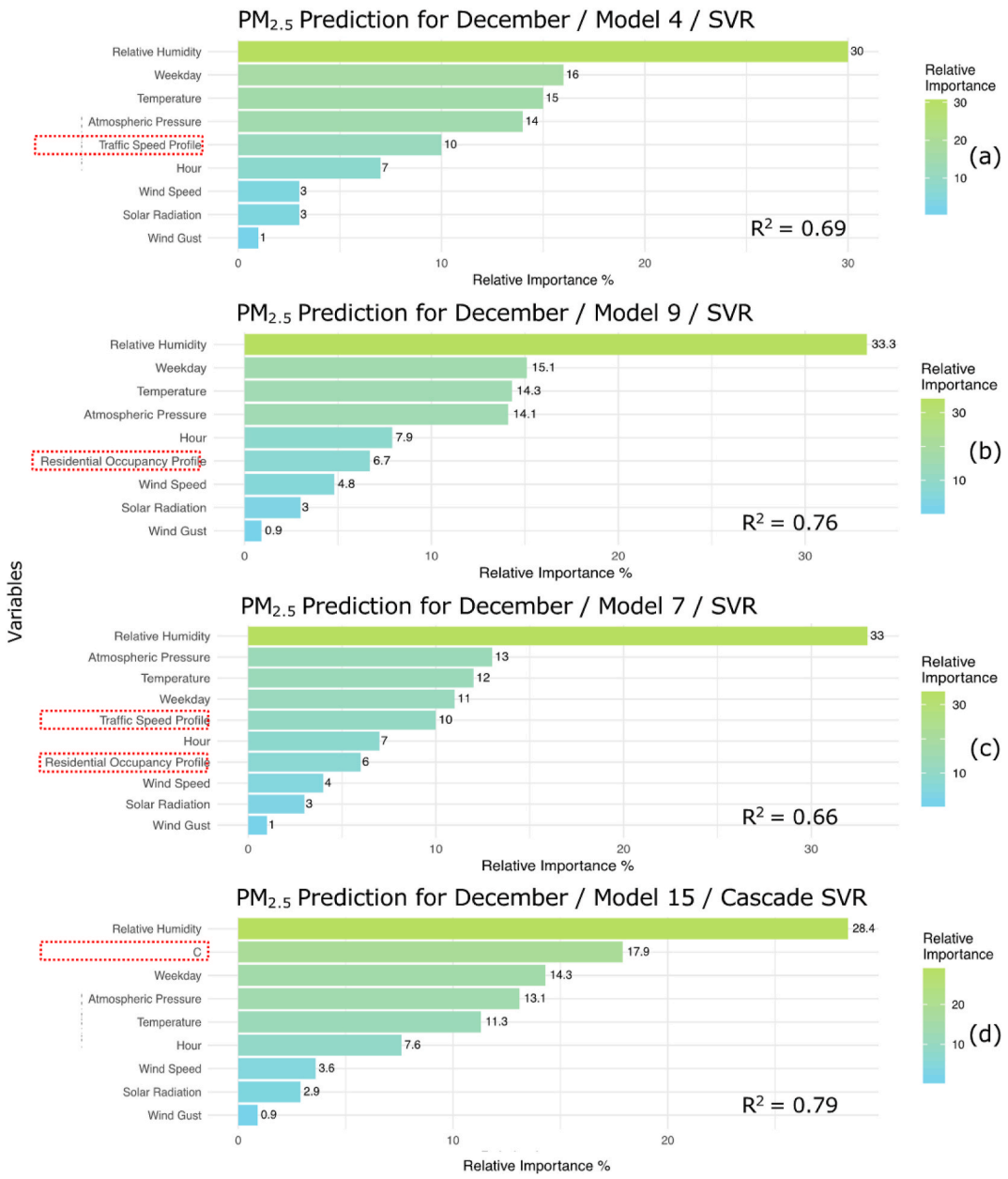

Fig. 4. Bar plots of relative VI for month-long hourly $PM_{2.5}$ prediction for Ashburn in December by regular SVR (a to c) and enhanced SVR (d) based for intra-model sensitivity analysis. Dotted red lines show factors for human activity patterns. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

model-3 (base model) was the best regular model, indicating that human activity patterns did not increase the model effectiveness using the regular predictive approach for summer data. It should be noted that model-3 and the best regular models are the same in July for both neighborhoods; thus, only one line (red) is shown. Unlike the regular models, the enhanced approach handled summer data better. In December, regular models with human activity patterns provided better fitting in the plots than the base model, meaning than in this month, outdoor $PM_{2.5}$ is more impacted by anthropogenic activities. As Fig. 5 shows, the best models, which are based on enhanced approach and C factors, handle extreme values more effectively than the regular approach.

3.3.2. *Models with auxiliary pollutants (NO_x and O_3)*

Tables 14 and 15 show the performance of both regular and enhanced models for predicting $PM_{2.5}$ concentrations at the Ashburn monitoring site in July 2017 and December 2017, respectively, considering additional auxiliary variables of concurrent ambient concentrations of ozone (O_3) and nitrogen oxides (NO_x) measured at the same site, which, as mentioned previously, are known precursor gaseous components for the formation of $PM_{2.5}$ (Anenberg et al., 2012). Only the Ashburn location was investigated for co-pollutant influences because it was the only monitoring site in our sample with data for all three constituents (i.e., $PM_{2.5}$, O_3, and NO_x). Tables 14 and 15 also include two additional models (model-16 and model-17) that include a new variable (C4) that accounts for O_3, solar radiation, and traffic variables; the models with null pollutants are the same as Table 11 for Ashburn. The results indicate that adding these two pollutants slightly increased model accuracy for both regular and enhanced models in July (i.e., R^2 increased by 0.03–0.04 compared to null models) but slightly decreased model accuracy in December (i.e., R^2 decreased by ~0.02 in enhanced models).

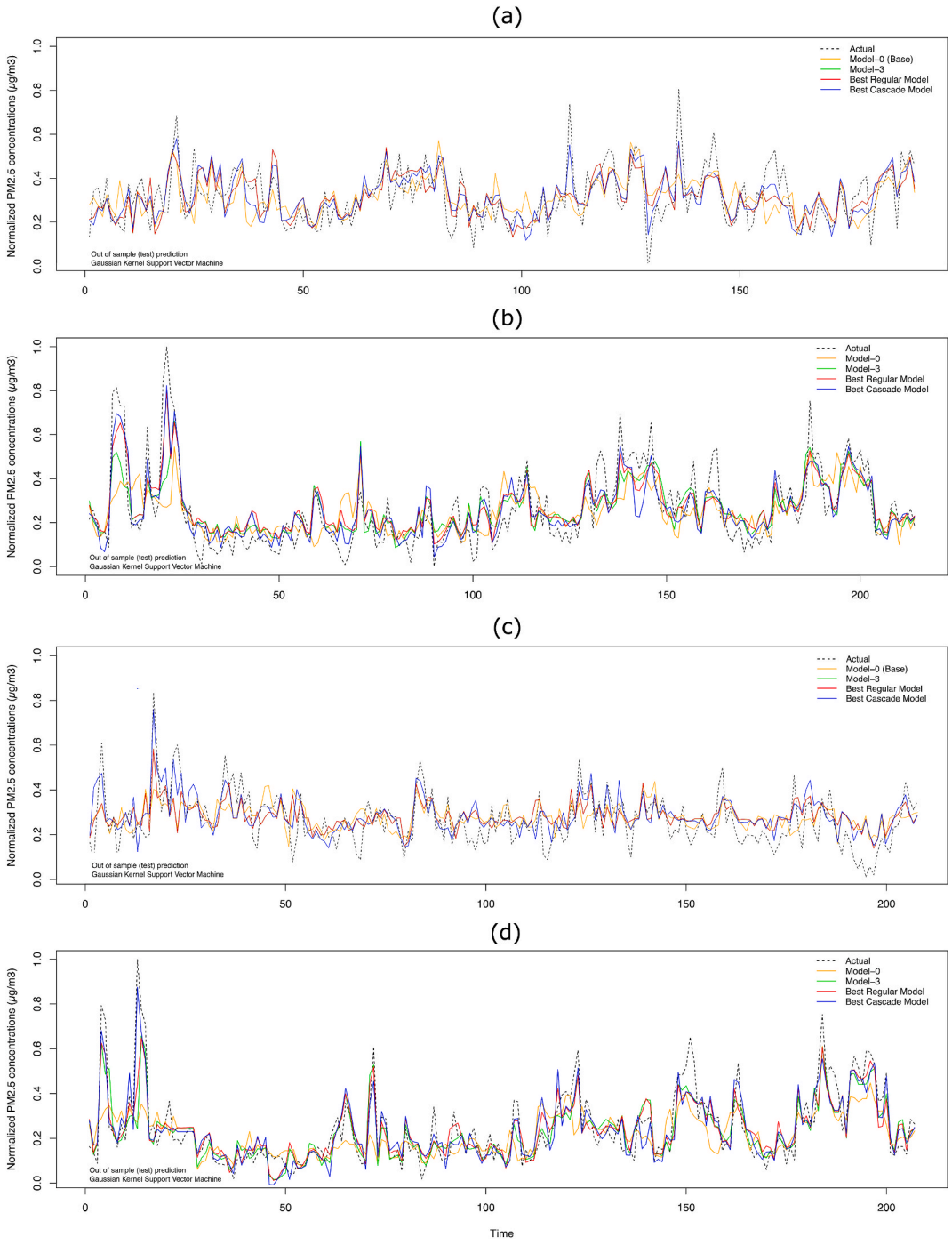

Fig. 5. Plots of predicted out of sample (test) $PM_{2.5}$ concentrations for month-long data in Logan Square and Ashburn during July (a) and (c) and December (b) and (d), 2017, comparing variations between actual data and the model-0, model-3 (base model), best regular model, and best enhanced models.

Regarding the p-models that include only pollutants as the explanatory variables, O_3 was found to be more effective than NO_x for hourly $PM_{2.5}$ prediction in July (Table 14), while NO_x explains $PM_{2.5}$ variations in December better than O_3 (Table 15). It should be noted that NO_x itself contributes as a precursor-pollutant for the formation of both urban O_3 (Khalil, 2018; Guo et al., 2019) and $PM_{2.5}$ (Fu et al., 2016; Sullivan et al., 2017). The accuracy of model-3, which includes all factors in the model, for July in Ashburn reduced from R^2 of 0.45 to R^2 of 0.37 and 0.39 by separately adding NO_x and O_3, respectively, indicating that adding auxiliary factors of pollutants imposed more complexity into the model and does not improve accuracy. However, considering these same pollutants but removing mobility (Tcount and Tspeed) and commercial occupancy (i.e., model-9) showed higher accuracy (R^2 = 0.48 for NO_x, 0.48 for O_3). Similar to July, these pollutants have interactions with human activity patterns in December (Table 15), as the model with no human activity patterns (model-3) was found to be the best model among regular models. This indicates that the auxiliary pollutants have interactions with human activity patterns. Thus, this added complexity into the model that reduced accuracy. Enhanced models, on the other hand, were able to keep the impacts of human activity patterns in the model while also slightly improving accuracy.

Model-16 in Table 14, which entails two new (calculated) variables (C3 and C4) is 3.5% more effective than model-9 in terms of R^2 because C3 reduced correlation between mobility and occupancy patterns (R =

Table 14
Results of inter-model sensitivity analysis based on R^2 metric obtained from regular and enhanced SVR models for month-long hourly data in July 2017 with various combinations of variables including auxiliary variables (O_3 and NO_x) for predicting intraurban $PM_{2.5}$ concentrations in Ashburn.

Model	Explanatory Variables	Ashburn											
		July 1–31, 2017											
		n = 697											
		Pollutant: Null				Pollutant: NOx				Pollutant: O_3			
		R^2	Adj. R^2	RMSE	MAE	R^2	Adj. R^2	RMSE	MAE	R^2	Adj. R^2	RMSE	MAE
Reguar													
0	Met, Pollutant	0.319	0.315	0.089	0.070	0.219	0.215	0.099	0.075	0.223	0.219	0.098	0.074
P	Pollutant	–	–	–	–	0.002	0.003	0.114	0.088	0.168	0.164	0.102	0.077
1	Hour, Pollutant	0.002	0.004	0.108	0.086	0.008	0.003	0.112	0.087	0.262	0.258	0.096	0.072
2	Hour, Pollutant, Met	0.373	0.369	0.086	0.069	0.306	0.302	0.096	0.074	0.300	0.297	0.094	0.070
3	Hour, Pollutant, Met, Weekday (Base Model)	**0.446**	**0.443**	**0.080**	**0.062**	0.373	0.369	0.086	0.069	0.386	0.383	0.092	0.074
4	Hour, Pollutant, Met, Weekday, Tcount, Tspeed	0.427	0.423	0.081	0.064	0.388	0.385	0.088	0.067	0.346	0.343	0.092	0.074
5	Hour, Pollutant, Met, Weekday, Comm	0.347	0.343	0.088	0.068	0.391	0.388	0.090	0.070	0.375	0.372	0.092	0.071
6	Hour, Pollutant, Met, Weekday, Tcount, Tspeed, Comm	0.359	0.355	0.086	0.069	0.374	0.371	0.090	0.072	0.384	0.381	0.089	0.070
7	Hour, Pollutant, Met, Weekday, Tcount, Tspeed, Comm, Res	0.357	0.353	0.086	0.069	0.374	0.371	0.090	0.073	0.389	0.386	0.089	0.071
8	Hour, Pollutant, Met, Weekday, Tcount, Speed, Res	0.414	0.410	0.082	0.066	0.421	0.418	0.086	0.068	0.406	0.403	0.087	0.070
9	Hour, Pollutant, Met, Weekday, Res	0.445	0.442	0.080	0.063	**0.475**	**0.473**	**0.083**	**0.066**	**0.475**	**0.472**	**0.084**	**0.068**
Enhanced													
10	Hour, Pollutant, Met, Weekday, C1	**0.455**	**0.451**	**0.079**	**0.062**	0.411	0.408	0.087	0.067	0.402	0.399	0.088	0.070
11	Hour, Pollutant, Met, Weekday, Comm, C1	0.380	0.376	0.085	0.067	0.446	0.444	0.085	0.066	0.450	0.448	0.085	0.066
12	Hour, Pollutant, Met, Weekday, Res, C1	0.436	0.432	0.079	0.063	0.467	0.464	0.084	0.065	0.459	0.456	0.084	0.066
13	Hour, Pollutant, Met, Weekday, C2	0.398	0.394	0.085	0.066	0.402	0.399	0.088	0.069	0.401	0.398	0.089	0.070
14	Hour, Pollutant, Met, Weekday, Res, C2	0.406	0.402	0.083	0.066	0.424	0.422	0.086	0.069	0.441	0.438	0.085	0.068
15	Hour, Pollutant, Met, Weekday, C3	0.450	0.446	0.083	0.065	0.471	0.468	0.084	0.066	0.436	0.433	0.087	0.070
16	Hour, Met, C3, C4	–	–	–	–	–	–	–	–	**0.480**	**0.477**	**0.082**	**0.064**
17	Hour, NO_X, Met, Weekday, C3, C4	–	–	–	–	**0.497**	**0.495**	**0.081**	**0.063**	–	–	–	–

Pollutant = O_3 or NO_x; C1 = Tcount & Tspeed; C2 = Comm & Tcount; C3 = Res & Tspeed; C4 = (O_3 & Sol) & Tcount.
Cells highlighted in bold lettering highlight the best performing model scenario in the table of model comparisons.

0.91) and C4 reduced correlation between solar radiation and O_3 (R = 0.97). Meanwhile, Model-17, which includes both NO_x and C4 in the same model was found to be 3.5% more effective than model-16 in July. Thus, C4, representing ozone, solar radiation, and traffic profiles, provided more accurate results than multiple correlated variables. Strong correlations between ozone and solar radiation (Chai and Draxler, 2014) has been well documented in the literature. Unlike Table 14, model-16 and model-17 in Table 15 were not the most effective models because the correlation between variables in creating C4 was not strong (R = 0.24). Thus, replacing ozone, solar radiation, and traffic profiles by C4 in December not only did not improve model performance but actually decreased accuracy below that of the regular models (i.e., model-3 and model-8). These results indicate that creating a new variable instead of a pair of variables can increase model accuracy only when they are highly correlated. Thus, the cascading approach based on the correlation criteria that are explored herein demonstrates the utility of the proposed workflow for improving predictive model accuracy through lowering model complexity.

3.4. Limitations and future work

There are a number of limitations to consider in this work. Results are first and foremost limited to the study locations in Chicago, IL; future work should expand to different cities. Although we focus on integrating dynamic building occupancy patterns as a way to potentially improve urban air quality prediction models (and we indeed observe some improvement), we rely on assumptions for hourly occupancy patterns for commercial and residential buildings from prototype building models and not actual building occupancy data from our study locations. As data becomes more accessible, future studies may benefit from using actual site-specific data such as mobile data (i.e., (Nyhan et al., 2016; Barbour et al., 2019)) for capturing city-specific and dynamic occupancy patterns. Additionally, we assumed 1000 m buffer radii for building occupancy patterns, which were drawn from prior studies for other cities in the U.S.; however, further investigation could provide city-specific buffer radii. Moreover, although the enhanced approach has shown to be capable of improving model accuracy and performance, it remains computational expensive; thus, more advanced optimization approaches such as evolutionary algorithm should be explored to handle large datasets and lower computational requirements.

4. Conclusion

This work was successful in developing and applying several machine learning (ML) approaches, including an enhanced ML approach – through hybridizing cascade and PCA approaches with ML algorithms – that incorporates dynamic human activity patterns (i.e., traffic mobility and building occupancy profiles), to improve the performance and accuracy of urban air quality prediction models compared to conventional approaches. Application of the modeling approaches to time series (hourly) $PM_{2.5}$ datasets from three air quality monitoring sites in different urban neighborhoods in Chicago, IL demonstrated that the proposed workflow is able to improve the accuracy of current urban air quality models by (i) handling large datasets with many urban factors, (ii) reducing dimensionality and complexity of many correlated factors, and (iii) using the most suitable ML algorithms for solving the problem with no pre-deterministic approach. Further, the results obtained through the inter-model sensitivity analysis demonstrated that there

Table 15
Results of inter-model sensitivity analysis based on R^2 metric obtained from regular and enhanced SVR models for month-long hourly data in December 2017 with various combinations of variables including auxiliary variables (O_3 and NO_x) for predicting intraurban $PM_{2.5}$ concentrations in Ashburn.

Model	Explanatory Variables	Ashburn											
		December 1–31, 2017											
		n = 375											
		Pollutant: Null				Pollutant: NOx				Pollutant: O_3			
		R^2	adj. R^2	RMSE	MAE	R^2	adj. R^2	RMSE	MAE	R^2	adj. R^2	RMSE	MAE
Regular													
0	Met, Pollutant	0.398	0.389	0.142	0.110	0.505	0.498	0.129	0.100	0.674	0.670	0.110	0.087
p	Pollutant	–	–	–	–	0.682	0.677	0.114	0.089	0.249	0.239	0.158	0.117
1	Hour, Pollutant	0.000	0.014	0.191	0.142	0.173	0.161	0.167	0.122	0.262	0.258	0.096	0.072
2	Hour, Pollutant, Met	0.486	0.479	0.133	0.101	0.504	0.497	0.137	0.105	0.616	0.611	0.113	0.085
3	Hour, Pollutant, Met, Weekday (Base Model)	0.790	0.790	0.085	0.069	**0.811**	**0.808**	**0.081**	**0.064**	0.790	0.787	0.086	0.070
4	Hour, Pollutant, Met, Weekday, Tcount, Tspeed	0.675	0.671	0.107	0.082	0.674	0.669	0.109	0.082	0.753	0.750	0.091	0.071
5	Hour, Pollutant, Met, Weekday, Comm	0.684	0.679	0.105	0.081	0.682	0.677	0.114	0.089	0.717	0.713	0.099	0.075
6	Hour, Pollutant, Met, Weekday, Tcount, Tspeed, Comm	0.651	0.646	0.109	0.085	0.662	0.658	0.110	0.084	0.704	0.699	0.100	0.075
7	Hour, Pollutant, Met, Weekday, Tcount, Tspeed, Comm, Res	0.703	0.699	0.102	0.081	0.697	0.692	0.105	0.082	0.739	0.735	0.094	0.075
8	Hour, Pollutant, Met, Weekday, Tcount, Speed, Res	0.772	0.769	0.088	0.067	0.757	0.754	0.092	0.071	**0.795**	**0.792**	**0.083**	**0.068**
9	Hour, Pollutant, Met, Weekday, Res	**0.798**	**0.795**	**0.084**	**0.066**	0.776	0.773	0.089	0.071	0.793	0.790	0.084	0.064
Enhanced													
10	Hour, Pollutant, Met, Weekday, C1	0.812	0.810	0.080	0.063	0.771	0.768	0.088	0.070	0.808	0.805	0.083	0.065
11	Hour, Pollutant, Met, Weekday, Comm, C1	0.726	0.722	0.096	0.072	0.691	0.687	0.105	0.081	0.743	0.739	0.094	0.074
12	Hour, Pollutant, Met, Weekday, Res, C1	0.813	0.810	0.080	0.064	0.789	0.786	0.084	0.068	0.808	0.805	0.083	0.067
13	Hour, Pollutant, Met, Weekday, C2\	0.797	0.794	0.083	0.066	0.767	0.763	0.089	0.071	0.822	0.820	0.079	0.059
14	Hour, Pollutant, Met, Weekday, Res, C2	0.811	0.809	0.080	0.063	**0.823**	**0.820**	**0.078**	**0.060**	**0.829**	**0.826**	**0.077**	**0.059**
15	Hour, Pollutant, Met, Weekday, C3	**0.842**	**0.84**	**0.074**	**0.058**	0.802	0.799	0.082	0.064	0.790	0.787	0.085	0.066
16	Hour, Pollutant, Met, C3, C4	–	–	–	–	–	–	–	–	0.756	0.752	0.096	0.079
17	Hour, Pollutant, Met, Weekday, C3, C4	–	–	–	–	0.809	0.807	0.080	0.064	–	–	–	–

Pollutant = O_3 or NO_x; C1 = Tcount & Tspeed; C2 = Comm & Tcount; C3 = Res & Tspeed; C4 = (O_3 & Sol) & Tcount.
Cells highlighted in bold lettering highlight the best performing model scenario in the table of model comparisons.

exists a correlation between the dominant human activity patterns (i.e., mobility and building occupancy) in urban zones and intraurban ambient $PM_{2.5}$ concentrations.

Credit author statement

Mehdi Ashayeri: conceptualization; methodology; software; formal analysis; investigation; data curation; writing original draft; visualization. Narjes Abbasabadi: methodology; software; investigation; data curation; writing (review and editing); visualization. Mohammad Heidarinejad: investigation; resources; writing (review and editing); supervision; project administration. Brent Stephens: conceptualization; investigation; resources; writing (review and editing); supervision; project administration.

Funding sources

M.A. was supported in part by the John Vinci Distinguished Research Fellowship in the College of Architecture and by the Armour College of Engineering at Illinois Institute of Technology. M.H. was supported in part by an ASHRAE New Investigator Award.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbasabadi, N., Ashayeri, Mehdi, 2019. Urban energy use modeling methods and tools: a review and an outlook. Build. Environ. 161, 106270. https://doi.org/10.1016/j.buildenv.2019.106270.

Abbasabadi, N., Ashayeri, M., Azari, R., Stephens, B., Heidarinejad, M., 2019. An integrated data-driven framework for urban energy use modeling (UEUM). Appl. Energy 253, 113550. https://doi.org/10.1016/j.apenergy.2019.113550.

AirData website File Download page n.d. https://aqs.epa.gov/aqsweb/airdata/download_files.html. (Accessed 17 June 2019).

Allison, P.D., 1999. Multiple Regression: a Primer. Pine Forge Press, Thousand Oaks, Calif.

Anenberg, S.C., Schwartz, J., Shindell, D., Amann, M., Faluvegi, G., Klimont, Z., et al., 2012. Global air quality and health Co-benefits of mitigating near-term climate change through methane and black carbon emission controls. Environ. Health Perspect. 120, 831–839. https://doi.org/10.1289/ehp.1104301.

Azimi, P., Zhao, H., Fazli, T., Zhao, D., Faramarzi, A., Leung, L., et al., 2018. Pilot study of the vertical variations in outdoor pollutant concentrations and environmental conditions along the height of a tall building. Build. Environ. 138, 124–134. https://doi.org/10.1016/j.buildenv.2018.04.031.

Barbour, E., Davila, C.C., Gupta, S., Reinhart, C., Kaur, J., González, M.C., 2019. Planning for sustainable cities by estimating building occupancy with mobile phones. Nat. Commun. 10 https://doi.org/10.1038/s41467-019-11685-w.

Beck, M.W., 2018. NeuralNetTools : visualization and analysis tools for neural networks. J. Stat. Software 85. https://doi.org/10.18637/jss.v085.i11.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., et al., 2017. Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmospheric Pollution Research 8, 652–659. https://doi.org/10.1016/j.apr.2016.12.014.

Björklund, J.A., Thuresson, K., Cousins, A.P., Sellström, U., Emenius, G., de Wit, C.A., 2012. Indoor air is a significant source of tri-decabrominated diphenyl ethers to outdoor air via ventilation systems. Environ. Sci. Technol. 46, 5876–5884. https://doi.org/10.1021/es204122v.

Building Footprints (current) | city of Chicago | data portal. Chicago n.d. https://data.cityofchicago.org/Buildings/Building-Footprints-current-/hz9b-7nh8. (Accessed 20 October 2018).

Burke, J.M., Zufall, M.J., Özkaynak, H., 2001. A population exposure model for particulate matter: case study results for PM2.5 in Philadelphia, PA. J. Expo. Sci. Environ. Epidemiol. 11, 470–489. https://doi.org/10.1038/sj.jea.7500188.

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C.A., et al., 2018. Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proc. Natl. Acad. Sci. Unit. States Am. 115, 9592–9597. https://doi.org/10.1073/pnas.1803222115.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. Geosci. Model Dev. (GMD) 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Chemel, C., Sokhi, R.S., Yu, Y., Hayman, G.D., Vincent, K.J., Dore, A.J., et al., 2010. Evaluation of a CMAQ simulation at high resolution over the UK for the calendar year 2003. Atmos. Environ. 44, 2927–2939. https://doi.org/10.1016/j.atmosenv.2010.03.029.

Chen, Z., Cai, J., Gao, B., Xu, B., Dai, S., He, B., et al., 2017. Detecting the causality influence of individual meteorological factors on local PM2.5 concentration in the Jing-Jin-Ji region. Sci. Rep. 7 https://doi.org/10.1038/srep40735.

Cheng, Y., He, K., Du, Z., Zheng, M., Duan, F., Ma, Y., 2015. Humidity plays an important role in the PM 2.5 pollution in Beijing. Environ. Pollut. 197, 68–75. https://doi.org/10.1016/j.envpol.2014.11.028.

Chicago traffic tracker - congestion estimates by Segments | city of Chicago | data portal n.d. https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Se/n4j6-wkkf. (Accessed 17 October 2018).

Chowdhury, S., Dey, S., Smith, K.R., 2018. Ambient PM2.5 exposure and expected premature mortality to 2100 in India under climate change scenarios. Nat. Commun. 9 https://doi.org/10.1038/s41467-017-02755-y.

Chu, N., Kadane, J.B., Davidson, C.I., 2010. Using statistical regressions to identify factors influencing PM 2.5 concentrations: the pittsburgh supersite as a case study. Aerosol. Sci. Technol. 44, 766–774. https://doi.org/10.1080/02786826.2010.490798.

Clougherty, J.E., Kheirbek, I., Eisl, H.M., Ross, Z., Pezeshki, G., Gorczynski, J.E., et al., 2013. Intra-urban spatial variability in wintertime street-level concentrations of multiple combustion-related air pollutants: the New York City Community Air Survey (NYCCAS). J. Expo. Sci. Environ. Epidemiol. 23, 232–240. https://doi.org/10.1038/jes.2012.125.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., et al., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet 389, 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6.

Commercial prototype building models | building energy codes program n.d. https://www.energycodes.gov/development/commercial/prototype_models. (Accessed 18 June 2019).

Cyrys, J., Heinrich, J., Hoek, G., Meliefste, K., Lewné, M., Gehring, U., et al., 2003. Comparison between different traffic-related particle indicators: elemental carbon (EC), PM2.5 mass, and absorbance. J. Expo. Sci. Environ. Epidemiol. 13, 134–143. https://doi.org/10.1038/sj.jea.7500262.

Deru, M., Field, K., Studer, D., Benne, K., Griffith, B., Torcellini, P., et al., 2011. U.S. Department of energy commercial reference building models of the national building stock. https://doi.org/10.2172/1009264.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., Schwartz, J., 2016. Assessing PM 2.5 exposures with high spatiotemporal resolution across the continental United States. Environ. Sci. Technol. 50, 4712–4721. https://doi.org/10.1021/acs.est.5b06121.

Dons, E., Van Poppel, M., Kochan, B., Wets, G., 2013. Int Panis L. Modeling temporal and spatial variability of traffic-related air pollution: hourly land use regression models for black carbon. Atmos. Environ. 74, 237–246. https://doi.org/10.1016/j.atmosenv.2013.03.050.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., et al., 2012. Development of Land Use Regression models for PM(2.5), PM(2.5) absorbance, PM (10) and PM(coarse) in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46, 11195–11205. https://doi.org/10.1021/es301948k.

Elbir, T., 2003. Comparison of model predictions with the data of an urban air quality monitoring network in Izmir, Turkey. Atmos. Environ. 37, 2149–2157. https://doi.org/10.1016/S1352-2310(03)00087-6.

Energy usage 2010 | city of Chicago | data portal n.d. https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp. (Accessed 15 October 2018).

Fan, Z., Meng, Q., Weisel, C., Laumbach, R., Ohman-Strickland, P., Shalat, S., et al., 2009. Acute exposure to elevated PM2.5 generated by traffic and cardiopulmonary health effects in healthy older adults. J. Expo. Sci. Environ. Epidemiol. 19, 525–533. https://doi.org/10.1038/jes.2008.46.

Fann, N., Lamson, A.D., Anenberg, S.C., Wesson, K., Risley, D., Hubbell, B.J., 2012. Estimating the national public health burden associated with exposure to ambient PM2.5 and ozone. Risk Anal. 32, 81–95. https://doi.org/10.1111/j.1539-6924.2011.01630.x.

Fann, N., Coffman, E., Timin, B., Kelly, J.T., 2018. The estimated change in the level and distribution of PM2.5-attributable health impacts in the United States: 2005–2014. Environ. Res. 167, 506–514. https://doi.org/10.1016/j.envres.2018.08.018.

Fischer, M.L., Chan, W.R., Delp, W., Jeong, S., Rapp, V., Zhu, Z., 2018. An estimate of natural gas methane emissions from California homes. Environ. Sci. Technol. 52, 10205–10213. https://doi.org/10.1021/acs.est.8b03217.

Franceschini, S., Gandola, E., Martinoli, M., Tancioni, L., Scardi, M., 2018. Cascaded neural networks improving fish species prediction accuracy: the role of the biotic information. Sci. Rep. 8 https://doi.org/10.1038/s41598-018-22761-4.

Friedman, J.H., 2001. Greedy function approximation: a gradient. Boosting Machine 29, 44.

Fritsch, S., Guenther, F., Wright, M.N.. Neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet. 2019.

Fu, X., Wang, S., Chang, X., Cai, S., Xing, J., Hao, J., 2016. Modeling analysis of secondary inorganic aerosols over China: pollution characteristics, and meteorological and dust impacts. Sci. Rep. 6 https://doi.org/10.1038/srep35992.

Gao, M., Cao, J., Seto, E., 2015. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. Environ. Pollut. 199, 56–65. https://doi.org/10.1016/j.envpol.2015.01.013.

Garg, A., Gupta, D., 2008. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinf. 9 https://doi.org/10.1186/1471-2105-9-62.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2013. Peeking inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. ArXiv:13096392 [Stat].

Graf HP, Cosatto E, Bottou L, Durdanovic I, Vapnik V. Parallel Support Vector Machines: the Cascade SVM n.d.:8.

Greenwell, B.M., 2017. Pdp: an R package for constructing partial dependence plots. The R Journal 9, 421. https://doi.org/10.32614/RJ-2017-016.

Greenwell, B.M., Boehmke, B.C., McCarthy, A.J., 2018. A Simple and Effective Model-Based Variable Importance Measure. ArXiv:180504755 [Cs, Stat].

Greenwell, Brandon, Bradley, Boehmke, Cunningham, Jay. GBM Developers. Gbm: Generalized Boosted Regression Models. R package version 2.1.5. https://CRAN.R-project.org/package=gbm. 2019.

Grolemund, G., Wickham, H., 2011. Dates and times made easy with lubridate. J. Stat. Software 40. https://doi.org/10.18637/jss.v040.i03.

Gulia, S., Shiva Nagendra, S.M., Khare, M., Khanna, I., 2015. Urban air quality management-A review. Atmospheric Pollution Research 6, 286–304. https://doi.org/10.5094/APR.2015.033.

Guo, H., Gu, X., Ma, G., Shi, S., Wang, W., Zuo, X., et al., 2019. Spatial and temporal variations of air quality and six air pollutants in China during 2015–2017. Sci. Rep. 9 https://doi.org/10.1038/s41598-019-50655-6.

Hallquist, M., Wenger, J.C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., et al., 2009. The formation, properties and impact of secondary organic aerosol: current and emerging issues. Atmos. Chem. Phys. 82.

Hassan, R., Li, M., 2010. Urban air pollution forecasting using artificial intelligence-based tools. In: Villanyi, V. (Ed.), Air Pollution. Sciyo. https://doi.org/10.5772/10049.

He, H., Lu, W.-Z., Xue, Y., 2014. Prediction of particulate matter at street level using artificial neural networks coupling with chaotic particle swarm optimization algorithm. Build. Environ. 78, 111–117. https://doi.org/10.1016/j.buildenv.2014.04.011.

Heidarinejad, M., Mattise, N., Dahlhausen, M., Sharma, K., Benne, K., Macumber, D., et al., 2017. Demonstration of reduced-order urban scale building energy models. Energy Build. 156, 17–28. https://doi.org/10.1016/j.enbuild.2017.08.086.

Heiple, S., Sailor, D.J., 2008. Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. Energy Build. 40, 1426–1436. https://doi.org/10.1016/j.enbuild.2008.01.005.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., et al., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42, 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057.

Hou, P., Wu, S., 2016. Long-term changes in extreme air pollution meteorology and the implications for air quality. Sci. Rep. 6 https://doi.org/10.1038/srep23792.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A Practical Guide to Support Vector Classification.

Isukapalli, S.S., Brinkerhoff, C.J., Xu, S., Dellarco, M., Landrigan, P.J., Lioy, P.J., et al., 2013. Exposure indices for the national children's study: application to inhalation exposures in queens county, NY. J. Expo. Sci. Environ. Epidemiol. 23, 22–31. https://doi.org/10.1038/jes.2012.99.

Ito, K., Thurston, G.D., Silverman, R.A., 2007. Characterization of PM2.5, gaseous pollutants, and meteorological interactions in the context of time-series health effects models. J. Expo. Sci. Environ. Epidemiol. 17, S45–S60. https://doi.org/10.1038/sj.jes.7500627.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., et al., 2005. A review and evaluation of intraurban air pollution exposure models. J. Expo. Sci. Environ. Epidemiol. 15, 185–204. https://doi.org/10.1038/sj.jea.7500388.

Jing, P., O'Brien, T., Streets, D.G., Patel, M., 2016. Relationship of ground-level ozone with synoptic weather conditions in Chicago. Urban Climate 17, 161–175. https://doi.org/10.1016/j.uclim.2016.08.002.

Karagulian, F., Belis, C.A., Dora, C.F.C., Prüss-Ustün, A.M., Bonjour, S., Adair-Rohani, H., et al., 2015. Contributions to cities' ambient particulate matter (PM): a systematic review of local source contributions at global level. Atmos. Environ. 120, 475–483. https://doi.org/10.1016/j.atmosenv.2015.08.087.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., et al., 2019. Evaluation of different machine learning approaches to forecasting PM2.5 mass concentrations. Aerosol and Air Quality Research 19, 1400–1410. https://doi.org/10.4209/aaqr.2018.12.0450.

Karppinen, A., Kukkonen, J., Elolähde, T., Konttinen, M., Koskentalo, T., 2000a. A modelling system for predicting urban air pollution. Atmos. Environ. 34, 3735–3743. https://doi.org/10.1016/S1352-2310(00)00073-X.

Karppinen, A., Kukkonen, J., Elolähde, T., Konttinen, M., Koskentalo, T., Rantakrans, E., 2000b. A modelling system for predicting urban air pollution: model description and applications in the Helsinki metropolitan area. Atmos. Environ. 34, 3723–3733. https://doi.org/10.1016/S1352-2310(00)00074-1.

Kelley, M.C., Brown, M.M., Fedler, C.B., Ardon-Dryer, K., 2020. Long-term measurements of PM2.5 concentrations in lubbock, Texas. Aerosol and Air Quality Research 20, 1306–1318. https://doi.org/10.4209/aaqr.2019.09.0469.

Khalil, M.A.K., 2018. Steady states and transport processes in urban ozone balances. Npj Climate and Atmospheric Science 1. https://doi.org/10.1038/s41612-018-0035-7.

Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M., Switzer, P., et al., 2001. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. J. Expo. Sci. Environ. Epidemiol. 11, 231–252. https://doi.org/10.1038/sj.jea.7500165.

Kramer, O., 2015. Cascade support vector machines with dimensionality reduction. Applied Computational Intelligence and Soft Computing 2015, 1–8. https://doi.org/10.1155/2015/216132.

Leung, L.R., 2005. Potential regional climate change and implications to U.S. air quality. Geophys. Res. Lett. 32 https://doi.org/10.1029/2005GL022911.

Levy, Jonathan I., Clougherty, Jane E., Lisa, K., Baxter, E., Andres, Houseman, Paciorek, Christopher J., 2010. Evaluating Heterogeneity in Indoor and Outdoor Air Pollution Using Land-Use Regression and Constrained Factor Analysis. Health Effects Institute, Boston, MA.

Li, L., Chen, B., Zhang, Y., Zhao, Y., Xian, Y., Xu, G., et al., 2006. Retrieval of daily PM2.5 concentrations using nonlinear methods: a case study of the beijing–tianjin–hebei region, China. Remote Sensing 2018 10. https://doi.org/10.3390/rs10122006.

Li, Y., Chen, L., Ngoc, D.M., Duan, Y.-P., Lu, Z.-B., Wen, Z.-H., et al., 2015. Polybrominated diphenyl ethers (PBDEs) in PM2.5, PM10, TSP and gas phase in office environment in shanghai, China: occurrence and human exposure. PloS One 10, e0119144. https://doi.org/10.1371/journal.pone.0119144.

Lin, C.Q., Liu, G., Lau, A.K.H., Li, Y., Li, C.C., Fung, J.C.H., et al., 2018. High-resolution satellite remote sensing of provincial PM2.5 trends in China from 2001 to 2015. Atmos. Environ. 180, 110–116. https://doi.org/10.1016/j.atmosenv.2018.02.045.

Liu, X.-H., Zhang, Y., Cheng, S.-H., Xing, J., Zhang, Q., Streets, D.G., et al., 2010. Understanding of regional air pollution over China using CMAQ, part I performance evaluation and seasonal variation. Atmos. Environ. 44, 2415–2426. https://doi.org/10.1016/j.atmosenv.2010.03.035.

Liu, W., Guo, G., Chen, F., Chen, Y., 2019. Meteorological pattern analysis assisted daily PM2.5 grades prediction using SVM optimized by PSO algorithm. Atmospheric Pollution Research. https://doi.org/10.1016/j.apr.2019.04.005.

Local Weather Forecast. News and conditions | weather Underground n.d. https://www.wunderground.com/. (Accessed 11 November 2019).

Lu, W.-Z., Wang, W.-J., 2005. Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. Chemosphere 59, 693–701. https://doi.org/10.1016/j.chemosphere.2004.10.032.

Lujan, I.R., 2012. A Practical View of Large-Scale Classification: Feature Selection and Real-Time Classification.

Ma, Z., Hu, X., Sayer, A.M., Levy, R., Zhang, Q., Xue, Y., et al., 2016. Satellite-based spatiotemporal trends in PM 2.5 concentrations: China, 2004–2013. Environ. Health Perspect. 124, 184–192. https://doi.org/10.1289/ehp.1409481.

Mage, D., Ozolins, G., Peterson, P., Webster, A., Orthofer, R., Vandeweerd, V., et al., 1996. Urban air pollution in megacities of the world. Atmos. Environ. 30, 681–686. https://doi.org/10.1016/1352-2310(95)00219-7.

Masiol, M., Zíková, N., Chalupa, D.C., Rich, D.Q., Ferro, A.R., Hopke, P.K., 2018. Hourly land-use regression models based on low-cost PM monitor data. Environ. Res. 167, 7–14. https://doi.org/10.1016/j.envres.2018.06.052.

Mayer, H., 1999. Air pollution in cities. Atmos. Environ. 33, 4029–4037. https://doi.org/10.1016/S1352-2310(99)00144-2.

McDonald, B.C., de Gouw, J.A., Gilman, J.B., Jathar, S.H., Akherati, A., Cappa, C.D., et al., 2018. Volatile chemical products emerging as largest petrochemical source of urban organic emissions. Science 359, 760–764. https://doi.org/10.1126/science.aaq0524.

Meyer, David, Dimitriadou, Evgenia, Kurt, Hornik, Weingessel, Andreas, Leisch, Friedrich. e1071: misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-2. https://CRAN.R-project.org/package=e1071. 2019.

Moltchanov, S., Levy, I., Etzion, Y., Lerner, U., Broday, D.M., Fishbain, B., 2015. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. Sci. Total Environ. 502, 537–547. https://doi.org/10.1016/j.scitotenv.2014.09.059.

US EPA, 2009. Integrated Science Assessment for Particulate Matter. National Center for Environmental Assessment, Research Triangle Park, NC.

Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., et al., 2016. "Exposure track"—the impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. Environ. Sci. Technol. 50, 9671–9681. https://doi.org/10.1021/acs.est.6b02385.

Nyhan, M.M., Kloog, I., Britter, R., Ratti, C., Koutrakis, P., 2019. J. Expo. Sci. Environ. Epidemiol. 29, 238–247. https://doi.org/10.1038/s41370-018-0038-9.

WHO | Air pollution. WHO n.d. http://www.who.int/airpollution/en/(accessed February 2, 2019).

Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Model. 178, 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013.

Overview. World Bank n.d. http://www.worldbank.org/en/topic/urbandevelopment/overview (accessed February 2, 2019).

Perkins, S., 2019. Major U.S. cities are leaking methane at twice the rate previously believed. Science. https://doi.org/10.1126/science.aay8122.

Philibert, A., Loyce, C., Makowski, D., 2013. Prediction of N2O emission from local information with Random Forest. Environ. Pollut. 177, 156–163. https://doi.org/10.1016/j.envpol.2013.02.019.

Residential prototype building models | building energy codes program n.d. https://www.energycodes.gov/development/residential/iecc_models. (Accessed 17 June 2019).

Russo, A., Lind, P.G., Raischel, F., Trigo, R., Mendes, M., 2015. Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmospheric Pollution Research 6, 540–549. https://doi.org/10.5094/APR.2015.060.

Ryan, P.H., LeMasters, G.K., 2007. A review of land-use regression models for characterizing intraurban air pollution exposure. Inhal. Toxicol. 19, 127–133. https://doi.org/10.1080/08958370701495998.

Saint-Vincent, P.M.B., Pekney, N.J., 2019. Beyond-the-Meter: unaccounted sources of methane emissions in the natural gas distribution sector. Environmental Science & Technology. https://doi.org/10.1021/acs.est.9b04657.

Singh, K.P., Gupta, S., Kumar, A., Shukla, S.P., 2012. Linear and nonlinear modeling approaches for urban air quality prediction. Sci. Total Environ. 426, 244–255. https://doi.org/10.1016/j.scitotenv.2012.03.076.

Singh, K.P., Gupta, S., Rai, P., 2013. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ. 80, 426–437. https://doi.org/10.1016/j.atmosenv.2013.08.023.

Sohn, M.D., Dunn, L.N., 2019. Exploratory analysis of energy use across building types and geographic regions in the United States. Frontiers in Built Environment 5. https://doi.org/10.3389/fbuil.2019.00105.

Solomon, P.A., Sioutas, C., 2008. Continuous and semicontinuous monitoring techniques for particulate matter mass and chemical components: a synthesis of findings from EPA's particulate matter supersites program and related studies. J. Air Waste Manag. Assoc. 58, 164–195. https://doi.org/10.3155/1047-3289.58.2.164.

Solomon, P.A., Crumpler, D., Flanagan, J.B., Jayanty, R.K.M., Rickman, E.E., McDade, C. E.U.S., 2014. National PM 2.5 chemical speciation monitoring networks—CSN and IMPROVE: description of networks. J. Air Waste Manag. Assoc. 64, 1410–1438. https://doi.org/10.1080/10962247.2014.956904.

Su, X., Gough, W., Shen, Q., 2016. Correlation of PM 2.5 and Meteorological Variables in Ontario Cities: Statistical Downscaling Method Coupled with Artificial Neural Network. Crete, Greece, pp. 215–226. https://doi.org/10.2495/AIR160201.

Sullivan, R.C., Levy, R.C., da Silva, A.M., Pryor, S.C., 2017. Developing and diagnosing climate change indicators of regional aerosol optical properties. Sci. Rep. 7 https://doi.org/10.1038/s41598-017-18402-x.

Tai, A.P.K., Mickley, L.J., Jacob, D.J., 2010. Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: implications for the sensitivity of PM2.5 to climate change. Atmos. Environ. 44, 3976–3984. https://doi.org/10.1016/j.atmosenv.2010.06.060.

Tenenbaum, J.B., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323. https://doi.org/10.1126/science.290.5500.2319.

Terry, Therneau, Atkinson, Beth. Rpart: recursive partitioning and regression trees. R package version 4.1-13. https://CRAN.R-project.org/package=rpart. 2018.

Traffic count database system (TCDS) n.d. https://idot.ms2soft.com/tcds/tsearch.asp?loc=Idot&mod=. (Accessed 10 December 2019).

Tripathy, S., Tunno, B.J., Michanowicz, D.R., Kinnee, E., Shmool, J.L.C., Gillooly, S., et al., 2019. Hybrid land use regression modeling for estimating spatio-temporal exposures to PM2.5, BC, and metal components across a metropolitan area of complex terrain and industrial sources. Sci. Total Environ. 673, 54–63. https://doi.org/10.1016/j.scitotenv.2019.03.453.

Tunno, B.J., Dalton, R., Michanowicz, D.R., Shmool, J.L.C., Kinnee, E., Tripathy, S., et al., 2016a. Spatial patterning in PM2.5 constituents under an inversion-focused sampling design across an urban area of complex terrain. J. Expo. Sci. Environ. Epidemiol. 26, 385–396. https://doi.org/10.1038/jes.2015.59.

Tunno, B.J., Michanowicz, D.R., Shmool, J.L.C., Kinnee, E., Cambal, L., Tripathy, S., et al., 2016b. Spatial variation in inversion-focused vs 24-h integrated samples of PM2.5 and black carbon across Pittsburgh, PA. J. Expo. Sci. Environ. Epidemiol. 26, 365–376. https://doi.org/10.1038/jes.2015.14.

van Donkelaar, A., Martin, R.V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., et al., 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ. Health Perspect. 118, 847–855. https://doi.org/10.1289/ehp.0901623.

van Donkelaar, A., Martin, R.V., Brauer, M., Boys, B.L., 2014. Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter. Environmental Health Perspectives. https://doi.org/10.1289/ehp.1408646.

van Donkelaar, A., Martin, R.V., Brauer, M., Hsu, N.C., Kahn, R.A., Levy, R.C., et al., 2016. Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.5b05833.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, forth ed. Springer, New York.

von Schneidemesser, E., Monks, P.S., Allan, J.D., Bruhwiler, L., Forster, P., Fowler, D., et al., 2015. Chemistry and the linkages between air quality and climate change. Chem. Rev. 115, 3856–3897. https://doi.org/10.1021/acs.chemrev.5b00089.

Wang, W., Zhao, S., Jiao, L., Taylor, M., Zhang, B., Xu, G., et al., 2019. Estimation of PM2.5 concentrations in China using a spatial Back propagation neural network. Sci. Rep. 9 https://doi.org/10.1038/s41598-019-50177-1.

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.

Wickham, H., 2009. ggplot2. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-98141-3.

Xu, G., Jiao, L., Zhang, B., Zhao, S., Yuan, M., Gu, Y., et al., 2017. Spatial and temporal variability of the PM2.5/PM10 ratio in wuhan, Central China. Aerosol and Air Quality Research 17, 741–751. https://doi.org/10.4209/aaqr.2016.09.0406.

Yang, D., Wang, X., Xu, J., Xu, C., Lu, D., Ye, C., et al., 2018. Quantifying the influence of natural and socioeconomic factors and their interactive impact on PM2.5 pollution in China. Environ. Pollut. 241, 475–483. https://doi.org/10.1016/j.envpol.2018.05.043.

You, Y., Fu, H., Song, S.L., Randles, A., Kerbyson, D., Marquez, A., et al., 2015. Scaling support vector machines on modern HPC platforms. J. Parallel Distr. Comput. 76, 16–31. https://doi.org/10.1016/j.jpdc.2014.09.005.

Yousefian, F., Faridi, S., Azimi, F., Aghaei, M., Shamsipour, M., Yaghmaeian, K., et al., 2020. Temporal variations of ambient air pollutants and meteorological influences on their concentrations in Tehran during 2012–2017. Sci. Rep. 10. https://doi.org/10.1038/s41598-019-56578-6.

Yuan, Y., Liu, S., Castro, R., Pan, X., 2012. PM 2.5 monitoring and mitigation in the cities of China. Environ. Sci. Technol. 46, 3627–3628. https://doi.org/10.1021/es300984j.

Zhai, B., Chen, J., 2018. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. Sci. Total Environ. 635, 644–658. https://doi.org/10.1016/j.scitotenv.2018.04.040.

Zhang, H., Wang, Y., Hu, J., Ying, Q., Hu, X.-M., 2015. Relationships between meteorological parameters and criteria air pollutants in three megacities in China. Environ. Res. 140, 242–254. https://doi.org/10.1016/j.envres.2015.04.004.

Zhang, Y., West, J.J., Mathur, R., Xing, J., Hogrefe, C., Roselle, S.J., et al., 2018. Long-term trends in the ambient PM 2.5 - and O 3 -related mortality burdens in the United States under emission reductions from 1990 to 2010. Atmos. Chem. Phys. 18, 15003–15016. https://doi.org/10.5194/acp-18-15003-2018.

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., et al., 2019. Drivers of improved PM 2.5 air quality in China from 2013 to 2017. Proc. Natl. Acad. Sci. Unit. States Am. 116, 24463–24469. https://doi.org/10.1073/pnas.1907956116.