Using Mobile Monitoring and Vehicle Emissions to Develop and Validate Machine Learning Empirical Models of Particulate Air Pollution

by

Asmaa Salem Alazmi

Dissertation submitted to the faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Civil Engineering

Hesham A. Rakha, Chair

Steve C. Hankey, Co-Chair

Linsey C. Marr

Wenwen Zhang

July 8th , 2021

Blacksburg, VA

Keywords: Machine learning, Land use regression, Emission factors, Black carbon, Particle Number, spatial and temporal variation, Air pollution

Using Mobile Monitoring to Develop and Validate Machine Learning Empirical Models of Particulate Air Pollution in a Rural Appalachian Community

Asmaa Salem Alazmi

Abstract

Increasing levels of air pollution are prompting researchers to develop more reliable air pollution modeling approaches in order to protect the public and the environment from toxic contaminants and airborne pathogens. Although land use regression has long been used to assess exposure to air pollution, researchers are increasingly using machine learning algorithms to quantify the concentration of harmful pollutants—for this study black carbon (BC) and particle number (PN). Additionally, researchers are moving away from using fixed-site data in favor of using mobile monitoring data in a variety of locations to develop hourly empirical models of particulate air pollution.

This study uses secondary data describing BC and PN pollutant levels, which are obtained from roads that bikers share in the more rural location of Blacksburg (VA). Machine learning (ML) algorithms are then built to develop accurate and reliable short-term empirical prediction models. Different pre-processing methods for the mobile monitoring data and various input variables are tested to assess how ML can be used effectively in this process. Three types of time-average models are developed (daytime, hourly average, and one second models). Various combinations of spatial and temporal input variables are used in the short-term models. The impact of adding more spatiotemporal variables (e.g., emissions) to machine learning models to improve model performance is assessed in the short-term models. Incorporating spatial and temporal autocorrelation is intended to develop more sophisticated validation approaches for identifying ML performance patterns—the goal of which is to predict concentration levels more accurately in comparison to using raw data without data reprocessing. The results show that the model developed using refined disaggregated data is able to detect the spatial distribution of the pollutant concentration at equivalent levels as the smoothed data models, although the latter display fewer errors. The performance of the short-term model including all variables is equivalent to the model omitting emissions. The ML results are compared to earlier stepwise regression model results, suggesting that ML has the ability to improve both long-term and short-term model accuracy.

Our findings indicate that ML demonstrates higher predictive capacity in comparison to stepwise regression. The results from this study may be useful in enhancing the performance of ML through the incorporation of different data preprocessing tasks, as well as showing how different input variables contribute to the ML modeling process. The findings from this study could be used toward the development of environmental/eco-friendly routes that would decrease the risk for exposure to harmful vehicle-related emissions.

Using Mobile Monitoring to Develop and Validate Machine Learning Empirical Models of Particulate Air Pollution in a Rural Appalachian Community

Asmaa Salem Alazmi

General Audience Abstract

Air pollution is a major environmental threat to human health, claiming the lives of millions of people each year, primarily as a result of fine particulate matter entering the respiratory system. As such, it is important to develop reliable and accurate air pollution modeling approaches in order to protect the public and the environment from toxic contaminants and pathogens in the air. Although an approach known as land use regression has long been used to assess exposure to air pollution, researchers are increasingly using machine learning (ML) algorithms to quantify the concentration of harmful pollutants—for this study black carbon and particle number, which is a generic assessment that captures a number of known airborne hazards. Additionally, researchers are moving away from using fixed-site data in favor of using mobile monitoring data in a variety of locations to develop hourly empirical models of particulate air pollution.

In this study, machine learning algorithms are developed using secondary data collected from roads that bikers share, which are representative of pollution levels of particle number and black carbon in the more rural location of Blacksburg (VA), in order to develop accurate and reliable short-term empirical prediction models. Different pre-processing methods of the mobile monitoring data and various input variables are tested to assess how machine learning can be efficiently used in this process. Our findings indicate that machine learning demonstrates higher predictive capacity in comparison to stepwise regression. The results from this study are expected to be useful in enhancing the performance of machine learning through the incorporation of different data preprocessing tasks, as well as how different input variables contribute to the machine learning modeling process. The findings from this study could assist transportation planners and other stakeholders better assess pollution risks for bike riders and pedestrians. As such, this study's findings could be used toward the development of environmental/eco-friendly routes that would decrease the risk for exposure to harmful vehicle-related emissions.

Dedication

To My Parents, Salem and Jamla

To My Sisters and Brothers

To My Kids

To My friends

To Every Single Independent Mother around the World

Acknowledgements

My success achievement is only by Allah, Thanks to Allah for giving us countless blessings without us even asking.

Many thanks go to my parents, Salem and Jamla; without their steadfast support I would not be here pursuing my passion. They were both (and remain) pivotal in making me a stronger and more independent women. Mom, you spent so many hours teaching me when I was child, I know how much you wanted me to pursue this degree. Father, you are always proud—even going so far as to tell me that I was the most successful woman that you ever had ever seen. This dissertation is entirely representative of your contributions to my education and sense of purpose. I could not have done this without your support. The Messenger of Allah (ﷺ) said: "When a person dies, his deeds are cut off except for three: Continuing charity, knowledge that others benefited from, and a righteous son who supplicates for him". May Allah accept this work as "knowledge that others benefited from" for me and my family. May Allah help me return part of what you did for me.

Many thanks go to my sisters and brothers. I am so grateful to my sisters for supporting me financially and academically.

I also must acknowledge my children for understanding how much time this journey took and supporting me along every step of the way. I hope one day you understand that I did this for you. I hope you are proud of your mom for persevering! And I hope you follow in my footsteps.

Special thanks go to my best friend in academia and in my social sphere, Dr. Huda Alazmi. You were always with me during good moods and bad—always willing to help me overcome any challenge I was facing. Thanks also to my friends in Kuwait, Madison-Wisconsin and Blacksburg.

I extend my deepest gratitude to my advisor Prof. Hesham Rakha for guiding me through my research and helping me develop the skills needed to complete this doctorate. On those particularly challenging days you were always there to remind me that I DID have the ability to complete this degree—even when I doubted I did. You always supported me and knew when I needed an extra push. I am so fortunate that I was able to work with you—you taught me so much and I will remember every single word I learnt from you.

Many thanks go to Prof. Steven Hankey. I will always be grateful for the time you spent helping me navigate life in academia and in fine-tuning this dissertation! I will value our association forever. Thanks to my committee members, Prof. Wenwen Zhang and Prof. Lindsey Marr, for their time and the gift of their knowledge. I also thank the students and employees of the Center for Sustainable Mobility (CSM): You made me feel that the Virginia Tech Transportation Institute (VTTI) was my second home! I also owe a great debt to Jinghui Wang for running the simulations needed to calculate the emissions factors that were used in this research.

Finally, I thank and acknowledge every single independent mother around the world. Life is not easy, and you are not alone. Let us make the world different and let's work together on our own two feet to make our generation special and stronger. And remember when you fall, "Fall Forward"!

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Introduction

Land use regression modeling (LUR) has been used globally to classify air pollution exposure, predict the presence and level of traffic-related pollutants, and identify the potential health impacts on people living in urban areas. This study utilized a number of land-use variables (e.g., road type, traffic count, elevation, and vehicle emissions) as input variables in regression analysis to build an empirical predictive model of pollutant concentration.

Recently, researchers have used mobile monitoring data to characterize emissions (Pétron et al., 2012), assess pollutant intake/uptake (Götschi et al., 2015), identify immediate sources of pollutants (Hudda et al., 2014), assess nearby sources of pollution (Hagler et al., 2012), and develop regression models to assess traffic-related air pollution (Hankey & Marshall, 2015). Kerckhoffs et al. (2017) investigated the effect of monitoring type by comparing the performance of LUR for both stationary and mobile monitoring of ultrafine particle (UFP) and black carbon levels. Their results indicated that the ability of mobile UFP models to predict particle concentration increased from 36% for stationary measurements to 57% for outdoor measurements; in contrast, the accuracy of the black carbon remained the same between both monitoring scenarios. More recently, (Liu et al., 2019) explored the degree to which black carbon mobile data could explain concentration levels, confirming that land use regression using mobile data measurement could explain 68% of the variability in black carbon levels.

Mobile monitoring has the ability to cover a larger area within a specified time, which can deliver more data at high spatial resolution—the end result of which can be a large dataset. However, traditional regression may not have the ability to capture non-linear interactions accurately spatially when using data that are highly temporal resolution; One second

measurement data. In contrast, one of the advantages of machine learning is that it can be effective even when data are gathered in the absence of a carefully controlled experimental design and in the presence of complicated nonlinear environments. Recently, combined machine learning modeling with mobile monitoring data has been used to improve the accuracy and capture the effects of applying advanced algorithms. (Su et al., 2015b) applied a combination of mobile monitoring data and D/S/A machine learning algorithms to improve predictive capacity involving a small area of particle matter concentration. The combined model successfully maximized prediction accuracy and reduced overfitting. (Hasenfratz et al., 2015) derived high-resolution, urban air-pollution maps using Generalized Additive Models (GAMs). This model demonstrated distinct advantages in constructing high-resolution spatial maps, which successfully helped to define optimal travel routes that would reduce exposure to particles by an average of 7.1%.

Studies show that the magnitude of explained variance depends on the independent variables used in the model. Spatial data has the ability to reveal spatial non-stationarity and spatial autocorrelation (Anselin, 1988). Temporal variables (meteorological variables) have been shown to increase the coefficient of determination to an average of 15% in a nitrogen dioxide prediction model (Bertazzon et al., 2021). Additionally, (Rakha & Ahn, 2004) used the INTEGRATION model to quantify the environmental impacts of light duty vehicles. Therefore, it is useful to explore the degree to which adding spatiotemporal variables (emissions) as input variables will improve the predictive capacity of models. Moreover, combining static and dynamic independent variables with LUR models may facilitate detailed exposure estimates and exposure misclassifications over both short- and long-term delays. However, integrating mobile

measurement data offers the opportunity to estimate high spatial resolution concentrations with greater accuracy.

Despite the growing application of machine learning algorithms in environmental monitoring, the development of algorithms in spatiotemporal data applications able to forecast a specific concentration for an unknown location remains challenging due to spatial and temporal autocorrelation. Therefore, it is important to define a precise validation approach to assess model reliability. Studies have shown that splitting the fold to systematic (patterned) assignment of blocks helps to avoid over-optimistic error (Oliveira et al., 2019). In the presence of spatial and temporal structures, researchers typically evaluate a given model by treating either the spatial data (Haberlandt, 2007) or the temporal data (Ceci et al., 2016)—but not both. However, this blocking strategy must be fully reviewed. Splitting data in space and time, while also considering dependencies, can induce extrapolations by limiting the intervals of predictor variables available for model training, resulting in the overestimation of interpolation errors (Roberts et al., 2017). In addition to making it difficult to define an accurate validation approach, the dependence between observations that have spatial-temporal boundaries highlights the challenge of developing reliable performance evaluations, as discussed in prior work (Meyer et al., 2018). Although the applicability of cross-validation strategies that consider temporal and spatial autocorrelation merits further examination, no modeling design was found to be appropriate--particularly for capturing spatial-temporal mobile monitoring data in air pollution. This research deficit motivated us to use a more sophisticated validation approach, which is expected to lead to identifying performance patterns in machine learning, thereby augmenting the ability to predict concentration levels more accurately.

1.2 Research Problem Statement

Despite recent research initiatives, the bias of short-term prediction estimation remains a point of contention. In particular, the deployment of spatial-temporal context in the data may significantly affect a model's performance. Researchers have discussed how mobile monitoring data improves a model's predictive ability (Wang et al., 2020). Other studies have pointed out how including temporal variables (e.g., meteorological information and time of day) and mobile monitoring data affects model performance. However, adding more data will increase a model's complexity, thereby increasing the risk for degraded model performance (Abernethy et al., 2013). Therefore, it is important to investigate how the inclusion of spatial and temporal variables will increase a model's accuracy. Estimating the reliability of a pollutant-prediction model featuring different variables with differing variability factors remains a challenging target due to the complexity of the environment's features. Therefore, it is necessary to show how a prediction model can forecast pollution concentrations in an unseen area. It must be noted that validation approaches for assessing a model's reliability still need improvement. Moreover, it can be challenging to conduct a robust investigation due to the difficulty of identifying and implementing generalized machine learning models in air pollution research. While prior studies have discussed different validation approaches for estimating model overfitting, few studies have discussed the application of systematic cross-validation techniques for developing air pollution prediction models.

1.3 Research Objective and Contributions

This study was designed to develop an accurate and reliable pollution-prediction model to assess residential exposure to pollutants by applying machine learning algorithms to mobile measurement data. Machine learning models, which have long been known to maximize the

prediction accuracy of a target property, are increasingly being applied to estimating ambient air pollution exposure via the pre-processing of mobile monitoring data. During the machine learning modeling design phase, various input variables were tested to assess efficiency improvements using different data-preprocessing strategies. To achieve this goal, three types of time-averaging models were developed: (a) Daytime, involving 12 hour aggregated data; (b) Hourly-average, whereby the data were aggregated temporally at one hour and spatially at 100 m increments; and (c) One-second, whereby unrefined one-second measurement data were applied to investigating the ability of the machine learning to handle raw air-pollution measurements. Moreover, different combinations of time-dependent variables were used to build short-term models to gauge the level of improvement provided by the spatiotemporal factors.

Nowadays, simulations have the ability to identify the pollution-emission footprint of light duty vehicles based on fuel consumption levels. Therefore, it was beneficial to use emissions factors as an input variable; this approach generated both spatial and temporal information to the model training process that could maximize the prediction accuracy. In this research, the INTEGRATION software, a program being developed in CSM that provides simulated fuel output as an input variable for the emissions model for generating specific emission factors, was used as one source of the input variables. Any improvements linked to emissions factors were tested throughout this study by comparing different combinations of input variables.

The main advantage of collecting mobile data that it is presents better spatial coverage at the expense of temporal information. Therefore, in this research, the ability of the machine learning algorithm to predict pollution concentration in an area for which there was no directly collected data for constructing a model was investigated through spatial cross validation. This

approach is expected to be useful for exploring how machine learning might be advantageous for predicting pollution levels in diverse locations without direct monitoring; in other words, data collected from one location can be applied to another with similar characteristics for prediction purposes.

In contrast to the random cross-validation approach that randomly assesses performance, the more systematic approaches of spatial cross-validation and temporal cross-validation represent important strategies for assessing model performance outside a specific development area. Accordingly, the reliability of the model was tested by applying three different systematic cross-validation approaches: random, spatial, and spatial-temporal. By applying these three types of cross-validation techniques, the principles of validation approaches were defined, and model overestimation was explained. This knowledge will enhance our understanding of how the validation process can be improved to enhance and empower land-use regression modeling.

In conclusion, the results from this study could be useful in testing the performance of machine learning through the incorporation of different data preprocessing task, as well as how different input variables contribute to the machine learning modeling process. Importantly, the findings from this study could assist transportation planners and other stakeholders in developing more environmentally/eco-friendly bicycle and pedestrian routes that would decrease the risk for exposure to harmful emissions.

1.4 Dissertation Layout

This dissertation thesis consists of six chapters. Chapter 1, the Introduction, briefly discusses the research problem statement and research objectives and contributions; this chapter also includes a comprehensive overview of the literature applicable to this study's topical focus. Chapter 2 provides a review of the literature, with a focus on the state-of-the-art regarding land

use regression modeling using mobile monitoring data techniques and application. Chapter 3 discusses empirical modelling of particle air pollutants, as well as details the materials and methods used to develop the model implemented for this investigation. Chapter 4 provides the statistical results from the modeling analyses and demonstrates the ability of machine learning to predict concentrations of air pollutants using mobile monitoring data, and in particular, how the ML model developed for mobile monitoring data can be viewed in light of previous research and the working hypotheses. Finally, Chapter 5 provides a summary of this study's findings, as well as discusses suggestions for future research.

Chapter 2: Literature Review

2.1 Health Effects of Air Pollution

Estimating the impacts of long-term exposure to ambient air pollution is crucial for determining a range of possible health-related risks, such as cardiovascular and respiratory diseases. The World Health Organization (WHO) published a report in 2018 indicating that outdoor air pollution resulted in approximately 4.2 million premature deaths worldwide. (Khorrami et al., 2021) recently investigated the correlation between air pollution and lung cancer in 22 districts of Tehran, Iran, with a focus on long-term average exposure to a range of pollutants, notably PM_{10} (particulate matter with a diameter of 10 microns or less), SO_2 (sulfur dioxide), NO (nitric oxide), NO_2 (nitrogen dioxide), and NO_X (nitrogen oxide). Their data indicated that districts with higher levels of ambient air pollution were linked to a higher incidence of lung cancer. Furthermore, particulate matter is associated with numerous health issues and negative health outcomes, including respiratory and cardiovascular disorders (Brauer et al., 2016), lung cancer (Raaschou-Nielsen et al., 2013), asthma (Anenberg et al., 2018), and an overall shorter lifespan (Correia et al., 2013).

Particulate matter (PM) refers a range of potential pollutants (e.g., dust, soot, dirt, and smoke) of varying sizes and shapes. PM can also represent hundreds of different chemicals that can pose health risks for humans. The levels and severity of particulate matter are typically assessed according to particle mass and particle number (PN)—the latter being relevant for the present study. Black carbon (BC), also known as soot, forms as a result of the partial combustion of fossil fuels, wood, and other fuels. BC contributes to fine particulate air pollution ($PM_{2.5}$) and is known to impact climate, agriculture, and human health.

PN and BC represent two important primary emissions factors that can lead to poor air quality, negative health impacts, and global climate change. As such, levels of BC and PN serve as vital air pollution indicators for assessing the health risks associated with combustion particles emitted locally. BC and PN are thought to act as common carriers of various chemicals that are harmful to the human body; when inhaled, BC and PN can penetrate the circulatory system and cause complex biological responses (Oberdörster, 2000; Petzold et al., 2013). For instance, BC and PN have been linked to adverse neonatal risks, such as low birth weight and preterm birth (Bové et al., 2019; Gaspar et al., 2018). And, of course, the presence these pollutants in the atmosphere are known contributors to global warming (Brewer, 2019; (Paasonen et al., 2013)).

In particular, urban dwellers are routinely exposed to a potentially damaging combination of toxins from the outdoor environment. In order to identify high-risk areas for more intensive remediation efforts, policymakers and scientists must have accurate estimates of the spatial trends and behavior of pollutant emissions. Moreover, effective methods for capturing spatiotemporal variations in the distribution of pollutants are essential. Due to significant small-scale spatial heterogeneity, however, estimating exposure for epidemiological studies of short-term exposure to ambient air pollution continues to be a challenging task.

2.2 Land Use Regression

The use of land-use regression (LUR) models represents a widely used approach for assessing urban air pollution. In a multivariate regression model, LUR uses a measurement of air pollution concentration as the dependent variable, and variables such as traffic, meteorological, and other spatial variables as the independent variables (Gilbert et al., 2005). Parameter estimates derived from a regression model can then be used to estimate pollution levels for any area, even

within an individual's home. The use of a LUR model continues to be a cost-effective way to quantify air pollution exposure (Meng et al., 2015).

Land-use regression is a popular method widely used for analyzing, explaining, and predicting air pollution concentrations—particularly in areas that are densely populated. According to Morley and Gulliver (2018), the model relies on predictable air-pollution patterns to help estimate pollution patterns within a particular area. Regression equations are then used to describe the relationship between environmental variables and the target locations. In a study designed to review and evaluate the application of LUR models in the characterization of intra-urban exposure to air pollution, Ryan and LeMasters (2007) conducted a systematic literature review of six previous studies encompassing a total of twelve LUR models. Their investigation revealed that land-use regression models provide an accurate means of air pollution assessment for unmonitored locations. The researchers concluded that LUR models epitomize a powerful tool for integrating geographical and traffic information as a way of characterizing the variability in air pollution exposure.

Similarly, Hoek et al. (2008) conducted an empirical review of the use of LUR models in the assessment of intra-urban air pollution, confirming that this approach has been increasingly utilized over the last few decades. According to the researchers, land use regression models normally apply a wide range of predictor variables—typically population density, climate, land use, and physical geography. It should be noted that Hoek and colleagues stressed that LUR models require robust validation through the use of personal exposure monitoring.

According to Larkin et al. (2017), land use regression model can also be used effectively in various parts of the world for assessing nitrogen dioxide air pollution. Based on an experimental research approach, the researchers developed a LUR model for estimating the

global nitrogen exposure for the year 2011. The model was shown to be effective in monitoring global nitrogen dioxide variations with more than 10 variables being captured by the model. The researchers concluded that LUR models are effective in assessing global nitrogen dioxide exposure, which is critically important in global health studies and risk assessments. In a Shanghai-based study, Meng et al. (2015) concluded that the LUR NO_2 model fit better than the kriging and inverse distance weighed (IDW) interpolation approaches. In a significantly broader study, Knibbs et al. (2014) determined that LUR models were effective for predicting the spatial distribution of NO_2 in Australia; specifically, his data findings indicated that the model was able to explain 81% of annual spatial variations in NO_2, and 76% of monthly spatial variations in NO_2. In their study of semi-volatile organic contaminants (SVOCs), Meylmuk et al. (2013) reported that the use of LUR was able to explain an average of 81% of the variability in polychlorinated biphenyls and polycyclic musks, and 63% of polycyclic aromatic hydrocarbons and polybrominated diphenyl ether.

The spatial variance of annual mean concentrations for a variety of traffic-related contaminants—notably nitrogen oxide, particulate matter, and black carbon—has been successfully modeled using land use regression. (Wu et al., 2015) applied a land-use regression model to define spatial variations in $PM_{2.5}$ (fine particulate matter of 2.5 microns or less in diameter), confirming that $PM_{2.5}$ was shown to exhibit significant temporal variations in comparison to spatial variations. Su et al. (2015) showed that the LUR model is an appropriate modeling approach for assessing contaminant levels in densely populated developing urban areas; the model was found to be accurate in predicting levels of NO_2, NO, $PM_{2.5}$, and BC concentrations.

Wang et al. (2013) studied the temporal stability of land-use regression models in their investigation of traffic-generated air pollution in Metro Vancouver. The researchers applied LUR to predict nitrogen oxide and nitrogen dioxide levels—first in 2003, and then in 2010 (in the same location)—to compare the performance of the two models. The researchers reported that both models were found to be highly correlated, with a correlation factor of 0.87 for NO and 0.74 for NO_2. Further, Wang et al. determined that the models were temporally stable, showing consistent spatial variation over the two years: $R^2 = 0.59$ and $R^2 = 0.58$ for NO; $R^2 = 0.52$ and $R^2 = 0.63$ for NO_2, in 2003 and in 2010, respectively. Beelen et al. (2013) utilized LUR models to determine the spatial variations in yearly average NO_2 and NO_X concentrations in 36 study areas in Europe, reporting significant differences between the concentrations for the areas studied. Specifically, the median of R^2 of the LUR models was 82% for NO_2 and 78% for NO_X.

Other researchers have also used LUR models to define correlations between pollutant concentrations. For example, Abernethy et al. (2013) examined the correlations between nitrogen oxide and particle number concentration, determining an average of 0.65. Furthermore, a Hong Kong-based case study conducted by Lee et al. (2017) indicated that LUR modeling is particularly effective in predicting the potential spatial variability of air pollution in high-rise, high-density cities. Moreover, their findings regarding the spatial variation of air pollution concentration in Hong Kong were consistent with other results describing emissions data for the city. Lee et al. concluded that LUR modeling is particularly suitable for high-density cities due to the fact that substantial spatial variations in pollution exposure usually occurs in such locations.

Finally, Johnson et al. (2010) argued that land-use regression modeling is one of the most accurate strategies for estimating human health effects resulting from exposure to urban air pollutions, principally because it makes it easier to effectively estimate individual exposure

levels to ambient air pollution. However, the researchers noted that LUR models feature a number of limitations when it comes to extensive air pollution monitoring, as well as the applicability of their data to other different locations.

2.3 The Impact of Independent Variables on Model Performance

Land use regression for developing hourly empirical models requires detailed exposure measurements to avoid exposure misclassification. Therefore, incorporating time-activity factors to the land use model can enhance prediction modeling accuracy. The amount of explained variation has been found to be dependent on the independent variables used in modeling studies (Shaban et al., 2016); specifically, the authors indicated that modeling using more independent variables tended to yield better forecasting performance. In their study, the root mean square error (RMSE) of an SO_2 prediction model decreased from 62.4 for the univariate modeling to 31.4 for the multivariate modeling.

Wu et al. (2015) demonstrated that changes in the concentration of particulate matter ($PM_{2.5}$) were more evident over time in comparison to spatial changes. However, a recent report indicated that the coefficients for temporal variables are similar to those for spatial variables, which means that both variables are important in developing short-term modeling approaches (Hankey et al., 2019). According to Chen et al. (2019), satellite observations represent the most accurate approach for modeling particulate matter ($PM_{2.5}$) levels; in contrast, when undertaking nitrogen dioxide modeling, traffic variables were found to be highly relevant. Indeed, Beleen et al. (2013) reported that traffic-intensity data were crucial in modeling nitrogen dioxide and nitrogen oxide; specifically, their study showed that there was a 10% drop in the explanation of variance when local traffic intensity was not used in model development.

An innovative solution used buffers that represent catchment areas based on meteorology as an alternative to typical circular buffers. However, models that included wind rose data were unable to describe any additional variation in particulate number measurements (Abernethy et al., 2013). Conversely, another study showed that the model fit better when wind speed and direction were included as variables (Bertazzon et al., 2021). Specifically, Bertazzon and colleagues indicated that when the model included meteorological variables (e.g., wind speed), R^2 findings improved an average of 15% compared to conventional models. Additionally, their model showed that population density and industrial emissions were highly correlated with NO_2 concentration; indeed, these variables were found to be the most critical variables for predicting NO_2 concentrations.

2.4 Air Pollution Monitoring Campaign

Increasingly, the concentration of air pollution in a given area has been investigated using mobile air quality monitoring tools instead of focusing a fixed site, which allow for more accurate and comprehensive assessments of air quality (Hasenfratz et al., 2015). Mobile monitoring can be transitioned around various sites, thereby facilitating an assessment of air quality in an area by utilizing temporary sites. Typically, air quality has been measured using fixed-site platforms (Messier et al., 2018). As such, multiple sites are required to develop a comprehensive view of how air quality tends to vary in a particular region on a spatial and temporal scale. Although the fixed-site approach can be effective, it is essential that the researcher understand the data's constraints and magnitude of representativeness (Hankey & Marshall, 2015). Air quality patterns vary widely in terms of space (e.g., densely populated cities versus less urbanized locations), local weather patterns, and time (e.g., intermittent emission cycles). Therefore, a fixed network may not be able to take into account these variabilities.

Kerckhoffs et al. (2016) compared prediction estimates from mobile and short-term stationary land-use regression models. Using both mobile and stationary monitoring data, the researchers confirmed the strong correlation between ultrafine particle and black carbon concentration surfaces predicted by LUR models. However, the researchers also reported that predicted concentrations based on mobile measurements were consistently higher.

In some cases, a system may already exist in the proper location, but it could be restricted or have gaps that need to be addressed (Baets, 2019). Kerckhoffs et al. (2017) investigated the effects of monitoring type, comparing the performance of LUR with respect to both stationary and mobile monitoring of ultrafine particle (UFP) and BC concentrations. For ultrafine particles, the authors stated,

> . . . [the ability of] UFP models to predict measurements with longer averaging time increased substantially from 36% for short-term stationary measurements to 57% for home outdoor measurements. In contrast, the mobile BC model only predicted 14% of the variation in the short-term stationary sites and also 14% of the home outdoor sites. (p. 500)

Furthermore, Hasenfratz et al. (2015) investigated the extent to which a BC mobile data prediction model could explain concentration levels, concluding that land-use regression on mobile data measurement could explain 68% of BC variability.

Intra-urban air pollution is distinguished by high geographical diversity of contaminants, as well as rapid destruction from the source (Su et al., 2015a). Sulfur concentrations, for example, have been shown to decrease by 50% between 50 and 150 meters from a highway (Hasenfratz et al., 2015). Nitrogen dioxide levels have been shown to be approximately 2.5 times higher with 50m of the source compared to beyond that range Within 50 m, nitrogen dioxide

levels have been shown to have approximately two and a half-fold (Messier et al., 2018). (Hankey & Marshall, 2015)reported that ultrafine particles (UFPs) have been discovered to have a substantial impact on the local environment up to 300 meters from highways. In contrast to mobile environmental monitoring, small-scale spatial differences in air pollution concentrations cannot be detected using traditional fixed-location approach techniques.

These are just a few examples of how mobile measurement can help to provide a fuller understanding of air quality. It should also be noted that mobile monitoring represents a more cost-effective assessment approach in comparison to fixed monitoring. Specifically, fixed sites tend to require a variety of components to work—notably power connections, launching, and security procedures—all of which can incur fixed costs and are difficult to relocate (Su et al., 2015a). In contrast, mobile environmental monitoring entails deploying instruments in a temporary location for a short period of time before relocating them to another location. Instruments are frequently placed in or on a vehicle (Hasenfratz et al., 2015), a bicycle (Hankey & Marshall, 2015; Messier et al., 2018), a robotic device (Reggente et al., 2010), or backpack (Baets, 2019). Advances in air quality instrumentation, such as higher spatial resolution and portability, have enabled these platforms. However, given the need to fully monitor and understand an air quality-related incident, constructing a permanent site that will be operational indefinitely may simply not be appropriate or affordable.

2.5 Machine Learning

Nowadays, the costs associated with obtaining mobile monitoring data from an airborne device have decreased. Coupled with the growing availability of environmental data, the number of pollution-related databases available for study has exploded. These large datasets are complicated due to mixed degrees of data dependencies, making it challenging to examine them

using traditional regression models. As a result, developing novel and improved methods of analysis are essential for building an empirical model. Machine learning and data mining methods that have emerged from the computer science domain are providing reliable and scalable methods that have performed well in analyzing big data. Furthermore, it is essential to examine the ability of machine learning to handle the variability that exists in the mobile monitoring data.

Machine learning is becoming increasingly important for epidemiological and environmental research targeting levels of indoor and outdoor air pollution. There are three basic areas of research for which machine learning algorithms have been applied: source apportionment, air pollution concentration forecasting, and hypothesis generation (Bellinger et al., 2017). Furthermore, four algorithm types have been applied to regression and classification, clustering, and outlier detection (Bellinger et al.). Classification and regression have been used for prediction and forecasting purposes, while clustering and association mining are more common for testing hypotheses and source apportionment. One study confirmed that a clustering algorithm can be used to improve the accuracy of source apportionment, which not only divides all of the sample data into distinct groups based on pollution characteristics (e.g., pollution source and concentration), but also identifies outliers at the same time, which enhances the task of recognizing and interpreting pollution sources (Chen et al., 2015).

Analyzing and formalizing the research objective is the first step in data mining in order to implement the appropriate learning algorithm paradigm. Toward that end, there are several factors that the user should consider when applying machine learning. While there are many different types of machine learning, Domingos (2012) focused on the most widely used type: classification; he also stressed that learning must entail representation, evaluation, and

optimization. Bellinger et al. (2017) pointed out that in order for a machine learning algorithm to perform optimally, one must apply different sets of algorithms from the paradigm of appropriate approaches. By stressing a different approach, researchers have demonstrated that applying a diverse set of models to form an ensemble of estimators is an effective solution in both theory and practice (Xi et al., 2015).

Cleaning data represents a crucial step in any science data endeavor. When data is collected, it is common for some missing values to be present in the collection. However, advanced machine learning has demonstrated its ability to handle missing data in predicting daily concentrations of PM. For instance, the use of the geographically-weighted gradient boosting machine learning evidenced 76% of the concentration variance even in the absence of partial aerosol optical depth data (Zhan et al., 2017). (Hu et al., 2017) compared the output from seven machine learning algorithms applied in mobile and site monitoring data to predict carbon monoxide concentrations. Their results confirmed that the Random Forest (RF) algorithm has the ability to generate predictions with high accuracy and interpretability.

Twenty-five classification algorithms relied on a fixed-site particle indicator, vehicle, and meteorological data to demonstrate that they can predict submicron particle levels accurately; indeed, the research shows that tree-based classification models provide the most powerful performance outcomes. It must be noted, however, that meteorological variables also play an important role in forecasting $PM_{1.0}$ and UFP levels. The most common algorithms that have been utilized in the area of air pollution epidemiology are Decision trees, K-means, Bayesian, Support Vector Machines, and Artificial Neural Networks; typically, their use can be classified as prediction-based or knowledge discovery. (Kleine Deters et al., 2017) evaluated six years of temporal data (meteorological data) and air pollution measurements to predict the concentrations

of $PM_{2.5}$ based on precipitation levels and wind (speed and direction). Boosted Trees (BTs), Neural Networks (NN), and Linear Support Vector Machines (L-SVM) classification algorithms were applied to identify the concentration of $PM_{2.5}$ using a three-class classification system (low, moderate, and high), which was derived using three basic meteorological parameters: wind direction, wind speed, and precipitation level. The results showed that the classification algorithms could perform well up to 20 $\mu g/m^3$ of $PM_{2.5}$; further, Kleine Deters and coworkers demonstrated that additional meteorological parameters should be used when $PM_{2.5}$ levels increased. Xi et al. (2015) combined numeric model -WRF Chem- and a classification machine learning algorithm in order to identify the most powerful model for each of 70 different cities in China. Their findings indicate that a combined model is better than a singular model, and that the use of more features will likely enhance the accuracy of the model.

Various modeling strategies have recently been developed in order to overcome the shortcomings of LUR in capturing non-linear interactions between contaminants and predictors. Meanwhile, several researchers have recently been collecting stationary data at various geographic scales in order to compare the performance of machine learning approaches with LUR. In a recent study conducted in Uganda, researchers applied eight different machine learning algorithms to predict $PM_{2.5}$ concentrations, demonstrating that the non-parametric machine learning approaches to LUR modeling were clearly superior to ordinary least squares regression methodology (Coker et al., 2021).

Another study applied machine learning to map asthma-prone areas in Tehran (Iran) to define the impact of six air pollution factors (CO, PM_{10}, $PM_{2.5}$, NO_2, SO_2, and O_3) across the different seasons (Razavi-Termeh et al., 2021). The researchers determined that the RM model exhibited adequate performance. Hu et al. (2017) conducted a study to compare the performance

between various machine learning algorithms in predicting air pollution concentrations. Their findings indicate that support vector regression was able to define the polluted area boundary, adding that it performed as well as random forest and decision tree regression, and better than extreme gradient boosting, adaptive boosting regression, and multi-layer perceptron. Another group of researchers used machine learning to determine the degree to which air pollution was linked to early cognitive skills among children in the U.S. (Stingone et al., 2017). Using data from 6900 children, the researchers concluded that children in urban and highly populated urban areas were generally at greater risk for learning deficits.

According to a study comparing linear regression and a machine learning approach for ambient ultrafine particles, researchers found that standard multivariable regression represented 62% of the spatial variation of ultrafine particles, while the machine learning algorithm explained 79% of the variance (Weichenthal et al., 2016). A more recent study used stationary and non-stationary ultrafine particle data obtained in the Netherlands to evaluate linear and non-linear regression; the authors demonstrated that performance deteriorated when external spatial validation was applied (Kerchhoffs et al., 2019).

Brokamp and coworkers (2017) also studied whether machine learning was superior to LUR models in predicting particle matter concentration. Their comparison of regression and random forest approaches showed that the prediction error decreased 15% when using RF over traditional land use regression. In their study, Lim et al. (2019) utilized three algorithms—linear regression, stacked ensemble (SE), and random forest (RF)—toward investigating and mapping street-level particulate matter $PM_{2.5}$ in Seoul, South Korea. Their results showed that two machine learning models (RF and SE) demonstrated higher coefficient of determination ($R^2 = 0.73$ and 0.8, respectivily). In their effort to develop Europe-wide spatial models of fine

particulate matter and nitrogen dioxide, Chen et al. (2019) found that using machine learning to develop annual average particle matter and nitrogen dioxide concentrations was not significantly different compared to linear stepwise regression and regularization modeling.

In conclusion, researchers have demonstrated that, in general, machine learning models outperform LUR models when assessing the same contaminants (Wang et al., 2020). However, the performance of land-use regression remains stable when the data size is relatively small, which is not the case with machine learning. The performance of the machine learning was found to decrease when aggregating 200 m segmentation (Wang et al.). This result confirms that machine learning algorithms perform better when larger datasets are to be considered.

2.6 Systematic Validation Approach

The performance of LUR models can be impacted by spatial autocorrelation and non-stationarity, which will increase the model error. The random leave-one out cross-validation approach revealed different model performance outcomes using the LUR model in 36 study areas in Europe (Beelen et al., 2013). The authors indicated a drop in the magnitude of the explained variance when applying cross-validation. The model $R^2$was less than 0.1 higher than the cross-validation R^2 Indicating the existence of model overestimation. Meng et al. (2015) reported that the coefficient of determination R^2 decreased from 0.82 to 0.75 when the leave-one-out-cross-validation was applied. The majority of LUR models are designed for specific cities, limiting their application to other regions (Knibbs et al., 2014).

Machine learning algorithms for use in spatio-temporal data applications that are designed to forecast a specific concentration for unknown locations enable ambient air pollution monitoring. However, the specific validation strategy is important for assessing a given model's performance to avoid model overestimation. Validation strategies have been applied differently

in spatio-temporal modeling. For example, researchers have applied both spatial validation and temporal validation to define machine learning model uncertainty of nitrogen dioxide (NO_2) prediction model (Di et al., 2020). Chen et al. (2019) applied both cross-validation and external validation to define the performance of the annual average fine particle matter and nitrogen dioxide concentration prediction models. The coefficient of determination (R^2) was found to decrease from ~0.63 to ~0.59 for fine particle concentration, and from ~0.59 to 0.50 for nitrogen dioxide. Earlier researchers have applied machine learning toward developing prediction models that have spatial and temporal structure introduce of spatio-temporal autocorrelation (Meyer et al., 2018). Therefore, there is a need to assess BC and PN concentration model uncertainty with high spatiotemporal resolution.

2.7 Chapter Summary

In conclusion, the literature review conducted for this study confirmed the deleterious health effects of air pollution across the human lifespan, but notably for younger individuals during the development of their respiratory systems. As shown in the literature, land use regression represents a powerful technique for mapping the spatial distribution of pollutant concentrations. Available reports also indicate that modeling efforts could be improved by taking into account more specific pollutant-related data (e.g., meteorological variables). It must be cautioned, however, that while software advances modeling vehicle emissions factors can improve the accuracy of prediction models, adding additional data points to the model will increase the chance for model overfitting. Therefore, there is a need to overcome this obstacle while still incorporating variables that will increase the model's predictive capacity. Furthermore, the development of the monitoring platform led to an increase in data available for

modeling. There is also a need to assess how enhanced data processing can be applied using land use regression methods.

The literature confirmed the potentially powerful use of machine learning to provide more accurate prediction models, particular in the case of large, noisy datasets. Nonetheless, there is a pressing need to further assess how machine learning modeling can be applied in presence of a growing range of available data. Researchers have also applied random cross validation to validate the land use model, which represents a powerful technique for decreasing model overestimation. However, the literature shows the existence of spatio-temporal autocorrelation within air pollution data. To address this issue, researchers have used machine learning in data with defined spatial and temporal characteristics, thus reinforcing the need to improve validation strategies to define more reliable model. As a possible approach for addressing these issues, this study incorporated both spatial and spatio-temporal cross validation to improve the model validation process and to better assess model overestimation.

Chapter 3: Methodology

For this study, measurements of particulate air pollution were collected via mobile monitoring within a small rural college town (Blacksburg, VA; Population: 181,863) during the daylight hours (7 am to 7 pm) during the summer and fall of 2016. These data were then used to develop several machine learning models to test the ability of ML to handle noisy, unprocessed mobile monitoring data vs. using spatially and temporally smoothed inputs. I focused on comparing how model derived short-term (Hourly_average: hourly aggregated data spatially 100 m and temporally 1 hour, and One_second: the disaggregated raw data) concentration estimates compare to long-term (Daily_average model: 12-hour average) estimates to evaluate how the design of a short-term modeling approach affects exposure assessment. Additionally, I also compared the performance of these models to the models produced by stepwise regression model to allow the comparison of performance of the machine learning models to analogous data obtained from traditional stepwise regression models. I used three types of cross-validation in the short-term models (random, spatial, and spatial-temporal cross validation) to assess the performance of the reliable model in order to forecast the pollutant concentration in unseen location. All the models were built using Python 3.7 and scikit-learn library and implemented in Advanced Research Computing at Virginia Tech.

3.1. Data Collection

The bicycle-based mobile measurements of Particle Number (PN) and Black Carbon (BC) concentrations in Blacksburg, VA, were used to develop empirical machine learning models. PN and BC are pollutants that are tracers of traffic emissions (Thurston et al., 2016), have a high degree of spatial variability in urban areas (Abernethy et al., 2013), and are linked to a range of health disparities (Aguilera et al., 2016).

Short-term mobile monitoring data were collected for a prior study (Hankey et al., 2019), and the details for that data collection campaign can be found there. In summary, however, a mobile monitoring platform was set up during the summer and fall of 2016. Microaethalometers (AE51; AethLabs) mounted on bicycles were used to measure BC concentrations, and condensation particle counters (CPC 3007; TSI, Inc.) were used to measure PN concentrations.

This monitoring effort was undertaken with four goals in mind: (a) to test the ability of noisy, unprocessed mobile monitoring data to be used in the development of empirical machine learning models capable of estimating short-term concentrations; (b) to test the impact of using the vehicle emissions factors collected from INTEGRATION software simulation to assess the model's performance; (c) to compare the performance of random, spatial, and spatial temporal CV models; and (d) to introduce a systematic validation approach which could help to improve model validation strategies. In summary, this research applied these different pre-processing approaches using secondary data collected via monitoring (Hankey et al. 2019), coupled with testing different input variables, to compare the performance of the machine learning models to analogous data obtained from mobile monitoring measurements.

(Hankey et al., 2019) experimented with various methods for adjusting mobile monitoring data for black carbon (BC) and particle number (PN) background concentrations in Blacksburg. For model development, I used concentrations adjusted by the multiplicative method, which attempts to best approximate long-term air pollution concentrations. In brief, the multiplicative background adjustment method computed adjustment factors according to the ratio of daily concentration to hourly concentration at a central site where background concentrations were measured. Then, the adjustment factors were multiplied to all mobile monitoring

observations (based on the hour the mobile measurements were collected) to account for differences in background concentrations.

3.2. Emission Factors Modelling

Land use regression modeling incorporating land-use data and meteorological factors was already well defined (Pandey et al., 2013b). However, this research was designed to integrate an emissions models with high spatial and temporal resolution. Therefore, one spatiotemporal variable, emissions, was used as an input variable. There are several sources of ambient air pollution emissions (e.g., power plants, restaurants, and vehicles emissions). In this study, emissions from light-duty vehicles were calculated using the INTEGRATION software (a software being developed in the Center for Sustainable Mobility). These emissions data were used as the spatiotemporal independent input variables.

The emissions discussed in this study include $PM_{2.5}$, black carbon, and the particle number measured by condensation nucleus counter (CNC) and optical particle counter (OPC) respectively. CNC measures the particle number concentrations for particles with diameters of larger than 0.01 μm, and OPC counts and optically sizes particles with diameters between 0.1 and 2 μm.

The four emissions are all fuel-based pollutant emissions that can be estimated by relating total carbon emissions to the carbon content of fuel using equation (3.1), where E_p is the emissions factor indicating how many grams of emissions or how many particles produced per unit mass of the fuel, in the unit of g/g for $PM_{2.5}$ and black carbon and in $\#/g$ for CNC and OPC, ΔP is the emission concentration in $\mu g/m^3$ for $PM_{2.5}$ and black carbon and in $\#/cm^3$ for CNC and OPC, $\Delta CO_2 + \Delta CO$ are carbon emission concentration of the fuel in ppm, w_c is mass

fraction of carbon in fuel (0.85 for gasoline), and C is the constant for unit conversion with 10^{-6} for $PM_{2.5}$ and black carbon and 10^{6} for CNC and OPC.

$$E_p = \frac{\Delta P}{\Delta CO_2 + \Delta CO} w_c C \qquad 3.1$$

Using equation (3.1), one can easily capture the pollutant emission footprint with the availability of fuel consumption. In this study, the emission factors were derived based on engine performance parameters (i.e., speed and acceleration) in a lab. Specifically, the INTEGRATION software simulates traffic conditions, notably the speed and acceleration of vehicles on the network, to generate fuel consumption data, which were then used in equation (3.1) to model emissions factors. Table 3.1 presents an example of the four emissions estimates for light-duty vehicles, in which the emission factors for CO, CO_2, $PM_{2.5}$, Black Carbon, CNC, and OPC were provided by the literature (Kirchstetter et al. 1999). It should be noted that the concentration data shown in this table were collected over a three-hour time span over the course of four days; emission factors were calculated for each day using equation (3.1). The final emission factors were then averaged over 4 days. The emissions were then tabulated in a 100-meter grid on an hourly average period to use these emissions as input variables to our machine learning model.

Table 3.1

Emission Concentrations and Factors

	Emission Concentrations						Emission Factors			
	CO(ppm)	CO_2(ppm)	$PM_{2.5}$ ($\mu g/m^3$)	Black carbon ($\mu g/m^3$)	CNC (#/cm^3)	OPC (#/cm^3)	$PM_{2.5}$	Black carbon	CNC	OPC
Day1	27.5	1008	56.1	15.5	2.10E+05	5.70E+03	4.61E-08	1.27E-08	1.72E+08	4.68E+06
Day2	26.1	946	52.5	12.1	1.90E+05	5.10E+03	4.59E-08	1.06E-08	1.66E+08	4.46E+06
Day3	27.5	1053	56.6	16.2	1.80E+05	5.60E+03	4.45E-08	1.27E-08	1.42E+08	4.41E+06
Day4	27.6	1090	53.7	16.8	1.60E+05	5.70E+03	4.08E-08	1.28E-08	1.22E+08	4.34E+06
Average emission factors							4.43E-08	1.22E-08	1.50E+08	4.47E+06

Note: Data obtained from Jianhe Du, VTTI lab, not yet published.

3.3. Modeling Approach

Previous research indicates that using more than one machine learning model is superior to using a single model (Xi et al., 2015). Thus, I employed nine machine learning algorithms based on the size of the datasets: Decision Tree, Extra Tree, KNeighbors, Support Vector Machine, Ridge, ElasticNet, Lasso, Random Forest, and Gradient Boosting. This approach assisted in the development of three sets of machine learning models to compare the short- and long-term concentrations calculated from the models developed using mobile monitoring data. As noted, three models were developed: (a) One_second models that included the raw disaggregated data; and (b) two spatially and temporally aggregated models—namely, Daily_average and Hourly average. For spatial aggregation, the mobile monitoring measurements were first spatially aggregated at 100-meter intervals along the monitoring path, after which the median of concentration measurements was determined at each aggregated location. This aggregation approach was applied for each hour of the day starting at 7:00am and ending at 7:00pm. For the Daily_average models, all mobile monitoring data temporally aggregated across all hours (7:00am to 7:00pm) for each location. For the Hourly_average models, all mobile measurement data were re-aggregated for each hour of the day for each aggregation location, after which the hour of the day was added as candidate dummy independent variable. Machine learning has the advantage of being accurate even when data is collected in the absence of a carefully controlled experimental design or when tasked to assess complicated nonlinear interactions. The aim of One_second models was to test the ability of machine learning to capture all patterns without the need for data preprocessing, which is typically required using conventional statistical models.

Three types of input variables were incorporated into the model-building process: (a) spatial component (land use, transportation, and natural environment); (b) temporal component (meteorological parameters and hour of the day); and (c) spatiotemporal component (emissions data). The hour of the day was added as a dummy variable (6−7 pm represents referent data). The meteorological variables for hourly temporal resolution were obtained from the Real-Time Mesoscale Analysis database. The emissions factors were generated from INTEGRATION, which provides simulated fuel output as the feed of the emission model. In this study, the effects of adding spatiotemporal variables to the model were assessed. According to (Xi et al., 2015), the more variables used to build the model, the greater the possibility of improving accuracy. It must be noted, however, that the addition of more variables will increase the model's potential for overfitting. Thus, for the One_second and Hourly_average models, I developed seven types of models based on the input variables fed into the model: (a) land use only; (b) weather only, (c) land use and weather; (d) weather and emissions, land use, weather, and emissions, land use, weather, and hour of the day, and (g) all variables and for the Daily_average model only the land-use data (Table 3.2). The set of algorithms applied to each dataset varied. For example, I did not employ either KNeighboor or SVM in the models featuring all the input variables—principally because they are very computationally expensive to train using large datasets. I used three types of cross validation in model validation process: random, spatial, spatial-temporal CV. I applied random and spatial CV in the Daily_average and Hourly_average models for both pollutants. On the One_second models, I applied the random CV and spatial CV, and the spatial temporal CV (Table 3.2).

Table 3.2.

Summary of Variables Used and CV Applies to Develop the Daily_average, One_second, and Hourly_average Empirical Models.

Pullutant	LUR Models	Input Variables	Temporal resolution	Number of observations	Input variables				Cross_validation		
					Land use, transportation, natural envirnoment	Weather	Emission	Hour of the day	Random_holdout	Spatial_holdout	Spatial_temporal_hold out
BC: µg/m^3 , PN: pt /cm³	Daily_ average	Lu	12 hr	423	☑				☑	☑	
	One_second	Lu	1 s	BC: 319,489- PN: 354,715	☑				☑	☑	☑
		W			☑	☑			☑	☑	☑
		Lu_W			☑	☑	☑		☑	☑	☑
		W_Em			☑	☑	☑	☑	☑	☑	☑
		Lu_W_Em			☑	☑		☑	☑	☑	☑
		Lu_W_Hr				☑			☑	☑	☑
		Lu_W_Em_Hr				☑	☑		☑	☑	☑
	Hourly_average	Lu	1 hr	5074	☑				☑	☑	
		W			☑	☑			☑	☑	
		Lu_W			☑	☑	☑		☑	☑	
		W_Em			☑	☑	☑	☑	☑	☑	
		Lu_W_Em			☑	☑		☑	☑	☑	
		Lu_W_Hr				☑			☑	☑	
		Lu_W_Em_Hr				☑	☑		☑	☑	

3.3.1. Daily_Average Models

I developed models designed to pool mobile measurement data obtained throughout the day to compare strategies for estimating long-term averages. I spatially aggregated all mobile monitoring data across all hours (7 am to 7 pm) as an input to our models. The resulting models show the median concentrations throughout the day. The Daily_average model was developed based on land-use data only. By aggregating data across all hours of the day, the mobile measurement and data aggregation approach was intended to remove temporal variables (Hankey et al., 2019).

3.3.2. One_Second Models.

I developed the one_second model using raw disaggregated measurement data. This approach facilitated the prediction of pollutants across small temporal resolution to determine the ability of machine learning to build an empirical model using raw data. I built seven types of One_second models using different combinations of input variables: (a) One_second_Lu (model based on land-use data only); (b) One_second_W (model based on weather data only); (c) One_second_Lu_W (model based on the land use and weather data; (d) One_second_W_Em (model based on weather and emissions data); (e) One_second_Lu_W_Em (model based on land use, weather, and emissions data); (f) One_second_Lu_W_Hr (model based on land use, weather, and hourly data); and (g) One_second_Lu_W_Em_Hr (model based on incorporating all independent variables). Applying various combinations of input variables was intended to help determine the efficacy of ML model performance.

3.3.3. Hourly_Average Models.

I developed the same seven types for the One_second models: Hourly_average_Lu, Hourly_average_W, Hourly_average_Lu_W, Hourly_average_W_Em,

Hourly_average_Lu_W_Em, Hourly_average_ Lu_W_Hr, Hourly_average_ Lu_W_Em_Hr. This approach facilitated the comparison of the short-term models to determine the ability of ML to handle the noisy, unprocessed mobile monitoring data model (One_second model) vs. spatially and temporally smoothed data models (Daytime and Hourly_average model).

3.4. Modeling Implementation

All the regression models were implemented using Python and scikit-learn (Pedregosa et al., 2011), which is an open-source machine learning library for the Python programming language. All hyper-parameters were tuned using the random 10-fold cross-validation method and the GridSearchCV function on the model training process. I defined the estimator/algorithm type and the sequence of hyper-parameter. I then set the negative mean absolute error as a scoring method as an input to the GridSearch function. The data were standardized prior to applying the algorithm. I then validated the models using three types of cross-validation on the test data, which is described more fully in Section 3.5. Grid-SearchCV is a function used to define the optimal hyper-parameter of the model; it can exhaustively search over specified parameter values defined by the user and automatically detect the hyper-parameter that provides an accurate prediction.

3.5. Model Validation

Three types of k-fold cross-validation strategies were applied according to the type of dataset: (a) randomly split up into 10 groups; (b) spatially split up into 17 spatial clusters based on several spatial sets defined by k-mean clustering using spatial variables (land use, transportation); and (c) spatially-temporally split up into 17 clusters using the spatial set and the run number (detailed in Section 3.5.3). For the k-fold cross validation, the quality and reliability of the models were measured using mean absolute error (MAE) and root mean squared error

(RMSE); this approach enabled comparisons across different model sets and for the different number of variables for the different cross-validation approaches. Comparisons were conducted to determine the best performance algorithms for each model.

In order to compare the robustness and suitability of the machine learning model with traditional statically baesed models, the coefficient R^2 for both Daily_average and Hourly_average Random CV models were determined. Additionally, random cross validation was applied for all model types (Daily_average, Hourly_average, and One_second) to assess the model overestimation. I applied three strategies of CV for the One_second models (random, spatial, and spatial-temporal CV) to facilitate comparing the performance of the spatial and spatial temporal CV to the random CV using very small temporal resolution data. It should also be noted that I did not apply the spatial-temporal CV in Daily_average and Hourly average models—mainly due to the fact than when we increased the restriction of the hold-out test data we lost observations in the test set. Thus, any test errors should not be considered to be representative.

3.5.1 Random Cross-Validation.

Random cross-validation is intended to determine a model's performance at a random time and location. For this study, random CV was calculated as follows. First, air pollution measurements were randomly split into 10 groups. Then, I selected one group as the hold-out group, trained the model using the remaining available data, and then predicted air pollution concentrations for the hold-out group. This process was repeated separately for each of the 10 groups.

3.5.2 Spatial Cross-Validation.

Spatial cross-validation is used to determine a model's performance for areas that do not share the same spatial characteristics. Spatial cross validation was conducted in a similar fashion to random CV, except for the fact that 17 groups were identified as spatial clusters. The data were split to train/test the dataset based on spatial clusters defined previously using K-mean clustering. I applied GridSearch's random 10-fold cross validation in the training dataset to tune the hyper-parameters. Then, the defined model was fit in the test dataset to define the prediction values and the evaluation matrix. This process was repeated 17 times. Thus, the resulting evaluation matrix was calculated based on averaging the predicted values from each iteration.

With Blacksburg (VA) being a smaller town with somewhat similar spatial features, splitting the datasets based on specific locations within Blacksburg (e.g., according to census tract id number), would not be representative. Instead, I chose to use spatial variables (land use and transportation) to segment observational data and define the resulting spatial-split performance. For each set of models (Daily_average, One_second, and Hourly_average), I clustered the observations based on spatial data using k-means clustering. Unfortunately, no general theoretical solution exists for determining the optimal number of clusters for any given dataset. As such, a straightforward approach is to compare the results of multiple runs with different k classes and select the optimal one based on a predefined criterion. Still, we must exercise caution—not only because k-means clustering can result in smaller error function values, but also because there is a greater risk of overfitting. To overcome this possibility, I first reduced the number of spatial variables to define the most representative variables, and then applied the elbow method (or elbow clustering). The elbow method is based on determining the optimal number of clusters by training the model with a range of different numbers of clusters,

and then defining the elbow indicating the model's best fit. Clustering is a "brute force" technique that requires the computational power of computers to conduct the analysis. To overcome this restriction, I first applied the recursive feature elimination cross-validation RFECV techniques, which identifies several spatial variables that fit a model and removes the weakness features until the most representative feature number is reached. I then applied the "RFECV" provided within the "yellowbrick" library in python using lasso regression. However, prior to this step, I first removed all features that had a correlation coefficient above 0.8 (i.e., the most highly correlated features) from the data using a correlation matrix and removed. Then, I defined the best alpha using the "LassoCV" from the same library, after which I used the variables defined by the RFECV to "KElbowVisualizer" from the same library to define the optimal number of clusters. Finally, I applied the "KMeans" module from the scikit-learn (Pedregosa et al., 2011) library using the most representative variables and the number of defined features.

3.5.3. Spatial-Temporal Cross-Validation.

To cluster the data, I used the spatial clusters defined previously and the number of runs for each hour of the day. Specifically, for each hour of day, I conducted at least five runs for each route. For each spatial cluster, the final run of each hour for each route was eliminated. Then, the data from the same spatial and temporal cluster was removed from the training data set during the model training process. I applied the same process used for spatial clustering; this process was repeated 17 times.

Chapter 4: Results and Discussion

As detailed earlier, I implemented three time-interval models (two short-term models and one long-term model) to estimate pollution concentration involving two pollutants in Blacksburg using mobile-monitoring data. Machine learning algorithms are known to be effective in evaluating multi-dimensional and multi-variety data, and they can do so in complex or uncertain situations. Therefore, it is essential to ascertain the degree to which machine learning is effective in building prediction models using large datasets of mobile monitoring data—especially when data are used without data preprocessing. Accordingly, I varied the input variables for the short-term models. Since many machine learning models were assessed, I focused on evaluating trends that emerged from varying specific parameters, and then explored any differences between the short-terms models to help define how well machine learning algorithms can derive models using row mobile monitoring data.

This section focuses on trends in model performance when temporal and spatial variables were applied to the mobile data models to determine the effects of different input variables. This section also presents the observed trends from the time-term models from the refined one_second measurements model (disaggregated data), the one-hour term model (temporally aggregated 1 hr), and the long-term aggregated model (12-hr aggregated model). In addition, three types of validation approaches were used to investigate model overestimation by comparing the oriented-validation approach with random cross-validation models.

Based on the fact that some machine learning models require the dependent variables to be similarly scaled, I standardized the independent variables. I also log transformed all mobile measurement-concentration data to fit a normal distribution more closely for modeling as a model input. It should be noted that a small number of black carbon concentration estimates

evidenced zero or negative values due to noise in the micro-aethalometer results. To overcome this issue, I add 1 μg/m³ to the black carbon BC concentration in the model training process; all model results were then transformed back to their original values for estimating and reporting model performance statistics (e.g., error).

Mobile measurement concentrations were averaged, which were determined to be 10,293 pt /cm³ for particle number PN and 1.08 μg/m³ for BC (Table 4.1). For all monitoring runs, the average temperature was maintained at 20.8 °C; the relative humidity was kept at 69 percent relative humidity, and wind speed was 1.6 m/s.

Table 4.1.

Summary of Pollutant Descriptive Statistics

Pollutant	Models	Count	Mean	Median	S.D.	Min	Q1	Q3	Max
Black Carbon (μg/m³)	Daily_average	423	0.67	0.62	0.27	0.17	0.47	0.81	1.63
	Hourly_average	5074	0.74	0.63	0.59	-0.64	0.41	0.95	16.76
	One_second	319489	1.08	0.69	2.49	-6.56	0.34	1.21	108.01
Particle Number (pt /cm³)	Daily_average	423	6388	5924	1491	4305	5430	6925	14382
	Hourly_average	5074	6834	6044	3205	1104	4996	7679	41281
	One_second	354715	10293	5950	30139	4	3420	11338	4447494

4.1. Importance of Adding Spatiotemporal Variables on Mobile Monitoring Models

Improving the spatial modeling of traffic-related air pollution by adding spatial variables and temporal factors has been explored. Nonetheless, less is known about the degree to which land-use regression models can be enhanced by using spatiotemporal variables (e.g., vehicle emissions data) on machine learning models that incorporate mobile monitoring measurements. Therefore, this section presents comparative performance data using models that have been fed with different variables with the goal of determining any variable-driven improvements. The

impact of incorporating different sets of variables was investigating by comparing the errors in each short-term model. Such comparisons were applied in each cross-validation approach model.

4.1.1. Random Cross-Validation Results

Random CV was applied to assess any performance differences between short-term and long-term models and to quantify the overfitting by comparing it with spatial and spatial-temporal CV data. The One_second and Hourly_average models were constructed using seven different input variables. The random CV models highlight the importance of the input variables, as indicated by improvements shown in a model's performance according to each variable component (see Table 4.2).

Table 4.2.

Summary of Model Performance for Each Pollutant and Model Type in the Random CV

Models	**Input variables**	**BC**			**PN**		
		Best ML algorithm	MAE (μg/m³)	RMSE (μg/m³)	Best ML algorithm	MAE (pt /cm³)	RMSE (pt /cm³)
Daily_average	Lu	Gradient Boosting	0.11	0.15	Ridge	515.21	765.48
One_second	Lu	Gradient Boosting	0.78	2.35	Random Forest	6590.91	31280.53
	W	Decision Tree	0.78	2.35	Decision Tree	4938.55	30793.68
	Lu_W	Gradient Boosting	0.43	1.16	Gradient Boosting	3370.64	28343.28
	W_Em	Random Forest	0.74	2.29	Random Forest	5011.75	30778.15
	Lu_W_Em	Gradient Boosting	0.43	1.16	Gradient Boosting	3394.32	28709.03
	Lu_W_Hr	Gradient Boosting	0.44	1.18	Gradient Boosting	3391.27	28854.74
	Lu_W_Em_Hr	Gradient Boosting	0.41	1.07	Gradient Boosting	3328.32	28423.16
Hourly_average	Lu	Random Forest	0.30	0.42	Random Forest	1698.33	2613.62
	W	Random Forest	0.33	0.48	SVR	1699.50	2857.61
	Lu_W	Gradient Boosting	0.25	0.37	Gradient Boosting	1420.29	2319.62
	W_Em	K Neighbors	0.26	0.39	K Neighbors	1356.25	2367.64
	Lu_W_Em	Gradient Boosting	0.24	0.35	Gradient Boosting	1323.55	2155.56
	Lu_W_Hr	Gradient Boosting	0.23	0.34	Gradient Boosting	1204.00	2038.31
	Lu_W_Em_Hr	Gradient Boosting	0.23	0.34	Gradient Boosting	1216.85	2058.40

Daily_average models. The Daily_average (12-h average) models demonstrated good performance, which is consistent with previous mobile monitoring efforts (Hankey et al., 2019) (Table 4.3).

Table 4.3.

Summary of Random CV Model Performance for Daily_average Models

BC (μg/m³)			**PN (pt /cm³)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.15	0.19	Decision Tree	609.63	1085.84
Extra Tree	0.14	0.19	Extra Tree	591.31	946.74
K Neighbors	0.13	0.17	K Neighbors	505.62	865.64
SVR	0.15	0.18	SVR	589.35	841.91
Ridge	0.13	0.18	Ridge	515.21	765.48
Elastic Net	0.13	0.17	Elastic Net	568.20	978.91
Lasso	0.17	0.21	Lasso	634.36	1147.87
Random Forest	0.12	0.16	Random Forest	515.77	1002.66
Gradient Boosting	0.11	0.15	Gradient Boosting	461.64	930.17

The MAE (RMSE) of the best BC machine learning model was found to be 0.11 μg/m³ (0.15 μg/m³) for the Gradient Boosting; in contrast, the worst performance was found to be 0.17 μg/m³ (0.21 μg/m³) for the Lasso model (Table 4.3). The MAE (RMSE) of the best PN machine learning models was found to be 461.64 pt /cm³ (930.17 pt /cm³) for the Gradient Boosting; conversely, the worst performance was found to be 634.36 pt /cm³ (1147.89 pt /cm³) for the Lasso model (Table 4.3).

One_second models. Seven types of the One_second model were constructed based on the different input variables. The goal for this particular dataset (1 sec) was not only to define the model in high temporal resolution, but also to show the ability of machine learning to capture all patterns without the need for data preprocess that are usually applied in traditional statistical models. The average of the MAE (RMSE) across the seven types of the One_second best models was 0.43 μg/m³ (1.26 μg/m³) for BC, and 3703.8 pt /cm³ (28723.04 pt /cm³) for PN (Table 4.2).

Table 4.4.

Summary of Random CV Model Performance for One_second_Land_Use Models

BC (µg/m³) (One_second_Lu)			**PN (pt /cm³) (One_second_Lu)**		
Models	Test_MAE	Test_RMSE	Models	TEST_MAE	TEST_RMSE
Decision Tree	0.79	2.35	Decision Tree	6598.13	31286.34
Extra Tree	0.79	2.35	Extra Tree	6602.49	31293.71
K Neighbors	0.84	2.37	K Neighbors	6720.78	31288.53
Ridge	0.79	2.36	Ridge	6612.34	31301.13
Elastic Net	0.79	2.36	Elastic Net	6621.46	31316.65
Lasso	0.80	2.40	Lasso	6641.62	31350.75
Random Forest	0.78	2.35	Random Forest	6590.91	31280.53
Gradient Boosting	0.78	2.35	Gradient Boosting	6592.56	31282.66

The One_second model constructed for the land use (Lu) dataset corresponds to the worst performance model. The best performance MAE (RMSE) was evidenced by the Random Forest and Gradient Boosting, whereas the worst one was the K Neighbors algorithm (Table 4.4). This finding indicates that constructing this model using only raw spatial data should not be considered to be representative of estimating the pollutant concentration.

Table 4.5.

Summary of Random CV Model Performance for One_second_Weather Models

BC (µg/m³) (One_second_W)			**PN (pt /cm³) (One_second_W)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.78	2.35	Decision Tree	4938.55	30793.68
Extra Tree	0.79	2.38	Extra Tree	5546.41	30969.37
Ridge	0.79	2.40	Ridge	5925.56	31469.16
Elastic Net	0.79	2.40	Elastic Net	5892.94	31389.32
Lasso	0.80	2.41	Lasso	5873.43	31325.41
Random Forest	0.77	2.36	Random Forest	4932.29	30795.23
Gradient Boosting	0.80	2.40	Gradient Boosting	6263.65	31304.68

As indicated in Table 4.5, the best performance algorithm was the Decision Tree with MAE (RMSE) findings of 0.77 μg/m³ (2.36 μg/m³) and 4932.29 pt /cm³ (30795.23 pt /cm³) for black carbon and particle number, respectively; in contrast, the worst algorithm was found to correspond to Gradient Boosting. For particle number, the error dropped significantly from the model developed based on the land-use data compared to the model build based on weather data. This finding indicates the importance of the temporal variable (weather) in building the PN prediction model.

Table 4.6.

Summary of Random CV Model Performance for One_second_Land_Use_Weather Models

BC (μg/m³) (One_second_Lu_W)			**PN (pt /cm³) (One_second_Lu_W)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.72	1.99	Decision Tree	5187.29	30493.89
Extra Tree	0.77	2.29	Extra Tree	6151.65	30992.75
Ridge	0.75	2.34	Ridge	5778.34	31477.07
Elastic Net	0.75	2.34	Elastic Net	5740.48	31393.21
Lasso	0.77	2.39	Lasso	5732.34	31352.95
Random Forest	0.65	1.87	Random Forest	4656.67	30508.10
Gradient Boosting	0.43	1.16	Gradient Boosting	3370.64	28343.28

According to Table 4.6, the One_second model constructed for land use and weather data generated the best fit in comparison to both the model based on the weather data only, as well as the model based on land-use data only. This indicates the importance of incorporating both variables. As indicated in Table 4.6, Gradient Boosting evidenced the best algorithm for the two pollutants. The MAE (RMSE) was found to be 0.43 μg/m³ (1.16 μg/m³) for black carbon, and 3370.64 pt /cm³ (28343.28 pt /cm³) for particle number.

This enhanced reliability indicates the importance of developing a model with appropriate input variables. This finding should be taken in account when applying ML models.

The best performing ML algorithm in the One_second model constructed for weather and emissions variables was found to be for Random Forest (Table 4.7). The MAE (RMSE) was found to be 0.74 μg/m³ (2.29 μg/m³) for black carbon, and 5011.75 pt /cm³ (30778.15 pt /cm³) for particle number; conversely, the worst performing ML algorithm corresponds to Gradient Boosting. This model shows the best performance models among the One_second models, indicating the importance of including spatiotemporal variables (emissions) in the short-term data model.

Table 4.7.

Summary of Random CV Model Performance for One_second_Weather_Emissions Models

BC (μg/m³) (One_second_W_Em)			**PN (pt /cm³) (One_second_W_Em)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.77	2.36	Decision Tree	5103.15	30819.35
Extra Tree	0.78	2.39	Extra Tree	6005.28	31117.99
Ridge	0.77	2.38	Ridge	5847.02	31420.90
Elastic Net	0.77	2.38	Elastic Net	5810.91	31345.80
Lasso	0.79	2.40	Lasso	5796.69	31292.90
Random Forest	0.74	2.29	Random Forest	5011.75	30778.15
Gradient Boosting	0.79	2.39	Gradient Boosting	6269.21	31311.63

Table 4.8.

Summary of Random CV Model Performance for One_second_Land_Use_Weather_Emissions Models

BC (μg/m³) (One_second_Lu_W_Em)			**PN (pt /cm³) (One_second_Lu_W_Em)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.66	1.74	Decision Tree	5087.97	30361.74
Extra Tree	0.73	2.25	Extra Tree	5314.47	30680.13
Elastic Net	0.75	2.34	Elastic Net	5758.12	31433.53
Lasso	0.77	2.39	Lasso	5735.61	31357.27
Random Forest	0.64	1.77	Random Forest	4613.81	30456.75
Gradient Boosting	0.43	1.16	Gradient Boosting	3394.32	28709.03

Table 4.9.

Summary of Random CV Model Performance for

One_second_Land_Use_Weather_Hour_of_Day Models

BC (µg/m³) (One_second_Lu_W_Hr)			**PN (pt /cm³) (One_second_Lu_W_Hr)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.67	1.76	Decision Tree	5301.30	30673.95
Extra Tree	0.77	2.23	Extra Tree	6185.16	31109.89
Ridge	0.74	2.33	Ridge	5577.72	31101.78
Elastic Net	0.74	2.34	Elastic Net	5605.19	31136.33
Lasso	0.77	2.39	Lasso	5662.80	31197.51
Random Forest	0.64	1.87	Random Forest	4708.96	30487.32
Gradient Boosting	0.44	1.18	Gradient Boosting	3391.27	28854.74

Table 4.10.

Summary of Random CV Model Performance for

One_second_Land_Use_Weather_Emissions_Hour_of_Day Models

BC (µg/m³) (One_second_Lu_W_Em_Hr)			**PN (pt /cm³) (One_second_Lu_W_Em_Hr)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.70	1.93	Decision Tree	5595.59	30595.02
Extra Tree	0.77	2.30	Extra Tree	5859.86	30899.57
Ridge	0.74	2.33	Ridge	5570.02	31091.23
Elastic Net	0.74	2.33	Elastic Net	5603.93	31135.05
Lasso	0.77	2.39	Lasso	5662.87	31197.34
Random Forest	0.64	1.84	Random Forest	4675.68	30506.77
Gradient Boosting	0.41	1.07	Gradient Boosting	3328.32	28423.16

The One_second models developed for (a) land use, weather, and emissions; (b) land use, weather, and hour of the day; and (c) all variables were similar in that they evidenced approximately the same mean absolute error and equivalent with the model developed for land use and weather data (Tables 4.8, 4.9, and 4.10). The Average MAE (RMSE) was found to be 0.43 µg/m³ (1.14 µg/m³) for black carbon, and 3371.14 pt /cm³ (28,582.6 pt /cm³) for particle

number among the last three models. The models with Lu + W + Hr did almost as well as the Lu + W + Em; this finding suggests that it would be feasible to develop a model without the emissions factors that could be costly to implement. However, the models developed based on both emissions and weather variables only did pretty well too for areas that lack quality land-use data.

The distinction between the impact of land use and weather variables in predicting pollutant levels in short-term modeling is documented in Table 4.1, which is in line with previous mobile monitoring efforts (Hankey et al., 2019). Specifically, a combined effect of land use and other meteorological factors may play a vital role in the concentration of pollutant particles. However, the data presented therein demonstrate that model include weather and emissions variables are more indicative of pollutant concentration when compared to land use factors model. The impact of particle number, as indicated in the One_second models, revealed that short-term model cannot be developed using spatial data alone. Table 4.1 shows the robustness of the ML in predicting pollutant concentration, as evidenced by the 1 sec measurement data with only weather and emissions data. These findings are very informative and worthwhile in the context of using machine learning techniques to predict pollutant concentration from raw datasets.

Hourly_average models. Similar to the One_second models, seven types of Hourly_average models were developed, which facilitated comparison of the short-term model across different term intervals. Overall, the average of the MAE (RMSE) across the seven types of Hourly_average best models was 0.26 μg/m³ (0.39 μg/m³) for BC, and 1417 pt /cm³ (2344.4 pt /cm³) for PN.

The average MAE (RMSE) of the Hourly_average models dropped to 38.7% (69%) for BC and 61.7% (92%) for PN when compared to the average MAE (RMSE) of the One_second models, indicating that the Hourly_average aggregation data fit better than the raw data. This finding attribute to the fact that the data were spatially and temporally smoothed and the noise was removed. Indeed, smoothing techniques are extremely useful in machine learning because conditional expectations can be viewed as trends of unknown shapes that must be predicted when uncertainty parameters are present.

Table 4.11.

Summary of Random CV Model Performance for Hourly_average_Land_Use Models

BC (μg/m³) (Hourly_average_Lu)			**PN (pt /cm³) (Hourly_average_Lu)**		
Models	TEST_MAE	TEST_RMSE	Models	TEST_MAE	TEST_RMSE
Decision Tree	0.31	0.44	Decision Tree	1770.99	2681.53
Extra Tree	0.31	0.44	Extra Tree	1787.09	2665.32
K Neighbors	0.30	0.42	K Neighbors	1760.95	2613.73
Ridge	0.30	0.42	Ridge	1795.71	2668.26
Elastic Net	0.30	0.42	Elastic Net	1721.93	2635.43
Lasso	0.31	0.45	Lasso	1748.71	2733.27
Random Forest	0.30	0.42	Random Forest	1698.33	2613.62
Gradient Boosting	0.30	0.42	Gradient Boosting	1700.97	2623.10

The Hourly_average models based on land-use data only and weather data only were identified the worst performing models (Table 4.1). Based on ML algorithms applied on Hourly_average_Lu, the MAE (RMSE) for the best was found to be 0.30 μg/m³ (0.42 μg/m³) for BC and 1698.33 pt /cm³ (2613.62 pt /cm³) for PN (Table 4.11). Based on ML algorithms on Hourly_average_W, the MAE (RMSE) for the best was found to be 0.33 μg/m³ (0.48 μg/m³) for BC and 1699.50 pt /cm³ (2857.61 pt /cm³) for PN (Table 4.12).

Table 4.12.

Summary of Random CV Model Performance for Hourly_average_Weather Models

BC (µg/m³) (Hourly_average_W)			**PN (pt /cm³) (Hourly_average_W)**		
Models	TEST_MAE	TEST_RMSE	Models	TEST_MAE	TEST_RMSE
Decision Tree	0.33	0.48	Decision Tree	1758.55	2826.09
Extra Tree	0.34	0.49	Extra Tree	1799.46	2886.88
K Neighbors	0.33	0.49	K Neighbors	1716.21	2767.10
SVR	0.33	0.49	SVR	1699.50	2857.61
Elastic Net	0.35	0.50	Elastic Net	1939.84	3035.09
Random Forest	0.33	0.48	Random Forest	1763.42	2829.62

The best Hourly_average_Lu_W model was obtained using Gradient Boosting, with MAE (RMSE) results of of 0.25 µg/m³ (0.37 µg/m³) for black carbon, and 1420.29 pt /cm³ (2319.62 pt /cm³) for particle number (Table 4.13). Lasso and Decision Tree presented the worst algorithms for both pollutants.

Table 4.13.

Summary of Random CV Model Performance for Hourly_average_Land_Use_Weather Models

BC (µg/m³) (Hourly_average_Lu_W)			**PN (pt /cm³) (Hourly_average_Lu_W)**		
Models	TEST_MAE	TEST_RMSE	Models	TEST_MAE	TEST_RMSE
Decision Tree	0.31	0.44	Decision Tree	1793.21	2681.17
Extra Tree	0.30	0.42	Extra Tree	1773.98	2671.75
K Neighbors	0.29	0.41	K Neighbors	1735.93	2591.91
Ridge	0.29	0.42	Ridge	1776.03	2638.53
Elastic Net	0.29	0.41	Elastic Net	1707.04	2621.73
Lasso	0.31	0.45	Lasso	1738.77	2718.91
Random Forest	0.28	0.41	Random Forest	1673.82	2558.07
Gradient Boosting	0.25	0.37	Gradient Boosting	1420.29	2319.62

The best Hourly_average _W_Em model was obtained using K Neighbors, with MAE (RMSE) results of 0.26 µg/m³ (0.39 µg/m³) for black carbon and 1356.25 pt /cm³ (2367.64 pt /cm³) for particle number (Table 4.14).

Table 4.14.

Summary of Random CV Model Performance for Hourly_average_Weather_Emissions Models

BC (µg/m³) (Hourly_average_W_Em)			**PN (pt /cm³) (Hourly_average_W_Em)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.32	0.47	Decision Tree	1655.72	2639.03
Extra Tree	0.32	0.47	Extra Tree	1754.04	2800.65
K Neighbors	0.26	0.39	K Neighbors	1356.25	2367.64
Ridge	0.33	0.48	Ridge	1817.95	2870.20
Elastic Net	0.33	0.48	Elastic Net	1806.50	2877.54
Lasso	0.33	0.48	Lasso	1807.05	2890.50
Random Forest	0.30	0.44	Random Forest	1579.11	2586.31
Gradient Boosting	0.33	0.48	Gradient Boosting	1794.11	2883.49

Table 4.15.

Summary of Random CV Model Performance for

Hourly_average_Land_Use_Weather_Emissions Models

BC (µg/m³) (Hourly_average_Lu_W_Em)			**PN (pt /cm³) (Hourly_average_Lu_W_Em)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.29	0.42	Decision Tree	1698.75	2620.45
Extra Tree	0.30	0.42	Extra Tree	1776.01	2700.05
K Neighbors	0.29	0.41	K Neighbors	1703.85	2543.82
Ridge	0.29	0.41	Ridge	1756.61	2625.98
Elastic Net	0.28	0.41	Elastic Net	1697.71	2613.56
Lasso	0.30	0.44	Lasso	1732.02	2713.59
Random Forest	0.27	0.40	Random Forest	1602.53	2489.02
Gradient Boosting	0.24	0.35	Gradient Boosting	1323.55	2155.56

The Hourly_average model based on land use, weather, and emissions resulted in errors that were almost similar to the Hourly_average_Lu_W and Hourly_average_W_Em models (Table 4.13, 4.14, and 4.15). The best Hourly_average_Lu_W_Em were obtained using Gradient Boosting, with MAE (RMSE) results of 0.24 µg/m³ (0.35 µg/m³) for black carbon and 1323.55 pt

/cm³ (2155.56 pt /cm³) for particle number (Table 4.15). The best performance models were found to be the models based on all input variables built by Gradient Boosting, with MAE (RMSE) results of 0.23 μg/m³ (0.34 μg/m³) for black carbon and 1204 pt /cm³ (2038.3 pt /cm³) for particle number (Table 4.17). The performance of the model based on the same variable except the emission (Hourly_average_Lu_W_Hr) was the most similar (Table 4.16); in other words, the emissions variable did not make a significant difference in performance across the Hourly_average models.

Table 4.16.

Summary of Random CV Model Performance for

Hourly_average_Land_Use_Weather_Hour_of_Day Models

BC (μg/m³) (Hourly_average_Lu_W_Hr)			**PN (pt /cm³) (Hourly_average_Lu_W_Hr)**		
Models	TEST_MAE	TEST_RMSE	Models	TEST_MAE	TEST_RMSE
Decision Tree	0.31	0.43	Decision Tree	1712.87	2579.69
Extra Tree	0.30	0.42	Extra Tree	1588.06	2451.72
K Neighbors	0.26	0.39	K Neighbors	1543.28	2440.11
SVR	0.26	0.37	SVR	1385.87	2314.83
Ridge	0.28	0.40	Ridge	1514.54	2393.25
Elastic Net	0.27	0.40	Elastic Net	1468.70	2384.65
Lasso	0.29	0.43	Lasso	1555.62	2532.63
Random Forest	0.28	0.41	Random Forest	1588.10	2449.36
Gradient Boosting	0.23	0.34	Gradient Boosting	1216.85	2058.4

Table 4.17.

Summary of Random CV Model Performance for

Hourly_average_Land_Use_Weather_Emissions_Hour_of_Day Models

BC (µg/m³) (Hourly_average_Lu_W_Em_Hr)			**PN (pt /cm³) (Hourly_average_Lu_W_Em_Hr)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.31	0.45	Decision Tree	1750.81	2631.54
Extra Tree	0.31	0.44	Extra Tree	1777.90	2661.11
K Neighbors	0.26	0.39	K Neighbors	1535.55	2419.91
SVR	0.26	0.37	SVR	1427.23	2543.49
Ridge	0.28	0.40	Ridge	1512.40	2367.46
Elastic Net	0.27	0.40	Elastic Net	1481.31	2380.23
Lasso	0.29	0.43	Lasso	1555.62	2532.63
Random Forest	0.27	0.39	Random Forest	1515.70	2385.10
Gradient Boosting	0.23	0.34	Gradient Boosting	1204.00	2038.31

In the Hourly_average models for PN, both spatial and temporal variables were found to be important for enhancing the model's accuracy. The MAE (RMSE) results for both Hourly_average_Lu and Hourly_average_W models were almost the same, approximately equal to 1,700 pt /cm³ (2,700 pt /cm³), which is consistent with the findings of Pandey et al. (Pandey et al., 2013b). The MAE (RMSE) results dropped to 1420.3 pt /cm³ (2319.6 pt /cm³) when both variables were incorporated; however, the data improved to 1204 pt /cm³ (2038.3 pt /cm³) when hour of the day was added.

4.1.2. Effect of Adding Spatiotemporal Variables on Random Cross Validation Models.

Differences in the improvement provided by time-dependent variables for the One_second models were significant, whereas this trend was not observed in the Hourly_average model. This variance can be attributed to the data aggregation approach that restricted temporal and spatial factors; when aggregation temporally one hour and spatially 100 meters (Figures 4.1 and 4.2).

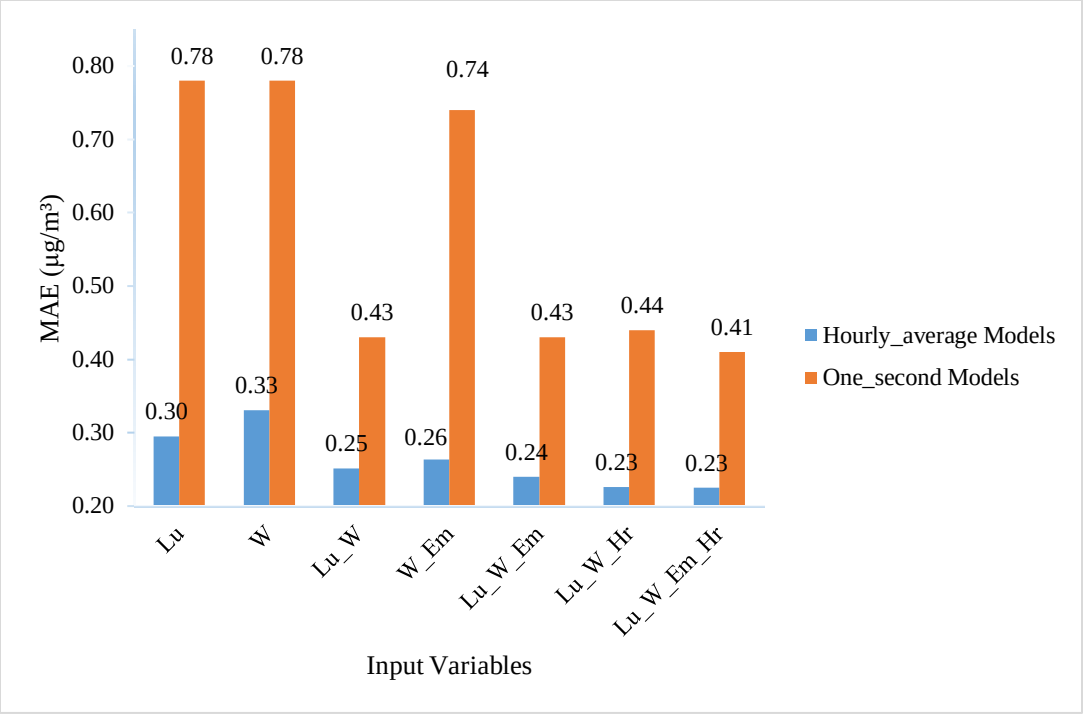

Figure 4.1. Black carbon random cross-validation model performance

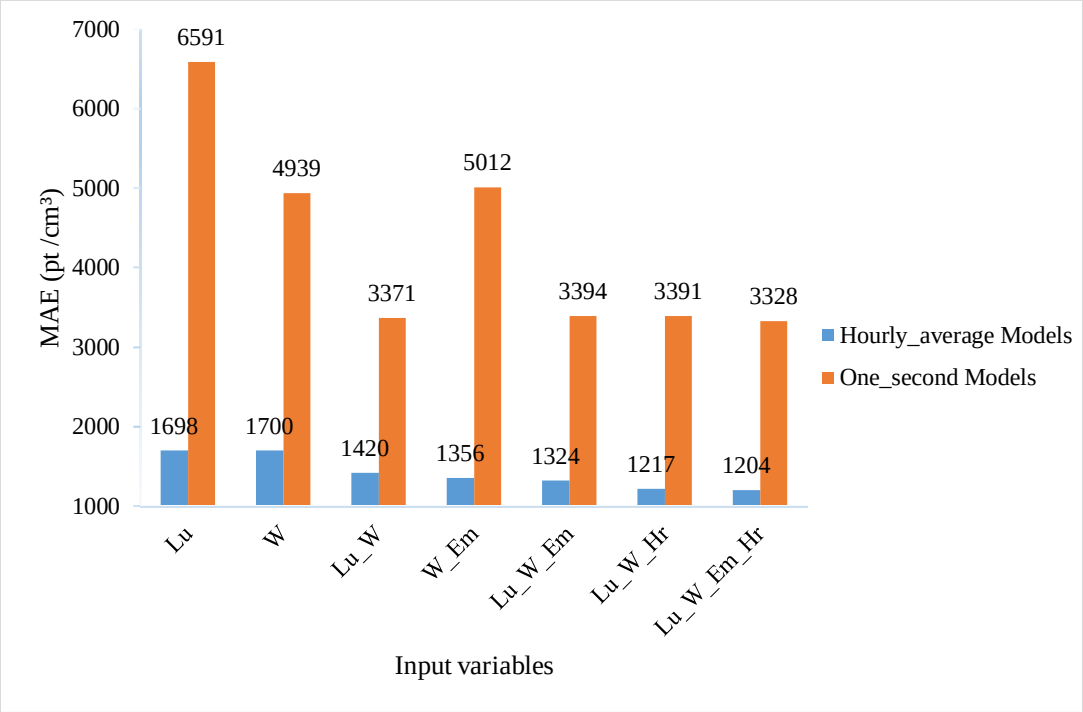

Figure 4.2. Particle number random cross-validation model performance

The Hourly_average models for BC, the model based on spatial factors (i.e., Hourly_average_Lu) shows less error than models based on weather factors (i.e., Hourly_average_W). Note that this trend is different in comparison to findings obtained for the One_second model where both of these variables evidenced similar performance. However, the errors dropped when both variables were incorporated in both types of models. For the Hourly_average models, the MAE (RMSE) results dropped to 0.25 μg/m³ (0.37 μg/m³) after incorporating both variables in the Hourly_average model, and improved to 0.23 μg/m³ (0.34 μg/m³) when hour of the day was factored in. For the One_second models, the MAE (RMSE) results dropped to 0.43 μg/m³ (1.16 μg/m³) after incorporating both variables, and improved to 0.44 μg/m³ (0.1.18 μg/m³) when hour of the day was factored in.

With respect to the Hourly_average models for PN, both land-use only and weather-only models performed similarly, where the MAE (RMSE) was almost equal to 1700 pt /cm³ (2857.61 pt /cm³). This trend is different in comparison to the One_second models, for which the weather-only model showed significant improvement in comparison to the land-use-only model. This outcome indicates the improvement generated by incorporating temporal variables in the PN model developed by refined data. On the other hand, the performance was found to fit better when both land use and weather variables were incorporated together in both Hourly_average and One_second models. Furthermore, including the hour of the day improved the Hourly_average model, which supports prior findings (Hankey et. al., 2019), where all the hours where selected in the stepwise regression PN model. However, no improvements were found in the One_second model.

For One_second model for both pollutants, the weather and emissions model was found to be similar to the weather-only model. For the Hourly_average model, the weather and

emissions variables could used to build model with error, as well as the land-use and weather model. However, the model incorporating all variables (except emissions) performed similarly to the model that did include the emissions factors. This finding indicates that the spatiotemporal variables would not be necessary to include if the hour-of-the-day variable was included in the model-development process.

Finally, incorporating time-dependent variables did not result in any significant improvements in the aggregated models—likely due to the fact that these variables were temporally aggregated 1 hour, which led to smoothing the data and decreasing the hour-by-hour variance. It should be noted that these data were collected during the summer and the fall months; thus, there were no extreme changes in the time-dependent factors during the monitoring campaign.

4.2. Spatial and Spatial-temporal Cross Validation Approach for Reliable Model Development

Three types of cross-validation approaches were applied to show the ability of machine leaning to forecast pollution levels in area not previously assessed. These types of CV approaches could help researchers to define the accurate CV strategies and validate model reliability. The performance of machine learning using different types of cross validation strategies will, of course, vary. For instance, as noted by Meyer et al. (2018), the relationship between time of observation and spatial-temporal boundaries can make it difficult to define an accurate validation approach, thereby jeopardizing reliable performance evaluations (Meyer et al., 2018). This factor motivated to use more sophisticated validation approaches that would help identify reliable performance patterns in machine learning with the goal of predicting concentration levels more accurately. Therefore, this section will show the spatial and spatial-

temporal cross validation model result. Furthermore, the summery will show the comparison across the three cross validation results to assess the model overestimated.

4.2.1. Spatial Cross-Validation Models Results

Spatial cross-validation was used to assess the performance of the prediction model in unseen areas by eliminating locational information in the model training process. The spatial CV was applied in similar models that provided in the random CV; thus, one Daily_average model and seven One_second and Hourly_average models were developed (Table 4.18).

Table 4.18.

Summary of Model Performance for Each Pollutant and Model Type in the Spatial CV

Models	Input variables	BC			PN		
		Best ML algorithm	MAE (μg/m³)	RMSE (μg/m³)	Best ML algorithm	MAE (pt /cm³)	RMSE (pt /cm³)
Daily_average	Lu	K Neighbors	0.20	0.27	K Neighbors	886.91	1338.33
One_second	Lu	Lasso	0.84	2.50	Random Forest	6831.35	30366.65
	W	Elastic Net	0.80	2.48	Random Forest	5388.70	29815.66
	Lu_W	Elastic Net	0.80	2.48	Random Forest	5660.47	29970.90
	W_Em	Elastic Net	0.81	2.48	Random Forest	5629.49	29874.96
	Lu_W_Em	Elastic Net	0.81	2.48	Random Forest	5579.23	29900.35
	Lu_W_Hr	Lasso	0.81	2.48	Random Forest	5611.13	29929.87
	Lu_W_Em_Hr	Lasso	0.81	2.48	Elastic Net	5696.60	29974.55
Hourly_average	Lu	Gradient Boosting	0.36	0.59	Random Forest	2002.01	3204.07
	W	Elastic Net	0.36	0.60	Ridge	2035.99	3265.50
	Lu_W	Lasso	0.36	0.59	K Neighbors	1979.40	3139.51
	W_Em	Gradient Boosting	0.36	0.60	Gradient Boosting	1982.75	3228.51
	Lu_W_Em	Random Forest	0.35	0.58	Random Forest	1939.55	3140.89
	Lu_W_Hr	Gradient Boosting	0.35	0.58	Lasso	1862.72	3022.66
	Lu_W_Em_Hr	Gradient Boosting	0.36	0.6	Lasso	1862.27	3023.05

Daily_average models. The performance of the Daily_average model was found to deteriorate from random CV to spatial CV. The MAE (RMSE) results for the best-performing BC machine learning model was found to be 0.2 μg/m³ (0.27 μg/m³) for the K Neighbors; in contrast, the poorest-performing model was found to be the Ridge model, at 0.26 μg/m³ (0.34 μg/m³). The MAE (RMSE) results for the best PN machine learning models were found to be 886.91 pt /cm³ (1338.33 pt /cm³) for the K Neighbors; in contrast, the poorest-performing model was the Ridge model at 1224.92 pt /cm³ (1615.2 pt /cm³) (Table 4.19).

Table 4.19.

Summary of Spatial CV Model Performance for Daily_average Models

BC (μg/m³)			PN (pt /cm³)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.24	0.30	Decision Tree	1132.43	1593.33
Extra Tree	0.24	0.30	Extra Tree	1152.54	1696.72
K Neighbors	0.20	0.27	K Neighbors	886.91	1338.33
SVR	0.21	0.28	SVR	1041.54	1415.10
Ridge	0.26	0.34	Ridge	1224.92	1615.21
Elastic Net	0.21	0.27	Elastic Net	985.90	1391.02
Lasso	0.21	0.27	Lasso	999.08	1390.64
Random Forest	0.21	0.27	Random Forest	976.83	1384.41
Gradient Boosting	0.21	0.27	Gradient Boosting	1019.84	1424.12

One_second models. Similar to the models constructed for random CV, seven types of the One_second models were developed based on the different input variables. Across the seven One_second models, the average MAE (RMSE) results were found to be 0.81 μg/m³ (2.48 μg/m³) for black carbon, and 5771 pt /cm³ (29976.13 pt /cm³) for particle number. The One_second models based on land-use dataset demonstrated the worst performance across all the One_second models. However, the different was not significant.

The best performance model for MAE (RMSE) were confirmed for the Random Forest and Elastic Net models, namely 0.84 μg/m³ (2.49 μg/m³) for black carbon; the best results for PN were associated with the Random Forest model, with MAE (RMSE) results of 6831.35 pt /cm³ (30366.65 pt /cm³). Conversely, the Ridge algorithm demonstrated the poorest performance (Table 4.20).

Table 4.20.

Summary of Spatial CV Model Performance for One_second_Land_Use Models

BC (μg/m³) (One_second_Lu)			PN (pt /cm³) (One_second_Lu)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.91	2.53	Decision Tree	6904.57	30397.43
Extra Tree	0.88	2.51	Extra Tree	6964.44	30377.66
Ridge	1.49	2.97	Ridge	8170.86	32009.87
Elastic Net	0.84	2.49	Elastic Net	6910.54	30380.15
Lasso	0.84	2.50	Lasso	6842.79	30372.20
Random Forest	0.84	2.49	Random Forest	6831.35	30366.65

Table 4.21.

Summary of Spatial CV Model Performance for One_second_Weather Models

BC (μg/m³) (One_second_W)			PN (pt /cm³) (One_second_W)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.81	2.48	Decision Tree	5713.95	29866.03
Extra Tree	0.82	2.49	Extra Tree	6117.63	30055.32
K Neighbors	1.10	3.54	K Neighbors	5774.51	30042.60
Ridge	0.80	2.48	Ridge	5877.45	30185.07
Elastic Net	0.80	2.48	Elastic Net	5700.16	29948.42
Lasso	0.81	2.49	Lasso	5715.65	29945.61
Random Forest	0.80	2.48	Random Forest	5388.70	29815.66
Gradient Boosting	0.82	2.49	Gradient Boosting	6352.47	30245.01

The One_second model based on weather data performed with the fewest errors among all the One_second models (Table 4.21); specifically, the MAE (RMSE) results were 0.8 μg/m³ (2.48 μg/m³) for BC, and 5388.7 pt /cm³ (29815.66 pt /cm³) for PN. On the other hand, the best One_second_Lu_W model build using Elastic Net shows the MAE (RMSE) of 0.8 μg/m³ (2.48 μg/m³) for BC and Random Forest shows the MAE (RMSE) of 5660.5 pt /cm³ (29970.9 pt /cm³) for PN (Table 4.22).

Table 4.22.

Summary of Spatial CV Model Performance for One_second_Land_Use_Weather Models

BC (μg/m³) (One_second_Lu_W)			PN (pt /cm³) (One_second_Lu_W)		
Models	TEST_MAE	TEST_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.91	2.54	Decision Tree	6573.20	30227.16
Extra Tree	0.88	2.50	Extra Tree	6824.26	30286.47
Ridge	1.42	2.91	Ridge	7871.31	32522.37
Elastic Net	0.80	2.48	Elastic Net	5740.41	30034.93
Lasso	0.81	2.48	Lasso	5860.50	30261.99
Random Forest	0.82	2.48	Random Forest	5660.47	29970.90

Table 4.23.

Summary of Spatial CV Model Performance for One_second_Weather_Emissions Models

BC (μg/m³) (One_second_W_Em)			PN (pt /cm³) (One_second_W_Em)		
Models	TEST_MAE	TEST_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.84	2.49	Decision Tree	6026.44	30032.97
Extra Tree	0.83	2.49	Extra Tree	6177.36	30066.57
Ridge	0.82	2.48	Ridge	5971.31	30205.60
Elastic Net	0.81	2.48	Elastic Net	5684.62	29952.97
Lasso	0.81	2.49	Lasso	5707.57	29943.58
Random Forest	0.81	2.48	Random Forest	5629.49	29874.96
Gradient Boosting	0.82	2.49	Gradient Boosting	6391.90	30251.33

Table 4.24.

Summary of Spatial CV Model Performance for One_second_Land_Use_Weather_Emissions Models

BC (μg/m³) (One_second_Lu_W_Em)			PN (pt /cm³) (One_second_Lu_W_Em)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.85	2.50	Decision Tree	6047.03	30092.61
Extra Tree	0.91	2.58	Extra Tree	6154.71	30010.21
Ridge	1.38	2.88	Ridge	9168.19	52714.23
Elastic Net	0.81	2.48	Elastic Net	5727.12	30010.69
Lasso	0.81	2.48	Lasso	5804.75	30097.46
Random Forest	0.83	2.48	Random Forest	5579.23	29900.35

Table 4.25.

Summary of Spatial CV Model Performance for One_second_Land_Use_Weather_Hour_of_Day Models

BC (μg/m³) (One_second_Lu_W_Hr)			PN (pt /cm³) (One_second_Lu_W_Hr)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.84	2.51	Decision Tree	6567.61	30284.32
Extra Tree	0.88	2.53	Extra Tree	6586.31	30261.41
Ridge	1.44	2.93	Ridge	7921.65	32069.72
Elastic Net	0.81	2.48	Elastic Net	5697.73	29974.11
Lasso	0.81	2.48	Lasso	5766.65	30027.63
Random Forest	0.82	2.47	Random Forest	5611.13	29929.87

Table 4.26.

Summary of Spatial CV Model Performance for

One_second_Land_Use_Weather_Emissions_Hour_of_Day Models

BC (μg/m³) (One_second_Lu_W_Em_Hr)			PN (pt /cm³) (One_second_Lu_W_Em_Hr)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.88	2.58	Decision Tree	6449.85	30166.41
Extra Tree	0.90	2.53	Extra Tree	6470.41	30243.07
Ridge	1.40	2.90	Ridge	9449.42	68041.50
Elastic Net	0.81	2.48	Elastic Net	5696.60	29974.55
Lasso	0.81	2.48	Lasso	5774.32	30036.49
Random Forest	0.82	2.48			

Among the One_second models demonstrating similar performance were those based on (a) land use; (b) land use and weather; (c) weather and emissions; (d) land use, weather, and hour of the day; (e) land use, weather, and emissions; and (f) the model that incorporated all variables. The average results for the best MAE (RMSE) performance were 0.81 μg/m³ (2.48 μg/m³) for black carbon and 5629 pt /cm³ (29,919 pt /cm³) for particle number among the last four models (Table 4.20, 4.21, 4.22, 4.223, 4.24, 4.25 and 4.26).

Hourly_average models. Like the One_second models, seven types of Hourly_average models were developed. Overall, the average MAE (RMSE) results across the seven types of Hourly_average best models were found to be 0.4 μg/m³ (0.86 μg/m³) for BC, and 1952 pt /cm³ (3147 pt /cm³) for PN (Table 4.18). When compared to analogous data for the One_second models, the average MAE (RMSE) results for the Hourly_average models fell to half that for BC, and 66% for PN.

Table 4.27.

Summary of Spatial CV Model Performance for Hourly_average_Land_Use Models

BC (μg/m³) (Hourly_average_Lu)			PN (pt /cm³) (Hourly_average_Lu)		
Models	TEST_MAE	TEST_RMSE	Models	TEST_MAE	TEST_RMSE
Decision Tree	0.41	0.63	Decision Tree	2492.20	3701.17
Extra Tree	0.42	0.65	Extra Tree	2462.84	3742.51
K Neighbors	0.38	0.60	K Neighbors	2016.49	3192.74
Ridge	0.58	0.87	Ridge	3706.32	5660.79
Elastic Net	0.41	0.63	Elastic Net	2123.79	3272.51
Lasso	0.36	0.59	Lasso	2026.28	3221.27
Random Forest	0.36	0.59	Random Forest	2002.01	3204.07
Gradient Boosting	0.36	0.59	Gradient Boosting	2049.63	3256.10

The MAE (RMSE) results for the best Hourly_average_Lu model was 0.36 μg/m³ (0.59 μg/m³) for BC and 2002.01 pt /cm³ (3204.07 pt /cm³) for PN (Table 4.27). In contrast, in the Hourly_average_W models the errors remain stable across all the algorithms with MAE (RMSE) of 0.36 μg/m³ (0.6 μg/m³) for BC and ~2042 pt /cm³ (~3220 pt /cm³) for PN (Table 4.28).

Table 4.28.

Summary of Spatial CV Model Performance for Hourly_average_Weather Models

BC (μg/m³) (Hourly_average_W)			PN (pt /cm³) (Hourly_average_W)		
Models	TEST_MAE	TEST_RMSE	Models	TEST_MAE	TEST_RMSE
K Neighbors	0.36	0.60	K Neighbors	2044.64	3173.63
Ridge	0.36	0.60	Ridge	2035.99	3265.50
Elastic Net	0.36	0.60	Elastic Net	2046.66	3281.15
Lasso	0.36	0.60	Lasso	2046.66	3281.15

Table 4.29.

Summary of Spatial CV Model Performance for Hourly_average_Land_Use_Weather Models

BC (µg/m³) (Hourly_average_Lu_W)			**PN (pt /cm³) (Hourly_average_Lu_W)**		
Models	TEST_MAE	TEST_RMSE	Models	TEST_MAE	TEST_RMSE
Decision Tree	0.44	0.67	Decision Tree	2419.64	3646.31
Extra Tree	0.43	0.66	Extra Tree	2256.68	3511.30
K Neighbors	0.37	0.60	K Neighbors	1979.40	3139.51
Ridge	0.58	0.86	Ridge	3713.82	5708.64
Elastic Net	0.41	0.63	Elastic Net	2024.84	3207.39
Lasso	0.36	0.59	Lasso	2017.88	3211.94
Random Forest	0.36	0.58	Random Forest	1981.40	3194.65
Gradient Boosting	0.36	0.59	Gradient Boosting	2140.75	3271.76

The best Hourly_average_Lu_W model MAE (RMSE) results were found to be 0.36 µg/m³ (0.59 µg/m³) for BC, and 1979.4 pt /cm³ (3139.51 pt /cm³) for PN (Table 4.29). Similarly, the best Hourly_average _W_Em model MAE (RMSE) results were found to be 0.36 µg/m³ (0.6 µg/m³) for BC, and 1982.75 pt /cm³ (3228.1 pt /cm³) for PN (Table 4.30).

Table 4.30.

Summary of Spatial CV Model Performance for Hourly_average_Weather_Emissions Models

BC (µg/m³) (Hourly_average_W_Em)			**PN (pt /cm³) (Hourly_average_W_Em)**		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.38	0.61	Decision Tree	2245.64	3423.78
Extra Tree	0.36	0.60	Extra Tree	2035.29	3215.94
K Neighbors	0.37	0.60	K Neighbors	1985.77	3136.78
Ridge	0.37	0.60	Ridge	2026.64	3231.94
Elastic Net	0.36	0.60	Elastic Net	2012.01	3234.99
Lasso	0.36	0.60	Lasso	2011.57	3234.73
Random Forest	0.37	0.60	Random Forest	2029.69	3191.67
Gradient Boosting	0.36	0.60	Gradient Boosting	1982.75	3228.51

According to Table 4.31, the black carbon Hourly_average best model based on land use, weather, and emissions afforded errors that were comparable to results for Hourly_average_Lu_W_Hr (Table 4.32). The MAE (RMSE) results for the best performing model MAE (RMSE) was found to be 0.35 μg/m³ (0.58 μg/m³). Meanwhile, the particle number Hourly_average best model based on land use, weather, and emissions MAE (RMSE) was 1939.55 pt /cm³ (3140.89 pt /cm³) (Table 4.31).

Table 4.31.

Summary of Spatial CV Model Performance for Hourly_average_Land_Use_Weather_Emissions Models

BC (μg/m³) (Hourly_average_Lu_W_Em)			PN (pt /cm³) (Hourly_average_Lu_W_Em)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.41	0.64	Decision Tree	2260.77	3517.29
Extra Tree	0.42	0.65	Extra Tree	2296.28	3463.43
K Neighbors	0.38	0.60	K Neighbors	1952.95	3111.29
Ridge	0.59	0.90	Ridge	3830.35	6038.86
Elastic Net	0.38	0.61	Elastic Net	2024.99	3222.56
Lasso	0.36	0.59	Lasso	2014.50	3209.45
Random Forest	0.35	0.58	Random Forest	1939.55	3140.89
Gradient Boosting	0.36	0.59	Gradient Boosting	1958.88	3117.46

Table 4.32.

Summary of Spatial CV Model Performance for Hourly_average_Land_Use_Weather_Hour_of_ Day Models

BC (µg/m³) (Hourly_average_Lu_W_Hr)			**PN (pt /cm³) (Hourly_average_Lu_W_Hr)**		
Models	TEST_MAE	TEST_RMSE	Models	TEST_MAE	TEST_RMSE
Decision Tree	0.45	0.66	Decision Tree	2374.10	3586.52
Extra Tree	0.39	0.63	Extra Tree	2227.96	3520.13
K Neighbors	0.36	0.59	K Neighbors	1903.67	3057.31
SVR	0.52	0.80	SVR	2729.56	6634.06
Ridge	0.57	0.86	Ridge	3488.46	5586.09
Elastic Net	0.36	0.59	Elastic Net	1951.56	3022.73
Lasso	0.35	0.59	Lasso	1862.72	3022.66
Random Forest	0.35	0.58	Random Forest	1949.06	3148.33
Gradient Boosting	0.35	0.58	Gradient Boosting	1917.43	3061.03

Table 4.33.

Summary of Spatial CV Model Performance for

Hourly_average_Land_Use_Weather_Emissions_Hour_of_Day Models

BC (µg/m³) (Hourly_average_Lu_W_Em_Hr)			PN (pt /cm³) (Hourly_average_Lu_W_Em_Hr)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.88	2.58	Decision Tree	2277.81	3374.79
Extra Tree	0.90	2.53	Extra Tree	2251.89	3409.99
K Neighbors	0.36	0.59	K Neighbors	1886.15	3036.79
SVR	0.52	0.86	SVR	2943.74	4837.44
Ridge	1.40	2.90	Ridge	3641.07	5884.33
Elastic Net	0.81	2.48	Elastic Net	1977.99	3045.34
Lasso	0.81	2.48	Lasso	1862.27	3023.05
Random Forest	0.82	2.48	Random Forest	1927.39	3115.56
Gradient Boosting	0.36	0.59	Gradient Boosting	1979.54	3168.35

The fewest errors for the Hourly_average_Lu_W_Em_Hr spatial CV models were linked to the Gradient Boosting algorithm for black carbon, and to the Lasso algorithm for particle number; specifically, the MAE (RMSE) results were found to be 0.36 μg/m³ (0.59 μg/m³) for BC, and 1862.27 pt /cm³ (3023.05 pt /cm³) for PN (Table 4.32).

4.2.2. Effect of Adding Spatiotemporal Variables on Spatial Cross Validation Models

This section discusses how time-dependent variables impacted model performance based on its ability to predict hidden location. Furthermore, we also investigated the extent to which a given model was able to predict pollutant concentration in a location that was not used in the model-building process. Figures 4.3 and 4.4 show the spatial CV across all models taking into account the different input variables.

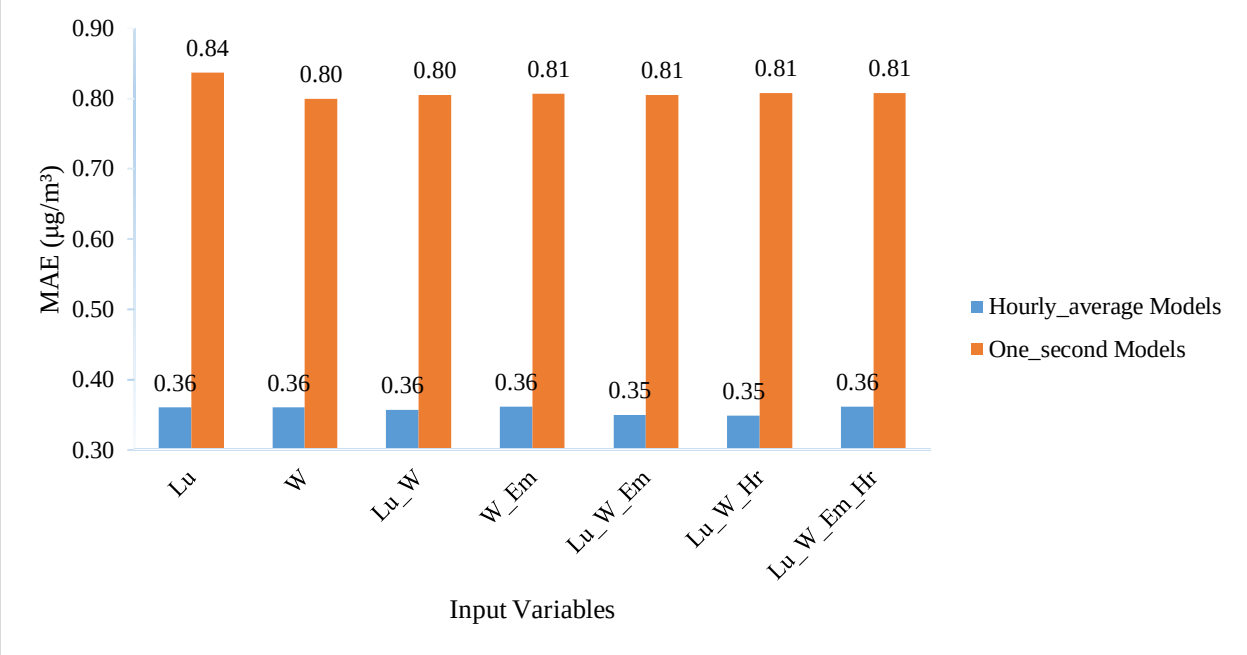

Figure 4.3. Black carbon spatial cross-validation model performance

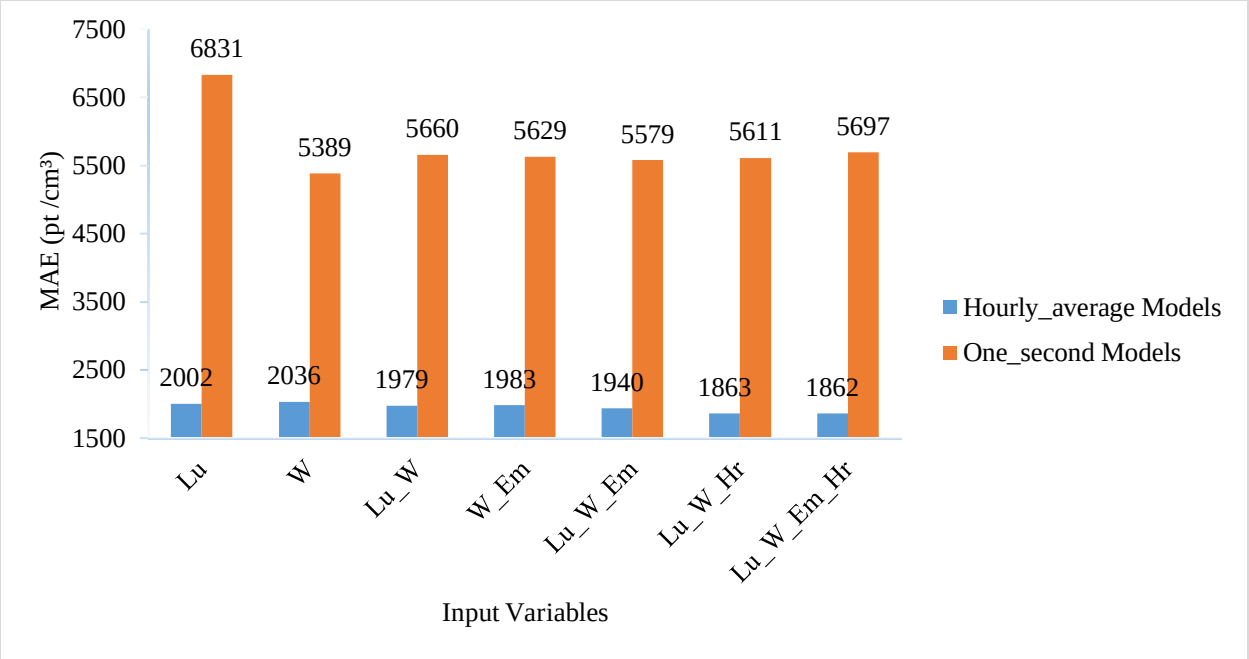

Figure 4.4. Particle number spatial cross-validation model performance

With respect to the One_second model, holding out one location in the training process was not improved by adding temporal or spatiotemporal variables. This finding was expected since meteorological data and emissions factors are time-dependent variables; thus, holding out one location was incorporated to eliminate the time-factors effect. However, the land-use data (spatial data) model performed most poorly when location was removed, which can be attributed to the fact that the model relies on spatial data; thus, hiding part of this data will affect the model's performance. Note, however, that the temporal variation still presents in the locations used in the training process. However, these trends were found to be quite different in the Hourly_average model; thus, smoothing the data spatially will help to develop robust models. It should be noted that the models were not impacted by spatial variables when location was hidden; additionally, aggregating the data temporally will aid in reducing or eliminating the influence of temporal factors.

4.2.3. Spatial-Temporal Cross-Validation Models Result.

Spatial-temporal CV was applied to assess the ability of machine learning to build a forecasting system for pollutant concentration in unseen location and time. Since not all hourly data were collected on the same day, holding out the last run of each hour of the day will force the data to be balanced. This approach for holding out data was applied only for One_second models. The spatial-temporal CV was not applied in the Hourly_average models because those data points were temporally one-hour aggregated and spatially 100 m aggregated; as such, a forced balancing cannot be applied to this type of aggregation data.

Table 4.34.

Summary of Model Performance for Each Pollutant and Model Type in the Spatial-temporal CV

Models	Input variables	BC			PN		
		Best ML algorithm	MAE (μg/m³)	RMSE (μg/m³)	Best ML algorithm	MAE (pt /cm³)	RMSE (pt /cm³)
One_second	Lu	Lasso	0.71	1.35	RandomForest	10145.75	34931.82
	W	ElasticNet	0.67	1.33	DecisionTree	8542.29	34110.93
	Lu_W	Lasso	0.67	1.33	Lasso	8867.18	34622.21
	W_Em	Lasso	0.67	1.32	RandomForest	8623.05	34231.66
	Lu_W_Em	Lasso	0.67	1.33	ElasticNet	8871.73	34415.29
	Lu_W_Hr	Lasso	0.67	1.33	ElasticNet	8890.51	34371.46
	Lu_W_Em_Hr	Lasso	0.67	1.33	ElasticNet	8888.30	34370.81

One_second models. Seven varieties of the One_second model were developed according to the different input variables. The One_second model based on the land-use dataset evidenced the worst performance among the seven models. The MAE (RMSE) results for the best-performing of One_second_Lu model was 0.71 μg/m³ (1.34 μg/m³) for BC, and 10145.75 pt /cm³ (34931.82 pt /cm³) for PN (Table 4.34). The errors were similar for the remaining six models with an average of 0.67 μg/m³ (1.33 μg/m³) for BC, and 8,780.5 pt /cm³ (34,353.7 pt /cm³) for PN. It must be noted, however, that the difference in error rate was not significant

across the pollutant models. In summary, averaging the MAE (RMSE) data across the seven types of One_second best models resulted in the following findings: 0.43 μg/m³ (1.26 μg/m³) for BC, and 3703.8 pt /cm³ (28723.04 pt /cm³) for PN (Tables 4.35 through 4.41).

Table 4.35.

Summary of Spatial-temporal CV Model Performance for One_second_Land_Use Models

BC (μg/m³) (One_second_Lu)			PN (pt /cm³) (One_second_Lu)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.88	1.49	Decision Tree	10295.46	34880.21
Extra Tree	1.04	2.04	Extra Tree	10324.67	35018.02
Ridge	1.38	2.10	Ridge	11115.64	35595.92
Elastic Net	0.71	1.34	Elastic Net	10199.22	34981.08
Lasso	0.71	1.35	Lasso	10148.96	34967.34
Random Forest	0.72	1.34	Random Forest	10145.75	34931.82

Table 4.36.

Summary of Spatial-temporal CV Model Performance for One_second_Weather Models

BC (μg/m³) (One_second_W)			PN (pt /cm³) (One_second_W)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.70	1.35	Decision Tree	8542.29	34110.93
Extra Tree	0.71	1.37	Extra Tree	9248.20	34108.03
Ridge	0.67	1.33	Ridge	9245.11	34777.57
Elastic Net	0.67	1.33	Elastic Net	9005.20	34756.92
Lasso	0.67	1.33	Lasso	8742.05	34191.46
Random Forest	0.72	1.36	Random Forest	8874.13	34307.90
Gradient Boosting	0.70	1.35	Gradient Boosting	9903.97	34863.65

Table 4.37.

Summary of Spatial-temporal CV Model Performance for One_second_Land_Use_Weather Models

BC (μg/m³) (One_second_Lu_W)			PN (pt /cm³) (One_second_Lu_W)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.78	1.38	Decision Tree	10476.85	34972.17
Extra Tree	0.78	1.40	Extra Tree	10047.96	34841.46
Ridge	1.28	2.00	Ridge	10983.51	38599.12
Elastic Net	0.67	1.32	Elastic Net	8998.73	35164.15
Lasso	0.67	1.33	Lasso	8867.18	34622.21
Random Forest	0.72	1.33	Random Forest	9667.81	34751.32

Table 4.38.

Summary of Spatial-temporal CV Model Performance for One_second_Weather_Emissions Models

BC (μg/m³) (One_second_W_Em)			PN (pt /cm³) (One_second_W_Em)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.75	1.35	Decision Tree	9054.29	34453.62
Extra Tree	0.69	1.34	Extra Tree	9214.70	34539.14
Ridge	0.69	1.33	Ridge	9347.99	34815.88
Elastic Net	0.68	1.33	Elastic Net	8973.93	34737.35
Lasso	0.67	1.32	Lasso	8828.61	34445.46
Random Forest	0.71	1.33	Random Forest	8623.05	34231.66
Gradient Boosting	0.70	1.33	Gradient Boosting	9983.33	34905.20

Table 4.39.

Summary of Spatial-temporal CV Model Performance for

One_second_Land_Use_Weather_Emissions Models

BC (μg/m³) (One_second_Lu_W_Em)			PN (pt /cm³) (One_second_Lu_W_Em)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.70	1.43	Decision Tree	9174.82	34483.71
Extra Tree	0.68	1.33	Extra Tree	9520.39	34544.28
Ridge	1.26	1.97	Ridge	12098.57	41710.61
Elastic Net	0.67	1.32	Elastic Net	8871.73	34415.29
Lasso	0.67	1.33	Lasso	8877.59	34606.42
Random Forest	0.73	1.34	Random Forest	9498.20	34677.38

Table 4.40.

Summary of Spatial-temporal CV Model Performance for

One_second_Land_Use_Weather_Hour_of_Day Models

BC (μg/m³) (One_second_Lu_W_Hr)			PN (pt /cm³) (One_second_Lu_W_Hr)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.82	1.78	Decision Tree	10361.44	35042.39
Extra Tree	0.81	2.30	Extra Tree	10020.68	34890.73
Ridge	1.31	2.03	Ridge	10999.58	35655.79
Elastic Net	0.69	1.34	Elastic Net	8890.51	34371.46
Lasso	0.67	1.33	Lasso	8916.41	34526.65
Random Forest	0.72	1.34	Random Forest	9867.21	34867.71

Table 4.41.

Summary of Spatial-temporal CV Model Performance for

One_second_Land_Use_Weather_Emissions_Hour_of_Day Models

BC (µg/m³) (One_second_Lu_W_Em_Hr)			PN (pt /cm³) (One_second_Lu_W_Em_Hr)		
Models	Test_MAE	Test_RMSE	Models	Test_MAE	Test_RMSE
Decision Tree	0.78	1.60	Decision Tree	9917.01	34665.18
Extra Tree	0.73	1.37	Extra Tree	10933.37	34839.51
Ridge	1.29	2.01	Ridge	12243.14	39647.06
Elastic Net	0.70	1.34	Elastic Net	8888.30	34370.81
Lasso	0.67	1.33	Lasso	8915.07	34524.39
Random Forest	0.72	1.33	Random Forest	9733.67	34817.46
Decision Tree	0.78	1.60	Decision Tree	9917.01	34665.18

4.2.4. Effect of Adding Spatiotemporal Variables on Spatial-Temporal Cross Validation Models

Results for spatial-temporal hold out were calculated when data values for location and time were hidden. This section will summarize the impact of time-dependent variables on the forecasting the pollutant concentration in unseen locations. This discussion will be enhanced by considering the disaggregated estimation model.

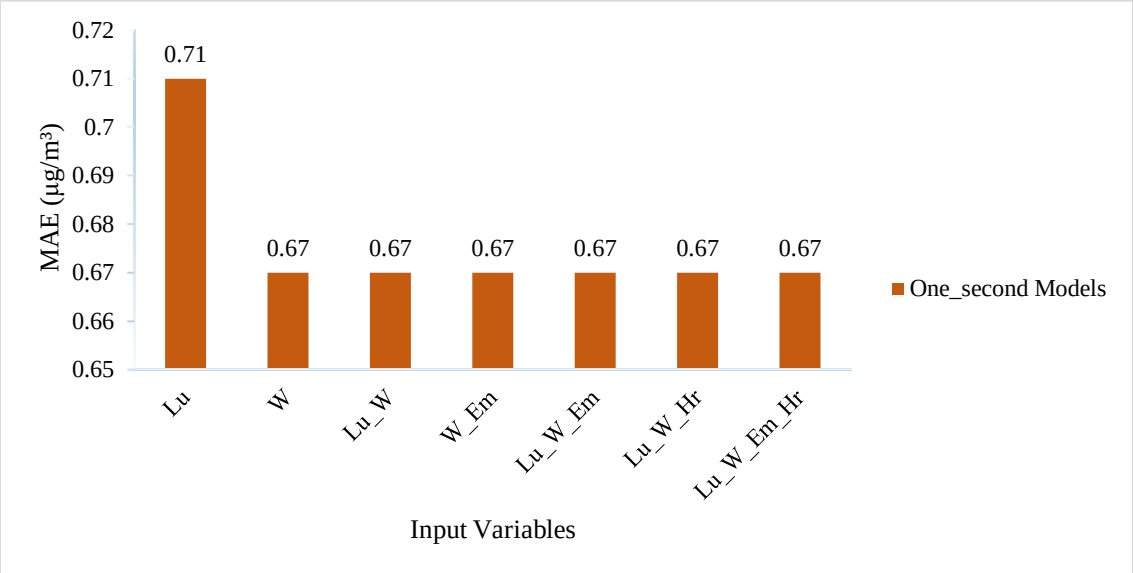

Figure 4.5. Black carbon spatial-temporal cross-validation model performance

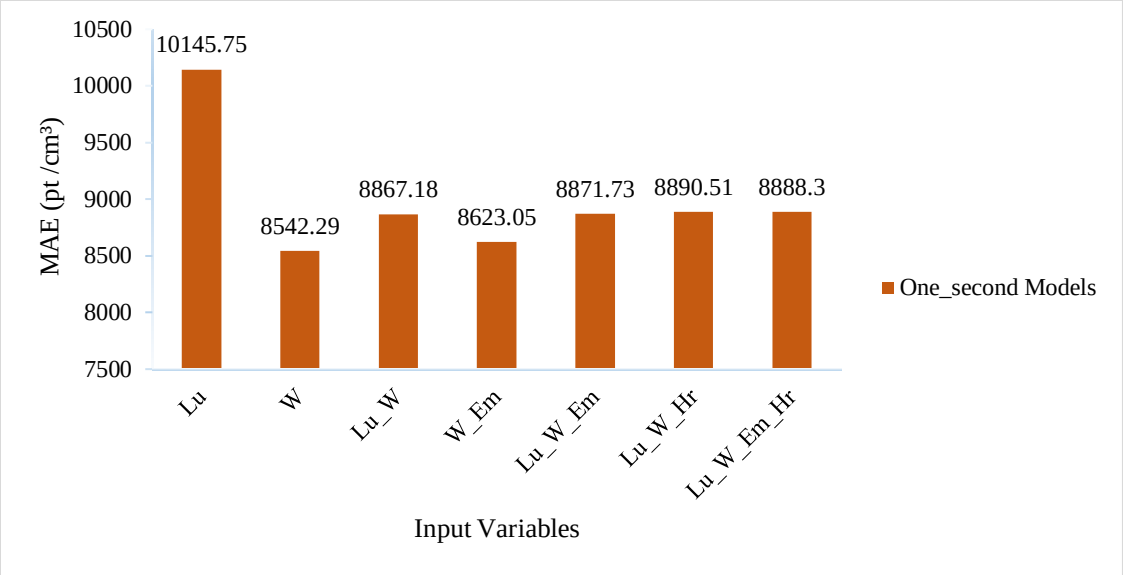

Figure 4.6. Particle number spatial-temporal cross-validation model performance

According to Figure 3.5 and 3.6, the One_second model developed using the spatial variables only (land-use data) shows the worst-performing model among spatial-temporal hold-out models. Hiding spatial variables in the location and time out model take off important the data in the training process. Additionally, because the spatial cluster was based on spatial data, one would expect to see the model's performance drop in the location or location/time out model. The temporal hold out was based on eliminating data for the final hour of the approximately six-hour monitoring period for each day. In order to develop a model that is robust to temporal hold outs, sufficient runs must be considered in data analysis. Each cluster in the spatial-temporal hold out validation appeared to exhibit the same temporal variation. Moreover, the performance were found to be consistent with the addition of time-dependent variables, which upholds prior research findings (Meyer et al., 2018). The discrepancies between random CV and spatial CV clearly demonstrated spatial over-fitting, as the models should accurately predict subsets of the time series of the training locations but failed to predict

unknown locations. This outcome became evident when applying different time-dependent variables in take time out. The model was found to be more robust based on temporal variations.

4.2.5. The Comparison Across the Three Validation Approach

Among all type of cross validation approaches assessed in this study, there was a significant difference in error generated by random CV in comparison to the other cross-validation methods. Random and spatial cross validation were used for the Daily_average and Hourly_average models. The number of observations in the test data obtained from spatial-temporal CV was found to decrease with an overlap between spatial cluster and last run on the data. Thus, the spatial-temporal CV approach was applied to the One_second models only in instances when more data were available for the test set. In general, comparisons were developed based on the best performance ML algorithm among the ML algorithms used.

Table 4.42.

Summary of the Three Types of Cross Validation for BC and PN.

Models	Input variables	BC						PN					
		Random CV		Spatial CV		Spatial_Temporal		Random CV		Spatial CV		Spatial_Temporal	
		MAE (μg/m³)	RMSE (μg/m³)	MAE (μg/m³)	RMSE (μg/m³)	MAE (μg/m³)	RMSE (μg/m³)	MAE (pt/cm³)	RMSE (pt /cm³)	MAE (pt/cm³)	RMSE (pt /cm³)	MAE (pt/cm³)	RMSE (pt/cm³)
Daily_avera	Lu	0.11	0.15	0.2	0.27			515.21	765.48	886.91	1338.33		
One_second	Lu	0.78	2.35	0.84	2.50	0.71	1.35	6590.91	31280.53	6831.35	30366.65	10145.75	34931.82
	W	0.78	2.35	0.80	2.48	0.67	1.33	4938.55	30793.68	5388.70	29815.66	8542.29	34110.93
	Lu_W	0.43	1.16	0.80	2.48	0.67	1.33	3370.64	28343.28	5660.47	29970.90	8867.18	34622.21
	W_Em	0.74	2.29	0.81	2.48	0.67	1.32	5011.75	30778.15	5629.49	29874.96	8623.05	34231.66
	Lu_W_Em	0.43	1.16	0.81	2.48	0.67	1.33	3394.32	28709.03	5579.23	29900.35	8871.73	34415.29
	Lu_W_Hr	0.44	1.18	0.81	2.48	0.67	1.33	3391.27	28854.74	5611.13	29929.87	8890.51	34371.46
	Lu_W_Em_Hr	0.41	1.07	0.81	2.48	0.67	1.33	3328.32	28423.16	5696.60	29974.55	8888.30	34370.81
Hourly_average	Lu	0.30	0.42	0.36	0.59			1698.33	2613.62	2002.01	3204.07		
	W	0.33	0.48	0.36	0.60			1699.50	2857.61	2035.99	3265.50		
	Lu_W	0.25	0.37	0.36	0.59			1420.29	2319.62	1979.40	3139.51		
	W_Em	0.26	0.39	0.36	0.60			1356.25	2367.64	1982.75	3228.51		
	Lu_W_Em	0.24	0.35	0.35	0.58			1323.55	2155.56	1939.55	3140.89		
	Lu_W_Hr	0.23	0.34	0.35	0.58			1204.00	2038.31	1862.72	3022.66		
	Lu_W_Em_Hr	0.23	0.34	0.36	0.60			1216.85	2058.40	1862.27	3023.05		

Daily_average and Hourly_average models. A random 10-fold CV revealed a lower model error with only minor differences between observed and predicted values, reflecting that the data represented a near "best fit." Nevertheless, when it came to unknown locations (spatial 17-fold CV), the performance dropped significantly. Table 4.42 presents the random and spatial CV validation results for the entire study period and area, as well as the MAE, and RMSE values between the CV predicted values and observations. Overall, the results indicate that MAE (RMSE) values increased from random to spatial hold out.

For particle number, the MAE results for daily predictions increased to almost 72% from 515.2 pt /cm³ for random CV to 886.9 pt /cm³ for spatial CV; to contrast, the average MAE values for the seven models of Hourly_average only increased approximately 33%. For the black carbon, the MAE results for daily predictions increased to almost 81% from 0.11 μg/m³ for random CV to 0.20 μg/m³ for spatial CV; to contrast, the average MAE values for the seven models of Hourly_average only increased approximately 38%.

In general, this difference indicates that the model became less capable of predicting beyond than the location of the training data in comparison to what might have been expected given the random CV error. Moreover, the ability of the Daily_average long-term model to predict the unknown location remained higher when compared to the Hourly_average short-term model. It should be noted that the percentage of error increase differed between the Daily_average and Hourly_average models. This finding indicates that the long-term average models were more sensitive to the type of CV approach; nonetheless, both random and spatial CV approaches still demonstrated best fit in comparison to the short-term model.

One_second models. The shifting dynamics of traffic-related air pollutants such as PN can result in significant differences in pollution levels over time and space (van den Bossche et

al., 2015). However, these differences can also take place on a much smaller scale, making it important to define accurate exposure assessments. Mobile monitoring is known to be able to capture pollutant data with high spatial resolution; at the same time, determining the representativeness of this data to acquire air quality at high temporal resolution remains a challenge. Due to the high temporal variability of mobile measurements, the spatial and spatial-temporal CV approaches were applied to develop and validate a model with high spatial-temporal resolution using mobile measurement data. The One_second models were used to determine the ability of ML to forecast the concentration of pollutants in different areas not yet assessed. This approach for cross validation can help to test the reliability of the model. Another strength of the spatial-temporal CV approach is that it provides a unique technique for holding out test data. Based on the route that bikers use, I held out the last run of each hour for each spatial cluster to show the ability of the ML algorithm to build a reliable model. This approach could help engineers and planners involved in health and environmental impact assessments to implement robust non-motorized traffic monitoring programs.

Among the three types of cross-validation approaches that were applied to PN models, it was found that the One_second model incorporating all data for random CV and incorporating weather only for spatial and spatial-temporal CV was the best fit model (Table 4.42). Note that the MAE (RMSE) values increased from 3328.32 pt /cm^3 (28423.16 pt /cm^3) using the random CV approach, to 5388.7 pt /cm^3 (29815.66 pt /cm^3) using the spatial CV approach, and to 8542.3 pt /cm^3 (34110.93 pt /cm^3) using the spatial-temporal CV approach from random CV. The significant increase in error data between the random CV and spatial-temporal CV approaches indicates spatial-temporal overfitting. We attribute this outcome to the strategy of removing data from the same spatial-temporal cluster in the training set, which causes the results to be less

robust compared to a spatial CV where more data are available for training the model. Different algorithms demonstrated better performance outcomes depending on each type of cross validation. For example, the Gradient Boosting algorithm was identified as most efficient for random CV; however, when I leveled up restrictions in the test set data, the Lasso and ElasticNet algorithms performed better.

These trends were found to deviate in the case of the black carbon models; specifically, the spatial CV models performed worse in comparison to the random hold-out models. In contrast, the spatial-temporal CV models performed better in comparison to the spatial CV models, which represents an unexpected finding. This outcome indicates that the spatial CV test dataset featured noise, which was reduced when I increased the limitation of the hold-out set; thus, with a reduction in test set size, the noise decreased. This finding can be attributed to the inherent noise associated with the microaethalometer that was used to collect black carbon concentration data.

In conclusion, holding out spatial and temporal data may be helpful in understanding the correlation between CV bias and error underestimation; it must be noted that the systematic validation approach may be very sensitive to test size. However, predictions may be enhanced by carefully considering the spatial-temporal dimension in training. The results detailed herein support the importance of defining validation strategies; indeed, the results obtained from this investigation confirm that the type of validation technique(s) plays a critical role in evaluating a model's reliability.

Finally, the effects of time-dependent variables on a model's performance were most pronounced in the random cross validation models. However, the use of spatial and spatial-

temporal cross validation was intended to eliminate the autocorrelation problems caused by autocorrelation variables (see Sections 4.1.2, 4.2.2, and 4.2.4).

4.3. Model Results by Averaging Time

In order to investigate the effectiveness and efficiency of machine learning to detect the pollutant concentration using high resolution data and assess its ability to use raw measurement data without data reprocessing, I varied averaging times of the mobile measurements. Therefore, several time-averaging models were implemented to compare the performance between the 1 sec measurement models and the longer duration averaging models. Specifically, one type of Daily_average model was developed based on spatial data (i.e., land-use data) and seven types of short-term models were developed according to different combinations of time-dependent variables. Therefore, the conclusions discussed in this section emerged from comparing the two model types. The first comparison targets the models that were developed from the same input variables, while the second comparison is based on long-term models versus short-term models that were built using all input variables.

4.3.1. Long-Term Vs Short-Term Spatial Data Models

This section describes the methodology used to achieve comparative consistency in evaluating the models. For the Daily_average model, averaging the temporal data across 12 hours helped to eliminate any variations caused by weather factors. On both short-term models (i.e., Hourly_average and One_second models), the model based on land use as an input data was identified as the poorest-performing model. However, in order to thoroughly explore the differences between the time-averaging models, it is essential to compare models based on the same input variables. Since the three types of time-averaging models were developed according

to two type of cross validation techniques (random and spatial CV), the conclusions discussed herein are limited to those two approaches.

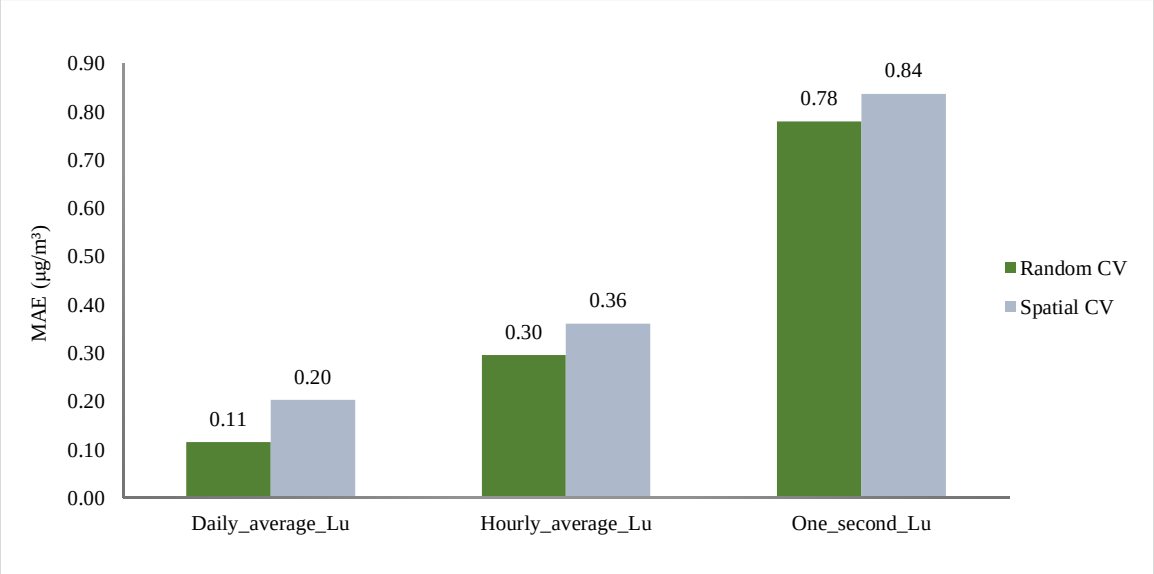

Figure 4.7. Black carbon model performance build based on spatial data only.

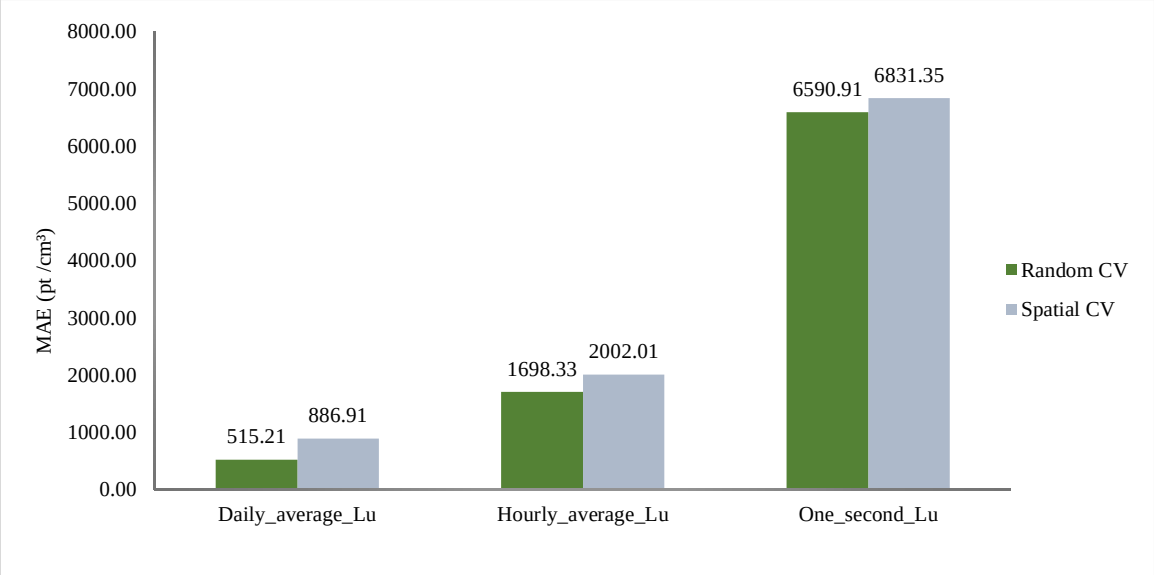

Figure 4.8. Particle number model performance build based on spatial data only.

As indicated in Figures 4.7 and 4.8, based on cross-validation data for both pollutants, the Daily_average model clearly performed better in comparison to the Hourly_average model, followed by the One_second model. It should be noted that the error differences between the

Daily_average model and Hourly_average model were not significant. In contrast, there were notable differences between the Hourly_average model and the One_second model. This trend was not surprising since both Daily_average and Hourly_average models underwent time series preprocessing or smoothing in order to eliminate any random variance that tends to be present in data collected over time. It should also be noted that sample size and a lack of spatiotemporal predictor variables may have contributed to differences in model fit between the time-average models.

4.3.2. Long-Term Vs Short-Term Models Based on All Input Variables.

Comparing the long-term model with the short-term model built using all input variables was undertaken to determine the advantages of developing one or more models based on the best representative variables—even in the absence of data preprocessing that should be used in traditional statistical analyses.

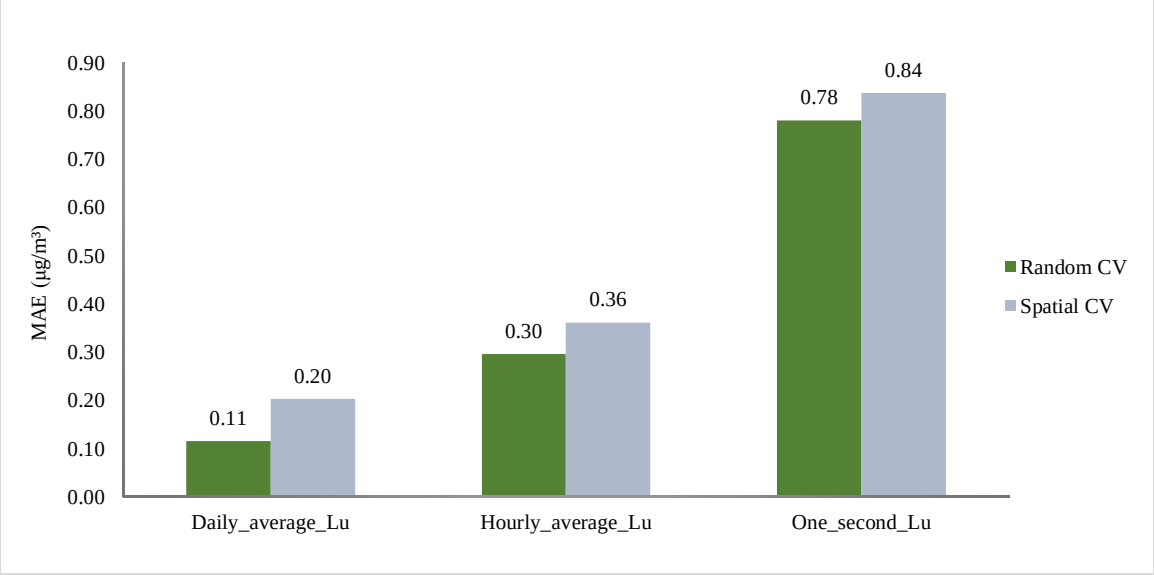

Figure 4.9. Black carbon models include all input variables.

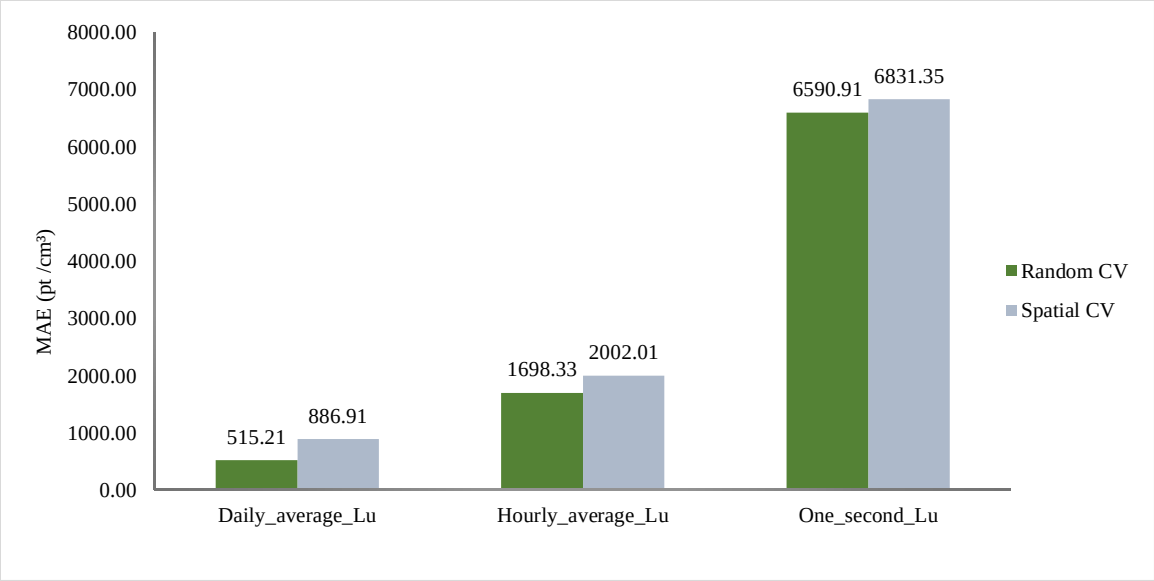

Figure 4.10. Particle number models includes all input variables.

According to Figures 4.9 and 4.10, the performance trends for the particle number models were found to be consistent with results detailed in Section 4.3.1. In particular, the Daily_average model demonstrated the best model performance, followed by the Hourly_average model, and then the One_second model. This finding is consisting with previous research conducted by Hankey et al. (2019), who indicated that model fit for the Hourly_average models (10-fold CV R2: 0.4 for PN and 0.27 for BC) was worse than the Daily_average average models (PN: 0.695; BC: 0.57) for both pollutants.

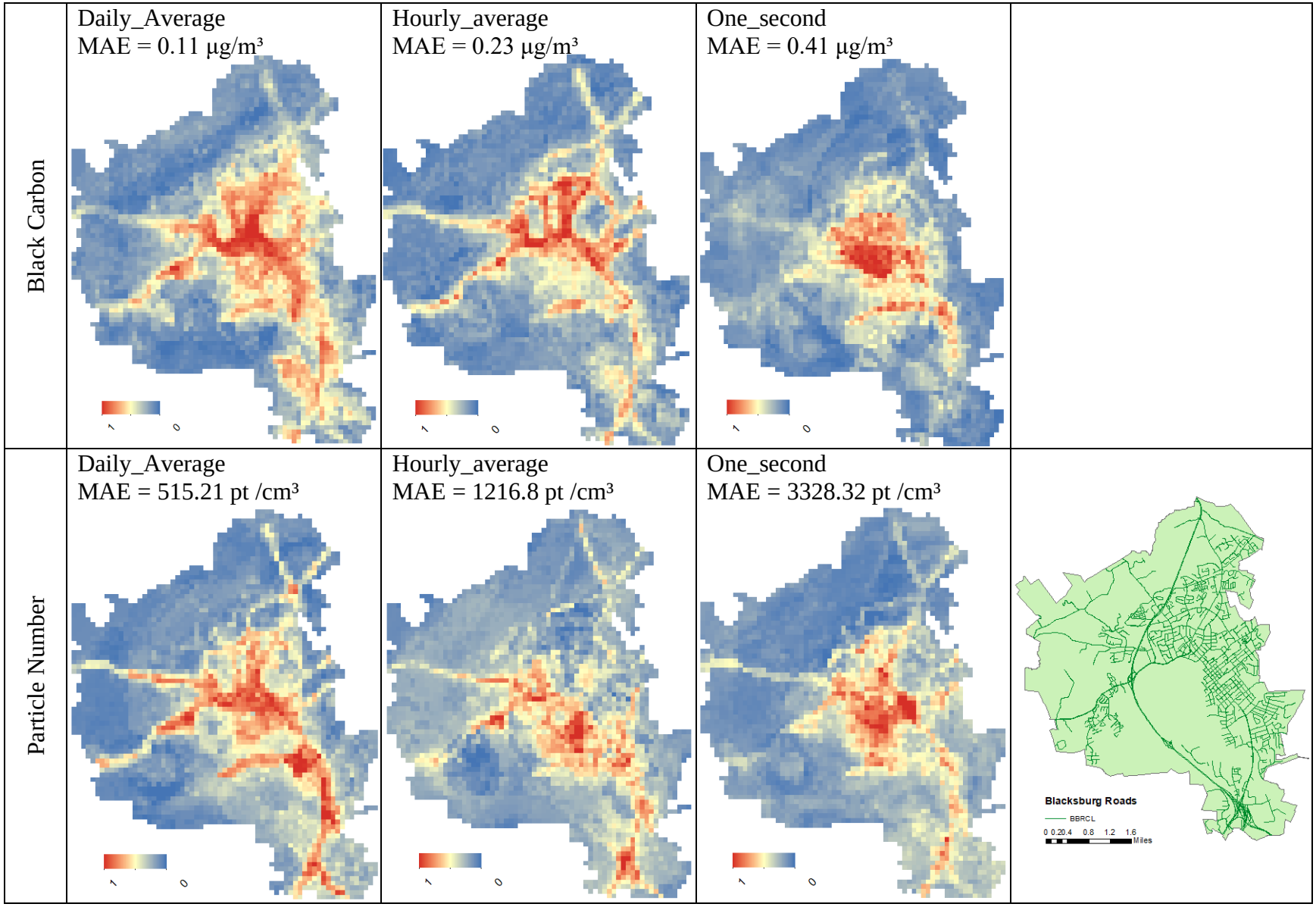

Figure 4.11. Models estimated standardized concentrations from the three time-averaging models for each pollutant.

According to Figure 4.11, all three time-averaging (Daily_average, Hourly_average, and One_second) produced models with similar spatial patterns; however, the disaggregated model (One_second) could detect hotspot areas. However, the One_second models have difficulty to detect the road pollutant concentration that the Daily and Hourly average got.

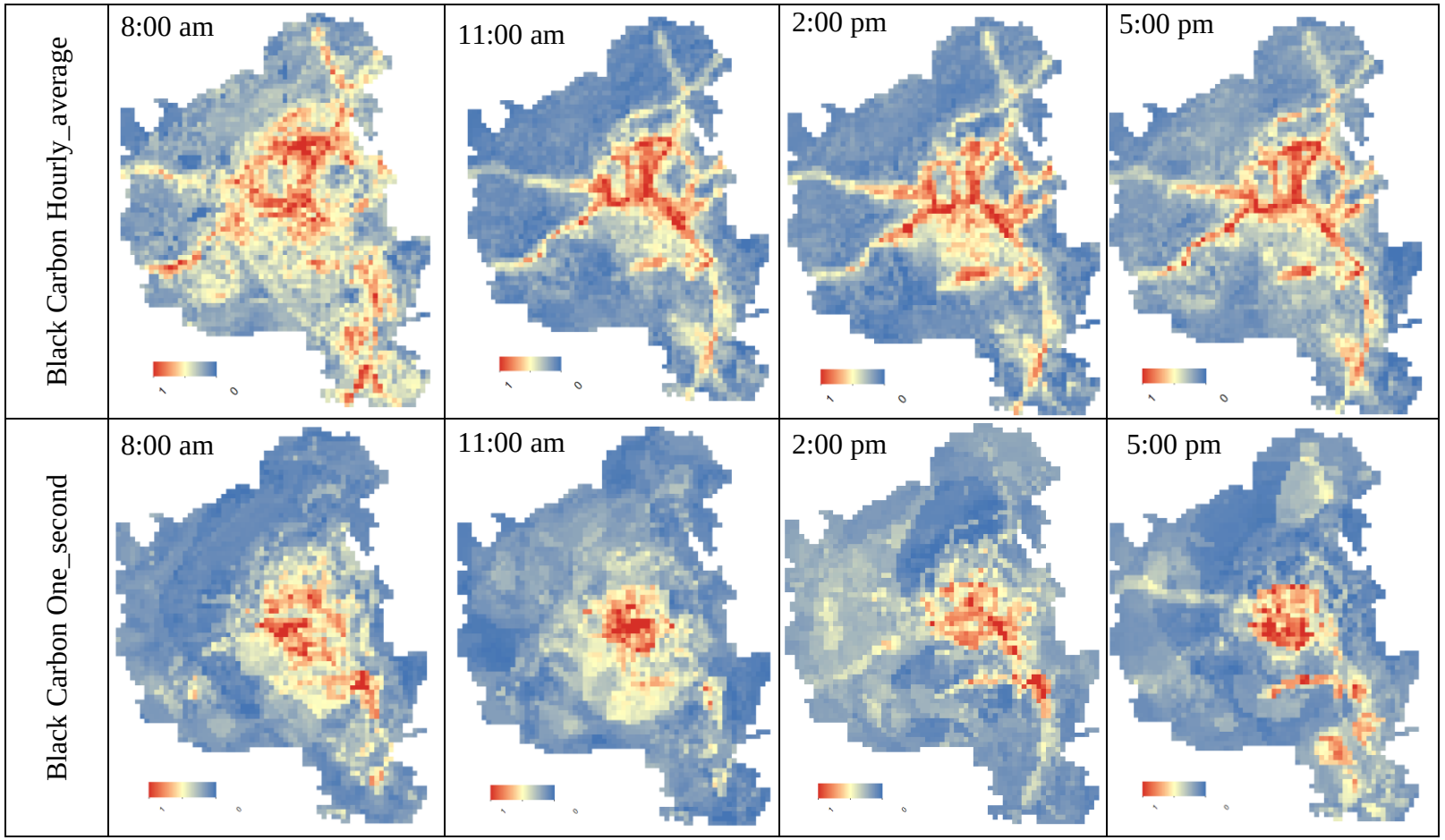

Figure 4.12. Model estimated BC standardized concentrations from the Hourly_average and One_second models for select hours of day

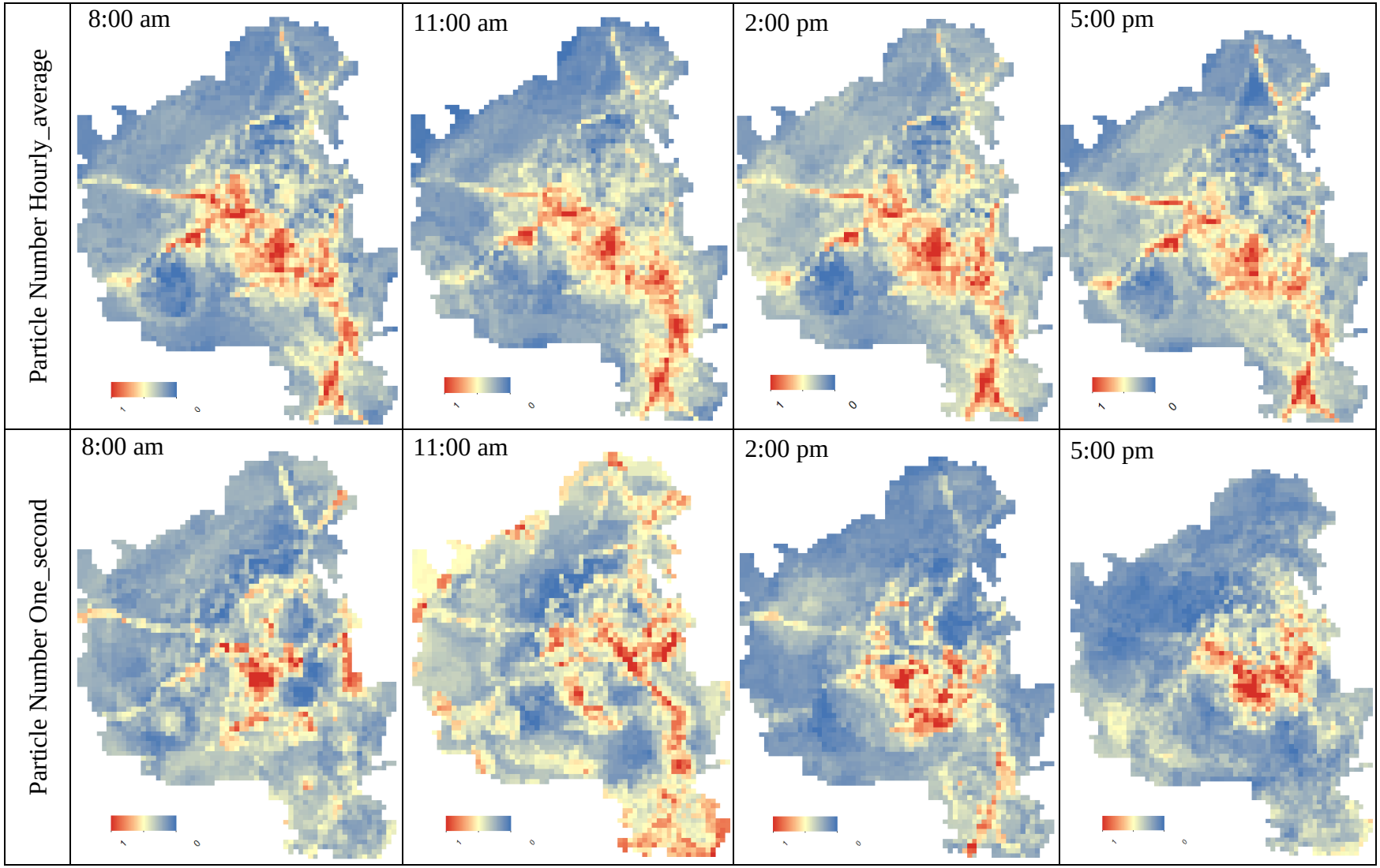

Figure 4.13. Model estimated PN standardized concentrations from the Hourly_average and One_second models for select hours of day

To illustrate spatial prediction from the short-term models, concentration estimates were mapped for selected hours on a 100 m × 100 m grid in Blacksburg (Figure 4.12 and 4.13). The red areas of the figures show that elevated concentrations of both BC and PN could be found in crowded residential and industrial districts, as well as at road interchanges and along larger interstates. These figures also illustrate how pollutant concentration may be higher at certain times of the day than at others.

In conclusion, incorporating smoothing techniques has the advantage of removing noise and outliers in time-series data. This type of data smoothing is based on the concept of identifying simplified changes that will help to predict various trends and patterns with greater accuracy. By implementing the spatial-temporal smoothing process in the Hourly_average models, this type of aggregation facilitates model fitting even when the number of observations is large.

4.4. Machine Learning vs Statistical Model

The primary distinction between non-parametric machine learning and parametric statistics is its intended use. Machine learning models use input data to make accurate predictions based on that data. In contrast, statistical models are mathematical equations designed to develop inferences about the relationships between variables. Employing 10-fold random CV R^2, Hankey et al. (2019) compared the performance between machine learning model and stepwise regression models using both Daily_average and Hourly_average models. The main goal of their comparative study was to determine the ability of machine learning to build more accurate models without relying on rules-based programming. In order to maintain consistency, the comparison of the Hourly_average model was based on the

Hourly_average_Lu_W_Hr models, which featured the same input variables used to build the stepwise regression model.

Table 4.43.

Comparison between Machine Learning and Stepwise Regression Performance for Each Pollutant

Pollutant	Models	Stepwise R^2	ML R^2
BC: µg/m³	Daily_average (Lu)	0.58	0.734
	Hourly_average (Lu+W+Hr)	0.27	0.54
PN: particles/cm³	Daily_average (Lu)	0.7	0.78
	Hourly_average (Lu+W+Hr)	0.42	0.54

Overall, it is important to note that the machine learning models performed better than the stepwise regression across all models (Table 4.43), which supports the potential utility of machine learning models. Performance improvements varied between BC and PN data, as well as whether short-term or long-term models were used for each pollutant. The stepwise 10-fold random CV R^2 of the Daily_average_Lu model was found to be 0.57 for the BC model, and 0.7 for the PN model. Thus, the use of machine learning models improved the 10-fold CV R^2 results to 26.5% and 11.4% for BC and PN, respectively.

The stepwise regression Hourly_average model fit better for PN concentration (10-fold CV R2 is 0.42) in comparison to BC concentration (10-fold CV R2 is 0.27), but worse when compared to the Daily_average model results for both pollutants (BC: 0.57; PN: 0.7). Also, the machine learning model was found to determine BC performance with greater precision in comparison to analogous findings for PN; specifically, ML increased the model 10-fold CV R^2 to twice for BC and 30% for the PN, which ended up with the same 10-fold CV R^2 results (0.54 for

both BC and PN). In terms of additional BC-related findings, the machine learning model for the Hourly_average model was as good as the stepwise regression of the Daily_average model. This finding demonstrates the ability of ML to produce more accurate short-term models that are equivalent with traditional long-term models. When comparing the two models, machine learning was shown to be more effective in developing BC prediction model in comparison to the PN prediction model.

Chapter 5: Conclusions and Future Work

Three time-interval models (two short-term models and one long-term model) were used to estimate the concentrations of the pollutants PN and BC in Blacksburg (VA) utilizing mobile-monitoring data involving several spatiotemporal variables. The main goal of this study was to test different data reprocessing approaches for the mobile monitoring data, coupled with various combinations of input variables, to assess the ability of machine learning to improve prediction accuracy using high-temporal resolution monitoring data (1-sec).

Researchers have been investigating how to obtain accurate exposure data by applying machine learning algorithms to available mobile monitoring data (Su et al., 2015b). Therefore, it is important to improve the spatial and temporal resolution of pollutant concentration to enhance the predictive capacity of machine learning (Hankey et al., 2019). Multiple studies have used machine learning to produce accurate prediction with high spatial-temporal resolution (Bellinger et al., 2017). Nonetheless, while mobile monitoring is able to produce data with high spatial resolution, its ability to provide similarly rich temporal resolution information for the same monitoring location is limited, especially over shorter monitoring periods. Therefore, researchers must continue to explore the performance of machine learning in the presence of different data pre-processing approaches in order to develop accurate models using high temporal resolution mobile monitoring data (one-second measurement data).

This study compared the performance of the machine learning models of the refined versus the spatially and temporally smoothed input data. The results demonstrated the consistency of the three time-average models in detecting the spatial distribution of PN and BC. Specifically, the One_second model was able to capture hotspot areas with the same equivalency as the Daily_average and the Hourly_average models. Furthermore, there were no significant

differences in model error across the three time-average models, indicating the ability of machine learning to handle noisy, unprocessed mobile monitoring data. It must be noted that this study used secondary data captured during the summer and fall within a smaller, more rural town with less urban diversity. Thus, the differences in spatial and temporal factors were not found to be significant. A follow-on study is recommended to expand these findings by conducting a similar study in a more urban locale and for a longer period.

Studies have suggested adding more independent variables to improve a model's performance (Shaban et al., 2016). However, the more variables feeding to the model, the greater the likelihood that overfitting will occur. However, Bertazzon et al. (2021) have recently shown that the predictive capacity of ML models can be improved by adding meteorological variables (Bertazzon et al., 2021). An essential contribution of this study pertains to how spatiotemporal variability can impact model performance. Specifically, we used the INTERGRATION simulation software to calculate light-duty vehicle emissions and used this as an input variable to estimate Hourly pollutant concentration. Results showed that incorporating this particular spatiotemporal variable, vehicle emissions, significantly enhanced the prediction accuracy of the algorithm. However, various combinations of input variables, including vehicle emissions, were tested to assess the how machine learning could be more efficient. The findings from this study showed that no significant improvements could be found from incorporating the different variables to the spatial-temporal models (Daily_average and Hourly_average). Specifically, time-dependent variables did not result in any significant improvements in the aggregated models, likely due to the fact that these variables were temporally aggregated 1 hour, which led to smoothing the data and decreasing the hour-by-hour variance.

In contrast, with respect to the highly temporal One_second resolution models, its predictive capacity improved significantly by adding time-dependent variables in the highly temporal resolution models. However, in both cases, the performance of the model that incorporated all variables was equivalent to that of the model that does not use vehicle emissions as an input variable. The correlation between land-use data and meteorological variables is a powerful predictor for modeling pollutant concentration, which is supported by prior findings (Hankey et al., 2019; Pandey et al., 2013a). It is important to note that the emission factor was calculated only with regard to light-duty vehicles. Thus, a similar study using emissions data from heavy duty vehicles could be used to expand our knowledge of the predictive capacity of machine learning models.

Many of the land-use studies have applied random validation to test the reliability and uncertainty of the model. However, it has been shown that applying another strategy—namely cross-validation—could help to overcome spatio-temporal autocorrelation (Ceci et al., 2017). Therefore, three types of cross-validation approaches were applied to show the ability of machine learning to forecast pollution levels in an area not previously assessed. The different cross-validation approaches could help to address problems of overestimation that probably occurred in the random cross-validation model. In general, prediction mean absolute errors increase from random to spatial cross-validation. This difference indicates that the model became less capable of predicting beyond the training data's immediate location compared to what might have been expected using random CV. The model’s mean absolute error rate increased from using spatial cross-validation to using spatial-temporal cross-validation. This significant increase in error data between the random CV and spatial-temporal CV approaches indicates spatial-temporal overfitting. Accordingly, this study also explored improvements resulting from using

independent spatiotemporal variables on spatial and spatial-temporal cross-validation. Based on results associated with the short-term model, holding out a location and specific time in the training process was not improved by adding temporal or spatiotemporal variables. This finding was expected since meteorological data and emissions factors are time-dependent variables; thus, holding out one location/time was incorporated to eliminate the time factors effect. However, the use of spatial and spatial-temporal cross-validation was intended to eliminate the autocorrelation problems caused by autocorrelation variables. However, the number of bike routes used to collect pollutant concentration levels was limited to the lack of students on bikes during the period of data collection. Therefore, it would be helpful to determine how many routes could be used to achieve stable prediction concentration. Moreover, during summer there is less traffic during the summer months in a collage town therefor, the data was not representative of other periods with higher pollutant concentrations. This research could be expanded to apply in more seasons to detect the pollutant concentration in different level.

Finally, comparing machine learning models and traditional statistical models (stepwise regression) was undertaken to assess the ability of using machine learning to model accuracy. Our findings indicate that machine learning demonstrated higher predictive capacity with respect to stepwise regression. Additionally, models used to predict black carbon levels were more accurate that those applied to predicting particle number.

REFERENCES

Abernethy, R. C., Allen, R. W., McKendry, I. G., & Brauer, M. (2013). A land use regression model for ultrafine particles in Vancouver, Canada. *Environmental Science and Technology, 47*(10), 5217–5225. https://doi.org/10.1021/es304495s

Anenberg, S. C., Henze, D. K., Tinney, V., Kinney, P. L., Raich, W., Fann, N., Malley, C. S., Roman, H., Lamsal, L., Duncan, B., Martin, R. v., van Donkelaar, A., Brauer, M., Doherty, R., Jonson, J. E., Davila, Y., Sudo, K., & Kuylenstierna, J. C. I. (2018). Estimates of the global burden of ambient PM2:5, ozone, and NO2 on asthma incidence and emergency room visits. *Environmental Health Perspectives, 126*(10), 1–14. https://doi.org/10.1289/EHP3766

Anselin, L. (1998). *Spatial econometrics: Methods and models*. Kluwer: New York.

Baets, B. de. (2019). Development of a land use regression model for black carbon using mobile monitoring data and its application. *Environmental Research,* 108619. https://doi.org/10.1016/j.envres.2019.108619

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., ... & de Hoogh, K. (2013). Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe–the ESCAPE project. Atmospheric Environment, 72, 10-23.

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health, 17(1), 1–19. https://doi.org/10.1186/s12889-017-4914-3

Bertazzon, S., Couloigner, I., & Mirzaei, M. (2021). Spatial regression modelling of particulate pollution in Calgary, Canada. *GeoJournal, 7*(Epa 2016). https://doi.org/10.1007/s10708-020-10345-7

Bertazzon, S.; Johnson, M.; Eccles, K.; & Kaplan, G.G. (2015). Accounting for spatial effects in land use regression for urban air pollution modeling. *Spatial and Spatio-temporal Epidemiology*, 14-15, 9-21. doi:10.1016/j.sste.2015.06.002

Bové, H., Bongaerts, E., Slenders, E., Bijnens, E. M., Saenen, N. D., Gyselaers, W., van Eyken, P., Plusquin, M., Roeffaers, M. B. J., Ameloot, M., & Nawrot, T. S. (2019). Ambient black carbon particles reach the fetal side of human placenta. *Nature Communications, 10*(1), 1–7. https://doi.org/10.1038/s41467-019-11654-3

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R. v., Dentener, F., Dingenen, R. van, Estep, K., Amini, H., Apte, J. S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P. K., Knibbs, L. D., Kokubo, Y., Liu, Y., … Cohen, A. (2016). Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. *Environmental Science and Technology, 50*(1), 79–88. https://doi.org/10.1021/acs.est.5b03709

Brokamp, C., Jandarov, R., Rao, M. B., LeMasters, G., & Ryan, P. (2017). Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmospheric Environment,* 151, 1-11.

Ceci, M., Corizzo, R., Fumarola, F., Malerba, D., & Rashkovska, A. (2016). Predictive modeling of PV energy production: How to set up the learning task for a better prediction?. *IEEE Transactions on Industrial Informatics, 13*(3), 956-966.

Chen, M., Wang, P., Chen, Q., Wu, J., & Chen, X. (2015). A clustering algorithm for sample data based on environmental pollution characteristics. *Atmospheric Environment, 107*, 194–203. https://doi.org/10.1016/j.atmosenv.2015.02.042

Coker, E. S., Amegah, A. K., Mwebaze, E., Ssematimba, J., & Bainomugisha, E. (2021). A Land Use Regression Model using Machine Learning and Locally Developed Low Cost Particulate Matter Sensors in Uganda. *Environmental Research*, 111352.

Correia, A. W., Pope III, C. A., Dockery, D. W., Wang, Y., Ezzati, M., & Dominici, F. (2013). The effect of air pollution control on life expectancy in the United States: an analysis of 545 US counties for the period 2000 to 2007. Epidemiology (Cambridge, Mass.), 24(1), 23.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., ... & Schwartz, J. (2019). Assessing NO2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. *Environmental Science & Technology, 54*(3), 1372-1384.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *5*5(10), 78–87. https://doi.org/10.1145/2347736.2347755

Gaspar, F. W., Maddalena, R., Williams, J., Castorina, R., Wang, Z. M., Kumagai, K., McKone, T. E., & Bradman, A. (2018). Ultrafine, fine, and black carbon particle concentrations in California child-care facilities. *Indoor Air, 28*(1). https://doi.org/10.1111/ina.12408

Gilbert, N. L., Goldberg, M. S., Beckerman, B., Brook, J. R., & Jerrett, M. (2005). Assessing spatial variability of ambient nitrogen dioxide in montréal, canada, with a land-use regression model. *Journal of the Air and Waste Management Association, 55*(8), 1059–1063. https://doi.org/10.1080/10473289.2005.10464708

Götschi, T., Garrard, J., & Giles-corti, B. (2015). Cycling as a Part of Daily Life : A Review of Health Perspectives Transport Reviews : A Transnational Cycling as a Part of Daily Life : A Review of Health Perspectives. June. https://doi.org/10.1080/01441647.2015.1057877

Haberlandt, U. (2007). Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event. *Journal of Hydrology, 332*(1–2), 144–157. https://doi.org/10.1016/j.jhydrol.2006.06.028

Hagler, G. S. W., Lin, M., Khlystov, A., Baldauf, R. W., Isakov, V., Faircloth, J., & Jackson, L. E. (2012). Field investigation of roadside vegetative and structural barrier impact on near-road ultrafine particle concentrations under a variety of wind conditions Science of the Total Environment Field investigation of roadside vegetative and structural barrier i.

Science of the Total Environment, *419*(March), 7–15. https://doi.org/10.1016/j.scitotenv.2011.12.002

Hankey, S., & Marshall, J. D. (2015). On-bicycle exposure to particulate air pollution: Particle number, black carbon, PM2.5, and particle size. Atmospheric Environment, 122, 65–73. https://doi.org/10.1016/j.atmosenv.2015.09.025

Hankey, S., Sforza, P., & Pierson, M. (2019). Using Mobile Monitoring to Develop Hourly Empirical Models of Particulate Air Pollution in a Rural Appalachian Community. Environmental Science and Technology, 53(8), 4305–4315. https://doi.org/10.1021/acs.est.8b05249

Hasenfratz, D., Saukh, O., Walser, C., Hueglin, C., Fierz, M., Arn, T., Beutel, J., & Thiele, L. (2015). Deriving high-resolution urban air pollution maps using mobile sensor nodes. *Pervasive and Mobile Computing*, *16*, 268–285. https://doi.org/10.1016/j.pmcj.2014.11.008

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*(33), 7561-7578.

Hu, K., Rahman, A., Bhrugubanda, H., & Sivaraman, V. (2017). HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation from Fixed and Mobile Sensors. *IEEE Sensors Journal*, *17*(11), 3517–3525. https://doi.org/10.1109/JSEN.2017.2690975

Hudda, N., Gould, T., Hartin, K., Larson, T. v., & Fruin, S. A. (2014). Emissions from an international airport increase particle number concentrations 4-fold at 10 km downwind. *Environmental Science and Technology*, *48*(12), 6628–6635. https://doi.org/10.1021/es5001566

Johnson, M., Isakov, V., Touma, J. S., Mukerjee, S., & Özkaynak, H. (2010). Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment*, *44*(30), 3660-3668.

Kerckhoffs, J., Hoek, G., Vlaanderen, J., van Nunen, E., Messier, K., Brunekreef, B., ... & Vermeulen, R. (2017). Robustness of intra urban land-use regression models for ultrafine particles and black carbon based on mobile monitoring. *Environmental research*, *159*, 500-508.

Khorrami, Z., Pourkhosravani, M., Rezapour, M., & Etemad, K. (2021). Multiple air pollutant exposure and lung cancer in Tehran , Iran. *Scientific Reports*, 1–11. https://doi.org/10.1038/s41598-021-88643-4

Kirchstetter, T. W., et al. (1999). "On-road measurement of fine particle and nitrogen oxide emissions from light-and heavy-duty motor vehicles." Atmospheric Environment 33(18): 2955-2968.

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. *Journal of Electrical and Computer Engineering*, 2017. https://doi.org/10.1155/2017/5106045

Knibbs, L. D., Hewson, M. G., Bechle, M. J., Marshall, J. D., & Barnett, A. G. (2014). A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environmental Research, 135*, 204-211.

Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., ... & Hystad, P. (2017). Global land use regression model for nitrogen dioxide air pollution. *Environmental science & technology*, *51*(12), 6957-6964.

Lim, C. C., Kim, H., Vilcassim, M. R., Thurston, G. D., Gordon, T., Chen, L. C., ... & Kim, S. Y. (2019). Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. *Environment international*, *131*, 105022.

Liu, M., Peng, X., Meng, Z., Zhou, T., Long, L., & She, Q. (2019). Spatial characteristics and determinants of in-traffic black carbon in Shanghai, China: Combination of mobile monitoring and land use regression model. *Science of the Total Environment, 658*, 51–61. https://doi.org/10.1016/j.scitotenv.2018.12.135

Meng, X., Chen, L., Cai, J., Zou, B., Wu, C. F., Fu, Q., ... & Kan, H. (2015). A land use regression model for estimating the NO2 concentration in Shanghai, China. *Environmental Research*, 137, 308-315.

Messier, K. P., Chambliss, S. E., Gani, S., Alvarez, R., Brauer, M., Choi, J. J., Hamburg, S. P., Kerckhoffs, J., Lafranchi, B., Lunden, M. M., Marshall, J. D., Portier, C. J., Roy, A., Szpiro, A. A., Vermeulen, R. C. H., & Apte, J. S. (2018). Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression. *Environmental Science and Technology, 52*(21), 12563–12572. https://doi.org/10.1021/acs.est.8b03395

Melymuk, L., Robson, M., Helm, P. A., & Diamond, M. L. (2013). Application of land use regression to identify sources and assess spatial variation in urban SVOC concentrations. Environmental science & technology, 47(4), 1887-1895.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling and Software*, 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Morley, D. W., & Gulliver, J. (2018). A land use regression variable generation, modelling and prediction tool for air pollution exposure assessment. Environmental Modelling & Software, 105, 17-23. Retrieved from

Oberdörster, G. (2000). Pulmonary effects of inhaled ultrafine particles. In *International Archives of Occupational and Environmental Healt*h (Vol. 74, Issue 1). https://doi.org/10.1007/s004200000185

Oliveira, M., Torgo, L., & Santos Costa, V. (2019). Evaluation procedures for forecasting with spatio-temporal data. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11051 LNAI(Cv), 703–718. https://doi.org/10.1007/978-3-030-10925-7_43

Paasonen, P., Asmi, A., Petäjä, T., Kajos, M. K., Äijälä, M., Junninen, H., Holst, T., Abbatt, J. P. D., Arneth, A., Birmili, W., van der Gon, H. D., Hamed, A., Hoffer, A., Laakso, L., Laaksonen, A., Richard Leaitch, W., Plass-Dülmer, C., Pryor, S. C., Räisänen, P., … Kulmala, M. (2013). Warming-induced increase in aerosol number concentration likely to moderate climate change. *Nature Geoscience, 6*(6). https://doi.org/10.1038/ngeo1800

Pandey, G., Zhang, B., & Jian, L. (2013). Predicting submicron air pollution indicators: A machine learning approach. *Environmental Sciences: Processes and Impacts, 15*(5), 996–1005. https://doi.org/10.1039/c3em30890a

Pétron, G., Frost, G., Miller, B. R., Hirsch, A. I., Montzka, S. A., Karion, A., Trainer, M., Sweeney, C., Andrews, A. E., Miller, L., Kofler, J., Bar-Ilan, A., Dlugokencky, E. J., Patrick, L., Moore, C. T., Ryerson, T. B., Siso, C., Kolodzey, W., Lang, P. M., … Tans, P. (2012). Hydrocarbon emissions characterization in the Colorado Front Range: A pilot study. *Journal of Geophysical Research Atmospheres, 117*(4), 1–19. https://doi.org/10.1029/2011JD016360

Petzold, A., Ogren, J. A., Fiebig, M., Laj, P., Li, S. M., Baltensperger, U., Holzer-Popp, T., Kinne, S., Pappalardo, G., Sugimoto, N., Wehrli, C., Wiedensohler, A., & Zhang, X. Y. (2013). Recommendations for reporting black carbon measurements. *Atmospheric Chemistry and Physics, 13*(16), 8365–8379. https://doi.org/10.5194/acp-13-8365-2013

Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., ... & Hoek, G. (2013). Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *The lancet oncology, 14*(9), 813-822.

Rakha, H., & Ahn, K. (2004). Integration modeling framework for estimating mobile source emissions. *Journal of Transportation Engineering, 130*(2), 183–193. https://doi.org/10.1061/(ASCE)0733-947X(2004)130:2(183)

Razavi-Termeh, S. V., Sadeghi-Niaraki, A., & Choi, S. M. (2021). Effects of air pollution in Spatio-temporal modeling of asthma-prone areas using a machine learning model. *Environmental* Research, 111344.

Reggente, M., Mondini, A., Ferri, G., Mazzolai, B., Manzi, A., Gabelletti, M., Dario, P., & Lilienthal, A. J. (2010). The DustBot system: Using mobile robots to monitor pollution in pedestrian area. *Chemical Engineering Transactions, 23*, 273–278. https://doi.org/10.3303/CET1023046

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Ryan, P. H., & LeMasters, G. K. (2007). A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation Toxicology, 19*(sup1), 127-133. doi:10.1080/08958370701495998

Shaban, K. B., Kadri, A., & Rezk, E. (2016). Urban air pollution monitoring system with forecasting models. *IEEE Sensors Journal*, *16*(8), 2598-2606.

Stingone, J. A., Pandey, O. P., Claudio, L., & Pandey, G. (2017). Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among us children. *Environmental Pollution*, *230*, 730-740.

Su, J. G., Hopke, P. K., Tian, Y., Baldwin, N., Thurston, S. W., Evans, K., & Rich, D. Q. (2015). Modeling particulate matter concentrations measured through mobile monitoring in a deletion / substitution / addition approach. *Atmospheric Environment*, *122*, 477–483. https://doi.org/10.1016/j.atmosenv.2015.10.002

Wang, A., Xu, J., Tu, R., Saleh, M., & Hatzopoulou, M. (2020). Potential of machine learning for prediction of traffic related air pollution. *Transportation Research Part D: Transport and Environment*, 88, 102599. https://doi.org/10.1016/j.trd.2020.102599

Weichenthal, S., Van Ryswyk, K., 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. *Environmental research*, *146*, 65-72.

Wu, J., Li, J., Peng, J., Li, W., Xu, G., & Dong, C. (2015). Applying land use regression model to estimate spatial variation of PM2.5 in Beijing, China. *Environmental Science and Pollution Research*, *22*(9), 7045–7061. https://doi.org/10.1007/s11356-014-3893-5

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., ... & Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM2. 5 concentrations across China using a spatially explicit machine learning algorithm. *Atmospheric Environment*, *155*, 129-139.