Neighborhood-level NO_2 Machine Learning Models Toward Sustainable Monitoring Systems

by Janay A.W. Frazier-Duncan

B.S. in Optical Engineering, May 2016, Norfolk State University
M.S. in Electrical Engineering, May 2018, Virginia Polytechnic Institute and State University

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

January 6th, 2023

Praxis directed by

Joseph Blackford
Professorial Lecturer of Engineering Management and Systems Engineering

The School of Engineering and Applied Science of The George Washington University certifies that Janay Frazier-Duncan has passed the Final Examination for the degree of Doctor of Engineering as of December 2, 2022. This is the final and approved form of the praxis.

Neighborhood-level NO_2 Machine Learning Models Toward Sustainable Monitoring Systems

Janay A. W. Frazier-Duncan

Praxis Research Committee:

Joseph Blackford, Professorial Lecturer in Engineering and Applied Science, Praxis Director

Steven Biemer, Professorial Lecturer in Engineering and Applied Science, Committee Member

Amir Etemadi, Associate Professor of Engineering and Applied Science, Committee Member

© Copyright 2022 by Janay A.W. Frazier-Duncan
All rights reserved

Dedication

I dedicate my research to my family, who has supported me through every phase of my matriculation through earning my doctoral degree. To the love of my life, my loving husband, Henry Duncan, who has been a constant source of encouragement every day, keeping me on schedule academically and in life, for your love has kept me afloat. To the loving memory of my brother, Jamil Rashad Hamilton Frazier, this one is for your dreams. To my loving parents, Mr. George Frazier, and Dr. Deborah Hamilton Frazier, who have given me the building blocks to pursue and push through the times that may get difficult, if it was an uplifting word or academic advice, you have always been there. To my beautiful and loving sister, Jasmine Monet' Brown, who has inspired me every day with her educational journey, and brother-in-law Emmett Brown, God's Blessings to you both for my niece's, Faith, Hope, and Grace for sharing their lively anecdotes in times of struggle. Ultimately, I want to thank my Lord and Savior, Jesus Christ, who has given me the strength to finish my dream of earning my doctorate!

Acknowledgments

I would like to recognize and thank my praxis director, Dr. Joseph Blackford, for his encouragement and support throughout the research process and his belief in my ability to accomplish and complete this journey. I want to thank Martha Kile, Principal Data Analyst, and Team from MWCOG for providing traffic volume count data. I would also like to thank Dr. Rama Tangirala, Chief Air quality Monitoring Branch and Team from Air Quality Division DOEE, for the emissions data. Special gratitude and thanks to my beautiful Hamilton Family, cousins, aunts, and uncles. A sincere appreciation and love to Dr. Akilah Hugine-Elmore, Rachelle Holmes Perkins, J.D., Seirra and Marcia Hamilton, Renee N. Hamilton, and Theresa Hamilton, all of whom have had a significant influence on me in the completion of my research. May God Bless them with joy, peace, and happiness always.

Abstract of Praxis

Neighborhood-level NO_2 Machine Learning Models Toward Sustainable Monitoring Systems

Although many researchers have studied aspects of monitoring and modeling vehicle traffic emissions, more attention is needed for smaller than regional scale monitoring. National Ambient Air Quality Standards (NAAQS) are set by the Environmental Protection Agency (EPA) for six criterion pollutants and use a system of federally recognized analytical techniques to track compliance with those requirements known as Federal Equivalent Methods or Federal Reference Methods (FEMs or FRMs). FEM monitoring networks may be effective for assessing air pollution at a regional spatial resolution, which may not be suitable for monitoring at the neighborhood level due to their high cost and technological complexity. This research aims to bridge the knowledge gap in understanding the types of modeling sufficient for neighborhood-level emissions monitoring and how environmental and traffic engineers use them to enhance regional monitoring system networks. This quantitative study investigates the knowledge and comprehension gap of what is perceived from historical traffic data about the level of emissions occurring from on-road pollution and the ability to forecast levels without purchasing new systems but utilizing existing traffic count stations. The conceptual framework was based on emissions modeling measurements using meteorological data, volume count, and other emissions to assess the ability to build a forecasting model. The research questions were developed with the goal of creating a model that forecasts Nitrogen Dioxide (NO_2) levels in Washington, D.C. Using historical data collected from data produced by the Department of Energy & Environment (DOEE) in support of the

Environmental Protection Agency EPA and the Metropolitan Washington Council of Governments (MWCOG), which describes real-world transportation and traffic data for Washington, D.C., to support the District Department of Transportation (DDOT). Data were modeled using WEKA (a data mining machine learning tool) and analyzed using Python and Minitab. The main conclusions observed are which multiple statistical tests and machine learning (ML) algorithms evaluated had the highest performance, what model features were determined to be the most significant in forecasting NO_2 levels using traffic data and other features for Washington D.C., and based on those models can ranges of NO_2 and their corresponding levels can be forecasted for other smart cities based on the framework developed in this praxis. Based on these findings, the key recommendations are to have Washington, D.C., utilize the model in this praxis by using existing systems to improve the monitoring network.

The study's findings could further advance the understanding of monitoring and modeling of forecasting of hazardous NO_2. This contribution has implications for positive neighborhood-level monitoring system improvements by identifying reactive emissions monitoring practices, which can lead to high-quality continuous reduction in vehicle NO_2 emission impacts.

Table of Contents

Dedication ..... **iv**

Acknowledgments ..... **v**

Abstract of Praxis ..... **vi**

List of Figures ..... **xi**

List of Tables ..... **xiii**

List of Symbols ..... **xiv**

List of Acronyms ..... **xv**

Chapter 1—Introduction **1**

1.1 Background 1

1.2 Research Motivation 1

1.3 Problem Statement 2

1.4 Thesis Statement 3

1.5 Research Objectives 4

1.6 Research Questions and Hypotheses 4

1.7 Scope of Research 5

1.8 Research Limitations 5

1.9 Organization of Praxis 5

Chapter 2—Literature Review **7**

2.1 Introduction 7

2.2 Traffic congestion 7

What is traffic congestion? 7

Why do we have traffic congestion? 8

Bottlenecks and Capacity 9
Work Zones 9
Traffic Incidents 10
Variation in traffic flow 10
Weather 10
Traffic congestion ranking for Washington D.C. 11
2.3 Machine Learning 11
Unsupervised Learning 12
Semi-Supervised Learning 13
Supervised Learning 14
2.4 Feature selection methods for Pollution 18
2.5 Forecasting Vehicle Traffic Emissions 21
2.6 Impacts of Traffic Congestion on Emissions and Health 23
Related Health Effects from on-road pollution 23
What is Particulate Matter? 24
The Harmful Effects of Smog 24
2.7 The Need and Requirements for Traffic Emission Monitoring 25
2.8 Summary and Conclusion 26
Chapter 3—Methodology 27
3.1 Introduction 27
3.2 Data Source Description and Use 29
3.3 Training, Testing, and Cross-Validation 35
Training and Cross-Validation 35

Training and Testing 38

Volume Data 38

3.4 Cross-Validation Methodology 43

Chapter 4—Results 45

4.1 Introduction 45

4.2 Data Descriptive Statistics and Missing Values 45

4.3 Feature Selection Results 47

4.4 Model Results 55

4.5 Improving Results 59

4.6 Testing on Richmond, VA 62

Chapter 5—Discussion and Conclusions 65

5.1 Discussion 65

5.2 Conclusions 65

5.3 Contributions to Body of Knowledge 67

5.4 Recommendations for Future Research 68

References 70

Appendix 84

List of Figures

Figure 2-1. Machine Learning paradigms and algorithms, and applications 12

Figure 2-2. Unsupervised Learning Descriptive Image 13

Figure 2-3. Semi-supervised Learning Descriptive Image 13

Figure 2-4. Supervised Learning of Descriptive Image 14

Figure 2-5. Random Forest Algorithm Diagram 16

Figure 2-6. Bagging and Boosting 17

Figure 2-7. Bagging Ensemble Classifier Methodology 17

Figure 2-8. SHAP Value from a NO_2 forecasting model 19

Figure 3-1. Research Methodology Diagram 28

Figure 3-2. Air Quality Monitoring Stations in Washington, D.C. 30

Figure 3-3. Traffic Monitoring Stations in Washington, D.C. 31

Figure 3-4. 5-fold cross-validation image 43

Figure 4-1. NO_2 vs. Month of the year 48

Figure 4-2. NO_2 versus Days of the Week 49

Figure 4-3. NO_2 vs SO_2 concentration 51

Figure 4-4. Temperature vs. NO_2 52

Figure 4-5. NO2 vs. Volume for 2017 53

Figure 4-7. The output of Correlation Attribute Evaluation 54

Figure 4-6. NO_2 vs. Wind Direction 54

Figure 4-8. Linear Regression Model Equation from WEKA and Minitab 57

Figure 4-9. NO_2 ppm vs. Actual and Predicted levels for 2018 using Multiple Linear Regression 58

Figure 4-10. NO_2 ppm vs. Actual and Predicted levels for 2018 using Random Forest Algorithm 58

Figure 4-11. Improved Regression Equation for (a) Spring (b) Summer (b) Fall (c) Winter 61

Figure 4-12. scatter plots from Richmond, VA, and Washington, D.C. model..... 63

Figure 5-1. a) Reference monitor locations in washinton D.C.with Ward map 69

b) Traffic monitors with Ward map. 69

Figure A-1. Input file for the 2017 Training Year Excel and WEKA format 95

List of Tables

Table 2-1. Emissions Modeling Comparison of performance NO_2 23

Table 2-2. Effects of Nitrogen Dioxide on Human Health (Hussain, 2007) 25

Table 3-1. Performance Metric Evaluation Measures 36

Table 3-2. File Format: Hourly Traffic by State by Year CSV Files 39

Table 3-3. Hourly Summary Files Format (US EPA Download Files AirData., 2022b) . 42

Table 4-1. Monitoring Data Stations Available for Modeling and Analysis 45

Table 4-2a. Descriptive Statistics for Numeric Parameters for 2017 training data set 47

Table 4-2b. Descriptive Statistics for Numeric Parameters for 2018 testing 47

Table 4-3. Feature Selection Optimization Techniques 55

Table 4-4. Feature Selection and Model Numbers 55

Table 4-5. Performance Metric 56

Table 4-6a Descriptive statistics of Richmond, VA 2018 62

Table 4-6b Descriptive statistics of Washington, D.C. 62

Table A-1. Annual Summary Files Format (US EPA Download Files AirData., 2022b) 84

Table A-2 Daily Summary Files Format (US EPA Download Files AirData, 2022) 91

Table B-1. Traffic Volume Stations used in 2017 and 2018 Location and Distance from Monitoring Stations 96

Table B-2. Traffic Volume Stations used in 2017 and 2018 Location and Distance from Monitoring Stations 97

List of Symbols

k	Number of folds in cross-validation
y	Target value
$\hat{y}$	the predicted value of y
$\bar{y}$	mean value of y
x	features or predictors
x_i	values of the x-variables
n	represents the number of observations
lat	Latitude
lon	Longitude
R	Correlation Coefficient
R^2	R-squared

List of Acronyms

AQM	Air Quality Monitoring
CAA	Clean Air Act
CBSA	Core-based statistical area
CO_2	Carbon Dioxide
COPD	Chronic obstructive pulmonary disease
D.C.	District of Columbia
DDOT	District Department of Transportation
DOEE	Department of Energy & Environment
DTs	Decision Trees
E.B.	East Bound
EPA	Environmental Protection Agency
FEMs	Federal Equivalent Methods
FRMs	Federal Reference Methods
GAs	Genetic Algorithms
GHG	greenhouse gas emissions
GPS	global positioning system
HUTRC	Howard University Transportation Research Center
LOS	Level of Service
mph	miles per hour
MWCOG	Metropolitan Washington Council of Governments
N	North

NAAQS	National Ambient Air Quality Standards
N.B.	North Bound
NE	North East
NO_2	Nitrogen Dioxide
NO_X	Nitrogen Oxides
NW	North West
PHF	peak-hour factor
PM	Particulate Matter
RCI	Relative Congestion Indices
S	South
S.B.	South Bound
SE	South East
SL	Supervised learning
SO_2	Sulfur Dioxide
SPI	Speed Performance Index
SW	South West
U.S.	United States
V/C	Volume-to-Capacity ratio
VOC	volatile organic compounds
VPD	Vehicles Per Day
W	West
W.B.	West Bound
WHO	World Health Organization

Chapter 1—Introduction

1.1 Background

In urban areas, increased vehicle traffic congestion has led to significant vehicle emissions. Factors such as increased motor vehicles on city roadways, outdated infrastructures, larger population densities, and rideshare and food delivery have played a critical role in pollutant levels (Reed, 2019). Transportation-related air pollution is a factor in smog and leads to poor air quality, negatively impacting U.S. citizens' wellbeing and health. Pollutants that come from vehicle emissions include particulate matter (PM), Nitric Oxide (NO), and Nitrogen Dioxide (NO_2), which is abbreviated as Nitric Oxides (NOx), Volatile organic compounds (VOCs), and Sulfur Dioxide (SO_2) (US EPA, 2022a). Over 55% of the total NOx emissions inventory in the U.S. are caused by transportation (U.S. EPA, 2022a). This study investigates the impacts of managing and forecasting emissions from traffic congestion to increase the ability for emissions level monitoring to upkeep these standards. In addition, this study hopes to bridge the gap by understanding how system-wide emissions forecasting can enhance air quality in other cities with similar urban populations. Finally, this research will provide a novel emissions forecasting model to monitor all areas where reduced levels are needed in Washington, D.C.

1.2 Research Motivation

National Ambient Air Quality Standards (NAAQS) are set by the Environmental Protection Agency (EPA) for six criteria pollutants and use a system of federally recognized analytical techniques to track compliance with those requirements known as

Federal Equivalent Methods or Federal Reference Methods (FEMs or FRMs). Regional spatial resolution FEM monitoring networks are efficient for assessing air pollution but may not be suitable for monitoring at the neighborhood level due to their high cost and technological complexity. While evidence suggests that pollution varies significantly at sub-regional and even neighborhood- or block-level spatial dimensions, FEM networks are most successful in monitoring pollution at the regional scale. As a result, even if a region may meet the NAAQS, the data gathered through the FEM network may not accurately reflect the air quality in particular areas. In reality, despite the District's compliance with federal air quality standards, public health trends, uneven tree canopy distribution, and socioeconomic statistics, some neighborhoods in Washington, D.C., may be at high risk for nonattainment (Saliby et al., 2020). There are many benefits when it comes to monitoring vehicle traffic congestion. One is being able to forecast emissions on roadways. There are only five District air monitoring network stations in Washington D.C., as seen in Chapter 3: **Table 3.1**. While there are only five stations, many traffic monitoring stations can be used instead of adding expensive air monitors to other locations where emissions monitors are not present in Washington, D.C. The creation of this model will contribute to air quality monitoring and policy change in the city of Washington, D.C. In addition, other cities with similar urban populations can apply to reporting air quality without spending budgets.

1.3 Problem Statement

Vehicle emissions due to traffic congestion have been shown to degrade air quality in urban areas (Hurwitz et al., 2012); subsequently, the need for more air quality

monitoring (AQM) is essential and involves robust monitors that are expensive (Saliby et al., 2020).

A network of air quality monitors is run by the Department of Energy and Environment (DOEE). The DOEE measures outdoor air pollutants and surface meteorological conditions critical in determining adherence to the air quality guidelines set out by the Environmental Protection Agency (EPA) for Washington, D.C. (DOEE Ambient Air Quality Monitoring, 2022). They run on continuous measurements that support reporting to the EPA and then to the Metropolitan Washington Council of Governments (MWCOG) to monitor long-term air quality and assess the success of control and reduction in emission mitigation strategies. Washington D.C Stations consist of five monitoring stations within an ambient air quality monitoring network (Specifications of their location information and what pollutant they measure are located in Chapter 4 Table 4-1.). These monitors are comprised of multiple instruments summarized in Chapter 3, Table 3-1.: Data Sources section.

1.4 Thesis Statement

A forecasting model built from historical data and using more readily available traffic monitoring station data is needed to avoid the purchase of expensive air quality measuring devices.

Constructing a forecasting model for one of the six criteria pollutants of National Ambient Air Quality Standards (NAAQS) will support the Environmental Protection Agency's (EPA) goal of improving neighborhood-level spatial scales for air quality monitoring (US EPA, 2022a). In addition, forecasting will help improve public health trends where reference monitors are not located (Patton et al., 2022).

1.5 Research Objectives

This praxis aims to create a model of vehicle traffic emissions levels to forecast using various predictors. This model can be utilized in traffic management and identifying which areas need policymakers' and transportation management's attention in terms of whether a permanent station is required or whether the traffic estimate for that area will be accurate enough to estimate emission levels. Moreover, this model will be created to assess attributes that can predict vehicle traffic emissions using historical meteorological data, vehicle traffic data, and other emissions. In addition, this study will determine which algorithm or statistical test can be used to develop a forecasting emissions model. Lastly, this study assesses the effectiveness of various models and algorithms used to create this forecasting model.

1.6 Research Questions and Hypotheses

Although this research focuses on modeling the amount of traffic congestion associated with certain emission levels, other factors impact their behavior in the modeled roadway. Understanding how these factors play a role in traffic system flow affects morbidity. The following research questions specifically drive this praxis research:

RQ1: What data predictors can be used to forecast NO_2 levels in Washington, D.C.?

RQ2: Which method provides the best performance to predict NO_2 Levels for different seasons in Washington, D.C.?

RQ3: Can this model be developed to forecast NO_2 levels in D.C. to be used in other cities to predict vehicle pollutants?

H1: Predictors such as Relative Humidity, Temperature, Traffic Volume, Wind Direction, Wind Speed, Month, Day, and Hour are significant in developing a NO_2 forecasting model.

H2: Machine learning algorithm Random Forest provides the best performance model in forecasting NO_2 levels in Washington, D.C.

H3: This model can predict NO_2 levels for different seasons in Richmond, Virginia.

1.7 Scope of Research

This research aims to determine at which levels of NO_2 by other predictors to utilize other sensors in the Washington D.C. areas to avoid the cost of adding more Emission monitors to the area. The data used in this research is on an hourly scale, making the model an appropriate solution for monitoring measurements as small as neighborhood-level emissions by utilizing the traffic volume sensors that are more easily accessible in Washington, D.C.

1.8 Research Limitations

The research is limited to the available data used in developing this praxis. The Department of Transportation Planning for the Metropolitan Washington Council of Governments (MWCOG) and the Department of Energy and Environment (DOEE) provided the data for the praxis. The traffic volume information is not taken at every station available in D.C. for the specified year of 2017 training data and tested by 2018 data. One emissions station was used as a proxy for this study as proof of concept of other stations in the D.C. area.

1.9 Organization of Praxis

The praxis is structured as follows: Chapter 2 describes current Emissions forecasting models using different predictors and the research gap of using available sensors in other cities and modeling for their performance. Chapter 3 explains the methodology by which the different predictors were selected based on feature selection, the various model's performance metrics, and its testing with other smart cities with volume count data. Chapter 4 describes the experimental results and findings. Lastly, conclusions, takeaways, contributions to the body of knowledge, and suggestions for potential future work are all included in Chapter 5.

Chapter 2—Literature Review

2.1 Introduction

This literature review will explore the impacts of managing traffic congestion to improve emission levels by emission forecasting. There continues to be a lack of understanding of how traffic system management affects air quality in urban areas. The knowledge gained from this literature review may assist researchers in continuously investigating emission levels related to vehicle congestion and other environmental factors.

The research has suggested that there are vital concerns surrounding the origin of traffic congestion in urban areas. The issues that continue to be investigated are as follows: the increased volume of traffic, the influence of poor weather conditions, random unplanned events on the roadway, work zones, variations in traffic flow, and bottlenecks (Nellore & Hancke, 2016). The questions that the literature review addresses are: What is traffic congestion? Why does traffic congestion occur? What are the different pollutants caused by vehicle traffic emissions? Machine Learning algorithms? How has ambient air quality been forecasted? What are the concentrations of harmful emission levels that come from traffic congestion? Lastly, the chapter will cover emissions monitoring management needs and requirements.

2.2 Traffic congestion

What is traffic congestion?

According to Downs (2004), traffic congestion occurs when vehicles go more slowly than a particular stretch of road can support. Some parameters that help engineers

measure the state of traffic congestion are time spent traveling, average speed, the average number of vehicles stopping, degree of overload, slower speed of flow, and total delay (Wang et al., 2018). Traffic congestion is related to an overabundance of vehicles on a roadway segment that may result in slower speeds (Barth & Boriboonsomsin, 2008). According to Heaslip et al. (2020), congestion could cause discontinuous traffic or stop the traffic flow completely. Traffic congestion is a significant issue in urban areas due to steadily increasing populations (Çolak et al., 2016). Traffic congestion is a phenomenon that occurs around the globe (Afrin & Yodo, 2020). Afrin & Yodo, 2020 also contended that population growth, increased vehicles on the road, lack of improved infrastructure, and more delivery services and ridesharing, discontinuous traffic or complete stop might lead to traffic congestion. Ridesharing affects traffic congestion by creating more traffic volume due to the increase in on-demand ridesharing because individuals are not taking public transportation (Li et al., 2022).

Why do we have traffic congestion?

Traffic congestion on roadways may result from one of the seven sources: insufficient infrastructure, bottlenecks, capacity, work zones, traffic incidents, variation in traffic flow, or weather (Vasudevan et al., 2004).

Insufficient infrastructure

Due to the higher population rates in urban areas like Washington, D.C., there is an increase in vehicles on the roadways (U.S. Department of Transportation Federal Highway Administration, 2022). As a result, existing infrastructures do not meet the demands of the rising number of cars on the roadways. An infrastructure that affects traffic congestion is the number of entrances and exits off and on highways (Wang et al.,

2018). Also, Li et al. (2022) assert that traffic lights that are poorly timed and slow-moving vehicles interrupt the traffic flow upstream.

Bottlenecks and Capacity

Bottlenecks occur on all types of roadways. However, highway bottlenecks are unlike traffic jams due to traffic control systems, traffic incidents, and work zones. Highway Bottlenecks occur when a road is overcapacity (Margiotta & Spiller, 2012). Yang et al. (2000) attested that the capacity of a transportation system indicates the maximum achievable throughput of that given system. These systems provide control and demand management for crucial adequate flow information. Capacity can be considered the highest amount of traffic capable of being managed by a roadway segment. Engineers of transportation managers have researched and focused on the physical capacity of the roads. Factors such as the shoulder and lane width, merge areas in interchanges, and roadway alignment (curves and grades) determined capacity (Taylor & Jeffrey Davis, 1999). Another case of bottlenecks is toll booths due to the physical restriction of traffic flow. Drivers on congested roadways may find themselves at closer driving distances than drivers driving on roads that are not as congested (U.S. Department of Transportation Federal Highway Administration, 2022).

Work Zones

Changes to the highway environment reduce flow due to roadway construction activities known as work zones (Hasan et al., 2019). The environmental changes caused by work zones may include reducing the number of narrow travel lanes, lane "shifts," and lane detours, as well as temporarily removing shoulders and short-term road closures (Wilson & Cales, 2008).

Traffic Incidents

Traffic incidents may disturb the normal capacity flow through physical impedances in lanes traveled (Azeem et al., 2019). For example, activities such as vehicular breakdowns beside the roadway, vehicle crashes, and fragments impeding travel lanes; are common incidents (Dowling et al., 2004). In addition, incidents on the roadway's shoulders distract drivers and cause roadway flow conditions to change (Mensah et al., 2014).

Variation in traffic flow

Rush hour can cause a high influx of cars on highways due to motorists traveling simultaneously to and from work (Ben-Elia & Ettema, 2011). In addition, regular traffic fluctuations and day-to-day demand variability lead to days with more significant traffic volumes (Zhou & Mahmassani, 2007). Even without any events, fluctuations in demand volumes overlaid on transportation systems with fixed capacity also result in unpredictable travel time (Ogunrinde & Lebile, 2015).

Weather

Changes in drivers' behavior due to Environmental and weather conditions can affect traffic flow (Mahmassani et al., 2009). When precipitation occurs, drivers generally lower their speeds and increase their headlights due to reduced visibility during bad weather (Hammad et al., 2019). When fog, smoke, and bright sunlight on the horizon are present, they can reduce travel speeds (Ahmed & Ghasemzadeh, 2018). Snowy, icy, and wet roadway surface conditions will have effects several hours after precipitation has ended (U.S. Department of Transportation Federal Highway Administration, 2022).

Understanding all of the above root causes of traffic congestion is why emissions monitoring is essential due to vehicle traffic congestion.

Traffic congestion ranking for Washington D.C.

According to the TomTom Traffic index 2021, which covers 404 cities across 58 countries on six continents, Washington, D.C. is 15th among 80 cities in the USA and 212th globally out of 404 other cities, with 48 hours per year spent in traffic, and has had a 21% congestion level in 2021, 8% less than 2019, and 4% increase from 2020 (Washington traffic report, 2022).

2.3 Machine Learning

Machine learning, a subfield of artificial intelligence, is used to simulate human decision-making using models and algorithms that learn based on experience from data (Aggarwal et al., 2022). Different Machine Techniques can be categorized into three different categories unsupervised, semi-supervised, and supervised learning, as seen in

(Van Engelen & Hoos, 2020). The different categories of machine learning techniques, as examples of their applications, are seen in **Figure 2-1**.

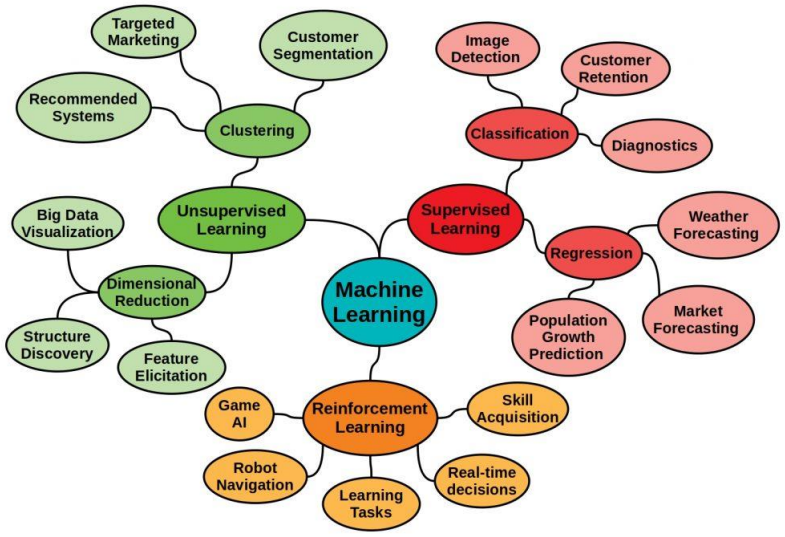

Figure 2-1. Machine Learning paradigms and algorithms, and applications (Miraftabzadeh et al.)

Unsupervised Learning

Pattern recognition with unsupervised data learning is done without using a target characteristic (Alloghani et al., 2019). Unsupervised learning is used in clustering, dimension reduction, probability distribution estimation, and finding feature associations. **Figure 2-2** below shows that the input data is unlabeled and clustered into groups based

on the algorithm used for unsupervised learning.

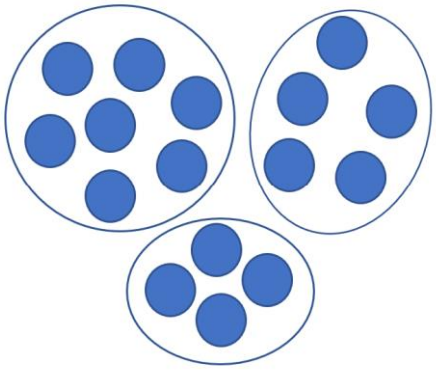

Figure 2-2. Unsupervised Learning Descriptive Image

Semi-Supervised Learning

Learning from a set of cases with few labels and plenty of unlabeled examples is known as semi-supervised learning (Pise & Kulkarni, 2008). Semi-supervised learning is best when there is a mixture of unlabeled data for the training of that model, typically with a more considerable amount of unlabeled data than data that is available that is labeled.

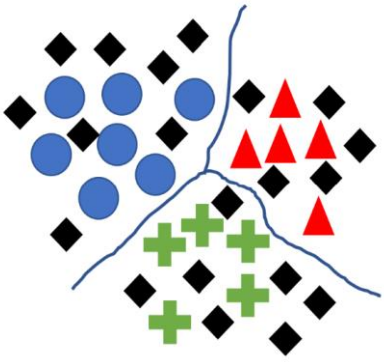

Figure 2-3. Semi-supervised Learning Descriptive Image

Supervised Learning

In the machine learning approach known as supervised learning (SL), all the data that make up the model's training set have been labeled. Supervised learning is used in prediction, classification problems with discrete labels, and Regression problems with numeric values (Zhou et al., 2015). For example, for **Figure 2-4,** the classification problem has discrete labels of shape being circles (blue), triangles (red), and plus signs (green).

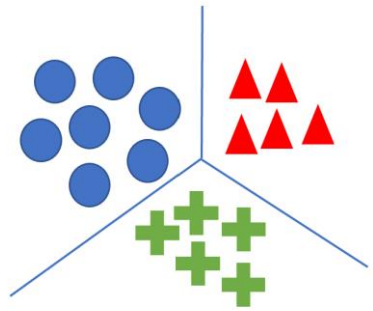

Figure 2-4. Supervised Learning of Descriptive Image

Machine Learning models include tree-based models (Sect. 2.3.1), linear and partial least squares regression (Sect. 2.3.2), support vector machines (Sect. 2.3.3), artificial neural networks (Sect 2.3.4), and k-nearest neighbors (Zhou et al., 2015).

2.3.1 Tree-based models

The fundamental building block of tree-based regression is decision trees (DTs). A root node and leave nodes connected by branches make up DTs (Twala, 2009). The fundamental concept is to divide the training dataset into subgroups at each branch based on the input features. Types of decision trees include continuous variable decision trees

and categorical variable decision trees (Twala, 2009). Categorical trees are decision trees with categorical target variables (Kim & Hong, 2017).

While Continuous Variable Decision Trees have continuous target variables (Kim & Hong, 2017); for example, if the amount of NO_2 levels are known for some previous years but want to determine missing years, there is not enough information to predict. Since the essential variables are known, decision trees can be constructed to forecast NO_2 levels based on meteorological variables such as Temperature, Wind Speed, and several other characteristics. For NO_2 forecasting studies, the model will be continuous, not categorical.

2.3.2 Random Forest

Another versatile supervisor machine learning technique, random forest, is utilized for classification and regression (Breiman, 2001). The "forest" refers to a set of independent decision trees combined to lower variation and produce more precise data predictions (Cutler et al., 2012). A Random Forest classifier uses many decision trees on different subsets of the input data to increase the dataset's prediction accuracy. It is based on ensemble learning, which integrates many classifiers to solve a challenging issue and enhance the model's performance (Cutler et al., 2012).

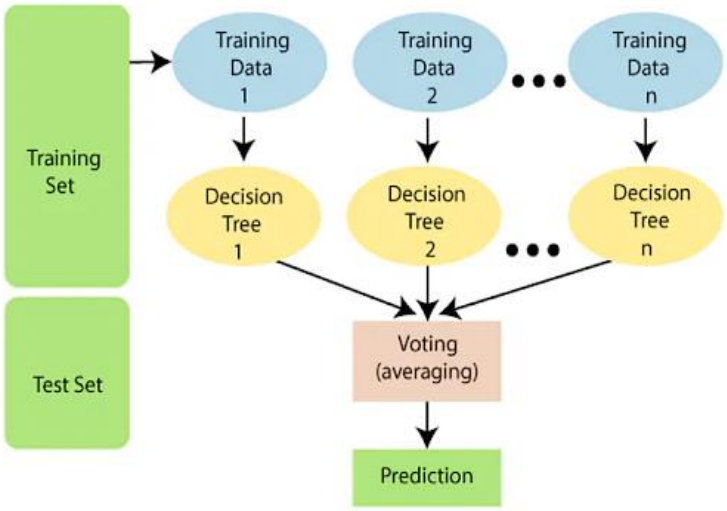

Figure 2-5. Random Forest Algorithm Diagram

(JavaTpoint Machine Learning Random Forest Algorithm, 2022)

The first step in the random forest algorithm is to choose random samples from a specified data collection or training set. Then the algorithm will build a decision tree for each training set of data. Next, voting will be conducted using an average of the decision tree. Lastly, select the prediction outcome that garnered the most significant support as the final prediction outcome (Breiman, 2001).

The ensemble is the name given to this merging of many models. Ensemble employs two techniques: Bagging and Boosting, as seen below

in Figure 2-6. Bagging generates a distinct training subset via replacement from a sample training dataset (Sagi & Rokach, 2018). A majority vote decides the outcome.

While Boosting is turning weak learners into strong ones by building successive models to achieve the maximum accuracy possible (Natekin & Knoll, 2013).

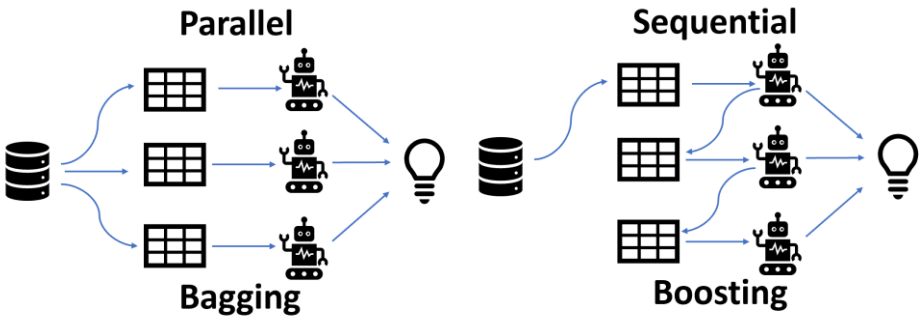

Figure 2-6. Bagging and Boosting

Random forest uses the Bagging method. With any initial random data, the procedure starts. It is then arranged into samples called Bootstrap Sample. Bootstrapping is the term for this procedure which is also known as Bootstrap Aggregation used by random forest (Gislason et al., 2006). In addition, since each model is trained separately, aggregated results are produced. The last stage combines all the findings, and the

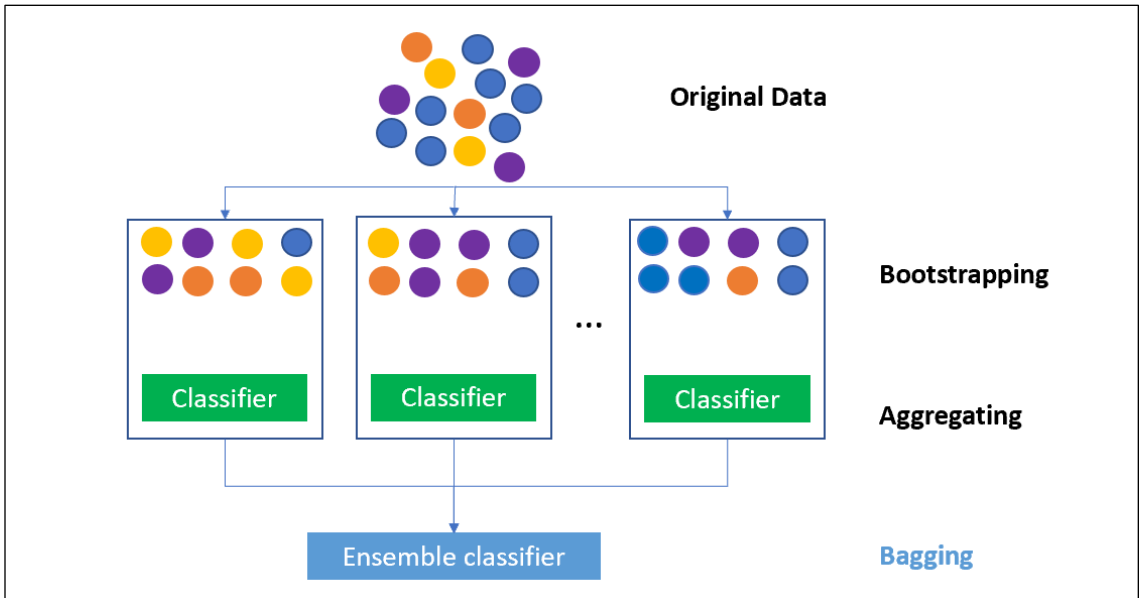

Figure 2-7. Bagging Ensemble Classifier Methodology

resultant outcome is based on a majority vote. Using an Ensemble Classifier, this procedure, known as bagging, is carried out as seen in Figure 2-7 above.

2.3.2 Linear and partial least squares regression

The relationship between one or more independent variables and a dependent variable is known as Multiple Linear regression (Mac Nally, 2000). Simple linear regression is used when there is only one independent variable and one dependent variable.

2.3.3 Support Vector Machines (SVM)

Vladimir Vapnik created the well-known supervised learning model known as the support vector machine (SVM), which is used for data classification and regression (Khorshid et al., 2015). SVM is typically leveraged for classification problems that use a generated hyperplane where the distance between two classes of data points is at its highest. It is often used to solve classification difficulties. The hyperplane is used as the separating decision boundary to separate classes of data points on either side.

2.4 Feature selection methods for Pollution

When creating a predictive model, the feature selection process involves lowering the number of input variables to maximize the model's performance (Lin et al., 2015). In some circumstances, reducing the number of input variables might enhance the model's efficiency while lowering the computing cost of modeling. When employing statistical feature selection techniques, each input variable's relationship to the target variable is assessed, and the input variables with the most significant relationships are chosen. Although the kind of data for both the input and output variables affects the statistical

measures chosen, these techniques can be quick and efficient. Different feature selection types include Shapley Additive Explanations (SHAP) importance, genetic algorithm, forward selection, backward elimination, and multi-objective optimization.

SHAP feature selection

Simple logic underlies the significance of SHAP features: It is crucial to have features with high absolute Shapley values. The average marginal contribution of a particular instance of a characteristic over all potential coalitions is known as the Shapley value (Michalak et al., 2013). The features are then sorted and plotted in decreasing order of relevance.

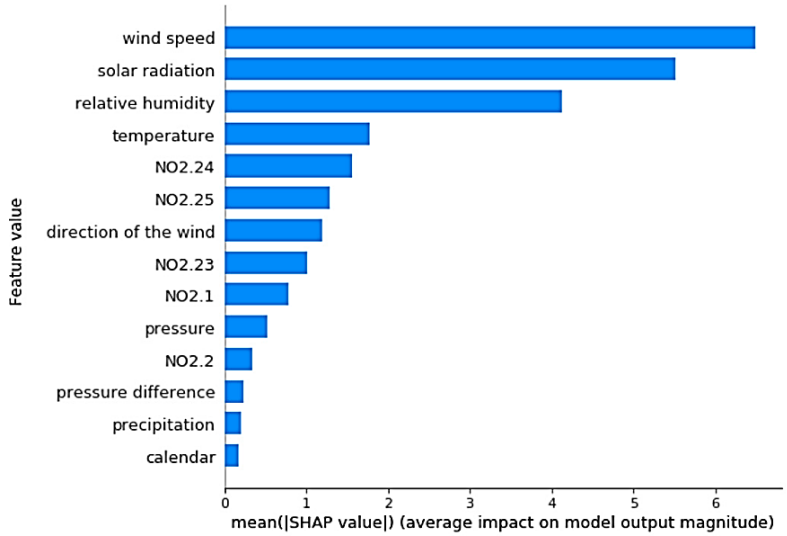

Figure 2-8. SHAP Value from a NO_2 forecasting model

(García & Aznarte, 2020)

The mean absolute Shapley values are used to measure the relevance of a SHAP feature from the LSTM model (García & Aznarte, 2020). The most significant factor, which altered the projected absolute NO_2 chance by an average of 6, was the number of Wind speeds used.

Genetic algorithm

When used for approximation issues, genetic algorithms (GAs) minimize the number of predictors since they are usually employed to choose input variables in these situations (Abdella & Marwala, 2005). The GA is a machine learning model whose behavior is derived from a simulation of natural evolution. Other sources have information on the precise structure of a typical genetic algorithm (Kalogirou, 2009).

Forward selection

Forward selection adds features one by one until the model's performance peaks (Koprinska et al., 2015). With no selected features and adding one at a time, forward selection selects the features that lead to the model’s best performance. Although this approach is straightforward, it can be computationally expensive and may not identify the best combination of attributes.

Backward Elimination

A more methodical method called backward elimination starts with a complete set of features and eliminates each feature one at a time until the model's performance achieves its peak (AlNuaimi et al., 2020). Although this method is more computationally efficient, it might not also locate the best combination of attributes.

Multi-Objective Optimization

Multi-objective optimization is a branch of multiple-criteria decision-making that deals with mathematical optimization problems requiring simultaneous optimization of multiple objective functions (Kannan et al., 2013). Objectives that may conflict have

been used to make decisions in many scientific fields, including engineering (Dhaubanjar et al., 2017). Many real-world engineering applications require designers to decide between conflicting objectives (Khalilpourazari & Arshadi Khamseh, 2019). For instance, it can be necessary to enhance the prediction of NO_2 by optimizing multiple features (Zhai & Chen, 2018). Therefore, multi-objective optimization studies are advised when providing various solutions highlighting the trade-offs between the different objective functions is needed (Vallerio et al., 2015).

Machine Learning Summary and tools

The algorithms and statistical methods described above can be applied using WEKA. WEKA uses a set of machine learning algorithms and statistical techniques that have been used for mining association rules, grouping, regression, classification, and data visualization (Frank et al., 2016).

2.5 Forecasting Vehicle Traffic Emissions

While there have been many studies that forecasted NO_2 emissions, the selection of features for different cities there have not ways that NO_2 has been forecasted using volume count stations in the thousands for the use to monitor in Washington D.C. has yet to be applied (Mishra & Goyal, 2016). Elangasinghe et al., 2014 attested that using the Artificial Neural Networks (ANN) modeling technique makes it possible to pinpoint the critical meteorological factors needed to accurately depict the temporal variability in air pollution concentrations for a given scenario. Using features selection to select down to features Month (M), Hour (H), Day (D), Wind Speed (WS), and Wind Direction (WD), the model was able to achieve an R=0.80, these five parameters can explain 80% of the variability in NO_2. When the model used the optimized feature selection through Wind

Speed (WS) and Wind Direction (WD), 77% of the variability could be explained. However, there was no use of vehicle traffic count in developing this model, and the research was conducted with meteorological and NO_2 monitors at the same location. Iskandaryan et al., 2022 were able to use different meteorological emissions data to forecast NO_2 emissions levels. A Bidirectionational Convolutional Long Short-Term Memory (LSTM) architecture was used to explore features associated with NO_2 forecasting. With the selection of 7 out of all the features by using maximum relevance-minimum redundancy (mRMR) seen in Table 2.1, the model achieved an RMSE=3.44 $\mu g/m^3$ and a MAE=2.87 $\mu g/m^3$. LSTM was also used by García & Aznarte, 2020, when nitrogen dioxide levels were forecasted for Madrid and features selection using Shapley Additive Explanation (SHAP). SHAP is a method based on game theory to find optimal Shapley values. The model showed good RMSE values and had no R^2 presented. The model used features of Solar Radiation (SR), Wind Speed (WS), Relative Humidity (RH), Temperature (TEMP), Wind Direction (WD), Pressure (P), Month (M), Day (Day), Year (Y) to forecast NO_2. Then, model features were down-selected to SR, RH, and TEMP by the highest SHAP values. There were able to forecast NO_2 using only meteorological data (García & Aznarte, 2020). Lucena-Sanchez et al., 2020 used 2015-2017 data to develop NO_2 forecasting using Multiple Linear Regression and Random Forest algorithms using features such as meteorological parameters; however, there is only one reference monitor available for the biases for prediction.

Table 2-1. Emissions Modeling Comparison of performance NO_2

Model Method	Pollutant	Predictors	Feature Selection Method	R^2	Training/Testing data Method	Source
ANN	NO_2	WS, WD, RH, TEMP, SR, H, D, M	genetic algorithm, forward selection, and backward elimination	0.77	2010 data/2011 data	(Elangasinghe et al., 2014)
LSTM	NO_2	WS, SR, RH, TEMP, WD, BP, P, M, D, Y	SHAP	N/A	Both 2017-2018 data	(García & Aznarte, 2020)
RF and LR	NO_2	AP, RH, solar duration, TEMP, WS	multi-objective optimization	RF 0.62 LR 0.51	2015-2017 data/10-fold 2015-2017	(Lucena-Sánchez et al., 2020)
BiConvLSTM	NO_2	UVR, WS, WD, TEMP, RH, BP, SI, P, I, OT, L, Avg. Speed	MI, mRMR	N/A	2019 data/10-fold and 2020 data	(Iskandaryan et al., 2022)

2.6 Impacts of Traffic Congestion on Emissions and Health

The categories for vehicle traffic congestion emissions are on-road sources and non-road sources (Kheirbek et al., 2016). Heavy machinery, airplanes, ships, trains, all-terrain vehicles, snowmobiles, and small engines like lawnmowers are examples of non-road sources (Dallmann & Menon, 2016). On-road sources include motorcycles, passenger vehicles, trucks, and commercial trucks and buses (Wang et al., 2008). Air pollution comes from on-road sources such as buses, trucks, and cars (Huang et al., 2020).

Related Health Effects from on-road pollution

Morbidity is being ill or symptomatic by a disease or illness (Hernandez & Kim, 2021). In contrast, the number of deaths caused by the health event under investigation is known as Mortality (Hernandez & Kim, 2021). Oxides of nitrogen (NO_X) combined with

Volatile organic compounds (VOC) in the presence of sunlight can create a chemical reaction that is known as "bad" ozone (Evuti, 2013). Health problems can be triggered, such as lung disease and asthma (US EPA, 2022a).

What is Particulate Matter?

Particulate Matter (PM) is a mixture of a particle in the form of airborne solids and liquids (Valavanidis et al., 2008). There are two types of PM which are PM_{10} and $PM_{2.5}$. PM_{10} particles that are inhalable and have a diameter of 10 micrometers or less. $PM_{2.5}$ are finer sized in diameter, less than 2.5 micrometers. PM comprises hundreds of different chemicals (Alves et al., 2018). Smog is an event defined as a time when the $PM_{2.5}$ concentration was >65 $\mu g/m\text{^}3$ (Kim et al., 2012). One primary source of PM is from automobiles which results from complex chemical reactions such as sulfur dioxide (SO_2) and nitrogen oxides (NO_X) (Swift, 2000).

Nitrogen Dioxide NO_2

The primary sources of Nitrogen oxide emissions into the environment are nitric oxide and nitrogen dioxide from motor vehicle exhaust. Nitrogen dioxide NO_2 is created when nitric oxide NO and Ozone (O_3) combine. Vehicle emissions are the primary source of nitrogen oxide in Washington, D.C., as they are in most large cities.

The Harmful Effects of Smog

There are two primary components of smog: PM and ground-level ozone (O_3). Breathing in smog can affect human health in the following ways: it makes it harder to take deep breaths, which causes upper respiratory pain causing shortness of breath; causing coughing and sore or sore throat; swollen and injured airways; aggravating lung

function, especially for individuals with chronic bronchitis, increase the frequency, asthma, emphysema (U.S. EPA, 2022a). As a result, those with these conditions are more susceptible to asthma attacks, chronic obstructive pulmonary disease (COPD), lung damage, and lung infections that persist even after the symptoms have subsided (Dharmage et al., 2019). The average passenger vehicle emits roughly 4.6 tons of carbon dioxide annually (U.S. EPA, 2022a). NO_2 is a form of smog because it indirectly impacts the visibility of air pollution. Below is a table of associated concentrations of NO_2 and their related health effects.

Table 2-2. Effects of Nitrogen Dioxide on Human Health (Hussain, 2007)

Conc. of NO_2 (ppm)	Time of Exposure	Effects
0.06 – 0.10	2 – 3 years	Increase in acute respiratory disease.
Upto 0.1	6 months	Increase in acute bronchitis in School Children.
0 – 12	<24 hrs.	Human olfactory threshold.
5	10 minutes	Increase airway resistant
90	30 minutes	Pulmonary edema

2.7 The Need and Requirements for Traffic Emission Monitoring

The EPA created the National Ambient Air Quality Standard (NAAQS) for contaminants that are often detected outside (US EPA, 2022a; Turner et al., 2020). It focuses on the levels of these pollutants that may damage the environment and public health. Primary and secondary standards comprise the national ambient air quality standards (Jin et al., 2016). Primary standards are NAAQS standards designed to safeguard the public's health while leaving enough room for error (Keehan, 2018). Secondary standards are NAAQS regulations created to safeguard the public's welfare from harmful impacts, such as those relating to damage to property, weather, visibility,

climate, soil, water, crops, vegetation, artificial (anthropogenic) materials, animals, wildlife, and transportation (Wang et al., 2020).

2.8 Summary and Conclusion

Traffic congestion will continuously grow due to population growth. There needs to be traffic management put in place to reduce traffic in urban areas. The literature describes measuring vehicle traffic emission modeling. There have also been many models that have utilized these main components to help successfully measure NO_2 levels. However, none have utilized this methodology to produce a no-cost system for emissions monitoring via Traffic Volume count stations. Once traffic congestion has improved, it is likely to reduce pollution caused by transportation (Beaudoin et al., 2015). While there have been models developed, a limited amount of them has addressed that there is a NAAQS need for more neighborhood-level modeling, as seen in the previous models summarized in **Table 2-1: Emissions Modeling Comparison of performance NO_2**.

Chapter 3—Methodology

3.1 Introduction

Since traffic conditions in the Washington, D.C. area have continued to be a concern (Nagy & Simon, 2018), it was determined that using Washington, D.C., may act as a stand-in to demonstrate the methods created inside this praxis. The methodology can be applied to other highly congested cities with similar mean congestion emissions and population sizes. The general methodology design uses a data-centric approach, where emissions from road traffic are determined by the number of vehicles occupying a segment (volume) by determining the concentrations of NO_2 to be estimated from traffic volume data at 1,316 volume count stations and five emissions data monitoring stations in Washington D.C. To improve the ability to estimate emissions where monitoring stations are not present. The model will assess and predict the amount of NO_2 for Washington, D.C., at a station based on traffic congestion and other meteorological data. This data-centric approach is visualized by the methodology below. The prediction models are prepossessed, developed, verified, and evaluated using historical data, as shown in the high-level diagram below in **Figure 3-1**.

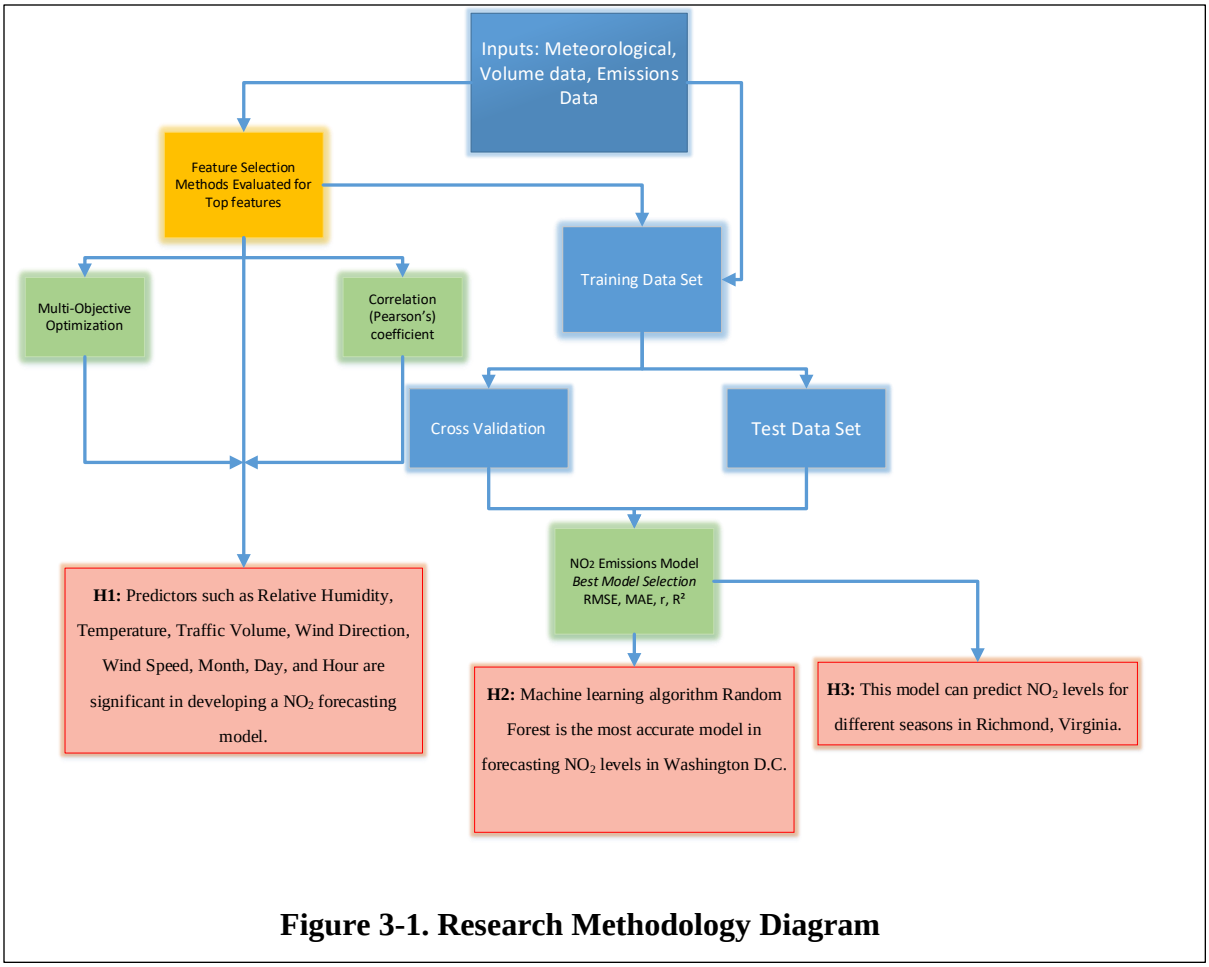

Figure 3-1. Research Methodology Diagram

Therefore, the methodology presented in this chapter will layout the procedures are taken to evaluate the three hypotheses stated in Chapter 1. Section **3.2** is led by Hypothesis 1 and focuses on the feature engineering and data preparation methods chosen to discover the key characteristics required to construct the NO_2 emissions models. The NO_2 Emissions models are then tested and trained using various ML algorithms and statistical techniques considered in this study in Section **3.3**, which is driven by Hypothesis 2. The 2018 test data was adjusted and assessed by looking at their performance indicators. This section seeks to establish whether a prediction model for NO_2 can be created using the methods described in the section. Lastly, section **3.4** will

describe the methodology utilizing the model developed as a basis to predict NO_2 for other cities like Washington D.C. through using the highest performing model in section **3.2**. The primary tool used to evaluate this methodology is WEKA.

3.2 Data Source Description and Use

Two primary data sources used in this research are the emissions and meteorological data DOEE in support of the EPA and the MWCOG for traffic volume data. The dataset includes hourly volume count data and meteorological data from stations within Washington, D.C. The data was collected from two years, 2017 for building the model and 2018 for testing the model. These are complete datasets before the COVID-19 pandemic due to the unpredictable nature of traffic patterns during the pandemic (Mousazadeh et al., 2021). For this study, stations were evaluated, as seen in **Figure 3-1**. The data source for the emissions data is the Department of Energy & Environment (DOEE). DOEE collects, maintains, and conducts an analysis of emissions data that is needed by the D.C. government and the Environmental Protection Agency (EPA). The DOEE emissions and meteorological data consist of hourly data, as shown in **Figure 3-2**. There is a five-station network in the district Run by DOEE at McMillian Reservoir, River Terrace Elementary School, Takoma School, Verizon Center, and Haines Point. The five stations that measure the pollutants are made of advanced sensors listed with the measurements they have taken. In comparison, there are 1316 volume count stations in Washington, D.C. A visual representation of those stations is presented in **Figure 3-3**.

The District's air monitoring network contains monitoring stations located at the McMillan Reservoir, River Terrace Elementary School, Takoma School, Verizon Center, and Haines Point.

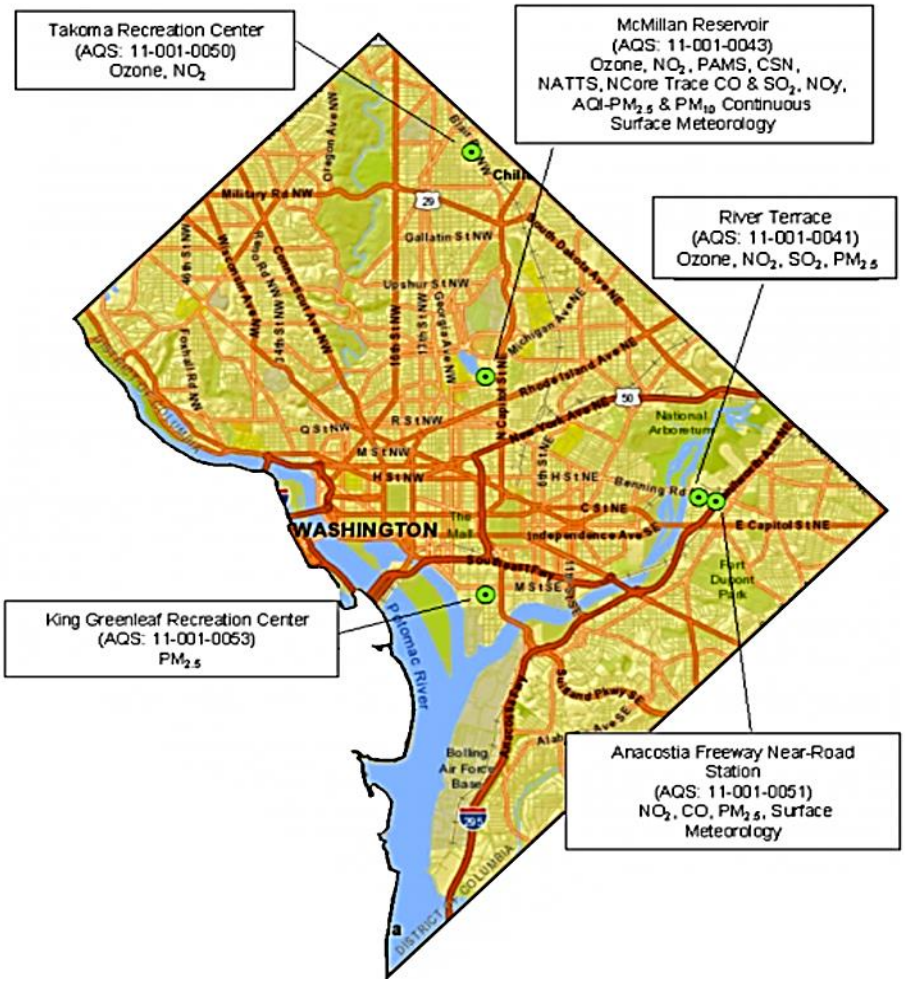

Figure 3-2. Air Quality Monitoring Stations in Washington, D.C.

(DOEE Ambient Air Quality Monitoring, 2022)

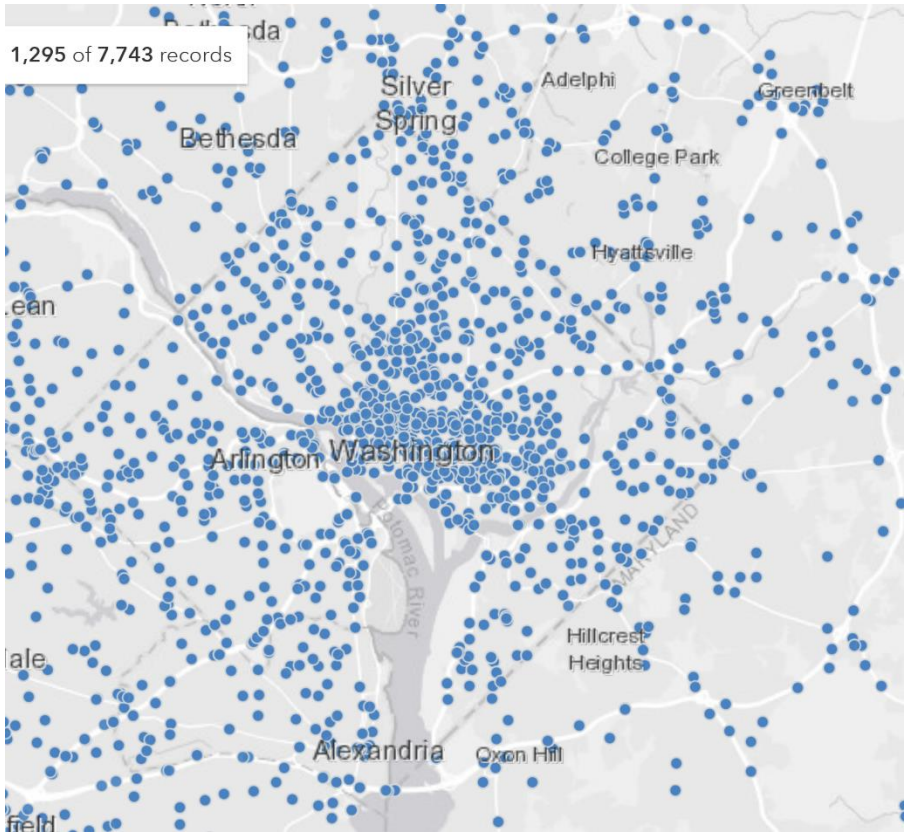

Figure 3-3. Traffic Monitoring Stations in Washington, D.C.

3.2.1 Data Preparation and Visualization

Before conducting feature selection, the Raw data from both Vehicle volume count, Meteorological data, and Emissions data had to be sorted by the following criteria using Python due to the size of each data set being more than 1 million points.

1. First, all the raw data is loaded into python through CSV files.
2. Then, emissions and meteorological data sets are sorted based on just the 'District Of Columbia' data, separating them from all the other reported stations in the U.S.

3. Next, the emissions data, meteorological and volume data sets, and the local date and local time column are joined into a single column called “DateTime.”
4. Now the station is selected for volume count (i.e., 6002_E reference Appendix Table B-1)
5. Lastly, all data sets are merged into one CSV file by merging “inner” on “Latitude, Longitude, and DateTime,” meaning only the columns with those common values are kept.
6. Next, the CSV files are opened, and the Method Detection Limit (MDL) for each parameter is used to replace all values lower than that limit.
7. The data is then visualized to view the statistical properties of the 2017 and 2018 data variables that will be read into the Models in Table 4-1.
8. There are eight columns in each data file for both 2017 and 2018, as seen in Appendix **Figure A-1**.
9. Next, these data files are loaded into the WEKA explorer for preprocessing.
    a. First, the 2017 data file is opened and filtered by normalizing all features except for the target data set of NO_2.
    b. Then the file was resaved as another file.
    c. The exact process was repeated for the test dataset of 2018.

These steps were taken to ensure that the month, day, and hour variables for all datasets’ were aligned due to random hourly data being missing from different datasets due to sensor malfunctions. Data from 2017 will be used to construct the model, and 2018 will serve as the test dataset.

3.2.2 Feature Selection

The data preprocessing in the Section 3.2.2 is used in the feature selection. The next stage in developing a model is choosing the optimal subset of inputs. The main goal of this work is to create a model that is easy to use for better-generalized performance when producing forecasts for NO_2.

3.2.2.1 Multi-objective Optimization Methodology

This section describes the multi-objective optimization techniques used in this research: forward selection, backward selection, and genetic algorithm. These feature selection methods were implemented using Weka Explorer under the "Select attributes" Tab; the following steps were taken to perform feature selection:

1. The normalized data file for 2017 was loaded into the WEKA explorer window again.
2. Then the WEKA “Select attributes” table was opened to feature selection.
3. Following CfsSubsetEval, feature selection was conducted.
    a. Description of CfSebsetEval: By considering each feature's unique predictive power and the degree of interdependence among them, determine the value of a subset of qualities. Preferable subsets of characteristics have little intercorrelation and are substantially associated with the class (Frank et al., 2016).
4. A Greedy Stepwise Forward Selection was used to perform a search within the model.
    a. Greedy Stepwise: Performs a greedy by forward or backward search selection. It can begin at any point in space with no or all qualities. It stops when the evaluation decreases due to adding or removing any remaining

characteristics. Moving around the area from one side to the other and noting the order in which the traits are chosen may also create a ranked list of attributes (Frank et al., 2016).

5. The search was conducted using the entire training set.
6. Then steps 1-5 were repeated, except using search methods of backward propagation and the Genetic Algorithm search.

Tables 4-3., describe the outcomes of the three common input optimization approaches used in this case study forward, backward, and genetic algorithm. The primary tool used to evaluate the methodology was WEKA.

3.2.2.2 Correlation Coefficient

1. The normalized data file for 2017 was loaded into the WEKA explorer window again.
2. Then the WEKA "Select attributes" table was opened to feature selection.
3. Following CorrelationAttributeEval, feature selection was conducted.
    a. Description of CorrelationAttributeEval: Determines the value of a characteristic by calculating its Pearson correlation with the class. Each value of a nominal property is treated as an indication, and each value is examined separately. A weighted average is used to determine a nominal attribute's overall correlation.
    b. A Ranker was used to perform a search in the model.
        i. Ranker: Sorts qualities according to user ratings. Together with attribute evaluators of GainRatio, ReliefF, or Entropy (Frank et al., 2016).
4. The search finished and was done using all of the training data.

5. **Table 4-3** describes the outcomes of the Correlation Coefficient selection attributes.

3.3 Training, Testing, and Cross-Validation

Training and Cross-Validation

This section is dedicated to the development of the model and how it was tested and validated. The following steps were used to develop the model and test it using the feature-selected and original data for comparison purposes, as shown in Table 4-3.

1. The normalized data file for 2017 was loaded into the WEKA explorer window again.
2. Based on **Table 4-3**., the remove features were removed one by one.
3. Then multiple linear regression was performed on the data set.
    a. Linear Regression: Using linear regression for prediction is referred to as this class. For model selection, it employs the Akaike criteria and supports weighted instances. Based on a single target variable and several predictors.
4. Training data was Tested against Training data, Test data, and Cross-Validated and Evaluated based on the performance metrics presented in **Table 3-1**. The classifier output is presented with the MLR model and model performance metrics such as RMSE, R, R^2, MAE, and RRSE, as seen in **Table 3-1**. The results of these metrics are presented in **Table 4-5** model results and again in Appendix B **Table B-2** improved model results.

Table 3-1. Performance Metric Evaluation Measures

Metric	Equation	Description
RMSE	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \lvert y_i - \hat{y} \rvert}$	The square root of the MSE.
r	$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \, (y_i - \bar{y})^2}}$	The measure of how strong a relationship is between two variables.
R^2	$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$	Represents the correlation between the values and the starting values, or the coefficient of fit. The range 0 to 1 is translated into percentages. The model is better the higher the value.
MAE	$MAE = \frac{1}{N}\sum_{i=1}^{N} \lvert y_i - \hat{y} \rvert$	Represents the difference between the extracted absolute difference over the whole data set and the original and forecasted values.
RRSE	$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y})^2}}$	The root relative squared error indicates how well a model performs in relation to the mean of the real values. Better if it is lower.

Note: $\hat{y} = predicted\ value\ of\ y$, $\bar{y} = mean\ value\ of\ y, x_i = values\ of\ the\ x - variables$, *n: represents the number of observations*

5. The settings of MLR were toggled to improve the results final configuration Model Configurations.
6. Steps 1-2 were repeated with Multilayer Perceptron, also known as Artificial Neural Network (ANN) and Random Forest (RF) Algorithms.
    a. Multilayer Perceptron (ANN): A classifier that categorizes examples by learning a multi-layer perceptron using backpropagation. The network may be set up manually or with a straightforward heuristic. During training, the network settings may also be checked and adjusted. This network's nodes are all sigmoid (except for when the class is numeric, in which case the output nodes become unresolved linear units).
7. The Settings of ANN were toggled to improve the results' final configuration.
8. Then steps 1-2 were repeated, and the RF Algorithm was run.
    a. The class for constructing a forest of random trees operates in four steps in WEKA:
        i. In bootstrap samples, the first instances from X are randomly chosen with replacement, and the process is repeated until the sizes of B_i and X are equal.
        ii. Next, to train tree t_i in the forest, a random selection of features is chosen for each B_i.
        iii. Unpruned decision trees are cultivated using an information gain measure.
        iv. The most typical individual tree forecast is the final regression result.

Training and Testing

Testing with 2018 data for prediction, as seen in Table 4-2, the descriptive statistics for 2017 to see if the concentration and meteorological parameters show similar Distributions throughout the years. Testing steps are as follows:

1. Steps 1-3 were repeated in the Training and Cross-Validation section.
2. Then the test option to supply a test set was selected, and the normalized 2018 data was run.
3. Then steps 5-7 in the Training and Cross-Validation section are repeated for each algorithm and statistical method.
4. The setting was toggled to improve performance.
5. The results from testing with the 2018 test set in Chapter 4, **Table 4-5**, and the plot of predicted 2018 vs. actual 2018 values are shown in **Figure 4-9.**

Model Inputs

The model inputs include Month, Day, SO_2, Wind Direction, Relative Humidity, Temperature, Traffic Volume Count, and NO_2 for 2017 as the training data. As for the test data, the same variables for 2018 were used as the 2018 testing data.

Volume Data

MWCOG provided VA, MD, and DC hourly traffic volume records for 2016, 2017, and 2018. The Hourly Traffic count is a tubular data set containing detailed hourly traffic counts from continuous and program count stations in the District, Maryland, and Virginia provided to Transportation Planning Board by the respective state DOTs. The hourly counts are processed, listed by direction, and matched to the appropriate

directional network link. Separate detailed volume and classification traffic tables are provided for 2016, 2017, and 2018. The tables are in CSV format. **Table 3-2** lists the file format for the volume tables in separate CSV files by state and year (MWCOG).

Table 3-2. File Format: Hourly Traffic by State by Year CSV Files

Item	Format	Description
station	Quoted character string	Original count station identifier from the states. This is a quotes string because some values are numeric-only while others are alphanumeric
Direction	Text - N, S, E, W	The direction was converted from codes to letters N - North, S - South, E - East, W - West
Volume	Long Integer	Hourly directional volume at the station
DateTime	MM/DD/YYYY mm:hh:ss Zero-filled date string	This field is zero-filled and contains both the date and the time that the count began
AB Pair	Quoted character string AAAAABBBBB	The Anode-Bnode pair of the 2018 network link is associated with the station. This is used to join with the DualTraffic_10to18 shapefile.
Date	MM/DD/YYYY Zero-filled date string	This field contains only the count date for ease of import into various formats
Hour	Short integer	This field contains only the count hour for ease of import into various formats
Sta_DIR	Text SSSSSSS_D	This is a concatenation of the Station ID

More about Emissions and Meteorological Data

The Emission and Meteorological data source is from DOEE data delivered to the EPA for 2017 and 2018. This hourly data is pre-generated data files available for download from the files updated twice a year in June and December from reference monitors around the country. For this study on files generated from reference monitors controlled by the DOEE in Washington, D.C. The data are large data files of criteria gases, particulates, meteorological, toxics, ozone precursors, lead, and Blank, in which empty canisters are measured for quality assurance. These data sets include the following temporal scales annual, daily, and hourly.

Annual Summary Files

Every monitor in the EPA database for a given year is represented in a one-yearly summary file. For the contaminant, sample durations have been estimated. For instance, the EPA computes 24-hour averages for $PM_{2.5}$, occasionally given as 1-hour samples. There are several pollution standards that call for several sampling times. Some measurements are affected by extraordinary occurrences, which the EPA agency has omitted from the annual database. The annual summary CSV file format is presented in Appendix **Table A-1**.

Daily Summary Files

Data for all monitors in the EPA database is included in each daily summary file. These files are divided per parameter (or parameter group). There are daily summary records in this file that is:

1. The total of all measurements obtained at the monitor on a sub-daily basis.

2. The single sample value if the monitor only collects one sample each day only one sample within 24 hours.

The mean and maximum daily sample will be equal in this situation. Lastly, at least one record in the daily summary files for each station provides data for a given day. The format for the Daily Summaries is presented in Appendix **Table A-2**.

Hourly Summary Files

The hourly data is presented in the hourly summary files composed of samples and measurements divided by parameter.

Some samples that EPA receives are in eat intervals other than hourly. For instance, specific samples of SO_2 are reported as 5-minute samples, whereas some samples of dangerous chemicals are recorded. Most PM data are presented as 24-hour samples, while 3-hour measurements are used for most other data. The averages of these samples are present in the daily files even if they are not in the hourly files. No hourly data for a given parameter or group was collected if a file is empty (record count = 0).

Less geographic descriptive information is included to reduce the raw data files' size. The Annual file for the identical monitor contains the Local Site Name, core-based statistical area (CBSA), and other information that may be required is obtainable. The format for each hourly summary is in **Table 3-3** below. Note that the ones presented in the table below are the values used in this research; therefore, some “Field positions” may be missing.

Table 3-3. Hourly Summary Files Format (US EPA Download Files AirData., 2022b)

Field Position	Field Name	Description
1	State Code	State where the monitor is located.
2	County Code	County where the monitor is located
3	Site Num	Unique number developed by the county to distinguish between monitors
6	Latitude	Location coordinate in angular distance north of the equator measured from 0 to 90 degrees.
7	Longitude	Location coordinate in angular distance west of the prime meridian -180 to 0 degrees.
9	Parameter Name	Name of the pollutant or metrological data.
10	Date Local	The calendar date.
11	Time Local	The Local Standard Time.
14	Sample Measurement	The measured standard units of measure for the parameter.
15	Units of Measure	The unit of measure for the parameter.
16	MDL	The Method Detection Limit. The minimum sample concentration that is detectable for the monitor and method.
17	Uncertainty	The total measurement uncertainty associated with a reported measurement.
21	Method Name	A short description of the processes, equipment, and protocols used in gathering and measuring the sample.
22	State Name	The name of the state where the monitoring site is located.
23	County Name	The name of the county where the monitoring site is located.

3.4 Cross-Validation Methodology

Cross-Validation is when the training data is split into k equal parts; in the methodology, the data was split into ten equal parts. Then the model is fitted on k-1 parts, and the test error is calculated using the fitted model on the kth part. Then this is repeated k (10) times, employing a single test set for each data subset. K is usually set to any value between 5 and 20. Below is an image of what k-fold cross-validation looks like visually.

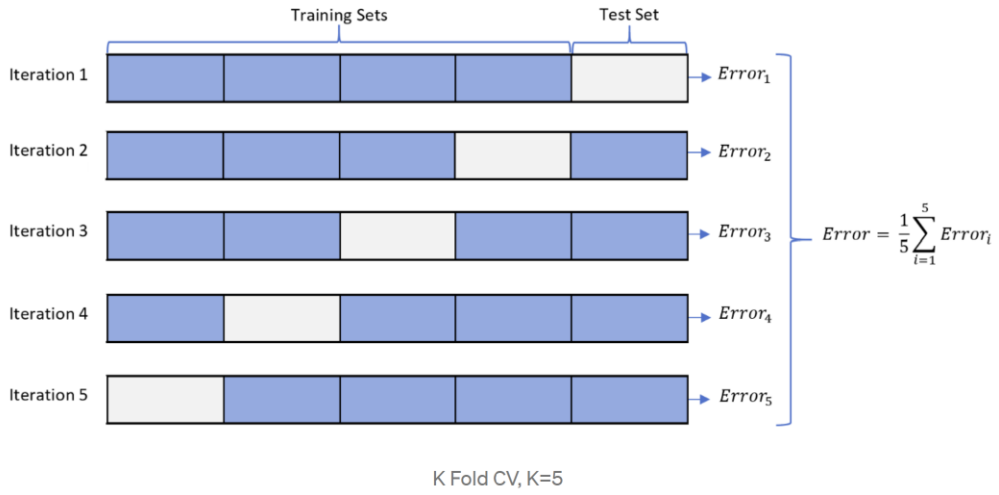

Figure 3-4. 5-fold cross-validation image

(Cross-Validation: K Fold vs. Monte Carlo | by Rebecca Patro | Towards Data Science, 2021)

3.5 Testing with other cites

Utilizing the original methodology Virginia Reference Monitor in Richmond, VA, located at The Richmond 2018 data was run through the model results from testing these values through the model are included in **Table 4-6**. Data from 2018 Richmond, VA:

SO_2, Volume, TEMP, RH, WD, Day, and Month are inputs to the model from the VA reference monitor station and the closest available traffic station.

Chapter 4—Results

4.1 Introduction

The methodology explained in the previous chapter was applied to two out of the five stations in Washington, D.C. Two of the five stations were considered due to their data availability. The results in this chapter of the research questions (R1, R2, R3) will answer and prove or disprove the corresponding hypotheses (H1, H2, H3).

4.2 Data Descriptive Statistics and Missing Values

Table 4-1. Monitoring Data Stations Available for Modeling and Analysis

Station	Year	Pollutant	Site Number and Name	Lat	Lon
1	2022-2016	NO_2, $PM_{2.5}$, SO_2	110010041 River Terrace	38.896	-76.958
2	2017-2015	$PM_{2.5}$	110010042 Hains Point	38.876	-77.034
3	2022-2015	NO_2, $PM_{2.5}$, SO_2	110010043 McMillian Revisor	38.922	-77.013
4	2022-2015	NO_2	110010050 Takoma School	38.970	-77.017
5	2022-2015	NO_2, $PM_{2.5}$	110010051 Anacostia Freeway	38.895	-76.953
6	2022-2018	$PM_{2.5}$	110010053 King Greenleaf Rec Center	38.875	-77.013

As seen in **Table 4-1** above, 4 out of 5 air pollution stations have NO_2 monitors, and only two have SO_2. Therefore, to serve as a proxy for this research, the Station River Terrace and McMillian Revisor were used in developing this analysis. The distance calculation to find the closest volume count stations to the emissions monitor station is located below, with a list of the closest stations in Appendix **Table B-1.**

Equation 4.1

$$Distace\ between\ Stations\ (miles) =$$

$$= ACOS\left(COS\big(RADIANS(90 - Lat1)\big) * COS\big(RADIANS(90 - Lat2)\big)\right.$$

$$+ SIN\big(RADIANS(90 - Lat1)\big) * SIN\big(RADIANS(90 - Lat2)\big)$$

$$\left. * COS\big(RADIANS(Long1 - Long2)\big)\right) * 3958.756$$

Where latitude1 and longitude1 are represented by the emissions monitoring station, and latitude2 and longitude2 are represented by the volume count station. Where 3958.756 miles is the radius of the earth. The Excel file of the stations and their distance from the monitoring stations are seen in Appendix A-3

As seen in Table 4-2a, Descriptive statistics for 2017 training data, and Table 4-2b, Descriptive statistics for 2018 test data and 2017 training data, there is not much variation between years for predictors. Except for Volume and SO_2 but as we can see from the descriptive statistics, the Maximum SO_2 is not a function of vehicle traffic volume because SO_2 is mainly produced when fossil fuels are burned in power stations and other industrial facilities, is the leading cause of SO_2 in the atmosphere. Smaller sources of SO_2 emissions include natural phenomena like volcanoes, industrial activities like metal extraction from ore, and large machinery like locomotives and ships that consume fuel with high sulfur content. NO_2, at least from the descriptive statistics, the NO_2 Maximum increases as a function of Vehicle traffic Volume because the combustion of fuel is the primary source of NO_2 in the atmosphere. NO_2 is created because of emissions from buses, automobiles, power plants, trucks, and off-road vehicles. The distributions of the concentration of the meteorological parameters are comparable across 2017 and 2018; therefore, these two years will serve as a proxy for this research

methodology. For 2017, most of the minimums are the minimum detection limits for the sensor where the data was collected; the same applies to 2018 data.

Table 4-2a. Descriptive Statistics for Numeric Parameters for 2017 training data set

Statistics

Variable	Mean	Minimum	Median	Maximum
X1 SO2_2017	0.8623	0.2000	0.8000	17.6000
X2 Volume_2017	241.33	40.20	245.20	425.70
X3 TEMP_2017	58.978	15.300	60.600	94.600
X4 WD_2017	182.36	8.00	181.00	350.00
X5 RH_2017	57.039	13.300	55.900	92.100
Y NO2_2017	10.665	0.200	8.100	52.500

Table 4-2b. Descriptive Statistics for Numeric Parameters for 2018 testing data set

Statistics

Variable	Mean	Minimum	Median	Maximum
X1 SO2_2018	0.79467	0.20000	0.70000	6.10000
X2 Volume_2018	3954.8	246.0	4339.0	7640.0
X3 TEMP_2018	57.720	9.500	57.400	94.800
X4 WD_2018	180.10	6.00	179.00	344.00
X5 RH_2018	61.400	11.700	60.300	96.300
Y NO2_2018	10.256	1.000	7.700	56.600

4.3 Feature Selection Results

The correlation (Pearson's) coefficient in statistics attributes is dropped by those qualities with low correlation. Suppose there is a medium to highly negative or positive correlation that is close to -1 or 1; show a better correlation. Using 0.2 as the threshold

for essential characteristics, derived by executing the Correlation Attribute Evaluation technique that requires the Ranker search method, the remaining attributes could be

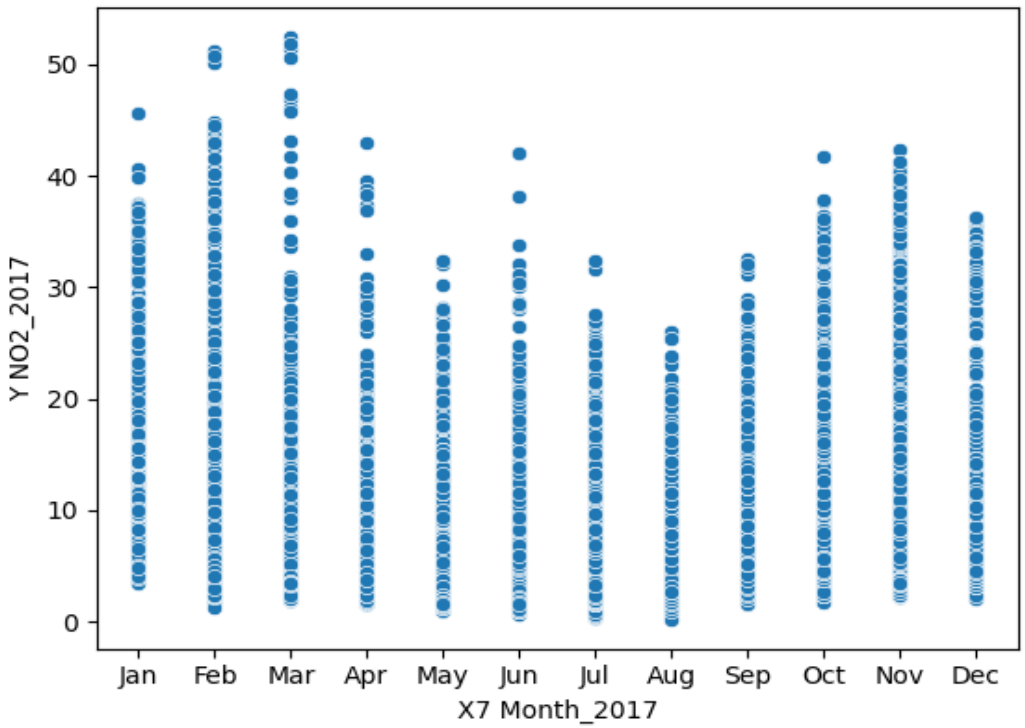

Figure 4-1. NO_2 vs. Month of the year

removed from the model are Month, RH, and Day. Features Volume, Temp, W.D., and SO_2 for significant attributes as chosen by the correlation as seen from the output below in Figure 4-1. Volume once again has the highest correlation of -0.3021. Due to the nature of the data being over some time, it makes sense that there is a negative correlation between volume and NO_2 because, at the beginning of the year, the NO_2 levels are higher in the winter months than lower in the summer months as seen in Figure 4-2.

The feature's scatter plot above shows that NO_2 concentrations are higher during winter. Thirdly, a distinct seasonal pattern is noticeable. Peak NO_2 is substantially greater in winter due to increased usage of combustion-based energy sources for

residential heating and the more prolonged winter airborne NO_2 concentrations (Niu et al., 2022). Finally, NO_2 atmospheric lifespan is mainly influenced by photochemical processes (Matandirotya et al., 2021; Niu et al., 2022). Processes that break down NO_2 are not as rapid in the winter; therefore, it gets trapped atmosphere

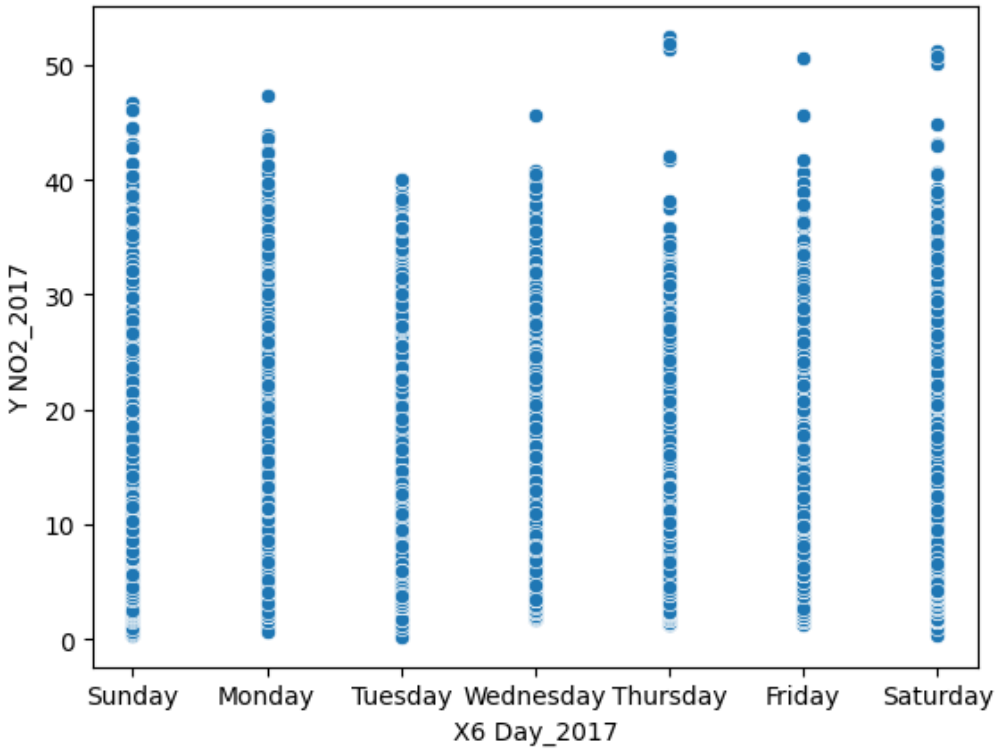

Figure 4-2. NO_2 versus Days of the Week

(US EPA, 2022a).

According to NOAA, the most significant NO_2 emissions are from urban centers; therefore, from the volume vs. days of the week graph, we see that as the volume increases, Thursday, Friday, and Saturday, with Thursday having the highest levels of NO_2 (Winchester et al., 2021).

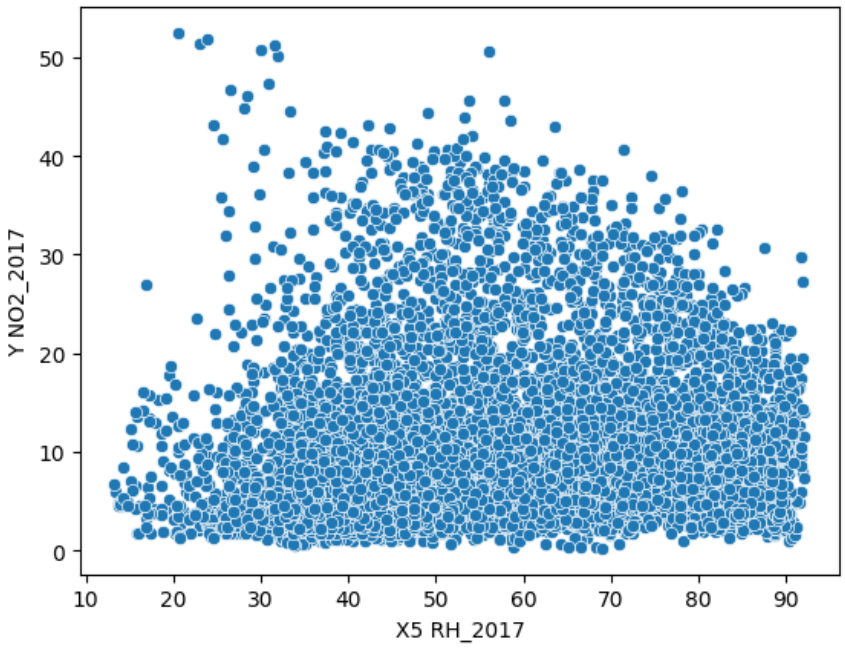

Figure 4-2. NO_2 versus Relative Humidity

As relative humidity is inversely proportional, as the RH increases, the NO_2 concentration levels decrease, as seen in Figure 4-2.

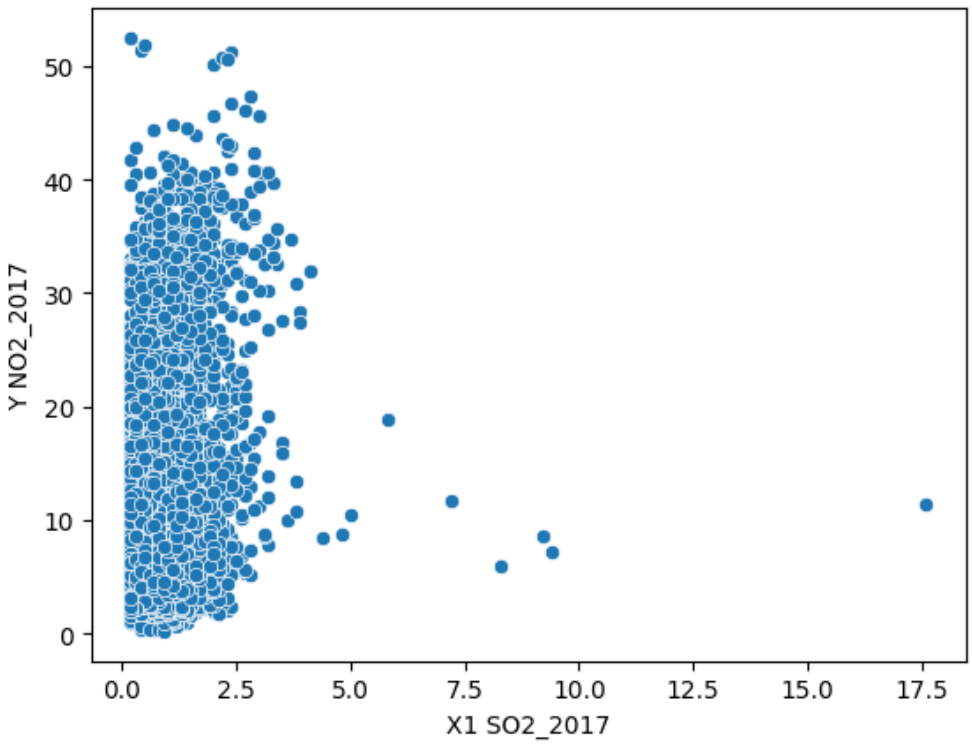

Figure 4-3. NO_2 vs SO_2 concentration

Usually, nitrogen dioxide is not emitted into the atmosphere directly. Instead, Nitrogen dioxide is created through a chemical reaction of nitrogen oxide (NO) other nitrogen oxides (NOx) and other airborne molecules mix; therefore, it is no surprise that NO_2 does not change much as NO_2 increases. The burning of fossil fuels (coal, gas, and oil), particularly the fuel used in automobiles, is the largest source of nitrogen dioxide from human activity. In comparison, power plants and other industrial facilities that use fossil fuels are the leading cause of SO_2 in the atmosphere. Smaller sources of SO_2 emissions include natural phenomena like volcanoes, industrial activities like metal extraction from ore, and large machinery like locomotives and ships that consume fuel with high sulfur content. Therefore, SO_2

and NO_2 are not a result of the same processes as seen in the scatter plot above in Figure 4-3 the SO_2 says constant as NO_2 levels increase.

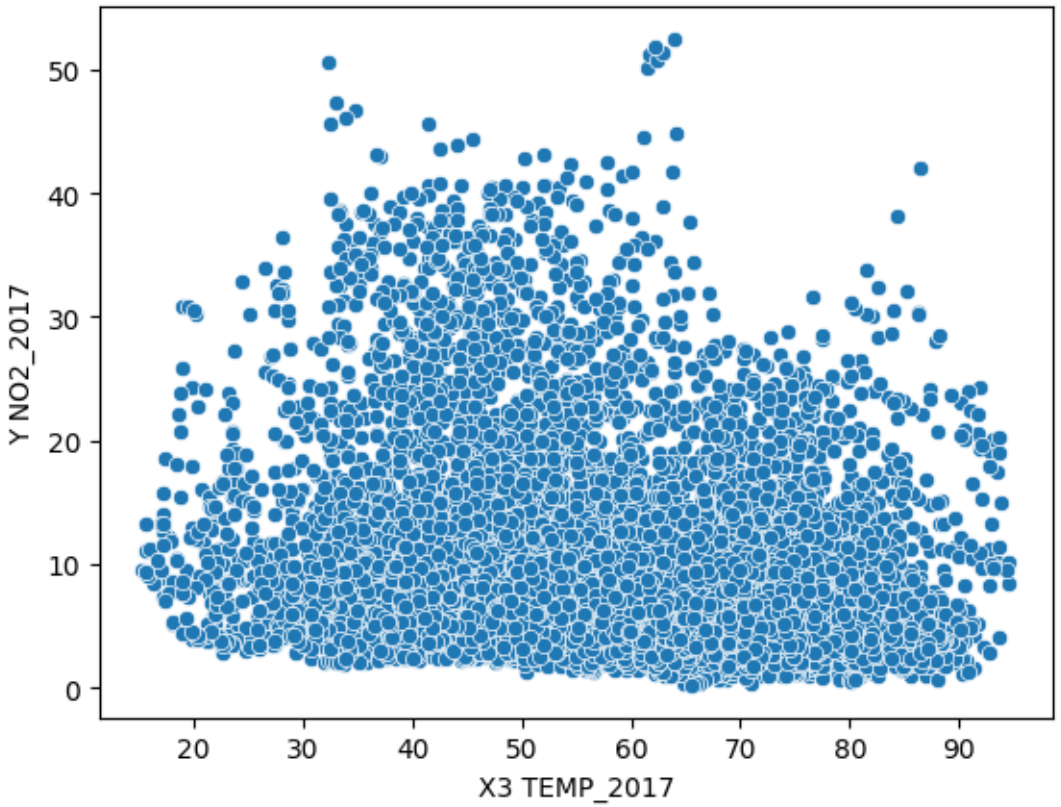

Figure 4-4. Temperature vs. NO_2

Based on the scatter plot for Temperature in 2017, NO_2 and Temperature or negatively correlated; the lower the temperature, the higher the NO_2 emissions, as previously explained for the Month vs. NO_2 plot in Figure 4-1., During colder temperature months which for

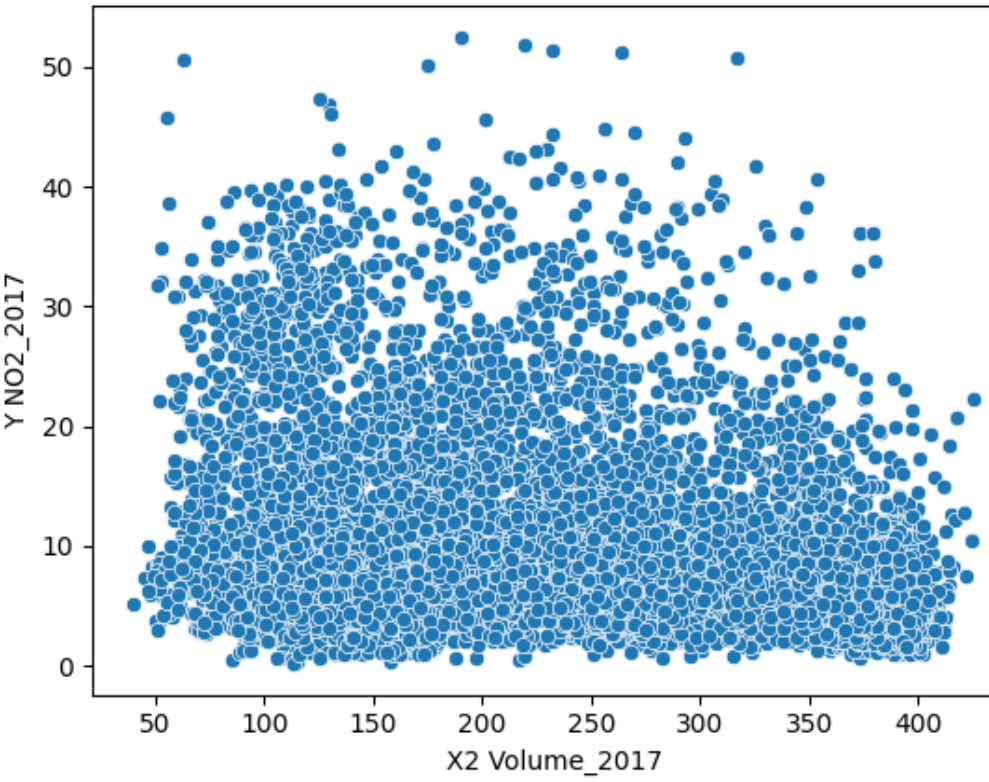

Figure 4-5. NO2 vs. Volume for 2017

Washington D.C. is from December-March the NO_2 emission levels are higher. The distribution of this scatter plot for NO_2 and the variation in traffic from season to season still affects NO_2 despite it not showing up in the scatter plot shown in **Figure 4-5**.

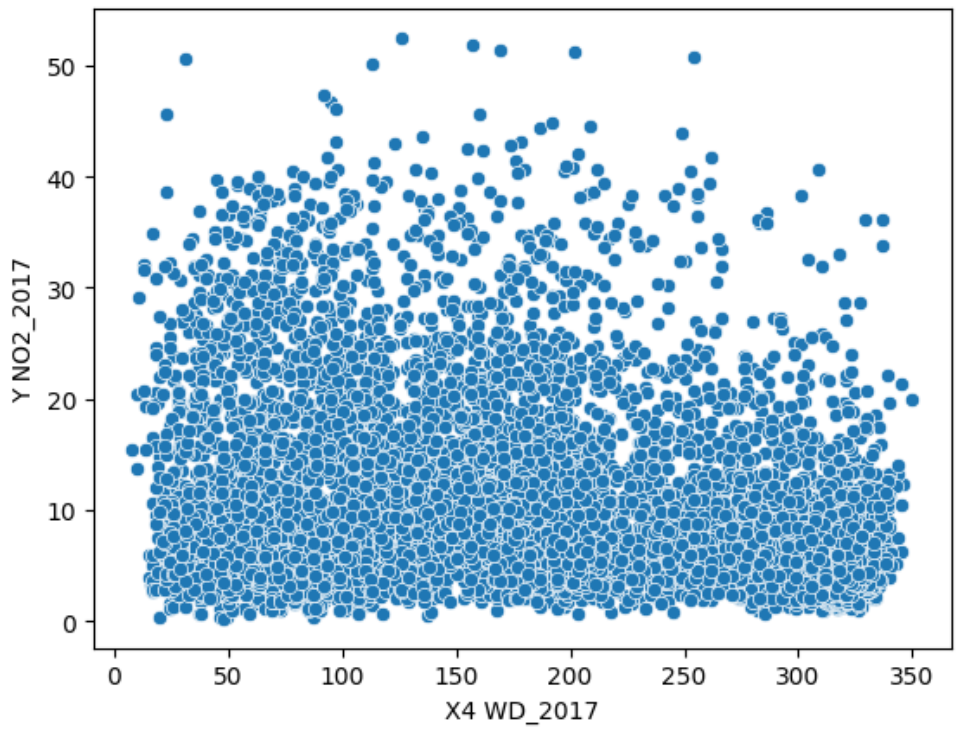

Figure 4-6. NO_2 vs. Wind Direction

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 8 NO2):
        Correlation Ranking Filter
Ranked attributes:
 0.2474    1 SO2
 0.0987    7 Month
 0.0647    5 RH
 0.042     6 Day
-0.2607    4 WD
-0.2611    3 TEMP
-0.3021    2 VOLUME

Selected attributes: 1,7,5,6,4,3,2 : 7
```

Figure 4-7. The output of Correlation Attribute Evaluation

Table 4-3. Feature Selection Optimization Techniques

Optimization Technique	Attribute Evaluator	Inputs Included	Inputs Excluded
Genetic Algorithm	Correlation-based Feature Subset Selection	SO2, Volume, TEMP, Day	R.H., W.D., Month
Forward Stepwise	Correlation-based Feature Subset Selection	SO2, Volume, TEMP, Day	R.H., W.D., Month
Backward Stepwise	Correlation-based Feature Subset Selection	SO2, Volume, TEMP, Day	R.H., W.D., Month
Correlation Ranker	Correlation (Pearson's)	Volume, TEMP, W.D., SO2	RH, Day, Month

Table 4-4. Feature Selection and Model Numbers

Model #	Features	MLR	ANN	RF
Model 1	SO_2, VOLUME, TEMP, WD, RH, Day, Month	0.5586	0.4364	0.4238
Model 2	SO_2, VOLUME, TEMP, WD, RH, Day	0.5233	0.4681	0.4511
Model 3	SO_2, VOLUME, TEMP, WD, RH	0.5235	0.4954	0.4917
Model 4	SO_2, VOLUME, TEMP, WD	0.5228	0.4512	0.4537
Model 5	SO_2, VOLUME, TEMP	0.4333	0.2487	0.3508
Model 6	VOLUME, TEMP	0.2381	0.2118	0.1965
Model 7	VOLUME	0.0639	0.0684	0.02

Selection evaluation was conducted based on the previous feature selection results and other feature configurations without feature selection. From the results presented in Table 4-4. for Multiple Linear Regression (MLR), all features are significantly based on the correlation coefficient, which is Model 1. For ANN, Model 3 contains the highest correlation coefficient, which describes a model that excludes Month and Day categorical variables. The same applies to R.F. Model 3 is strongly positively correlated.

4.4 Model Results

The average model prediction error is expressed using the units of the relevant variable by both MAE and RMSE. The direction of mistakes has no bearing on either metric, ranging from 0 to ∞. Given MAE and RMSE's negative orientation, lower values

are preferable, as seen in the Performance Metric **Table 4-5.**, Performance Metrics Table for Model 1 configuration. Model 1 configuration was selected because through evaluation of Feature Removal with different Model Numbers, **Table 4-4.** Linear regression was performed on the other two methods with all features selected.

Table 4-5. Performance Metric

Correlation coefficient R					
Regression Method	**10-fold**	**9-fold**	**8-fold**	**Test**	**Training**
Linear Regression	**52%**	**52%**	**52%**	**56%**	**53%**
Multilayer Perceptron	**59%**	**56%**	**57%**	**47%**	**70%**
Random Forest	**85%**	**85%**	**85%**	**46%**	**99%**
R-squared					
Regression Method	**10-fold**	**9-fold**	**8-fold**	**Test**	**Training**
Linear Regression	**27%**	**27%**	**27%**	**31%**	**28%**
Multilayer Perceptron	**35%**	**31%**	**33%**	**22%**	**48%**
Random Forest	**72%**	**72%**	**72%**	**21%**	**97%**
Mean absolute error MAE					
Regression Method	**10-fold**	**9-fold**	**8-fold**	**Test**	**Training**
Linear Regression	**5.1981**	**5.1983**	**5.1973**	**6.1314**	**5.1814**
Multilayer Perceptron	**5.1689**	**5.5895**	**5.4286**	**7.3582**	**4.3305**
Random Forest	**2.9248**	**2.8753**	**2.8677**	**6.4647**	**1.1475**
Root mean squared error RMSE					
Regression Method	**10-fold**	**9-fold**	**8-fold**	**Test**	**Training**
Linear Regression	**7.0341**	**7.0335**	**7.0306**	**7.8166**	**7.004**
Multilayer Perceptron	**6.9148**	**7.2921**	**7.136**	**9.2161**	**6.0079**
Random Forest	**4.3977**	**4.3768**	**4.3578**	**7.9864**	**1.7067**
Relative absolute error RAE					
Regression Method	**10-fold**	**9-fold**	**8-fold**	**Test**	**Training**
Linear Regression	**0.83**	**0.83**	**0.83**	**1.04**	**0.83**
Multilayer Perceptron	**0.83**	**0.90**	**0.87**	**1.25**	**0.69**
Random Forest	**0.47**	**0.46**	**0.46**	**1.10**	**0.18**
Root relative squared error RRSE					
Regression Method	**10-fold**	**9-fold**	**8-fold**	**Test**	**Training**
Linear Regression	**0.85**	**0.85**	**0.85**	**1.00**	**0.85**
Multilayer Perceptron	**0.84**	**0.88**	**0.87**	**1.17**	**0.73**
Random Forest	**0.53**	**0.53**	**0.53**	**1.02**	**0.21**

```
NO2 =

     47.8687 * SO2 +
     -7.5731 * WD +
      1.7325 * RH +
      0.9597 * Day=Saturday,Wednesday,Monday,Thursday,Friday,Tuesday +
      1.4527 * Day=Wednesday,Monday,Thursday,Friday,Tuesday +
     -0.5766 * Day=Monday,Thursday,Friday,Tuesday +
      0.7651 * Day=Thursday,Friday,Tuesday +
      0.6603 * Month=August,September,April,March,July,October,January,November,February,December +
      0.6806 * Month=September,April,March,July,October,January,November,February,December +
      0.6051 * Month=April,March,July,October,January,November,February,December +
      1.5748 * Month=March,July,October,January,November,February,December +
      3.2394 * Month=January,November,February,December +
      2.1815 * Month=February,December +
      6.0048
```

Regression Equation

Y NO2_2017 = 10.355 + 47.88 X1 SO2_2017 - 8.548 X2 Volume_2017 + 1.606 X3 TEMP_2017
+ 1.767 X5 RH_2017 + 0.0 X6 Day_2017_Friday - 0.860 X6 Day_2017_Monday
- 1.750 X6 Day_2017_Saturday - 2.700 X6 Day_2017_Sunday
- 0.255 X6 Day_2017_Thursday - 0.038 X6 Day_2017_Tuesday
- 0.290 X6 Day_2017_Wednesday + 0.0 X7 Month_2017_April
- 1.265 X7 Month_2017_August + 7.021 X7 Month_2017_December
+ 6.889 X7 Month_2017_February + 4.680 X7 Month_2017_January
+ 1.761 X7 Month_2017_July - 1.850 X7 Month_2017_June
+ 1.675 X7 Month_2017_March - 2.013 X7 Month_2017_May
+ 4.871 X7 Month_2017_November + 1.309 X7 Month_2017_October
- 0.599 X7 Month_2017_September

Figure 4-8. Linear Regression Model Equation from WEKA and Minitab

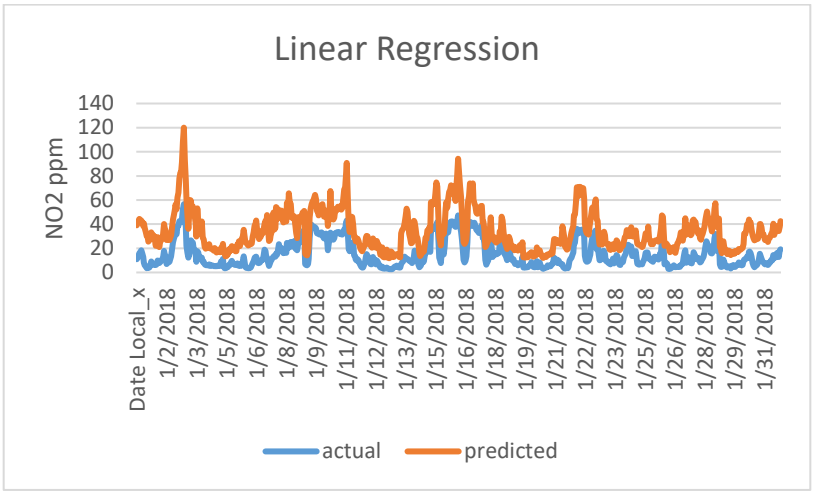

Figure 4-9. NO_2 ppm vs. Actual and Predicted levels for 2018 using Multiple Linear Regression

In **Figure 4-9.,** the multiple linear regression model represents the results from the Model 1 features selected in **Table 4-4**. In the temporal pattern of NO_2 for the first month of

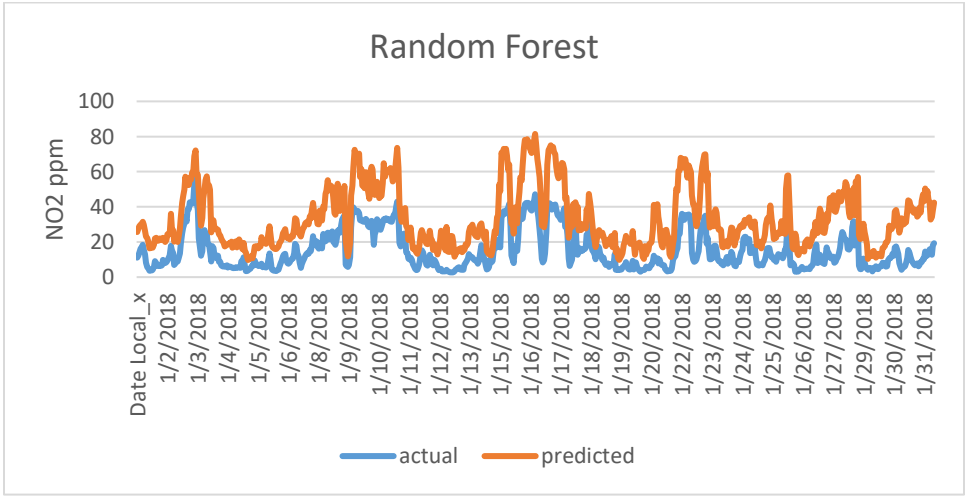

Figure 4-10. NO_2 ppm vs. Actual and Predicted levels for 2018 using Random Forest Algorithm

2018, the blue line represents the NO_2 values reported by DOEE to EPA. The orange line represents the predicted 2018 NO_2 levels forecasted by the model.

In **Figure 4-10.,** the Random Forest Algorithm model represents the results from the Model 3 features selected in **Table 4-4**. In the temporal pattern of NO_2 for the first month of 2018, the blue line represents the NO_2 values reported by DOEE to EPA. The orange line represents the predicted 2018 NO_2 levels forecasted by the model. The mean absolute error MAE is 7.5491, and the RMSE is 9.4603. With 10-fold cross-validation with Model 3 and Random Forest algorithm, the correlation coefficient R is .67 making the R^2 .45. In all model parameters are reinstated for Model 1 with Random Forest algorithm with 10-fold-cross validation, the correlation coefficient R is .85, R^2 is .72 with an MAE of 2.9 and an RMSE of 4.4. Both are strong correlations between the model and the predicted NO_2 2018 target value.

4.5 Improving Results

Results were improved based on the following metrics: (i) shorting the distance to monitor, (ii) adjusting the wind direction from numeric to nominal, (iii) adjusting the hyperparameters of the ML algorithms (iv) Evaluating by season.

In the results presented in 4.4, the evaluation station had a location 6 miles away from the reference monitor. This station was selected due to its large amount of data for each hour and day. The number of data points for 2017 training and 2018 test data was cross-referenced for each volume station by assessing if it had both 2017 and 2018; 1316 traffic count stations were down-selected into 13 Volume Count stations, as seen in **Table B-1**. By evaluating distances by **Equation 4.1** in **Table B-1**, we can see that the

Anacostia Freeway Reference Monitor that measures NO_2 has a Volume Count station with available data for 2017 and 2018 with the closest proximity for hourly data

Next Wind Direction parameter was adjusted from numerical degree compass to nominal wind directions by the following Excel Equation 4.2:

Equation 4.2

$$Wind\ Direction\ (Directions) = CHOOSE(1 + ABS(ROUND(\ Wind\ Direction\ (Degree\ Compass)\ /45,0)), "N", "NE", "E", "SE", "S", "SW", "W", "NW", "N")$$

Hyperparameters are used to configure aspects of ML algorithms. However, they can considerably impact the final model's performance (Yang & Shami, 2020). Therefore, finding the optimal values can be difficult. Therefore, in this study, various hyperparameters were varied to optimize results.

As mentioned previously, NO_2 has a seasonal nature; therefore, the data was split into four datasets for 2017 and 2018 for Winter, Spring, Summer, and Fall seasons. The Spring Model from Minitab's available data had an R-squared of 53.08%. The Summer model with an R-squared of 54.08%. The Fall model with an R-squared of 64.71%. The Winter Model had an R-squared of 53.68%.

Regression Equation

NO2 = 39.18 - 0.000783 Volume - 0.3403 TEMP + 0.0 Month_Jul - 1.864 Month_Jun
+ 0.0 Day_Friday + 1.784 Day_Monday - 1.873 Day_Saturday - 2.053 Day_Sunday
+ 0.359 Day_Thursday + 2.452 Day_Tuesday + 3.005 Day_Wednesday + 0.0 WD_E - 5.94 WD_N
- 2.84 WD_NE - 5.44 WD_NW + 7.64 WD_S + 2.33 WD_SE + 4.78 WD_SW + 0.39 WD_W

Regression Equation

NO2 = 14.17 + 19.62 RH + 0.000981 Volume + 0.0 WD_E - 13.99 WD_N - 4.66 WD_NE - 13.17 WD_NW
- 0.57 WD_S - 0.17 WD_SE - 6.32 WD_SW - 8.51 WD_W + 0.0 Day_Friday - 2.07 Day_Monday
- 6.194 Day_Saturday - 0.67 Day_Sunday + 1.420 Day_Thursday - 1.102 Day_Tuesday
- 1.955 Day_Wednesday

Regression Equation

NO2 = 1.133 + 0.0737 TEMP + 0.12334 RH + 0.000253 Volume + 0.0 Month_Nov - 3.828 Month_Oct
+ 0.0 Day_Friday + 0.146 Day_Monday + 0.362 Day_Saturday - 1.390 Day_Sunday
+ 0.796 Day_Thursday - 0.438 Day_Tuesday + 1.242 Day_Wednesday + 0.0 WD_1_E
- 4.523 WD_1_N - 1.212 WD_1_NE - 4.434 WD_1_NW + 1.152 WD_1_S + 0.910 WD_1_SE
+ 0.761 WD_1_SW - 1.977 WD_1_W

Regression Equation

NO2 = 78.5 - 0.725 TEMP - 13.87 RH + 0.000224 Volume + 0.0 WD_E - 12.25 WD_N - 9.97 WD_NE
- 11.44 WD_NW + 4.58 WD_S - 1.85 WD_SE - 0.35 WD_SW - 3.65 WD_W + 0.0 Day_Friday
- 4.72 Day_Monday + 4.23 Day_Saturday - 0.56 Day_Sunday - 0.83 Day_Thursday
- 0.66 Day_Tuesday - 3.27 Day_Wednesday

Figure 4-11. Improved Regression Equation for (a) Spring (b) Summer (b) Fall (c) Winter

4.6 Testing on Richmond, VA

Table 4-6a Descriptive statistics of Richmond, VA 2018

Statistics

Variable	Mean	Minimum	Median	Maximum
SO2_2018	0.27549	0.20000	0.20000	5.00000
TEMP_2018	66.716	22.800	71.400	96.300
WD_2018	161.04	1.00	152.00	359.00
RH_2018	70.710	13.000	75.000	98.000
Volume_2018	1473.1	55.0	1443.0	3513.0
NO2_2018	5.4418	0.4000	3.7000	36.4000

Table 4-6b Descriptive statistics of Washington, D.C.

Statistics

Variable	Mean	Minimum	Median	Maximum
SO2_2017	0.8188	0.2000	0.7000	17.6000
TEMP_2017	65.035	15.600	68.400	94.600
WD_2017	176.13	8.00	173.00	350.00
RH_2017	58.788	20.100	58.200	91.700
Volume_2017	4223.2	484.0	4581.0	8169.0
NO2_2017	10.124	1.000	7.400	42.300

The descriptive statistics for 2017 and 2018 are relatively different because they are two different cities, one being Richmond, Virginia, and the other in Washington, D.C. however, they have similar climates.

Following the same steps in the method using 2018 Richmond VA as a test data set from MLR, the correlation coefficient is 32%. For Multilayer perceptron, the Correlation coefficient decreases to 27%. With RF, the sets the number of randomly chosen attributes. If 0, $int(\log_2(\#predictors) + 1)$ is used to 2, the correlation coefficient is 32%. There is a decrease in performance due to the difference in the amount of traffic volume in Richmond, VA and Washington, D.C. as seen for the descriptive statistics shown in Table 4-6 a and b for above.

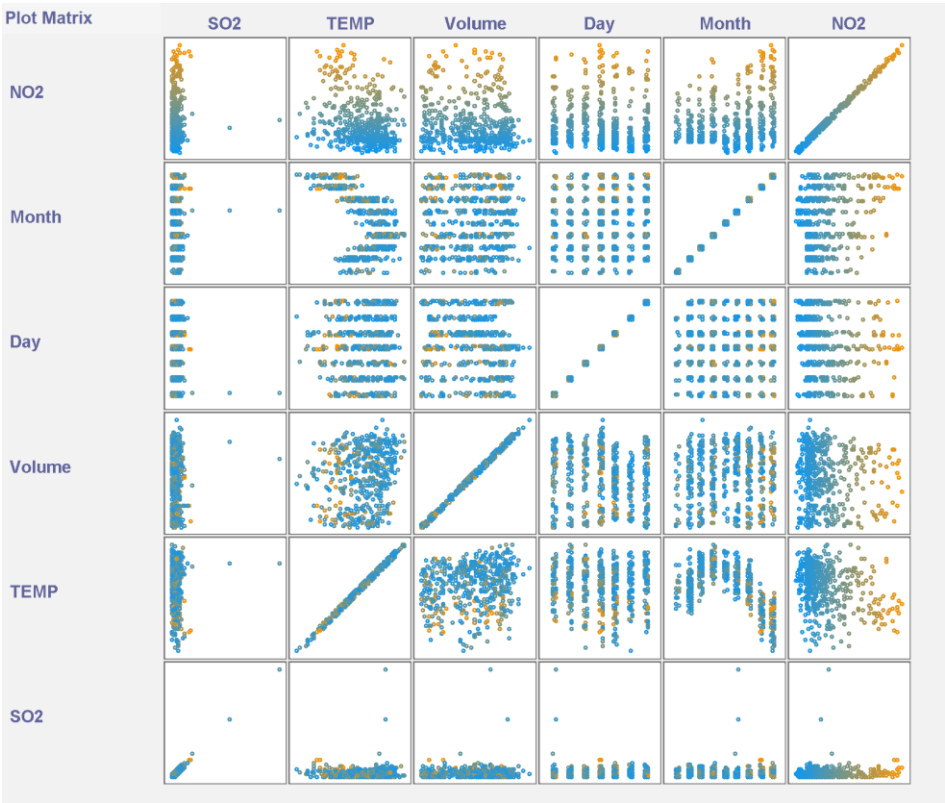

Figure 4-12. scatter plots from Richmond, VA, and Washington, D.C. model

2018 data was inputted into the spring 2017 data to test on another smart city. Richmond, VA, data was used to try out the model presented in **Figure 4-11**(a). The NO_2 temporal pattern follows NO_2 Truth well for Richmond, VA input values, as seen below.

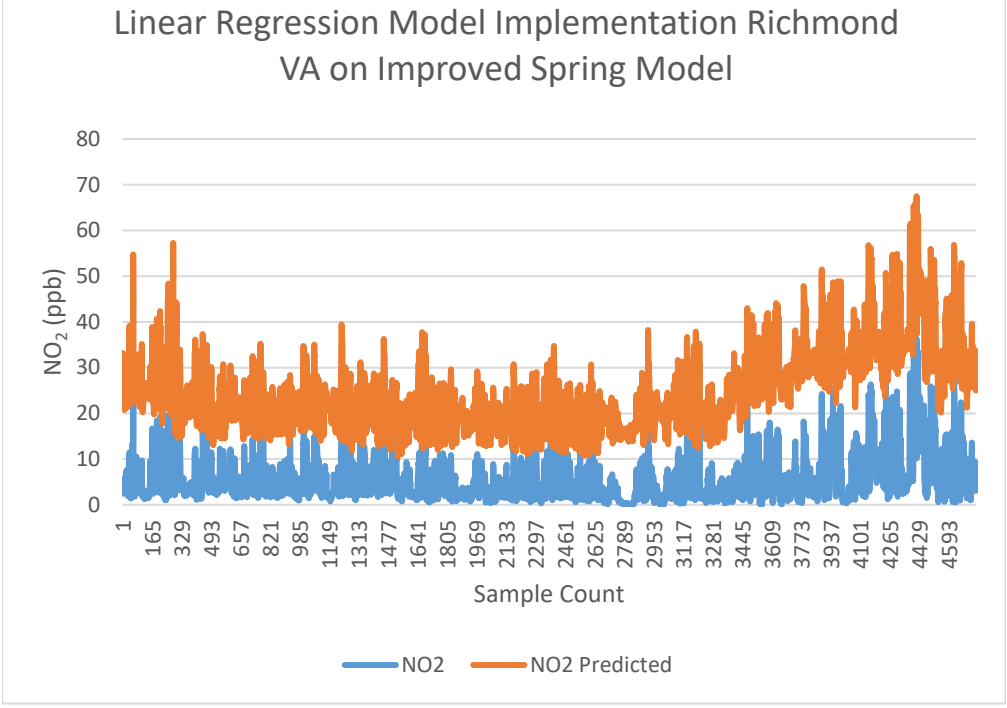

Figure 4-13. Linear Regression Implementation Using Richmond, VA

Chapter 5—Discussion and Conclusions

5.1 Discussion

Emission levels are challenging to analyze due to the unknown specific volumes in the outdoor air. Due to different motivations in data collection from year-to-year, workarounds were needed for the framework of the model. In addition, the model had to be flexible enough to adjust to different variations in years Maximum, Minimum, Mean, and Median. For Transportation monitoring systems, the context of finding the level of emissions and its association with a station location was achieved.

The model research addressed the issues by providing the neighborhood-level emissions model to use in areas with volume stations available, where no FEM/FRM model station is available.

5.2 Conclusions

The integrated Emissions monitoring model, using traffic volume data and meteorological data, was developed in this praxis. Variation in NO_2 predicted from 2018 can be explained by the fact that the traffic monitor itself is not located right on top of the monitor used to build the model but is within 6 miles to .5 miles away from the monitor. In addition, the model's traffic modeling portion showed that traffic was occurring in that area for all six stations. However, the roadway emissions would become more dangerous for the residents of these areas within Washington, D.C., being evaluated.

RQ1: What data predictors can be used to forecast NO_2 levels in Washington, D.C.?

H1: Predictors such as Relative Humidity, Temperature, Traffic Volume, Wind Direction, Wind Speed, Month, Day, and Hour are significant in developing a NO_2 forecasting model.

As seen in the results in chapter 4 for feature selection, SO_2, Volume, and Temp were the common predictors that improved model performance for the ML algorithms of Random Forest and Multilayer Perceptron—using a correlation coefficient cut of 0.2. While hypothesis 1 was that all the predictors as significant. For the improved performance results, with a p-value <0.5, Relative Humidity, Volume, and Wind direction in N, NE, NW, SW, W. Day =Monday, Saturday are significant.

RQ2: Which method provides the best performance to predict NO2 Levels for different seasons in Washington, D.C.?

H2: Machine learning algorithm Random Forest provides the best performance model in forecasting NO2 levels in Washington, D.C.

While Random Forest performed well for testing by the cross-validation at 85% and Training at 98%, Multiple Linear Regression performed better with the use of the test from a different year set at a correlation coefficient of 56%, which was shown to be higher than some of the reported values in the literature as seen in Table 2-1 from the Lucena-Sanchez et al., 2020 that showed a value of 51%. RF performed ANN values from 77%, where there were no values used from volume count stations, and the meteorological stations were located at the exact location of the NO_2 monitor. This model is created for an entire year span, not just a season, unlike the models presented in the literature. At the same time, the improved model showed performance R-squared=74% for improved testing with 2018 data for Random Forest still performs Linear Regression

at 56% and Multilayer Perceptron at 53% for significant features of Wind Direction, Day, Volume, Temperature, and Relative Humidity.

RQ3: Can this model be developed to forecast NO_2 levels in D.C. and be used in other cities to forecast vehicle pollutants?

H3: This model can predict NO_2 levels for different seasons in Richmond, Virginia.

The values put into the model include volume and meteorological data. At the same time, the Richmond, Virginia truth NO_2 is compared to the predicted data using the model developed by using the Washington D.C. model results of correlation coefficient of with Random forest 32% and Linear regression 27% with improved values to 65% and 56% when using both 2017 and 2018 data from VA stations.

5.3 Contributions to Body of Knowledge

This praxis impacts the body of knowledge in the following crucial aspects:

1. Emissions Forecasting for NO_2 Models

   This praxis provides a model to examine and implement the usage of historical traffic volume, and meteorological data sources for creating predictive models can be used to forecast NO_2. These models use existing volume count stations for monitoring emissions within Washington, D.C. Without having to pay for any new monitoring stations in Washington, D.C.; the region and the individual areas may use these models to increase the ability to meet the NAAQS and not rely solely on the data gathering of emissions through the FEM network. It is expanding the monitors from five monitoring stations to over a thousand.

2. Usable forecasting model for cities that do not have existing monitors

This praxis concludes that by not only implementing no-cost systems for Washington D.C., which will need some form of maintenance in the future by maintaining the five systems in place. However, they can also serve as truth for the areas with those systems without downtime for creating new high-cost FEMs systems.

3. Features needed to predict NO_2.

   Due to the multiple features explored in this study, the model's flexibility can be used in different smart cities with volume count stations but does not currently monitor emission levels.

5.4 Recommendations for Future Research

While the values are highly correlated, the research model can be improved by decreasing the distance between the emissions monitoring station and the vehicle traffic count station. There are stations closer to the emissions monitoring station; however, for 2017, that station did not collect data. Since the volume station already exists, it can be utilized in future research. Also, this methodology can be applied to other pollutants that are one of the five NAAQS pollutants. Also, improvement of data optimization and be used to improve results. In the future, this developed forecasting no-cost pollutant forecasting model can extend the five reference monitors to 1316 monitors in Washington, D.C., as shown in Figure 5-1a to 5-1b. Increasing the amount of pollutant monitoring stations in D.C. wards that do not have FEM and FRM monitors can help better support the EPA’s goal of improving neighborhood-level monitoring.

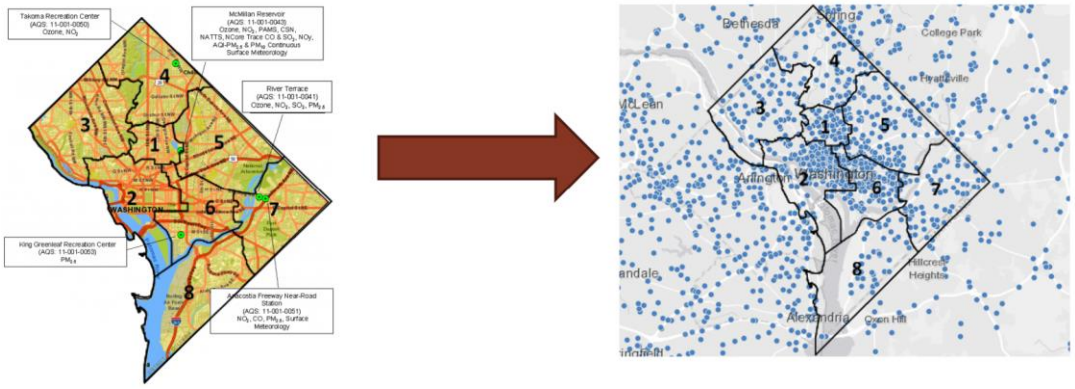

Figure 5-1. a) Reference monitor locations in washinton D.C.with Ward map

b) Traffic monitors with Ward map.

References

Abdella, M., & Marwala, T. (2005, April). The use of genetic algorithms and neural networks to approximate missing data in database. In IEEE 3rd International Conference on Computational Cybernetics, 2005. ICCC 2005. (pp. 207-212). IEEE.

Afrin, T., & Yodo, N. (2020). A survey of road traffic congestion measures towards a sustainable and resilient transportation system. Sustainability, 12(11), 4660. https://doi.org/10.3390/su12114660

Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A. H., Alomari, S., Gök, M., Alaabdin, A. M. Z., & Abdulrhman, S. H. (2022). Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning. Iraqi Journal for Computer Science and Mathematics, 3(1), 115-123.

Ahmed, M. M., & Ghasemzadeh, A. (2018). The impacts of heavy rain on speed and headway behaviors: an investigation using the SHRP2 naturalistic driving study data. *Transportation research part C: emerging technologies*, *91*, 371-384.

Alloghani, M., Al-Jumeily, D., Aljaaf, A. J., Khalaf, M., Mustafina, J., & Tan, S. Y. (2020). The application of artificial intelligence technology in healthcare: a systematic review. In International conference on applied computing to support industry: Innovation and technology (pp. 248-261). Springer, Cham.

AlNuaimi, N., Masud, M. M., Serhani, M. A., & Zaki, N. (2020). Streaming feature selection algorithms for big data: A survey. Applied Computing and Informatics.

Alves, C. A., Evtyugina, M., Vicente, A. M. P., Vicente, E. D., Nunes, T. v., Silva, P. M. A., Duarte, M. A. C., Pio, C. A., Amato, F., & Querol, X. (2018). Chemical profiling of PM10 from urban road dust. *Science of the Total Environment*, 634, 41–51. https://doi.org/10.1016/j.scitotenv.2018.03.338

Azeem, S., Imran, W., & Amin, G. (2019). Mix Use Traffic and Added Impedance Affecting Traffic Flow Variables. *SSRG International Journal of Civil Engineering*, 6, 48-50.

Barth, M., & Boriboonsomsin, K. (2008). Real-world carbon dioxide impacts of traffic congestion. *Transportation research record*, *2058*(1), 163-171.

Beaudoin, J., Farzin, Y. H., & Lawell, C. Y. C. L. (2015). Public transit investment and sustainable transportation: A review of studies of transit's impact on traffic congestion and air quality. *Research in Transportation Economics*, 52, 15-22.

Ben-Elia, E., & Ettema, D. (2011). Rewarding rush-hour avoidance: A study of commuters' travel behavior. *Transportation Research Part A: Policy and Practice*, 45(7), 567-582.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Çolak, S., Lima, A., & González, M. C. (2016). Understanding congested travel in urban areas. *Nature communications*, 7(1), 1-8.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In Ensemble machine learning (pp. 157-175). Springer, Boston, MA.

Dallmann, T., & Menon, A. (2016). Technology pathways for diesel engines used in non-road vehicles and equipment. *International Council on Clean Transportation* (ICCT): Washington, DC, USA.

Dharmage, S. C., Perret, J. L., & Custovic, A. (2019). Epidemiology of asthma in children and adults. *Frontiers in pediatrics*, 7, 246.

Dhaubanjar, S., Davidsen, C., & Bauer-Gottwein, P. (2017). Multi-objective optimization for analysis of changing trade-offs in the Nepalese water–energy–food nexus with hydropower development. Water, 9(3), 162.

DOEE Ambient Air Quality Monitoring. (2022). Retrieved November 27, 2022, from https://doee.dc.gov/service/ambient-air-quality-monitoring

Dowling, R., Skabardonis, A., Carroll, M., & Wang, Z. (2004). Methodology for measuring recurrent and nonrecurrent traffic congestion. *Transportation Research Record*, 1867(1), 60-68.

Downs, A. (2004). *Stuck in traffic: Coping with peak-hour traffic congestion.* Brookings Institution Press. http://www.jstor.org/stable/10.7864/j.ctt1vjqprt

Elangasinghe, M. A., Singhal, N., Dirks, K. N., & Salmond, J. A. (2014). Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmospheric pollution research, 5(4), 696-708.

Evuti, A. M. (2013). A synopsis on biogenic and anthropogenic volatile organic compounds emissions: hazards and control. *Int. J.* Eng, 2(5), 145.

Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan

García, M. V., & Aznarte, J. L. (2020). Shapley additive explanations for NO_2 forecasting. Ecological Informatics, 56, 101039.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. Pattern recognition letters, 27(4), 294-300.

Hammad, H. M., Ashraf, M., Abbas, F., Bakhat, H. F., Qaisrani, S. A., Mubeen, M., Fahad, S., & Awais, M. (2019). Environmental factors affecting the frequency of road traffic accidents: a case study of sub-urban area of Pakistan. *Environmental Science and Pollution Research*, 26(12), 11674-11685.

Hasan, U., Whyte, A., & Al Jassmi, H. (2019). Critical review and methodological issues in integrated life-cycle analysis on road networks. Journal of Cleaner Production, 206, 541-558.

Hassan, R., Rahman, M., & Hamdan, A. (2022, May). Assessment of air quality index (AQI) in Riyadh, Saudi Arabia. In IOP Conference Series: Earth and Environmental Science (Vol. 1026, No. 1, p. 012003). IOP Publishing.

Heaslip, K., Goodall, N. J., Kim, B., & Aad, M. A. (2020). Assessment of Capacity Changes Due to Automated Vehicles on Interstate Corridors (No. FHWA/VTRC 21-R1). Virginia. Dept. of Transportation.

Hernandez, J. B., & Kim, P. (2021). *Epidemiology morbidity and mortality*. StatPearls.

Huang, Y., Mok, W. Chen, Yam, Y. Shing, Zhou, J. L., Surawski, N. C., Organ, B., Chan, E. F. C., Mofijur, M., Mahlia, T. M. I., & Ong, H. C. (2020). Evaluating in-use vehicle emissions using air quality monitoring stations and on-road remote sensing systems. Science of the Total Environment, 740. https://doi.org/10.1016/j.scitotenv.2020.139868

Hussain, A. (2007). Ambient air quality monitoring of Aligarh city: A case study. no. January, 2016.

Hurwitz, D., Heaslip, K., & Moore, D. (2012). Relating transportation systems management and operations strategies to policy goals: A framework for quantitative decision making. Engineering Management Journal, 24(3), 32-42. 2.

Iskandaryan, D., Ramos, F., & Trilles, S. (2022). Bidirectional convolutional LSTM for the prediction of nitrogen dioxide in the city of Madrid. PloS one, 17(6), e0269295.

JavaTpoint Machine Learning Random Forest Algorithm. (2022) “Random Forest Algorithm.” https://www.javatpoint.com/machine-learning-random-forest-algorithm

Jin, Y., Andersson, H., & Zhang, S. (2016). Air pollution control policies in China: a retrospective and prospects. International journal of environmental research and public health, 13(12), 1219.

Kalogirou, S. A. (2009). Artificial neural networks and genetic algorithms in energy applications in buildings. Advances in Building Energy Research, 3(1), 83-119.

Kannan, D., Khodaverdi, R., Olfat, L., Jafarian, A., & Diabat, A. (2013). Integrated fuzzy multi criteria decision making method and multi-objective programming approach for supplier selection and order allocation in a green supply chain. Journal of Cleaner production, 47, 355-367.

Keehan, C. J. (2018). Lessons from cancer alley: How the clean air act has failed to protect public health in Southern Louisiana. Colo. Nat. Resources Energy & Envtl. L. Rev., 29, 341.

Khalilpourazari, S., & Arshadi Khamseh, A. (2019). Bi-objective emergency blood supply chain network design in earthquake considering earthquake magnitude: a comprehensive study with real world application. Annals of Operations Research, 283(1), 355-393.

Kheirbek, I., Haney, J., Douglas, S., Ito, K., & Matte, T. (2016). The contribution of motor vehicle emissions to ambient fine particulate matter public health impacts in New York City: a health burden assessment. *Environmental Health*, 15(1), 1-14.

Khorshid, M., Abou-El-Enien, T., & Soliman, G. (2015). A comparison among support vector machine and other machine learning classification algorithms. IPASJ International Journal of Computer Science (IIJCS).

Kim, K., & Hong, J. S. (2017). A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis. Pattern Recognition Letters, 98, 39-45.

Kim, H. S., Kim, D. S., Kim, H., & Yi, S. M. (2012). Relationship between mortality and fine particles during Asian dust, smog–Asian dust, and smog days in Korea. *International journal of environmental health research*, 22(6), 518-530.

Koprinska, I., Rana, M., & Agelidis, V. G. (2015). Correlation and instance based feature selection for electricity load forecasting. Knowledge-Based Systems, 82, 29-40.

Li, M. L., Bouardi, H. T., Lami, O. S., Trikalinos, T. A., Trichakis, N., & Bertsimas, D. (2022). Forecasting COVID-19 and analyzing the effect of government interventions. *Operations Research*.

Li, Z., Liang, C., Hong, Y., & Zhang, Z. (2022). How do on-demand ridesharing services affect traffic congestion? The moderating role of urban compactness. *Production and Operations Management*, 31(1), 239-258.

Lin, L., Wang, Q., & Sadek, A. W. (2015). A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transportation Research Part C: Emerging Technologies, 55, 444-459.

Lucena-Sánchez, E., Jiménez, F., Sciavicco, G., & Kamińska, J. (2020, May). Simple versus composed temporal lag regression with feature selection, with an application to air quality modeling. In 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS) (pp. 1-8). IEEE.

Mac Nally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: the distinction between–and reconciliation of–'predictive'and 'explanatory'models. Biodiversity & Conservation, 9(5), 655-671.

Mahmassani, H. S., Dong, J., Kim, J., Chen, R. B., & Park, B. (2009). *Incorporating weather impacts in traffic estimation and prediction systems. U.S. Department of Transport, Washington, 108*.

Margiotta, R. A., & Spiller, N. C. (2012). Recurring Traffic Bottlenecks: A Primer. Focus.

.

Matandirotya, N. R., & Burger, R. P. (2021). Spatiotemporal variability of tropospheric NO_2 over four megacities in Southern Africa: implications for transboundary regional air pollution. Environmental Challenges, 5, 100271.

Mensah, J., Annan, J., & Andoh-Baidoo, F. (2014). Assessing the Impact of Vehicular Traffic on Energy Demand in the Accra Metropolis. *Journal of Management Policy & Practice, 15*(4).

Michalak, T. P., Aadithya, K. V., Szczepanski, P. L., Ravindran, B., & Jennings, N. R. (2013). Efficient computation of the Shapley value for game-theoretic network centrality. Journal of Artificial Intelligence Research, 46, 607-650.

Miraftabzadeh, S. M., Longo, M., Foiadelli, F., Pasetti, M., & Igual, R. (2021). Advances in the Application of Machine Learning Techniques for Power System Analytics: A Survey. Energies, 14(16), 4776. MDPI AG. Retrieved from http://dx.doi.org/10.3390/en14164776

Mishra, D., & Goyal, P. (2016). Neuro-fuzzy approach to forecast NO2 pollutants addressed to air quality dispersion model over Delhi, India. Aerosol and air quality research, 16(1), 166-174.

Mousazadeh, M., Paital, B., Naghdali, Z., Mortezania, Z., Hashemi, M., Karamati Niarah, E., Aghababaei, M., Ghorbankhani, M., Lichtfouse, E., Sillanpää, M., Hashim, K.S, & Mahdi Emamjomeh, M. (2021). Positive environmental effects of the coronavirus 2020 episode: a review. *Environ Dev Sustain* 23, 12738–12760 https://doi.org/10.1007/s10668-021-01240-3

Nagy, A. M., & Simon, V. (2018). Survey on traffic prediction in smart cities. Pervasive and Mobile Computing, 50, 148–163. https://doi.org/10.1016/j.pmcj.2018.07.004

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

Nellore, K., & Hancke, G. P. (2016). A survey on urban traffic management system using wireless sensor networks. *Sensors*, 16(2), 157. https://doi.org/10.3390/s16020157

Niu, H., Zhang, C., Hu, W., Hu, T., Wu, C., Hu, S., ... & Fan, J. (2022). Air Quality Changes during the COVID-19 Lockdown in an Industrial City in North China: Post-Pandemic Proposals for Air Quality Improvement. Sustainability, 14(18), 11531.

Ogunrinde, R. B., & Lebile, O. (2015). On mathematical model of traffic control. *Mathematical Theory and Modeling*, 5(1), 58-69. complex adaptive systems perspective. *Decision sciences*, 38(4), 547-580.

Patton, A., Datta, A., Zamora, M. L., Buehler, C., Xiong, F., Gentner, D. R., & Koehler, K. (2022). Non-linear probabilistic calibration of low-cost environmental air pollution sensor networks for neighborhood level spatiotemporal exposure assessment. Journal of Exposure Science & Environmental Epidemiology, 1-9.

Patro, R. (2021). Cross-Validation: K Fold vs Monte Carlo |Towards Data Science. Retrieved November 27, 2022, from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

Pise, N. N., & Kulkarni, P. (2008, December). A survey of semi-supervised learning methods. In 2008 International conference on computational intelligence and security (Vol. 2, pp. 30-34). IEEE.

Reed, T. (2019). Inrix global traffic scorecard.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

Saliby, S., Tong, A., Ciesielski, S., Lim, P., & Francis, R. (2020, April). Analysis of Policy Factors Impacting the Use of Low-Cost Air Monitoring Networks in Washington, DC. In 2020 Systems and Information Engineering Design Symposium (SIEDS) (pp. 1-4). IEEE. doi: 10.1109/SIEDS49339.2020.9106694.

Swift, B. (2000). How environmental laws work: An analysis of the utility sector's response to regulation of nitrogen oxides and sulfur dioxide under the Clean Air Act. Tul. Envtl. L.J., 14, 309.

Taylor, D., & Jeffrey Davis, W. (1999). Review of basic research in bicycle traffic science, traffic operations, and facility design. Transportation Research Record, 1674(1), 102-110.

Turner, M.C., Andersen, Z.J., Baccarelli, A., Diver, W.R., Gapstur, S.M., Pope III, C.A., Prada, D., Samet, J., Thurston, G. and Cohen, A., (2020). Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. CA: a cancer journal for clinicians, 70(6), 460-479.

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. Applied Artificial Intelligence, 23(5), 373-405.

U.S. Department of Transportation Federal Highway Administration. (2022). Active Traffic Management: Approaches: Active Transportation and Demand Management - FHWA Operations. https://ops.fhwa.dot.gov/atdm/approaches/atm.htm

US EPA. (2022a). How Mobile Source Pollution Affects Your Health. Retrieved April 27, 2022, from https://www.epa.gov/mobile-source-pollution/how-mobile-source-pollution-affects-your-health

US EPA Download Files AirData. (2022b). Retrieved November 27, 2022, from https://aqs.epa.gov/aqsweb/airdata/download_files.html

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine Learning, 109(2), 373-440.

Valavanidis, A., Fiotakis, K., & Vlachogianni, T. (2008). Airborne particulate matter and human health: toxicological assessment and importance of size and composition of particles for oxidative damage and carcinogenic mechanisms. Journal of Environmental Science and Health, Part C, 26(4), 339-362.

Vallerio, M., Hufkens, J., Van Impe, J., & Logist, F. (2015). An interactive decision-support system for multi-objective optimization of nonlinear dynamic processes with uncertainty. Expert Systems with Applications, 42(21), 7710-7731.

Vasudevan, M., Wunderlich, K. W., Shah, V. P., & Larkin, J. (2004). Effectiveness Of Advanced Traveler Information Systems (Atis) Under Extreme Congestion: Findings From A Washington, D.C., Case Study. In At the Crossroads: Integrating Mobility Safety and Security. ITS America 2004, *14th Annual Meeting and ExpositionITS America.*

Wang, H., Chen, C., Huang, C., & Fu, L. (2008). On-road vehicle emission inventory and its uncertainty analysis for Shanghai, China. *Science of the total environment*, 398(1-3), 60-67. https://doi.org/10.1016/j.scitotenv.2008.01.038

Wang, M. H. S., Wang, L. K., & Shammas, N. K. (2020). Glossary of acid rain management and environmental protection. In Handbook of environment and waste management: Acid rain and greenhouse gas pollution control (pp. 719-749).

Wang, W. X., Guo, R. J., & Yu, J. (2018). Research on road traffic congestion index based on comprehensive parameters: Taking Dalian city as an example. *Advances in Mechanical Engineering*, 10(6). https://doi.org/10.1177/1687814018781482

Washington traffic report (2022). TomTom Traffic Index. https://www.tomtom.com/en_gb/traffic-index/washington-traffic/

Wilson, R. D., & Cales, B. (2008). Geographic information systems, evacuation planning and execution. Communications of the IIMA, 8(4), 2. https://scholarworks.lib.csusb.edu/ciima/vol8/iss4/2

Winchester, A. K., Peterson, R. A., Carter, E., & Sammel, M. D. (2021). Impact of COVID-19 Social Distancing Policies on Traffic Congestion, Mobility, and NO2 Pollution. Sustainability, 13(13), 7275.

World Health Organization. (2015). Global status report on road safety 2015. World Health Organization.

Yamazaki, K., Vo-Ho, V. K., Bulsara, D., & Le, N. (2022). Spiking neural networks and their applications: A Review. *Brain Sciences*, 12(7), 863.

Yang, H., Bell, M. G. H., & Meng, Q. (2000). Modeling the capacity and level of service of urban transportation networks. www.elsevier.com/locate/trb

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295-316.

Zhai, B., & Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average PM2. 5 concentrations in Beijing, China. Science of The Total Environment, 635, 644-658.

Zhou, J., Li, X., & Mitri, H. S. (2015). Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction. Natural Hazards, 79(1), 291-316.

Zhou, X., & Mahmassani, H. S. (2007). A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. Transportation Research Part B: Methodological, 41(8), 823–840. https://doi.org/10.1016/j.trb.2007.02.004

Appendix

Table A-1. Annual Summary Files Format (US EPA Download Files AirData., 2022b)

Field Position	Field Name	Description
1	State Code	The FIPS code of the state in which the monitor resides.
2	County Code	The FIPS code of the county in which the monitor resides.
3	Site Num	A unique number within the county identifying the site.
4	Parameter Code	The AQS code corresponding to the parameter measured by the monitor.
5	POC	This is the "Parameter Occurrence Code" used to distinguish different instruments that measure the same parameter at the same site.
6	Latitude	The monitoring site's angular distance north of the equator measured in decimal degrees.
7	Longitude	The monitoring site's angular distance east of the prime meridian measured in decimal degrees.
8	Datum	The Datum associated with the Latitude and Longitude measures.
9	Parameter Name	The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants.

10	**Sample Duration**	**The length of time that air passes through the monitoring device before it is analyzed (measured). So, it represents an averaging period in the atmosphere (for example, a 24-hour sample duration draws ambient air over a collection filter for 24 straight hours). For continuous monitors, it can represent an averaging time of many samples (for example, a 1-hour value may be the average of four one-minute samples collected during each quarter of the hour).**
11	**Pollutant Standard**	**A description of the ambient air quality standard rules used to aggregate statistics. (See description at beginning of document.)**
12	**Metric Used**	**The base metric used in the calculation of the aggregate statistics presented in the remainder of the row. For example, if this is Daily Maximum, then the value in the Mean column is the mean of the daily maximums.**
13	**Method Name**	**A short description of the processes, equipment, and protocols used in gathering and measuring the sample.**
14	**Year**	**The year the annual summary data represents.**
15	**Units of Measure**	**The unit of measure for the parameter. QAD always returns data in the standard units for the parameter. Submitters are allowed to report data in any unit and EPA converts to a standard unit so that we may use the data in calculations.**

16	**Event Type**	**Indicates whether data measured during exceptional events are included in the summary. A wildfire is an example of an exceptional event; it is something that affects air quality, but the local agency has no control over. No Events means no events occurred. Events Included means events occurred and the data from them is included in the summary. Events Excluded means that events occurred but data form them is excluded from the summary. Concurred Events Excluded means that events occurred but only EPA concurred exclusions are removed from the summary. If an event occurred for the parameter in question, the data will have multiple records for each monitor.**
17	**Observation Count**	**The number of observations (samples) taken during the year.**
18	**Observation Percent**	**The percent representing the number of observations taken with respect to the number scheduled to be taken during the year. This is only calculated for monitors where measurements are required (e.g., only certain parameters).**
19	**Completeness Indicator**	**An indication of whether the regulatory data completeness criteria for valid summary data have been met by the monitor for the year. Y means yes, N means no or that there are no regulatory completeness criteria for the parameter.**
20	**Valid Day Count**	**The number of days during the year where the daily monitoring criteria were met, if the calculation of the summaries is based on valid days.**

21	**Required Day Count**	**The number of days during the year which the monitor was scheduled to take samples if measurements are required.**
22	**Exceptional Data Count**	**The number of data points in the annual data set affected by exceptional air quality events (things outside the norm that affect air quality).**
23	**Null Data Count**	**The count of scheduled samples when no data was collected and the reason for no data was reported.**
24	**Primary Exceedance Count**	**The number of samples during the year that exceeded the primary air quality standard.**
25	**Secondary Exceedance Count**	**The number of samples during the year that exceeded the secondary air quality standard.**
26	**Certification Indicator**	**An indication whether the completeness and accuracy of the information on the annual summary record has been certified by the submitter. Certified means the submitter has certified the data (due May 01 the year after collection). Certification not required means that the parameter does not require certification or the deadline has not yet passed. Uncertified (past due) means that certification is required but is overdue. Requested but not yet concurred means the submitter has completed the process, but EPA has not yet acted to certify the data. Requested but denied means the submitter has completed the process, but EPA has denied the request for cause. Was certified but data changed**

		means the data was certified but data was replaced, and the process has not been repeated.
27	**Num Obs Below MDL**	**The number of samples reported during the year that were below the method detection limit (MDL) for the monitoring instrument. Sometimes these values are replaced by 1/2 the MDL in summary calculations.**
28	**Arithmetic Mean**	**The average (arithmetic mean) value for the year.**
29	**Arithmetic Standard Dev**	**The standard deviation about the mean of the values for the year.**
30	**1st Max Value**	**The highest value for the year.**
31	**1st Max DateTime**	**The date and time (on a 24-hour clock) when the highest value for the year (the previous field) was taken.**
32	**2nd Max Value**	**The second highest value for the year.**
33	**2nd Max DateTime**	**The date and time (on a 24-hour clock) when the second highest value for the year (the previous field) was taken.**
34	**3rd Max Value**	**The third highest value for the year.**
35	**3rd Max DateTime**	**The date and time (on a 24-hour clock) when the third highest value for the year (the previous field) was taken.**
36	**4th Max Value**	**The fourth highest value for the year.**

37	**4th Max DateTime**	**The date and time (on a 24-hour clock) when the fourth highest value for the year (the previous field) was taken.**
38	**1st Max Non Overlapping Value**	**For 8-hour CO averages, the highest value of the year.**
39	**1st NO Max DateTime**	**The date and time (on a 24-hour clock) when the first maximum non-overlapping value for the year (the previous field) was taken.**
40	**2nd Max Non Overlapping Value**	**For 8-hour CO averages, the second-highest value of the year does not share any hours with the 8-hour period of the first max non-overlapping value.**
41	**2nd NO Max DateTime**	**The date and time (on a 24-hour clock) when the second maximum non-overlapping value for the year (the previous field) was taken.**
42	**99th Percentile**	**The value from this monitor for which 99 percent of the rest of the measured values for the year are equal to or less than.**
43	**98th Percentile**	**The value from this monitor for which 98 percent of the rest of the measured values for the year are equal to or less than.**
44	**95th Percentile**	**The value from this monitor for which 95 percent of the rest of the measured values for the year are equal to or less than.**
45	**90th Percentile**	**The value from this monitor for which 90 percent of the rest of the measured values for the year are equal to or less than.**

46	**75th Percentile**	**The value from this monitor for which 75 percent of the rest of the measured values for the year are equal to or less than.**
47	**50th Percentile**	**The value from this monitor for which 50 percent of the rest of the measured values for the year are equal to or less than (i.e., the median).**
48	**10th Percentile**	**The value from this monitor for which 10 percent of the rest of the measured values for the year are equal to or less than.**
49	**Local Site Name**	**The name of the site (if any) given by the State, local, or tribal air pollution control agency that operates it.**
50	**Address**	**The approximate street address of the monitoring site.**
51	**State Name**	**The name of the state where the monitoring site is located.**
52	**County Name**	**The name of the county where the monitoring site is located.**
53	**City Name**	**The name of the city where the monitoring site is located. This represents the legal incorporated boundaries of cities and not urban areas.**
54	**CBSA Name**	**The name of the core bases statistical area (metropolitan area) where the monitoring site is located.**
55	**Date of Last Change**	**The date the last time any numeric values in this record were updated in the AQS data system.**

Table A-2 Daily Summary Files Format (US EPA Download Files AirData, 2022)

Field Position	Field Name	Description
1	State Code	The FIPS code of the state in which the monitor resides.
2	County Code	The FIPS code of the county in which the monitor resides.
3	Site Num	A unique number within the county identifying the site.
4	Parameter Code	The AQS code corresponding to the parameter measured by the monitor.
5	POC	This is the "Parameter Occurrence Code" used to distinguish different instruments that measure the same parameter at the same site.
6	Latitude	The monitoring site's angular distance north of the equator measured in decimal degrees.
7	Longitude	The monitoring site's angular distance east of the prime meridian measured in decimal degrees.
8	Datum	The Datum associated with the Latitude and Longitude measures.
9	Parameter Name	The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants.
10	Sample Duration	The length of time that air passes through the monitoring device before it is analyzed (measured). So, it represents an averaging period in the atmosphere (for example, a 24-hour sample duration draws ambient air over a collection filter for 24 straight

		hours). For continuous monitors, it can represent an averaging time of many samples (for example, a 1-hour value may be the average of four one-minute samples collected during each quarter of the hour).
11	**Pollutant Standard**	**A description of the ambient air quality standard rules used to aggregate statistics. (See description at beginning of document.)**
12	**Date Local**	**The calendar date for the summary. All daily summaries are for the local standard day (midnight to midnight) at the monitor.**
13	**Units of Measure**	**The unit of measure for the parameter. QAD always returns data in the standard units for the parameter. Submitters are allowed to report data in any unit and EPA converts to a standard unit so that we may use the data in calculations.**
14	**Event Type**	**Indicates whether data measured during exceptional events are included in the summary. A wildfire is an example of an exceptional event; it is something that affects air quality, but the local agency has no control over. No Events means no events occurred. Events Included means events occurred and the data from them is included in the summary. Events Excluded means that events occurred but data form them is excluded from the summary. Concurred Events Excluded means that events occurred but only EPA concurred exclusions are removed from the summary. If an event occurred for the parameter in question, the data will have multiple records for each monitor.**

15	**Observation Count**	**The number of observations (samples) taken during the day.**
16	**Observation Percent**	**The percent representing the number of observations taken with respect to the number scheduled to be taken during the day. This is only calculated for monitors where measurements are required (e.g., only certain parameters).**
17	**Arithmetic Mean**	**The average (arithmetic mean) value for the day.**
18	**1st Max Value**	**The highest value for the day.**
19	**1st Max Hour**	**The hour (on a 24-hour clock) when the highest value for the day (the previous field) was taken.**
20	**AQI**	**The Air Quality Index for the day for the pollutant, if applicable.**
21	**Method Code**	**An internal system code indicating the method (processes, equipment, and protocols) used in gathering and measuring the sample. The method name is in the next column.**
22	**Method Name**	**A short description of the processes, equipment, and protocols used in gathering and measuring the sample.**
23	**Local Site Name**	**The name of the site (if any) given by the State, local, or tribal air pollution control agency that operates it.**
24	**Address**	**The approximate street address of the monitoring site.**
25	**State Name**	**The name of the state where the monitoring site is located.**

26	**County Name**	**The name of the county where the monitoring site is located.**
27	**City Name**	**The name of the city where the monitoring site is located. This represents the legal incorporated boundaries of cities and not urban areas.**
28	**CBSA Name**	**The name of the core bases statistical area (metropolitan area) where the monitoring site is located.**
29	**Date of Last Change**	**The date the last time any numeric values in this record were updated in the AQS data system.**

	A	B	C	D	E	F	G	H
1	SO2	NO2	TEMP	WD	RH	Volume	Day	Month
2	2.1	10.5	16.5	334	51.4	2709	Monday	Jan
3	2.9	11	16	335	53.2	2883	Monday	Jan
4	3	12.2	15.6	334	54.1	2465	Monday	Jan
5	3	13.9	15.4	336	55.6	2269	Monday	Jan
6	2.9	16.1	14.9	340	58	1541	Monday	Jan
7	2.7	17	14.5	340	58.3	1132	Monday	Jan
8	2.4	18.4	14.2	313	57.8	1136	Monday	Jan
9	2.4	15.6	14	261	57.4	1179	Monday	Jan
10	2.5	12.9	15.1	305	55.9	1370	Monday	Jan

```
@relation NO2-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0

@attribute SO2 numeric
@attribute VOLUME numeric
@attribute TEMP numeric
@attribute WD numeric
@attribute RH numeric
@attribute Day {Sunday,Monday,Tuesday,Wednesday,Thursday,Friday,Saturday}
@attribute Month {January,February,March,April,May,June,July,August,September,October,November,December}
@attribute NO2 numeric

@data
0.051724,0.583917,0.37831,0.619883,0.280457,Sunday,January,10
0.045977,0.566018,0.366961,0.602339,0.30203,Sunday,January,13.9
0.04023,0.584955,0.358134,0.625731,0.321066,Sunday,January,16.8
0.04023,0.50013,0.349306,0.532164,0.340102,Sunday,January,16.3
0.034483,0.236835,0.330391,0.239766,0.38198,Sunday,January,20.9
0.057471,0.356161,0.330391,0.374269,0.389594,Sunday,January,31.1
0.04023,0.511284,0.340479,0.546784,0.385787,Sunday,January,26
0.045977,0.561089,0.355612,0.599415,0.374365,Sunday,January,22.9
0.045977,0.73904,0.375788,0.795322,0.390863,Sunday,January,22.8
0.045977,0.685084,0.403531,0.72807,0.409898,Sunday,January,13.2
0.051724,0.633204,0.428752,0.663743,0.378173,Sunday,January,6.7
0.051724,0.657069,0.443884,0.687135,0.360406,Sunday,January,4.6
0.04023,0.713878,0.455233,0.748538,0.350254,Sunday,January,4.6
0.051724,0.724514,0.469105,0.75731,0.340102,Sunday,January,4.2
0.051724,0.631128,0.469105,0.652047,0.347716,Sunday,January,4.3
0.051724,0.198703,0.460277,0.166667,0.36802,Sunday,January,6.8
```

Figure A-1. Input file for the 2017 Training Year Excel and WEKA format

Table B-1. Traffic Volume Stations used in 2017 and 2018 Location and Distance from Monitoring Stations

Station	ROADWAY NAME	COUNTY	Direction of Traffic	Site Latitude	Site Longitude	MM	RT	TP	AF
1203	I-296N north of Chesapeake St SW	District of Columbia	NB and SB	38.833	-77.013	6.11	5.22	9.44	5.32
1403	INTERSTATE 695 – 11th Street Bridge	District of Columbia	NB and SB	38.872	-76.990	3.68	2.35	6.94	2.50
2204A	SOUTH CAPITOL ST between I ST SW and K ST SW	District of Columbia	NB and SB	38.873	-77.009	3.37	3.15	6.71	3.35
6003	INDEPENDENCE AVE SW between 14TH ST SW and RAOUL WALLENBERG PL SW	District of Columbia	EB and WB	38.888	-77.033	2.58	4.06	5.75	4.30
6002	EAST CAPITOL ST between C ST /INDEPENDENCE AVE and Ramps to Anacostia Freeway	District of Columbia	EB and WB	38.890	-76.965	3.42	0.54	6.21	0.70
3606	13TH ST NW between I ST NW and K ST NW	District of Columbia	NB and SB	38.902	-77.030	1.63	3.87	4.76	4.13
1503	I-395, South of New York Avenue	District of Columbia	NB and SB	38.904	-77.015	1.27	3.12	4.60	3.38
2401	Kenilworth Avenue at DC Line	District of Columbia	NB and SB	38.911	-76.935	4.26	1.62	5.99	1.48
3808	MASSACHUSETTS AVE NW between WATERSIDE DR NW and ROCK CREEK & POTOMAC PKY NW	District of Columbia	NB and SB	38.918	-77.059	2.46	5.63	4.24	5.89
2502	NEW YORK AVE NE between 36TH PL NE and 36TH PL NE	District of Columbia	EB and WB	38.919	-76.948	3.50	1.69	5.11	1.68
3707	WISCONSIN AVE NW between Jenifer ST NW and Ingomar ST NW	District of Columbia	NB and SB	38.959	-77.085	4.61	8.08	3.73	8.32
7032	NEW HAMPSHIRE AVE NE AT B&O Railroad Overpass	District of Columbia	NB and SB	38.961	-77.006	2.74	5.21	0.85	5.38
7041	16TH ST NW between PORTAL DR NW and NORTH PORTAL DR NW	District of Columbia	NB and SB	38.991	-77.036	4.96	7.84	1.81	8.02

Note: MM is McMillion Resivor, RT is River Terrace, TP is Takoma Park, AF is Anacostia Freeway

Table B-2. Traffic Volume Stations used in 2017 and 2018 Location and Distance from Monitoring Stations

MLR	Mean_absolute error	Root_mean squared_error	Relative absolute_error	Root_relative squared_error	Correlation coefficient	R-squared
8	4.38	5.92	0.67	0.76	74%	55%
9	4.49	6.01	0.64	0.79	77%	59%
10	4.48	6.25	0.61	0.79	79%	63%
Training	5.00	6.36	0.66	0.72	77%	59%
Test	5.63	7.37	0.72	0.79	75%	56%
ANN	Mean_absolute_error	Root_mean_squared_error	Relative_absolute_error	Root_relative_squared_error	Correlation_coefficient	R-squared
8	5.72	7.63	0.69	0.79	77%	59%
9	5.71	7.33	0.73	0.80	80%	64%
10	5.48	6.96	0.67	0.73	80%	63%
Training	4.72	6.21	0.60	0.66	82%	67%
test	3.10	3.65	0.23	0.26	73%	53%
RF	Mean_absolute_error	Root_mean_squared_error	Relative_absolute_error	Root_relative_squared_error	Correlation_coefficient	R-squared
8	3.87	5.53	0.47	0.58	81%	66%
9	3.91	5.14	0.49	0.55	84%	70%
10	3.82	5.24	0.47	0.54	85%	72%
Training	1.43	2.02	0.18	0.22	98%	97%
Test	3.25	4.58	0.43	0.52	86%	74%