

Texas Southern University

Digital Scholarship @ Texas Southern University

Dissertations (2016-Present)

Dissertations

12-2021

The Temporal and Frequent Pattern Mining Analysis and Machine Learning Forecasting on Mobile Sourced Urban Air Pollutants

Jianbang Du

Follow this and additional works at: <https://digitalscholarship.tsu.edu/dissertations>

Recommended Citation

Du, Jianbang, "The Temporal and Frequent Pattern Mining Analysis and Machine Learning Forecasting on Mobile Sourced Urban Air Pollutants" (2021). *Dissertations (2016-Present)*. 28.
<https://digitalscholarship.tsu.edu/dissertations/28>

This Dissertation is brought to you for free and open access by the Dissertations at Digital Scholarship @ Texas Southern University. It has been accepted for inclusion in Dissertations (2016-Present) by an authorized administrator of Digital Scholarship @ Texas Southern University. For more information, please contact haiying.li@tsu.edu.

THE TEMPORAL AND FREQUENT PATTERN MINING ANALYSIS
AND MACHINE LEARNING FORECASTING ON MOBILE SOURCED
URBAN AIR POLLUTANTS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate School
of Texas Southern University

By

Jianbang Du, B.S., M.S.

Texas Southern University

2021

Approved By

Fengxiang Qiao, Ph.D.
Chairperson, Dissertation Committee

Gregory H. Maddox, Ph.D.
Dean, The Graduate School

Approved By

Fengxiang Qiao, Ph.D.
Chairperson, Dissertation Committee

08/27/2021
Date

Hyun-Min Hwang, Ph.D.
Committee Member

08/27/2021
Date

Qing Li, Ph.D.
Committee Member

08/27/2021
Date

Lei Yu, Ph.D.
Graduate School Representative

08/27/2021
Date

© Copyright by Jianbang Du 2021

All Rights Reserved

THE TEMPORAL AND FREQUENT PATTERN MINING ANALYSIS
AND MACHINE LEARNING FORECASTING ON MOBILE SOURCED
URBAN AIR POLLUTANTS

By

Jianbang Du, Ph.D.

Texas Southern University, 2021

Professor Fengxiang Qiao, Advisor

Ground-level ozone and atmospheric fine particles ($PM_{2.5}$) have been recognized as critical air pollutants that act as important contributors to the toxicity of anthropogenic air pollution in urban areas. To limit the adverse impacts on public health and ecosystems of ground-level ozone and $PM_{2.5}$, it is necessary and imperative to identify a practical and effective way to predict the upcoming pollution concentration levels accurately. Under this need, various research was conducted aiming to perform the forecasting of ground-level ozone and $PM_{2.5}$ that mainly utilized the time-series and neural network analysis. In the meantime, machine learning is also adopted in analysis and forecasting in existing research, which is, however, associated with some limitations that are not easily overcome. (1) The majority of existing forecasting models are highly dependent on time-series inputs without considering the influencing factors of the air pollutants. While a relatively accurate prediction may be provided, the influencing factors of the air pollution level caused by real-world complexity are neglected. (2) The existing forecasting models are mainly focused on the short-term estimation, while some of them need to use the previous

prediction as a part of the input, which increased the system complexity and decreased the computational efficiency and accuracy. (3) The accurate annual hourly air pollution level forecasting ability is seldomly achieved. The objective of this research is to propose a systematical methodology to forecast the long-term hourly future air pollution concentration levels through historical data considering the concentration influencing factors. To achieve this research goal, a series of methodologies to analyze the historical air pollution concentration by temporal characteristics and frequent pattern data mining algorithms are introduced. The association rules of air pollution concentration levels and the influencing factors are revealed. A systematical air pollution level forecasting approach based on supervised machine learning algorithms with the ability to predict the annual hourly value is proposed and evaluated. To quantify and validate the results, a case study was conducted in the Houston region with the collection and analysis of ten years of historical environmental, meteorological, and transportation-related data. From the results of this research, (1) the complex correlations between the influencing factors and air pollution concentration levels are quantified and presented. (2) The association rules between each dependant and independent parameters are calculated. (3) The supervised machine learning algorithm pool is created and evaluated. And (4), an accurate long-term hourly air pollution level machine learning forecasting procedure is proposed. The innovative methodology of this research is advanced in computation complexity with high accuracy when compared with the existing models, which could be easily applied to similar regions for various types of air pollution concentration level forecasting.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vii
LIST OF FIGURES	ix
VITA.....	x
ACKNOWLEDGEMENTS	xi
CHAPTER 1 INTRODUCTION	1
1.1 Overview	1
1.2 Motivation	2
1.3 Research Gaps and Objectives	4
CHAPTER 2 LITERATURE REVIEW	6
2.1 Overview of Pollution	6
2.2 Air Pollution and Pollutants	9
2.2.1 Ground-level Ozone	10
2.1.2 PM _{2.5}	12
2.1.3 Nitrogen Dioxide/ Oxides of Nitrogen	16
2.3 Influencing Factors of Air pollution.....	17
2.3.1 Meteorological Measurements	17
2.3.2 Transportation.....	21
2.4 Air Pollution Forecasting Technologies.....	23
CHAPTER 3 DESIGN OF THE STUDY	27
3.1 Data Collection.....	29
3.1.1 Air Pollution Concentration Data Collection	29

3.1.2 Meteorological Measurements Data Collection	32
3.1.3 Traffic Situation Data Collection	33
3.2 Raw Data Preprocessing and Statistical Analysis.....	34
3.2.1 Data Preprocessing.....	34
3.2.2 Statistical Analysis	36
3.3 Frequent Pattern Mining.....	37
3.3.1 Input Datasets Integrating and Data Fitting.....	38
3.3.2 Understanding Frequent Pattern Mining	39
3.3.3 Frequent Pattern Mining Algorithms.....	41
3.4 Machine Learning and Forecasting	45
3.4.1 Machine Learning Forecasting Concepts	46
3.4.2 Machine Learning Models.....	48
3.4.2.1 Linear Regression Algorithm.....	48
3.4.2.2 Polynomial Regression Algorithm.....	50
3.4.2.3 Multilayer Perceptron Algorithm.....	51
3.4.2.4 XGBoost Algorithm.....	54
3.4.2.5 Support Vector Machine Algorithm	57
3.4.2.6 Random Forest Algorithm	59
3.4.2.7 K-Nearest Neighbors Algorithm.....	61
3.4.3 Machine Learning Model Evaluation.....	62
CHAPTER 4 RESULTS AND DISCUSSION.....	66
4.1 Temporal Characteristics.....	66

4.1.1 Annual Characteristics of Air Pollution and Meteorological Measurements...	67
4.1.2 Monthly Characteristics of Air Pollution and Meteorological Measurements.	70
4.1.3 Day-of-Week Characteristics of Air Pollution	73
4.1.4 Hourly Characteristics of Air Pollution and Meteorological Measurements ...	74
4.1.5 Temporal Characteristics of Traffic Speed.....	77
4.2 Correlation Analysis.....	79
4.3 Frequent Pattern Mining Analysis.....	86
4.3.1 Data Preprocessing and Binning.....	86
4.3.2 Ground-level Ozone Frequent Pattern Mining.....	89
4.3.3 PM _{2.5} Frequent Pattern Mining	92
4.3.4 NO ₂ Frequent Pattern Mining.....	95
4.3.5 NO _x Frequent Pattern Mining.....	98
4.4 Machine Learning Prediction Models for Air Pollution	103
4.4.1 Machine Learning Model Selection and Training.....	103
4.4.1.1 Model Selection for Ground-level Ozone.....	103
4.4.1.2 Model Selection for PM _{2.5}	105
4.4.1.3 Model Selection for NO ₂	107
4.4.1.4 Model Selection for NO _x	109
4.4.2 Air Pollution Forecasting Based on the Models Selected	112
4.4.2.1 Ground-level Ozone Concentration Prediction.....	113
4.4.2.2 PM _{2.5} Concentration Prediction.....	116
4.4.2.3 NO ₂ Concentration Prediction	119

4.4.2.4 NO _x Concentration Prediction.....	122
4.5 Discussion	125
CHAPTER 5 SUMMARY AND RECOMMENDATIONS	129
5.1 Summary	129
5.2 Contributions.....	133
5.3 Recommendations.....	134
APPENDIX.....	136
A. MACHINE LEARNING MODEL SELECTION FOR GROUND-LEVEL OZONE	136
B. MACHINE LEARNING MODEL SELECTION FOR PM _{2.5}	138
C. MACHINE LEARNING MODEL SELECTION FOR NO ₂	140
D. MACHINE LEARNING MODEL SELECTION FOR NO _x	142
REFERENCES	144

LIST OF TABLES

Table	Page
1. STUDIES ON THE ADVERSE EFFECTS OF PM _{2.5}	14
2. PM _{2.5} REGULATIONS	15
3. AIR POLLUTANTS COLLECTED FROM TCEQ DATABASE	31
4. METEOROLOGICAL MEASUREMENTS COLLECTED FROM TCEQ DATABASE.....	32
5. RESULTANT WIND DIRECTIONS SHOWN ON COMPASS	33
6. INVALID MEASUREMENTS OF RAW DATA.....	34
7. EXAMPLE OF A RECORD IN THE INTEGRATED DATASET.....	38
8. EXAMPLE ENTRIES IN THE TRANSFORMED DATASET	39
9. CONFUSION MATRIX FOR CLASSIFICATION MODELS	62
10. ONEWAY ANOVA TEST RESULTS THROUGH TEN YEARS.....	66
11. BINS AND LABELS FOR PARAMETERS	87
12. FREQUENT PATTERNS WITH THE HIGHEST SUPPORT FOR GROUND- LEVEL OZONE	90
13. FREQUENT PATTERNS WITH HIGHEST SUPPORT FOR PM _{2.5}	93
14. FREQUENT PATTERNS WITH HIGHEST SUPPORT FOR NO ₂	97
15. FREQUENT PATTERNS WITH HIGHEST SUPPORT FOR NO _x	100
16. MODEL SELECTED FOR GROUND-LEVEL OZONE.....	103
17. MODEL SELECTED FOR PM _{2.5}	106
18. MODEL SELECTED FOR NO ₂	108

19. MODEL SELECTED FOR NO _x	110
20. THE MODELS SELECTED FOR EACH AIR POLLUTANT	111
21. PREDICTION OF GROUND-LEVEL OZONE CONCENTRATION.....	113
22. PREDICTION OF PM _{2.5} CONCENTRATION.....	116
23. PREDICTION OF NO ₂ CONCENTRATION.....	120
24. PREDICTION OF NO _x CONCENTRATION.....	123

LIST OF FIGURES

Figure	Page
1. Flow Chart of This Research	28
2. Air Monitoring Sites in Houston-Galveston Area and Site 403	31
3. Flow Chart of The Apriori Algorithm (Han, Pei, & Kamber, 2011)	41
4. Flow Chart of The FP-Growth Algorithm (H. Li, Wang, Zhang, Zhang, & Chang, 2008).....	43
5. Forecasting by Machine Learning	47
6. Layers and Structures of MLP	52
7. Annual Characteristics	68
8. Monthly Characteristics	71
9. Day of Week Characteristics	73
10. Hourly Characteristics	75
11. Annual Average Traffic Speed Profile	77
12. Interrelationship Matrix and Distribution Chart for All Parameters.....	80
13. Pearson's r Correlation Test Matrix.....	83
14. Distributions of Influencing Factors of Ground-level ozone	90
15. Distributions of Influencing Factors of PM _{2.5}	93
16. Distributions of Influencing Factors of NO ₂	96
17. Distributions of Influencing Factors of NO _x	99

VITA

- 2006 - 2010..... B.S., Information and Computation
Science, Changchun University of
Science and Technology, Changchun,
China
- 2011 - 2013..... M. Eng., Industrial Engineering, Texas
A&M University, College Station, Texas
- 2013 - 2016..... Handset Engineer, Brook Consultant
Inc., Dallas, Texas
- 2016 – 2018..... M.S., Environmental Toxicology, Texas
Southern University, Houston, Texas
- 2018- Present..... Ph.D. Candidate, Environmental
Toxicology, Texas Southern University,
Houston, Texas

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my dissertation advisor Dr. Fengxiang Qiao who offered me this great opportunity of researching under his supervision. Dr. Qiao was very kind to provide his constant support, advice, and guidance from the beginning throughout all the odds that I faced. He consistently steered me in the right direction whenever he thought I needed it.

Besides, I thank the rest of my committee members: Dr. Lei Yu, Dr. Hyun-Min Hwang, and Dr. Qing Li for their encouragement and contributions. I would also like to thank my fellow labmates and friends in Houston for the many ways that, they were always beside me to help and inspire me during my study at TSU.

Finally, I must express my profound gratitude to my father Zhiping Du, my mother Dong Mu, my wife Dao Xiang, and my son Leon for their selfless support and love throughout my life. This accomplishment would not be possible without their incredible efforts and sacrifices. I'm grateful to them for shaping my beautiful world the way it is.

CHAPTER 1

INTRODUCTION

1.1 Overview

With the rapid growth of the global population and technologies that depend on fossil fuels and petrochemicals, various critical environmental dilemmas have emerged. One of the most harmful environmental issues that human being has encountered is air pollution, which is a significant inducing factor of a series of life-shortening diseases such as Cardio Vascular Disease, Stroke, Chronic Obstruction Pulmonary Disease, Lung Cancer, and even death (Zumla et al., 2015). The World Health Organization (WHO) indicated that 2.4 million human death per year is attributed to air pollution-related diseases (Xing, Xu, Shi, & Lian, 2016). Furthermore, air pollution also plays an important role in adversely affecting the ecosystems, as well as flora and fauna species.

Many chemicals have been identified as air pollutants, which include the six criteria pollutants (carbon monoxide (CO), lead (Pb), nitrogen oxides (NO_x), ground-level ozone (O₃), particulate matter (PM), and sulfur oxides (SO_x)) defined by the United States Environmental Protection Agency (US. EPA) and other air pollutants such as asbestos and total petroleum hydrocarbons (CDC, 2021). Some of these air pollutants are emitted into the atmosphere directly, which are called primary pollutants, such as CO, Pb, NO_x, and SO_x. On the contrary, the ground-level ozone is formed in the atmosphere from other chemicals include NO_x and volatile organic compounds (VOC) rather than emitted from the sources directly, which is called the secondary air pollutant (Du, Li, Qiao, & Yu, 2018; Shao et al., 2009). PMs in the air can be either the primary PM from wind transport,

combustion, and human activities, or the secondary PM by a series of complex chemical reactions by, for example, SO₂, NO_x, ammonia (NH₃), and VOC (Breysse et al., 2013).

Some of the air pollutants can be naturally sourced. For instance, ozone can be formed naturally in the upper atmosphere that protects the earth from the harmful ultraviolet rays from the sun. The PM can also be naturally formed by volcanic emissions, dust storms, forest wildfires, and ocean salt spray (Omidvarborna, Kumar, & Kim, 2015). However, the majority of these air pollutants are anthropogenic, in another word, human-sourced. For example, fossil fuel combustions such as vehicle exhausts, power plants emitters, and oil refineries are the primary sources of ground-level ozone precursor chemicals (Cardelino & Chameides, 1995). Similarly, anthropogenic PMs that are sourced from human activities such as coal and petroleum production burning, construction sites, and unpaved roads account for around ten percent of the total atmospheric PM mass (Hardin & Kahn, 1999).

1.2 Motivation

Among various types of air pollutants, ground-level ozone and PM, especially PM_{2.5} that has an aerodynamic diameter smaller than 2.5 micrometers, are the most critical and hazardous (ATSDR, 2021). A high concentration of ground-level ozone is known to have severe adverse health effects on the high-risk population by inducing asthma, respiratory system irritation, lung function decline, and lung lining damages (Lippmann, 1989). It is reported that ninety US urban communities, which count for approximately 40% of the US total population, are suffering from ground-level ozone pollution hazards

(Bell, McDermott, Zeger, Samet, & Dominici, 2004). Ground-level ozone pollution can also induce economic loss by ozone cracking (Lake, 1970), and several ecosystem consequences including greenhouse effects (Shindell, Rind, & Loneragan, 1998) as ground-level ozone is a greenhouse gas, of which the radiative forcing effect is 1,000 times stronger than that of carbon dioxide (CO₂) (Curran, 2012). Because ozone is a strong oxidant, it is able to influence the growth rate and seed production of plants (Manes et al., 2012). Previous studies revealed that several flora species including kinds of crops are particularly sensitive to ground-level ozone (Reich, 1987). Given so many hazards are triggered, strategies to decrease the ground-level ozone concentration have been established by different authorities (Ryerson et al., 2001).

In the meantime, due to its extremely small in size, PM_{2.5} is able to penetrate deep into the lungs and even access the bloodstream, and further impair lung function (Xing et al., 2016). Epidemiological studies have been extensively conducted to provide scientific evidence of PM_{2.5}'s public health risks and revealed the relationship between ambient PM_{2.5} concentration and cardiopulmonary mortality (EPA, 2010; Pope III, 2000; SCHWARTZ, 1991, 2004). What's more, PM_{2.5} can be transported for long periods in an airborne manner and travel hundreds of miles. To regulate the harmful effects, the WHO has firstly published the Air Quality Guidelines (AQG) in 1987, which recommends thresholds of 25 µg/m³ average daily concentration (ADC) and 10 µg/m³ average annual concentration (AAC) for PM_{2.5} (Pope III et al., 2002; WHO, 2006). The highly toxic nature of ground-level ozone and PM_{2.5} is threatening the environment and public health. Thus, the forecasting of these air pollution concentration levels in a predetermined future period is in urgent need.

1.3 Research Gaps and Objectives

A large number of studies have been conducted focusing on the air pollution concentration variation pattern in order to forecast the future concentration level. Most of the studies utilized the classic time-series approach to analyze the historical air pollution concentration data and make the forecasting. This approach can be accurate, which, however, lacks the ability to predict the air pollution level for special events, such as the pandemic throughout the year 2020 that influenced the transportation activities and the vehicles' emission along with the altered traffic pattern. Special weather events such as hurricanes and flooding can also reduce the accuracy of the prediction based on the time-series approach. Thus, the time-series approach is usually solely accurate in several special cases. Some recent research employed neural network algorithms to perform air pollution level forecasting, which may yield a more accurate prediction and relatively higher computation efficiency. However, there are still issues when using neural network algorithms. (1) Most existing neural network research is highly dependent on the time-series inputs. (2) The complexity of the algorithm requires more tuning of the parameters of the model. Machine learning is a state-of-the-art analytical tool that is utilized to forecast the air pollution concentration level in recent years. The existing research is limited and mainly has the following limitations: (1) the inputs for the models are solely air pollution data, which are still time-series data inherently, (2) most research does not employ the influencing factors of air pollutants, which restricts the accuracy and robustness of the models, and (3) the existing machine learning forecasting models were mainly designed to perform a short period air pollution forecasting.

To fill these research gaps, the objectives of this research are: (1) to analyze the temporal characteristics of ground-level ozone and its precursor NO_x/NO_2 and $\text{PM}_{2.5}$ concentration patterns within the target coastal industrial urban area; (2) to find the inherent pattern between air pollution concentration levels and their influencing factors that include several meteorological parameters and traffic situations through data mining technology; and (3) to perform the hourly air pollution concentration forecasting over a year by different supervised machine learning algorithms with the suitable assessment of accuracy.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Pollution

Environmental pollution has been widely recognized as the substance that adversely affects the global climate, flora and fauna species, and public health of human beings (Haines, Kovats, Campbell-Lendrum, & Corvalán, 2006). There are several types and forms of pollution, which mainly include water pollution, soil pollution, and air pollution (Brunekreef & Holgate, 2002). Other types of pollution have also been recognized such as light, noise, thermal, electromagnetic and radiation pollutions (Stansfeld & Matheson, 2003).

Among these pollutions, water pollution that includes surface water and groundwater pollution is the contamination of water bodies such as lakes, rivers, oceans, aquifers, and groundwater by mainly human activities, which is one of the leading causes of some human diseases, especially water-borne diseases (Moe & Rheingans, 2006). Organic and inorganic chemicals and pathogens may occur in polluted water (Parry, 1998). There are two kinds of sources of water pollution, which are point sources such as plant and refinery discharge, and non-point sources such as boating and atmospheric fallout (Moss, 2008). The primary federal law in the USA to regulate water pollution is the Clean Water Act that is implemented by the EPA. Soil pollution is also called soil contamination that alters the natural environment of soil by industrial, agricultural, improper waste disposal, and other human activities (Means, 1989). The main chemicals that appear in soil contamination include heavy metals, hydrocarbons, pesticides, herbicides, and solvents (Cunningham, Berti, & Huang, 1995), which may show the health effects such as pollution-

related diseases include cancer and congenital disorders, and ecosystem effects (Merry, Tiller, & Alston, 1986). Light pollution that may influence animal behavior and ecology usually appear during the night by artificial lights, which is one of the major side effects of urbanization (Longcore & Rich, 2004). Unlike the water pollution that is severer in developing countries, light pollution in developed countries is generally severer. The sources of light pollution include but are not limited to the direct glare, sky glow that is the light reflected from the sky, buildings and street lights, etc. (Longcore & Rich, 2004). Based on previous research, approximately only 40% of Americans live in areas that the night is dark sufficiently, and 18.7% of the earth's land surface is polluted by artificial light (Cinzano, Falchi, & Elvidge, 2001). Current practices to reduce light pollution are to control the light from the sources by technologies (Falchi, Cinzano, Elvidge, Keith, & Haim, 2011). Noise pollution that is also called sound pollution is mainly coming from transportation and industry activities, which may be disturbing and influence social behaviors, human development, and mental health such as hypertension (Stansfeld & Matheson, 2003). Noise pollution is also harmful to the ecosystem, especially the invertebrates that use antennae or hairs on particle motion detection (Nedelec, Campbell, Radford, Simpson, & Merchant, 2016). Furthermore, research indicated that noise pollution is closely related to human's increased catecholamine secretion and possibly high blood pressure in children (Stansfeld & Matheson, 2003). In the USA, the noise pollution in low-income and racial minority communities is relatively higher (Casey et al., 2017). To solve this issue, the Noise Control Act was established in 1972 by EPA, and the Recommended Exposure Limit (REL) was published by the National Institute for Occupational Safety and Health (NIOSH) at the Centers for Disease Control and

Prevention (CDC) (Kirchner et al., 2012). Thermal pollution is a type of water pollution that changes water temperature by the coolant water use of manufacturers and power plants (Davidson & Bradshaw, 1967). In the USA, 75 to 82% of thermal pollution is sourced from power plants (Laws, 2017). Thermal pollution may impact aquatic organisms such as fish and amphibians by decreasing the level of dissolved oxygen and increasing the organisms' metabolic rates (Goel, 2006). More severe consequences induced by thermal pollution appear including global warming (Nordell, 2003). Electromagnetic and radiation is a type of pollution that is developed with the spread of some of the technologies such as microwave and RF technologies (Dhami, 2012). The source of electromagnetic and radiation pollution include electric power, electronic surveillance system, and wireless communications (Ahlbom & Feychting, 2003). The electromagnetic and radiation pollution may induce acute symptoms such as burns, sleep disturbance, depression, headache, nausea, visual disorders, respiratory problems, nervousness, and agitation (Santini, Santini, Danze, Le Ruz, & Seigne, 2002), and other chronic symptoms such as the possible cardiovascular disease, brain tumors, leukemia, and breast cancer (Caplan, Schoenfeld, O'Leary, & Leske, 2000; Sastre, Cook, & Graham, 1998). Due to the potential hazards of electromagnetic and radiation pollution, it is categorized as Group 2B by the WHO and the International Agency for Research on Cancer (IARC) (Baan et al., 2011). In the USA, the Occupational Safety and Health Act that was published in 1970 and the Radiation Control for Health and Safety Act that was published in 1968 are the regulations of the nonionizing radiation.

Environmental pollution has become a global problem and costs a lot. A report by the Lancet Commission on Pollution and Health revealed that in 2015, nine million death

globally are caused by air, water, and soil pollution, which counts for 16% of all death. This number is more than the death caused by smoking, hunger, natural disasters, war, malaria, or AIDS, 97% happens in the less developed countries (Landrigan et al., 2018). As stated in the report, 6.2% of global economic output is lost because of pollution, which is \$4.6 trillion per year. In the meantime, the benefits of pollution control significantly outweighed the pollution costs. The cost and benefit ratio of air pollution control reaches 1:30 in the USA. \$65 billion were invested by the USA to air pollution control and received around \$1.5 trillion benefits. In another word, every dollar spent on air pollution control will receive 30 dollars benefit (Landrigan et al., 2018).

Among these hazardous pollutions, air pollution is one of the most critical issues that adversely influence the environment and human health and caused large costs worldwide every year.

2.2 Air Pollution and Pollutants

Air pollution is the harmful substances that are present in the air in the forms of liquid droplets, solid and gaseous states, which may induce diseases, allergy, and death of humankind, sickness of animals, crops, and livestock, and further threaten the ecosystem (Organization, 2014). For human beings, air pollution may induce respiratory diseases, heart diseases, chronic obstructive pulmonary disease (COPD), stroke, or even cancer (Sunyer et al., 2015). The World Health Organization (WHO) pointed out that 2.4 million human mortality is caused by air pollution per year. Furthermore, based on various research, air pollution is related to around 9% of the total death worldwide (Page, 2019),

and 100 thousand in the US each year (Neuhauser, 2019). Based on the property of the air pollutants, they can be categorized as gases, particulates, and biological molecules (Stern, 1977). Based on the type of pollution sources, air pollutants can be categorized as stationary sourced pollutants that are emitted from settled sources such as power plants, manufacturers, and oil refineries, and mobile sourced pollutants that are emitted from moving sources such as vehicles, ships, and air crafts (Colvile, Hutchinson, Mindell, & Warren, 2001). Based on the chemical formation type of air pollutions, they can be categorized as the primary air pollutants that are emitted from the sources directly, and the secondary air pollutants that are formatted in the atmosphere by the emitted chemicals from the sources (Stern, 1977).

Air pollution sourced from human activities also called anthropogenic air pollution is considered the most critical environmental issue nowadays (Vitousek, Mooney, Lubchenco, & Melillo, 1997). Mobile sourced air pollution especially on-road vehicle emission occupies greater than 50% of total air pollution in the USA (EPA, 2019) due to fossil fuel combustion, which counts for 50% to 90% of total urban air pollutions (Shinar, 2017).

2.2.1 Ground-level Ozone

One of the most notorious air pollutions that threaten human health is ground level O₃ (Ozone). Ozone is a gas composed of triple oxygen atoms and distributed on both the upper atmosphere level and ground level of the earth. This carries out two types of ozone in the atmosphere: stratospheric ozone and tropospheric ozone (Benedick, 1998).

Stratospheric ozone is naturally formed and beneficial to the earth's environment by blocking out most of the ultraviolet rays from the sun. On the contrary, tropospheric ozone or ground-level ozone is a type of harmful air pollution (Guttorp, Meiring, & Sampson, 1994) that causes various deleterious consequences. While there is a small amount of nature-sourced ground-level ozone, fossil fuel combustion such as vehicle exhaust, power plants, and oil refineries are the primary source of ground-level ozone precursor chemicals (Cardelino & Chameides, 1995).

Unlike most other air pollutions, the major amount of ground-level ozone pollution is a type of secondary air pollution, which is formed in the atmosphere by several chemicals rather than emitted directly from the source (Guttorp et al., 1994). Those chemicals that are related to the formation of ground-level ozone include nitrogen oxide (NO_x) and volatile organic compounds (VOC) mainly hydrocarbon (HC) (Shao et al., 2009) are also called ozone precursors (Du, Qiao, & Yu, 2019). There are three levels of adverse effects of ground-level ozone: (1) climate effects, (2) health effects, and (3) ecosystem effects. Ground-level ozone is a kind of greenhouse gas that could drastically affect the earth's climate by absorbing radiation and producing heatwaves. Based on a study by Curran et al. (2012), the radiative forcing effect from ground-level ozone is 1,000 times stronger than from carbon dioxide (CO_2), which is another kind of greenhouse gas that attracts the most attention from the public (Curran, 2012). When being exposed to the ground-level ozone, human' health will be adversely and significantly affected, especially for the elders, children, and those who are with asthma (Lippmann, 1989). Acute exposure under ozone could induce shortness of breath, wheezing, coughing, asthma, and other pulmonary and lung diseases (Gent et al., 2003). Chronic exposure under ozone could induce

cardiovascular effects, COPD, and even death (Bell et al., 2004). As ozone is a strong oxidant, ground-level ozone has the ability to damage the ecosystem by impacting the growth rate and seed production of plants (Manes et al., 2012). Previous studies revealed that several flora species are particularly sensitive to ozone, including some agricultural crops (Reich, 1987).

Ozone can be measured by remote sensing technology due to its UV spectrum absorption property (Marenco et al., 1998). The concentration of ground-level ozone is regulated by the EPA, which for the eight-hour average mole fraction concentrations, 116-404 nmol/mol is very unhealthy, 96-115 nmol/mol is unhealthy, and 76-95 nmol/mol is unhealthy to sensitive groups (Warneck, 1999). Based on the previous research, 95 US urban communities, of which approximately 40% of the US population are facing ground-level ozone pollution hazards (Bell et al., 2004). Given the hazards triggered by ground-level ozone, several strategies to decrease the ground-level ozone concentration have been established (Ryerson et al., 2001). Ground-level ozone is been identified as one of the six common air pollutants by the Clean Air Act, which is also been called criteria air pollutants by the EPA (EPA, 2020).

2.1.2 PM_{2.5}

Particulate matter (PM) is also called particulates, is a class of air pollutants that is composed of a mixture of solid and aqueous suspended substances, which includes both organic and inorganic particles such as fine dust, soot, smoke, and liquid droplets. The sizes of the PM are critical to their property due to there are impacts on their aerodynamic

properties by the size. The aerodynamic properties are also associated with the particles' chemical composition and sources, they could influence the transport and removal of particles in the air and the deposition of the particles in human and animals' respiratory systems. The aerodynamic properties of particles are summarized by the term aerodynamic diameter which is the size of a unit-density sphere with the same aerodynamic characteristics. The particles' aerodynamic diameter is usually called particle size (Xing et al., 2016). The aerodynamic sizes of PM may range from 0.1 micrometers (μm) to 10 μm , for instance, PM_{10} , $\text{PM}_{2.5}$, PM_1 , and $\text{PM}_{0.1}$ are referred to the fraction of particles with the diameter smaller than or equal to 10 μm , 2.5 μm , 1 μm and 0.1 μm (Querol et al., 2004). The boundary between the coarse particles and fine particles usually lies between 1 μm and 2.5 μm , it used to be set as 2.5 μm in the convention. The components of coarse particles and fine particles vary from earth crust materials, road and architecture fugitive dust to combustion particles and metal vapors from vehicles and industries. Through coarse acid droplets can present in fog, fine particles contain most of the acidity and mutagenic activity of PM. The fine particles (particles between 100 nm and 2.5 μm) count for most of the mass and the largest number of particles are found in extremely small sizes even less than 100 nm (Zumla et al., 2015). Exposure to high concentration PM has the potential to induce various chronic diseases and may lead to life expectancy reduction.

When compared to coarse particles, fine particles especially $\text{PM}_{2.5}$ are extremely harmful to human health due to their small diameters and large surface areas that can mix with various toxic chemicals and pass through the human respiratory system deeply without being filtered by the nose hair (Xing et al., 2016). As a result, the lower respiratory system includes the pulmonary alveoli can be easily reached, penetrated, and accumulated

by PM_{2.5}. Various previous studies showed the adverse effects of PM_{2.5} on human health, which are summarized in TABLE 1.

TABLE 1. STUDIES ON THE ADVERSE EFFECTS OF PM_{2.5}

Study	Author	Results
Six cities Study, USA	Schwartz et al. (1997)	Around 190,000 deaths were observed over years in six towns in the United States of America. In this study, PM _{2.5} was associated with mortality.
Santiago, Chile	Cifuentes et al. (2000)	Both PM _{2.5} and PM ₁₀ were associated with mortality.
Philadelphia, USA	Lipfert et al. (2000)	Both PM _{2.5} and PM ₁₀ were associated with mortality but the association with coarse PM ₁₀ was mostly not significant.
Eight cities, Canada	Burnett et al. (2000; 2003)	Both PM _{2.5} and PM ₁₀ are associated with mortality, the correlation of PM _{2.5} is much higher than coarse PM ₁₀ .
Santa Clara, California, USA	Fairley et al. (1999; 2003)	Found mortality to be associated with PM _{2.5} but not PM ₁₀ .
West Midlands Conurbation, UK	Anderson et al. (2001)	Found no association between mortality and PM _{2.5} and PM ₁₀ . However, in season-specific analyses, there was a significant association with PM _{2.5} but not coarse PM ₁₀ in the warm season.
Mexico City, Mexico	Castillejos et al. (2000)	Both PM _{2.5} and PM ₁₀ were associated with mortality, but in a two-pollutant model, coarse mass was clearly dominant. The authors speculated that there was much biogenic contamination in the coarse mass fraction.
Wayne County, Michigan, USA	Lippmann et al. (2000)	This research found PM _{2.5} and PM ₁₀ were both not significantly associated with mortality. The effect estimate for PM ₁₀ was somewhat larger than for PM _{2.5} .

Coachella Valley, California, USA	Ostro et al. (2000, 2003)	Found the cardiovascular mortality was significantly associated with PM ₁₀ but not PM _{2.5} although the effect estimate for fine particles was still much larger than for coarse PM.
Phoenix, Arizona, USA	Mar et al. (2000, 2003)	Both PM _{2.5} and PM ₁₀ were found to be associated with cardiovascular mortality

Due to the mortality and pathogenicity of PM_{2.5}, it is regulated strictly by regions and countries worldwide as shown in TABLE 2.

TABLE 2. PM_{2.5} REGULATIONS

Country/Region	Type	PM _{2.5} Concentration
Australia	Yearly average	8 µg/m ³
	Daily average	25 µg/m ³
Mainland China	Yearly average	35 µg/m ³
	Daily average	75 µg/m ³
European Union	Yearly average	25 µg/ m ³
	Daily average	N/A
Hong Kong	Yearly average	35 µg/m ³
	Daily average	75 µg/m ³
Japan	Yearly average	15 µg/m ³
	Daily average	35 µg/m ³
South Korea	Yearly average	15 µg/m ³
	Daily average	35 µg/m ³
Taiwan	Yearly average	15 µg/m ³
	Daily average	35 µg/m ³
United States	Yearly average	12 µg/m ³
	Daily average	35 µg/m ³

2.1.3 Nitrogen Dioxide/ Oxides of Nitrogen

Oxides of Nitrogen (NO_x) is a group of poisonous and highly reactive gases that are mainly generated by fossil fuel combustion (EPA, 2021a) through on-road transportation such as automobiles and trucks, and industrial sources such as power plants, refineries, and turbines. The predominant chemicals of NO_x are nitrogen monoxide (NO), which is a colorless gas, and nitrogen dioxide (NO_2), which is a reddish-brown gas that has acid and pungent odor. NO_x can be nature sourced mainly by the extreme heat of lightning of thunderstorms and wildfires (Tagle, 2021). It can also be generated by agricultural fertilization and nitrogen-fixing plants. Based on previous research, on-road transportation counts for 40% of urban NO_x , commercial institutional and households count for 14%, energy production and use in industry count for 34% (EPA, 2021a).

NO_x may interact with water, oxygen, and other chemicals to form acid rain that damage the waterbody as well as the organisms (EPA, 2021a) by negatively affecting the vegetation and making them more susceptible to disease and frost. It may also react with ammonia and other compounds to form smog and acid vapor and cause damage to lung tissues. While NO is not normally considered hazardous to human health at a typical ambient condition, it may still induce several adverse effects such as respiratory, metabolic, and blood pressure disorders and diseases (Tagle, 2021). Inflammation of the upper airways and other respiratory problems such as wheezing, coughing, and bronchitis may be triggered by the acute contact of a high concentration of NO_2 , especially for the population with asthma. Chronic contact with NO_2 at a high level may cause irreversible damages to the respiratory system (Amr & Hadidi, 2001). Furthermore, as introduced previously, NO_x/NO_2 is responsible for the formation of ground-level ozone, which is a

secondary air pollutant. An oxide ion is formed by gaseous NO_2 in the presence of sunlight, and the oxide ion further combines with the oxygen molecule (O_2) to form ozone (Tagle, 2021). When ozone is present, NO can be converted to NO_2 in the atmosphere.

Due to the adverse effects of NO_x , it is regulated by the current national ambient air quality standards (NAAQS) published on April 6th, 2018, that 1-hour standard level at 100 ppb, and annual standard level at 53 ppb, which could protect the public health, especially for the high-risk individuals include elders, children, and people with asthma (EPA, 2021b).

2.3 Influencing Factors of Air pollution

Air pollution concentration levels of an area are not constant values, which vary by time and can be influenced by different factors. In this research, the influencing factors of the air pollution concentration are categorized into two classes, which include the nature influencing factors that are meteorological situations such as solar radiation, outdoor temperature, pressure, precipitation, relative humidity, wind speed and direction, and human activity factor that is mainly on-road transportation.

2.3.1 Meteorological Measurements

Ground-level ozone is formed by the presence of sunlight. Thus, meteorological conditions influence ozone formation largely. The meteorological influencing factors of ozone concentration have been extensively studied. Ding et al., 2013 and Gao et al., 2005

indicated that the formation of ground-level ozone is highly related to the tropical cyclones and continental anticyclones (Ding, Wang, Zhao, Wang, & Li, 2004; Gao, Wang, Ding, & Liu, 2005). The tropical cyclones are low pressure, closed low-level atmospheric circulation, and strong wind rapid rotating storm system that accompanied with thunderstorms and heavy rains (Emanuel, 2003). On the contrary, the anticyclones are associated with the large circulation of winds around a high-pressure core that bring clear skies and cooler and drier air (Rodwell & Hoskins, 2001). Based on Ding's research, the sunny and low-wind weather created by anticyclones provides a favorable situation for ozone formation. The peripheral of the tropical cyclones in the Western Pacific region produces a low-pressure system that brings high temperature and sunlight as well as light wind, which also contribute to the formation of ozone (Ding et al., 2004). It is also revealed that the ground-level ozone can be influenced by several local meteorological situations.

Ding et al., 2004 conducted research on the formation of ground-level ozone within an economic region. The results showed that the days with strong sunlight and low winds are beneficial to the formation and accumulation of ozone along with its precursors, which is correlated with higher ozone concentration (Ding et al., 2004). Research also indicated that the wind direction could affect the ozone concentration by influenced pollution transportation (Duan, Tan, Yang, Wu, & Hao, 2008). However, different effects of wind directions occur based on the specific location. For instance, the upslope wind in a valley area may transport ozone from the bottom upward to the peak (Gao et al., 2005). In research conducted in Beijing, which is one of the most ozone-polluted cities, the ozone was transported up to the surrounding mountains in afternoons during summer and was transported back in the evening (T. Wang, Ding, Gao, & Wu, 2006). The seashore and

offshore winds may impact ozone accumulation, especially the cycling wind pattern that traps the ozone pollution in the city areas (Tie, Geng, Peng, Gao, & Zhao, 2009).

Based on previous research, meteorological conditions show significant effects on PM_{2.5} concentration, especially in urban areas. A study conducted in Japan showed that the outdoor temperature is positively correlated with local PM_{2.5} pollution concentration, while it is negatively correlated with wind speed and relative humidity (J. Wang & Ogawa, 2015). As stated in the research, among the meteorological factors, wind speed and relative humidity play a more important role in influencing PM_{2.5} concentration than temperature. There was a threshold for the correlation coefficients between wind speed, relative humidity, and PM_{2.5} concentration because of the geography profile of Japan, which is location-specific. A study conducted by Xiao et al., 2001 analyzed the correlations between PM_{2.5} concentration level and several selected meteorological parameters that include wind speed, temperature. The correlation coefficient is 0.32 and 0.36, which means these three factors are correlated (Z.-m. Xiao et al., 2011).

The chemical transport models (CTMs) driven by general circulation models (GCMs) developed by Climate Research Community are commonly used to simulate the PM_{2.5} concentration variation trends that are influenced by weather factors (Liao et al., 2006; Racherla & Adams, 2008). While the CTMs and GCMs models aim at different targets when they were designed, both are capable of performing global atmospheric chemistry modeling in a complementary manner (Jeuken, 2000). Amos et al., 2010, conducted research based on CTMs and GCMs models that analyzed the relationship between PM_{2.5} concentration and climate conditions (Tai, Mickley, & Jacob, 2010). In the research, it was revealed that variant chemical components in PM_{2.5} pollution are

influenced by temperature, relative humidity, and wind with different correlation coefficients. To be more specific, the outdoor temperature is positively correlated with sulfate, organic carbon, elemental carbon, and negatively correlated with nitrate; while the relative humidity is positively correlated with sulfate, nitrate, and negatively correlated with organic carbon, elemental carbon; and as other research indicated, the wind effects on $PM_{2.5}$ concentration vary by locations.

Atmospheric NO_x concentration is related to the meteorological status as well. A study targeting a valley area in Nepal that is conducted by Pudasainee et al., 2006, indicates that the NO and NO_2 concentration met their peak value following the presence of sunlight, and the mid-day peak of NO_x occurred with lower nocturnal concentration (Pudasainee et al., 2006). In the meantime, the NO_x concentration level is slightly lower in the monsoon season. Ocak et al., 2008 proposed a statistical model analyzing the relationship between air pollutants and meteorological factors (Ocak & Turalioglu, 2008). In the research, it is indicated that the daily CO , NO_x , and O_3 air pollution levels are influenced not only by the meteorological parameters that include wind speed, temperature, and relative humidity but also by the level of the previous day. The historical meteorological measurements are analyzed by the multiple linear regression algorithm. Based on the analysis, the level of NO_x shows a negative relationship with wind speed and temperature. Several previous research utilized the hourly and seasonal factors to analyze the NO_x concentration variance. David et al., 2011, analyzed the association between NO_x concentration and the mesoscale synoptic meteorological measurements by temporal variances in a tropical coastal area (David & Nair, 2011). It is found that the diurnal NO_x concentration pattern is closely related to the mesoscale circulation that includes mainly the wind.

2.3.2 Transportation

On-road vehicle emissions are responsible for lots of air pollutions such as CO, PM, and ozone precursor chemicals. Among all the on-road transportations, highway transportation is playing an important role in vehicle emissions. As of 2016, one-fourth of all vehicle miles traveled in the USA are on the highway system (FHWA, 2017). Based on a recent report by EPA, transportation is responsible for over 55% of total NO_x emissions, around 10% of VOCs emissions, and around 10% of PM emissions in the USA. By proper transportation emission management, it is expected to reduce 40,000 premature death, 34,000 hospitalization visits, and 4.8 million workdays lost by the year 2030 (EPA, 2019, 2021c). In the meantime, on-road transportation accounts for 18.4% of total PM emissions worldwide (Xia et al., 2015). It is revealed that long-term exposure to traffic-related air pollution may reduce life expectancy (Zhang, Khlystov, Norford, Tan, & Balasubramanian, 2017).

The pandemic caused by the novel coronavirus, which is abbreviated as COVID-19, has been outbreaked since the end of the year 2019 worldwide that became a global medical problem. Other than the pathogenicity of the pandemic, it also impacts human activities. Most regions have issued the so-called Stay-at-Home or quarantine orders to limit the spread of the COVID-19 pandemic. Some major cities and urban areas even practiced the lockdown policy that only essential businesses remain in operation (Gray, 2020). The living styles along with the traffic pattern were altered due to those situations. The main impacts of the pandemic outbreak on on-road transportation include the travel

demand, the transportation mode, and the land use or even urban planning (Du, Wang, & Qiao, 2020).

Based on a United Nations Educational, Scientific and Cultural Organization (UNESCO) report, due to the work from home order, the work-related traffics has been significantly reduced (UNESCO, 2020). On the contrary, it is reported that the non-work-related travels are increasing such as shopping and delivering (Wu, Chen, & Chan, 2020). It is further observed that the total trips in 2020 were 2.8 billion fewer than that of the year 2019 (Du et al., 2020), especially, the trips longer than three miles are reduced from December 2019 to April 2020, however, it was slightly increased after that till August 2020 (Chinazzi et al., 2020). This phenomenon was also true for other major cities worldwide. For instance, the travel demands of the first half of the year 2020 have reduced by 40% in Taipei, 80% in London, and 90% in Milan (Gössling, Scott, & Hall, 2020). Freight transportation that including the railway volume has been decreased by 20% due to business inactivity. However, the enhanced e-commerce or online shopping mode has resulted in 30% more freight traffic volumes in the US (Newport, 2020). The concentration levels of air pollutants that come from fossil fuel combustion were extensively impacted by the pandemic given the on-road transportation is one of the most significant contributors to them. Thus, considering the transportation factor for air pollution analysis is meaningful, especially during a pandemic situation.

2.4 Air Pollution Forecasting Technologies

Due to the non-linear and complex nature of air pollution levels, forecasting them became a tough task and usually only within a short target time period. However, various approaches are utilized to perform the air pollution concentration level forecasting and estimation. One of the most widely used algorithms is the neural network (NN) model. Based on a previous review work conducted by Cabaneros et al., 2019, the majority of air pollutants that NN models are used to predict are PM₁₀, PM_{2.5}, NO_x, and ozone. Some of them utilized the meteorological and source emission predictors as part of the inputs of the multilayer perceptron and ensembled models (Cabaneros, Calautit, & Hughes, 2019). The time-series data as inputs is usually associated with the NN models that are utilized to predict the air pollution concentration level.

The research conducted by Niska et al., 2004 designed a NN model to forecast hourly NO_x concentration in Helsinki (Niska, Hiltunen, Karppinen, Ruuskanen, & Kolehmainen, 2004). In this research, a parallel genetic algorithm (GA) was built for NN input selection and high-level architecture of a multi-layer perceptron model. However, it is shown that the evaluation process was computationally expensive and limited the search technique. A predictive model based on NN targeting SO₂, PM₁₀, and CO levels in the Greater Istanbul area was built by Kurt et al., 2008 (Kurt, Gulbagci, Karaca, & Alagha, 2008). The results showed that relatively accurate predictions could be provided by NN that using the historical 3-15 days' value as the training set to forecast the future three days. However, the forecasting period of this research is only three days, which is relatively short. In the meantime, the forecasting of the second and the third day requires the previous day's value, which limited the accuracy and practicability. Other than that, this research

involved the day of week data as an input parameter, which improved the forecasting accuracy. A case study conducted by Azid et al., 2014, focused on the air quality of Malaysia analyzed eight air pollutants in ten years (Azid et al., 2014). An artificial NN was developed with the principal component analysis (PCA) that requires only a few variables.

Many researchers have considered integrating meteorological measures in forecasting to improve accuracy. Wen et al., 2019, exhibited a spatiotemporal convolutional long short-term memory (C-LSTME) NN extended model to perform the air pollution prediction (Wen et al., 2019). The historical air pollution data were utilized in the model as well as the k-nearest neighboring stations and the integrated meteorological data. It was claimed that the model was well performed for different time predictions at different region scales. Maleki et al., 2019, conducted research to evaluate the hourly air pollution (ozone, NO₂, PM₁₀, PM_{2.5}, SO₂, and CO) concentration prediction ability of an artificial NN algorithm in Ahvaz, Iran, over a year (Maleki et al., 2019). The inputs of the 30-neurons model include five meteorological parameters, time, date, and three hours and six hours previous pollution concentrations. The results demonstrated the correlation coefficient and root-mean-square error values of the prediction. However, there is no intercomparison to show the advantage or disadvantages of the artificial NN algorithm. Elangasinghe et al., 2014, proposed a protocol to extract the key information from meteorological parameters and the emission patterns of a year to build an artificial NN model (Elangasinghe, Singhal, Dirks, & Salmond, 2014). A case study was conducted in Auckland, New Zealand targeting NO₂ pollution with eight input variables: meteorological parameters such as wind speed, wind direction, solar radiation, temperature, relative humidity, and time factors such as the hour of the day, day of the week and month of the

year. A simplified model was also created that will not significantly decrease the performance, which outperformed the linear regression model with the same inputs.

As a cutting-edge technology, machine learning has shown its abilities in air pollution forecasting for a predefined future period, which could voluntarily learn from the input data to improve the analysis. When compared to other forecasting methods, the advantages of the machine learning technique are significant to perform more accurate forecasting: the data precession speed is accelerated, automatic forecasting updates, more data capacity, hidden pattern identification of the data, and increased adaptability (Taranenko, 2019). In the research conducted by Shaban et al., 2016, univariate and multivariate machine learning algorithms that include support vector machines and M5P model trees were adopted to build one-step and multi-step ahead forecasting models targeting ozone, NO₂, and SO₂ (Shaban, Kadri, & Rezk, 2016). Results showed that the M5P algorithm yielded more accurate predictions when using different features. A study conducted in China compared different machine learning classification algorithms for 74 cities (Xi et al., 2015). It is revealed in the research that the accuracy is positively related to the number of features selected to use, and the combined model is usually better than a unique model. Srivastava et al., 2018, implemented various classification and regression models that include linear regression, SDG regression, random forest regression, decision tree regression, support vector regression, artificial NN, gradient boosting regression and adaptive boosting regression to forecast the air pollution concentration such as PM_{2.5}, PM₁₀, CO, NO₂, SO₂, and ozone (Srivastava, Singh, & Singh, 2018). After a case study of New Delhi and evaluation, the support vector regression and artificial NN outperformed other algorithms in the forecasting.

Besides air pollution level forecasting, machine learning technology can often be used in other air pollution-related analyses. For instance, Bellinger et al., 2017 reviewed the utilization of machine learning and data mining on air pollution epidemiology (Bellinger, Mohamed Jabbar, Zaïane, & Osornio-Vargas, 2017). In this research, the air pollutants, public health, and environmental factors were merged and imported to find the patterns, extract information and make predictions. It is concluded that the early studies are more focused on artificial NN, and the decision trees, support vector machines, k-means clustering, and the Apriori algorithm are extensively used in more recent works. From the literature review of previous research, the meteorological parameters and transportation data can be used to perform the air pollution forecasting, however, the year-round or other long-term forecasting is always underperformed with relatively lower accuracy. On the other hand, new technologies such as machine learning and data mining were usually conducted based on time-series analysis, the meteorological parameters were occasionally considered as improving factors. Thus, the use of data mining and machine learning to analyze the air pollution concentration pattern and further forecast the pollution level in the long-term future is meaningful and practical.

CHAPTER 3

DESIGN OF THE STUDY

This research involves using ten years of historical air pollution concentration records and the air pollution influencing factors, including meteorological measurements and travel activity status data for temporal statistical analysis and frequent pattern mining analysis. Based on the inherent relationships between the air pollution concentration and influencing factors, a pool of supervised machine learning models will be developed, compared, and evaluated, and the most fitted models for each pollutant will be selected to perform the air pollution concentration level prediction for the year 2020. The main analytical tools used in this research include by not limited to Python 3.7, Pandas 1.2.4, Scikit-Learn 0.24.1, Mlxtend 0.18.0, Scipy 1.6.2, Matplotlib 3.3.4, and Microsoft Excel VBA.

To achieve the research objectives, various datasets are collected and utilized and analytical analyses are conducted as shown in Figure 1. Figure 1 is the flow chart of this research, in which, the magnetic disk objects are datasets, the block arrow objects are processes, and the rectangle objects are analysis. As shown in the flow chart, there are mainly four modules that include: (1) data processing that is shown in the blue block, (2) temporal analysis that is shown in the red block, (3) frequent pattern mining analysis that is shown in the gray block, and (4) machine learning forecasting that is shown in the green block.

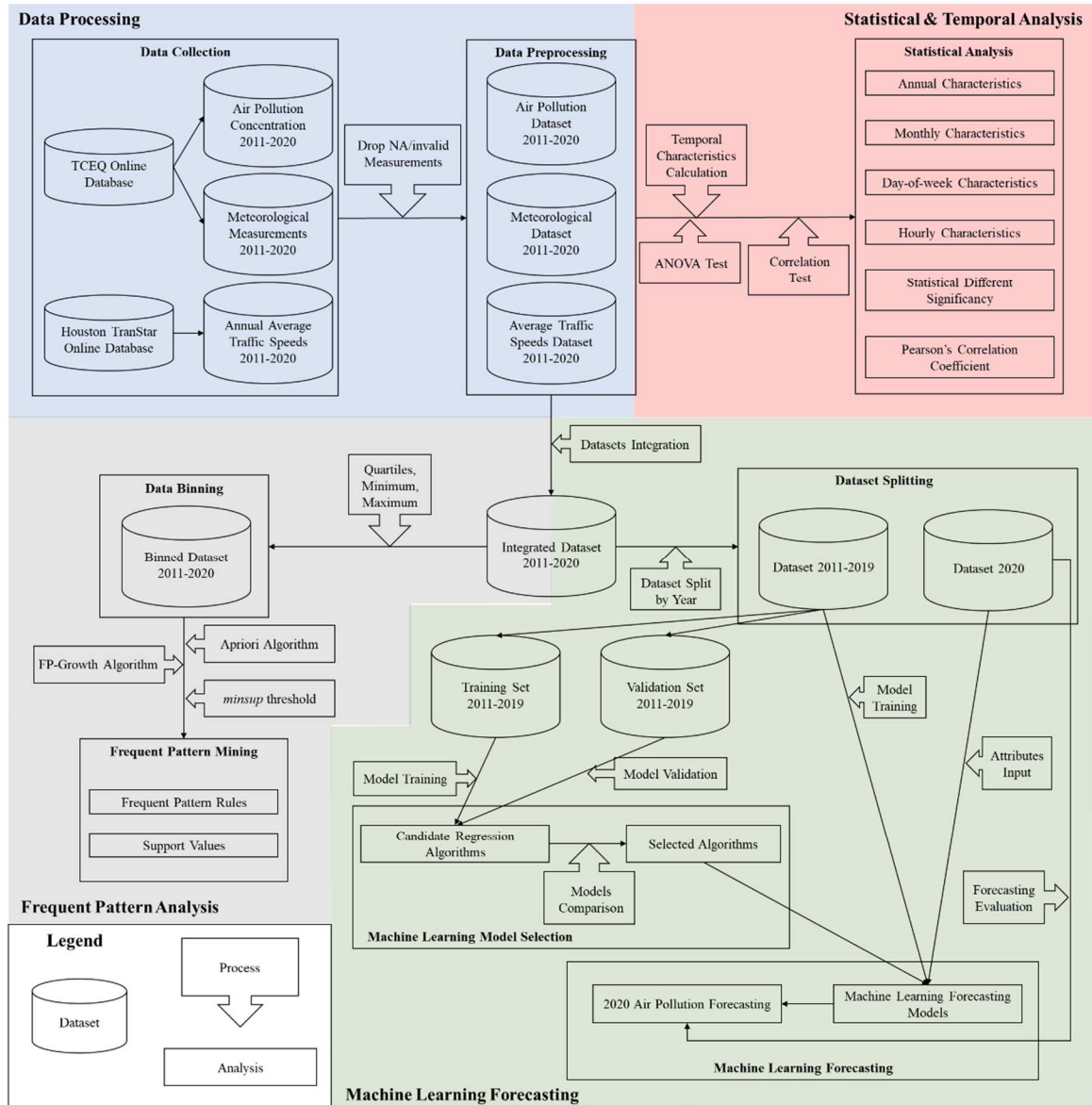


Figure 1. Flow Chart of This Research

As Figure 1 indicates, data collection and processing are fundamentals of this research, while all calculations and analyses are based on the data collected. Three parts of data from the year 2011 to 2020 are utilized in this research, which includes air pollution concentration data, meteorological measurements data, and travel activity data. All three parts of data are preprocessed and transformed into datasets of the same format that can be

analyzed. The statistical and temporal analyses are performed by a series of statistical tests such as the ANOVA test and correlation test, and temporal characteristics analysis. The datasets are then integrated into a uniform dataset that each air pollution record is associated with a series of meteorological and traffic attributes. The frequent pattern mining is performed on the binned integrated dataset by Frequent Pattern (FP)-Growth and Apriori algorithms, which yields the frequent pattern rules with corresponding support values. The machine learning forecasting process is conducted on the split integrated dataset. The proper machine learning models are selected through a model selection process, and the selected models are trained to perform air pollution forecasting. The forecasting results are validated and evaluated by the real data from the dataset. The details of the processes, algorithms, modeling, and analysis are shown in the rest of this chapter.

3.1 Data Collection

3.1.1 Air Pollution Concentration Data Collection

The air pollution and meteorology data are collected from the Texas Commission on Environmental Quality. The TCEQ is the fourth largest environmental agency in the US, which has sixteen regional offices and the headquarter is located in Austin. The TCEQ is currently operating more than 200 air monitoring stations serving over 25 million statewide areas in Texas including industrial and large population regions. There are different types of air toxins The TCEQ online database is called the Texas Air Monitoring Information System (TAMIS). The air quality data values for parameters include criteria pollutants, hazardous air pollutants (HAPs), volatile organic compounds (VOCs), and

meteorological data. Different air pollutants are monitored by different networks in TCEQ. The networks include: (1) a Community Air Toxics Monitoring Network that collects every six days from urban and industrial areas and analyzed by a gas chromatograph-mass spectrometer, (2) Automated Gas Chromatography (AutoGC) Samplers that are located in major cities collect 40 minutes of data per hour and analyzed automatically on-site, (3) Carbonyl samplers that collect carbonyl compounds by high-performance liquid chromatography every six days in major cities, and (4) Air toxic metal monitors collect twenty-four metallic PM_{2.5} and PM₁₀ samples every six or three days. For this research, the main air pollutant data are collected by the AutoGC samplers.

The target air pollutants in this research include ground-level ozone, PM_{2.5}, NO₂, and NO_x. The meteorological measurements include solar radiation, temperature, pressure, precipitation, relative humidity, resultant wind speed, and resultant wind direction. The time span of the historical data should be ten continuous years from 2011 to 2020. Currently, there are more than 20 air monitoring sites operated by TCEQ in the Houston-Galveston area. However, during the data collection process, most of the sites failed to provide qualified raw data from the year 2011 to 2020. Most of the sites have one of the following issues that cannot be selected as the data collection site by this research: (1) too much null data from the monitoring sites that significantly influence the data quality, (2) too much invalid data such as negative measurements for some parameters, (3) discontinues measurements that most of the monitoring sites are involving, some of them may be inactivated for more than two years, (4) located in the areas that cannot represent the Houston metropolitan area. Based on the above criteria, air monitoring site 403 (TCEQ site name: Clinton C403/C304/AH113, which is located at 9,525 ½ Clinton Dr. nearby

Highway 610 and the ship channel was selected to perform the data collection as shown in Figure 2.

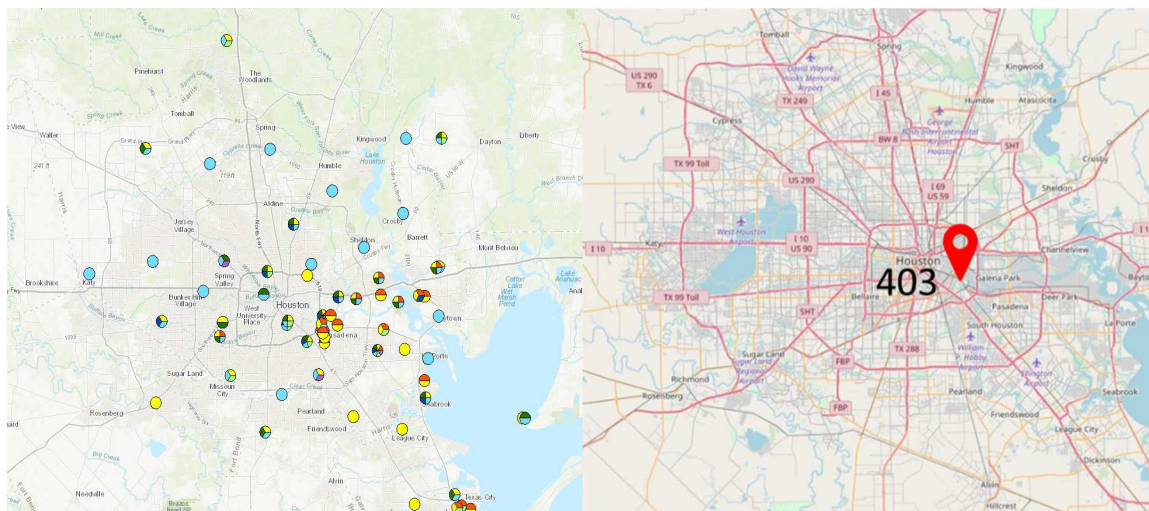


Figure 2. Air Monitoring Sites in Houston-Galveston Area and Site 403

TABLE 3. AIR POLLUTANTS CONCENTRATION COLLECTED FROM TCEQ DATABASE

Air pollutants		Unite
Ground-level ozone	O ₃	Parts per billion (ppb)
Fine Particulate Matter	PM _{2.5}	Micrograms per Cubic Meter (ug/m ³)
Nitrogen Dioxide	NO ₂	Parts per billion (ppb)
Nitrogen Oxides	NO _x	Parts per billion (ppb)

The PM_{2.5} data is measured near real-time for particulates less or equal to 2.5 micros in size from the surrounding air, which is made at local conditions and not corrected for temperature or pressure. For NO_x measurement, all higher oxides of nitrogen are grouped. The hourly air pollution concentration raw data that is collected as shown in TABLE 3 was calculated by the average of the testing equipment reading by every five minutes.

3.1.2 Meteorological Measurements Data Collection

Different technologies are utilized to perform the monitoring for different meteorological measurements. All hourly meteorological measurements are collected from the same TCEQ monitoring site (site 403) as the air pollution concentration.

TABLE 4. METEOROLOGICAL MEASUREMENTS COLLECTED FROM TCEQ DATABASE

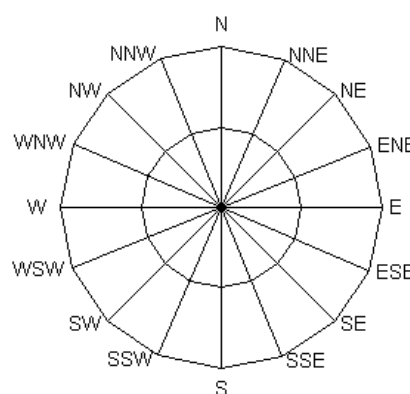
Meteorological Measurements	Unit
Solar Radiation	langleys per minute (Langleys/min)
(Outdoor) Temperature	degrees Fahrenheit (°F)
(Barometric) Pressure	millibars
Precipitation	inches
Relative Humidity	Percent (%)
Resultant Wind Speed	miles per hour (mph)
Resultant Wind Direction	degrees compass (°)

TABLE 4 shows the meteorological measurements collected from the TCEQ online database. Solar radiation is measured by the total electromagnetic radiation emitted by the sun and received by the monitoring site. The temperature is measured outside the monitoring site. Precipitation is the rainfall from the cloud to the ground in liquid or solid form. The relative humidity is the percentage measurement of the moisture in the air that ranges from 0% that means no humidity to 100% that means totally saturated air. The resultant wind speed and direction is a single vector that is measured by converting the five minutes wind speeds and directions. The resultant wind direction shows where the wind is

blowing, which is measured to the nearest degree on a 360degree compass that the 0° (360°) means the north, and 180° means the south. The details about the resultant wind direction are shown in TABLE 5.

TABLE 5. RESULTANT WIND DIRECTIONS SHOWN ON COMPASS

Cardinal Direction	Degree Direction
N	348.75 - 11.25
NNE	11.25 - 33.75
NE	33.75 - 56.25
ENE	56.25 - 78.75
E	78.75 - 101.25
ESE	101.25 - 123.75
SE	123.75 - 146.25
SSE	146.25 - 168.75
S	168.75 - 191.25
SSW	191.25 - 213.75
SW	213.75 - 236.25
WSW	236.25 - 258.75
W	258.75 - 281.25
WNW	281.25 - 303.75
NW	303.75 - 326.25
NNW	326.25 - 348.75



3.1.3 Traffic Situation Data Collection

The ten years traffic situation data from the year 2011 to 2020 is collected in form of annual speed averages from the online database that is operated by Houston TranStar: <http://traffic.houstontranstar.org/hist/histmain.aspx>. The data is recorded and collected in 15 minutes intervals from 5:00 to 19:00 during full years' worth of weekdays. Various technologies are used to measure the average speeds of on-road vehicles, which are also utilized by the Houston Transtar speed maps, roadside travel time message signs, and radio and television media traffic condition broadcasting, etc. The main technology for average

on-road vehicle speed detecting is the Anonymous Wireless Address Matching (AWAM) which is supported by Bluetooth™.

Currently, more than 20 highway segments' annual average speed can be achieved from the Houston TranStar database. In this research, the traffic situation of the highway that is nearest to the air monitoring site 403 is critical, which is Highway 610 East Loop. There are two series of annual average speeds include the Southbound from Wayside Drive to Broadway Street (4.9 miles) and the Northbound from SH-225 to Gellhorn Drive (4.9 miles). From the traffic speed data collection, ten years of historical data with each year containing 56 15-minutes time interval records for each direction are collected.

3.2 Raw Data Preprocessing and Statistical Analysis

3.2.1 Data Preprocessing

The raw data collected from the TCEQ contains some invalid measurements. Those invalid measurements may impact further analysis. During the data preprocessing step, the main job is to eliminate invalid measurements. There are six error types in the raw data as in TABLE 6.

TABLE 6. INVALID MEASUREMENTS OF RAW DATA

Error Codes	Detailed Information
NA	The average cannot be computed until all the measurements are received. This average will not be available until enough data has been received for the current or past hour.
AQI	Data rejected by TCEQ validators. The TCEQ validators have reviewed the data and determined that it is not valid.

FEW	Not enough five-minute measurements available to create an hourly average. There must be 45 minutes of data available each hour for a valid hourly average to be created.
LIM	Data exceeds automatic criteria for rejection. For meteorological parameters, this indicates that the measurements fall outside the EPA guidelines for meteorological collection of data. For pollution parameters, this indicates that the instrument failed a scheduled automatic calibration or span check.
LST	Lost data. Usually indicates that data was never collected. This may also be triggered by delays or breakdowns in data communications. Try retrieving the data later.
QAS	Quality Control audit in progress. TCEQ conducts periodic Quality Control audits on each monitoring site.

The calculation and analysis of the data in this research require numerical records or float and integer data types, thus, the above invalid data types need to be eliminated. There are several techniques to remove or convert the invalid data such as using the moving average, replacing it with the nearest valid data, and replacing it with a certain value. In this research, the data will be processed by frequent pattern mining and machine learning analysis, it is essential to keep the original data trends untouched and to avoid the subjective factors of the conductor. To meet this objective, the invalid records in the raw data are dropped through all parameters. To unify the data sets, if a record x of a parameter is invalid, the record x of all parameters will be dropped respectively. The raw data collected from the TCEQ database contains 87,360 hourly records originally. 14,124 invalid records are removed from the data preprocessing and the remaining 73,236 valid records are processed to the following analysis.

There is no invalid data in the annual average traffic speeds collected from the Houston TranStar, however, preprocessing is necessary to convert the data into the same

format as the air pollution concentration and meteorological measurements records. As introduced in the data collection section, the original average traffic speeds are recorded annually in the 15-minutes time interval from 5:00 to 19:00 that includes both traffic directions of eastbound and southbound. There are generally three steps in traffic speed data preprocessing. (1) For each 15-minutes time interval data, calculate the mathematical average between the northbound and southbound speed to get the on-road speed for the selected road segment. (2) The hourly average traffic speeds are calculated by the mathematical average of the four 15-minutes time intervals on-road speed from the previous step. (3) Assume the on-road speeds from 0:00 to 5:00 and 20:00 to 24:00 equal to the designed free-flow speed of the selected road segment, which is 60 mph. By the steps above, the traffic speed data are converted to hourly records.

3.2.2 Statistical Analysis

To ensure the data of each parameter are comparable and significantly different between the ten years, the one-way ANOVA test is performed with the significance level of 0.05. The temporal analysis is then performed on the preprocessed data collected from each agency's database are analyzed. Four types of temporal characteristics are analyzed for historical air pollution concentration levels and meteorological measurements from the year 2011 to 2020, which include the annual, monthly, day-of-week, and hourly characteristics. For details, the analysis of the annual characteristics provides the yearly variation trends of the air pollution concentration and meteorological measurements; the analysis of the monthly characteristics provides the monthly and seasonal variation trends of the air pollution concentration and meteorological measurements; due to the day of

weeks is an artificial definition that is not related to the earth movements and only related to the human activities, only air pollution concentration day-of-week characteristics are analyzed to provide the daily variation trends in a week; the analysis of the hourly characteristics provide the variation trends of air pollution concentration and meteorological measurements through a day. The temporal characteristics of the average traffic speeds are analyzed to provide the yearly and hourly traffic variation trends.

To measure and analyze the linear correlation between the parameters, Pearson's correlation coefficients r are calculated, which can be expressed by Equation 1.

$$\text{Pearson's } r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

where:

n = the sample size of the data collected,

x_i, y_i = the single sample indexed with i ,

\bar{x}, \bar{y} = the sample mean of the data collected.

The Pearson's r values range from -1 to +1 for negative and positive correlations, respectively. The r values that are closer to the -1 and +1 values present the stronger correlation between the entries, and closer to 0 present a lower correlation.

3.3 Frequent Pattern Mining

According to the theory of data mining, the concept of "pattern" is a set of items, subsequences, or substructures that occur frequently together (or strongly correlated) in a

data set. Patterns can represent intrinsic and important properties of datasets. To perform the frequent pattern mining analysis, the format of the input data needs to be converted.

3.3.1 Input Datasets Integrating and Data Fitting

The original data are in continuous data types. However, frequent pattern analysis is better performed on discrete data types such as integers and labels. In this research, the first step of frequent pattern mining analysis is to categorize the air pollution concentrations, meteorological measurements, and average traffic speeds into bins that can be represented as integers. The bins for each parameter are determined by their minimum, maximum, and quartile values.

After binning each parameter, four air pollutants datasets, seven meteorological factors' datasets, and one average traffic speed dataset are integrated into one dataset of records that contains four classes, eight attributes, and 73,236 records. The example of one record in the integrated dataset is shown in TABLE 7.

TABLE 7. EXAMPLE OF A RECORD IN THE INTEGRATED DATASET

Class				Attributes								
Ground-level ozone	PM _{2.5}	NO ₂	NO _x	Solar	Temperature	Pressure	Precipitation	Relative Humidity	Wind Speed	Wind Direction	Traffic Speed	
27.0	5.2	6.1	5.7	0.0	59.4	1009.0	0.0	53.6	4.1	130.0	60.0	

As stated above, there are 73,236 records in the integrated dataset. To perform frequent pattern mining on the input data, the values in the integrated dataset need to be converted into binary format. The process is to list all bin values of parameters and set

them as new attributes, and the parameters bin array will be transformed into a sparse array. The transformed dataset is also called a fitted dataset. For example, in the record shown in TABLE 7, the ground-level ozone and solar integer entries are transformed into binary entries in TABLE 8.

TABLE 8. EXAMPLE ENTRIES IN THE TRANSFORMED DATASET

Integrated Dataset Entries	Ground-level ozone				Solar			
	Ground-level ozone_bin_1	Ground-level ozone_bin_2	Ground-level ozone_bin_3	Ground-level ozone_bin_4	Solar_bin_1	Solar_bin_2	Solar_bin_3	Solar_bin_4
Fitted Dataset Entries	0	0	1	0	0	0	0	1

As shown in TABLE 8, within the fitted data, each air pollutant concentration level record contains a series of binary information of all related factors. If a factor is related to an air pollutant concentration level record, the corresponding input is one. Otherwise, the input is zero. In this research, different air pollutants are analyzed separately.

3.3.2 Understanding Frequent Pattern Mining

The process of pattern discovery is to find the inherent regularities in an air pollutant concentration level record, in which the influencing factors are considered as items. In this research, the term ‘Item’ is the listed attributes on each record (e.g., solar, temperature, pressure, precipitation...), and the term “Itemset” is a set of one or more items. A k -itemset can be represented as $X = (x_1, x_2, \dots, x_k)$. The absolute *Support* or count of X is the frequency or the number of occurrences of itemset X . The relative *Support* s is

the percentage of transactions that contain X , which is also the probability an air pollutant concentration level record contains X . An itemset X is frequent if the Support of X is no less than a minimum support (*minsup*) threshold (σ).

The *Support*, *Confidence*, and interestingness measurement LIFT can be calculated using Equations 2- 4 (Lin, Wang, & Sadek, 2015).

$$s(C, D) = s(C \cup D) = \frac{n(C \cup D)}{n(T)} \quad (2)$$

$$c(C, D) = \frac{s(C \cup D)}{s(C)} \quad (3)$$

$$l(C, D) = \frac{c(C \cup D)}{s(D)} = \frac{s(C \cup D)}{s(C) * s(D)} \quad (4)$$

where,

$s(C, D)$: the *Support* for air pollutant concentration C and influencing factor value D occurring together, ranging (0, 1),

$n(C, D)$: the number of events when C and D occur together,

$n(T)$: the number of total events,

$c(C, D)$: the *Confidence* for event D to occur when event C occurs, ranging (0, 1),

$l(C, D)$: the interestingness measurement LIFT (ranging (0, ∞)) for event D to occur when event C occurs, which tells how C and D are correlated,

if $l(C, D) = 1$, events C and D are independent,

if $l(C, D)$ in (1, ∞), events C and D are positively correlated, and

if $l(C, D)$ in (0, 1), events C and D are negatively correlated.

The *Support* $s(C, D)$ can provide the scale of an air pollutant concentration record occurring on a set of influencing items. The *Confidence* $c(C, D)$ is the likelihood of an item

occurring if another item happened. The LIFT illustrates the increase in an air pollutant concentration record when another item happened.

3.3.3 Frequent Pattern Mining Algorithms

Two typical frequent pattern mining algorithms are widely used: the Apriori algorithm and the Frequent Pattern (FP)-Growth algorithm (Aggarwal, Bhuiyan, & Al Hasan, 2014). Figure 3 shows the flow chart of the Apriori algorithm.

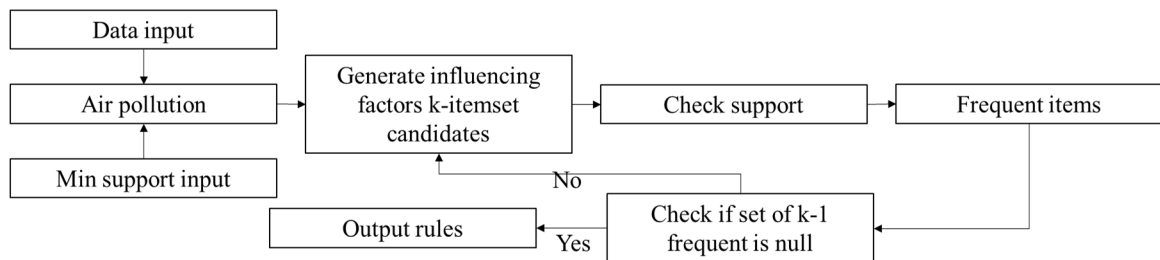


Figure 3. Flow Chart of The Apriori Algorithm (Han, Pei, & Kamber, 2011)

The Apriori algorithm scans all possible itemsets and conducts all calculations. As shown in Figure 3, the itemset candidates of air pollution factors are generated from the fitted dataset as inputs, which are compared with the *Support* that is set by the *minsup*. If the *Support* of the candidate itemset is greater than the *minsup*, the frequent items are recorded and the process goes through the null test. The output is then generated after passing null tests. The pseudo-code of the Apriori algorithm is shown as follows (Han, Pei, Yin, & Mao, 2004).

Pseudo-code of the Apriori algorithm

```

Ck: Candidate itemset of size k;
Fk: Frequent itemset of size k;
k := 1;
Fk := {frequent items}; // frequent 1 – itemset
While (Fk ≠ ∅) // when Fk is nonempty
    do { Ck+1 := candidates generated from Fk; // candidate generation
        Derive Fk+1 by counting candidates in Ck+1 with respect to TDB
        at min_support; k := k + 1 }
return  $\cup_k F_k$  // return Fk generated at each level

```

Unlike the Apriori algorithm, the FP-Growth algorithm does not consider all possible itemsets. There are generally two parts of the FP-Growth algorithm, which are creating the FP-Tree and applying the FP-Growth algorithm. The FP-tree is created by: (1) scanning the database once and collecting the dataset *F* along with its *Support* and sorting the dataset by descending sequence and saving it as a list of datasets; (2) creating the root

r of the FP-tree and note it as *null*; for each transaction in the database T_i , selecting frequent items and sorting the list, and calling $insert_tree(T_i, r)$; and (3) creating the function $insert_tree(T_i, r)$ by checking if node r has successive nodes N that $N.item-name=p.item-name$, N increase by 1 if true, or creating a new node N that links to its parent node and set its value to 1. Figure 4 shows the flow chart of the FP-Growth algorithm.

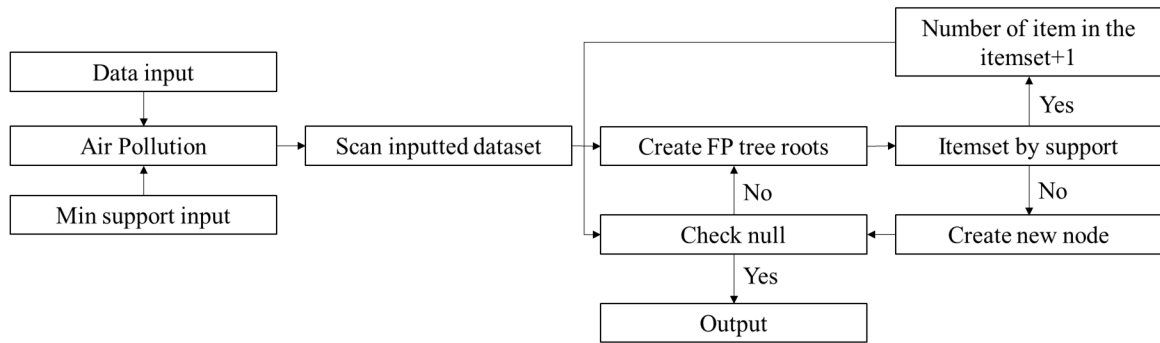


Figure 4. Flow Chart of The FP-Growth Algorithm (H. Li, Wang, Zhang, Zhang, & Chang, 2008)

As shown in Figure 4, the right part shows the construction of the FP-Tree and the left part shows the application of the FP-Growth algorithm. The pseudo-code of the FP-Growth algorithm is shown as follows (Han et al., 2004).

Pseudo-code of FP-Tree

```

T: List of transactions;
min_sup: minimum support value;
define F[] =  $\emptyset$ ;
for each transaction Ti in T
    for each item aj in Ti
        do{ F[aj] ++ }
sort F[];
define null root r;
for each transaction Ti in T
    do{ Ti ordered by F
        call insert_tree(Ti, r); }
for each transaction Ti in T
end for
// insert_tree(Ti, r)
if r contains node N & N.item_name
    = T.item_name ;
    then N.count ++;
    else creat new node N;
        N.item - name = p.item - name;
        N.count ++;
        T.parent = r;
        point N.nodek in k to the node
        with the same name
    end if
if T! =  $\emptyset$ ;
    then N.count ++;
    call insert_tree(Ti, r);
end if

```

Pseudo-code of FP-Growth

```

// (FP-Tree,  $\emptyset$ )
If tree contains single path P then
for nodes combination of path P noted
as  $\beta$ 
    do{ create pattern  $\beta \cup \alpha$ ,
        support = min_sup of nodes in  $\beta$  }
else for each  $\alpha_i$  in tree
    do{ create pattern  $\beta = \alpha \cup \alpha_i$ 
        support =  $\alpha_i$ .support
        construct  $\beta$  conditional
        database
        construct  $\beta$  conditional
        Tree $_{\beta}$ 
        if Tree $_{\beta}$ ! =  $\emptyset$ 
            then call FP
            - Growth(Tree $_{\beta}$ ,  $\beta$ )
            end if }
    end for
end if

```

To apply the FP-tree and perform the FP-Growth mining, the steps include (1) checking if the *tree* has a single pass P , if true, then creating pattern $\beta \cup \alpha$ and setting its *Support* counts as the *minsup* count of β ; (2) if false, for each a_i , creating a pattern $\beta = a_i \cup \alpha$ with *support* = a_i .*support*; (3) constructing β conditional tree as $tree_\beta$ and checking if it is not *null*, if true, then calling function $FP\text{-}Growth(tree_\beta, \beta)$. The FP-Growth algorithm only scans the dataset twice when creating the FP-tree for being utilized to store the information(Koh & Shieh, 2004), which avoids repeated scans in the Apriori algorithm for larger datasets. The inputs of the FP-Growth algorithm include all relevant air pollutant concentration records and the preset *minsup* to be finalized through multiple test runs.

While the mining results of the Apriori and FP-Growth algorithms are the same(Xin, Han, Yan, & Cheng, 2005), the FP-growth algorithm runs faster than the Apriori algorithm when the settled *minsup* is under a specific range. If the *minsup* is relatively small, it would be more efficient to use the Apriori algorithm. In this research, both algorithms are utilized in the analysis depending on the *minsup* threshold.

3.4 Machine Learning and Forecasting

Machine learning is capable to process a large amount of data including multiple attributes and making accurate predictions through a robust model. In this research, the meteorological measurements and the average traffic speed will be utilized as attributes to predict the corresponding air pollution concentration levels. Seven most representative machine learning algorithms are considered to build the forecasting models, and two models are selected for each air pollutant. The historical data from the year 2011 to 2019

will be imported to train the selected models that perform the 2020 air pollution prediction. The actual air pollution measurements from the year 2020 will be used to evaluate the prediction. As an innovation of this research, the temporal variances of each influencing parameter are considered to further improve the forecasting accuracy.

3.4.1 Machine Learning Forecasting Concepts

The concept of air pollution forecasting using machine learning models involves using meteorological measurements and traffic situations to predict the real-time air pollutants' concentration through a certain time frame. In this research, eight attributes are considered, which include the hourly solar radiation, temperature, pressure, precipitation, relative humidity, resultant wind speed, wind direction, and annual average traffic speeds on the nearby highway from the year 2011 to the year 2020. Four classes are expected that includes the hourly ground-level ozone concentration, PM_{2.5} concentration, NO₂ concentration, and NO_x concentration.

Machine learning models are able to learn automatically and perform self-improvement by using the data. With this ability, the more data that is fed to the machine learning models, the more accurate the forecasting results because it will enable the models to find even subtle patterns in the data and to use the patterns it identified to make better decisions. The machine learning prediction flow chart can be briefly described in Figure 5.

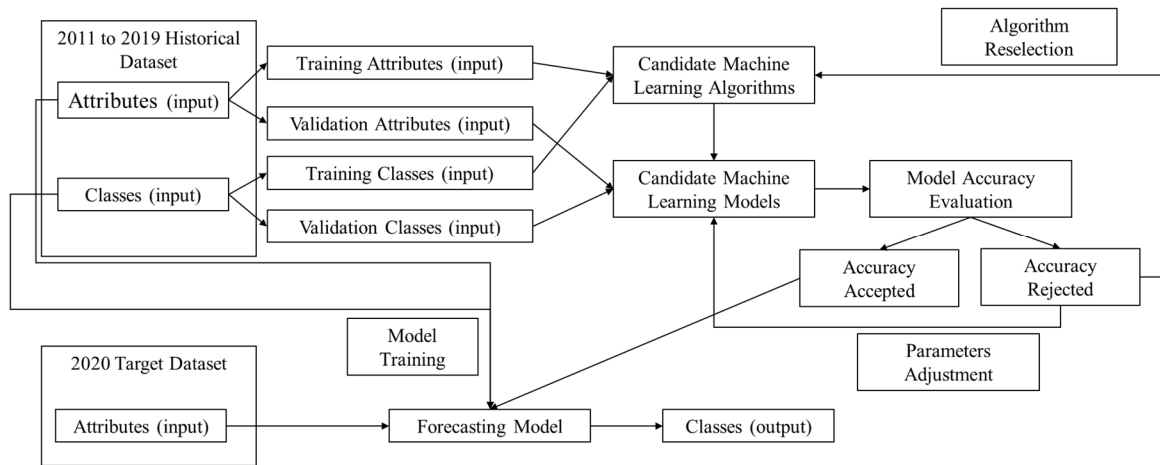


Figure 5. Forecasting by Machine Learning

As shown in Figure 5, the datasets that are preprocessed are separated into two subsets, the first one is the historical dataset that contains the data from the year 2011 to 2019, the second one is the target dataset that contains the data from the year 2020. The historical dataset is used to perform the model selection and model training, the target dataset is used to perform the forecasting. In addition, the historical dataset is further divided into the training dataset and validation dataset that both include attributes and classes. The candidate machine learning algorithms are trained by the training dataset, and the accuracy of the models that are built are evaluated by the validation dataset. If the accuracy is not accepted, then the parameters of the models will be adjusted or the algorithm will be rejected. If the accuracy is accepted, then the target dataset will be fed to the forecasting model, and the prediction results will be provided.

3.4.2 Machine Learning Models

The prediction using these inputs to achieve the outputs can be fulfilled by supervised machine learning. Due to the input data being continuous data types rather than discrete values, regressions algorithms are used in this research. Seven most representative supervised machine learning regression algorithms are utilized in this research, which includes Polynomial Regression (PR) algorithm, Multilayer Perceptron (MLP) algorithm, XGBoost (XGB) algorithm, Support Vector Machine (SVM) algorithm, Random Forest (RF) algorithm, Linear Regression (LR) algorithm, and K-Nearest Neighbors (KNN) algorithm.

3.4.2.1 Linear Regression Algorithm

The model selection process starts from one of the most widely used predictive models and is usually the first type of regression model to be analyzed, which is linear regression. LR model analyzes the linear relationship between the dependent and independent variables. When there is only a single independent variable to be considered, the model is called simple linear regression. On the contrary, when multiple independent variables are considered, that is multiple linear regression. The number of independent variables in this research is eight, thus, multiple linear regression should be used, which can usually be demonstrated by the function shown in Equation 5 (Pedregosa et al., 2011).

$$y_i = \beta_0 + \beta_1 * x_{i,1} + \beta_2 * x_{i,2} + \dots + \beta_n * x_{i,n} + \varepsilon_i \quad (5)$$

where,

n : numbers of the independent variables, 8 in this research,

i : numbers of records, 73,236 in this research,

$x_{i,n}$: independent variables/ attributes,

y_i : dependent variable/ classes,

β_0 : intercept,

β_n : coefficient of x ,

ε_i : error.

The objective of LR is to fit the line to the value of y by a given value of x by finding the best β values that the line fits the data best. Equation 5 can be further transformed into matrix form as shown in Equation 6.

$$y = X \cdot \beta + \varepsilon \quad (6)$$

where,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{pmatrix}, X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_i^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i,1} & \cdots & x_{i,n} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \end{pmatrix}.$$

One of the well-developed algorithms to solve linear regression problems is called the least-squares estimation algorithm, which aims to minimize the sum of the mean squared loss. The least-squares estimation algorithm can be calculated by Equation (7, 8).

$$\text{Find } \beta \text{ that } \min \sum_i (\beta \cdot x_i - y_i)^2$$

$$\begin{aligned}
&= \|X \cdot \beta - y\|^2 = (X \cdot \beta - y)^T (X \cdot \beta - y) \\
&= y^T \cdot y - y^T \cdot X \cdot \beta - \beta^T \cdot X^T \cdot y + \beta^T \cdot X^T \cdot X \cdot \beta \quad (7)
\end{aligned}$$

Thus, when

$$\begin{aligned}
&\frac{\partial(y^T \cdot y - y^T \cdot X \cdot \beta - \beta^T \cdot X^T \cdot y + \beta^T \cdot X^T \cdot X \cdot \beta)}{\partial(\beta)} \\
&= -2X^T \cdot y + 2X^T \cdot X \cdot \beta = 0 \quad (8)
\end{aligned}$$

And can be achieved by:

$$\beta = (X^T \cdot X)^{-1} \cdot X^T \cdot y$$

3.4.2.2 Polynomial Regression Algorithm

Polynomial regression analyzes the relationship between the independent variable and the dependent variable to a certain degree. The polynomial regression can be considered as a special case of linear regression that the data are fitted on a curve. The polynomial regression model can be built as Equation (9, 10) (Pedregosa et al., 2011).

$$y_i = \beta_0 + \beta_1 * x_i + \beta_2 * x_i^2 + \beta_3 * x_i^3 + \dots + \beta_n * x_i^n + \varepsilon_i \quad (9)$$

And in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ 1 & x_3 & x_3^2 & \dots & x_3^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_i & x_i^2 & \dots & x_i^n \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_i \end{bmatrix} \quad (10)$$

Then, vector β can be calculated by:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} = \left(\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ 1 & x_3 & x_3^2 & \cdots & x_3^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_i & x_i^2 & \cdots & x_i^n \end{bmatrix}^T \cdot \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ 1 & x_3 & x_3^2 & \cdots & x_3^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_i & x_i^2 & \cdots & x_i^n \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ 1 & x_3 & x_3^2 & \cdots & x_3^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_i & x_i^2 & \cdots & x_i^n \end{bmatrix}^T \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} \quad (11)$$

where,

y_i : the dependent variable/ classes,

x_i : the explanatory or independent variable/ attributes,

n : degree of the polynomial regression,

β_n : weight parameters of the equation,

ε : random error.

In the analysis, the degree of the polynomial regression is determined by test runs to avoid overfitting.

3.4.2.3 Multilayer Perceptron Algorithm

The multilayer perceptron (MLP) algorithm is a kind of supervised learning that can learn nonlinear function approximators through a training dataset. It can be inferred

from the name, there can be one or more nonlinear layers that are called hidden layers between the input and the output layer.

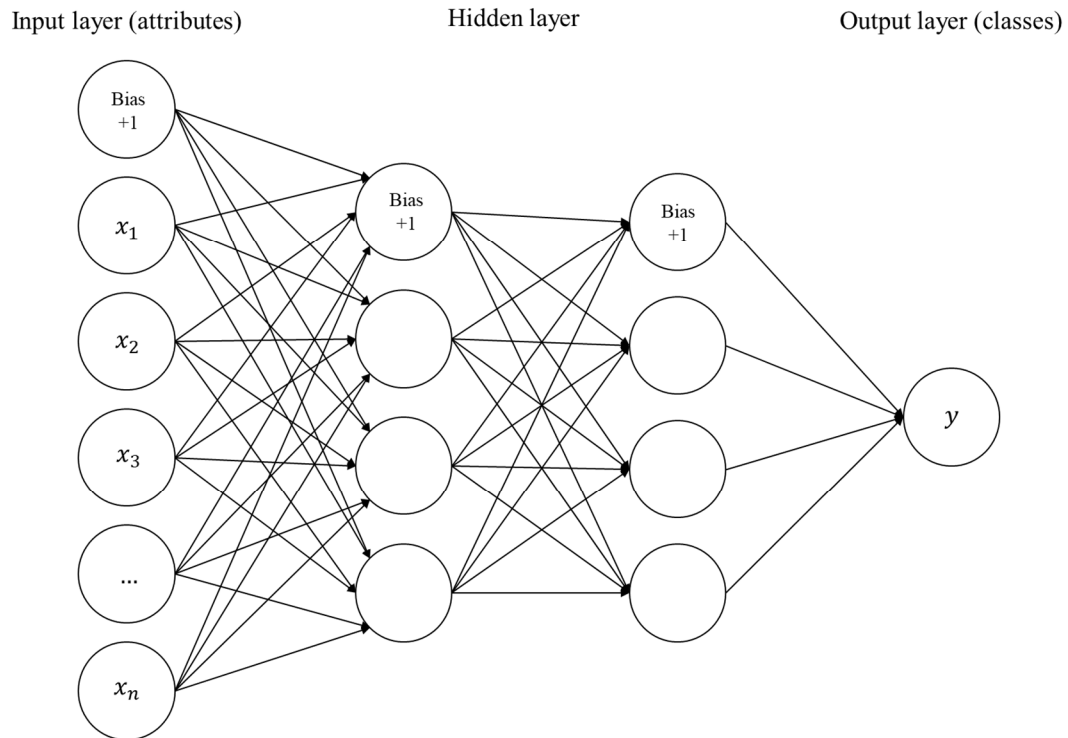


Figure 6. Layers and Structures of MLP

As shown in Figure 6, the left layer is the input layer, the right layer is the output layer, and the layers in between are the hidden layers. $x_1, x_2, x_3, \dots, x_n$ represent the input attributes, and $n = 8$ in this research. The neurons in the hidden layers are transformed from the values of previous layers by a weighted linear summation and a nonlinear activation function as shown in Equation (12) (Pedregosa et al., 2011).

$$\emptyset(w_1x_1 + w_2x_2 + \dots + w_nx_n) \quad (12)$$

where w_1, w_2, \dots, w_n are the weights of each attribute, and \emptyset is the nonlinear activation function that is widely used as:

$$\text{hyperbolic tangent } \phi = \tanh(a), \quad (13)$$

$$\text{logistic sigmoid } \phi = 1/(1 + \exp(-a)) \quad (14)$$

The hyperbolic tangent and logistic sigmoid activation functions have the relationship as:

$$(\tanh(a) + 1)/2 = 1/(1 + \exp(-a)) \quad (15)$$

The learning process of the MLP algorithm is performed in the perceptron involving changing connection weights based on the error. The degree of error is the difference between the target value d and the perceptron generated value y in an output node j in the n th data point, which can be represented by (Pedregosa et al., 2011):

$$e_j(n) = d_j(n) - y_j(n) \quad (16)$$

The objective is to minimize the error in the entire outputs:

$$\min \varepsilon(n) = 1/2(\sum_j e_j^2(n)) \quad (17)$$

By gradient descent,

$$\begin{aligned} \Delta w_{ji}(n) &= -\eta \frac{\partial \varepsilon(n)}{\partial v_j(n)} y_i(n) = -\eta (e_j(n) \phi'(v_j(n))) y_i(n) \\ &= -\eta (\phi'(v_j(n)) \sum_k (-\frac{\partial \varepsilon(n)}{\partial v_k(n)} w_{kj}(n))) y_i(n) \end{aligned} \quad (18)$$

where,

y_i : output of the previous neuron,

η : learning rate that is predetermined,

v_j : induced local field,

ϕ' : derivative of the activation function.

3.4.2.4 XGBoost Algorithm

The XGBoost means the Extreme Gradient Boosting, which is developed based on the gradient boosting library that is higher in efficiency and flexibility. The boosting is a term in learning technique that is capable to build a strong classifier and regressor and controls both bias and variance. Various boosting algorithms are developed such as adaptive boosting (AdaBoost), which was the first boosting algorithm aiming to perform binary classification; gradient boosting, which is one of the most powerful predictive models technique that minimizes the loss function by adding weak learners; and XGBoosting (Sreekanth, 2020).

The training of the XGBoost algorithm is to optimize the objective function like all supervised learning algorithms. The objective function of the XGBoost for each x in the dataset can be written as (J. Li, 2017):

$$\frac{\partial L(y, f^{m-1}(x) + f_m(x))}{\partial f_m(x)} = 0 \quad (19)$$

After the Taylor expansion around the current estimate $f^{m-1}(x)$:

$$\begin{aligned} L(y, f^{m-1}(x) + f_m(x)) \\ \approx L(y, f^{m-1}(x)) + g_m(x)f_m(x) + \frac{h_m(x)f_m(x)^2}{2} \end{aligned} \quad (20)$$

where,

$g_m(x)$: the gradient as the gradient boosting algorithm,

$$h_m(x) = \frac{\partial^2 L(Y, f(x))}{\partial f(x)^2}, f(x) = f^{m-1}(x) \quad (21)$$

while the loss function is:

$$\begin{aligned} L(f_m) &\approx \sum_{i=1}^n [g_m(x_i) f_m(x_i) + h_m(x) f_m(x)^2 / 2] + \text{constant} \\ &\propto \sum_{j=1}^{T_m} \sum_{i \in R_{jm}} [g_m(x_i) w_{jm} + h_m(x) w_{jm}^2 / 2] \end{aligned} \quad (22)$$

$$L(f_m) \propto \sum_{j=1}^{T_m} [G_{jm} w_{jm} + H_{jm} w_{jm}^2 / 2] \quad (23)$$

$$L(f_m) \propto -\frac{1}{2} \sum_{j=1}^{T_m} [G_{jm}^2 / H_{jm}] \quad (24)$$

which is the structure of the tree that the smaller, the better, where,

G_{jm} : sum of $g_m(x)$ in region j ,

H_{jm} : sum of $h_m(x)$ in region j , and

$$w_{jm} = -\left(\frac{G_{jm}}{H_{jm}} \right), j = 1, \dots, T_m \quad (25)$$

When splitting a leaf into two, the score it gains is calculated by ("Introduction to Boosted Trees," 2020):

$$\begin{aligned} \text{Gain} &= \frac{1}{2} \left[\frac{G_{jmL}^2}{H_{jmL}} + \frac{G_{jmR}^2}{H_{jmR}} - \frac{G_{jm}^2}{H_{jm}} \right] \\ &= \frac{1}{2} \left[\frac{G_{jmL}^2}{H_{jmL}} + \frac{G_{jmR}^2}{H_{jmR}} - \frac{(G_{jmL} + G_{jmR})^2}{(H_{jmL} + H_{jmR})} \right] \end{aligned} \quad (26)$$

When considering the regularization, the loss function can be written as:

$$\begin{aligned} L(f_m) &\propto \sum_{j=1}^{T_m} [G_{jm} w_{jm} + \frac{H_{jm} w_{jm}^2}{2}] + \gamma T_m + \frac{1}{2} \lambda \sum_{j=1}^{T_m} w_{jm}^2 + \alpha \sum_{j=1}^{T_m} |w_{jm}| \\ &= \sum_{j=1}^{T_m} [G_{jm} w_{jm} + \frac{H_{jm} w_{jm}^2}{2} + \alpha |w_{jm}|] + \gamma T_m \end{aligned} \quad (27)$$

where,

γ : penalization term on the number of terminal nodes,

α : regularization of L_1 ,

λ : regularization of L_2 . When the optimal weight for region j is defined as:

$$w_{jm} = \begin{cases} -\left(\frac{G_{jm} + \alpha}{H_{jm} + \lambda}\right), & G_{jm} < -\alpha \\ -\left(\frac{G_{jm} - \alpha}{H_{jm} + \lambda}\right), & G_{jm} > \alpha \\ 0, & \text{else} \end{cases} \quad (28)$$

The gain of each split is then calculated as:

$$\begin{aligned} \text{Gain} &= \frac{1}{2} \left[\frac{T_\alpha(G_{jmL})^2}{(H_{jmL} + \lambda)} + \frac{T_\alpha(G_{jmR})^2}{(H_{jmR} + \lambda)} - \right. \\ &\quad \left. \frac{T_\alpha(G_{jm})^2}{(H_{jm} + \lambda)} \right] - \gamma \end{aligned} \quad (29)$$

where,

$$T_\alpha(G) = \begin{cases} G + \alpha, & G < \alpha \\ G - \alpha, & G > \alpha \\ 0, & \text{else} \end{cases} \quad (30)$$

After the boosting tree is constructed by the equations above, the XGBoost algorithm is fully developed.

3.4.2.5 Support Vector Machine Algorithm

The Support Vector Machine (SVM) is a type of supervised machine learning algorithm that is capable of both classification and regression (Cortes & Vapnik, 1995). The SVM was developed based on statistical learning frameworks and robust in prediction by linear and nonlinear models. The advantages of SVM include high effectiveness on high dimensional spaces and when the number of dimensions is greater than that of samples, memory efficient and versatile and customizable Kernel functions. However, some disadvantages of SVM such as over-fitting may present (Pedregosa et al., 2011). In this research, the SVM regressor (SVR) is utilized to perform the air pollution prediction in the model selection section, in which the primal problem is presented in Equation (31, 32) (Pedregosa et al., 2011).

$$\min_{w,b,\zeta,\zeta^*} \left(\frac{1}{2} w^T w + C \sum_{i=1}^n (\zeta_i + \zeta_i^*) \right) \quad (31)$$

$$\text{subject to: } y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i, \quad (32)$$

$$w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*,$$

$$\zeta_i, \zeta_i^* \geq 0, i = 1, \dots, n$$

And for Linear SVR, the primal problem can be formulated as:

$$\min_{w,b} \left(\frac{1}{2} w^T w + C \sum_{i=1}^n \max(0, |y_i - (w^T \phi(x_i) + b)| - \varepsilon) \right) \quad (33)$$

where (Shalev-Shwartz, Singer, Srebro, & Cotter, 2011),

$$i=1, \dots, n,$$

w : normal vector to the hyperplane $w^T \phi(x) + b$,

b : intercept,

ϕ : identity function,

$\zeta_i = \max(0, 1 - y_i(w^T x_i - b))$: the smallest nonnegative number satisfying the equation,

C : penalty term that controls the strength of the penalty.

The predictions that are at least ε away from the true target value will be penalized by ζ_i, ζ_i^* depending on the predictions lie above or below the ε range (Pedregosa et al., 2011).

The dual problem can be written as:

$$\min_{\alpha, \alpha^*} \left(\frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \varepsilon e^T (\alpha + \alpha^*) - y^T (\alpha - \alpha^*) \right) \quad (34)$$

$$\text{subject to: } e^T (\alpha + \alpha^*) = 0, \quad (35)$$

$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \dots, n$$

where,

e : all ones vector,

Q : n by n semidefinite matrix that by the kernel $Q_{ij} \equiv K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$,

α_i : dual coefficients that are bounded by zero and C ,

And the prediction can be calculated by:

$$\sum_{i \in SV} (\alpha - \alpha^*) K(x_i, x) + b \quad (36)$$

3.4.2.6 Random Forest Algorithm

The Random Forest (RF) algorithm is a type of ensemble method that combines the prediction of several base estimators, which have several improvements in generalizability and robustness (Pedregosa et al., 2011). RF is one of the most used algorithms that's capable of classification and regression (Donges, 2020). Numbers of decision trees are built when constructing the RF algorithm, and the disadvantages of decision trees such as overfitting can be avoided by increasing the number of trees (Friedman, Hastie, & Tibshirani, 2001; "Random Forest Algorithm," 2020).

Two phases are involved in creating the RF algorithm, which include the decision tree creation phase and the tree prediction phase. By implementing these two phases, a number of random data points from the training set need to be selected to build the decision trees and repeat, find the predictions of each decision tree and assign the new data points to the category that have the majority votes ("Random Forest Algorithm," 2020). The node importance of each decision tree is calculated by Gini importance (Ronaghan, 2018):

$$i_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (37)$$

where,

i_j : the importance of node j ,

w_j : weight of samples reaching node j ,

C_j : the impurity value of node j , which is calculated by Mean Square Error variance reduction,

$left(j)/right(j)$: child node from left/right split on node j .

The importance for each feature on a decision tree is calculated as (Ronaghan, 2018):

$$f_i = \frac{\sum_{j: \text{node } j \text{ splits on feature } i} i_j}{\sum_{k \in \text{all nodes}} i_k} \quad (38)$$

where,

f_i : the importance of feature i .

i_j : the importance of node j .

The feature importance values are then be normalized to values between 0 and 1 by Equation (39):

$$\text{norm } f_i = \frac{f_i}{\sum_{j \in \text{all features}} f_j} \quad (39)$$

The final feature importance of the random forest is the average of all trees:

$$RF \ f_i = \frac{\sum_{j \in \text{all trees}} \text{norm } f_{ij}}{T} \quad (40)$$

where,

$RF \ f_i$: the importance of feature i of all trees in the model,

f_{ij} : normalized importance for feature i in tree j ,

T : total number of trees.

The features' importance is then calculated and decision trees are built, and the results of all trees are collected by the random forest algorithm to make the final prediction decision.

3.4.2.7 K-Nearest Neighbors Algorithm

The K-Nearest Neighbors algorithm (KNN) was first introduced in 1951 and developed since then (Altman, 1992; Fix, 1985) that can be used in both classification and regression problems, which processes the input of k closest training examples and output the property value of the project that is the average of the k nearest neighbors values. The algorithm finds and predicts the label of a predefined number of training samples that the distance to the new point is closest. Several distance measurements can be used include Euclidean distance, Manhattan distance, and Hamming distance. The standard Euclidean distance is the most common choice (Jaskowiak & Campello, 2011), which can be calculated by Equation (41) (Pandey, 2021):

$$dist((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (41)$$

where,

$(x_1, y_1), (x_2, y_2)$: the coordinates of two points that the distance is calculated.

Or, for more data points, the Euclidean distance can be calculated as (Zakka, 2016):

$$dist(x, y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (42)$$

While the Manhattan distance can be calculated as:

$$dist(x, y) = \sum_{i=1}^m |x_i - y_i| \quad (43)$$

The seven machine learning algorithms will be tested and evaluated based on ten-fold cross-validation, which splits the dataset into ten parts with nine for training and one for the test, repeating for all combinations (Brownlee, 2019). The ten-fold cross-validation is a representative k-fold cross-validation that the k value is preset as 10, which is normally

used to evaluate the performance of a machine learning model on unseen data. The k -fold cross-validation technique is able to predict how the model fit. A random seed was given to the dataset to ensure that each algorithm is evaluated on the same dataset splits.

3.4.3 Machine Learning Model Evaluation

There are several parameters that can be used for machine learning model evaluation. For classification models that are used on discrete data types, the confusion matrix and its associated scores are commonly used to evaluate the models.

TABLE 9. CONFUSION MATRIX FOR CLASSIFICATION MODELS

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

TABLE 9 shows the confusion matrix for classification models. For actual measurement and predicted values, there are positive and negative categories. When the prediction belongs to the same class as the actual measurement, it is the true positive (TP). When the prediction does not belong to the class and the actual measurement does not belong to the class either, it is the true negative (TN). When the prediction belongs to a class but the actual measurement does not, it is the false positive (FP). When the prediction does not belong to a class but the actual measurement does, it is the false negative (FN) (Muskan, 2020). From the confusion matrix, the accuracy score that provides the

percentage of right predictions can be calculated as the number of correct predictions divided by the total number of predictions.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (44)$$

Other commonly used classification model evaluation matrices include recall that shows the percentage of correctly predicted positive out of all positives; precision that shows the percentage of correctly predicted positive out of all predicted positives; F1 score that is the harmonic mean of the model's precision and recall (Olson & Delen, 2008).

$$Recall = \frac{TP}{TP+FN} \quad (45)$$

$$Precision = \frac{TP}{TP+FP} \quad (46)$$

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (47)$$

Several plots can also be used to evaluate the performances of classification models such as the receiver operating characteristic (ROC) curve and the area under the curve (AUC) that plots the relationship between the false positive rate and the true positive rate.

Unlike the classification model evaluation, different metrics are used for regression model evaluation. One of the most common metrics for regression evaluation is the mean squared error (MSE) that is the average of the squared difference between the prediction and actual measurement, and the root means squared error (RMSE) is the square root of MSE.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (48)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (49)$$

The mean absolute error (MAE) is also used for regression model evaluation, which is the average of the absolute difference between the prediction and actual measurements.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (50)$$

In this research, the concentration records of different air pollutants are using different units and scales, for example, PM_{2.5} concentrations are measured in ug/m³ and ground-level ozone concentrations are measured in ppb. As one of the results, the prediction accuracy of one air pollutant cannot be compared with other air pollutants prediction accuracies by the RMSE and MAE values. To present the relative error percentage that can be used for inter-comparison instead of the error value, the normalized root means squared error is commonly used on regression model evaluation, which can be compared between datasets and models with different scales (Nash & Sutcliffe, 1970; Ris, Holthuijsen, & Booij, 1999; Willmott et al., 1985).

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \quad (51)$$

There are no consistent means of normalization when computing NRMSE, it can be different for different research. For example, the tidal range is normally used to normalize water levels, and offshore wave height can be used to normalize wave heights. The range for normalization in this research utilizes the maximum and minimum values that are collected, in which the outliers are not counted (CIRP, 2020).

Other than the NRMSE, the prediction vs. actual (PVA) plot is utilized to analyze the regression predictions, which is a type of scatter plot that is one of the richest forms of

data visualization (Piñeiro, Perelman, Guerschman, & Paruelo, 2008). For a good prediction, the model should be aiming to have prediction = actual, thus, the line $y=x$ will be drawn in the PVA plot to show how much the predictions deviated from actual measurements.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Temporal Characteristics

With the hourly air pollution concentration and meteorological data collected, the annual, monthly, daily, and hourly temporal characteristics are analyzed. The temporal characteristics of the transportation situation in terms of annual average traffic speeds are analyzed separately. Before the analysis of the temporal characteristics, a one-way ANOVA test is performed for air pollution and meteorological parameters between ten years.

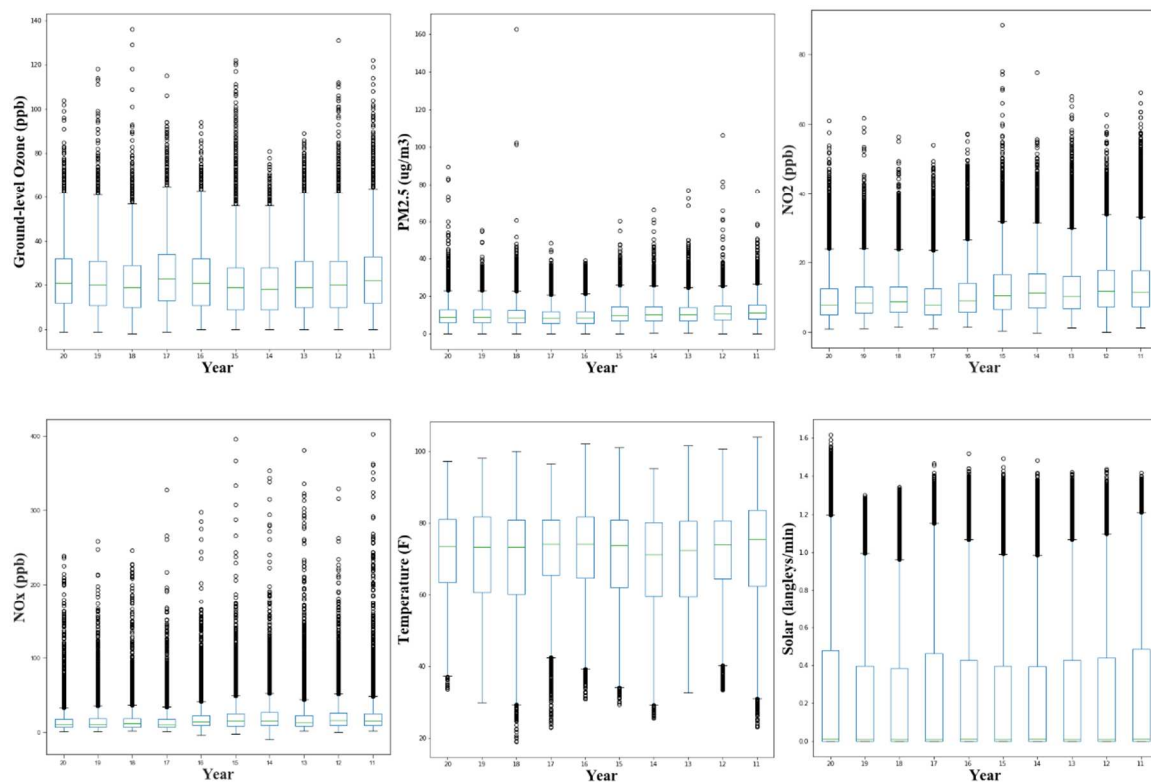
TABLE 10. ONEWAY ANOVA TEST RESULTS THROUGH TEN YEARS

Parameters	P-value
Ground-level Ozone	7.24E-121
PM2.5	0.00
NO2	0.00
NOx	1.60E-172
Temperature	1.87E-139
Solar	1.32E-17
Pressure	1.91E-208
Precipitation	4.81E-08
Relative Humidity	2.24E-291
Resultant Wind Speed	7.83E-182
Wind Direction	7.94E-67

TABLE 10 shows the p -values from the ANOVA test. From the table, the p -values for all parameters are smaller than the significance level of 0.05, which means the measurements for these parameters of each year are statistically significantly different and

comparable. For $PM_{2.5}$ and NO_2 , the p -values are shown as 0.00 in the calculation, which means they are so small and extremely close to zero.

4.1.1 Annual Characteristics of Air Pollution and Meteorological Measurements



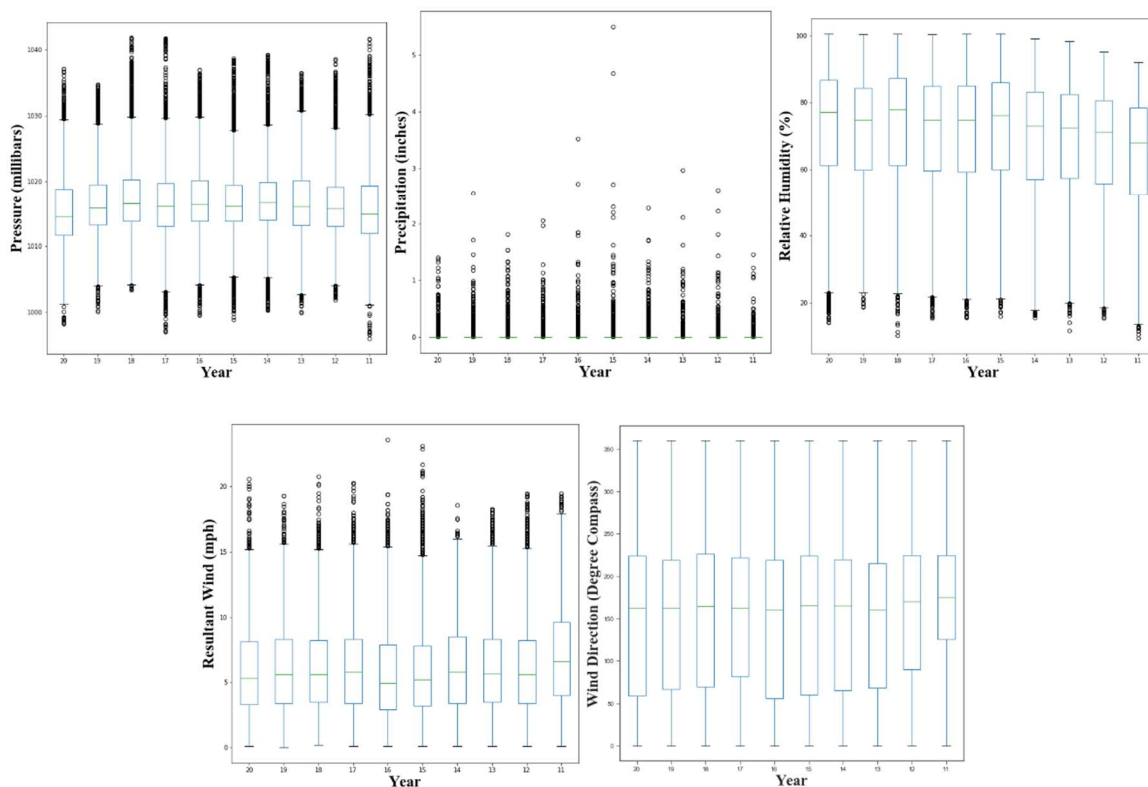


Figure 7. Annual Characteristics

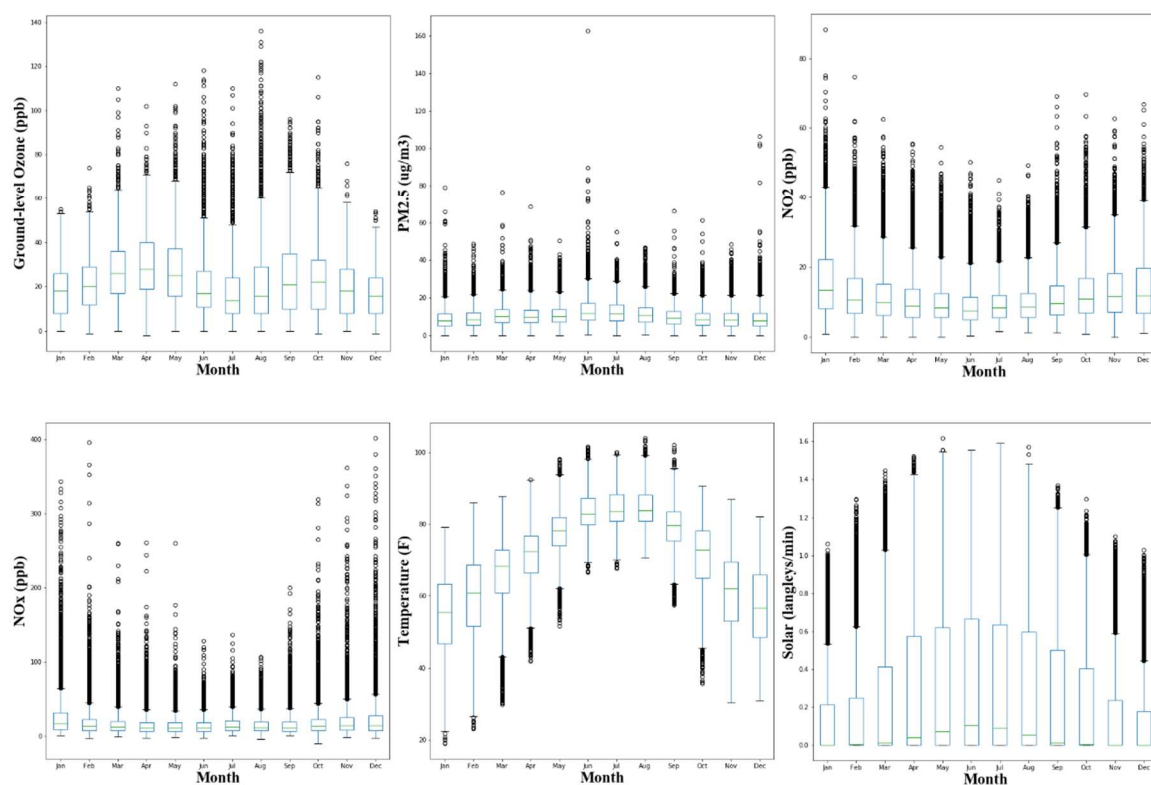
The annual air pollution concentration and meteorological measurements are shown in Figure 7 as box-whisker plots. In the plots, the x-axis shows the year from 2011 to 2020; the y-axis shows the measurements of each category; the boxes show the interquartile ranges (IQR); the upper and lower edges of the boxes show the upper and lower quartiles (Q_3 and Q_1); the upper and lower short bars show the maximum ($Q_3 + 1.5 * IQR$) and minimum ($Q_1 - 1.5 * IQR$); the short green bars in the boxes show the median; the black spots show the outliers. For the precipitation plot, the boxes are in extremely small ranges close to zero, which is a result of the climate nature in Houston area that for most of the days, there is no rain or only drizzle, and when the heavier rains occur, they are shown as the outliers in the plot.

For the annual air pollution concentrations, there is no decreasing or increasing trend between the ten years for ground-level ozone. For all years, the minimums of the concentration are around 0 ppb, the maximums of the concentration are around 60 ppb, the outliers are below 140 ppb. The year 2017 has the highest concentration range, the years 2014, 2015, and 2018 have the lowest concentration ranges. There is no annual decreasing or increasing trend found for PM_{2.5} concentration from the data. For all years, the minimums of the concentration are around 0 ug/m³ and the maximums are around 25 ug/m³. The outliers are below 100 ug/m³ except for the years 2018 and 2012. The year 2015 has the highest PM_{2.5} concentration ranges and the year 2017 has the lowest. The NO₂ concentration is decreased significantly from the year 2016. The pre-2016 NO₂ concentrations have minimums around 0 ppb and maximums around 30 ppb, which are overall higher than the post-2016 (including 2016) concentrations that have minimums around 0 ppb and maximums around 23 ppb. The outliers of pre-2016 NO₂ concentration are also higher than the post-2016 concentration. The annual NO_x concentrations show similar trends with the NO₂ that the pre-2016 concentrations are higher than post-2016 concentrations. However, the outlier ranges for NO_x concentrations are much higher than NO₂ concentrations.

For the annual meteorological measurements, the highest maximum temperature appeared in the year 2011, which is around 110 F, and the lowest minimum temperature appeared in the year 2018, which is around 32 F and some outliers are even below 20 F. The solar radiation for the year 2020 was relatively higher with a maximum of 1.2 langley/min, and that for the year 2018 was relatively lower with a maximum of 1.9 langley/min. The annual outdoor pressures are all in the range of 1000 to 1030 millibars.

From the precipitation and relative humidity measurements, the years 2015 and 2018 are relatively wetter and the year 2011 is relatively dryer. From the resultant wind speed and plots, the year 2011 has the stronger winds and the year 2016 has the weaker winds. From the wind direction plot, the medians of the wind direction of the monitoring site are all around 160-degree compass, which is SSE wind. Most of the wind directions are in the range of 60 to 220-degree compass, which means ENE to SW wind.

4.1.2 Monthly Characteristics of Air Pollution and Meteorological Measurements



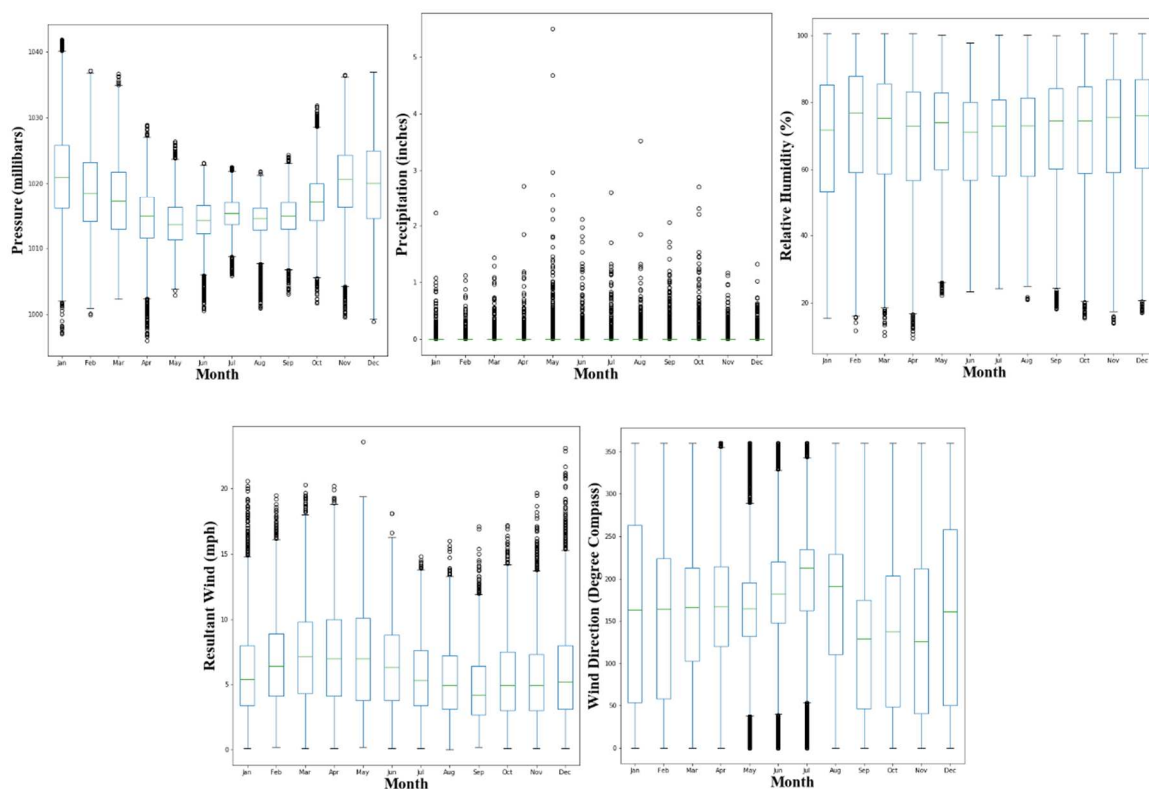


Figure 8. Monthly Characteristics

The monthly characteristics of air pollution concentrations and meteorological measurements are shown in Figure 8 as box-whisker plots. The x -axis shows the month of the years, and the y -axis shows the measurements of each category. For the air pollution concentrations, ground-level ozone concentration shows two significant peaks during a year with a maximum of around 70 ppb, which are April and September. In the meantime, the concentration meets its bottom values with the maximums around 50 ppb during July and December. In another word, for the monitoring site, the ground-level ozone concentration is higher in the spring and fall seasons and lower in the summer and winter seasons. The monthly trend of $PM_{2.5}$ shows that the concentration is higher in June with a maximum of around $30 \mu\text{g}/\text{m}^3$, and lower in January with a maximum of around $20 \mu\text{g}/\text{m}^3$. Thus, $PM_{2.5}$ is higher in summer seasons and lower in winter seasons. On the

contrary, NO₂ concentration is higher in winter seasons and meets its peak during January with the maximum of 43 ppb, and lower in summer seasons and meets its lowest during June with the maximum of 25 ppb. The NO_x concentration shows a similar trend with NO₂. The peak value is found during January with a maximum of 65 ppb, and the lower value is found during June and August with the maximums of around 45 ppb.

From the temperature monthly plot, the hottest month is August with the temperature in a range of 70 F to 100 F, the coolest month is January with the temperature in a range of 23 F to 80 F. From the solar radiation plot, June and July have the highest solar radiation with maximums of around 1.6 langley/min, December has the lowest solar radiation with a maximum of 0.5 langley/min. This shows that temperature and solar radiation have some relationship, but do not have the same trend. The outdoor pressure varies in a large range of 1,002 to 1,040 millibars in January, and in a small range of 1008 to 1,022 millibars in July. This means the outdoor pressure is more stable in summer and relatively unstable in winter. The precipitation is higher from April to October, and lower from November to March. However, the relative humidity is lower from April to August and higher from September to March. This means the relative humidity may be more related to temperature when compared to precipitation. The resultant wind plot shows that the wind season in the Houston area is from March to May, and the wind speed is lower in September. In December and January, the wind direction varies in a larger range, and in June and July, the wind direction varies in a smaller range.

4.1.3 Day-of-Week Characteristics of Air Pollution

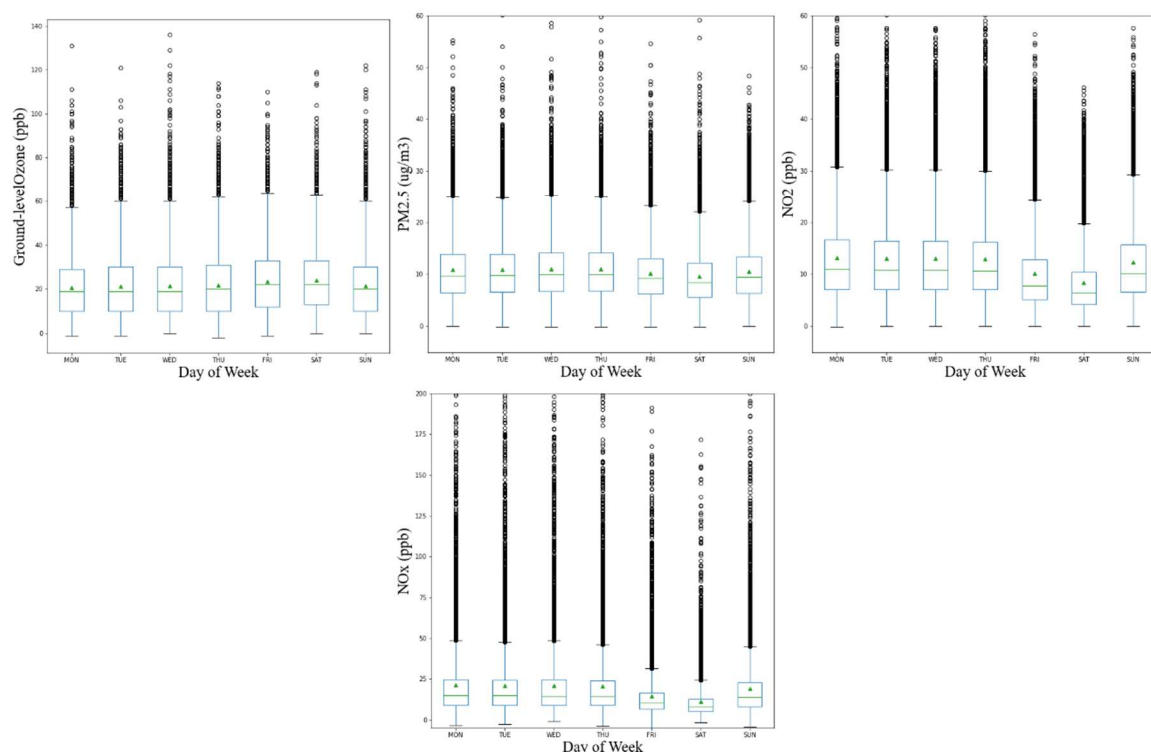


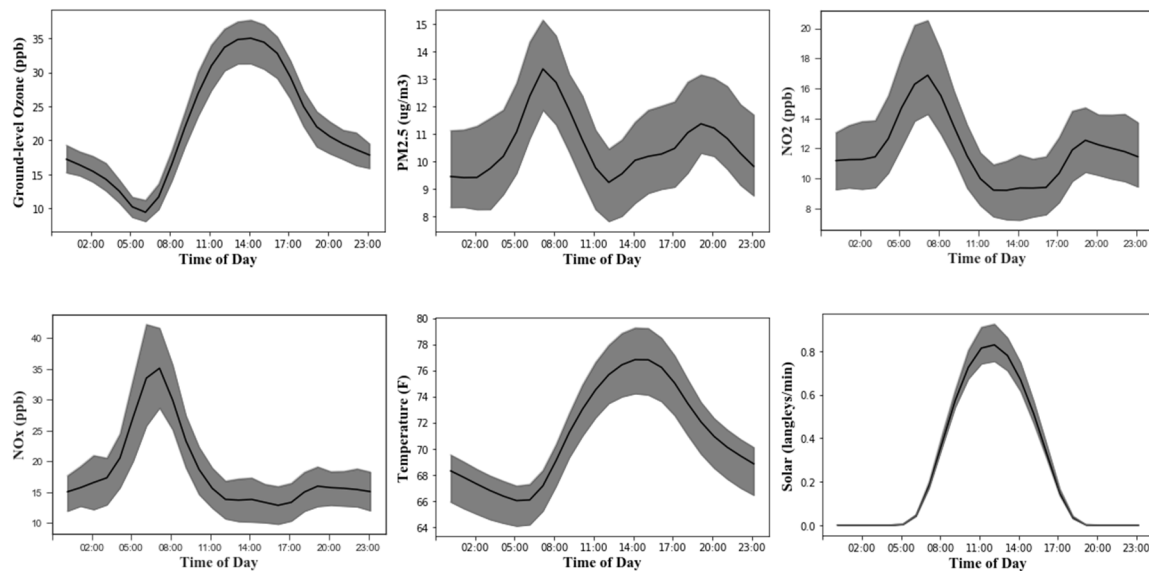
Figure 9. Day of Week Characteristics

Figure 9 shows the box plots of the air pollution concentrations of each day of the week. The x-axis shows the days of weeks, and the y-axis shows the measurements of each category. In addition to the previous box plots, the green solid triangles in the box show the mean values of each category. The PM_{2.5}, NO₂, and NO_x show similar trends in that the concentrations are lower during weekends and higher during weekdays, which meet their lowest value during Saturday. The PM_{2.5} concentration reaches its peak value on Thursday with an average concentration of 9.9 ug/m³ and decreases after that until Saturday, the level then increases on Sunday. The NO₂ concentration reaches its peak value on Monday with an average concentration of 11 ppb and decreases on Tuesday. A minor increase of NO₂ concentration level appears on Wednesday and decreases substantially from Friday and

returns to a relatively high level on Sunday. The NO_x concentration level reaches its peak value on Monday with an average level of 15 ppb and decreases start from Tuesday, which decreases substantially from Friday and returns to a relatively high level on Sunday.

On the contrary, ground-level ozone concentration shows a different trend, which is higher during Friday and Saturday with average concentrations of 22 ppb, and lower from Sunday to Thursday. The ground-level ozone concentration is maintained at a low level on Monday and Tuesday. All these trends show that the concentrations of these air pollutants are related to human activities positively such as $\text{PM}_{2.5}$, NO_2 , and NO_x , and negatively such as ground-level ozone.

4.1.4 Hourly Characteristics of Air Pollution and Meteorological Measurements



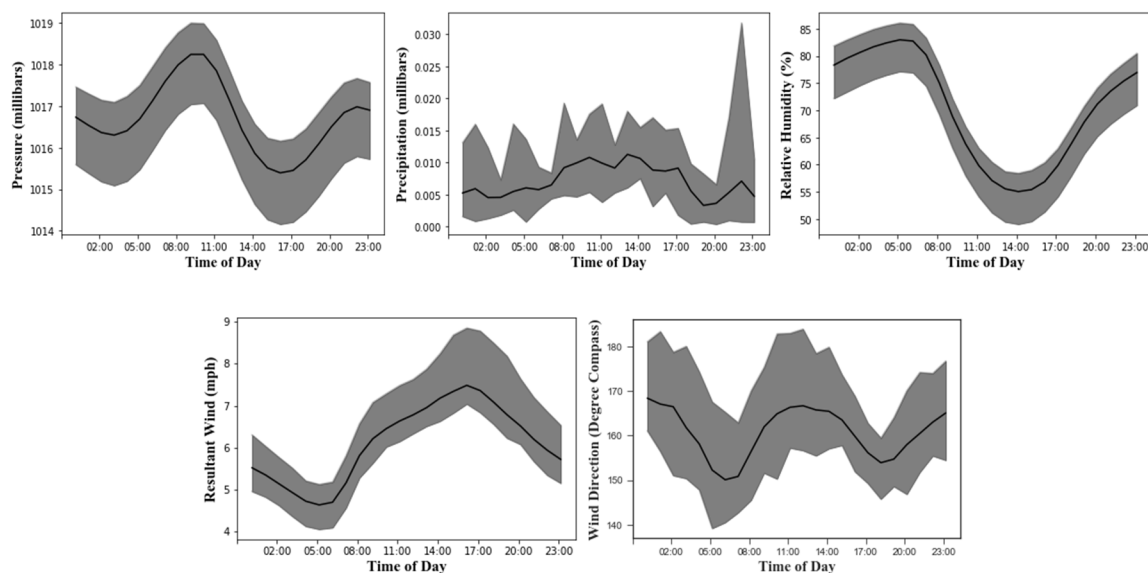


Figure 10. Hourly Characteristics

Figure 10 shows the hourly characteristics of the air pollution concentration and meteorological measurements. The x -axis shows the time of days, and the y -axis shows the measurements of each category. The gray areas show the variances of the average values of each category, and the black solid lines of each plot show the mean values of the variances.

For air pollution, ground-level ozone concentration during a day is lowest at around 7:00 and increases until it reaches its peak at around 14:00, then decreases. The variance in ground-level ozone throughout a day may range from 18 to 30 ppb. There are two peaks for $PM_{2.5}$ concentration in a day, which are the higher peak at around 6:30 and the lower peak at around 19:00. The $PM_{2.5}$ concentration reaches its lowest values at around 2:00 and 12:00. The variance in $PM_{2.5}$ concentration throughout a day may range from 1 to 9 $\mu g/m^3$. The NO_2 concentrations also have two peaks in a day, which are higher peak at around 7:00 and lower peak at around 19:00. This character is similar to the $PM_{2.5}$ concentration.

The lowest value of NO₂ concentration is reached from 11:00 to 16:00. The variance in NO₂ concentration throughout a day may range from 2 to 17 ppb. The NO_x concentrations in a day have a peak value at around 7:00 and the lowest value is reached between 15:00 to 17:00. The overall NO_x concentrations from 12:00 to the following day 3:00 are relatively stable. The variance in NO_x concentration throughout a day may range from 10 to 33 ppb.

For meteorological measurements, the temperature reaches its highest at around 14:00 and lowest around 6:00. The variance of temperature throughout a day may range from 6 to 15 F. The solar radiation rises from 5:00 and arrives at its peak at 12:30, which then decreases through 19:00 and reaches its lowest value. The outdoor pressure has two peaks, which are the higher peak at around 9:00 and the lower peak at around 21:30. The lowest pressure measure appears at around 16:00 during the day. The variance of outdoor pressure may range from 0.5 to 4.5 millibars. The precipitation during a day has no obvious trends. However, the average precipitations from 10:00 to 17:00 and 22:00 are higher based on the plot. The relative humidity is highest at 6:00 and lowest at 14:00, which is negatively related to temperature. The variance of humidity may range from 15 to 36 percent. The resultant wind speed is highest at 16:00 and lowest at 5:30, which is positively related to temperature. The variance of resultant wind speed may range from 1.7 to 4.4 mph. The wind direction plot shows that at around 12:00 and 24:00 the wind directions are closer to 170-degree compass, which means the wind comes from the south (S). At around 6:00 and 18:00, the wind directions are closer to 150 to 165-degree compass, which means SSW wind.

4.1.5 Temporal Characteristics of Traffic Speed

To consider the influences from human activities on the air pollution level, the annual average traffic speed of the nearby freeway is analyzed.

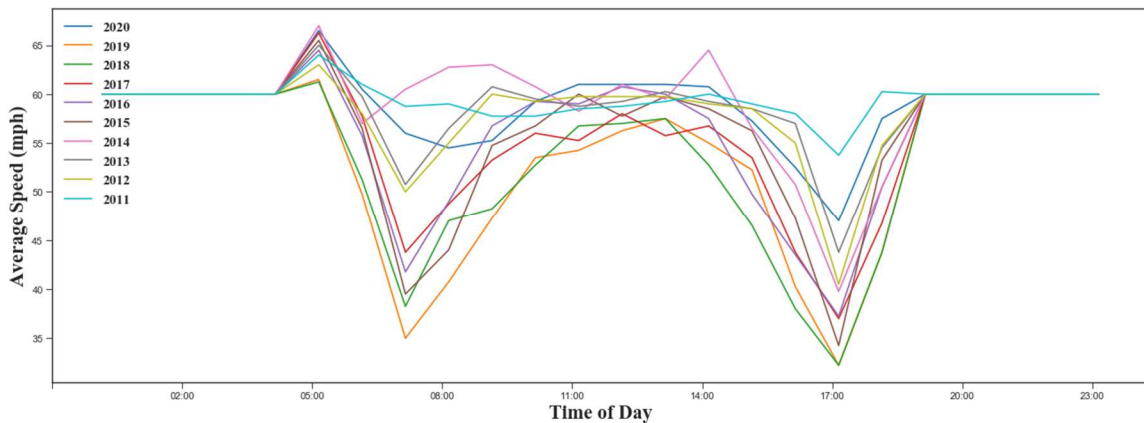


Figure 11. Annual Average Traffic Speed Profile

In Figure 11, the *x*-axis shows the time of the day and the *y*-axis shows the annual average traffic speed in mph. The traffic speed of each year is presented as colored lines in the plot as noted in the legend on the upper left corner. As shown in the figure, the average traffic speed of each year shows a similar trend. For all years, the traffic speeds have two significant low points, which are directly related to the rush hours of Houston. The first one is the morning rush hours around 6:00 to 9:00, the second one is the afternoon rush hours around 16:00 to 19:00. There is a peak value before the morning rush hours when the average traffic speed is even more than the speed limit. For the year 2014, there is another peak value before the afternoon rush hours that is higher than the speed limit. From Figure 11, the years 2019 and 2018 have the lowest average traffic speeds, which the speeds may be lower than 40 mph during rush hours. The years 2013 and 2020 have the relatively

highest average traffic speeds, which the morning rush hour speeds can be higher than 55 mph, and the afternoon rush hour speeds can be higher than 45 mph.

From the above temporal analysis, the relationship between air pollution concentration levels and their influencing factors can be inferred.

(1) The ground-level ozone and $PM_{2.5}$ concentration levels are not influenced by different years, however, the NO_2 and NO_x levels are lower since the year 2016, which may be related to changing human activity patterns or pollution reduction technologies due to the weather conditions do not have significant changing trends.

(2) The monthly/seasonal temporal patterns for all air pollutants show that ground-level ozone concentration is higher in spring and fall, $PM_{2.5}$ concentration is higher in summer, NO_2 and NO_x concentrations are higher in winter, which is a sign that these air pollutants are influenced by meteorology situations throughout a year.

(3) The day-of-week air pollution concentration patterns are almost the same for $PM_{2.5}$, NO_2 , and NO_x , which are opposite to the ground-level ozone concentration pattern. This means that increased human activities such as working and commuting on weekdays may result in higher $PM_{2.5}$, NO_2 , and NO_x concentration levels, however, ground-level ozone concentration levels may be negatively impacted by the intensity of human activities.

(4) The hourly temporal characteristics imply that the ground-level ozone concentration may be positively related to outdoor temperature and solar radiation and may be negatively related to the pressure. However, the $PM_{2.5}$, NO_2 , and NO_x concentrations show limited relationships with the meteorology status while they are significantly influenced by the time of the day.

(5) When considering the monitoring site's nearby traffic situation, it can be implied that the ground-level ozone concentration has only a limited relationship with the transportation activities. While the $PM_{2.5}$, NO_2 , and NO_x concentrations are negatively related to the average traffic speeds on the nearby highway and reach their peak values during the traffic rush hours.

To further discover the relationship between and inter air pollution concentration, meteorology measurements, and traffic speed, a correlation analysis is conducted.

4.2 Correlation Analysis

To determine the relationships between the total 12 parameters that are considered in this research, an interrelationship and distribution chart is created.

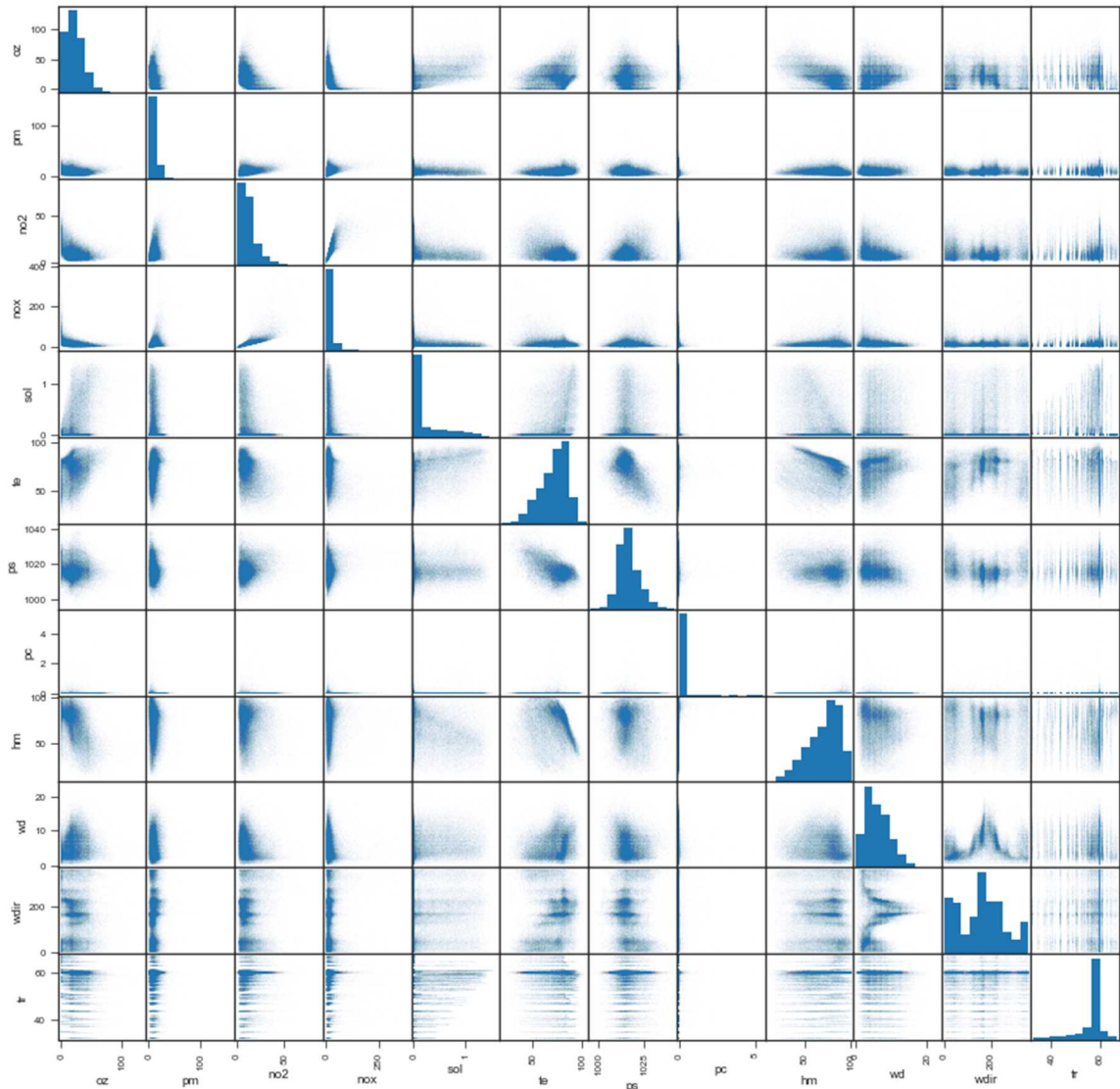


Figure 12. Interrelationship Matrix and Distribution Chart for All Parameters

Figure 12 shows the interrelationship scatter plot matrix between each parameter and the distribution of the values. In this figure, both the x -axis and y -axis show the parameters in the same sequence and the axis ticks show the values of each parameter. The diagonal of the matrix shows the histogram of each parameter. From the diagonal of Figure 12, air pollution concentrations include ground-level ozone that is denoted as ‘oz’, $PM_{2.5}$ that is denoted as ‘pm’, NO_2 and NO_x , and meteorological measurements include solar that

is denoted as 'sol', precipitation that is denoted as 'pc', relative humidity that is denoted as 'hm' and resultant wind speed that is denoted as 'wd' are following the Poisson distribution, and the large portion of measurements of these parameters is in lower value bins. In the meantime, meteorological measurements such as temperature that is denoted as 'te', outdoor pressure that is denoted as 'ps', and average traffic speed that is denoted as 'tr' are following the Normal distribution. The difference behind the distribution is that for the concentration of ground-level ozone, PM_{2.5}, NO₂, and NO_x, there is a lower bound of the measurements that are zero. While large portions of the measurements are close to the lower bound, none of them is smaller than zero because the air pollution concentration measurements cannot be negative. This is also true for solar radiation, precipitation, and resultant wind speed. The values of these meteorological measurements are close to zero but cannot be negative. On the contrary, the relative humidity values are following the Poisson distribution because the values have an upper bound that is 100 percent, and the measurements are close to that bound but cannot be more than that.

For the outdoor temperature, pressure, and average traffic speed, the measurements are clustered at median bins. For example, most temperature measurements are in the range of 70 to 90 F; most of the pressure measurements are in the range of 1,010 to 1,020 millibars; most of the average traffic speed measurements are in the range of 55 to 60 mph. As the values are higher or lower, the densities become smaller, thus, they are following the Normal distribution. One exception is the wind direction that is denoted as 'wdir' in the figure, due to the measurements of the directions being values from 0 to 360-degree compass, which are not accumulable and distributed dispersive in bins. Thus, the wind direction value is not following the Poisson distribution, nor the Normal distribution.

From the scatter plots of Figure 12, some air pollutants have relatively clearer relationships with several other parameters. For instance, (1) Ground-level ozone shows a positive relationship with solar radiation, a negative relationship with relative humidity, a positive relationship with resultant wind speed, and no clear relationship with other parameters. (2) $PM_{2.5}$ shows positive relationships with NO_2 and NO_x , and no clear relationship between other parameters. (3) NO_2 shows positive relationships with $PM_{2.5}$ and NO_x , positive relationships with outdoor pressure and humidity, and negative relationships with temperature and resultant wind speed. (4) NO_x shows positive relationships with $PM_{2.5}$ and NO_2 . (5) Other than air pollutants, solar radiation is positively related to temperature and negatively related to relative humidity. (6) Other than previous relationships, the temperature is negatively related to the pressure and relative humidity, and positively related to resultant wind speed. (7) No other parameter is obviously related to the pressure other than previous relationships. (8) Precipitation level remains in a low range and hard to find relationships with other parameters. (9) The relative humidity may be positively related to the resultant wind speed. (10) The resultant wind speed is positively related to wind direction when it's less than around 170-degree compass (N-E-SSE wind), and negatively related to wind direction when it's more than 170 degree compass (SSE-S-W-N wind), which means wind direction ranges from SSE to S is stronger.

To further analyze and quantify the relationships between each parameter, a Pearson's r correlation test is performed.

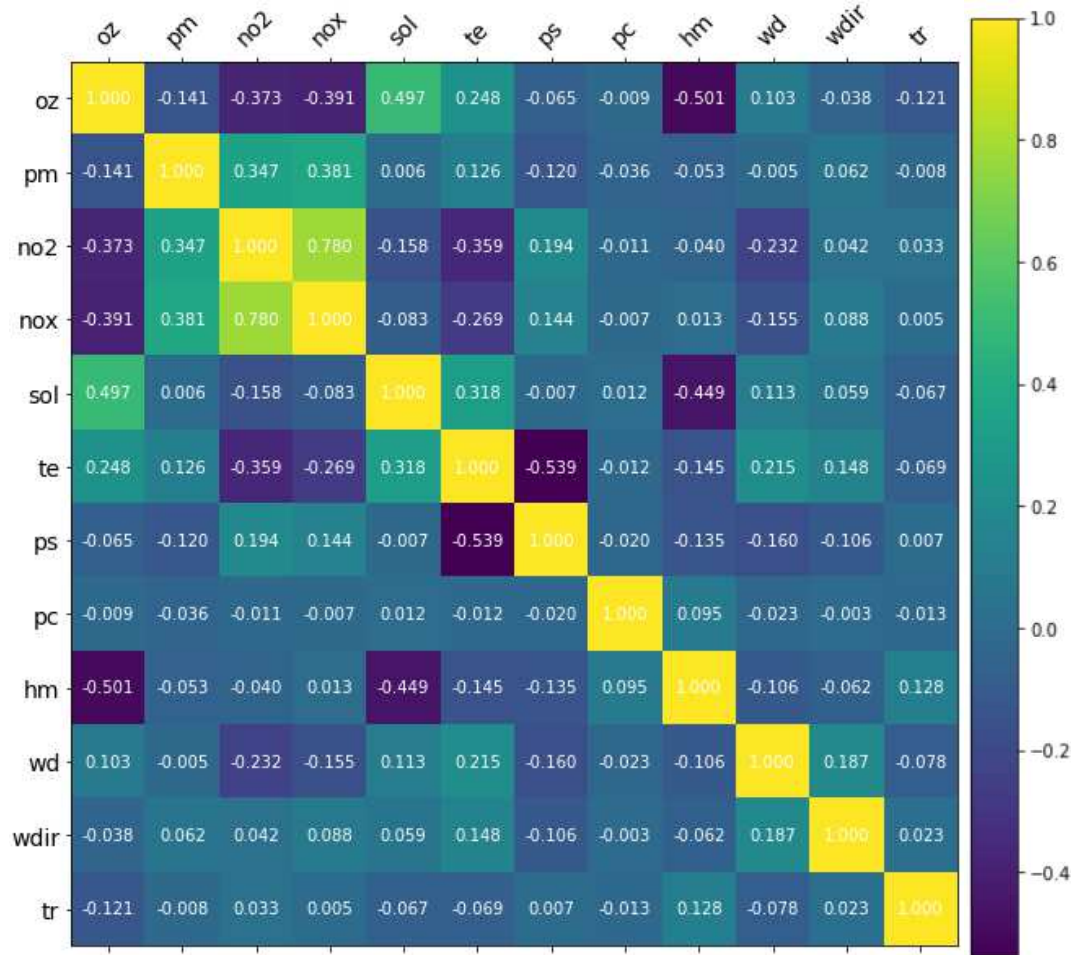


Figure 13. Pearson's r Correlation Test Matrix

Figure 13 shows the matrix plot of Pearson's r test results for each parameter. In the matrix table, colors are self-comparable, which means the colors can be only compared between the same column. The green hue (toward yellow color) means the positive correlation, and the yellower color the stronger the correlation, which means the r -value is closer to +1; the blue hue (toward purple color) means the negative correlation, and the deeper the color the stronger the correlation, which means the r -value closer to -1 . The details of the color are shown in the color bar under the matrix plot. The numbers in the matrix show Pearson's r values between every two parameters. The diagonal numbers of

the matrix are all equal to +1, which are the correlations between each parameter to themselves.

From the figure, the correlations can be described as follows.

(1) Ground-level ozone has relatively weaker negative correlations $[-0.1, 0)$ with pressure, precipitation, wind direction, and relatively strong negative correlations $(-0.1, -1)$ with $PM_{2.5}$, NO_2 , NO_x , relative humidity, and average traffic speed.

(2) Ground-level ozone has relatively strong positive correlations $(0.1, 1)$ with solar radiation, temperature, and resultant wind speed.

(3) $PM_{2.5}$ has relatively weaker negative correlations with precipitation, relative humidity, resultant wind speed, average traffic speed, and a relatively strong negative correlation with pressure.

(4) $PM_{2.5}$ has relatively stronger positive correlations with NO_2 , NO_x , and temperature, and relatively weak position correlations with solar radiation and wind direction.

(5) NO_2 has relatively weaker negative correlations with precipitation and relative humidity and relatively strong negative correlations with solar radiation, temperature, and resultant wind speed.

(6) NO_2 has relatively weaker positive correlations with wind direction and average traffic speed and relatively strong correlations with NO_x and pressure.

(7) NO_x has relatively weaker negative correlations with solar radiation and precipitation and relatively strong negative correlations with temperature, and resultant wind speed.

(8) NO_x has relatively weaker positive correlations with relative humidity, wind direction, and average traffic speed, and a relatively strong correlation with pressure.

(9) Solar radiation has relatively weaker negative correlations with pressure and average traffic speed and a relatively strong negative correlation with relative humidity.

(10) Solar radiation has relatively weaker positive correlations with precipitation and wind direction and relatively strong negative correlations with temperature and resultant wind speed.

(11) Outdoor temperature has relatively weaker negative correlations with precipitation and average traffic speed and relatively strong negative correlations with pressure and relative humidity.

(12) Outdoor temperature has relatively stronger positive correlations with resultant wind speed and wind direction.

(13) Pressure has a relatively weaker negative correlation with precipitation, relatively strong negative correlations with relative humidity, resultant wind speed, wind direction, and a relatively weak correlation with average traffic speed.

(14) Precipitation has relatively weaker negative correlations with resultant wind speed, wind direction, average traffic speed, and a relatively weak positive correlation with relative humidity.

(15) Relative humidity has a weaker negative correlation with wind direction, a relatively strong negative correlation with resultant wind speed, and a relatively strong positive correlation with average traffic speed.

(16) Resultant wind speed has a relatively weaker negative correlation with average traffic speed and a relatively strong positive correlation with wind direction.

(17) Wind direction has a relatively weaker position correlation with average traffic speed.

From the analysis above, some essential correlations between air pollutants and meteorological measurements and transportation situations can be summarized. Such as ground-level ozone is more related to solar radiation, temperature, relative humidity, resultant wind speed, and average traffic speed; $PM_{2.5}$ is more related to temperature and pressure; NO_2 is more related to solar radiation, temperature, pressure, and resultant wind speed; NO_x is more related to temperature, pressure, and resultant wind speed. Based on the counts of influencing factors and respective correlation levels of each air pollutant, it may be anticipated that ground-level ozone, NO_2 , and NO_x concentration prediction might be more accurate than that for $PM_{2.5}$. However, further analysis is performed and the frequent patterns are revealed.

4.3 Frequent Pattern Mining Analysis

4.3.1 Data Preprocessing and Binning

The first step to perform the frequent pattern mining analysis is to transfer the raw data into different bins based on the distribution of the data. As introduced in the Design

of the Study section, the four bins for each parameter are determined by quartiles. As introduced in Section 3, the first quartile is labeled as 1; the second quartile is labeled as 2; the third quartile is labeled as 3, and the last quartile is labeled as 4. For precipitation bins, any positive values are labeled as 1 and the value zero is labeled as 0.

TABLE 11. BINS AND LABELS FOR PARAMETERS

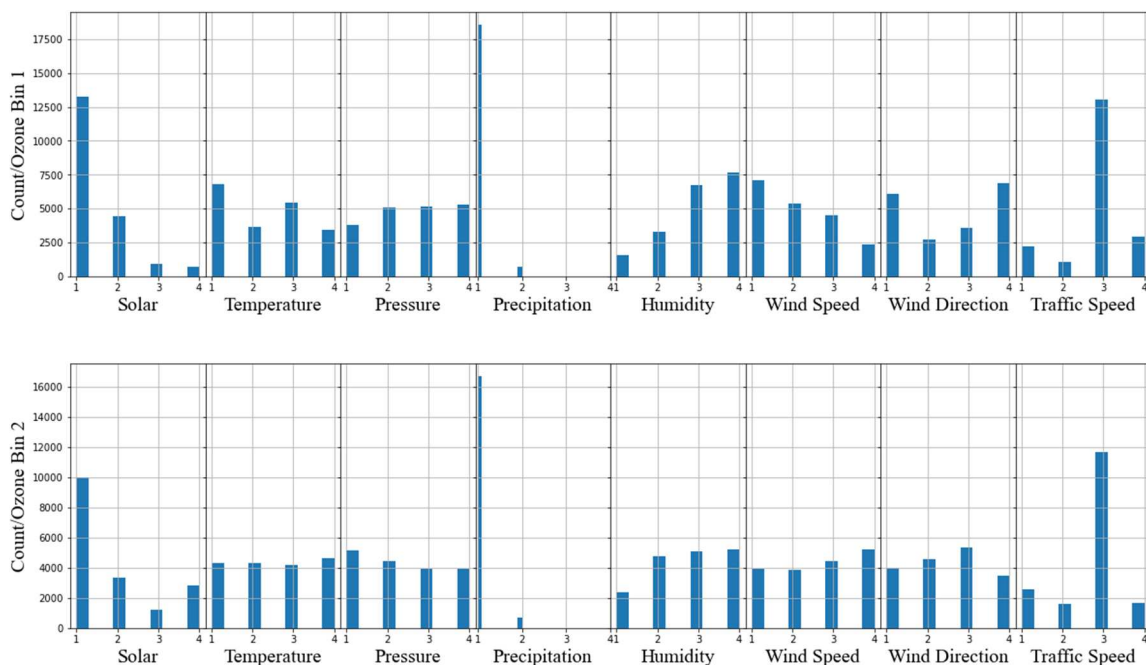
Measurements	Bins	Labels
Ground-level Ozone (ppb)	(0, 11]	1
	(11, 20]	2
	(20, 31]	3
	(31, +∞)	4
PM2.5 (ug/m3)	(0, 6.4]	1
	(6.4, 29.5]	2
	(9.5, 13.5]	3
	(13.5, +∞)	4
NO2 (ppb)	(0, 6.1]	1
	(6.1, 9.6]	2
	(9.6, 15.1]	3
	(15.1, +∞)	4
NOx (ppb)	(0, 7.6]	1
	(7.6, 12.6]	2
	(12.6, 21.4]	3
	(21.4, +∞)	4
Solar (Langleys/min)	[0, 0.01]	1
	(0.01, 0.25]	2
	(0.25, 0.414]	3
	(0.414, +∞)	4
Temperature (F)	(-∞, 62.5]	1
	(62.5, 73.8]	2
	(73.8, 81.4]	3
	(81.4, +∞)	4
Pressure (millibars)	(-∞, 1013.1]	1
	(1013.1, 1015.9]	2
	(1015.9, 1019.6]	3
	(1019.6, +∞)	4
Precipitation (inches)	[0]	0
	(0, +∞]	1

Relative Humidity (%)	(0, 57.9]	1
	(57.9, 73.8]	2
	(73.8, 83.8]	3
	(83.8, 100]	4
Resultant Wind (mph)	(0, 3.4]	1
	(3.4, 5.6]	2
	(5.6, 8.3]	3
	(8.3, $+\infty$)	4
Wind Direction (Degree Compass)	[0, 70]	1
	(70, 165]	2
	(165, 222]	3
	(222, 360)	4
Traffic Speed (mph)	(0, 50]	1
	(50, 55]	2
	(55, 60]	3
	(60, $+\infty$)	4

TABLE 11 shows the bins and labels of all parameters. As shown in the table, there are four bins for all parameters except precipitation, in which, Bin 1 is the first quarter that ranges from negative infinity to the first quartile; Bin 2 is the second quarter that ranges from the first quartile to the second quartile; Bin 3 is the third quarter that ranges from the second quartile to the third quartile; Bin 4 is the fourth quarter that ranges from the third quartile to the positive infinity. The binning for traffic speed utilizes 5 mph as ranges for each bin from 50 to 60 mph because most of the speeds are within this range, which is different from other parameters that use quartiles as boundaries. Only two bins are used to categorize the precipitation due to the large scale of its value is zero, which means as long as it is raining, no matter drizzling or storming, it belongs to precipitation Bin 1. At the same time, if there is no rain, it belongs to precipitation Bin 0. The frequent pattern mining processes in the following sessions are conducted based on the bins that are shown above.

4.3.2 Ground-level Ozone Frequent Pattern Mining

Figure 14 shows the bin distribution of each meteorological measurement and traffic speed of ground-level ozone. The trends can be found by comparing each column of the figure vertically. For ground-level ozone concentration level categories from bin 1 to bin 4, (1) the number of solar radiation bin 1 decreased and the number of solar radiation bin 4 increased, (2) the number of temperature bin 1 and bin 4 decreased significantly and the number of bin 3 decreased, (3) the number of humidity bin 1 increased and bin 4 decreased significantly, and (4) the resultant wind speed bin 1 decreased and bin 4 increased. These results are consistent with the previous correlation analysis.



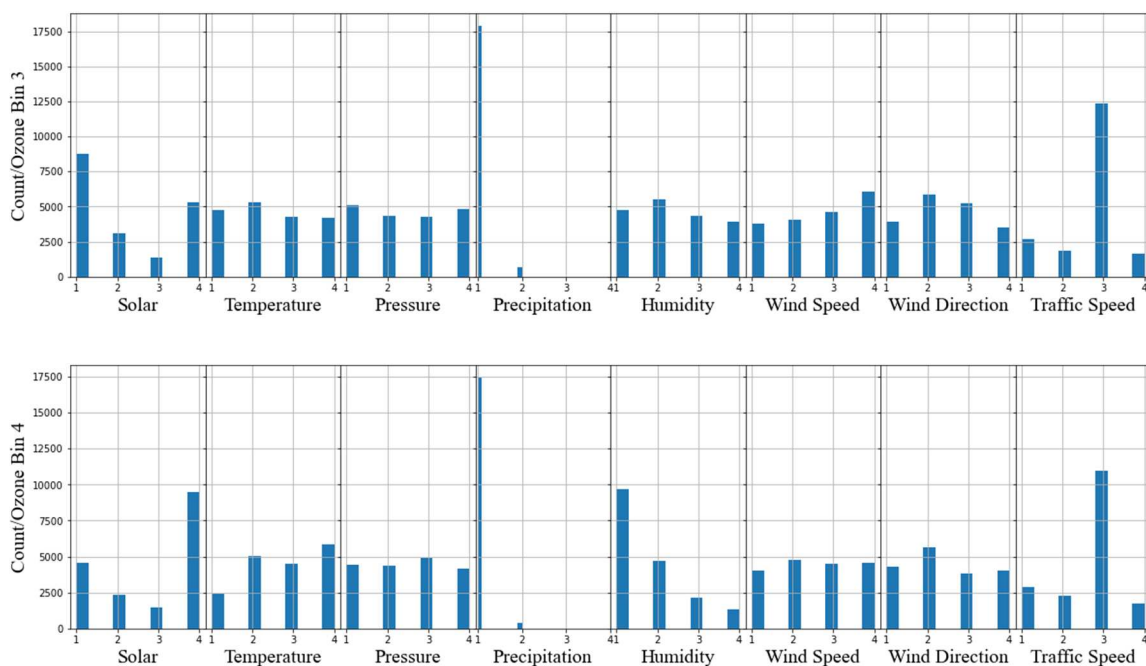


Figure 14. Distributions of Influencing Factors of Ground-level ozone

TABLE 12 shows the frequent pattern mining rules with the highest support for ground-level ozone. This table can be interpreted as follows.

TABLE 12. FREQUENT PATTERNS WITH THE HIGHEST SUPPORT FOR GROUND-LEVEL OZONE

Parameters				
Ground-level Ozone Level	1	2	3	4
Solar	1	1	1	4
Temperature	1	4	3	4
Pressure	4	2	1	1
Precipitation	0	0	0	0
Humidity	4	2	3	1
Wind Speed	1	4	4	4
Wind Direction	1	3	3	3
Traffic Speed	3	3	3	3
Support	0.00265	0.0024	0.00213	0.00132

(1) when the solar radiation level under bin 1 ($[0, 0.01]$ Langleys/min), the temperature level under bin 1 ($(-\infty, 62.5]$ F), the pressure level under bin 4 ($(-\infty, 1013.1]$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 4 ($(83.8, 100]$ percentage), the resultant wind speed level under bin 1 ($(0, 3.4]$ mph), the wind direction under bin 3 ($[0, 70]$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 1 ($(0, 11]$ ppb) with the support of 0.265%.

(2) When the solar radiation level under bin 1 ($[0, 0.01]$ Langleys/min), the temperature level under bin 4 ($(81.4, +\infty)$ F), the pressure level under bin 2 ($(1013.1, 1015.9)$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 2 ($(57.9, 73.8]$ percentage), the resultant wind speed level under bin 4 ($(8.3, +\infty)$ mph), the wind direction under bin 3 ($(165, 222]$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 2 ($(11, 20]$ ppb) with the support of 0.24%.

(3) When the solar radiation level under bin 1 ($[0, 0.01]$ Langleys/min), the temperature level under bin 3 ($(73.8, 81.4]$ F), the pressure level under bin 1 ($(-\infty, 1013.1]$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 3 ($(73.8, 83.8]$ percentage), the resultant wind speed level under bin 4 ($(8.3, +\infty)$ mph), the wind direction under bin 3 ($(165, 222]$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 3 ($(20, 31]$ ppb) with the support of 0.24%.

(4) When the solar radiation level under bin 4 ($(0.414, +\infty)$ Langleys/min), the temperature level under bin 4 ($(81.4, +\infty)$ F), the pressure level under bin 1 ($(-\infty, 1013.1]$

millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 1 ((0, 57.9] percentage), the resultant wind speed level under bin 4 ((8.3, $+\infty$) mph), the wind direction under bin 3 ((165, 222] degree compass), and the traffic speed level under bin 3 ((55, 60] mph), the ground-level ozone concentration level tend to be under bin 4 ((31, $+\infty$) ppb) with the support of 0.132%.

4.3.3 PM_{2.5} Frequent Pattern Mining

Figure 15 shows the bin distribution of each meteorological measurement and traffic speed of PM_{2.5}. For PM_{2.5} concentration level categories from bin 1 to bin 4, (1) the numbers of higher temperature bins increased, and the numbers of lower temperature bins decreased. (2) The numbers of lower pressure bins increased and the numbers of higher pressure bins decreased. (3) The number of relative humidity bin 4 decreased. (4) The numbers of higher wind direction bins increased, and the numbers of lower wind direction bins decreased. These results are consistent with the previous correlation analysis.

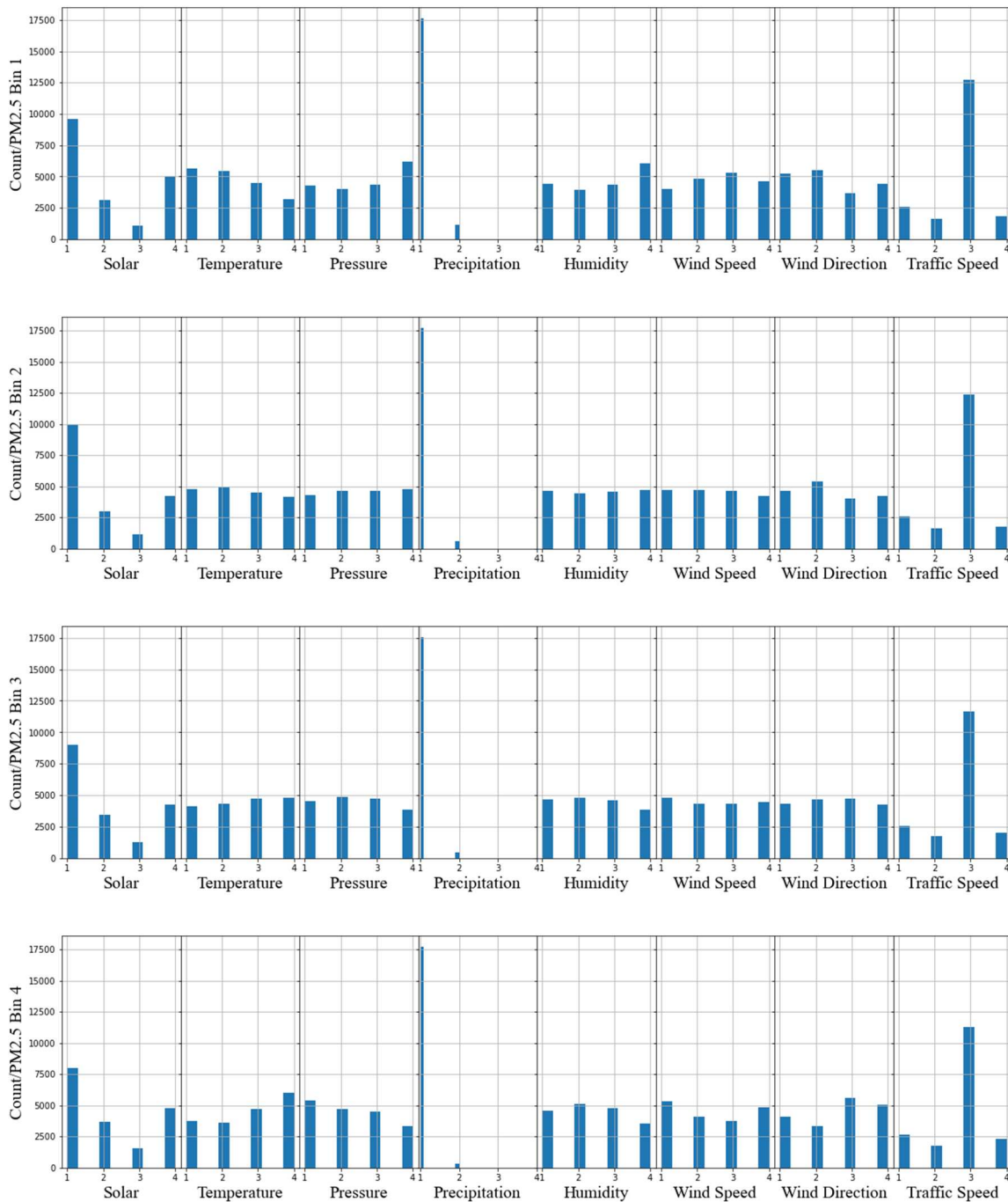


Figure 15. Distributions of Influencing Factors of PM_{2.5}

TABLE 13. FREQUENT PATTERNS WITH HIGHEST SUPPORT FOR PM_{2.5}

Parameters	1	2	3	4
PM2.5 Level	1	2	3	4

Solar	1	4	1	1
Temperature	3	1	4	3
Pressure	2	4	1	1
Precipitation	0	0	0	0
Humidity	4	4	2	3
Wind Speed	1	1	4	4
Wind Direction	1	1	3	3
Traffic Speed	3	3	3	3
Support	0.00141	0.00124	0.00137	0.00259

TABLE 13 shows the frequent pattern mining rules with the highest support for PM_{2.5}. The table can be interpreted as follows: (1) when the solar radiation level under bin 1 ([0, 0.01] Langleys/min), the temperature level under bin 3 ((73.8, 81.4] F), the pressure level under bin 2 ((-∞, 1013.1] millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 4 ((83.8, 100] percentage), the resultant wind speed level under bin 1 ((0, 3.4] mph), the wind direction under bin 1 ([0, 70] degree compass), and the traffic speed level under bin 3 ((55, 60] mph), the ground-level ozone concentration level tend to be under bin 1 ((0, 6.4] ug/m³) with the support of 0.141%.

(2) When the solar radiation level under bin 4 ((0.414, +∞) Langleys/min), the temperature level under bin 1 ((-∞, 62.5] F), the pressure level under bin 4 ((1019.6, +∞) millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 4 ((83.8, 100] percentage), the resultant wind speed level under bin 1 ((0, 3.4] mph), the wind direction under bin 1 ([0, 70] degree compass), and the traffic speed level under bin 3 ((55, 60] mph), the ground-level ozone concentration level tend to be under bin 2 ((6.4, 29.5] ug/m³) with the support of 0.124%.

(3) When the solar radiation level under bin 1 ($[0, 0.01]$ Langleys/min), the temperature level under bin 4 ($(81.4, +\infty)$ F), the pressure level under bin 1 ($(-\infty, 1013.1]$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 2 ($(57.9, 73.8]$ percentage), the resultant wind speed level under bin 4 ($(8.3, +\infty)$ mph), the wind direction under bin 3 ($(165, 222]$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 3 ($(9.5, 13.5]$ $\mu\text{g}/\text{m}^3$) with the support of 0.137%.

(4) When the solar radiation level under bin 1 ($[0, 0.01]$ Langleys/min), the temperature level under bin 3 ($(73.8, 81.4]$ F), the pressure level under bin 1 ($(-\infty, 1013.1]$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 3 ($(73.8, 83.8]$ percentage), the resultant wind speed level under bin 4 ($(8.3, +\infty)$ mph), the wind direction under bin 3 ($(165, 222]$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 4 ($(13.5, +\infty)$ $\mu\text{g}/\text{m}^3$) with the support of 0.259%.

4.4.4 NO₂ Frequent Pattern Mining

Figure 16 shows the bin distribution of each meteorological measurement and traffic speed of NO₂. For NO₂ concentration level categories from bin 1 to bin 4, (1) the numbers of higher temperature bins decreased and the numbers of lower temperature bins increased. (2) The numbers of lower higher pressure bins decreased and the numbers of higher pressure bins increased. (3) The number of relative humidity bin 4 decreased. (4) The numbers of higher resultant wind speed bins decreased and the numbers of lower resultant

wind speed bins increased. (5) The numbers of wind direction bins 2 and 3 decreased, and the numbers of wind direction bins 1 and 4 increased. These results are consistent with the previous correlation analysis.

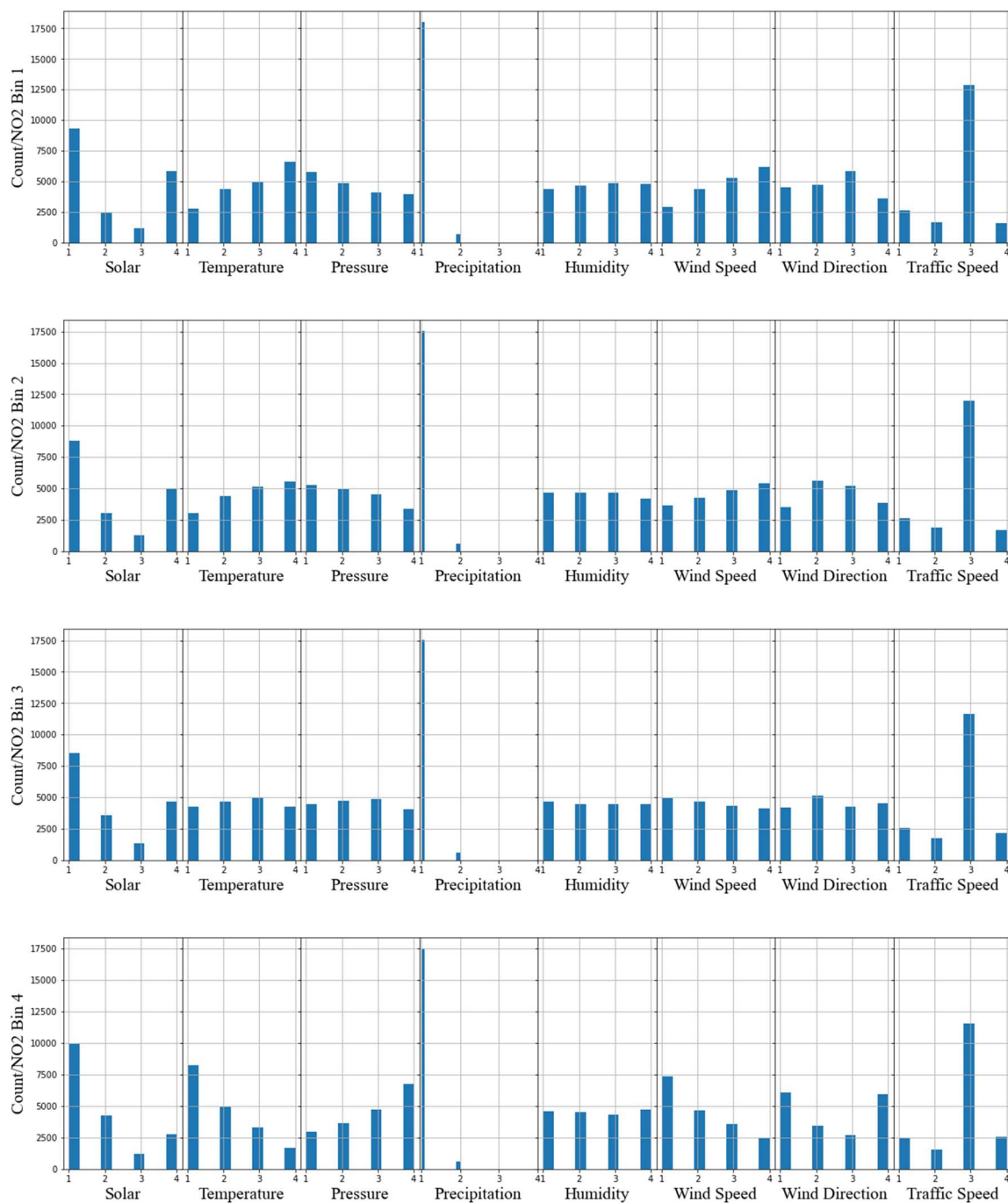


Figure 16. Distributions of Influencing Factors of NO₂

TABLE 14 shows the frequent pattern mining rules with the highest support for NO₂. TABLE 14 can be interpreted as follows.

TABLE 14. FREQUENT PATTERNS WITH HIGHEST SUPPORT FOR NO₂

Parameters	1	2	3	4
NO₂ Level	1	2	3	4
Solar	1	1	1	1
Temperature	3	3	3	1
Pressure	1	1	2	4
Precipitation	0	0	0	0
Humidity	3	3	4	1
Wind Speed	4	4	1	1
Wind Direction	3	3	1	4
Traffic Speed	3	3	3	3
Support	0.00289	0.00161	0.00128	0.00285

(1) when the solar radiation level under bin 1 ($[0, 0.01]$ Langleys/min), the temperature level under bin 3 ($(73.8, 81.4]$ F), the pressure level under bin 1 ($(-\infty, 1013.1]$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 3 ($(73.8, 83.8]$ percentage), the resultant wind speed level under bin 4 ($(8.3, +\infty)$ mph), the wind direction under bin 3 ($(165, 222]$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 1 ($(0, 6.1]$ ppb) with the support of 0.289%.

(2) When the solar radiation level under bin 1 ($[0, 0.01]$ Langleys/min), the temperature level under bin 3 ($(73.8, 81.4]$ F), the pressure level under bin 1 ($(-\infty, 1013.1]$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 3 ($(73.8, 83.8]$ percentage), the resultant wind speed level under bin 4 ($(8.3, +\infty)$

mph), the wind direction under bin 3 ((165, 222] degree compass), and the traffic speed level under bin 3 ((55, 60] mph), the ground-level ozone concentration level tend to be under bin 2 ((6.1, 9.6] ppb) with the support of 0.161%.

(3) When the solar radiation level under bin 1 ([0, 0.01] Langleys/min), the temperature level under bin 3 ((73.8, 81.4] F), the pressure level under bin 2 ((1013.1, 1015.9] millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 4 ((83.8, 100] percentage), the resultant wind speed level under bin 1 ((0, 3.4] mph), the wind direction under bin 1 ([0, 70] degree compass), and the traffic speed level under bin 3 ((55, 60] mph), the ground-level ozone concentration level tend to be under bin 3 ((9.6, 15.1] ppb) with the support of 0.128%.

(4) When the solar radiation level under bin 1 ([0, 0.01] Langleys/min), the temperature level under bin 1 ((-∞, 62.5] F), the pressure level under bin 4 ((1019.6, +∞) millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 1 ((0, 57.9] percentage), the resultant wind speed level under bin 1 ((0, 3.4] mph), the wind direction under bin 4 ((222, 360) degree compass), and the traffic speed level under bin 3 ((55, 60] mph), the ground-level ozone concentration level tend to be under bin 4 ((15.1, +∞) ppb) with the support of 0.285%.

4.4.5 NO_x Frequent Pattern Mining

Figure 17 shows the bin distribution of each meteorological measurement and traffic speed of NO_x.

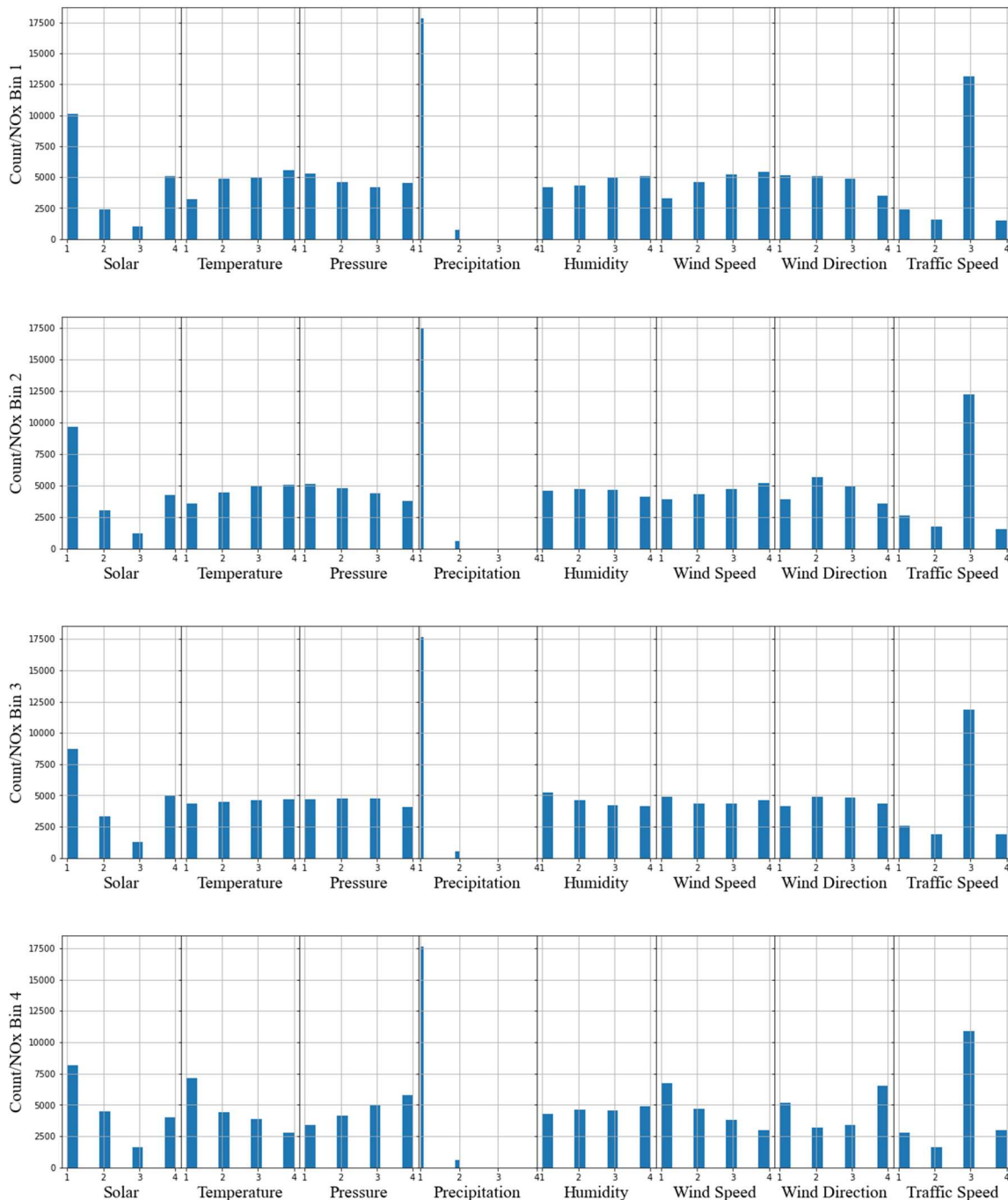


Figure 17. Distributions of Influencing Factors of NO_x

For NO_x concentration level categories from bin 1 to bin 4, (1) numbers of lower temperature bins decreased. (2) The numbers of higher pressure level bins decreased and then increased. (3) The numbers of lower resultant wind direction bins increased and then

decreased, the numbers of higher resultant wind speed bins decreased and then increased.

(4) The numbers of wind direction bins 2 and 3 decreased and then increased. These results are consistent with the previous correlation analysis.

TABLE 15 shows the frequent pattern mining rules with the highest support for NO_x.

TABLE 15. FREQUENT PATTERNS WITH HIGHEST SUPPORT FOR NO_x

Parameters	1	2	3	4
NO_x Level	1	2	3	4
Solar	1	1	4	1
Temperature	3	3	4	1
Pressure	1	1	1	4
Precipitation	0	0	0	0
Humidity	3	3	1	1
Wind Speed	4	4	4	1
Wind Direction	3	3	3	4
Traffic Speed	3	3	3	3
Support	0.00279	0.00158	0.00167	0.0024

TABLE 15 can be interpreted as follows: (1) when the solar radiation level under bin 1 ([0, 0.01] Langleys/min), the temperature level under bin 3 ((73.8, 81.4] F), the pressure level under bin 1 ((-∞, 1013.1] millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 3 ((73.8, 83.8] percentage), the resultant wind speed level under bin 4 ((8.3, +∞) mph), the wind direction under bin 3 ((165, 222] degree compass), and the traffic speed level under bin 3 ((55, 60] mph), the ground-level ozone concentration level tend to be under bin 1 ((0, 7.6] ppb) with the support of 0.279%.

(2) When the solar radiation level under bin 1 ($[0, 0.01]$ Langley/min), the temperature level under bin 3 ($(73.8, 81.4]$ F), the pressure level under bin 1 ($(-\infty, 1013.1]$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 3 ($(73.8, 83.8]$ percentage), the resultant wind speed level under bin 4 ($(8.3, +\infty)$ mph), the wind direction under bin 3 ($(165, 222]$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 2 ($(7.6, 12.6]$ ppb) with the support of 0.158%.

(3) When the solar radiation level under bin 4 ($(0.414, +\infty)$ Langley/min), the temperature level under bin 4 ($(81.4, +\infty)$ F), the pressure level under bin 1 ($(-\infty, 1013.1]$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 1 ($(0, 57.9]$ percentage), the resultant wind speed level under bin 4 ($(8.3, +\infty)$ mph), the wind direction under bin 3 ($(165, 222]$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 3 ($(12.6, 21.4]$ ppb) with the support of 0.167%.

(4) When the solar radiation level under bin 1 ($[0, 0.01]$ Langley/min), the temperature level under bin 1 ($(-\infty, 62.5]$ F), the pressure level under bin 4 ($(1019.6, +\infty)$ millibars), the precipitation level under bin 0 (no precipitation), the relative humidity level under bin 1 ($(0, 57.9]$ percentage), the resultant wind speed level under bin 1 ($(0, 3.4]$ mph), the wind direction under bin 4 ($(222, 360)$ degree compass), and the traffic speed level under bin 3 ($(55, 60]$ mph), the ground-level ozone concentration level tend to be under bin 4 ($(21.4, +\infty)$ ppb) with the support of 0.024%.

Based on the frequent pattern analysis above, each air pollutant concentration level can be roughly predicted and anticipated through meteorological measurements and

average traffic speeds. However, several issues arise from the analysis results and cannot be neglected. The first one is that for the rules of meteorological measurements and average traffic speeds, the corresponding frequent pattern mined air pollutant levels have support values lower than 3%, which is a result of the weak interrelationship between each parameter, and that may result in lower prediction accuracy. During the analysis, the higher support values mostly appeared in several certain single parameter rules. This shows that certain single parameters may be more related to the air pollution levels when compared to the full sets of parameters. The second one is that the bin values of some parameters in all rules are relatively constant. This is true for bin 3 ((55, 60] mph) of average traffic speed, and bin 0 (no precipitation) of precipitation. This phenomenon is a result of less variation of the input data for these parameters. Bin 3 of average traffic speed input data occupies more than 55% of all air pollutants bins, and bin 0 of precipitation input data occupies more than 94% of all air pollutants bins. In another word, the average traffic speeds on the monitoring site nearby highway are between 55 mph and 60mph for more than 55% of the time, and there is no rain near the monitoring site for 94% of the time, which can be resulted in for all air pollutants bins, the bins of these two parameters remain unchanged in the highest support value rules.

The frequent pattern analysis revealed the in-depth relationships between each air pollutant and parameters include the meteorological measurements and traffic status. Thus, air pollution concentration predictions using these parameters can be achieved in the following section.

4.4 Machine Learning Prediction Models for Air Pollution

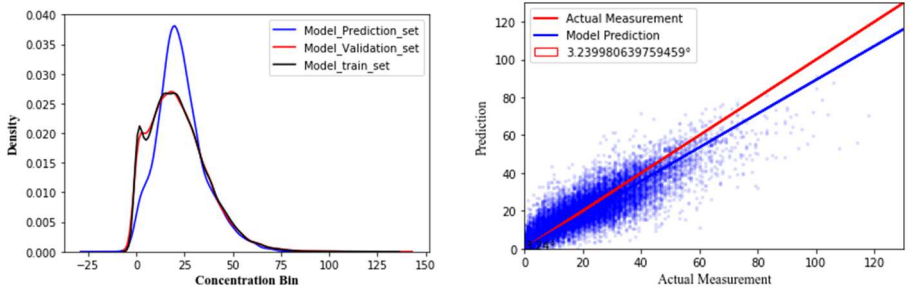
As there is no machine learning algorithm that fits all kinds of data sets, it is essential to select the proper algorithms for the air pollutants predictions, respectively. Thus, the machine learning models are trained, compared, and analyzed for each air pollutant. As stated in Chapter 3 on Design of the Study, the loaded dataset includes 2011 to 2019 historical air pollution concentration, meteorological measurements, and average traffic speed hourly data. 80% of the dataset are randomly selected to train the models, and the rest 20% of the dataset are used as validation to test the model accuracy. The stratified 10-fold cross-validation is used and the best result of each model is used for comparison between models.

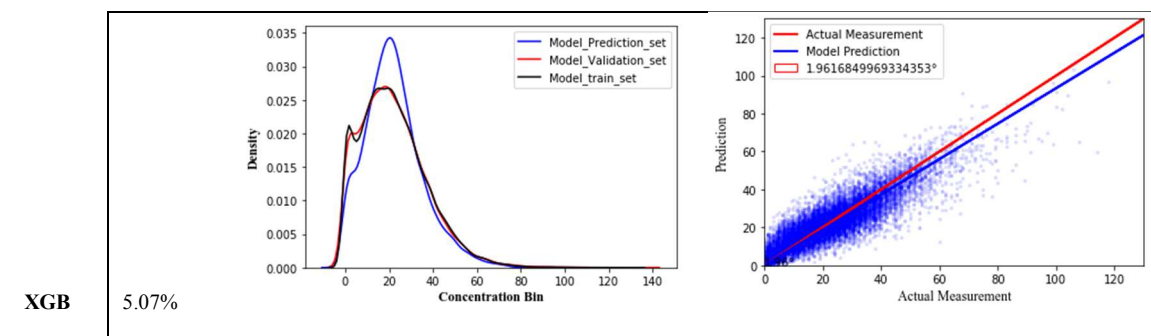
4.4.1 Machine Learning Model Selection and Training

4.4.1.1 Model Selection for Ground-level Ozone

TABLE 16 shows the selected two models for ground-level ozone concentration prediction. The full table for all seven candidate models is provided in Appendix A.

TABLE 16. MODEL SELECTED FOR GROUND-LEVEL OZONE

	NRMSE	Ground-level Ozone
MLP	4.81%	 <p>The figure contains two plots. The left plot is a density plot with 'Density' on the y-axis (0.000 to 0.040) and 'Concentration Bin' on the x-axis (-25 to 150). It shows three curves: a blue line for 'Model_Prediction_set', a red line for 'Model_Validation_set', and a black line for 'Model_train_set'. All curves peak around a concentration bin of 25. The right plot is a scatter plot with 'Prediction' on the y-axis (0 to 120) and 'Actual Measurement' on the x-axis (0 to 120). It shows blue dots representing data points, a red diagonal line for 'Actual Measurement', a blue diagonal line for 'Model Prediction', and a red box containing the value '3.239980639759459°'.</p>



In TABLE 16, the first column shows the name of the machine learning model and the second column shows the NRMSE value. Two types of plots are demonstrated that are line plots and scatter plots. The line plot shows the distribution of the ground-level ozone concentration. In which the model training set is shown in the black line, the model validation set is shown in the red line, and the model prediction set is shown in the blue line. Since the model training set is randomly selected 80% of the database and the model validation set is the rest 20%, the distribution of the validation should be close to the training set, which means the black line and the red line should be close to each other. In the meantime, the more accurate the model prediction, the blue line and the red line are closer. The distribution line plot only shows the distribution differences between the sets, but it doesn't show the degree of the error like the NRMSE value. Thus, even the gap between the blue line and the red line is relatively big, the NRMSE value may still be relatively small. The scatter plot shows the comparison between the model prediction set and the model validation set. In which the red line is a 45-degree line that shows the 100% accuracy match, the blue line is the trend line of the prediction scatters. The more accurate the model prediction, the smaller the prediction error angle between the red line and the blue line.

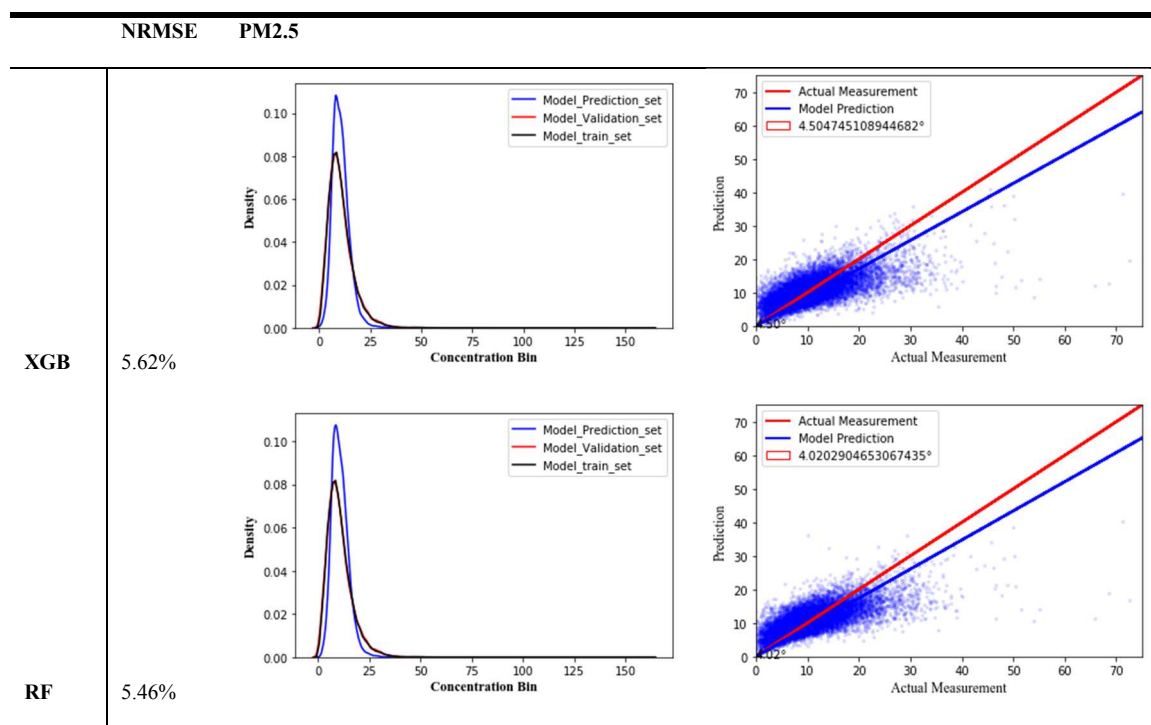
From the comparison, all models have NRMSE values lower than 7%, and the prediction error angles are smaller than 9° . The blue trend lines in the scatter plots are all below the red 45-degree lines, which means the concentrations predicted tend to be lower than the validation values. Among the models, the KNN and SVM models show relatively poorer fits for the dataset with the NRMSE value of 6.74% and 6.72%, and the prediction error angles are 3.44° and 8.87° . The PR, RF, and LR models are more accurate than the KNN and SVM models, while still not good enough. In the meantime, the MLP model prediction of ground-level ozone concentration has the lowest NRMSE of 4.81%. From the distribution plot, the prediction of the air pollution concentration over 30 ppb is very accurate, the prediction between 0 to 10 ppb has a lower density than the validation set and the prediction between 10 to 30 ppb has a higher density than the validation set. This trend is also true for all models except the SVM model, which has a higher density for predictions between 10 to 35 ppb, and a lower density for other predictions. The prediction error angle of the MLP model is 3.24° in the scatter plot, which means the prediction is very close to the validation, and this model fits the prediction needs for ground-level ozone. Another model XGB is also selected as a candidate model that has a good NRMSE value of 5.07%, and a small prediction error angle of 1.96° . Thus, the XGB and MLP models are selected to perform the ground-level ozone prediction.

4.4.1.2 Model Selection for PM_{2.5}

TABLE 17 shows the model comparison for PM_{2.5} concentration, which includes the models' names, the NRMSE values, the distribution plots, and the prediction scatter plots. The format and the meaning of the lines, angles, and colors used in this table are

the same as in TABLE 16. The full table for all seven candidate models is provided in Appendix B.

TABLE 17. MODEL SELECTED FOR PM_{2.5}

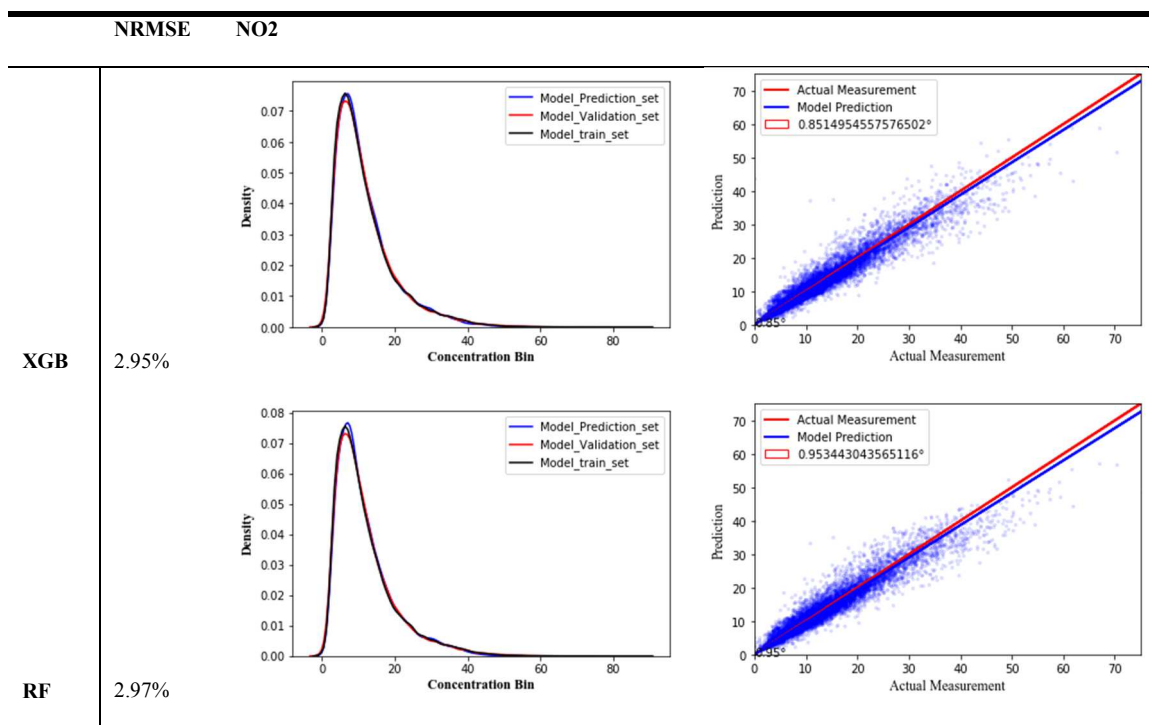


As shown in TABLE 17, all models have NRMSE values lower than 7%, and the prediction error angles are smaller than 10°, which means the predictions of PM_{2.5} are not as accurate as ground-level ozone. The blue trend lines in the scatter plots are all below the red 45-degree lines, which means the concentrations predicted tend to be lower than the validation values. Among them, the PR, SVM, and LR models show relatively poorer fits for the dataset with the NRMSE value of 6.58%, 6.91%, and 6.58% and the prediction error angles are 6.21°, 9.18°, and 6.21°. The MLP and KNN models are more accurate than the PR, SVM, and LP models but less accurate than the XGB and RF models.

The RF model prediction of PM_{2.5} concentration has the lowest NRMSE value of 5.46%. From the distribution plot, the prediction of the air pollution concentration between 7 to 19 ug/m³ has a higher density than the validation set, and other predictions have lower densities than the validation set. The XGB model shows almost the same trend as the RF model with the NRMSE value of 5.62%. Other models with lower NRMSE also have this trend, which is a sign that the predicted PM_{2.5} concentration values tend to be more concentrated in a certain range while the validation values tend to be more dispersed. The prediction error angle of the RF model is 4.02° in the scatter plot, which means the prediction is very close to the validation, and this model fits the prediction needs for PM_{2.5}. The prediction error angle of the XGB model is 4.5°, which is competitive when compared with other models. Thus, the XGB and RF models are selected to perform the PM_{2.5} prediction. However, based on the model validations, the PM_{2.5} concentration prediction is less accurate than the ground-level ozone concentration. This may be a result of the complex sources of the local PM_{2.5} that are not highly influenced by the meteorological measurements.

4.4.1.3 Model Selection for NO₂

TABLE 18 shows the model comparison for NO₂ concentration, which includes the models' names, the NRMSE values, the distribution plots, and the prediction scatter plots in columns from left to right. The format and the meaning of the lines, angles, and colors used in this table are the same as in TABLE 16 and 8. The full table for all seven candidate models is provided in Appendix C.

TABLE 18. MODEL SELECTED FOR NO₂

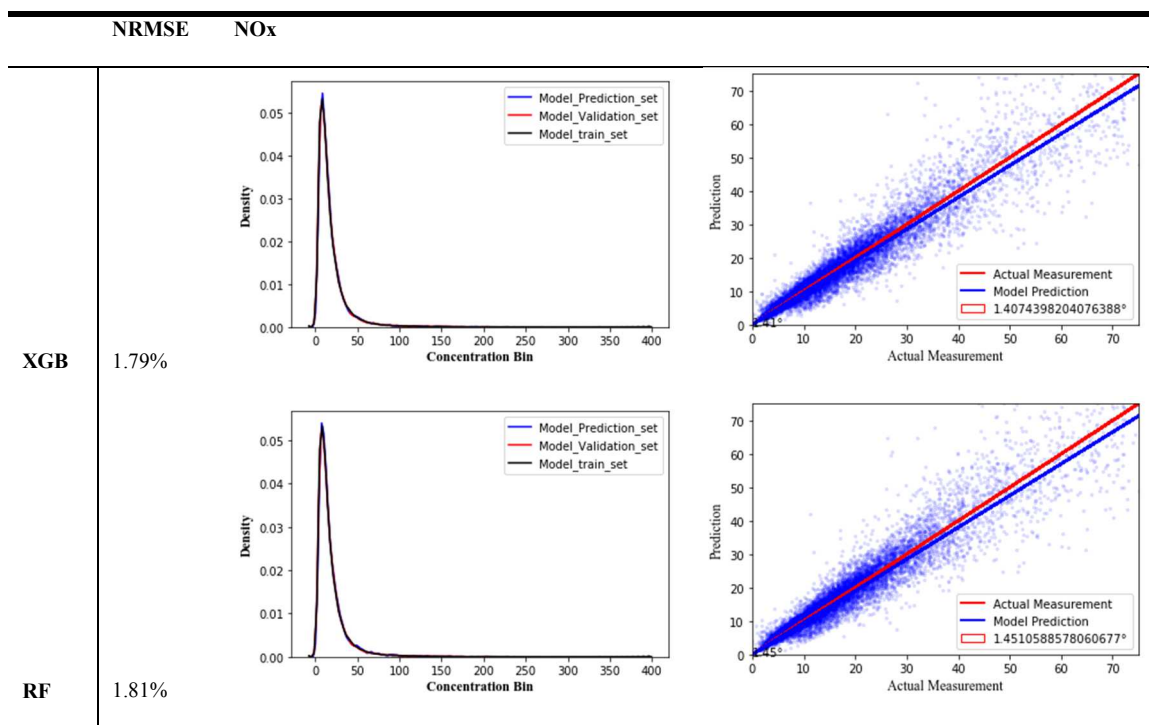
As shown in TABLE 18, all models have NRMSE values lower than 5%, and the prediction error angles are smaller than 3.8° , which indicates that the NO₂ predictions may be more accurate than the PM_{2.5} and ground-level ozone predictions. The blue trend lines in the scatter plots are all below the red 45-degree lines, which means the NO₂ concentrations predicted tend to be lower than the validation values even some models have extremely accurate predictions. Among the models, the PR, SVM, LR, and KNN models show relatively poorer fits for the dataset and the prediction accuracy. When compared to these models, the MLP, XGB, and RF models are more accurate.

The XGB model prediction of NO₂ concentration has the lowest NRMSE value of 2.95% based on the validation set. From the distribution plot, the predictions of the air pollution concentration are almost perfectly overlapping with the validation set and the

model training set. The prediction error angle of the RF model is 0.851° in the scatter plot, which means the prediction is very close to the validation, and this model fits the prediction needs for NO_2 . The RF model is also a good fitting model with an NRMSE value of 2.97% and the prediction error angle is 0.959° . This result is slightly lower than the XGB model but still relatively accurate. Another good fitting model in the selection process is the MLP model that has an NRMSE value of 3.12%, and the prediction error angle is 1.51° . However, due to the superior prediction ability of the XGB and RF models, they are selected to perform the NO_2 prediction. From the above analysis and comparison, the machine learning models have a more accurate prediction potential for NO_2 than for $\text{PM}_{2.5}$ and ground-level ozone.

4.4.1.4 Model Selection for NO_x

TABLE 19 shows the model comparison for NO_x concentration, which includes the models' names, the NRMSE values, the distribution plots, and the prediction scatter plots in columns from left to right. The format and the meaning of the lines, angles, and colors used in this table are the same as in TABLE 16, 8, and 9. The full table for all seven candidate models is provided in Appendix D.

TABLE 19. MODEL SELECTED FOR NO_x

As shown in TABLE 19, all models except the SVM have NRMSE values lower than 4%, and the prediction error angles are smaller than 6°, which indicates that the NO_x predictions may be more accurate than the PM_{2.5} and ground-level ozone predictions but less accuracy than the NO₂ prediction. The blue trend lines in the scatter plots are all below the red 45-degree lines, which means the NO₂ concentrations predicted tend to be lower than the validation values. Among the models, the PR, MLP, LR, and KNN models show relatively poorer fits for the dataset and the prediction accuracy. The SVM model has a huge error in the prediction performance with an NRMSE value of 5.02%, and a prediction error angle of 19.7°. When comparing these models, the XGB and RF models are more accurate.

The XGB model prediction of NO_x concentration has the lowest NRMSE value of 1.79% based on the validation set. From the distribution plot, the predictions of the air pollution concentration are almost perfectly overlapping with the validation set and the model training set. The prediction error angle of the RF model is 1.41° in the scatter plot, which means the prediction is very close to the validation, and this model fits the prediction needs for NO_x. The RF model is also a good fitting model with an NRMSE value of 1.81% and the prediction error angle is 1.49°. This result is slightly lower than the XGB model but still relatively accurate. As a result, the XGB and RF models are selected to perform the NO_x prediction. From the above analysis and comparison, the XGB and RF models have the ability to predict NO_x more accurate than NO₂, however, the SVM model performs the poorest for NO_x prediction among all models and datasets, which is a sign that the NO_x prediction is more model-specific than other air pollutants.

From the model selection process for all air pollutants, the outperformed models are selected and listed in TABLE 20.

TABLE 20. THE MODELS SELECTED FOR EACH AIR POLLUTANT

Air Pollutants	Ground-level ozone	PM_{2.5}	NO₂	NO_x
Models Selected	MLP, XGB	XGB, RF	XGB, RF	XGB, RF

As shown in TABLE 20, two models for each air pollutant are selected to perform the prediction. As shown in the table, the XGB model fits all the air pollutant datasets accurately, while the MLP and RF models are also good air pollution concentration prediction candidates. Based on the validation data set, the NO₂ concentration prediction can be relatively more accurate, the PM_{2.5} concentration prediction has relatively lower

accuracy, and the NO_x concentration prediction is more models specific, which can be very accurate when using the XGB and RF models, and inaccurate when using the SVM model. The following prediction processes are performed based on the models selected.

4.4.2 Air Pollution Forecasting Based on the Models Selected

As introduced in the Design of the Study section, the selected models are utilized to predict the air pollution concentration level of the year 2020 based on the historical data from 2011 to 2019. For which, the historical data is used to train the models, and the 2020 data is used as the prediction validation.

TABLE 21, TABLE 22, TABLE 23, and TABLE 24 show the air pollution concentration level predictions for the year 2020 and their validation. The first row shows the NRMSE values for the two selected models. The distribution plots show the distribution comparison between the prediction and the actual measurements, in which the black line shows the nigh years train data, the red line shows the actual measurements from the year 2020, and the blue line shows the model prediction. Unlike the model selection process, the black lines and the red lines do not overlap with each other on a large scale. The reason is that for the preprocess, the validation data set is randomly selected, thus, has the same distribution as the training data. The *x*-axis of the distribution plots means the air pollution concentration bins in ppb, and the *y*-axis means the distribution density of each bin. The PVA plots show the comparison between the prediction and the real data in scatters. The red line is a 45° line that shows the trend of the actual measurements, the blue lines show the linear trend of the prediction values, and the blue scatters show the prediction for each

actual data. In the PVA plots, the angles between the blue line and the red line are the prediction error angles that are negatively related to the prediction accuracy. The predictions are overall higher than the actual measurements if the blue line is above the red line, vice versa. The last two rows of tables show the real-time comparison plots between the prediction and the actual measurements in the hours of the year 2020. Due to the invalid and null records being dropped during the preprocess on the raw data, 6,912 hours of air pollution concentration levels are predicted for the year 2020 rather than the true total of 8,784 hours. The blue scatters in the comparison plots show the prediction of the year 2020, the red scatters show the actual measurements, and the gray scatters show the error between the real data and the prediction. The blue, red, and gray lines show the polynomial trend lines of the scatters respectively. When the gray error line is smoother and closer to the $y=0$ line, the more stable and accurate the model is.

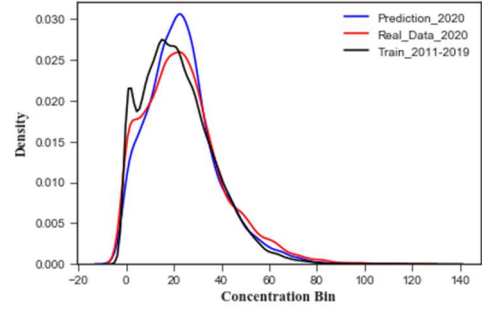
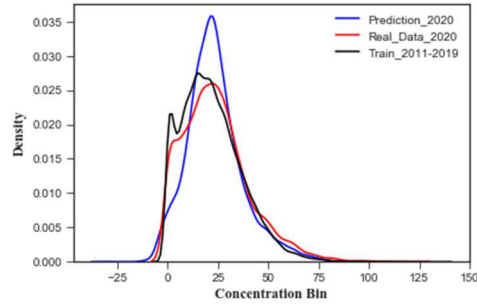
4.4.2.1 Ground-level Ozone Concentration Prediction

Based on the analysis and comparison from TABLE 21, the NRMSE value of the XGB model is 6.50%, which is higher than 5.42% of the MLP value. This means the MLP model is more accurate than the XGB model for ground-level ozone prediction.

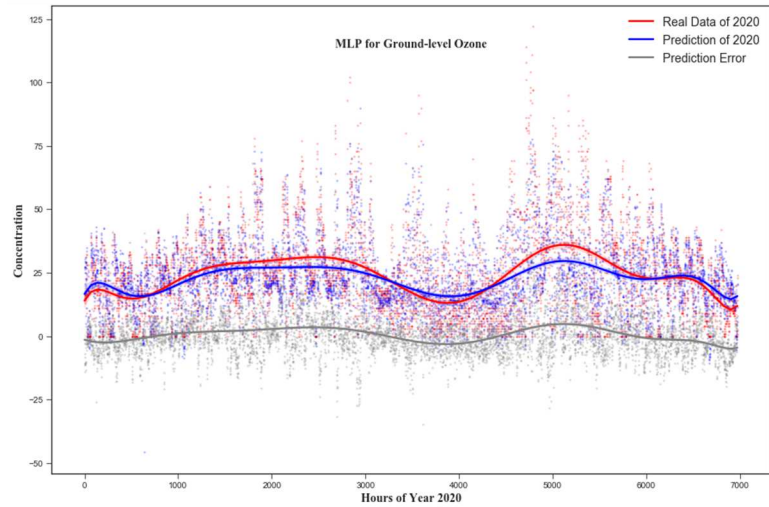
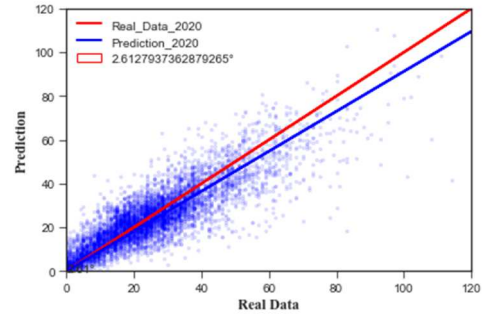
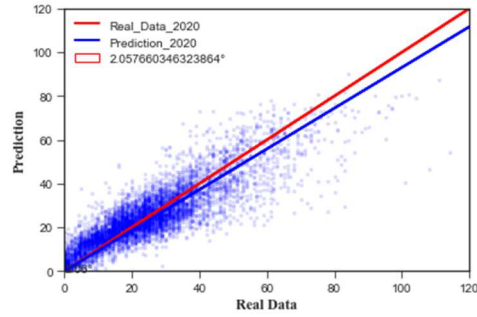
TABLE 21. PREDICTION OF GROUND-LEVEL OZONE CONCENTRATION

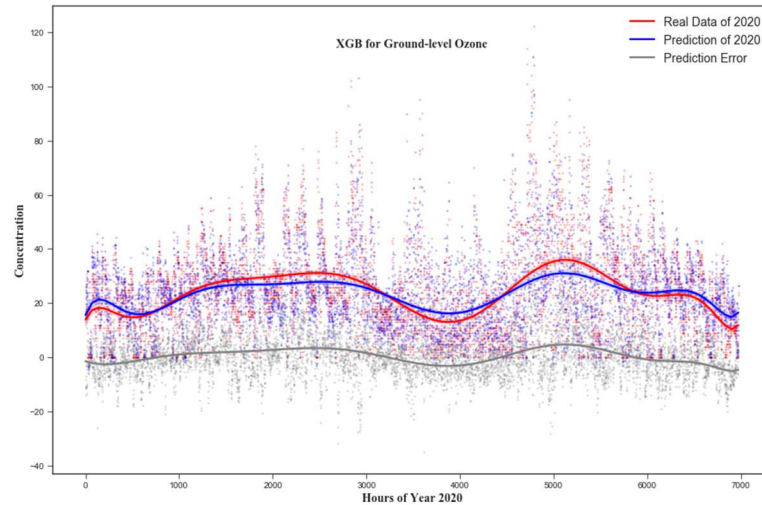
Ground-level Ozone Prediction		
Model	MLP	XGB
NRMSE	5.42%	6.50%

Distribution Plot



PVA Plot





The NRMSEs of the 2020 air pollution predictions are lower than those of the model selection process, which means the annual predictions by the models perform better than the random selection prediction. From the distribution plots, for both of the models, the prediction densities of the lower concentration value bins are lower than the real data; the prediction densities of the medium concentration value bins are higher than the real data; the prediction densities of the higher value bins tend to be more accurate and fits the real data better. Therefore, as a result, for these two models, the counts of the prediction values range from around 0 to 20 ppb is less than that of the actual measurements, and the counts of the prediction values range from around 20 to 40 ppb are more than that of the actual measurements.

From the PVA plots, the error angle of the MLP model is 2.06° , which is smaller than that of the XGB model of 2.61° . This is another sign that the MLP model is more accurate than the XGB model for ground-level ozone concentration prediction for the year 2020. The blue linear trend lines of both models are below the red lines, which means the prediction values of both models tend to be lower than the real data.

From the comparison plots, two ground-level ozone concentration seasonal peaks can be predicted by both models. For both models, the predictions tend to be smoother than the real data, which is presented as the vertical differences of the blue lines are less. The concentration trending in terms of increasing and decreasing can be perfectly predicted by both models. For 0 to 700, 3000 to 4500, and after 5800 hours, the predictions tend to be higher than the real values; for other hours, the predictions tend to be lower than the real values. Especially for the fall season concentration peak, the real values are significantly higher than the prediction. The error lines are smoother around zero before 3000 hours point, which means the predictions for the first half-year are more accurate and stable. From this result, the annual prediction of ground-level ozone based on the meteorological measurements and average traffic speeds is accurate when compared to other research.

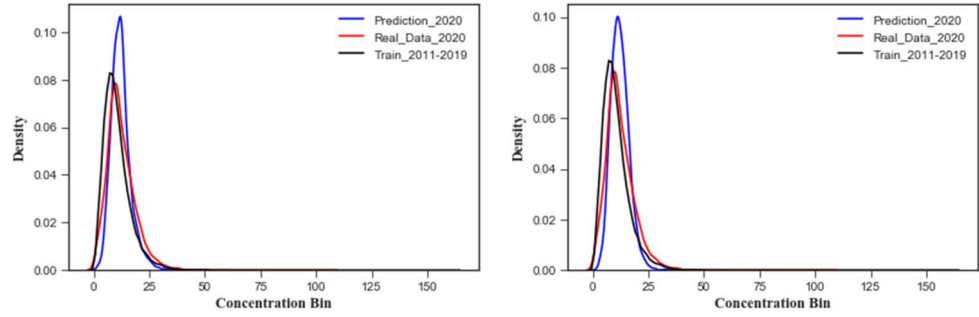
4.4.2.2 PM_{2.5} Concentration Prediction

TABLE 22 shows the PM_{2.5} concentration level prediction for the year 2020 and its validation. Based on the analysis and comparison, the NRMSE value of the RF model is 5.38%, which is lower than 5.48% of the XGB value. This means the RF model is more accurate than the XGB model for PM_{2.5} concentration level prediction.

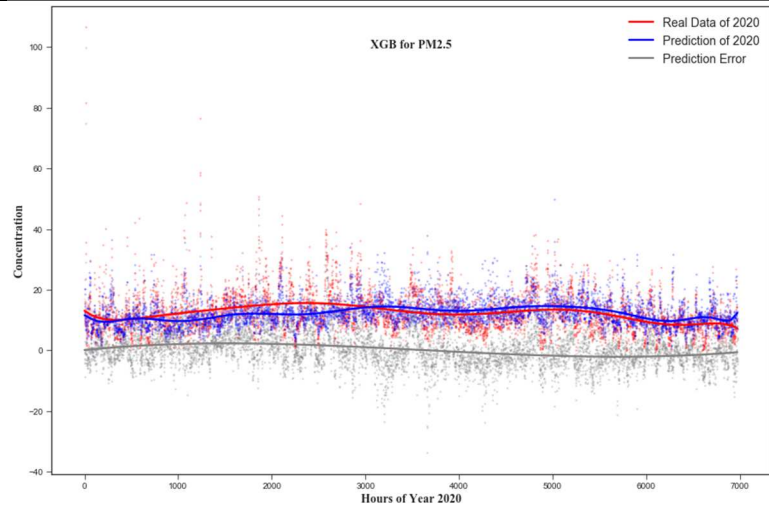
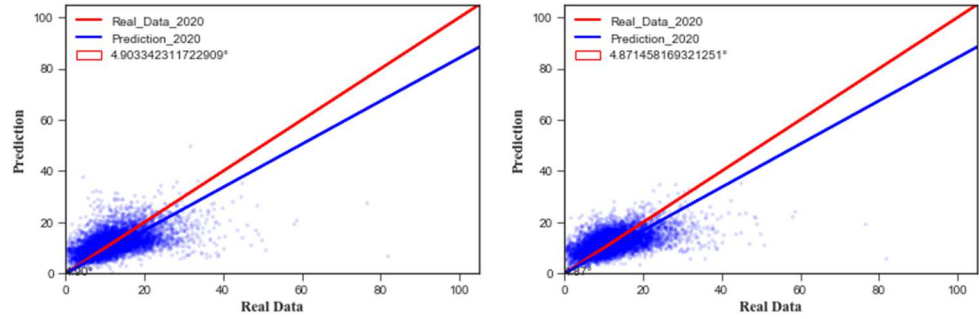
TABLE 22. PREDICTION OF PM_{2.5} CONCENTRATION

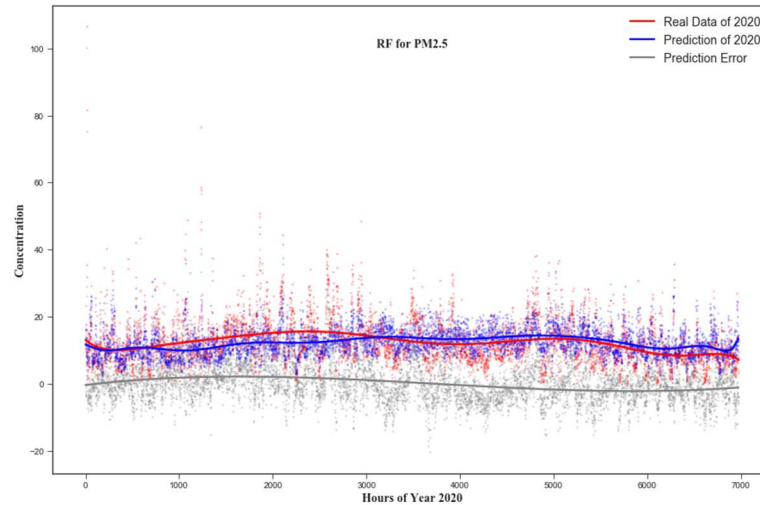
PM_{2.5} Prediction		
Model	XGB	RF
NRMSE	5.48%	5.38%

Distribution Plot



PVA Plot





The NRMSEs of the 2020 air pollution predictions are smaller than that of the model selection process, which means the annual predictions by the models perform better than the random selection prediction. From the distribution plots, for both of the models, the prediction densities of the lower concentration value bins are lower than the real data; the prediction densities of the medium concentration value bins are higher than the real data; the prediction densities of the higher value bins tend to be higher than the real data. Therefore, as a result, for these two models, the counts of the prediction values range from around 0 to 10 $\mu\text{g}/\text{m}^3$ and above 20 $\mu\text{g}/\text{m}^3$ are less than that of the actual measurements, and the counts of the prediction values range from around 10 to 15 $\mu\text{g}/\text{m}^3$ are more than that of the actual measurements.

From the PVA plots, the error angle of the RF model is 4.87° , which is smaller than that of the XGB model of 4.90° . The angles of these two models do not have big differences. However, the RF model is more accurate than the XGB model for $\text{PM}_{2.5}$ concentration prediction for the year 2020. The blue linear trend lines of both models are below the red lines, which means the prediction values of both models tend to be lower

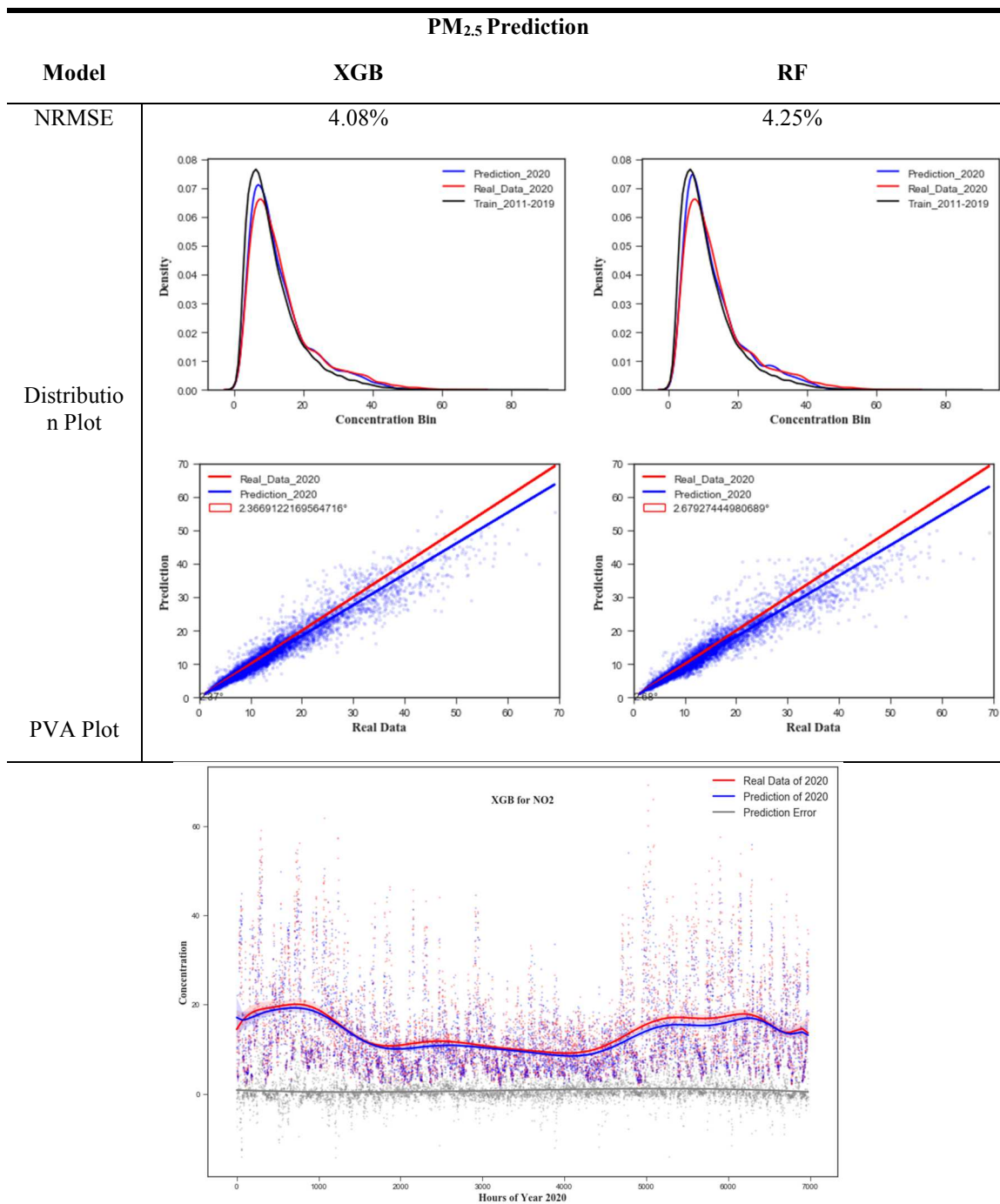
than the real data. From the scatters, the predictions tend to be lower than the real data for higher concentrations. Based on the comparison plots, the $PM_{2.5}$ concentration levels throughout the year 2020 do not have significant high or low peaks. The overall predictions are very close to the real data along the hours of the year, which can also be proved by the gray error lines that are relatively smooth and close to the $y=0$ line. However, from the error lines, the predictions of $PM_{2.5}$ for the first half of the year 2020 tend to be higher than the real data, and the second half is lower. In the meantime, the predictions of the second half of the year are more accurate with lower errors while both models failed to predict the decreasing concentration trend at the end of the year.

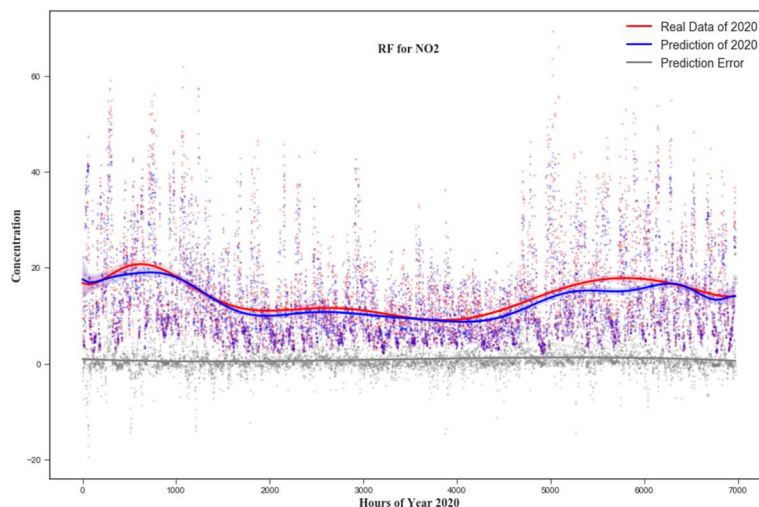
Unlike the results from the model selection process, the prediction of $PM_{2.5}$ is more accurate than the ground-level ozone. Based on the charts, the ground-level ozone concentration varies larger than the $PM_{2.5}$ concentration, which may influence the prediction accuracy. The $PM_{2.5}$ concentration levels are mainly gathered in a lower range, which can result in even there are more prediction errors, the degree of the errors can be relatively small. On the contrary, the ground-level ozone concentrations are distributed in a larger range, thus, a lower count of prediction errors can result in a relatively higher degree of error.

4.4.2.3 NO₂ Concentration Prediction

TABLE 23 shows the NO₂ concentration level prediction for the year 2020 and its validation. Based on the analysis and comparison, the NRMSE value of the XGB model is 4.08%, which is lower than 4.25% of the RF value, which is more accurate than ground-level ozone and $PM_{2.5}$ prediction.

TABLE 23. PREDICTION OF NO₂ CONCENTRATION





However, unlike the previous two air pollutants, the NRMSEs of the 2020 air pollution predictions are higher than that of the model selection process. One of the reasons is that the NO_2 pollution is largely related to on-road transportation, which was significantly influenced by the COVID-19 pandemic during the year 2020. Thus, the random selection data predictions by the models perform better than the 2020 prediction. From the distribution plots, there are relatively large differences between the 2020 real data and the historical training data for concentration bins ranging from 0 to 15 ppb. The prediction line of the XGB model lies between the training and real data lines, and the prediction line of the RF model is closer to the training data line, which is one of the reasons that the RF model is less accurate than the XGB model. For these two models, the counts of the prediction values range above 15 ppb are more accurate and closer to that of the actual measurements.

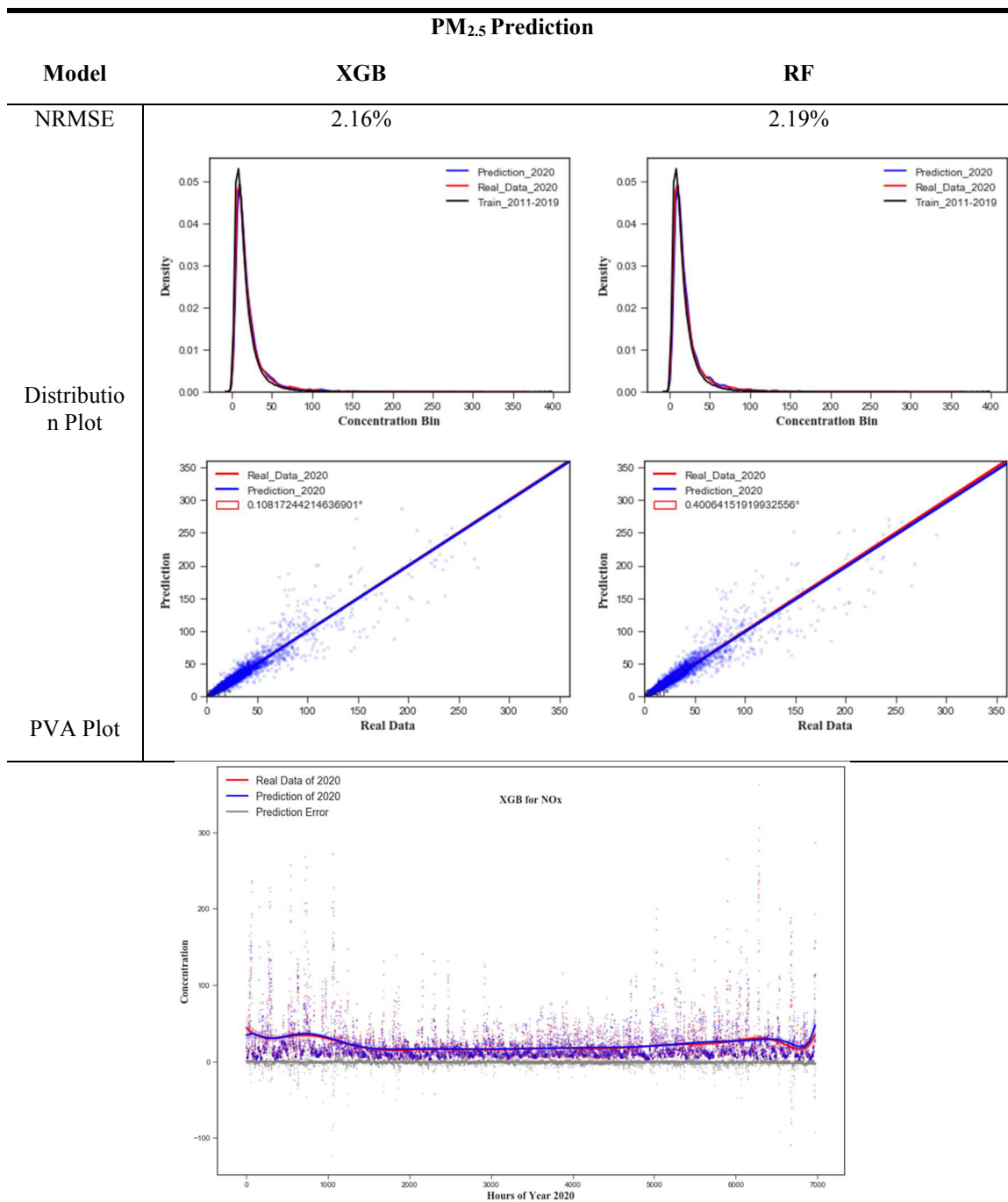
From the PVA plots, the error angle of the XGB model is 2.37° , which is smaller than that of the RF model of 2.68° . The XGB model prediction trend of NO_2 is closer to the real data than the RF model. The blue linear trend lines of both models are below the

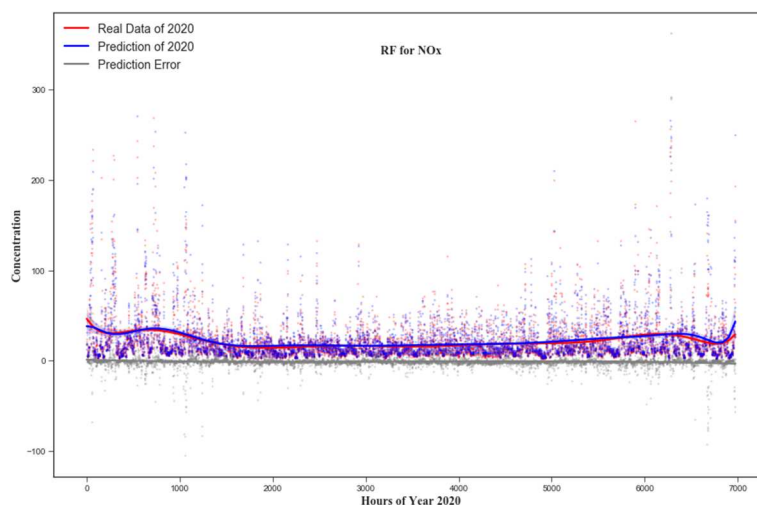
red lines, which means the prediction values of both models tend to be lower than the real data. From the scatters, the predictions tend to be lower than the real data for higher concentrations. From the hourly comparison charts, the overall NO₂ predictions are very close to the real data throughout the year. The gray error lines are almost straight lines that are close to the y=0 line. The spring and fall high peaks and the summer low values of NO₂ pollution can be accurately predicted. For the XGB model, the predictions from 0 to around 4100 hours and after 6200 hours, which include the spring high peak and the summer low value are very accurate with limited errors, while the predictions from 4100 to around 6200 hours, which include the fall peak are slightly higher than the real data. For the RF model, the predictions from around 0 to 100 hours, 1100 to 4000 hours, and after 6100 hours, which include the summer low value are relatively accurate, while the predictions from around 100 to 1100 hours and from 4000 to 6100 hours, which include the spring and fall peaks are lower than the real data. The error degree of the RF model is higher than the XGB model. This is consistent with the NRMSE values. Based on the analysis, the annual hourly NO₂ concentration can be predicted by machine learning models more accurately than ground-level ozone and PM_{2.5}.

4.4.2.4 NO_x Concentration Prediction

TABLE 24 shows the NO_x concentration level prediction for the year 2020 and its validation. Based on the analysis and comparison, the NRMSE value of the XGB model is 2.16%, which is lower than 2.19% of the RF value.

TABLE 24. PREDICTION OF NO_x CONCENTRATION





The differences between the XGB and RF models' NRMSE values are limited, which is the most accurate air pollutant concentration prediction. From the distribution plots, both of the models' predictions are extremely close to the training data, which are slightly higher than the real 2020 data at the medium concentration bins. From the PVA plots, the error angle of the XGB model is 0.108° , which is smaller than that of the RF model of 0.401° . The prediction lines of both models are very close to the 45° real data lines. From the plots, the error angles are smaller than that of the model selection process. However, the NRMSE values are higher. Based on the comparison plots, the hourly annual NO_x concentration level predictions are extremely close to the real data, which include the peak values during spring and low values during winter. The low errors can also be supported by the gray error line that is smooth and close to the $y=0$ line.

Based on the analysis and comparison, the prediction of NO_x concentration level throughout the year 2020 outperformed other air pollutants. In the meantime, ground-level ozone, $\text{PM}_{2.5}$, and NO_2 can also be predicted relatively accurately.

4.5 Discussion

The temporal characteristics that were analyzed in this dissertation are comparable with other previous studies. More specifically, part of the monthly characteristics of the target air pollutants is consistent with previous research. For instance, Logan, 1985, found a peak value of the ground-level ozone during Spring, and the minimum value in summer in the USA (Logan, 1985). Lal et al., 2000, found a higher level of ground-level ozone during Fall, in which the reason was claimed as the higher transportation activities (Lal, Naja, & Subbaraya, 2000). However, the PM_{2.5} concentration is higher in the winter season in China instead of in summer that was found in this research (Gehrig & Buchmann, 2003; Zheng et al., 2005). The differences occur due to the different sources in these two countries. The main source of PM_{2.5} in Houston is the on-road vehicle, which burns more fuel in Summer due to the extended operation of the air conditioner (Q. Li, Du, Qiao, & Yu, 2018). However, the main source of PM_{2.5} in China is the coal-burning for keeping warm in winter and the (Zheng et al., 2005). The monthly trends of NO_x/NO₂ are also consistent with previous research (Stevenson et al., 2004; Van Der A et al., 2008). The day of week characteristics is consistent with other previous research (Baan et al., 2011; Coman, Ionescu, & Candau, 2008; DeGaetano & Doherty, 2004), which show that the on-road transportation activities influence the pollution concentration on a large scale. The hourly ground-level ozone is found to be higher around the daytime (Pudasainee et al., 2006), and higher PM_{2.5}, NO_x/NO₂ concentrations during daily traffic peaks (Sonntag, Baldauf, Yanca, & Fulper, 2014; Streets & Waldhoff, 2000; Q. Xiao, Chang, Geng, & Liu, 2018), which are consistent with the results from this research. However, some subtle variances of the temporal characteristics between this research and others occurred, which are mainly due

to different region that the studies were conducted may have different influencing factors such as wind.

The correlation analysis and frequent pattern mining between the target air pollutants' concentration and the influencing factors showed some significant trends. The relationship between ground-level ozone and pressure and heavy rains that were found by the research conducted by Emanuel, 2003, in the west pacific area was not found in this research due to the different study locations (Emanuel, 2003). However, the strong correlation between ground-level ozone level and solar radiation and humidity was proved by this research. Wang et al., 2015, concluded PM_{2.5} concentration is positively correlated with temperature and negatively related with wind speed and relative humidity (J. Wang & Ogawa, 2015), which is consistent with the results from this research, while the correlation was found to be weak. The research conducted by Ocak et al., 2008, indicates the NO_x level is negatively related to wind speed and temperature that was also found in this research (Ocak & Turalioglu, 2008). The relationship between the target air pollution and transportation is relatively weak because the transportation data is not detailed enough to show the hourly and daily trends. The relatively higher correlation between PM_{2.5} and NO_x concentration that was found in this research was also found by previous studies. For instance, Song et al., 2019, Li et al., 2020, and Buccolieri et al., 2018 all found the same relationships between those two air pollutants (Buccolieri, Jeanjean, Gatto, & Leigh, 2018; L. Li et al., 2020; Song et al., 2019). Some previous research found a positive relationship between NO_x and ground-level ozone relationship in contrast to this research (David & Nair, 2011; Duan et al., 2008). In the meantime, a negative relationship was also found by some (Varotsos, Efstathiou, & Kondratyev, 2003). For example, Pancholi et al., 2018,

discovered that the O_3 and NO_x during daytime exhibited an inverse relationship (Pancholi, Kumar, Bikundia, & Chourasiya, 2018). Yu et al., 2020, concluded that the correlation between ground-level ozone and NO_x concentration can be both positive and negative due to the complexity of ozone precursors sensitivity (Yu et al., 2020).

The target air pollution forecasting based on machine learning outperformed most of the existing forecast models. Mallet et al., 2009, applied machine learning algorithms to perform ozone forecasting, in which the RMSE was relatively low (Mallet, Stoltz, & Mauricette, 2009). While the machine learning models utilized in that research require tuning on parameters that limit their efficiency. The forecasting model built by Tang et al., 2011, is able to provide a relatively accurate prediction for ozone, which, however, requires the input of NO_x and VOCs information (Tang, Zhu, Wang, & Gbaguidi, 2011). This research predicts the ground-level ozone by influencing factors, which eliminate the possible errors caused by other pollutants' predictions. The annual hourly forecasting ability of the models built in this research is also superior to the majority of the previous studies. For instance, the ozone forecasting model based on the MLP algorithm by Dutot et al., 2007, utilized a 24-hour lead time, which is not able to forecast a further future. The forecasting accuracy of $PM_{2.5}$ by the model built in this research is comparable if not more accurate than others. Zhou et al., 2018, utilized time-series input to develop an SVM model, in which the influencing factors were not considered (Zhou et al., 2019). This is a common practice for most models built by existing research (Mahajan, Chen, & Tsai, 2018; Perez, Menares, & Ramírez, 2020; Zhu et al., 2018). The lack of influencing factors in the forecasting may result in a less robust prediction. When comparing with the traditional

statistical forecasting method (Du, 2018), the accuracy of the models proposed in this research is more accurate, practical, and efficient, especially for special events.

CHAPTER 5

SUMMARY AND RECOMMENDATIONS

5.1 Summary

Ground-level ozone and PM_{2.5} are known for their severe adverse effects on human health and the ecosystems as critical air pollutants. Unlike the primary air pollutants that are emitted to the atmosphere directly from the sources, ground-level ozone is a type of secondary air pollutant that is formed in the air by reactions of chemicals such as NO_x that involves the presence of sunlight. When people are exposed to the high ground-level ozone or PM_{2.5} concentration levels, respiratory and cardio-pulmonary diseases, and lung cancer mortality might be triggered, or even death, especially for the high-risk population such as those who have asthma. Furthermore, flora and fauna species can be damaged by them. This research aims to analyze the ten years of ground-level ozone and PM_{2.5} historical concentration data along with the meteorological measurements and traffic situation and propose a practical forecasting approach.

For these research purposes, an in-depth literature review was extensively conducted. Different types, sources, and consequences of pollutions were reviewed and summarized following by the emphatical introduction of the four critical air pollutants, which include ground-level ozone, PM_{2.5}, and NO_x/NO₂ that are selected as the research targets. Previous research on the pathogenicity of these pollutants was concluded along with various regional regulations and laws, which gave information on how human health is threatened. The influencing factors of the air pollutants are reviewed and categorized into two classes, which include meteorological factors such as sunlight, temperature, wind, relative humidity, and the source factor that mainly on-road traffic situations. It is revealed

from previous research that these factors may impact the target air pollutants' concentration level to a certain degree, albeit the specific effects need to be analyzed case by case. The different approaches of air pollution forecasting technologies and approaches were thoroughly reviewed. Based on the review, the mainstream forecasting approaches utilized the time-series analysis on the historical air pollution data. More advanced analytical tools such as NN and machine learning in the existing research tend to overdependence on the time-series input, which limits the applicability, accuracy, and efficiency of the proposed models. However, the literature review provided an insight into various analytical and forecasting approaches that can be used in this research.

To achieve the research goals, four modules were purposed that include data collection and processing, temporal analysis, frequent pattern mining analysis, and machine learning forecasting. In the data collection and processing module, the ten years of air pollution concentration, meteorological, and traffic situation data were introduced along with the detailed data type and the information entries that were collected. The preprocessing of the raw data to eliminate the invalid records was also demonstrated. In the temporal analysis module, the statistical analysis that includes the ANOVA and correlation tests and the temporal analysis that includes the annual, monthly, day-of-week, and hourly characteristics were introduced. In the frequent pattern mining analysis, various datasets were integrated and transformed into a unified dataset. The dataset was then binned into binary entries only. The frequent pattern algorithms include the Apriori and FP-Growth were introduced in this module based on the detailed theory, parameters, and pseudo-code that were utilized in this research. In the machine learning forecasting module, the concepts of using it in data analysis were shown in a detailed flow chart. The seven

most representative machine learning algorithms including PR, MLP, XGB, SVM, RF, LR, and KNN were demonstrated with the detailed computation kernels. The evaluation methods that include the NRMSE and PVA plot of the forecasted values were provided in detailed calculation equations and concepts.

As the analytical results of this research, the yearly, monthly, day-of-week, and hourly temporal characteristics were demonstrated for ground-level ozone, PM_{2.5}, NO_x/NO₂, temperature, solar, pressure, precipitation, relative humidity, and resultant wind speed and direction. The yearly characteristics of the traffic situation in terms of average on-road speeds were analyzed. As shown in the intercorrelation matrix and the Pearson's *r* matrix, ground-level ozone is relatively more correlated with relative humidity, average traffic speed, solar radiation, temperature, and resultant wind speed. PM_{2.5} is relatively more correlated with pressure and temperature. NO₂ is relatively more correlated with solar radiation, temperature, resultant wind speed, and pressure. NO_x is relatively more correlated with temperature, resultant wind speed, and pressure. The intercorrelation of the air pollutants and meteorological parameters was also calculated. Based on the frequent pattern mining analysis, the highest concentration level of ground-level ozone appeared when the solar radiation was higher than 0.414 Langley/min, the temperature was higher than 81.4 F, the pressure below 1013.1 millibars, no precipitation, relative humidity below 57.9%, resultant wind speed higher than 8.3 mph with the direction of 165 to 222-degree compass and the average traffic speed between 55 and 60 mph. The highest concentration level of PM_{2.5} appeared when the solar radiation was below 0.01 Langley/min, the temperature between 73.8 and 81.4 F, the pressure below 1013.1 millibars, no precipitation, relative humidity between 73.8% and 83.8%, resultant wind speed higher than 8.3 mph

with the direction of 165 to 222-degree compass and the average traffic speed between 55 and 60 mph. The highest concentration level of NO_2 and NO_x appeared when the solar radiation was below 0.01 Langleys/min, the temperature below 62.5 F, the pressure higher than 1019.6 millibars, no precipitation, relative humidity below 57.9%, resultant wind speed below 3.4 mph with the direction of 222 to 360-degree compass and the average traffic speed between 55 and 60 mph.

Based on the algorithm selection for machine learning forecasting, MLP and XGB outperformed other algorithms with the lowest NRMSE and error angles in the PVA plots, and were selected as the candidate algorithm for ground-level ozone prediction, XGB and RF were selected to perform $\text{PM}_{2.5}$ prediction, XGB and RF were selected to perform NO_2 and NO_x prediction. From the prediction evaluation process that utilized the actual values from the year 2020, the MLP model provided accurate forecasting for annual hourly ground-level ozone forecasting with NRMSE of 5.42% and the error angle of 2.06° . The RF model provided relatively accurate forecasting for annual hourly $\text{PM}_{2.5}$ concentration with NRMSE of 5.38% and the error angle of 4.87° . While the forecasting of $\text{PM}_{2.5}$ concentration was not as accurate as of the other target air pollutants in terms of values, however, the trends prediction is of high accuracy and outperformed most of the previous research. The XGB model provided relatively accurate forecasting for annual hourly NO_2 and NO_x concentrations with NRMSE of 4.08% and 2.16% and the error angle of 0.16° , respectively. From the forecasting evaluation, the predictions of ground-level ozone and NO_x/NO_2 were very accurate in both values and trends.

In conclusion, when comparing with the previous studies, the main advantages of the machine learning forecasting models that were proposed in this research include: (1)

the relatively long-term (annual) detailed (hourly) prediction ability; (2) the higher accuracy in the trends and values forecasting; (3) the air pollution influencing factors are fully considered and utilized; (4) less dependency on the time-series inputs; and (5) the high-efficiency forecasting procedure that can be easily used in other similar areas.

5.2 Contributions

This research was conducted based on the cutting edge Frequent Pattern Data Mining and Machine Learning technologies to achieve the research goals. The methodologies and results of this research made contributions to the state of knowledge in air pollution characterization and forecasting in several aspects. Firstly, ten years' historical data is analyzed, which is superior to most other research. Furthermore, the raw data contains a complementing range of variables that include seven meteorological measurements (solar radiation, outdoor temperature, barometric pressure, precipitation, relative humidity, resultant wind speed, and resultant wind direction) and traffic situation data for each air pollution. This provides fundamental databases for an accurate model.

Secondly, the complex nature of the impacts of the influencing factors on air pollution concentration level determined that it is hard to be analyzed by a traditional statistic method. In this research, the binning method in the preprocessing transformed the float type raw data into discrete integer labels, and the relationships and trends between meteorological measurements, traffic situation, and the target air pollution concentration are revealed through the Frequent Pattern Data Mining algorithms.

Thirdly, a universal methodology to create and evaluate a pool of Machine Learning forecasting models for air pollution is proposed and tested. The input data is split and each forecasting model is self and cross-compared to boost the accuracy, which is high in efficiency and the computation complexity is greatly reduced. The forecasting models considered both anthropogenic and natural influencing factors of the air pollution concentration, thus, it will be more applicative and robust when compared with the forecasting models that only consider the air pollution data. As a result, the methodology proposed in this research can be used in other areas and other variables.

5.3 Recommendations

This research collected and analyzed ten years of ground-level ozone, PM_{2.5}, NO_x/NO₂ concentration, seven meteorological measurements, and traffic situation data from 2011 to 2020. The results revealed detailed pollution patterns associated with the influencing factors and their forecasting models based on machine learning algorithms. Albeit the achievements of this research met the objectives, there are several recommendations for future research.

(1) The data collected were in diverse formats and quality, which lower the preprocessing efficiency and precision. A good way to improve efficiency is to create and maintain a database through, for instance, Amazon AWS or MySQL, so that all data formats could be integrated automatically.

(2) The traffic situation data collected lacks the hourly measurements, which were not detailed enough. In future research, it is recommended to access a more detailed traffic database that may further improve the forecasting accuracy.

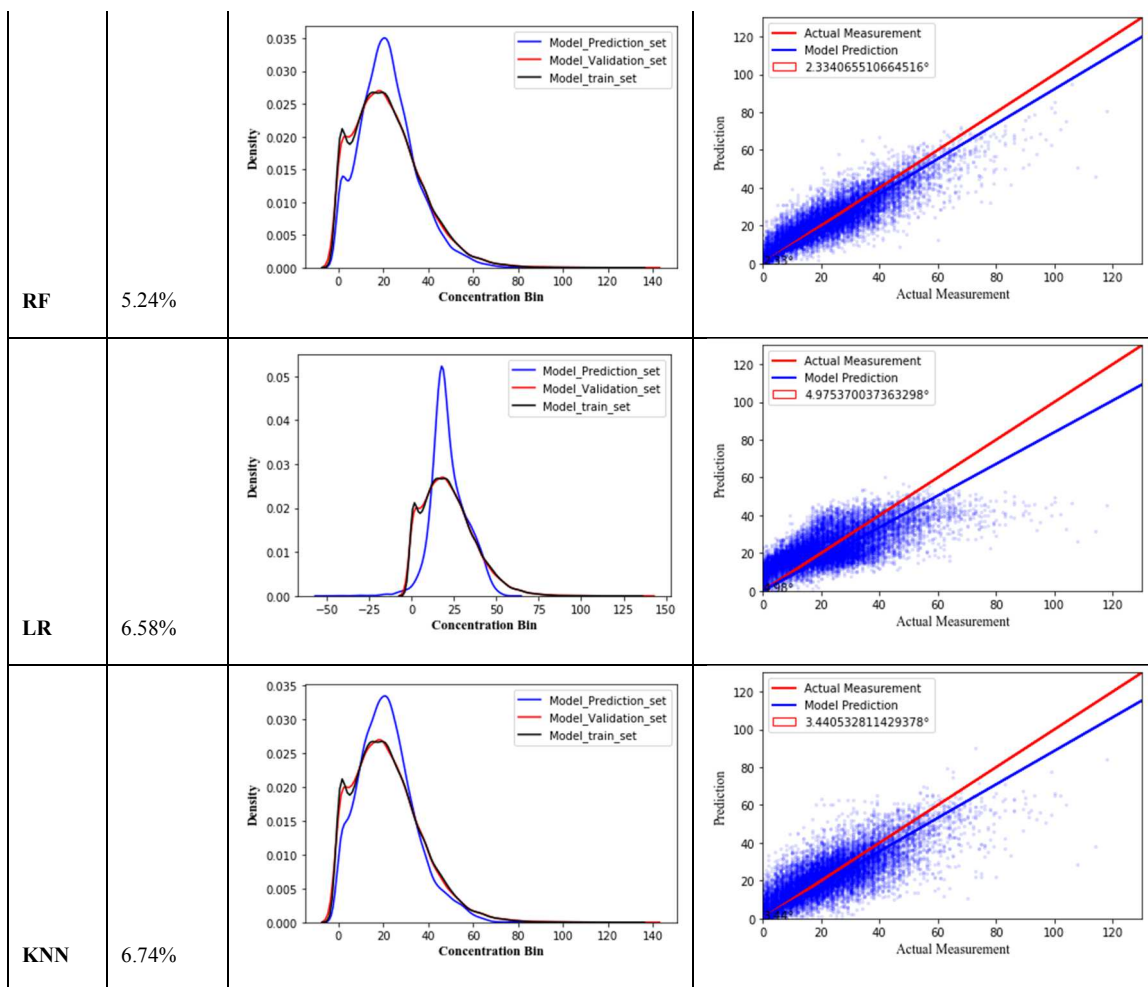
(3) The target air pollutants in this research are mainly anthropogenic sourced, thus, human activities play an important role in influencing their concentration. However, this research solely considered the traffic situation, which is not adequate. In future research, other human activities should be considered. For example, economic activities should be considered by electricity consumption parameters in an area.

(4) The forecasting of air pollution concentration based on machine learning models was evaluated on a large scale for the whole predicted dataset, the hour-by-hour accuracy is not measured. Thus, the single point accuracy may not be as good as the whole model. There are still rooms to improve the machine learning models, which may yield more accurate forecasting.

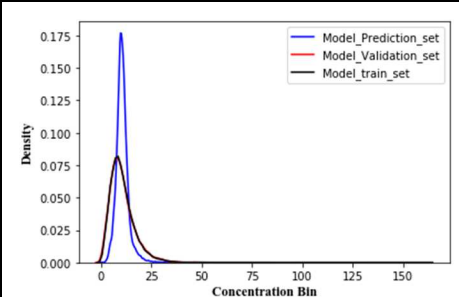
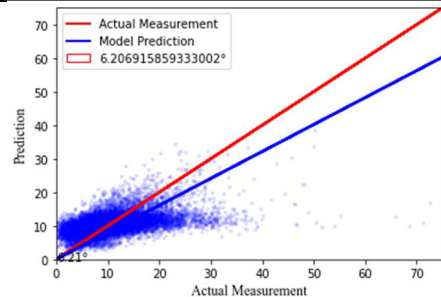
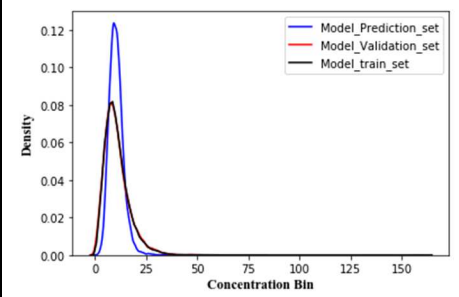
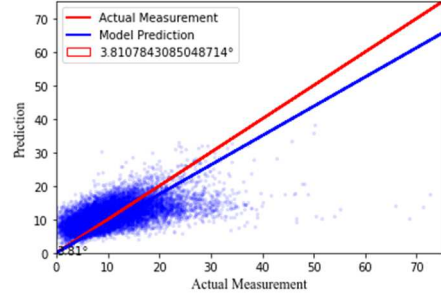
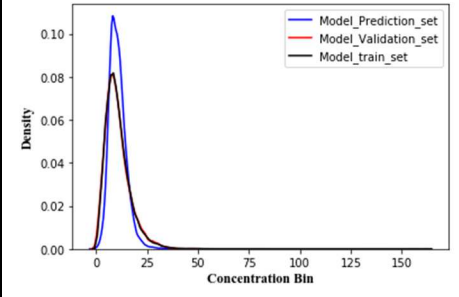
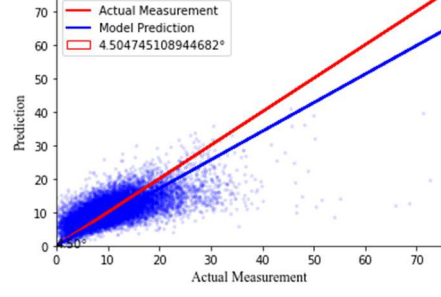
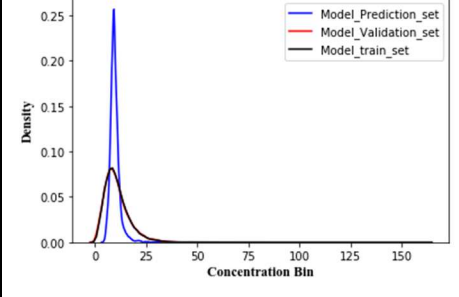
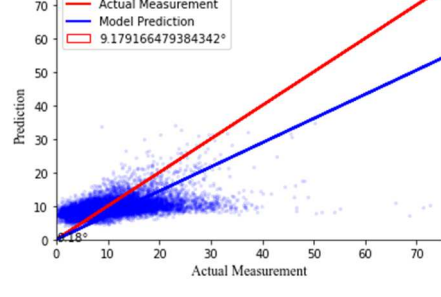
APPENDIX

A. MACHINE LEARNING MODEL SELECTION FOR GROUND-LEVEL OZONE

	NRMSE	Ground-level Ozone	
PR	5.68%		
MLP	4.81%		
XGB	5.07%		
SVM	6.72%		

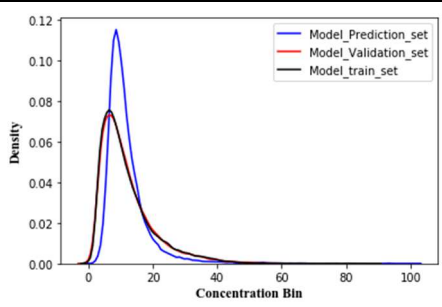
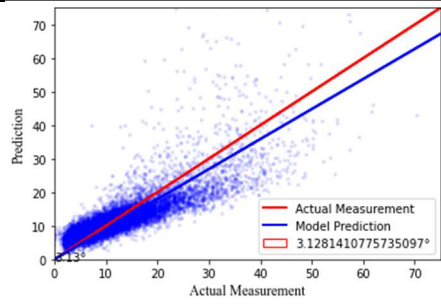
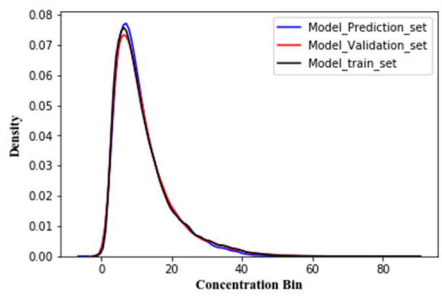
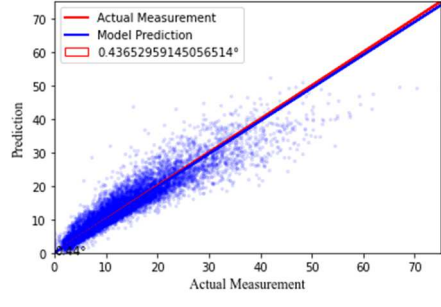
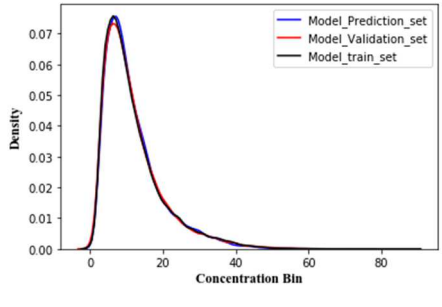
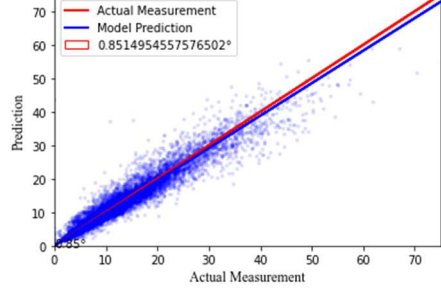
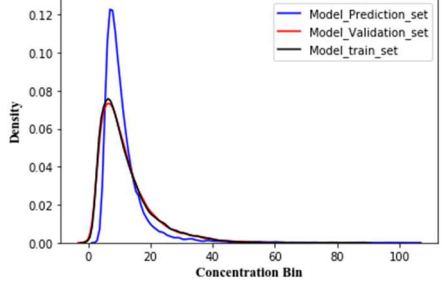
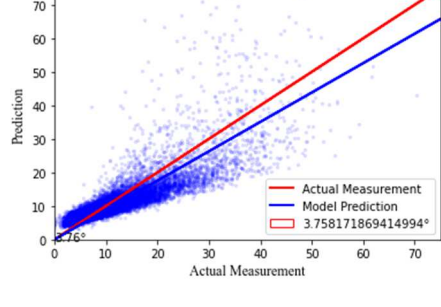


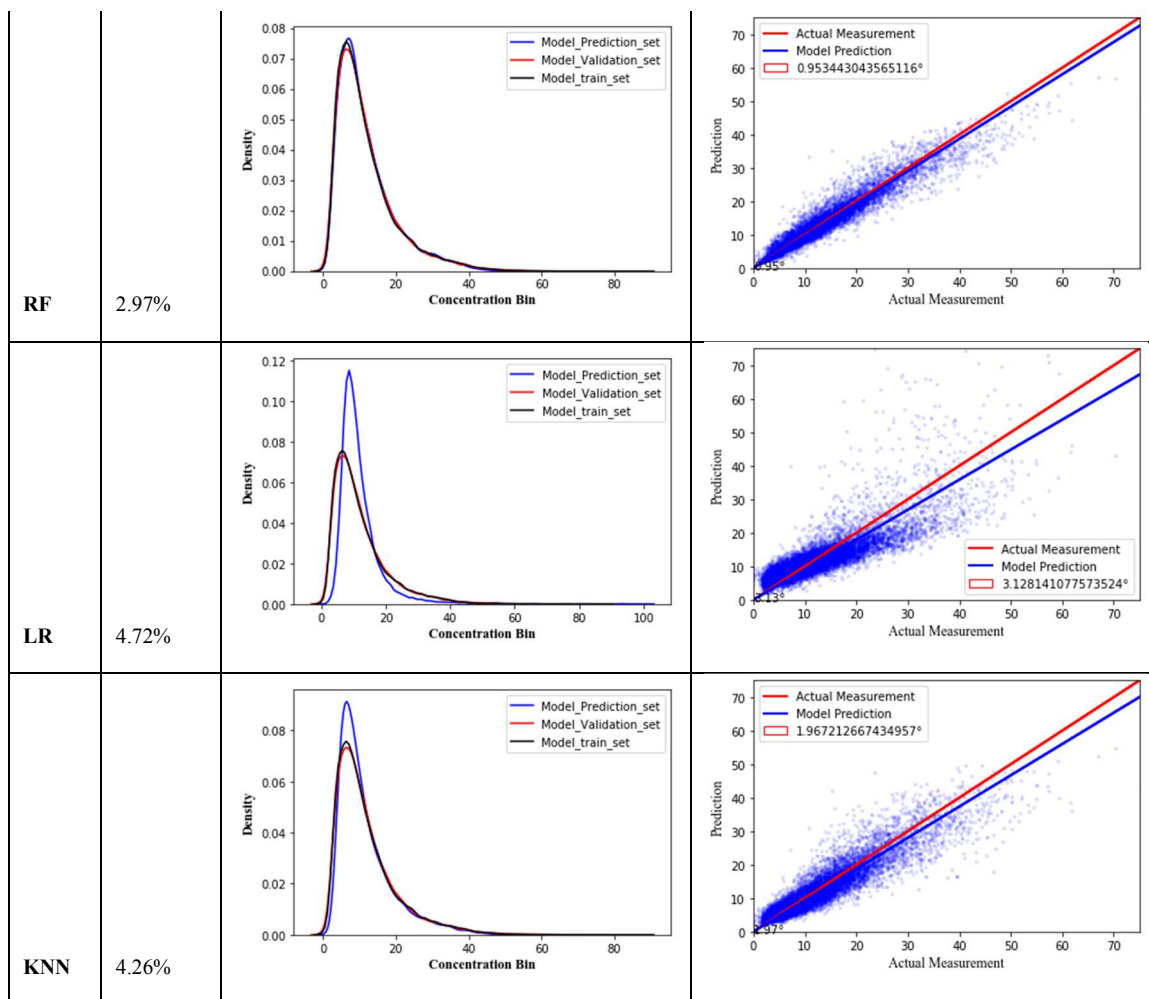
B. MACHINE LEARNING MODEL SELECTION FOR PM_{2.5}

	NRMSE	PM2.5
PR	6.58%	 
MLP	6.31%	 
XGB	5.62%	 
SVM	6.91%	 

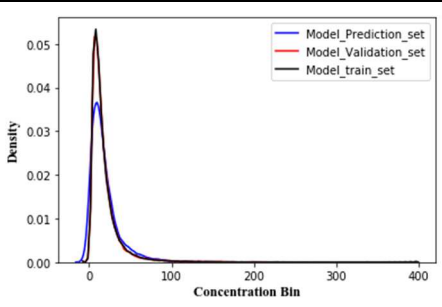
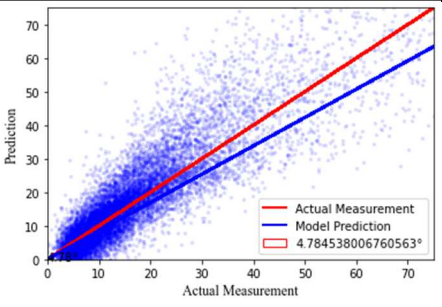
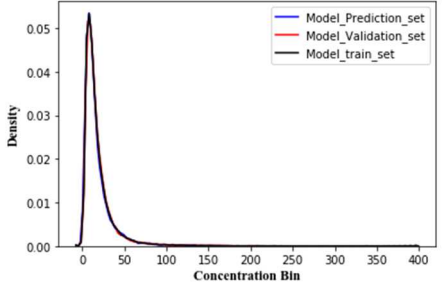
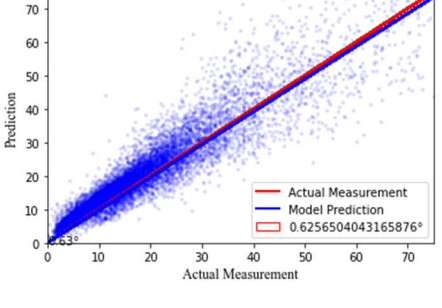
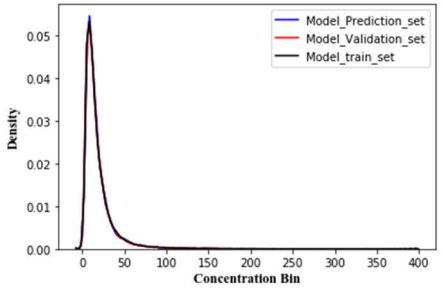
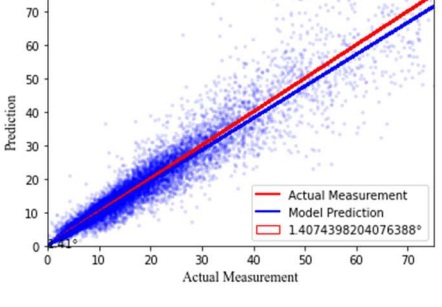
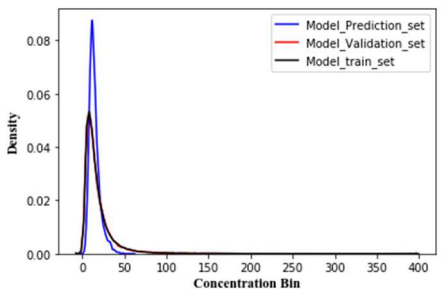
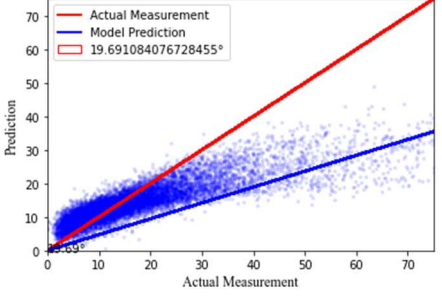
RF	5.46%		
LR	6.58%		
KNN	6.32%		

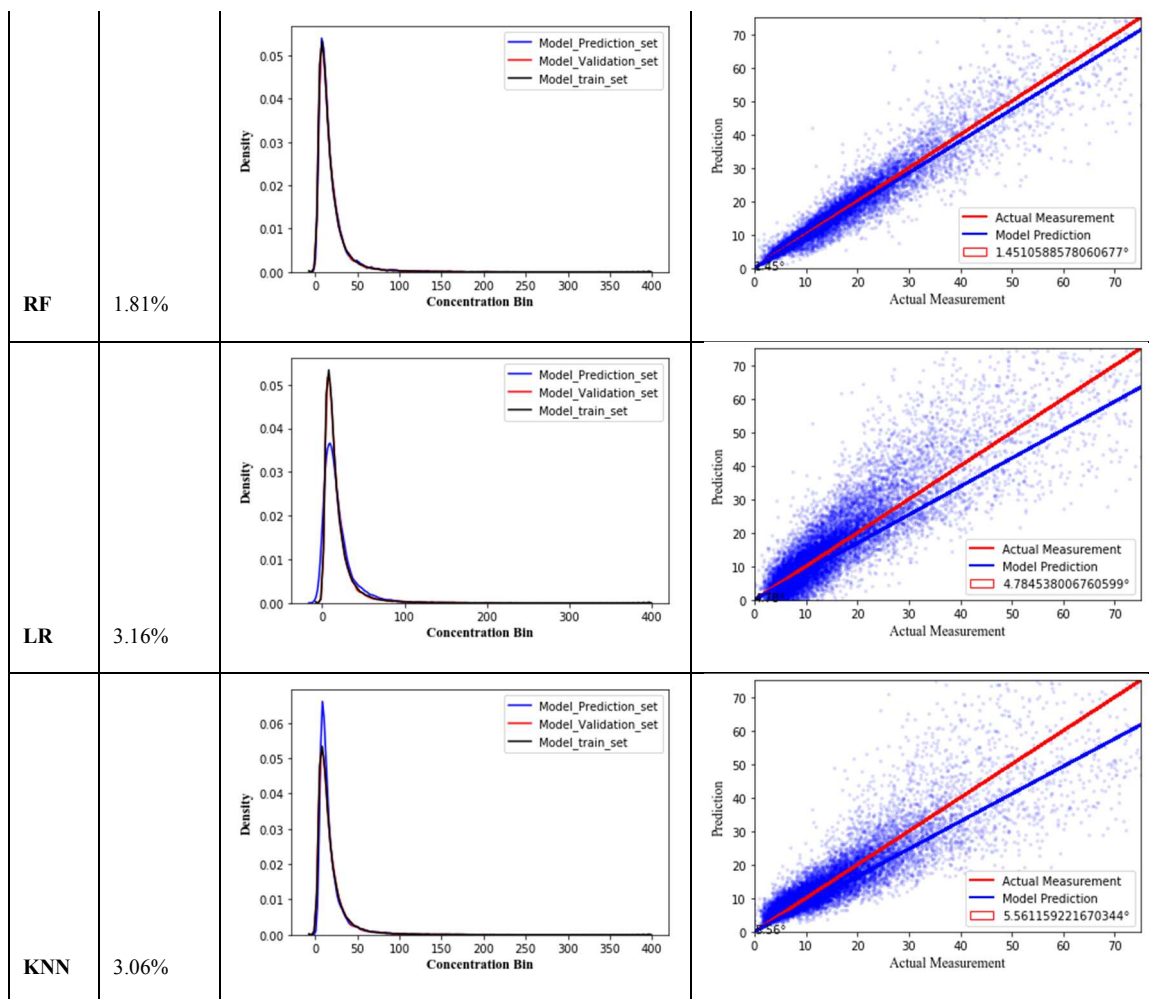
C. MACHINE LEARNING MODEL SELECTION FOR NO₂

	NRMSE	NO ₂
PR	4.72%	 
MLP	3.12%	 
XGB	2.95%	 
SVM	4.71%	 



D. MACHINE LEARNING MODEL SELECTION FOR NO_x

	NRMSE	NO _x
PR	2.19%	 
MLP	2.18%	 
XGB	1.79%	 
SVM	5.02%	 



REFERENCES

- Aggarwal, C. C., Bhuiyan, M. A., & Al Hasan, M. (2014). Frequent pattern mining algorithms: A survey. In *Frequent pattern mining* (pp. 19-64). Cham: Springer.
- Ahlbom, A., & Feychting, M. (2003). Electromagnetic radiation: Environmental pollution and health. *British medical bulletin*, *68*(1), 157-165.
- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.
- Amr, A., & Hadidi, N. (2001). Effect of cultivar and harvest date on nitrate (NO₃) and nitrite (NO₂) content of selected vegetables grown under open field and greenhouse conditions in Jordan. *Journal of food composition and analysis*, *14*(1), 59-67.
- ATSDR. (2021). *Carbon Monoxide*. Retrieved from <https://www.cdc.gov/TSP/substances/ToxSubstance.aspx?toxid=253>
- Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., . . . Zainuddin, S. F. M. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, & Soil Pollution*, *225*(8), 1-14.
- Baan, R., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., . . . Straif, K. (2011). Carcinogenicity of radiofrequency electromagnetic fields. *The lancet oncology*, *12*(7), 624-626.
- Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., & Dominici, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987-2000. *Jama*, *292*(19), 2372-2378.

- Bellinger, C., Mohamed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health, 17*(1), 907. doi:10.1186/s12889-017-4914-3
- Benedick, R. E. (1998). *Ozone diplomacy*: Harvard University Press.
- Breyse, P. N., Delfino, R. J., Dominici, F., Elder, A. C., Frampton, M. W., Froines, J. R., . . . Hopke, P. K. (2013). US EPA particulate matter research centers: summary of research results for 2005–2011. *Air Quality, Atmosphere & Health, 6*(2), 333-355.
- Brownlee, J. (2019). *Your First Machine Learning Project in Python Step-By-Step*. Retrieved from <https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>
- Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *The Lancet, 360*(9341), 1233-1242.
- Buccolieri, R., Jeanjean, A. P., Gatto, E., & Leigh, R. J. (2018). The impact of trees on street ventilation, NO_x and PM_{2.5} concentrations across heights in Marylebone Rd street canyon, central London. *Sustainable Cities and Society, 41*, 227-241.
- Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software, 119*, 285-304. doi:<https://doi.org/10.1016/j.envsoft.2019.06.014>
- Caplan, L. S., Schoenfeld, E. R., O'Leary, E. S., & Leske, M. C. (2000). Breast cancer and electromagnetic fields—a review. *Annals of Epidemiology, 10*(1), 31-44.

- Cardelino, C., & Chameides, W. (1995). An observation-based model for analyzing ozone precursor relationships in the urban atmosphere. *Journal of the Air & Waste Management Association*, 45(3), 161-180.
- Casey, J. A., Morello-Frosch, R., Mennitt, D. J., Fristrup, K., Ogburn, E. L., & James, P. (2017). Race/ethnicity, socioeconomic status, residential segregation, and spatial variation in noise exposure in the contiguous United States. *Environmental health perspectives*, 125(7), 077017.
- CDC. (2021). *Air Pollutants*. Retrieved from <https://www.cdc.gov/air/pollutants.htm>
- Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., . . . Sun, K. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489), 395-400.
- Cinzano, P., Falchi, F., & Elvidge, C. D. (2001). The first world atlas of the artificial night sky brightness. *Monthly Notices of the Royal Astronomical Society*, 328(3), 689-707.
- CIRP. (2020). Nondimensional Statistics. *Statistics*. Retrieved from https://cirpwiki.info/wiki/Statistics#Nondimensional_Statistics
- Colville, R., Hutchinson, E. J., Mindell, J., & Warren, R. (2001). The transport sector as a source of air pollution. *Atmospheric Environment*, 35(9), 1537-1565.
- Coman, A., Ionescu, A., & Candau, Y. (2008). Hourly ozone prediction for a 24-h horizon using neural networks. *Environmental Modelling & Software*, 23(12), 1407-1421.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

- Cunningham, S. D., Berti, W. R., & Huang, J. W. (1995). Phytoremediation of contaminated soils. *Trends in biotechnology*, 13(9), 393-397.
- Curran, M. A. (2012). *Life cycle assessment handbook: a guide for environmentally sustainable products*: John Wiley & Sons.
- David, L. M., & Nair, P. R. (2011). Diurnal and seasonal variability of surface ozone and NO_x at a tropical coastal site: Association with mesoscale and synoptic meteorological conditions. *Journal of Geophysical Research: Atmospheres*, 116(D10).
- Davidson, B., & Bradshaw, R. W. (1967). Thermal pollution of water systems. *Environmental Science & Technology*, 1(8), 618-630.
- DeGaetano, A. T., & Doherty, O. M. (2004). Temporal, spatial and meteorological variations in hourly PM_{2.5} concentration extremes in New York City. *Atmospheric Environment*, 38(11), 1547-1558.
- Dhami, A. (2012). Study of electromagnetic radiation pollution in an Indian city. *Environmental monitoring and assessment*, 184(11), 6507-6512.
- Ding, A., Wang, T., Zhao, M., Wang, T., & Li, Z. (2004). Simulation of sea-land breezes and a discussion of their implications on the transport of air pollution during a multi-day ozone episode in the Pearl River Delta of China. *Atmospheric Environment*, 38(39), 6737-6750.
- Donges, N. (2020). A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM. Retrieved from <https://builtin.com/data-science/random-forest-algorithm>

- Du, J. (2018). *Temporal Characteristics of Particulate Matter 2.5 Concentration and Their Correlations with Weather Condition and Traffic Volume*. Texas Southern University,
- Du, J., Li, Q., Qiao, F., & Yu, L. (2018). Estimation of vehicle emission on mainline freeway under isolated and integrated ramp metering strategies. *Environmental Engineering and Management Journal*, 17(5), 1237-1248.
- Du, J., Qiao, F., & Yu, L. (2019). Temporal characteristics and forecasting of PM2.5 concentration based on historical data in Houston, USA. *Resources, Conservation and Recycling*, 147, 145-156. doi:<https://doi.org/10.1016/j.resconrec.2019.04.024>
- Du, J., Wang, H., & Qiao, F. (2020). Transportation-Related Toxic Emissions Influenced by Public Reactions to the COVID-19 Pandemic. *Journal of Environmental and Toxicological Studies*, 4(1).
- Duan, J., Tan, J., Yang, L., Wu, S., & Hao, J. (2008). Concentration, sources and ozone formation potential of volatile organic compounds (VOCs) during ozone episode in Beijing. *Atmospheric Research*, 88(1), 25-35.
- 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. doi:<https://doi.org/10.5094/APR.2014.079>
- Emanuel, K. (2003). Tropical cyclones. *Annual review of earth and planetary sciences*, 31(1), 75-104.
- EPA. (2010). *Quantitative Health Risk Assessment for Particulate Matter Research*
Triangle Park

- EPA. (2019). *Air Emissions Sources*. Retrieved from <https://www.epa.gov/transportation-air-pollution-and-climate-change/smog-soot-and-local-air-pollution>
- EPA. (2020). *Ground-level Ozone Pollution*. Retrieved from <https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics>
- EPA. (2021a). *Nitrogen Oxides (NOx) Control Regulations*. Retrieved from <https://www3.epa.gov/region1/airquality/nox.html>
- EPA. (2021b). *Primary National Ambient Air Quality Standards (NAAQS) for Nitrogen Dioxide*. Retrieved from <https://www.epa.gov/no2-pollution/primary-national-ambient-air-quality-standards-naaqs-nitrogen-dioxide>
- EPA. (2021c). *Smog, Soot, and Other Air Pollution from Transportation*. Retrieved from <https://www.epa.gov/transportation-air-pollution-and-climate-change/smog-soot-and-local-air-pollution>
- Falchi, F., Cinzano, P., Elvidge, C. D., Keith, D. M., & Haim, A. (2011). Limiting the impact of light pollution on human health, environment and stellar visibility. *Journal of Environmental Management*, 92(10), 2714-2722.
- FHWA. (2017). *Annual Vehicle Distance Traveled in Miles and Related Data - 2016 by Highway Category and Vehicle Type*. Retrieved from <https://www.fhwa.dot.gov/policyinformation/statistics/2016/vm1.cfm>
- Fix, E. (1985). *Discriminatory analysis: nonparametric discrimination, consistency properties* (Vol. 1): USAF school of Aviation Medicine.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1): Springer series in statistics New York.

- Gao, J., Wang, T., Ding, A., & Liu, C. (2005). Observational study of ozone and carbon monoxide at the summit of mount Tai (1534 m asl) in central-eastern China. *Atmospheric Environment*, 39(26), 4779-4791.
- Gehrig, R., & Buchmann, B. (2003). Characterising seasonal variations and spatial distribution of ambient PM10 and PM2.5 concentrations based on long-term Swiss monitoring data. *Atmospheric Environment*, 37(19), 2571-2580.
- Gent, J. F., Triche, E. W., Holford, T. R., Belanger, K., Bracken, M. B., Beckett, W. S., & Leaderer, B. P. (2003). Association of low-level ozone and fine particles with respiratory symptoms in children with asthma. *Jama*, 290(14), 1859-1867.
- Goel, P. (2006). *Water Pollution-Causes, Effects and Control*. New Delhi: New Age International. Retrieved from
- Gössling, S., Scott, D., & Hall, C. M. (2020). Pandemics, tourism and global change: a rapid assessment of COVID-19. *Journal of Sustainable Tourism*, 1-20.
- Gray, R. S. (2020). Agriculture, transportation, and the COVID-19 crisis. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 68(2), 239-243.
- Guttorp, P., Meiring, W., & Sampson, P. D. (1994). A space-time analysis of ground-level ozone data. *Environmetrics*, 5(3), 241-254. doi:10.1002/env.3170050305
- Haines, A., Kovats, R. S., Campbell-Lendrum, D., & Corvalán, C. (2006). Climate change and human health: impacts, vulnerability and public health. *Public health*, 120(7), 585-596.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

- Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1), 53-87.
- Hardin, M., & Kahn, R. (1999). Aerosols & Climate Change. *Earth Observatory*.
- Introduction to Boosted Trees. (2020). Retrieved from <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>
- Jaskowiak, P. A., & Campello, R. (2011). *Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data*. Paper presented at the Proceedings of the Brazilian symposium on bioinformatics.
- Jeuken, A. B. M. (2000). *Evaluation of chemistry and climate models using measurements and data assimilation*: Technische Universiteit Eindhoven.
- Kirchner, D. B., Evenson, E., Dobie, R. A., Rabinowitz, P., Crawford, J., Kopke, R., & Hudson, T. W. (2012). Occupational noise-induced hearing loss: ACOEM task force on occupational hearing loss. *Journal of occupational and environmental medicine*, 54(1), 106-108.
- Koh, J. L., & Shieh, S. F. (2004). *An efficient approach for maintaining association rules based on adjusting FP-tree structures*. Paper presented at the International Conference on Database Systems for Advanced Applications, Springer, Berlin, Heidelberg.
- Kurt, A., Gulbagci, B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environment International*, 34(5), 592-598.

- Lake, G. (1970). Ozone cracking and protection of rubber. *Rubber Chemistry and Technology*, 43(5), 1230-1254.
- Lal, S., Naja, M., & Subbaraya, B. (2000). Seasonal variations in surface ozone and its precursors over an urban site in India. *Atmospheric Environment*, 34(17), 2713-2724.
- Landrigan, P. J., Fuller, R., Acosta, N. J., Adeyi, O., Arnold, R., Baldé, A. B., . . . Breyse, P. N. (2018). The Lancet Commission on pollution and health. *The Lancet*, 391(10119), 462-512.
- Laws, E. A. (2017). *Aquatic pollution: an introductory text*: John Wiley & Sons.
- Li, H., Wang, Y., Zhang, D., Zhang, M., & Chang, E. Y. (2008). *Pfp: parallel fp-growth for query recommendation*. Paper presented at the Proceedings of the 2008 ACM conference on Recommender systems.
- Li, J. (2017). Boosting algorithm: XGBoost. Retrieved from <https://towardsdatascience.com/boosting-algorithm-xgboost-4d9ec0207d>
- Li, L., Lu, C., Chan, P.-W., Zhang, X., Yang, H.-L., Lan, Z.-J., . . . Zhang, L. (2020). Tower observed vertical distribution of PM_{2.5}, O₃ and NO_x in the Pearl River Delta. *Atmospheric Environment*, 220, 117083.
- Li, Q., Du, J., Qiao, F., & Yu, L. (2018). Characterizing Particulate Matter 2.5 Concentration Pattern within a Transportation Network: A Case Study in the Port of Houston Region. *Journal of Pollution*, 2(1). Retrieved from <https://www.omicsonline.org/open-access/characterizing-particulate-matter-25-concentration-pattern-within-atransportation-network-a-case-study-in-the-port-of-houston-regi.pdf>

- Liao, D., Peuquet, D. J., Duan, Y., Whitsel, E. A., Dou, J., Smith, R. L., . . . Heiss, G. (2006). GIS approaches for the estimation of residential-level ambient PM concentrations. *Environmental health perspectives*, *114*(9), 1374.
- 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.
- Lippmann, M. (1989). Health effects of ozone a critical review. *Japca*, *39*(5), 672-695.
- Logan, J. A. (1985). Tropospheric ozone: Seasonal behavior, trends, and anthropogenic influence. *Journal of Geophysical Research: Atmospheres*, *90*(D6), 10463-10482.
- Longcore, T., & Rich, C. (2004). Ecological light pollution. *Frontiers in Ecology and the Environment*, *2*(4), 191-198.
- Mahajan, S., Chen, L.-J., & Tsai, T.-C. (2018). Short-Term PM_{2.5} Forecasting Using Exponential Smoothing Method: A Comparative Analysis. *Sensors*, *18*(10), 3223. Retrieved from <https://www.mdpi.com/1424-8220/18/10/3223>
- Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Birgani, Y. T., & Rahmati, M. (2019). Air pollution prediction by using an artificial neural network model. *Clean technologies and environmental policy*, *21*(6), 1341-1352.
- Mallet, V., Stoltz, G., & Mauricette, B. (2009). Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research: Atmospheres*, *114*(D5).
- Manes, F., Incerti, G., Salvatori, E., Vitale, M., Ricotta, C., & Costanza, R. (2012). Urban ecosystem services: tree diversity and stability of tropospheric ozone removal. *Ecological Applications*, *22*(1), 349-360.

- Marenco, A., Thouret, V., Nédélec, P., Smit, H., Helten, M., Kley, D., . . . Pyle, J. (1998). Measurement of ozone and water vapor by Airbus in-service aircraft: The MOZAIC airborne program, An overview. *Journal of Geophysical Research: Atmospheres*, 103(D19), 25631-25642.
- Means, B. (1989). *Risk-assessment guidance for superfund. Volume 1. Human health evaluation manual. Part A. Interim report (Final)*. Retrieved from
- Merry, R. H., Tiller, K., & Alston, A. (1986). The effects of contamination of soil with copper, lead and arsenic on the growth and composition of plants. *Plant and Soil*, 91(1), 115-128.
- Moe, C. L., & Rheingans, R. D. (2006). Global challenges in water, sanitation and health. *Journal of water and health*, 4(S1), 41-57.
- Moss, B. (2008). Water pollution by agriculture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1491), 659-666.
- Muskan. (2020). How to Choose Evaluation Metrics for Classification Models. Retrieved from <https://www.analyticsvidhya.com/blog/2020/10/how-to-choose-evaluation-metrics-for-classification-model/>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, 10(3), 282-290.
- Nedelec, S. L., Campbell, J., Radford, A. N., Simpson, S. D., & Merchant, N. D. (2016). Particle motion: the missing link in underwater acoustic ecology. *Methods in Ecology and Evolution*, 7(7), 836-842.

- Neuhauser, A. (2019). 100,000 Americans Die from Air Pollution, Study Finds.
Retrieved from <https://www.usnews.com/news/national-news/articles/2019-04-08/100-000-americans-die-from-air-pollution-study-finds>
- Newport, F. (2020). Religion and the COVID-19 Virus in the US.
- Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., & Kolehmainen, M. (2004). Evolving the neural network model for forecasting air pollution time series. *Engineering Applications of Artificial Intelligence*, 17(2), 159-167.
- Nordell, B. (2003). Thermal pollution causes global warming. *Global and planetary change*, 38(3-4), 305-312.
- Ocak, S., & Turalioglu, F. S. (2008). Effect of meteorology on the atmospheric concentrations of traffic-related pollutants in Erzurum, Turkey. *Journal of International Environmental Application and Science*, 3(5), 325-335.
- Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*: Springer Science & Business Media.
- Omidvarborna, H., Kumar, A., & Kim, D.-S. (2015). Recent studies on soot modeling for diesel combustion. *Renewable and Sustainable Energy Reviews*, 48, 635-647.
- Organization, W. H. (2014). 7 million premature deaths annually linked to air pollution. *World Health Organization, Geneva, Switzerland*.
- Page, M. L. (2019). Does air pollution really kill nearly 9 million people each year?
Retrieved from <https://www.newscientist.com/article/2196238-does-air-pollution-really-kill-nearly-9-million-people-each-year/>
- Pancholi, P., Kumar, A., Bikundia, D. S., & Chourasiya, S. (2018). An observation of seasonal and diurnal behavior of O₃–NO_x relationships and local/regional oxidant

- (OX= O₃+ NO₂) levels at a semi-arid urban site of western India. *Sustainable Environment Research*, 28(2), 79-89.
- Pandey, A. (2021). The Math Behind KNN. *Exploring the metric functions used in K-Nearest Neighbors (KNN) model*. Retrieved from <https://ai.plainenglish.io/the-math-behind-knn-7883aa8e314c>
- Parry, R. (1998). Agricultural phosphorus and water quality: A US Environmental Protection Agency perspective. *Journal of Environmental Quality*, 27(2), 258-261.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.
- Perez, P., Menares, C., & Ramírez, C. (2020). PM_{2.5} forecasting in Coyhaique, the most polluted city in the Americas. *Urban Climate*, 32, 100608.
doi:<https://doi.org/10.1016/j.uclim.2020.100608>
- Piñeiro, G., Perelman, S., Guerschman, J. P., & Paruelo, J. M. (2008). How to evaluate models: observed vs. predicted or predicted vs. observed? *Ecological Modelling*, 216(3-4), 316-322.
- Pope III, C. A. (2000). Invited commentary: particulate matter-mortality exposure-response relations and threshold. *American Journal of Epidemiology*, 152(5), 407-412.
- Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama*, 287(9), 1132-1141.

- Pudasainee, D., Sapkota, B., Shrestha, M. L., Kaga, A., Kondo, A., & Inoue, Y. (2006). Ground level ozone concentrations and its association with NO_x and meteorological parameters in Kathmandu valley, Nepal. *Atmospheric Environment*, 40(40), 8081-8087.
- Querol, X., Alastuey, A., Ruiz, C., Artiñano, B., Hansson, H., Harrison, R., . . . Bruckmann, P. (2004). Speciation and origin of PM₁₀ and PM_{2.5} in selected European cities. *Atmospheric Environment*, 38(38), 6547-6555.
- Racherla, P. N., & Adams, P. J. (2008). US ozone air quality under changing climate and anthropogenic emissions. In: ACS Publications.
- Random Forest Algorithm. (2020). Retrieved from <https://www.javatpoint.com/machine-learning-random-forest-algorithm>
- Reich, P. B. (1987). Quantifying plant response to ozone: a unifying theory. *Tree physiology*, 3(1), 63-91.
- Ris, R., Holthuijsen, L., & Booij, N. (1999). A third generation wave model for coastal regions: 2. Verification. *Journal of Geophysical Research: Oceans*, 104(C4), 7667-7681.
- Rodwell, M. J., & Hoskins, B. J. (2001). Subtropical anticyclones and summer monsoons. *Journal of Climate*, 14(15), 3192-3211.
- Ronaghan, S. (2018). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. Retrieved from <https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3>

- Ryerson, T., Trainer, M., Holloway, J., Parrish, D., Huey, L., Sueper, D., . . . Atlas, E. L. (2001). Observations of ozone formation in power plant plumes and implications for ozone control strategies. *Science*, *292*(5517), 719-723.
- Santini, R., Santini, P., Danze, J., Le Ruz, P., & Seigne, M. (2002). Study of the health of people living in the vicinity of mobile phone base stations: I. Influences of distance and sex. *Pathol Biol*, *50*(369), 369-373.
- Sastre, A., Cook, M. R., & Graham, C. (1998). Nocturnal exposure to intermittent 60 Hz magnetic fields alters human cardiac rhythm. *Bioelectromagnetics: Journal of the Bioelectromagnetics Society, The Society for Physical Regulation in Biology and Medicine, The European Bioelectromagnetics Association*, *19*(2), 98-106.
- SCHWARTZ, J. (1991). MORTALITY AND AIR-POLLUTION IN LONDON-A TIME-SERIES ANALYSIS-REPLY. *American Journal of Epidemiology*, *133*(6), 632-633.
- Schwartz, J. (2004). The effects of particulate air pollution on daily deaths: a multi-city case crossover analysis. *Occupational and Environmental Medicine*, *61*(12), 956-961.
- Shaban, K. B., Kadri, A., & Rezk, E. (2016). Urban Air Pollution Monitoring System With Forecasting Models. *IEEE Sensors Journal*, *16*(8), 2598-2606.
doi:10.1109/JSEN.2016.2514378
- Shalev-Shwartz, S., Singer, Y., Srebro, N., & Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, *127*(1), 3-30.
- Shao, M., Zhang, Y., Zeng, L., Tang, X., Zhang, J., Zhong, L., & Wang, B. (2009). Ground-level ozone in the Pearl River Delta and the roles of VOC and NO_x in its

production. *Journal of Environmental Management*, 90(1), 512-518.

doi:<https://doi.org/10.1016/j.jenvman.2007.12.008>

Shinar, D. (2017). *Traffic safety and human behavior*. Bingley: Emerald Publishing Limited.

Shindell, D. T., Rind, D., & Lonergan, P. (1998). Increased polar stratospheric ozone losses and delayed eventual recovery owing to increasing greenhouse-gas concentrations. *Nature*, 392(6676), 589.

Song, W., Wang, Y.-L., Yang, W., Sun, X.-C., Tong, Y.-D., Wang, X.-M., . . . Liu, X.-Y. (2019). Isotopic evaluation on relative contributions of major NO_x sources to nitrate of PM_{2.5} in Beijing. *Environmental Pollution*, 248, 183-190.

Sonntag, D. B., Baldauf, R. W., Yanca, C. A., & Fulper, C. R. (2014). Particulate matter speciation profiles for light-duty gasoline vehicles in the United States. *J Air Waste Manag Assoc*, 64(5), 529-545.

Sreekanth. (2020). Understanding XGBoost Algorithm | What is XGBoost Algorithm? Retrieved from <https://www.mygreatlearning.com/blog/xgboost-algorithm/>

Srivastava, C., Singh, S., & Singh, A. P. (2018, 28-29 Sept. 2018). *Estimation of Air Pollution in Delhi Using Machine Learning Techniques*. Paper presented at the 2018 International Conference on Computing, Power and Communication Technologies (GUCON).

Stansfeld, S. A., & Matheson, M. P. (2003). Noise pollution: non-auditory effects on health. *British medical bulletin*, 68(1), 243-257.

Stern, A. C. (1977). *Air Pollution: The effects of air pollution* (Vol. 2): Elsevier.

- Stevenson, D. S., Doherty, R. M., Sanderson, M. G., Collins, W. J., Johnson, C. E., & Derwent, R. G. (2004). Radiative forcing from aircraft NO_x emissions: Mechanisms and seasonal dependence. *Journal of Geophysical Research: Atmospheres*, 109(D17).
- Streets, D., & Waldhoff, S. (2000). Present and future emissions of air pollutants in China: SO₂, NO_x, and CO. *Atmospheric Environment*, 34(3), 363-374.
- Sunyer, J., Esnaola, M., Alvarez-Pedrerol, M., Forns, J., Rivas, I., López-Vicente, M., . . . Basagaña, X. (2015). Association between traffic-related air pollution in schools and cognitive development in primary school children: a prospective cohort study. *PLoS medicine*, 12(3), e1001792.
- Tagle, T. (2021). A Beginner's Guide to NO_x, NO and NO₂ as Air Pollutants. Retrieved from <https://www.aeroqual.com/meet-the-nitrogen-oxide-family>
- Tai, A. P., Mickley, L. J., & Jacob, D. J. (2010). Correlations between fine particulate matter (PM_{2.5}) and meteorological variables in the United States: Implications for the sensitivity of PM_{2.5} to climate change. *Atmospheric Environment*, 44(32), 3976-3984.
- Tang, X., Zhu, J., Wang, Z., & Gbaguidi, A. (2011). Improvement of ozone forecast over Beijing based on ensemble Kalman filter with simultaneous adjustment of initial conditions and emissions. *Atmospheric Chemistry and Physics*, 11(24), 12901-12916.
- Taranenko, L. (2019). HOW TO APPLY MACHINE LEARNING TO DEMAND FORECASTING. Retrieved from <https://mobidev.biz/blog/machine-learning-methods-demand-forecasting-retail>

- Tie, X., Geng, F., Peng, L., Gao, W., & Zhao, C. (2009). Measurement and modeling of O₃ variability in Shanghai, China: Application of the WRF-Chem model. *Atmospheric Environment*, 43(28), 4289-4302.
- UNESCO. (2020). Education: From disruption to recovery. Retrieved from <https://en.unesco.org/covid19/educationresponse>
- Van Der A, R., Eskes, H., Boersma, K., Van Noije, T., Van Roozendaal, M., De Smedt, I., . . . Meijer, E. (2008). Trends, seasonal variability and dominant NO_x source derived from a ten year record of NO₂ measured from space. *Journal of Geophysical Research: Atmospheres*, 113(D4).
- Varotsos, C. A., Efstathiou, M. N., & Kondratyev, K. Y. (2003). Long-term variation in surface ozone and its precursors in Athens, Greece: a forecasting tool. *Environmental science and pollution research international*, 10(1), 19-23.
- Vitousek, P. M., Mooney, H. A., Lubchenco, J., & Melillo, J. M. (1997). Human Domination of Earth's Ecosystems. *Science*, 277(5325), 494-499.
doi:10.1126/science.277.5325.494
- Wang, J., & Ogawa, S. (2015). Effects of Meteorological Conditions on PM_{2.5} Concentrations in Nagasaki, Japan. *International Journal of Environmental Research and Public Health*, 12(8), 9089-9101. doi:10.3390/ijerph120809089
- Wang, T., Ding, A., Gao, J., & Wu, W. S. (2006). Strong ozone production in urban plumes from Beijing, China. *Geophysical Research Letters*, 33(21).
- Warneck, P. (1999). *Chemistry of the natural atmosphere*: Elsevier.

- Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., & Chi, T. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Science of the Total Environment*, 654, 1091-1099.
- WHO. (2006). *Air quality guidelines: global update 2005*: World Health Organization.
- Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., . . . Rowe, C. M. (1985). Statistics for the evaluation and comparison of models. *Journal of Geophysical Research: Oceans*, 90(C5), 8995-9005.
- Wu, Y.-C., Chen, C.-S., & Chan, Y.-J. (2020). The outbreak of COVID-19: An overview. *Journal of the Chinese Medical Association*, 83(3), 217.
- Xi, X., Wei, Z., Xiaoguang, R., Yijie, W., Xinxin, B., Wenjun, Y., & Jin, D. (2015, 15-17 Nov. 2015). *A comprehensive evaluation of air pollution prediction improvement by a machine learning method*. Paper presented at the 2015 IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI).
- Xia, T., Nitschke, M., Zhang, Y., Shah, P., Crabb, S., & Hansen, A. (2015). Traffic-related air pollution and health co-benefits of alternative transport in Adelaide, South Australia. *Environment International*, 74, 281-290.
- Xiao, Q., Chang, H. H., Geng, G., & Liu, Y. (2018). An Ensemble Machine-Learning Model To Predict Historical PM_{2.5} Concentrations in China from Satellite Data. *Environmental Science & Technology*, 52(22), 13260-13269.
doi:10.1021/acs.est.8b02917
- Xiao, Z.-m., Zhang, Y.-f., Hong, S.-m., Bi, X.-h., Jiao, L., Feng, Y.-c., & Wang, Y.-q. (2011). Estimation of the main factors influencing haze, based on a long-term

- monitoring campaign in Hangzhou, China. *Aerosol and Air Quality Research*, 11(7), 873-882.
- Xin, D., Han, J., Yan, X., & Cheng, H. (2005). *Mining compressed frequent-pattern sets*. Paper presented at the Proceedings of the 31st international conference on Very large data bases.
- Xing, Y., Xu, Y., Shi, M., & Lian, Y. (2016). The impact of PM_{2.5} on the human respiratory system. *Journal of Thoracic Disease*, 8(1), E69-E74.
doi:10.3978/j.issn.2072-1439.2016.01.19
- Yu, D., Tan, Z., Lu, K., Ma, X., Li, X., Chen, S., . . . Qiu, P. (2020). An explicit study of local ozone budget and NO_x-VOCs sensitivity in Shenzhen China. *Atmospheric Environment*, 224, 117304.
- Zakka, K. (2016). A Complete Guide to K-Nearest-Neighbors with Applications in Python and R. Retrieved from <https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>
- Zhang, Z.-H., Khlystov, A., Norford, L. K., Tan, Z.-K., & Balasubramanian, R. (2017). Characterization of traffic-related ambient fine particulate matter (PM_{2.5}) in an Asian city: Environmental and health implications. *Atmospheric Environment*, 161, 132-143.
- Zheng, M., Salmon, L. G., Schauer, J. J., Zeng, L., Kiang, C., Zhang, Y., & Cass, G. R. (2005). Seasonal trends in PM_{2.5} source contributions in Beijing, China. *Atmospheric Environment*, 39(22), 3967-3976.
- Zhou, Y., Chang, F.-J., Chang, L.-C., Kao, I. F., Wang, Y.-S., & Kang, C.-C. (2019). Multi-output support vector machine for regional multi-step-ahead PM_{2.5}

forecasting. *Science of the Total Environment*, 651, 230-240.

doi:<https://doi.org/10.1016/j.scitotenv.2018.09.111>

Zhu, S., Lian, X., Wei, L., Che, J., Shen, X., Yang, L., . . . Li, J. (2018). PM2.5 forecasting using SVR with PSO-GSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. *Atmospheric Environment*, 183, 20-32.
doi:<https://doi.org/10.1016/j.atmosenv.2018.04.004>

Zumla, A., George, A., Sharma, V., Herbert, R. H. N., Oxley, A., & Oliver, M. (2015). The WHO 2014 global tuberculosis report—further to go. *The Lancet Global Health*, 3(1), e10-e12.