INVESTIGATING CORRELATIONS AMONG TRAFFIC FLOW CHARACTERISTICS, AIR POLLUTION, AND METEOROLOGICAL FACTORS ON URBAN FREEWAYS

A Thesis

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

In

Civil Engineering

By

Xuqing Liang

2021

SIGNATURE PAGE

THESIS:	INVESTIGATING CORRELATIONS AMONG TRAFFIC FLOW CHARACTERISTICS, AIR POLLUTION, AND METEOROLOGICAL FACTORS ON URBAN FREEWAYS
AUTHOR:	Xuqing Liang
DATE SUBMITTED:	Spring 2021
	Department of Civil Engineering

Dr. Xinkai Wu
Thesis Committee Chair
Civil Engineering Department _____

Dr. Yongping Zhang
Civil Engineering Department _____

Dr. Wen Cheng
Civil Engineering Department _____

ABSTRACT

Air pollution has been one of the most concerning issues for metropolitan cities, especially areas in the vicinity of urban freeways that deal with tremendous traffic flow every day. The need to evaluate the impact of freeway traffic on ambient air quality in a microscopic aspect has led this study to investigate the correlations between traffic flow characteristics and pollution concentrations using year-round hourly observations collected from a specific section of the Interstate 210 freeway in the Los Angeles County. In this thesis, carbon dioxide (CO2) representing the greenhouse gas, traffic-induced air pollutants including carbon monoxide (CO), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO2), Ozone (O3), and particulate matter (PM) were analyzed with emphasis. Some significant meteorological factors were also taken into account based on their notable influence on the dispersion and chemical reaction of pollutants. The data of traffic, air pollutants, and meteorological factors were jointly used for comprehensive univariate analysis to explore their interrelationships. In addition, three statistical models for prediction purposes, including multiple linear regression and stepwise regression, were conducted to predict the pollutant concentrations one hour ahead and identify the significant predictors of pollutants. A machine-learning algorithm, artificial neural network (ANN), was also applied to provide comparable results for prediction performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The United States is well known to be a "nation on wheels" over decades since the Interstate Highway System, an integrated network of controlled-access highways initiated by the Federal Aid Highway Act of 1956, profoundly changed the American way of life (Foster, 2003). Covering the contiguous states, the Interstate Highway System has played a significant role in fostering the trucking industry and prosperity in metropolitan areas (Weber, 2012; Federal Highway Administration, 2017). However, such explosive growth in traffic demand has raised critical concerns about air pollution to the public, especially in major cities where interstate freeways usually intersect. Numerous studies have illustrated the negative impacts of traffic-induced air pollution in urban communities, including environmental, economic, and human-health aspects (Tischer, Fountas, Polette, & Rye, 2019; Yan, et al., 2019; Volk, Lurmann, Hertz-Picciotto, & McConnell, 2012; Padula, et al., 2012).

Among a wide variety of sources, burning fossil fuels for transportation purposes is the most contributive one to the emissions of greenhouse gas (GHG) such as carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O). According to the United States Environmental Protection Agency (EPA), the transportation sector was responsible for 28% of total U.S. GHG emissions in 2018; within the transportation emissions, light-duty vehicles and medium-to-heavy trucks accounted for 59% and 23%, respectively (2020). Other regulatory work by EPA indicated that 99% of carbons in fuels are emitted as CO_2 during combustion and that CO_2 emissions from a vehicle overall account for 95-99% of the total GHG emissions weighted by the global warming potential of all GHGs (2018).

Given this knowledge, we merely consider CO_2 as the representative of GHG to be studied in this thesis.

Combustion of fossil fuels, mainly for road transport and power generation, is also a significant source of some classical air pollutants such as carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), volatile organic compounds (VOCs), and particulate matter (PM). In metropolitan areas like Los Angeles, from which industrial factories have been relocated, vehicle exhausts become more vital to the issues. It is of great importance for transportation professionals to examine how and to what extent the traffic volume will influence the density of pollutants.

It is an indisputable fact that automobile is a contributive factor of air pollution if one measures pollutant emissions from a running vehicular combustion engine in the laboratory, a controlled environment in which other contributive factors can be minimized. Nevertheless, it becomes quite challenging to prove such correlation through a field experiment since extraneous factors or systems may be involved in a natural setting. Researchers commonly acknowledged one unregulated variable to be concerned with outdoor air pollution epidemiology: the science of weather, known as meteorology. The variations in the physical and dynamic properties, such as temperature, relative humidity, and atmospheric pressure, may have a remarkable impact on the physical or chemical reactions between toxic pollutants previously mentioned (Maynard, Holgate, Koren, & Samet, 1999). Properties of wind, including wind speed and wind direction, influence the dispersion and diffusion of toxic gases (The State of Queensland, 2017). As

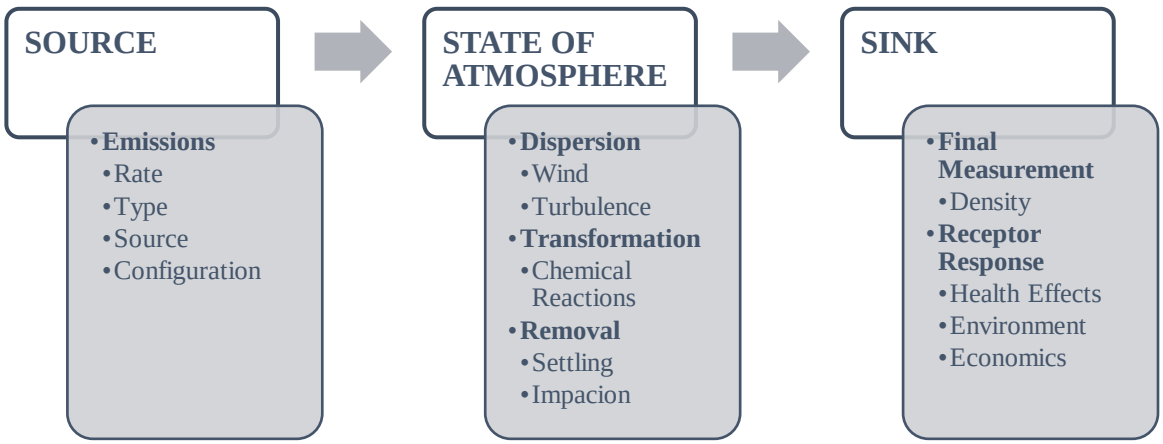


Figure 1: The Role of the Atmosphere in the Air Pollution Source-sink Relationship (Maynard, Holgate, Koren, & Samet, 1999)

shown in Figure 1, meteorological factors play a substantial part in the continuous process of air pollutants from being produced by sources to being received in the presence of the atmosphere.

Previous research regarding this topic varies diversely in method and scope. Many studies have examined relationships between air pollution and traffic by using data from national inventory on a relatively large scale regarding the observation interval or examined area. Other works have focused on developing simulation or visualization models to analyze similar issues. According to the best of our knowledge, until recent years, there are very few studies examining correlations between air pollution density and traffic that use data mining and machine learning methods due to no access to real-time measurement data of both fields. With the rapid development of technology today, the decreasing costs of traffic detectors and remote sensors for measuring air indices have made massive datasets available for analysis. Air-related datasets often contain a considerable number of samples and suffer from high dimensionality and multicollinearity when independent variables are plenty and highly correlated (Bellinger

C. , Jabbar, Zaiane, & Osornio-Vargas, 2017). Combining some computational science, including artificial intelligence (AI), statistics, mathematics, and machine learning, data mining methods are able to analyze this sort of dataset, discover patterns, and predict results.

In addition, speaking of the data collection site, we expect that taking measurements of ambient air pollution density and traffic flow rate at the roadside of a freeway, which often deals with the highest traffic volume within a city, should provide us precise data for a comprehensive analysis and regression modeling. Given all the context above, we intend to investigate the statistical correlation among traffic, air pollution, and meteorological factors using integrated data collected from a microscopic aspect. Data description will be provided in Chapter 3.

This thesis first reviews the literature on the formation and characteristics of common air pollutants induced by traffic (CO, NO, NO_2, O_3, and PM) and GHG (CO_2). We also review previous studies and research on correlations among air pollutants' ambient density and traffic-related parameters (flow, density, speed, truck flow, and non-truck flow). Meteorological data such as temperature, relative humidity, wind speed, and wind direction are also considered to obtain a comprehensive insight into air pollution epidemiology. The thesis then introduces the dataset in which data collected by multiple organizations next to a freeway is joint together for analysis. Next, we present a comprehensive data analysis using year-round real-time historical data to investigate correlations among traffic, air pollutant concentrations, and meteorological factors. We also develop three predictive regression models that estimate each air pollutant’s concentrations an hour in the future and evaluate their performance by comparing their

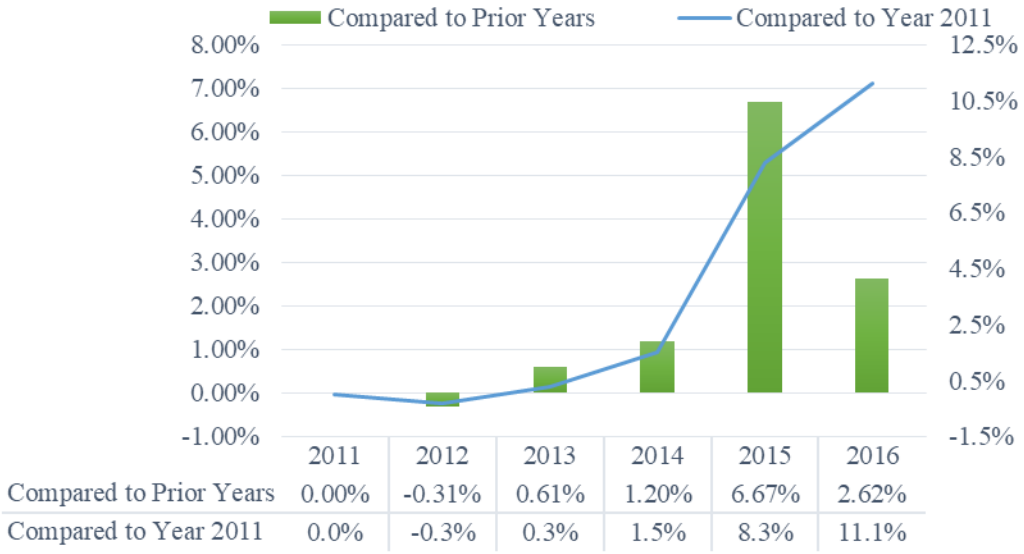
adjusted R-squared values and root mean squared error (RMSE). Finally, our conclusion is drawn based on the model outcomes and limitations we found along with the study.

CHAPTER 2: LITERATURE REVIEW

As stated in the previous chapter, the United States has been facing severe air pollution issues induced by high-volume traffic on its Interstate Highway System. According to the California Department of Transportation (Caltrans) statistics presented by a combined graph and summary table in Figure 2, vehicle miles traveled on California state highways in 2016 was increased by 2.62% compared with the prior year and 11.1% cumulatively compared to the year 2011 (2018). Therefore, analyzing the potential relationship between traffic and ambient air pollution under time- and season-varying weather conditions becomes an urgent need given the growing traffic demand. A systematic review of existing works related to this issue, employing statistical analysis, simulation and prediction modeling, or machine learning algorithms, is delivered in this chapter.

2.1 Studies of Traffic-Related Air Pollutants

CO, NO, NO_2, O3, and PM are commonly known as traffic-induced air pollutants and CO_2 as the representative of GHGs. It is of immense importance to understand how air pollutants are formed before analyzing their interrelationships. Therefore, following how they are produced and emitted into the atmosphere, pollutants may be categorized into primary and secondary pollutants (Maynard, Holgate, Koren, & Samet, 1999). Primary pollutants are emitted directly into the air, while secondary pollutants are formed when primary pollutants interact with each other in the atmosphere. A summary of pollutant classification and formation is provided in Table 1, with targeted pollutants of this study in bold. Some primary pollutants are precursors to certain secondary pollutants, such as NO to NO_2 and O_3, which may imply a considerable correlation between them.



	2011	2012	2013	2014	2015	2016
Compared to Prior Years	0.00%	-0.31%	0.61%	1.20%	6.67%	2.62%
Compared to Year 2011	0.0%	-0.3%	0.3%	1.5%	8.3%	11.1%

Figure 2: Comparison of Annual VMT Increased over Prior Years

Table 1: Summary Traffic-induced Air Pollutants in Categories (Scottish Environment Protection Agency; United States Environmental Protection Agency, 2020)

Air Pollutant	Primary	Secondary	General Description of Formation
CO	Yes		A small amount of carbon is emitted as CO due to the incomplete combustion process of fossil fuel.
CO_2	Yes		Most of the carbon in fuel is emitted as CO_2.
NO	Yes		NO is formed when N_2 reacts with O_2.
NO_2	Yes	Yes	NO_2 can be formed from fuels burned at high temperatures or oxidation of NO.
PM2.5	Yes	Yes	Fine PM is emitted directly from natural and anthropogenic sources.
PM10		Yes	PM10 is formed primarily from the oxidation of SO_2, NO_2, and ammonia.
SO_2	Yes		The oxidation of sulfur in fuel forms SO2.
Ground-Level O_3		Yes	O_3 is formed by chemical reactions between NO_x and VOCs in the presence of sunlight.

The amount of carbon in the fuel to be burned for transportation purposes is primarily emitted as CO_2 and CO into the air, with CO_2 accounting for 99% (United States Environmental Protection Agency, 2018). CO_2 is one of the primary contributors to global warming, and CO is a byproduct of incomplete combustion, which is a colorless, odorless, and poisonous gas. Therefore, the emission ratio of CO_2 to CO depends on combustion efficiency (Wang, et al., 2010). In addition, their research analyzed temporal variations of CO_2 and CO based on continuous measurements over the years and suggested a significant CO_2-CO correlation for cold seasons, while it was degraded by the impact of photosynthesis and respiration during spring and summer months. Other studies on fuel combustion showed that CO emissions increase when vehicle engines are operated at low temperatures and when traffic is congested (Vakkilainen, 2017; Miller, 2011).

NO and NO_2 contribute to the formation of photochemical smog, acid rain, and ground-level O_3 and henceforth are harmful to human health and the environment. They are collectively regarded as nitrogen oxides (NO_x) because they are the most common form of nitrogenous pollutants in the ambient air. A significant number of studies have confirmed the interconvertibility of NO, NO_2, and O_3 in photochemical smog reactions. The best understood atmospheric chemical process among these three problem pollutants is that NO, reacting with O_3, transforms into NO_2. NO can also convert itself into NO_2 through the reaction with O_2, but this occurs solely during the cold winter seasons or night times (Maynard, Holgate, Koren, & Samet, 1999). However, NO_2 and VOCs undergo photolysis in the presence of sunlight to reform the NO and O_3 from which it was created (Li, Zhou, & Tong, 2019; World Bank Group, 1998). The oxidation of NO

and photolysis of NO_2 form a cycle of reactions that continuously generates a mixture of NO_x in the sunlit atmosphere. Therefore, it is expected to observe a substantial variation in these pollutant concentrations on an hourly and seasonal basis. Given all the context above, it is highly suggested that there might be potential correlations among concentrations of NO, NO_2, and O_3.

A complex mixture of various particles and inorganic substances present in the ambient air is termed particulate matter (PM). Coarse particles such as dust, dirt, and soot, which usually have diameters of 10 μm and less, are referred to as PM10. Similarly, PM2.5 are those fine particles of 2.5 μm and less. PMs are incredibly harmful to human health as they can be inhaled into human lungs or even the bloodstream, causing heart attacks and asthma symptoms (Li, Zhou, & Tong, 2019). Some existing research has demonstrated a high correlation between PM2.5 and PM10 (Zhou, et al., 2016; Wang, Bi, Sheng, & Fu, 2006).

2.2 Correlations Between Pollutant Concentrations and Traffic Characteristics

In traffic flow theory, the traffic condition of a specific roadway can be defined by macroscopic characteristics such as flow (q), density (k), and space mean speed (u). The relationship among these characteristics has been well established in the traditional fundamental diagram that q = uk. Traffic parameters can be measured by loop detectors deployed on the roads, which are able to count the number of passing vehicles and measure the total time of them being occupied by vehicles given the desired period. In this thesis, we solely focus on the hourly parameters on an urban freeway of our interest. Hourly flow is the total throughput of traffic expressed in vehicles per hour, density is the average density of vehicles contained in a specific length of roadway, and speed is the

average speed at which vehicles travel during that hour. Traffic density will be significantly increased, and average speed decreased in congested traffic in which vehicles are iteratively operated under driving conditions of accelerating, decelerating, or idling. The emission levels of traffic-related pollutants during traffic congestion are expected to rise due to the degraded efficiency of fuel combustion in vehicle engines. However, the fundamental diagram is not a physical law; its shape can be widely influenced by other significant factors such as the road type, vehicle composition, and speed limits. Based on this information, we include hourly truck flow into our data analysis in addition to the traffic parameters mentioned above.

A massive number of works have focused on investigating ambient air pollution and other impacts induced by traffic. However, most of them either derive their pollutant data by multiplying the traffic volume from simulation models with emission factors from vehicle emission models or assume that pollutant production observed next to freeways and arterials are entirely accounted for by traffic. According to our current understanding, limited studies have examined the impacts of traffic flow characteristics on the concentration or spatial distribution of ambient air pollutants. For instance, Brief, Jones & Yoder studied atmospheric levels of CO at six locations in a United States city and established a positive relationship between CO concentration and traffic density (2012). In another research, stable emission rates of air pollutants for various driving conditions, including idle, acceleration, cruise, and deceleration, were developed with onboard measurements (Frey, Unal, Rouphail, & Colyar, 2003). Wang, Makino & Wu examined the correlations among traffic flow, air pollutant emissions, and meteorology using high-frequency data of both traffic and pollutant concentrations collected by

(2018). On-site observation samples in time series demonstrate the importance of taking real-world emissions and the nature of traffic flow dynamics into account.

2.3 Preexisting Statistical Prediction Models and Data Mining Applications

Statistical models, data mining algorithms, and simulation models are commonly adopted among studies concentrating on air pollution epidemiology, of which the first two will be primarily discussed in this section based on our similar interest of study. On the one hand, land-use regression (LUR) modeling is one of the most popular tools to predict spatial concentration levels of pollutants with potential transportation- and land-use-related predictors as independent variables within a multiple regression model.

Linear regression models are relatively simple to apply and provide coefficients with p-values indicating the statistical significance for each explanatory variable. However, regression models are often criticized for the ignorance of potential correlations between predictors and the inapplicability to categorical variables. Given the fact that datasets of air pollution usually contain a significant number of variables with a mixed degree of dependencies (Bellinger C. , Jabbar, Zaiane, & Osornio-Vargas, 2017), data mining and machine learning methods become favorable to researchers due to their capability of analyzing large datasets, discovering patterns, and predicting future outcomes.

Based on our review in this thesis, it is not rare to observe that both regression models and machine learning methods are applied and compared side by side in one study. Perez & Trier first built a prediction model of NO concentrations based on historical data observed on an avenue with heavy traffic and then forecasted NO_2 concentrations using linear regression models and a multi-layer neural network (Perez &

Trier, 2001). Wang, Keita & Wu utilized multiple linear regression with lagged inputs, k-nearest neighbor (k-NN), and feedforward artificial neural network (FANN) to predict pollutant concentrations 10 minutes and one hour in advance on urban arterials in Beijing, China. Kukkonen, et al. developed and evaluated five neural network models, a linear statistical model, and a deterministic modeling system (DET) for predicting NO_2 and PM_{10} concentrations in central Helsinki (2003). Furthermore, LUR models adopting multiple linear regression techniques (linear, stepwise, LASSO, elastic net, etc.) and data mining methods (neural network, SVM, random forest, extreme boosting, etc.) were presented along with a systematic evaluation of model performance (Kerckhoffs, Hoek, Portengen, Brunekreef, & Vermeulen, 2019). Focusing on machine learning techniques only, a decision tree algorithm was trained to predict the concentration of PM2.5 in morning peak hours of pollution (Zalakeviciute, Bastidas, Buenano, & Rybarczyk, 2020). Juhos, Makra & Toth performed predictions of NO and NO_2 partially based on their past values and partially based on temperature, humidity, and wind data (2008).

Given all the information above, we will perform prediction modeling using multiple linear regression, stepwise regression, and neural network techniques in this thesis. Differing from preexisting studies, we jointly used a unique dataset that constitutes one-year time-series data of traffic parameters, air pollutant concentration, and meteorologic factors on an hourly basis collected at an urban freeway in Pasadena, California.

CHAPTER 3: DATA COLLECTION

3.1 Traffic Data Collection

The traffic data utilized in this research was obtained from the Caltrans Performance Measurement System (PeMS), a published database that stores traffic-related data of the freeway system across all major metropolitan areas of California. The raw data is collected from the single lane detectors, which measure the flow and occupancy every 30 seconds. It is then transformed, stored, and processed through data grinding in the PeMS system, of which the mechanism is presented in Figure 3.

We extracted the hourly flow, hourly truck flow, density, and speed data in time series from PeMS based on our interest of this thesis. To specifically distinguish the

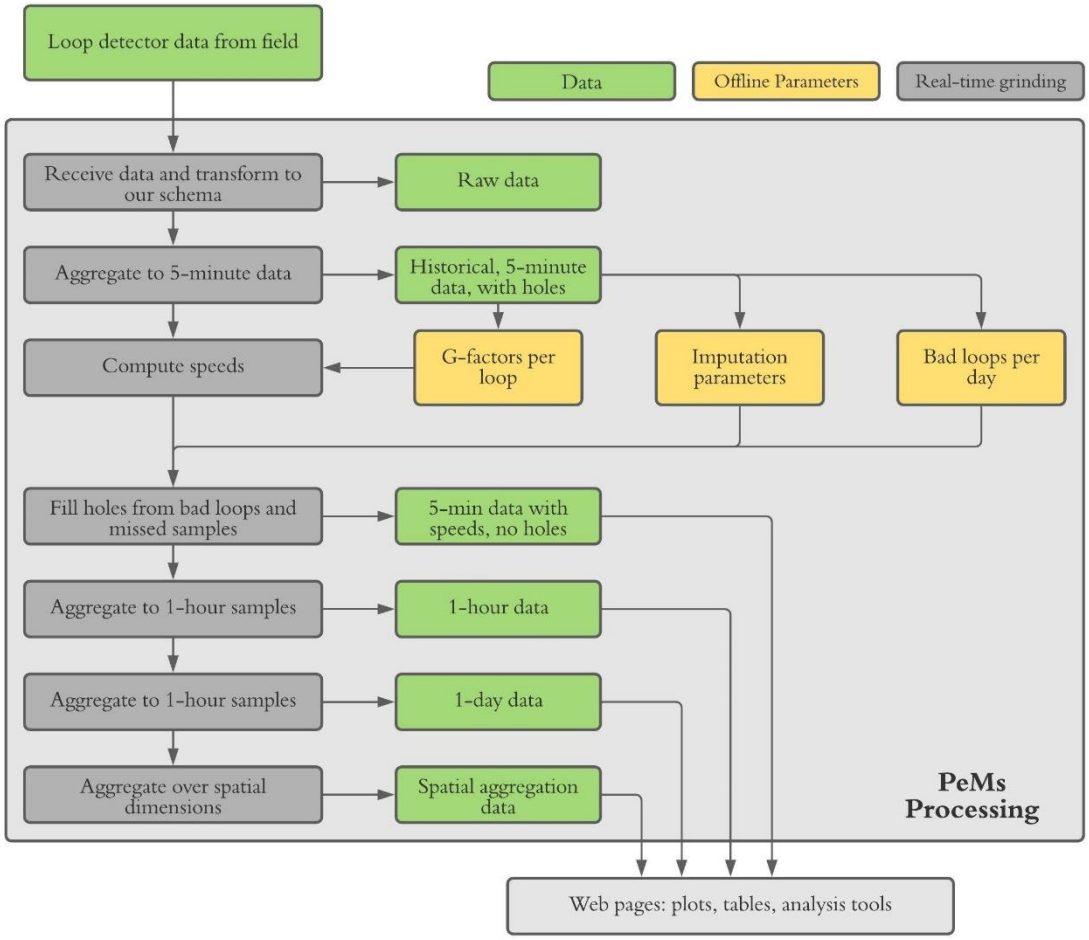


Figure 3: Introductory Flowchart of Data Grinding Process in the PeMS System.

impacts of truck flow and regular passenger vehicle flow on air pollutant emission rates, we obtained the non-truck flow by subtracting the truck flow from the total flow. Table 2 summarizes the traffic-related variables generated at this point.

Table 2: Overview of Hourly Traffic Data Attributes Retrieved from PeMS

Attribute Notation	Description	Unit
Flow	Total hourly traffic flow from all lanes	vph
Density	Average density derived from total flow divided by speed	vpm
Speed	Flow-weighted average of 5-minute station speeds	mph
Heavy	The hourly flow of trucks and buses	vph
Small	Traffic flow of regular passenger vehicles and vans	vph

Unit abbreviation: vph vehicles per hour, vpm vehicles per mile, mph miles per hour

3.2 Air Pollutant Data Collection

To measure air pollutants and particulate matters that are to be studied in this research, a customized, portable, and cost-effective air pollution monitoring unit has been installed previously by a research team from Cal Poly Pomona. Supported by solar power, this monitoring unit consists of detectors for CO_2, CO, NO_2, NO, O_3, $PM_{x,}$ and sensors for temperature and relative humidity and is able to take sample measurements of all variables listed above every two minutes. The data is then stored and transferred to the database of the research team. This device is also allowed to perform calibrations from 12 to 1 automatically am every day to ensure the accuracy of measurements.

The raw 2-minute pollutant data was transformed into a 1-hour base by averaging to be consistent with the time interval of traffic data. Table 3 presents the summary and units of all air pollutant variables and a portion of meteorological variables.

Table 3: Overview of Data Attributes Collected from the Air Pollution Monitoring Unit

Attribute Notation	Description	Unit
CO2	CO_2 concentration value	ppm
CO	CO concentration value	mg/m^3
NO2	NO_2 concentration value	$\mu g/m^3$
NO	NO concentration value	$\mu g/m^3$
O3	O_3 concentration value	$\mu g/m^3$
PM2.5	PM2.5 concentration value	$\mu g/m^3$
TEMP	Temperature	F
RH	Relative humidity	%

Unit Abbreviation: ppm parts per million

3.3 Wind Data Collection

The wind properties were retrieved from the EPA's Air Quality System (AQS) database, a public database that stores ambient air sample data and meteorological data collected by local and federal agencies from thousands of monitors. The wind speed is expressed in knots and the wind direction in degrees with 0 degrees indicating the north.

3.4 Data Collection Site

The roadside air pollutant measurement is mounted on a CCTV pole adjacent to the partial interchange of the I-210 freeway and North Altadena Drive in the city of Pasadena, CA. Represented by the blue icon in Figure 4, the device is located at the gap between the mainline of I-210 westbound and the on-ramp. The location of the traffic loop detectors indicated by the green icon is merely 10 feet from the air monitoring unit. Lastly, the EPA station shown as the red icon in the figure is approximately two miles from the other two collection points in real scale.

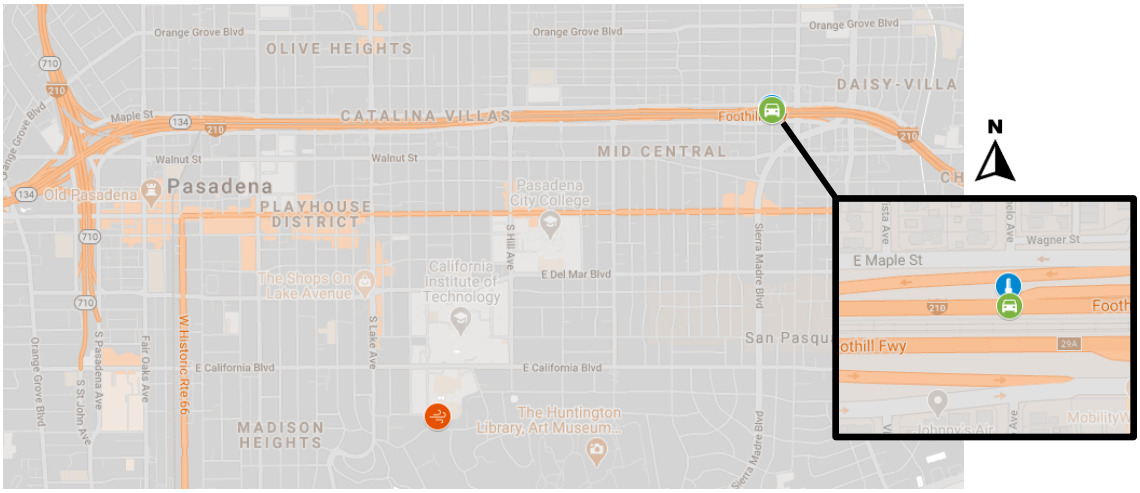


Figure 4: Data Collection Site of Traffic (Green), Air Pollutants (Blue), and Wind (Red).

CHAPTER 4: CORRELATION ANALYSIS

In this chapter, an initial analysis of traffic and air pollutant variables in time series, given the purpose to discover some daily patterns, will be presented. We also explore the seasonality in each air pollutant followed by an investigation of wind distribution regarding four different levels of pollutant concentrations. Finally, a correlation matrix is created to provide comprehensive knowledge of potential predictors of pollutants, based on which the selection of explanatory variables for prediction models is finalized.

4.1 Distribution in Time Series

The traffic detectors are situated on the westbound of the I-210 freeway, which connects principal valleys lying to the east of Los Angeles city with the central business districts in Los Angeles and the I-5 freeway, a major north-to-south Interstate highway in California. Therefore, the studied freeway section carries massive traffic volume daily and encounters severe traffic congestions during peak hours. All traffic samples were first grouped into weekdays or weekends based on the day of the observations, and then the average values were taken for each hour of the day, displayed in time series in Figure 5.

The time series plot of hourly flow demonstrates a long and continuous interval of heavy traffic from 5 am to 6 pm for weekdays, with traffic volumes ranging from 7800 to 9500 vehicles per hour. The weekend traffic shows a similar range of peak flows but with a shorter peak hour interval. The trend of heavy traffic on weekdays is more clearly observed in the density and speed plot. The sharp increase in density and drop in speed further certify the heavily congested traffic conditions from 6 to 9 am and 3 to 5 pm. Interestingly, the truck flow appears to have a slightly different daily pattern compared to

the total flow, which suggests a variation of traffic demand for different vehicle types regarding the time of day and sequentially emphasizes the necessity of incorporating truck flow and non-truck flow into our data analysis.

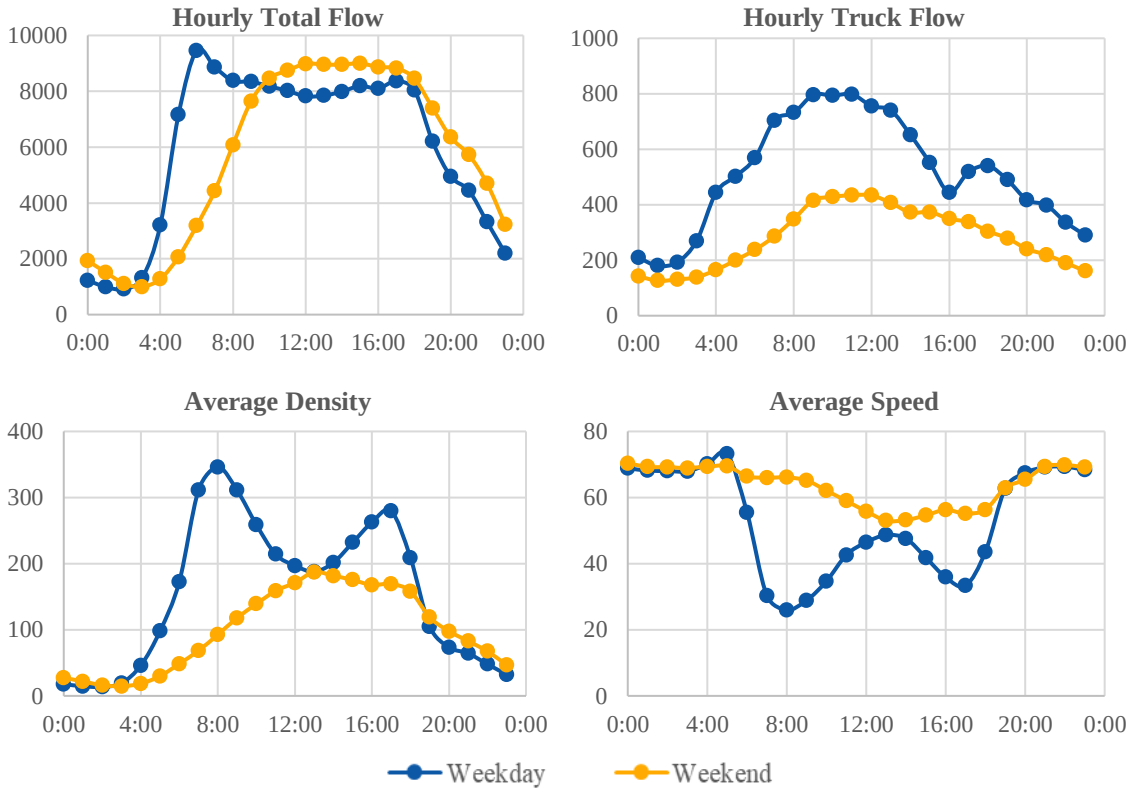


Figure 5: Time Series Plots of Annual Average Hourly Flow, Truck Flow, Density, and Speed.

Next, we adopted the same technique on the air pollutant data to produce time series plots shown in Figure 6. Note that the measurements taken during 12 to 1 am have shown certain degrees of bias due to the daily calibration process of the air monitoring unit thus were omitted from our dataset and analysis. The distribution of concentrations for each pollutant regarding the time of day shows different trends that are likely caused by traffic and meteorological conditions. It was interesting to discover a distinct pattern in CO_2 that its values are generally lower at night, reaching the lowest point before sunrise, gradually rising during the daytime, and reaching the highest average around 5

pm. On the contrary, NO_2/O_3/PM2.5 concentrations tended to be lower in the afternoon and higher at night and early morning. More importantly, an evident decline of NO_2 concentrations was clearly observed during the daytime, strongly implying a correlation with the time of day with which the temperature is known to fluctuate. It also confirms previous findings that NO_2 can be formed by oxidation of NO at low temperatures. (Maynard, Holgate, Koren, & Samet, 1999). No significant pattern was found in the plot of CO/NO, meaning their formation and emissions could be more complex.

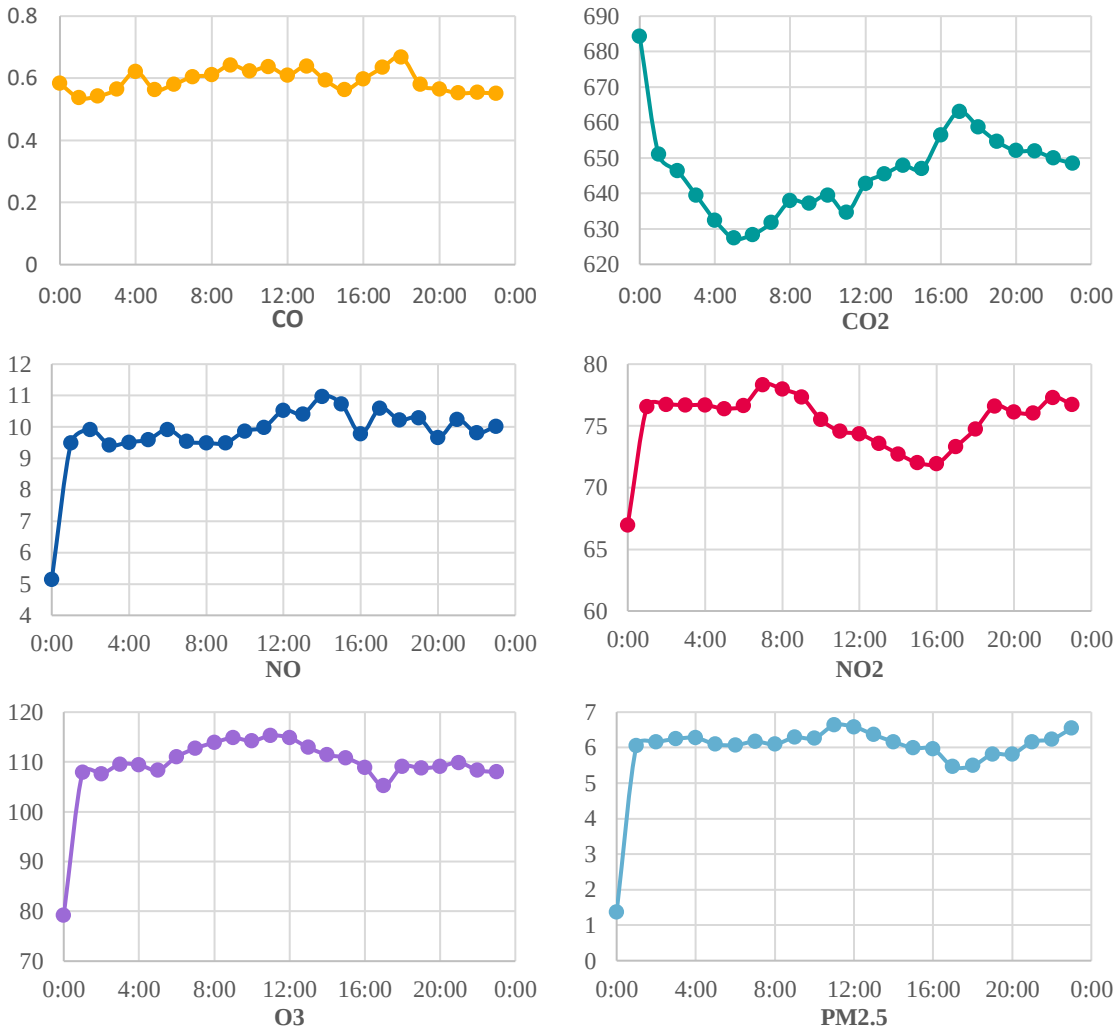


Figure 6: Time Series Plots of Annual Average Hourly Concentration for Air Pollutants

4.2 Distribution of Air Pollutants by Month

In this section, we aimed to explore the seasonality in the concentration variation of each pollutant by computing the mean values by month. The results are shown with bar plots in Figure 7. The CO_2 densities were relatively higher during the winter and spring months. NO concentrations tended to fluctuate intensely from month to month, which is not consistent with previous studies; however, the increasing trend in summer months could be caused by the reformation of NO from NO_2 in the condition of cumulative heat. As expected, the NO_2 concentrations appeared higher during the winter months while lower during the spring months. O_3/PM2.5 plots both showed a significant pattern with greater concentrations in the summer seasons. This outcome further confirmed that their concentrations are very likely correlated with ambient temperature. Meanwhile, CO did not feature any seasonal pattern in the plot.

4.3 Distribution of Wind

Wind plays a critical role in the dispersion and diffusion of air pollutants. We investigated the annual wind data to enhance our knowledge of how wind properties influence pollutant concentrations with the wind rose plots. The wind rose is an advantageous and effective technique for summarizing wind data. Wind samples are grouped by direction in the wind rose plot, typically with 45- or 30-degree increments, and represented by several paddles. The wind speed levels are expressed as different colors. The wind rose plot substantially is a stacked-column chart in which all columns stand for different angles of winds are arranged as a ring. Therefore, the annual wind data are summarized in Figure 8, with the paddle heights showing the counts of time that the wind is from a particular wind direction and wind speed range.

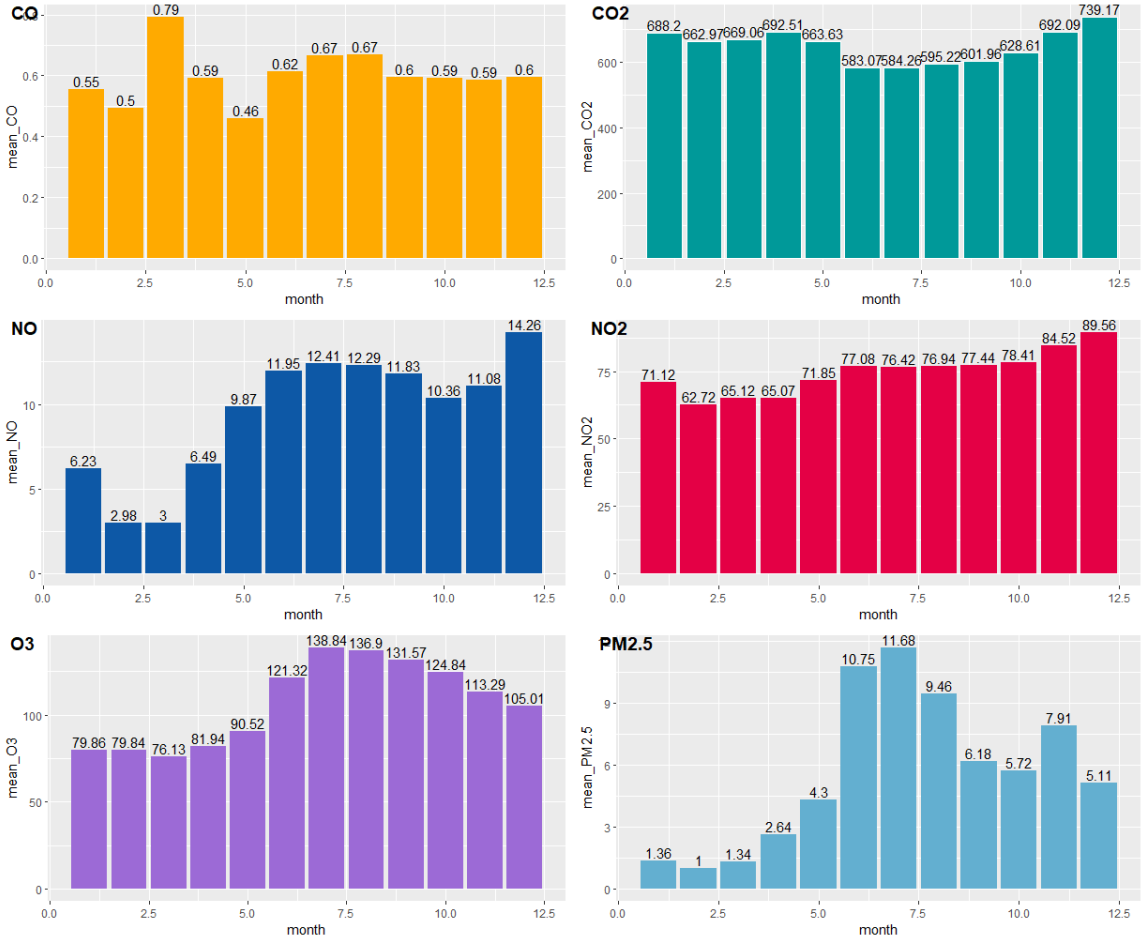


Figure 7: Bar Plots of Average Monthly Concentrations

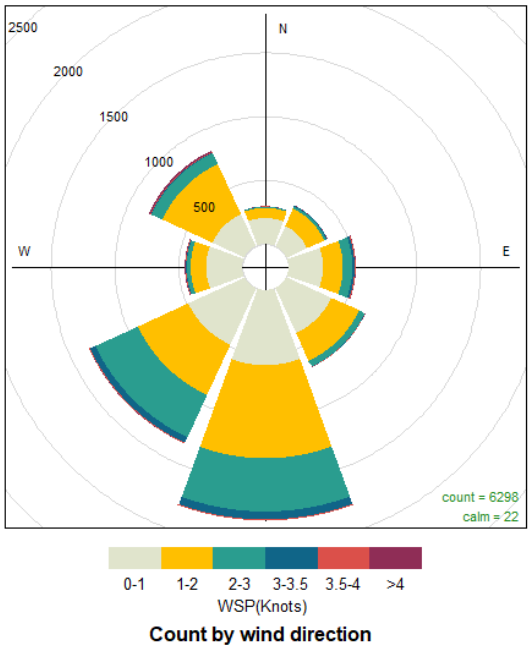


Figure 8: Summary of Wind Data in 2019.

According to the wind rose plot, it is of significant evidence that a substantial portion of wind occurrences in our study area come from the southern and southwestern directions. This phenomenon might indicate a high chance of pollutants being transported to our study site by the wind from the south. Moreover, we safely concluded that the wind traveled from the south and southwest at relatively high speeds by comparing the heights of the green portion in each direction.

We further recreated the wind roses plots by each pollutant at different emission levels by initially dividing all sample concentrations into four quartiles for each pollutant and then plotting them in separate wind roses, shown in Figure 9. These wind roses are very informative as they can reflect the potential impacts of wind properties on the concentrations of one pollutant. For example, wind roses created based on four different levels defined as the four quantiles of CO/NO_2 concentrations were almost identical to each other, suggesting insignificant impacts of wind speed and wind directions on CO. By contrast, southerly winds at high wind speeds were dominant at the time of lower CO_2 concentrations. Low NO concentrations were also concerned with high wind speeds because of the more frequent occurrences of strong wind in that wind rose plot. More interestingly, the majority of wind came from the south when high O_3 concentrations were detected. It might suggest a more prominent source of O3 at the south side of our study site in addition to road transport.

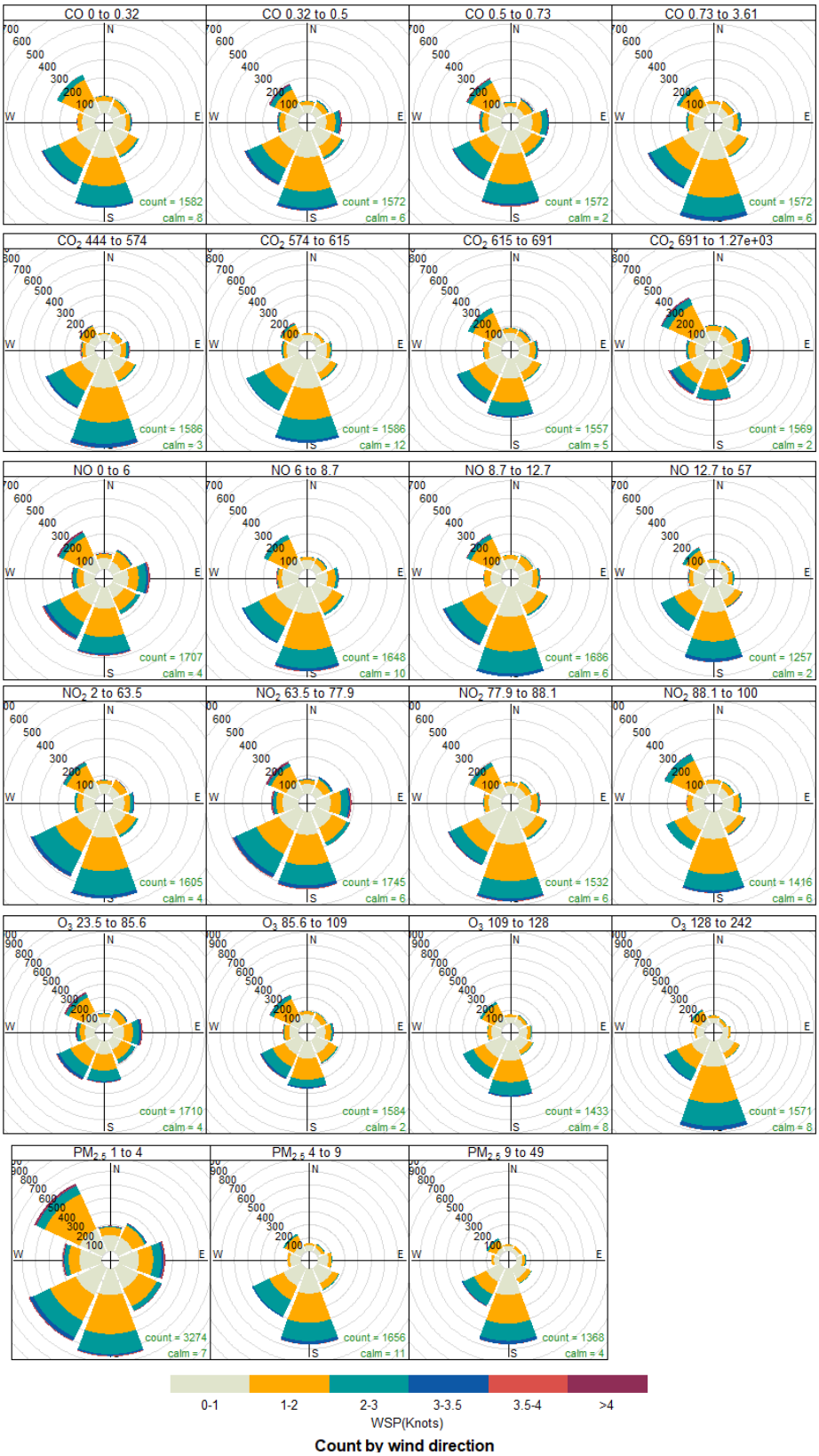


Figure 9: Plots of Wind Speed/direction Counts by Different Levels of Pollution Concentrations.

4.4 Correlation Matrices

In this section, the dependence between variables was investigated and assessed by the computed Pearson correlation coefficient. Correlation measures the monotonic association between variations of two numeric variables (Schover, Boer, & Schwarte, 2018). In other words, in correlated data, the change of value in one variable can be reflected in the change of value in another variable, either in the same (positive) or opposite (negative) direction. Correlation coefficients are scaled and thus ranging from -1 to +1, where the closer the value is to 0, the less correlation between the two variables is suggested. On the contrary, a correlation value closer to +1/-1 indicates a stronger positive/negative relationship in variables.

The final part of this correlation analysis on our data was conducted in three aspects: 1. the interrelationship of air pollutants; 2. the interrelationship of non-pollutant variables; and 3. the correlations between air pollutants and other variables.

The correlations between each two air pollutant variables are summarized in a correlation matrix displayed in Figure 10. The correlation matrix is a powerful statistical tool that incorporates informative graphs and coefficients in one place. With each variable occupying one row and one column of the matrix, the diagonal shows the density plot of each variable, the top shows the computed correlation values for each pair of variables, and the bottom shows the corresponding scatter plots. Moderate and positive correlations of O_3 – PM2.5 and O3 – NO_2 were indicated in the figure with a correlation value higher than 0.4. The correlation coefficients for the remaining pairs demonstrated a relatively weak or slight association between those variables.

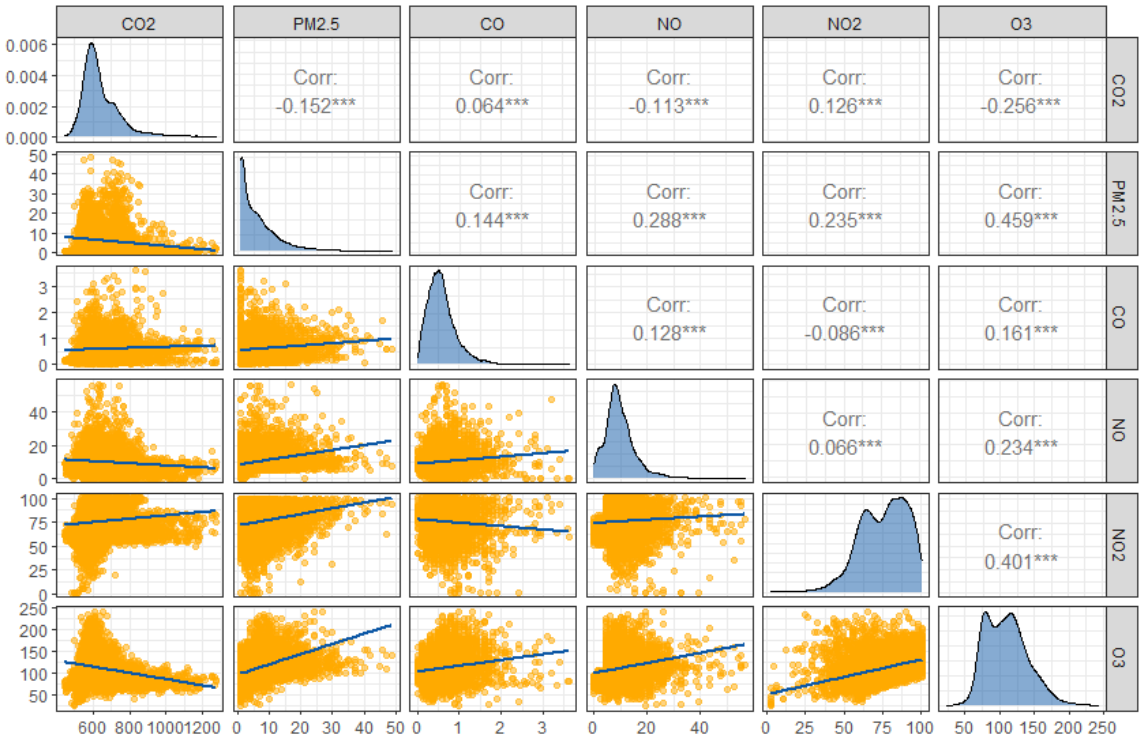


Figure 10: Correlation Matrix of Air Pollutants.

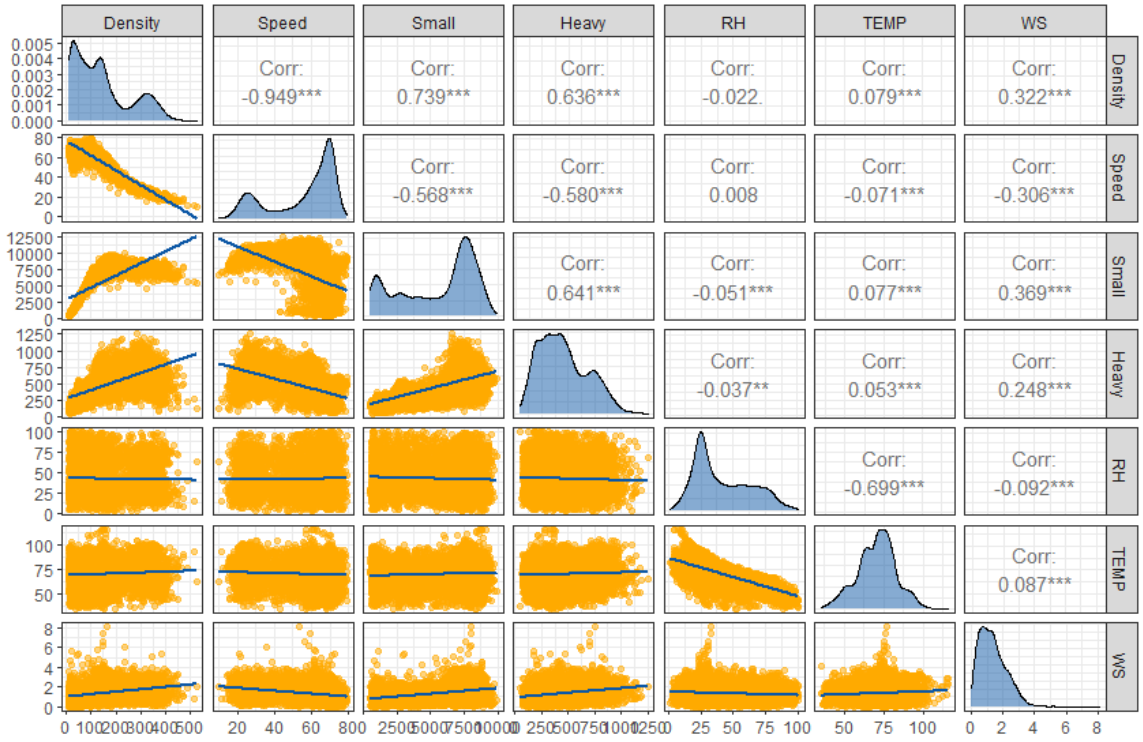


Figure 11: Correlation Matrix of Non-Pollutant Variables.

Next, we also created a correlation matrix for non-pollutant variables to estimate their associations since they would be potentially selected as the explanatory variables in the prediction models. Highly correlated predictors should be avoided because one of the fundamental assumptions for linear regression models is that all input variables are independent of each other or are weakly correlated. Using highly correlated predictors in regression models could result in multicollinearity, which reduces the precision of the estimated coefficients and thus weakens the statistical significance of the model (Frost, 2021). Therefore, it is essential to investigate the dependencies of all predictors and perform data screening prior to regression modeling.

Figure 11 shows the correlations of traffic-related and meteorological variables. Density/Speed was found to have a perfectly negative correlation with a correlation value of – 0.949. In addition, temperature (TEMP) was negatively associated with relative humidity (RH). Hence, we selected the speed and humidity factor to be left out of the dataset structure. Although the correlation value also indicated a relatively strong relationship between truck flow (Heavy) and non-truck flow (Small), we still kept them as explanatory variables for regression models considering that they are not dependent on each other by true definition. The relationship of these two kinds of flow is not casual; instead, in essence, they merely share some characteristics in common due to the involvement of human behaviors.

Lastly, the correlation coefficients for each pair of pollutant/non-pollutant variables are summarized in Table 4. The shaded shells indicated a moderate to strong correlation between the row variable and column variable, which were PM2.5/RH (+), NO_2/RH (+), O_3/RH (+), CO/TEMP (+), and NO_2/TEMP (-). However, to our surprise,

all the traffic flow characteristics were suggested to have slight linear correlations with pollutant emissions. These results could imply the nonlinearity in their associations and the complexity of pollutant emissions induced by traffic.

Table 4: Correlation Coefficients for Paired Pollutant/Non-Pollutant Variables.

	Flow	**Density**	**Speed**	**Small**	**Heavy**	**RH**	**TEMP**	**WS**
CO2	-0.020	-0.020	0.006	-0.020	-0.024	0.055	-0.236	0.037
PM2.5	0.017	0.052	-0.057	0.016	0.019	**0.354**	0.002	-0.021
CO	0.059	0.065	-0.061	0.058	0.052	-0.245	**0.347**	0.013
NO	0.048	0.048	-0.045	0.050	0.014	0.086	0.044	-0.034
NO2	-0.047	-0.030	0.018	-0.048	-0.017	**0.676**	**-0.724**	-0.083
O3	0.070	0.089	-0.092	0.067	0.080	**0.297**	-0.005	-0.019

CHAPTER 5: PREDICTIVE REGRESSION MODELS

In this chapter, two statistical regression models and one machine learning algorithm will be briefly introduced, and prediction outcomes presented systematically. As previously mentioned, the goals of applying these models are to predict the air pollutant concentrations one hour ahead of the present observations and retrieving better knowledge by looking into the estimated coefficients of models.

In practice, not all the amount of a pollutant observed at time t is newly produced during that time interval. Instead, it is constituted by partial residuals from the previous hour time t-1 due to the dispersion to a certain degree and a partial new production of pollutants from various sources during time t (Juhos, Makra, & Toth, 2008). It is of great necessity to assess the contribution of traffic characteristics to the concentrations of pollutants that are newly formed from hour to hour in the regression models. However, neither the residuals nor the new formation of pollutants is measurable; therefore, we simplify the model inputs by considering the past values of one hour before as the base value of residuals and include them into the models as predictors of future values. The fundamental structure of the equation for all prediction models can be summarized as below:

$$Y_{t+1} = f(X1_t, X2_t, \dots, Xp_t) = f(Density_t, Small_t, Heavy_t, TEMP_t, WS_t, Y_t) \quad (1)$$

where the subscript t stands for the time in one-hour increment, Y_t and Y_{t+1} represent the current concentration of one pollutant and predicted concentration in one-hour future, respectively. Note that there is a simple assumption underlying all regression models that the current concentration of one pollutant used for prediction is an independent variable to other predictors, although it is literally denoted by Y_t.

We performed the following prediction models to investigate the contributions of traffic parameters and meteorological factors on ambient air pollution: multiple linear regression, stepwise regression, and artificial neural network (ANN).

5.1 Multiple Linear Regression Model

The linear regression model has been extensively used in data science and prediction modeling because of its primitiveness. It attempts to model the relationship between multiple explanatory variables and one response variable under the assumption that the line of best fit to the observed data is linear. The equation of linear regression is stated as below:

$$Y_{t+1} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + \epsilon \qquad (2)$$

where, for $i = 1, 2, \ldots p$, $Y_{i\,t+1}$ = dependent variable at time $t+1$, $X_{i\,t}$ = independent variables at time t, β_0 = constant term (y-intercept), β_p = estimated coefficient for each independent variable, and ε = error term of the model.

We created one linear regression model for each target pollutant, of which coefficients are summarized in Table 5. Although all independent variables were inputted into the models and assigned an estimated coefficient, it did not necessarily mean that they were recognized as significant predictors. Coefficients shaded and in bold are those coefficients identified as statistically significant with a p-value less than 0.05. We then looked into some estimated coefficient values to understand how much the response variable will change based on the change in one explanatory variable.

Firstly, we started by examining the coefficients for traffic parameters. The density was determined statistically significant for NO_2 and O_3 concentrations. One unit increase in density caused an increase in NO_2 and O_3 concentrations in the next hour by

0.00723 μg/m^3 and 0.0173 μg/m^3, respectively. Similarly, one additional passenger vehicle in the traffic flow would raise the CO_2 concentrations by 0.00227 ppm. However, the model also indicated truck flow to be negatively contributive to NO_2, which is contrary to our common ideas.

In terms of meteorological factors, the temperature was indicated as a significant predictor for all models. It was positively correlated to CO, NO, and PM2.5 while negatively associated with the remaining pollutants, which perfectly matched previous studies and our findings in correlation analysis (Maynard, Holgate, Koren, & Samet, 1999). It could be concluded from this table that the higher wind speeds, the lower concentrations of NO and O_3, which is also partially aligned with our findings with the help of wind rose plots. Last but not least, the concentration at time t was statistically significant to the concentration at time t+1 for all air species, which is consistent with preexisting research by Juhos, Makra & Toth (2008). The actual-by-predicted plots shown in Figure 12 indicated a relatively outstanding performance in the O_3 and PM2.5 models based on their shape of scatter. CO_2 model predicted better on the lower end, and CO and NO models performed below satisfaction.

Table 5: Summary of Coefficients for Linear Regression Models.

MLM	CO_{t+1}	$CO2_{t+1}$	NO_{t+1}	$NO2_{t+1}$	$O3_{t+1}$	$PM2.5_{t+1}$
Intercept	**1.34E-01**	**1.98E+02**	**4.12E+00**	**9.93E+01**	**6.68E+01**	**1.14E+00**
$Density_t$	1.10E-04	-2.04E-02	6.90E-04	**7.23E-03**	**1.73E-02**	1.62E-03
$Small_t$	-3.31E-06	**2.27E-03**	3.05E-05	-1.42E-04	**-4.43E-04**	-7.06E-05
$Heavy_t$	-6.82E-05	-8.70E-04	5.89E-04	**-3.39E-03**	-1.64E-03	-7.42E-04
$TEMP_t$	**4.48E-03**	**-2.85E-01**	**4.30E-02**	**-5.86E-01**	**-3.92E-01**	**1.13E-02**

WS_t	5.25E-03	1.13E+00	**-3.24E-01**	-4.47E-01	**-1.44E+00**	-2.06E-02
CO_t	**3.21E-01**	-	-	-	-	-
$CO2_t$	-	**7.12E-01**	-	-	-	-
NO_t	-	-	**2.86E-01**	-	-	-
$NO2_t$	-	-	-	**2.69E-01**	-	-
$O3_t$	-	-	-	-	**6.74E-01**	-
$PM2.5_t$	-	-	-	-	-	**7.76E-01**

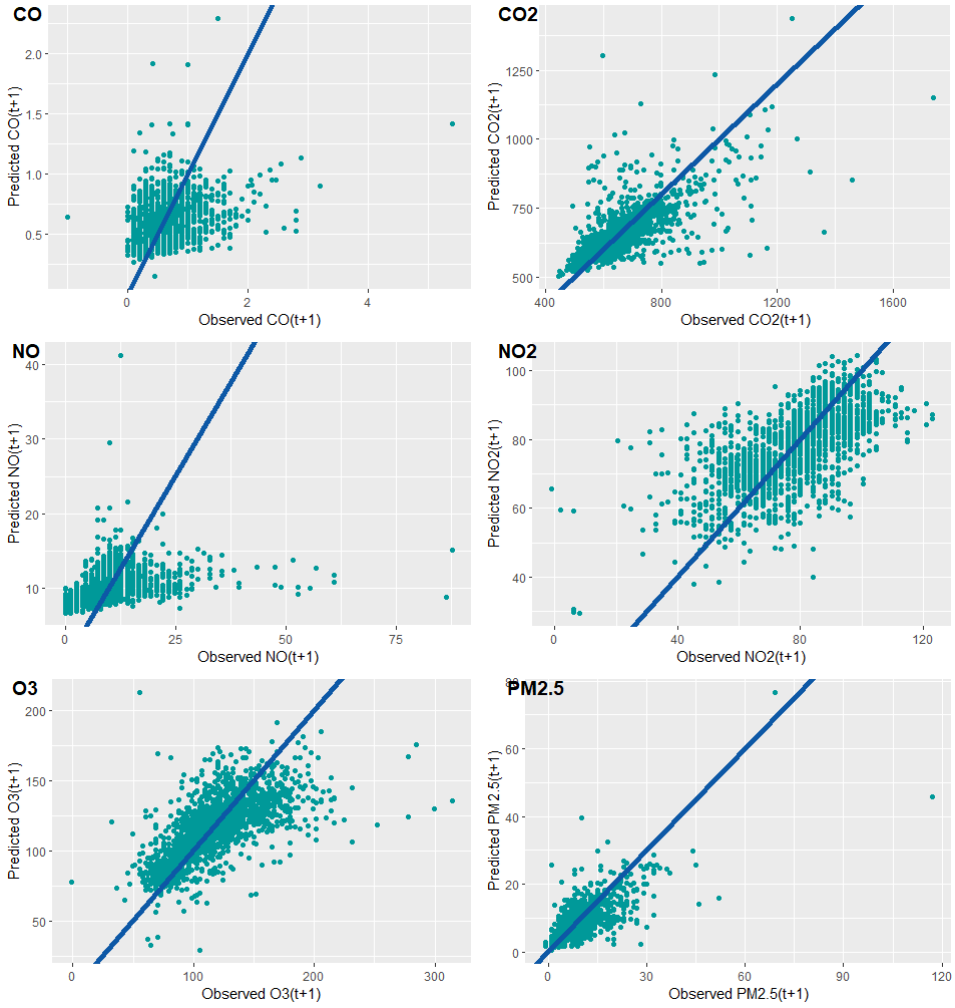


Figure 12: Actual by Predicted Plots for Linear Regression Models

5.2 Stepwise Regression Model

The mechanism within a stepwise regression is to iteratively add or remove predictors in the models and create subsets. After each step, the model itself would compute statistical metrics such as R-squared and AIC for comparing the model performance of each selected sub dataset. Stepwise regression has the ability to discern significant variables that give the best fit results without human intervention. There are three types of stepwise regressions. Forward selection, which starts with no predictors in the model, continuously adds the most contributive predictors until no improvement is statistically significant. Backward selection starts with all predictors in the models, iteratively remove the least contributive variables until the remaining predictors are proved statistically significant. Stepwise selection is a combined methodology of the forward and backward selection. In this section, only the backward selection was performed to screen for the most significant predictors for our target air pollutants. All coefficients are shown in Table 6.

Instead of incorporating all predictors into the model, the backward selection filtered out the insignificant variables, leaving their coefficients blank. In the table, we observed positive contributions of the traffic density to more air pollutants such as CO, NO_2, and O_3 except for a negative contribution on CO_2. This finding matched with Brief, Jones & Yoder's research that traffic density is correlated to CO emissions (2012). However, the regular passenger vehicle flow and truck flow were shown as negatively associated with concentrations of CO, NO_2, and O_3. It could be because NO_2 and O_3 established a decreasing trend during afternoon traffic peak hours. We also noticed something interesting in the table that the trained model for PM2.5 only used the y-

intercept and past concentration values as the significant predictors. This result could be potentially implying the sensitivity of concurrent concentrations to residuals in the past hours.

The actual-by-predicted plots presented in Figure 13 showed similar shapes of scattering to those of linear regression models. Remarkably, even the PM2.5 model merely used the past values of concentration to train the model, the prediction outcomes were still quite promising despite those extreme outliers. It could possibly indicate that a significant amount of residual PM2.5 did not experience much dispersion and diffusion in the ambient air on freeways. The emission of PM2.5 from other sources, exceptionally from the transportation sector, could become minor compared to massive residuals.

Table 6: Summary of Coefficients for Stepwise Regression Models

Stepwise	**CO_{t+1}**	**$CO2_{t+1}$**	**NO_{t+1}**	**$NO2_{t+1}$**	**$O3_{t+1}$**	**$PM2.5_{t+1}$**
Intercept	1.76E-01	2.01E+02	3.88E+00	9.95E+01	6.37E+01	1.32E+00
$Density_t$	9.57E-05	-4.31E-02		4.98E-03	2.10E-02	
$Small_t$	-5.97E-06	3.54E-03	7.84E-05		-7.01E-04	
$Heavy_t$	-3.51E-05			-3.55E-03		
$TEMP_t$	4.05E-03	-3.77E-01	4.25E-02	-6.02E-01	-3.62E-01	
WS_t			-2.79E-01		-1.17E+00	
CO_t	3.13E-01	-	-	-	-	-
$CO2_t$	-	7.10E-01	-	-	-	-
NO_t	-	-	3.07E-01	-	-	-
$NO2_t$	-	-	-	2.68E-01	-	-
$O3_t$	-	-	-	-	6.85E-01	-
$PM2.5_t$	-	-	-	-	-	7.94E-01

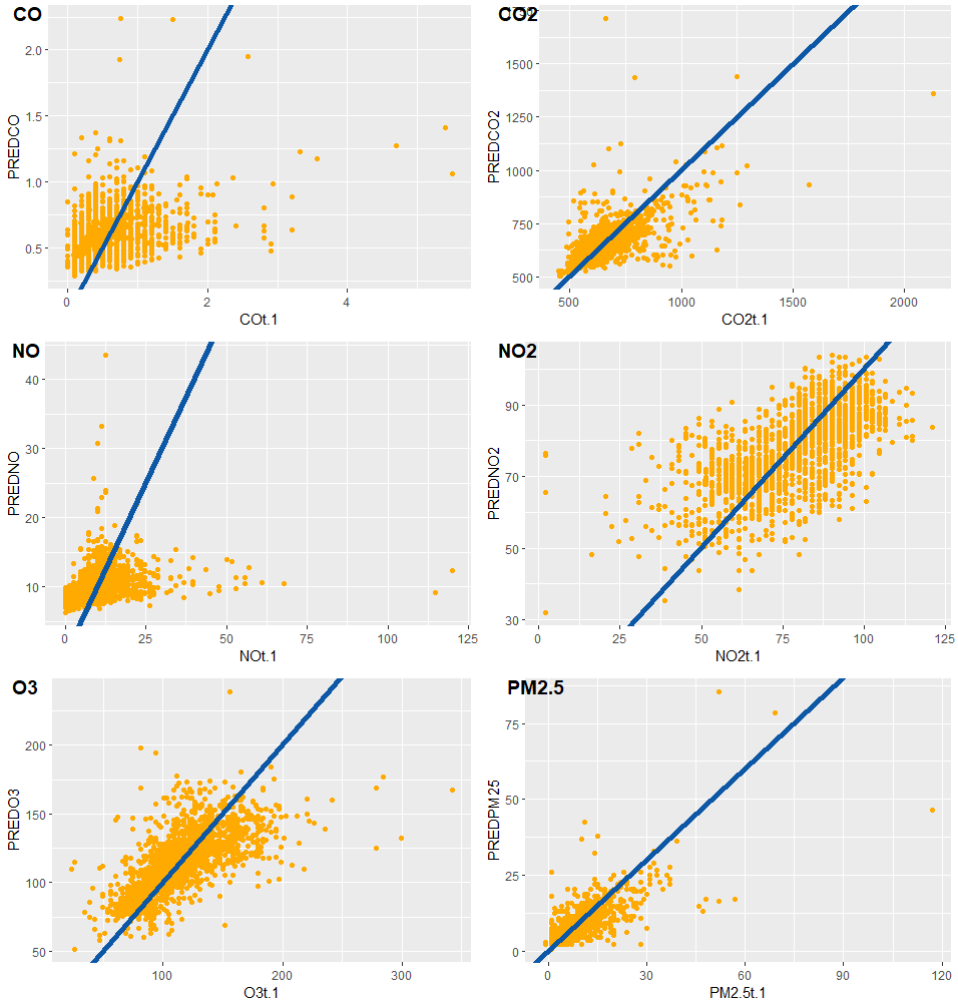


Figure 13: Actual by Predicted Plots for Stepwise Backward Regression Models.

5.3 Artificial Neural Network Model

Artificial Neural Networks (ANN) are a powerful machine learning algorithm that has been extensively used in data science. The architecture of a neural network, analogous to the human brain in which neurons interconnect with each other, comprises interconnected information processing units, shown in Figure 14. Also known as a multi-layer perceptron, a typical neural network consists of three or more layers: the first layer being the input layer (green units), and the final layer being the output layer (yellow unit). The layers between the input and output layers are named hidden layers, in which

the number of hidden units specified by the users. The inputs will initially enter the first layers and pass successively through all the hidden layers until the final layer is reached.

For each unit *i* of each hidden layer *l*, the value of the unit $h_i^{(l)}$ is computed as the values of the units connected to $h_i^{(l)}$ using the equation below:

$$h_j^{(l)} = \sum_{i=1}^{d} x_i \omega_{ji} + b_l \qquad (3)$$

where *i* is the number of units in the *l-1* layer, *j* is the specified number of unit in the current layer, ω_{ji} is the parametrized weights connecting layer *l-1* to the current layer, and b_l is the bias applied to the current layer (Bellinger C. , Jabbar, Zaiane, & Osornio-Vargas, 2017). Non-linear relationships between input and output variables can be established through those information processing units in such a neural network architecture.

Unlike the trained regression models in previous sections, where we were able to retrieve coefficient values for the models, the artificial neural network merely provides prediction results and neural network layouts (see Figure 16 – 21 in Appendix). The actual-by-predicted plots were presented in Figure 15 below. The predictions for NO_2 were significantly improved since more data points were scattered close to the actual = predicted line, except some of the high values were severely underestimated. CO_2, O_3, and PM2.5 models also provided satisfactory results, while the predictions for CO were estimated very poorly.

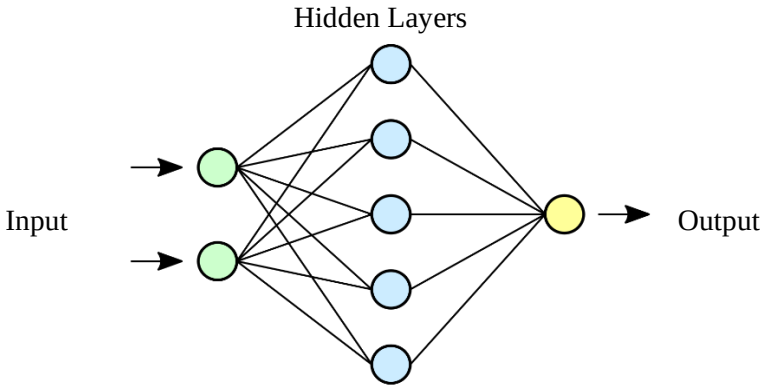


Figure 14: Example of Neural Network Architecture. (Source: online)

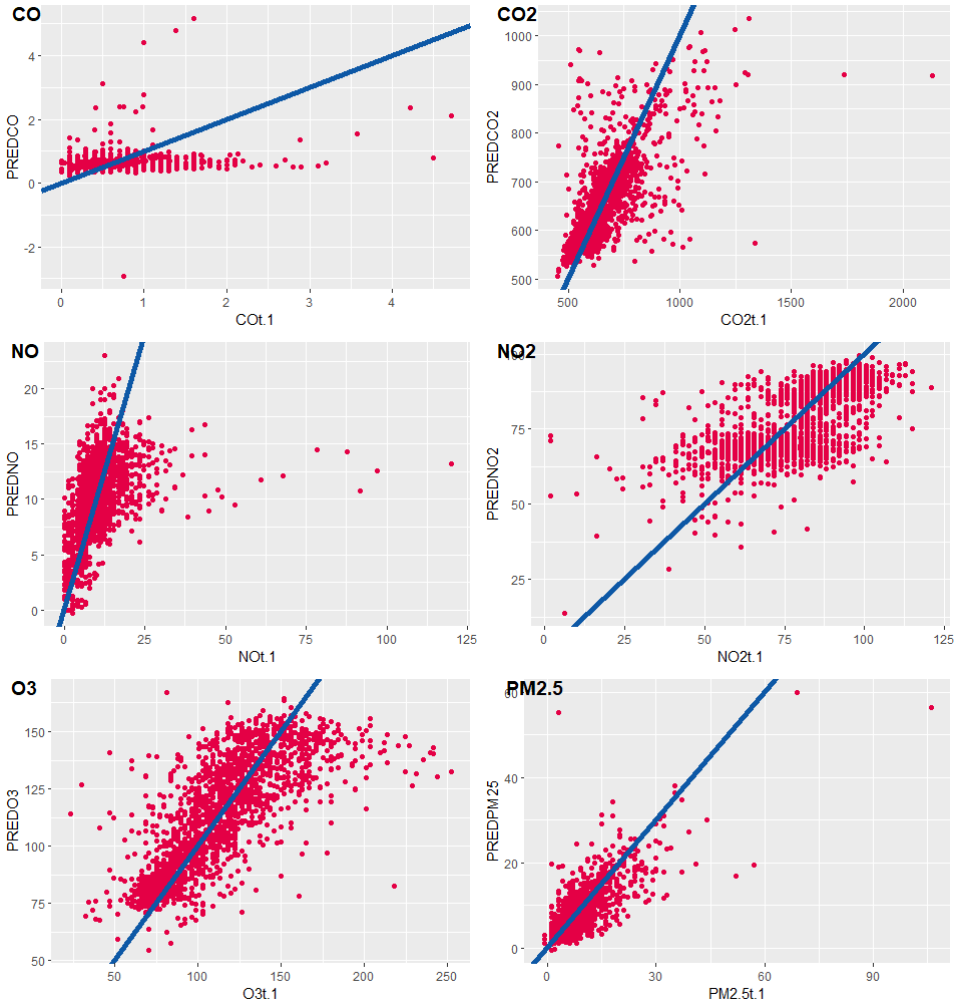


Figure 15: Actual by Predicted Plots for Artificial Neural Networks.

CHAPTER 6: OUTCOMES AND DISCUSSION

In this chapter, we inspected the model performance by comparing two statistical metrics: R-squared (R^2) value and root mean square error (RMSE). These metrics are calculated as below summarized in Table 7 and Table 8:

$$R^2 = 1 - \frac{\sum(y - y')^2}{\sum(y - \bar{y})^2} \qquad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{p=1}^{n}(y'_p - y_p)^2} \qquad (5)$$

where n is the number of observations, y'_p is the predicted value of y_p, the pth observation of the response variable. The higher R^2 and lower RMSE values, the better the model performance is. Overall, ANN accounted for most of the best-performed models among the three, with winning R^2 and MRSE values among NO_2, O_3, and PM2.5 prediction models. The multiple linear regression achieved both the best R^2 and RMSE values among CO_2 prediction models. However, it was difficult to determine the best prediction model for NO since their performance was average in many ways.

Table 7: R-squared Values for All 1-hour Prediction Models.

	CO	CO2	NO	NO2	O3	PM2.5
MLR	0.1078	**0.5174**	0.1208	0.4443	0.5017	0.6621
Stepwise	**0.1190**	0.4926	0.0858	0.4179	0.4958	0.6366
Neural Network	-0.1068	0.5021	**0.1680**	**0.4602**	**0.5513**	**0.6821**

Table 8: RMSE Values for All 1-Hour Prediction Models.

	CO	CO2	NO	NO2	O3	PM2.5
MLR	**0.3804**	**80.2468**	**6.6192**	12.3996	24.2644	4.1720
Stepwise	0.4177	83.2734	7.3360	12.5701	23.7313	4.3907
Neural Network	0.4429	87.5239	7.0653	**12.1340**	**22.2254**	**3.9933**

Speaking of models for each pollutant, the PM2.5 models overall were surprisingly well-performed because the R^2 values for all three prediction models were close and higher than 0.6, which means that more than 60% variance of observed PM2.5 can be explained by the predictors within each model. The predictor commonly considered as significant contributors to PM2.5 concentrations by MLM and Stepwise is the past concentrations of PM2.5. It might be because most PM2.5 in the ambient air tend to stay still regardless of other factors incorporated in this thesis.

Models trained for CO_2, NO_2, and O_3 also achieved satisfactory prediction outcomes in general. All CO_2 models demonstrated their accuracy in predicting low CO_2 concentrations, while it was downgraded when CO_2 concentrations were in a higher range. In actual-by-predicted plots for NO_2 and O_3 in all models, data points appeared scattered like a bunch while still lying along the regressed diagonal line. Moreover, NO_2 and O_3 concentrations were more precisely predicted using the ANN algorithm, indicating that it is more suitable to consider non-linear relationships between NO_2/O_3 and their potential predictors. The models are also expected to improve by considering the correlations between NO/NO_2/O_3 in the prediction modeling, as demonstrated in previous studies (Maynard, Holgate, Koren, & Samet, 1999).

Finally, CO and NO models performed poorly regardless of the methodology. The ANN model predicting CO concentrations scored worst, resulting in a negative R^2. The formation of CO and NO might be much more complicated that the most contributive factors for them other than traffic and meteorological variables were not considered in this study.

CHAPTER 7: CONCLUSION

7.1 Concluding Remarks

In this work, with the intention of investigating the correlations between ambient air pollution and traffic, we present the steps for correlation analysis and prediction modeling. A dataset containing hourly data of traffic flow characteristics (passenger vehicle flow, truck flow, density, and speed), air pollutant concentration samples (CO, CO_2, NO, NO_2, O_3, and PM2.5), and meteorological factors (temperature, humidity, wind speed, and wind direction) continuously collected on urban freeways for a year underlies this thesis.

We first begin with probing into the temporal distribution of traffic parameters and air pollutant concentrations, in which we understand the dynamics of the traffic and some significant daily patterns of air pollutants at the studied freeway. Next, we explore the seasonality of target pollutants, which essentially indicates the potential impacts of outdoor temperature on pollutant concentrations. It is notably suggested that NO, NO_2, O_3, and PM2.5 tend to fluctuate widely from month to month. We further explore the contribution of wind properties to air pollutant concentrations. From the wind rose plots, we obtain a basic knowledge that our study site is dominated by winds from the south and southwest at relatively high wind speeds. It is furthered noticed that low CO_2 but high O_3 concentrations are potentially concerned with high wind speeds from the south.

Since one of our goals is to apply different models to our data to predict future pollutant concentrations, it is essential to investigate the correlations and dependencies among the potential input variables. From the correlation matrices, the analysis indicates some moderate to strong correlations between O_3-PM2.5, O_3-NO_2, TEMP-RH, Truck-

Non-truck, and Density-Speed. We perform a data screening to ensure that the selected explanatory variables for prediction models are not highly correlated.

For each pollutant's prediction modeling, the density, passenger vehicle flow, truck flow, temperature, wind speed, and concentration of that pollutant at time t are used as the input variables to predict the concentration of that pollutant at time t+1. We perform prediction models on CO, CO_2, NO, NO_2, O_3, and PM2.5; for each pollutant species, a multiple linear regression, stepwise regression, and artificial neural network are applied to aim for the best fit.

In terms of model selection by pollutants, artificial neural networks give the best results for most pollutants such as NO_2, O_3, and PM2.5. Speaking of the goodness of fit performed by models on pollutants, it is sufficiently suggested that CO_2, NO_2, O_3, and PM2.5 can be predicted with regression models or neural networks with some satisfactory outcomes. In addition, the estimated coefficients of predictors for each model are very informative; for example, all models indicate the statistically significant impact of the pollutant's past concentrations on its future concentrations. Both regression models point out the positive contribution of density to CO, NO_2, and O_3 emissions. However, such contribution of traffic flow on pollutant emissions seems insignificant because traffic flow cannot substantially represent the intensity of traffic. For example, traffic flow can be low when only a few vehicles travel or during severe traffic congestion where only a few vehicles can move.

7.2 Limitations and Future Directions

We have identified some limitations reoccurring challenges along with our study progress. First of all, this is a study of traffic and air pollution using local data. Findings

from preexisting studies might be inconsistent with ours. Similarly, the finding in this paper is not necessarily applicable to new locations.

Secondly, we are unable to remove the noise or background air pollutant concentrations as they are unmeasurable. This results in potentially causing unavoidable bias in our trained prediction models. Last but not least, we merely applied the simplest regression models on our unique data, which few researchers have access to or analyze similar datasets in previous research. We believed wind direction could potentially have significant impacts on pollutant dispersion; however, we could not use it as one of the inputs for predictive regression models due to their inapplicability to categorical variables.

There is always a necessity of analyzing the environmental impacts of traffic for transporation planners and stakeholders. If traffic flow characteristics were statistically proved to have significant impacts on the environment and especially the air quality, then transportation planners and professionals would be able to ultimately reduce traffic-induced air pollution by applying various strategies on operational transportation systems to take controls of those traffic parameters.

REFERENCES

Bellinger, C., Jabbar, M. S., Zaiane, O., & Osornio-Vargas, A. (2017). A Systematic Review of Data Mining and Machine Learning for Air Pollution Epidemiology. *BMC Public Health*.

Bellinger, C., Jabbar, M. S., Zaiane, O., & Osornio-Vargas, A. (2017). A Systematic Review of Data Mining and Machine Learning For Air Pollution Epidemiology. *BMC Public Health, 17:907*.

Brief, R. S., Jones, A. R., & Yoder, J. D. (2012). Lead, Carbon Monoxide and Traffic. *Journal of the Air Pollution Control Association*, 384-388.

Division of Traffic Operations. (2018). *2016 Traffic Volumes on the California State Highway System.* Sacramento, CA: California Department of Transportation.

Federal Highway Administration. (2017, June 27). *History of the Interstate Highway System*.

Foster, M. S. (2003). *Nation on Wheels.* Belmont, CA: Thomson, Wadsworth.

Frey, H. C., Unal, A., Rouphail, N. M., & Colyar, J. D. (2003). On-Road Measurement of Vehicle Tailpipe Emissions Using a Portable Instrument. *Journal of the Air & Waste Management Association*, 992-1002.

Frost, J. (2021). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*.

Juhos, I., Makra, M., & Toth, B. (2008). Forecasting of Traffic Origin NO and NO2 Concentrations by Support Vector Machines and Neural Networks Using Principal Component Analysis. *Simulation Modelling Practice and Theory*, 1488-1502.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., & Vermeulen, R. C. (2019). Performance of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces. *Environmental Science & Technology*, 1413-1421.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., . . . Cawley, G. (2003). Extensive Evaluation of Neural Network Models for the Prediction of NO2 and PM10 Concentrations, Compared with a Deterministic Modelling System and Measurements in Central Helsinki. *Atmospheric Environment*, 4539-3550.

Li, L., Zhou, X., & Tong, W. (2019). Analysis of Exposure to Ambient Air Pollution. In L. Li, X. Zhou, & W. Tong, *Spatiotemporal Analysis of Air Pollution and Its Application in Public Health* (pp. 217-225). Elsevier.

Maynard, R. E., Holgate, S. T., Koren, H. S., & Samet, J. M. (1999). *Air Pollution and Health.* Elsevier.

Miller, B. G. (2011). The Effect of Coal Usage on Human Health and the Environment. In *Clean Coal Engineering Technology* (pp. 85-132). Butterworth-Heinemann Ltd.

Padula, A. M., Mortimer, K., Hubbard, A., Lurmann, F., Jerrett, M., & Tager, I. B. (2012). Exposure to Traffic-related Air Pollution During Pregnancy and Term Low Birth Weight: Estimation of Casual Associations in a Semiparametric Model. *Am J Epidemiol*, 815-824.

Perez, P., & Trier, A. (2001). Prediction of NO and NO2 Concentrations Near a Street with Heavy Traffic in Santiago, Chile. *Atmospheric Environment*, 1783-1789.

Schover, P., Boer, C., & Schwarte, L. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia, Volume 126, Issue 5*, 1763-1768.

Scottish Environment Protection Agency. (n.d.). The Chemistry of Air Pollution. Scottland: SEPA.

The State of Queensland. (2017, March 27). *Meteorological factors*. Retrieved from Queensland Government.

Tischer, V., Fountas, G., Polette, M., & Rye, T. (2019). Environmental and Economic Assessment of Traffic-related Air Pollution Using Aggregate Spatial Information: A Case Study of Balneário Camboriú, Brazil. *Journal of Transport & Health*.

United States Environmental Protection Agency. (2018, March). *Greenhouse Gas Emissions from a Typical Passenger Vehicle.*

United States Environmental Protection Agency. (2020, June). *Fast Facts: U.S. Transportation Sector Greenhouse Gas Emissions 1990-2018.*

Vakkilainen, E. K. (2017). Solid Biofuels and Combustion. In E. K. Vakkilainen, *Steam Generation from Biomass: Construction and Design of Large Boilers* (pp. 18-56). Butterworth-Heinemann Ltd.

Volk, H. E., Lurmann, F., Hertz-Picciotto, I., & McConnell, R. (2012, November 26). *Traffic-Related Air Pollution, Particulate Matter, and Autism.* Retrieved from JAMA Psychiatry.

Wang, G., Makino, K., & Wu, X. (2018). Investigating Correlations and Prediction Models among Roadside Air Pollutants, Meteorological Factors and Traffic Flow on Urban Arterials using High Frequency Data.

Wang, X., Bi, X., Sheng, G., & Fu, J. (2006). Chemical Composition and Sources of PM10 and PM2.5 Aerosols in Guangzhou, China. *Environ. Monit. Assess.*, 1-5.

Wang, Y., Munger, J., Xu, S., McElroy, M., Hao, J., Nielsen, C., & Ma, H. (2010). CO2 and its Correlation with CO at a Rural Site Near Beijing: Implications for Combustion Efficiency in China. *Atmospheric Chemistry and Physics*, 8882-8897.

Weber, J. (2012). The Evolving Interstate Highway System and the Changing Geography of the United States. *Journal of Transport Geography*, 70-86.

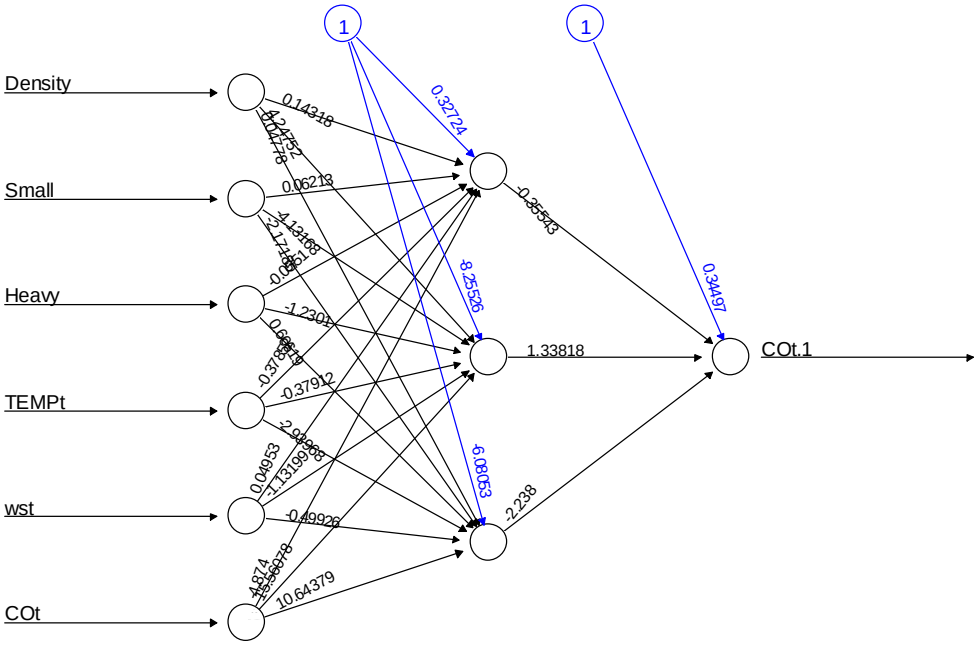
World Bank Group. (1998). Ground-Level Ozone. In W. B. Group, *Pollution Prevention and Abatement Handbook* (pp. 227-230).

Yan, Q., Liew, Z., Uppal, K., Cui, X., Ling, C., Heck, J. E., . . . Ritz, B. (2019). Maternal Serum Metabolome and Traffic-related Air Pollution Exposure in Pregnancy. *Environmental International*.

Zalakeviciute, R., Bastidas, M., Buenano, A., & Rybarczyk, Y. (2020). A Traffic-Based Method to Predict and Map Urban Air Quality. *Applied Science, Volume 10, Issue 6*.

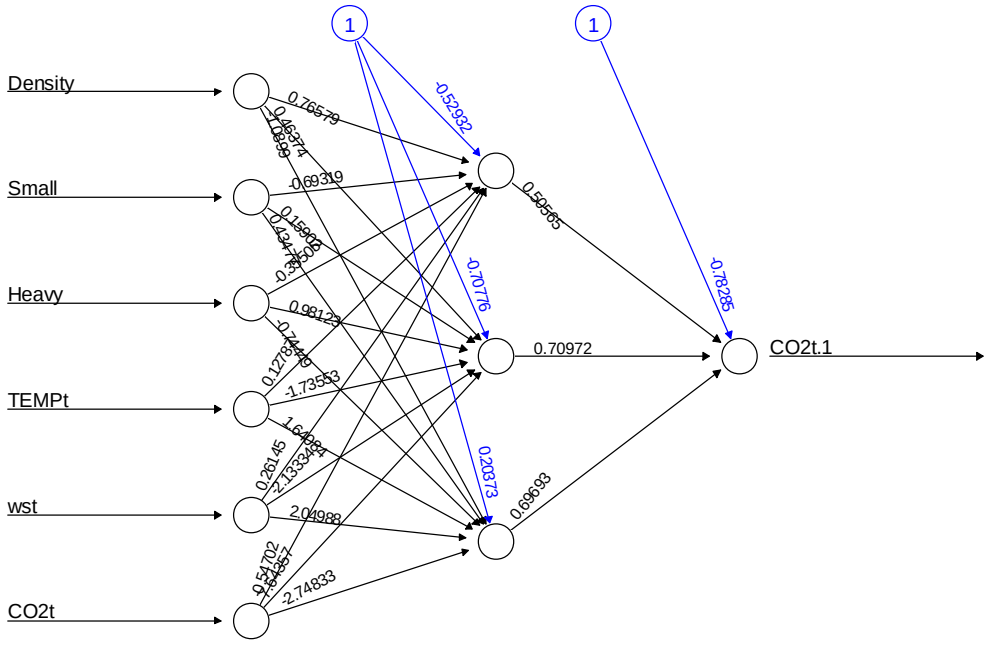
Zhou, X., Cao, Z., Ma, Y., Wang, L., Wu, R., & Wang, W. (2016). Concentrations, Correlations and Chemical Species of PM2.5/PM10 Based on Published Data in China: Potential Implications for the Revised Particulate Standard. *Chemosphere*, 518-526.

APPENDIX



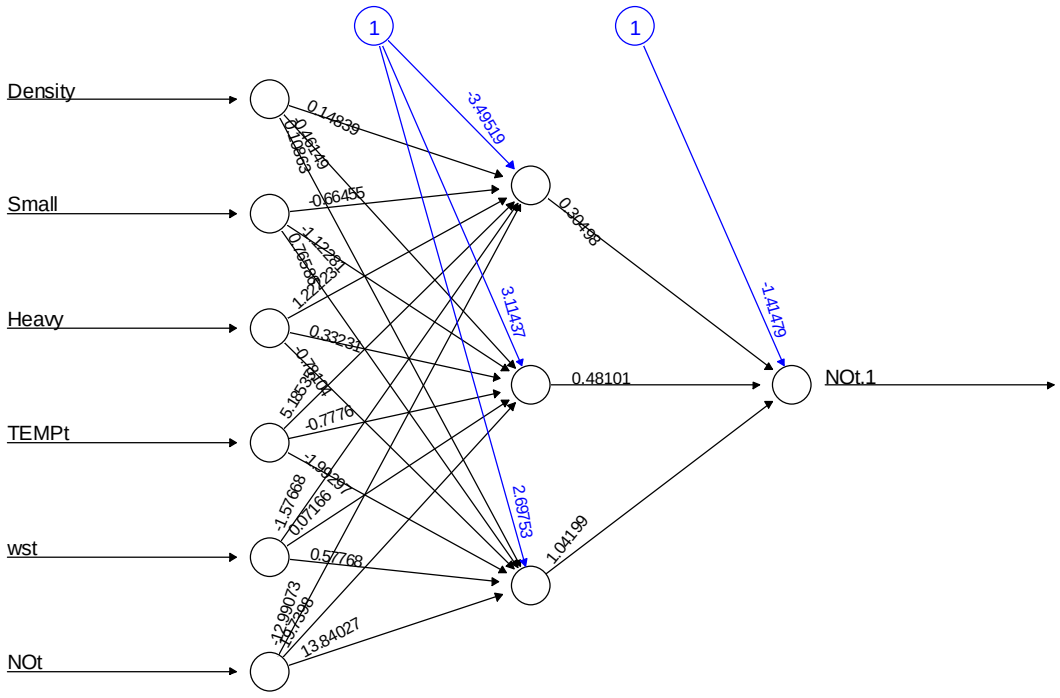
Error: 7.14604 Steps: 12271

Figure 16: Neural Network of CO.



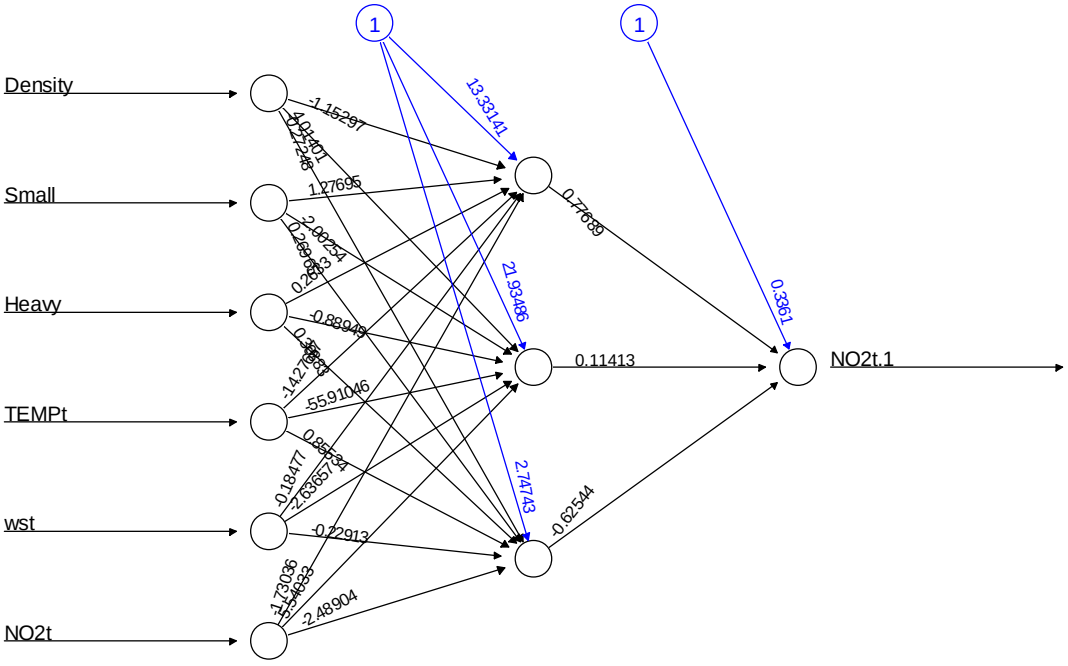
Error: 4.851369 Steps: 21127

Figure 17: Neural Network of CO_2.



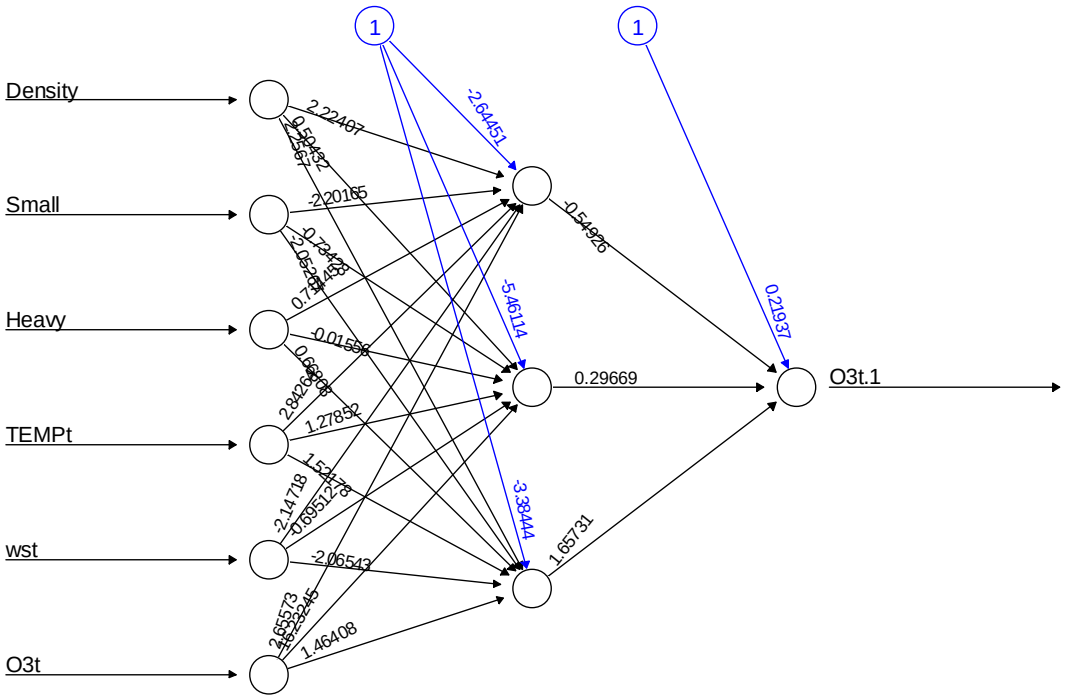
Error: 4.705349 Steps: 4114

Figure 18: Neural Network of NO.



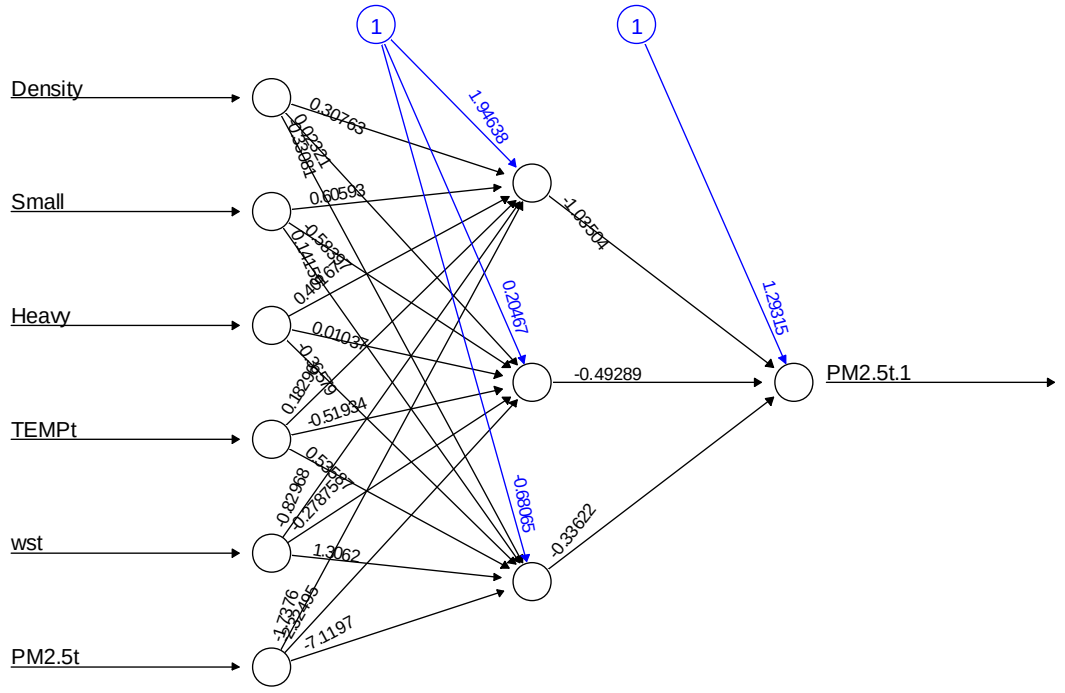
Error: 20.403048 Steps: 61664

Figure 19: Neural Network of NO_2.



Error: 9.109391 Steps: 39899

Figure 20: Neural Network of O_3.



Error: 2.748883 Steps: 2056

Figure 21: Neural Network of PM2.5.