Fusion of Real-Time Congestion Pattern to Land Use Regression Model for Air Pollution

A THESIS REPORT

Submitted in the partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY

IN

CIVIL ENGINEERING

(With specialization in Transportation Engineering)

Submitted by

Chinmay Agarwal

20524003

Under the Guidance of

Dr. Amit Agarwal

Assistant Professor
Civil Engineering Department

IIT Roorkee

Department of Civil Engineering

Indian Institute of Technology, Roorkee

May 2022

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled **"Fusion of Real-Time Congestion Pattern to Land Use Regression Model for Air Pollution "** submitted by **CHINMAY AGARWAL** bearing the **Enrollment No: 20524003**, in the partial fulfilment of the requirements for the award of degree of Master of Technology in Civil Engineering with specialization in Transportation Engineering, submitted to the Department of Civil Engineering, Indian Institute of Technology Roorkee, is an authentic record of my own work carried out under the supervision of **Dr. Amit Agarwal** Department of Civil Engineering, IIT Roorkee. The matter embodied in this report has not been submitted for the award of any other degree.

CHINMAY AGARWAL
20524003
Date: May 7, 2022
Place: Roorkee

CERTIFICATE

This is to certify that the above declaration made by **CHINMAY AGARWAL** is correct to the best of my knowledge.

Dr. Amit Agarwal
Assistant Professor
Transportation Engineering
Department of Civil Engineering
Indian Institute of Technology, Roorkee

Date: May 7, 2022
Place: Roorkee

Acknowledgement

"Collaboration allows to capture each other's fund of collective intelligence." –Mike Schmoker

I would like to express my sincere gratitude and honor to my respected supervisor **Dr. Amit Agarwal**, Assistant Professor, Department of Civil Engineering, Indian Institute of Technology Roorkee, for his valuable guidance and consistent encouragement throughout the work. I am very thankful to him for his kind and useful suggestions and valuable time during the work period. I would also like to extend my indebtness to the Head of the Department, **Prof. Praveen Kumar** for providing the administrative support. I would also like to thank **Prof. Manoranjan Parida** for his constant support, encouragement and guidance. My work would have not been completed without OIC M. Tech. **Dr. Indrajit Ghosh** for standing with the students giving moral support at each step and counsel at any time. I would also like to genuinely thank **Dr. Pushpa Chaudhry** for providing all the required facilities at Transportation System, Analysis and Research Laboratory and specially to help me in getting the funding for the period of work. I would also like to thank all the faculties for their insightful teaching, educative classes and constant advices.

I am also very grateful to **Mr. Vikram Singh**, Research Scholar, Department of Civil Engineering, Indian Institute of Technology Roorkee, for his support. He provided his valuable help in reviewing the report and presenting various ideas. I would also like to thank **Ms. Rashmi Choudhary**, Research Scholar, Department of Civil Engineering, Indian Institute of Technology Roorkee for helping with the work flow and mentoring me. I would also like thank all the Ph.D. students for constantly helping me with my doubts.

Finally and most importantly I would like to express my sincere thanks to my beloved family for their blessings, friends/classmates and people who are part of this work in various ways, direct or indirect for their help and wishes.

CHINMAY AGARWAL

(**20524003**)

Abstract

With unprecedented advancement in mobility technology, most of the vehicles in the world are still operating on fuels from natural resources. On burning these vehicles contribute to air pollution significantly affecting life of every individual with maximum concern for the older as well as to newest generations. Thus, it becomes extremely important to measure and model air pollution and take preventive actions as efficiently and quickly as possible. For modelling, traffic characteristics like volume, and density near fixed monitoring sites plays an important role. These flow characteristics can also be coupled with nearby land use to give a better spatio-temporally varied model for pollutant prediction. Real-time congestion data can provide a fast and accurate measure of various pollutants that a person can expect on a particular route. It can significantly help non-motorized transit users and active users to plan their route based on the greenest route available.

In the present study, real-time congestion information is fused in a land-use regression model. The former is obtained from HERE maps Traffic Flow API (Application programming interface). To integrate land use information in the model, each raster pixel for the data inside the buffer region can be converted to a point, and a value is assigned to it, which is based on its distance from the monitoring station, land use, and traffic flow. Using these point data, regression analysis can be done to obtain a predictive model which can be used along any route to give a better-estimated value of pollutant concentration experienced by the user. These results can be integrated with map services to give a greener and safer route for active and non-motorized users leading to sustainable development.

Keywords : Air pollution, Land Use, Real-time traffic, Regression modelling

Contents

List of Figures

List of Tables

List of Abbreviations

API	Application programming interface
ANOVA	Analysis of Variance
API	Application Programming Interface
AT	Atomspheric Temperature
AQI	Air quality index
BP	Barometric Pressure
CPCB	Central pollution control board
DPCC	Delhi Pollution Control Committee
FB	Fractional Bias
FCC	False Colour Composite
IMD	India Meteorological Department
IME	Integrated Meteorological Emission
GIS	Geographic Information System
GPS	Global Positioning System
LUR	Land Use Regression
MAPE	Mean Absolute Probability Error
MSE	Mean Square Error
NIR	Near Infra Red
NMSE	Normalized Mean Square Error
NO_2	Nitrogen dioxide
OSM	Open Street Map
$PM_{2.5}$	Particulate matter less than 2.5 micrometer
RH	Relative Humidity
ROI	Region Of Interest
RMSE	Root Mean Square Error
SCP	Semi-automatic Classification Plugin
SR	Solar Radiance
SWIR	Short Wave Infra Red
TCC	True Colour Composite
WS	Wind Speed

Chapter 1

Introduction

1.1 General

Air pollution is the presence of non-natural or harmful substances which alters proportion of various gases in the atmosphere. These pollutants can be in any form of gases (like CO, NH_3, SO_2, NO_x, CH_4, CFC's), biological molecules or particulate matter which can be subdivided on basis of diameter of pollutants. These may be of organic or inorganic matter (Manisalidis et al., 2020). These pollutants can harm the health of human beings and other living beings.

In the United States of America, United States Environmental Protection Agency (USEPA) have set upper limits for 6 major contributors of air pollution in its National Ambient Air Quality Standards (NAAQS). These criteria pollutants are Sulfur Oxides (SO_x), Nitrogen Oxides (NO_x), Particulate matter (PM), Carbon Monoxide (CO), Lead (Pb) and Ozone (O_3) (USEPA, 2022). Burning of fossil fuels and biomass yields Sulfur Oxides (SO_x), Nitrogen Oxides (NO_x) and Carbon Monoxide (CO) whereas metal smelting plants, petroleum refining and industrial effluents are major source for Sulfur Oxides (SO_x) and Lead (Pb) pollution (WHO, 2021). Lead (Pb) is also waste generated from incineration plants, automobile exhausts and lead-acid batteries. These hydrocarbons and Nitrogen Oxides (NO_x) react in presence of sunlight to form Ozone (O_3) (WHO, 2021). Major source of Particulate Matter (PM) pollution is industrial combustion and vehicular emission. From Internal Combustion engines Carbon Monoxide (CO) is also generated (WHO, 2021). In India, the Air (Prevention and Control of Pollution) Act, 1981 gave government power and rights to set standards for air pollution. The control and enforcement for vehicular emissions also come under this act. A central agency

Central Pollution Control Board (CPCB) along with BIS, MoEF, MoPNG, MoRTH gave a value for industrial area, residential areas, rural and other areas and ecologically sensitive areas which are notified by the government in National Ambient Air Quality Standards providing upper limit for 12 air pollutants (Envis CPCB, 2016). These are listed in Table 1.1 with corresponding limits. Among all of these particulate matter is the most common pollutant.

Table 1.1: National Ambient Air Quality Standards (NAAQS) India – 2009 (Envis CPCB, 2016)

Pollutant	Time weighted average	Concentration in ambient air: Industrial, residential, rural and other areas	Concentration in ambient air: Ecologically sensitive area
Sulfur dioxide (SO_2)	Annual	50	20
	24 hr	80	80
Nitrogen dioxide (NO_x)	Annual	40	30
	24 hr	80	80
PM_{10}	Annual	60	60
	24 hr	100	100
$PM_{2.5}$	Annual	40	40
	24 hr	60	60
Ozone (O_3)	8 hr	100	100
	1 hr	180	180
Lead (Pb)	Annual	0.5	0.5
	24 hr	1	1
Carbon Monoxide (CO) milli	8 hr	2	2
	1 hr	4	4
Ammonia (NH_3)	Annual	100	100
	24 hr	400	400
Benzene	Annual	5	5
Benzo α pyrene nano	Annual	1	1
Arsenic nano	Annual	6	6
Nickel nano	Annual	20	20

All these pollutants cause harmful effects to human being to such an extent that it can lead to death, prior to which it causes many diseases and allergies (Velasco et al., 2019). These symptoms may be classified as long term or short term symptoms. Old people and children are highly susceptible to harmful effect of these pollutants (Manisalidis et al., 2020). Epidemiological and toxicological studies have shown that the air pollution primarily affect functioning of heart and lungs leading to adverse complications by entering through respiratory tract and accumulating in lung cells (Manisalidis et al., 2020). From Figure 1.1 it can be seen that for long term diseases ischaemic heart diseases and strokes lead to more than half of the deaths caused by air pollution followed by lung related problems which can finally lead to end of life (WHO, 2014). Long term air pollution exposure can also lead to change in count of total blood cells (Manisalidis et al., 2020). Prolonged exposure to these pollutants results in lung cancer and chronic obstructive pulmonary disease (COPD) which is a proven fact (Polednik and Piotrowicz, 2020). Along with these decline in cognitive ability is also seen along with neuro-degenerative diseases like dementia (Sinharay et al., 2018).

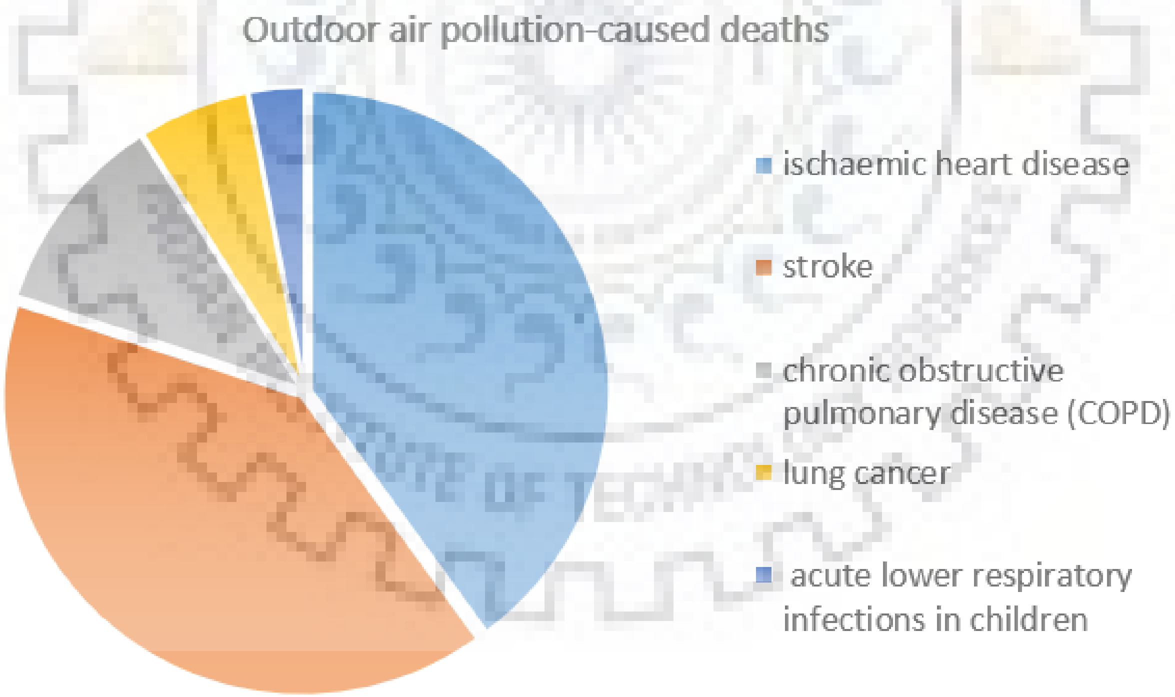


Figure 1.1: Outdoor air pollution caused deaths, source: (WHO, 2014)

While for short term effects problems related to eyes, nose and throat are most common. These may compile if exposed for long duration leading to bronchitis, pneumonia, asthma, and allergic reactions (Velasco et al., 2019). Smog, particulate matter, ozone, nitrogen dioxide and sulfur dioxide can all contribute to ear, nose and/or throat irritation. Eyes, nose or throat infections maybe results of either smog or particulate matters (Manisalidis et al., 2020). Smog is a mixture of fog with smoke leading to higher quantity of suspended small diameter particles. Due to smog the amount of particulate pollution increases manyfolds (EDF, 2017). Pollutants in the air cause oxidative stress, which harms the skin. Although human skin functions as a biological protection against pro-oxidative chemicals and physical air pollutants, repeated or protracted exposure to excessive amounts of these pollutants can have serious consequences for the skin (Brauer et al., 2015). Many of the short term effects caused due to air pollutants on the human skin may contribute to problems like skin aging, atopic dermatitis, skin cancer, psoriasis, and acne (Kathuria et al., 2017).

Not only living beings, it can cause damage to non-living agents like materials of structures or monuments and it also greatly affects on the environment. It is also the source for acid rain which damages crops, monuments and natural environment (Envis CPCB, 2016). Both natural activities as well as human activities can cause substantial damage in environmental ecosystem which may lead to air pollution. Some of these may directly be emitted by various sources into air while others may be formed by reaction between substances present in atmosphere. The former are known as primary pollutants while later is called secondary pollutants (Envis CPCB, 2016). Urbanization also plays an important role in determining type of pollutants (Wang et al., 2018).

Urbanization is a process composite of many changes started by population migration leading to economic development and change in land use (Wang et al., 2018). A general trend is also observed by scholars stating that the increase of population during urbanization is also causing a rise in serious air pollution concentrations (Larkin et al., 2016). Studies have been done to demonstrate correlation between air contamination and urban population density (Wang et al., 2020). Approximately a quarter of particulate pollution is emitted from the vehicles. With increase in vehicle ownership the amount of air pollution is ought to rise (Larkin et al., 2016). At present Indian capital and areas nearby gets affected by problem of air pollution every year. Indian cities, in general, are characterized by the high motor vehicle population as they are excessively dependent on their road network for freight and passenger traffic movements (Dutta and Jinsart, 2021). This problem of severe air pollution intensifies in the winter months. In the winters the dispersion rate of these pollutants changes due to which it takes more time to settle down (Tiwari et al., 2018). Other than these meteorological factors, land use also plays an important role in determining the dispersion of pollutants (Xu et al., 2021).

Till now most of the studies are towards finding the correlation between the land use and pollutants concentration (Briggs et al., 1997). Another subset of spatial distribution simulation focuses on finding the predictor variables based on GIS (Geographic Information System) techniques for different pollutants (Eeftens et al., 2012). With the increase in significance and accuracy of results obtained from these models, researchers have felt the land use to be among one of the major contributing factor in pollutant concentration detection (Yang et al., 2017; Shi et al., 2019; Xu et al., 2021).

The elementary principle of geo-spatial regression comes from idea of correlating any measured quantity like pollutant concentration at any location or water contamination level at a given location with the presence and intensity of influence of sources and sinks. These presence or intensity can act as explanatory variables due to which increase and decrease in the measured quantity can be seen (Mölter and Lindley, 2021).

The institutionalized method for Land Use Regression or the classical LUR approach is used by most of the air pollution studies done. It is based on the method developed for the European Study of Cohorts for Air Pollution Effects (ESCAPE) (Beelen et al., 2013). ESCAPE study was done utilising already done studies in approx 30 cohorts spanning pan-Europe. This was done to find correlation between long term air pollution and health effects (Jerrett et al., 2004).

Another study on Land use regression technique was also conducted in Europe under Small Area Variation in Air Pollution Health (SAVIAH) Project (Jerrett et al., 2004). In this study air pollution emissions from vehicular sources were taken into account for four European cities of Amsterdam (Netherlands), Huddersfield (U.K.), Prague (Czech Republic) and Poznan (Poland). In this study a methodology was developed based on the regression analysis and GIS tools were used for determining predictor variables. This study was done for four 2-week periods by taking oxides of nitrogen as marker variable for traffic related pollution. Later study by Briggs (2005) investigated the same for four U.K. cities of Huddersfield, Hammersmith and Ealing, Sheffield and Northampton. All these studies showed similar results as that found from other traffic pollution studies of the time (Brauer et al., 2003).

1.2 Need for study

Air pollution contributes considerably to morbidity and death especially in Asia. Roughly 60% of the world's population lives in Asia (Kaneda and Bietsch, 2015). And among them about 88% reside in low to middle-income nations with around 3 million worldwide excess fatalities as the result of exposure to external air pollution (Kumar et al., 2018). According to a report published by World Health Organization (WHO, 2018) in April 2018 encompassing 100 countries for a span of 5 years from 2011 to 2016, top 15 most polluted cities ranked by concentration of $PM_{2.5}$ from the various monitoring database, were situated in Asia with Delhi leading the list among the world's largest cities (Guttikunda et al., 2019). New Delhi, capital of India, is one of the world's ten largest metropolitan regions with a population estimated of 16.78 million in 2011 (Forstall et al., 2009; Apte et al., 2011). Although there are various fixed monitoring stations which record air quality data in real time but due to high spatial as well as temporal variability of pollutant concentration it is very difficult to get correct pollutant concentration data for everyplace. A proper air quality sampler with all the equipments is very difficult to place in a well maintained grid without incurring huge costs. Both installation and maintenance of such equipments would require a good amount of expense thus making the system non-feasible for the project to sustain for long duration.

This study can help people to understand air quality around them. The health of the entire community would consequently be benefited from reducing air pollution connected to transport. To tackle the problem of air pollution government has also taken various step to curb pollution causing activities. These include Graded Response Action Plan (GRAP) in 2017. GRAP were set of action plans for an emergency situation when air quality reaches very poor category and if air quality deteriorates to severe category strict rules like shutting down schools and implementing the odd-even road-space rationing scheme were exercised. It provided a step by step action plan for whole Delhi-National Capital Region (NCR) region by taking all different agencies responsible for crubbing air pollution on a single platform (Desk, 2020).

In general, choices about transit also affect physical activity. There is evidence that active transportation decreases the risk of cardiovascular, cancer and various other causes of death rather than driving an automobile (Mitsakou et al., 2021). Thus providing a greener route will benefit active commuters in their route choice. Also, assessment of real time pollutant concentration is essential for health effect studies and local policymaking (Manojkumar et al., 2021).

1.3 Objectives of the study

The aim of this study is to model an air pollution prediction system based on land use and traffic flow characteristic of any area. To carry out the study, this report is divided into following goals:

- To understand various pollution modelling techniques for practical applications.
- To develop a real-time congestion-based land use model for prediction of pollutant concentration.
- To perform a spatially varied regression analysis using Traffic API (Application Programming Interface) data.

1.4 Organization of Report

The report contains five chapters briefed as following: chapter 1 initiates with introduction of air pollution and land use regression, afterwards topics like the need and objectives of the study are highlighted. In chapter 2, the key findings from various literature studies is reviewed and the factors that are directly or indirectly affecting the modelling function are mentioned. Methodology adopted for analysis is discussed in chapter 3. Results from model analysis obtained are seen in chapter 4. In chapter 5, a general discussion and limitations are highlighted followed by future work which can be done in this field.

Chapter 2

Literature Review

2.1 General

Any air pollutant concentration predictive model is constructed on basis of many factor which can be meteorological in nature like temperature, pressure or relative humidity or based on demographics of the area like population and traffic characteristics. To understand the development of a spatio-temporal predictive models for pollutants taking into consideration real-time congestion patterns a methodological review was done. To identify articles related to this study keywords like - ("Land Use" OR "Air pollution" OR "India") were used in Google Scholar, Scopus and PubMed. Out of 527 research articles 52 were shortlisted based on abstract. These studies were used to understand land use modelling and regression modelling. In additions to these, references from these selected studies were used for further understanding. Only articles from last two decade (2001 - 2021) and in English were selected.

Urban $PM_{2.5}$ consists mostly of compounds of carbon, ions, and chemicals which are known to have come from traffic-related sources, industrial emissions, combustion of biomass and salt (Manojkumar et al., 2021). As $PM_{2.5}$ stays suspended in the atmosphere for a longer duration and travels a larger distance it have a greater impact on the lives of people. With a variety of toxic components $PM_{2.5}$ is the most prominent difficulty faced by developing country like India. This pollutant concentration can be found from two ways one by fixed monitoring while second by mobile monitoring. In fixed monitoring the outreach is very limited. To predict these pollutants' concentration at any position various different models are used like (a) proximity-based assessments (b) statistical interpolation (c) land use regression models (d) line dispersion models (e) integrated emission-meteorological models. These models are discussed in brief in next section.

2.2 Models used in past

Before going into detail on land use regression methods, this section deals with several other methods which can be used to model air pollution concentrations. There are several methods which can be used to model air pollution in larger areas, like 'interpolation of fixed-site government monitoring data, dispersion modelling, satellite remote sensing, land use regression (LUR), and proximity and deterministic methods'. The most important methods, in the context of this study, are discussed below to determine their characteristics and their strong and weak points. Fig. 2.1 shows the models that has been used in past to model air pollution concentrations.

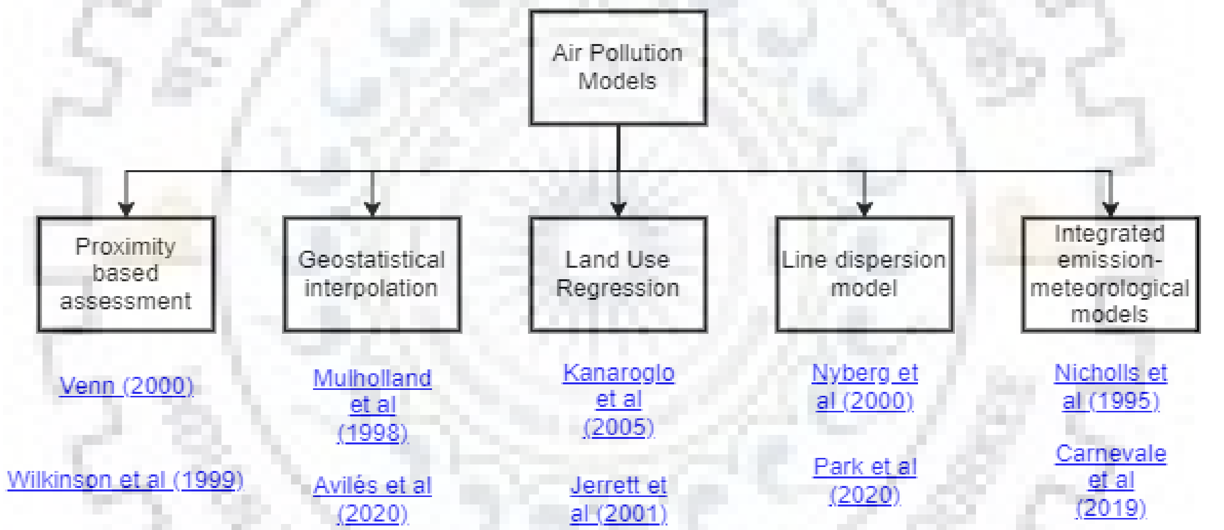


Figure 2.1: Different models used for Air pollution study

2.2.1 Proximity-based assessments

Measuring a subject's proximity to a pollution source is the most fundamental method for distinguishing intra-urban air pollution exposures. This strategy assists in the identification of correlations between atmospheric pollution and health outcomes based on the idea that proximity to emission sources may be used as proxies to indicate exposure in human populations. Venn (2000) used a traffic activity index. Traffic flows on roadways near a study schools were assessed as a continuous measure of traffic density for those 1-km^2 grid cells. Scholars found greater traffic counts or pollutants near the dwelling aggravate asthma symptoms (Vliet et al., 1997; Ciccone et al., 1998). But later studies statistically did not supported that pollution exposure affects asthma onsets (Wilkinson et al., 1999).

2.2.2 Geo-statistical interpolation

Geostatistical approaches based on both deterministic and stochastic methods, are used in interpolation models. The target pollutant is measured at a network of monitoring stations located across the study region. The goal is to obtain estimates of pollutant concentrations at locations other than monitoring station locations based on interpolation of results. “Kriging” is the most often used geostatistical approach in the field of air pollution (Jerrett et al., 2001). Kriging models use spatial dependency in data to create continuous pollution surfaces. The ability to provide both projected values and associated standard errors (kriging variance) in unsampled sites is a significant benefit. These standard errors measure the degree of uncertainty in spatial forecasts at unsampled locations, indicating where interpolation is less accurate (Mulholland et al., 1998).

2.2.3 LUR models

The robust strength of Land use regression lies in its empirical structure of the regression mapping. Because of this adaptation to new areas can be done without requirement of additional monitoring or data stations (Jerrett et al., 2004). LUR also assists in methods that identify areas requiring more intensive monitoring through the installation of additional stations (Kanaroglou et al., 2005). The main setback for this method comes from its area specificity (Jerrett et al., 2004).

2.2.4 Line dispersion models

Dispersion models are typically based on Gaussian plume equations (Bellander et al., 2001). They generate geographic exposure estimates of air pollution concentrations using assumptions about deterministic processes and data on emissions, meteorological conditions, and topography. In the research by Nyberg et al. (2000), when a number of factors were controlled, lung cancer was not substantially linked with simulated NO2. These models have the following drawbacks: (a) very expensive data input; (b) unrealistic assumptions regarding dispersion patterns (i.e., Gaussian dispersion); (c) extensive cross-validation required using monitoring data; and (d) temporal mismatches in data might potentially generate estimate biases (Jerrett et al., 2004).

2.2.5 Integrated emission-meteorological models

Meteorological and chemical modules are used together in integrated meteorological-emission (IME) models to study the dynamics of atmospheric pollutants (Nicholls et al., 1995). They have a lot of potential, especially in locations with a lot of people, where very little air pollution hazards may lead to significant and high secondary pollutant levels, which can cause a lot of disease and death (Jerrett et al., 2004). For installation and operation, IME models need high-end computing equipment, complex software, and highly skilled and experienced individuals leading to increase in cost of installation and operations.

2.3 Land use regression (LUR)

Land use describes how a part of land is being used by the humans. It represents the type of economic, social or environmental activity being practised at a particular position. Generally the land cover and use is used interchangeably but there is a difference between two as Land use indicates how people are using the land, whereas land cover indicates the physical land type. Land use does not describe the surface cover on the ground. Land cover refers to the surface cover on the ground like vegetation, urban infrastructure, water, bare soil or other. Land cover does not describe the use of land, and the use of land may be different for lands with the same cover type.

Land use affects how the pollutant will behave and disperse at different places due to which pollutant concentration may differ significantly. The spatial variability of pollutants have not been explained within city by fixed monitoring stations because of high dispersion rate and complex dispersion track because of land use (Sahsuvaroglu et al., 2006).

Measured pollution concentrations y at site s are used as the response variable, while land use types x within buffers surrounding location s are used as predictors of the measured concentrations as shown in Figure 2.2. The approach comprises using least-squares regression modelling to estimate pollutant surfaces based on pollution monitoring data and exogenous independent factors that already exist.

Figure 2.2: Land Use Regression *(Source : Jerrett et al. (2004))*

Several methods can explain such small-scale within-city variations such as geostatistical interpolation, dispersion models, and LUR models. Geostatistical interpolation of monitored concentrations is problematic whenever networks are not dense enough, and therefore fail to capture variability of concentrations over short distances. Dispersion models depend on detailed and spatially resolved input data if they are to capture small-scale spatial variations in air pollutants adequately. LUR modelling uses multiple linear regression to analyse associations between measured pollutant concentrations at a number of monitoring sites and predictor variables such as traffic, land use and topography. LUR models have been shown to be a cost-effective method to explain the spatial variation in air pollution in a number of studies (Beelen et al., 2013). Two of the major projects done using LUR methodology were European Study of Cohorts for Air Pollution Effects (ESCAPE) and Small Area Variation in Air Pollution Health (SAVIAH) Project both conducted in Europe.

2.3.1 ESCAPE project

The ESCAPE study was performed to evaluate the health implications of long-term exposure to polluted air. The investigation in the European environment was necessary since studies on the health impacts of air pollution exposure had previously been conducted primarily in North American region only (Beelen et al., 2013). Numerous air pollution concentrations were investigated as part of the ESCAPE study. Aside from particulate matter, the influence of nitrogenous oxides on health was also investigated, therefore NO_x and NO_2 were also considered in the project (Eeftens et al., 2012). Individuals enrolled in the ESCAPE experiment had their health data taken from previous cohort studies. Individuals' exposure to air pollution concentrations was assessed at their residences. A person's home location in a cohort explains a lot of the disparities in exposure across people (Cyrys et al., 2012).

The amounts of air pollution were measured in 36 places around Europe, although not all categories of air pollution were detected in all regions. Particulate matter and nitrogen oxides ($\mathrm{NO_2}$ and NO_x) were detected in 20 study sites, whereas just NO_x was recorded in 16 regions. Most of the time, the territories comprised of a significant city and its surrounds (Beelen et al., 2013). However, vast regions were also considered, such as in the Netherlands and Belgium, where the entire nation was simulated. Because the population of the research locations ranged from 100,000 to millions of people in big cities like London and Paris, there was a huge variance between them (Cyrys et al., 2012).

Beelen et al. (2013) discuss the data utilised in the ESCAPE study. The information is separated into two categories: central GIS data and local GIS data. The primary GIS data collection is made up of four datasets: a 1:10000 digital road network from Eurostreets, land use data from the CORINE land cover dataset, population density data on a 100m grid, and height data from SRTM 90m. Local datasets include a local digital road network combined with traffic intensity data, local land use data with more specific local land use types, population density data (which is not modelled, unlike the central GIS dataset on population density), altitude data (which was only used when local data was better than the central dataset), and local data specific to a specific study area (Eeftens et al., 2012).

2.3.2 SAVIAH project

In the SAVIAH project by Briggs et al. (1997); Lebret et al. (2000), road traffic volume was used for representing traffic conditions along with land use data and altitude as predictor variables. This study was done to develop a method for mapping NO_2 levels as an indicator for traffic-related pollution was a critical component of the study. A variety of approaches were used and compared as part of the study, including dispersion modelling (CALINE and CAR — exclusively in Huddersfield), spatial interpolation (contouring, kriging, and trend surface analysis), and regression mapping. Data on mean yearly NO_2 concentrations for a dense network of around 80 locations were gathered using passive diffusion tubes to aid in the development and calibration of these approaches (Lebret et al., 2000). 8 to 10 sites were used as reference in the study area for validation. They reported a good fit for mean annual concentration with R^2 values in the range of 0.79 to 0.87. Three important variables were employed as predictors in the regression model in the original SAVIAH study were traffic volume in the 300m buffer zone around each site, land cover in the 300m buffer zone, and surface height at the site. The way these were defined and the overall result of the multiple regression model were permitted to vary due to variances in data availability in the different research locations (Briggs, 2005). In summary, the study was a multicenter, EU-funded initiative that attempted to develop and evaluate methodologies for measuring the link between traffic-related air pollution and health at the small-area scale (Briggs et al., 2000).

In the study done by Brauer et al. (2003) in 3 European countries Netherlands, Germany, and Sweden, $PM_{2.5}$ filter absorbance and concentration of $PM_{2.5}$ matter was used as independent variables. These were predicted by using traffic volume as traffic characteristics and population as demographic variables. This study by Brauer et al. (2003) produced a good result with near similar trends for concentration of $PM_{2.5}$ obtained from this model. Half of the variability was explained by this study for Netherlands while for Germany and Sweden explained variability was around 60%. A similar study published earlier was carried out in Montreal, Quebec, Canada, in which the R^2 was 0.54 (Gilbert et al., 2005).

GIS softwares helps in process of linking data by providing means to combine geographic spatial information like pollution and population data. These will also help in modelling large scale data usually resulting in better accuracy (Briggs, 2005).

2.3.3 Advantages of LUR

From literature it is known that land use regression is less data demanding than from dispersion models (ROSS et al., 2007). Also as compared to proximity model higher number of explanatory factors can be included in land use study. Ryan and LeMasters (2007) affirmed the use of this model by using land use variables to differentiate exposure in a buffer area which can not be done in case of other models.

2.4 Different variables used in LUR

Predictor variables are derived from various factors like land use types, traffic characteristics and geography and meteorology of the area. This information is used by geographic information systems (GIS) to create integrated environment to work taking into account all the parameters based on spatial distribution (Sahsuvaroglu et al., 2006). In most of the studies common traffic parameters included were length of major roads in buffer, length of all roads in buffer and distance to nearest road or major road, while for land use parameters residential and commercial area, industrial area, green space and water area are taken in account (Meng et al., 2015; Lee et al., 2014). Similarly, from geographic point of view altitude and height of building in nearby location was seen and variables like temperature, wind speed, pressure and solar radiation is taken into account. For each different variable different buffer area is chosen as well as a direction is determined. Based on the studies one can say that possible direction of variables suggesting increase in traffic will have a positive direction like length and density of the roads in the buffer area, while increase in the variables like the distance of monitoring site with road results in negative direction (Wang and Ogawa, 2015). For the land use variables, increase in density of green area including forest area and park area will result in decrease of pollutant concentration thus the assumed direction tends to be negative while more commercial and residential area will result in higher concentration leading to a positive direction for the predictor variables (Lee et al., 2014; Meng et al., 2015). Land use variables are the possible predictors for the spatial data while for the temporal nature of pollutants meteorological data as well as traffic data can be used. Meteorological factors which heavily influence $PM_{2.5}$ concentration are wind speed and temperature. Strong variations are also observed along with change in seasons. Various studies suggested negative correlation between $PM_{2.5}$ and wind speed lower than 3m/s and positive correlation between $PM_{2.5}$ and wind speed higher than 3m/s (Wang and Ogawa, 2015; Li et al., 2017). This is due to the fact that wind can act as both dispersing force as well as transporting agent. When the wind speed is low, it can help in dispersing away the pollutants within a certain geographical range thus reducing pollution concentration but, when the wind speed is higher than 3m/s it becomes strong enough that it can transport large quantities of pollutants from far away thus increasing pollutant concentration. In relation to temperature, $PM_{2.5}$ has a strong positive correlation with temperature (Meng et al., 2015). This is due to the fact that temperature facilitates the formation of particles by providing them required energy thus, at higher temperature the photo-chemical reaction can take place which lead to formation of $PM_{2.5}$ thus increasing its concentration in the atmosphere.

Chapter 3

Methodology

3.1 Overview

The main idea behind this study is to develop a spatio-temporal model for development of a predictive technique for concentration of various pollutants. As discussed in the literature review, land use regression provides good results in predicting the pollutants' concentration. It also helps in incorporating the spatial data and by differentiating among buffer areas. For land use general predictor variables are line densities for different types and densities of roads, traffic volumes for private as well as public modes, types of land use and meteorological factors like, temperature, pressure, and relative humidity. The detailed methodology of the study is shown in Fig. 3.1.

The type of land use variables helps in this study to provide spatial variability in the model while the traffic flow conditions provides temporal variability in the study. In Land Use regression an iterative process is used for removal of variables to the model as shown in Figure 3.2. The first step in the modelling process starts with selecting a significance level for a variable to stay in the model. Generally this significance level is around 5%. Model is started by entering all the possible predictor variables. Now the predictor variable with least significance or highest p-value is seen. If the p-value of predictor variable is greater than the significance level then we say that the null hypothesis is accepted. The null hypothesis says that the coefficient of variable is not significantly different from 0. If the variable with highest p-value is less than the significance level then the iterative process is stopped and all the remaining variables are included in the model.

For this study, Quantum Geographic Information System (QGIS) is used to visualize and provide tools for spatial analysis by providing a way of capturing and linking spatial

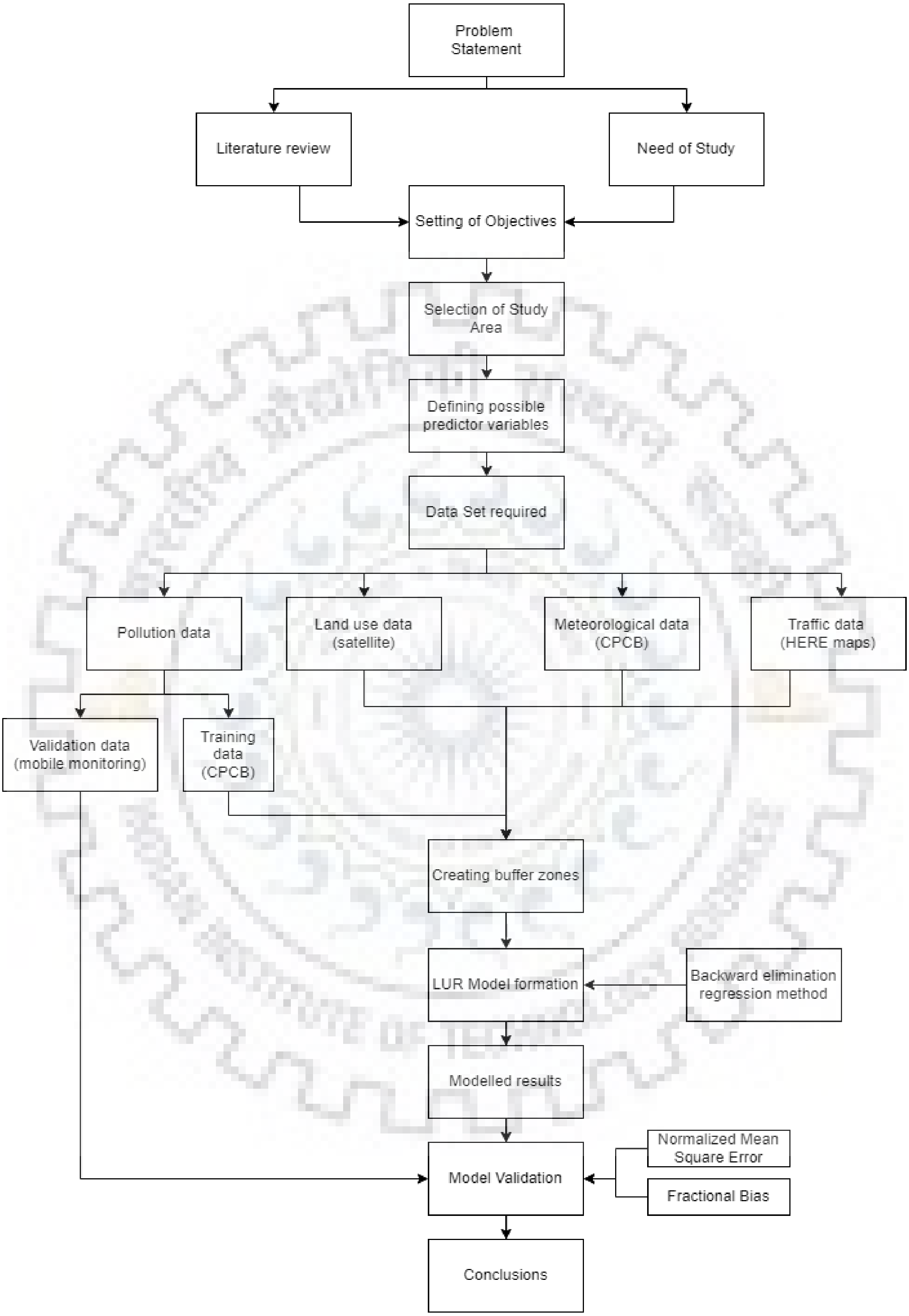


Figure 3.1: Methodology

data within a single geographical structure (Briggs, 2005). Python was used to handle the large amount of data generated while SPSS was uses to analyse the data.

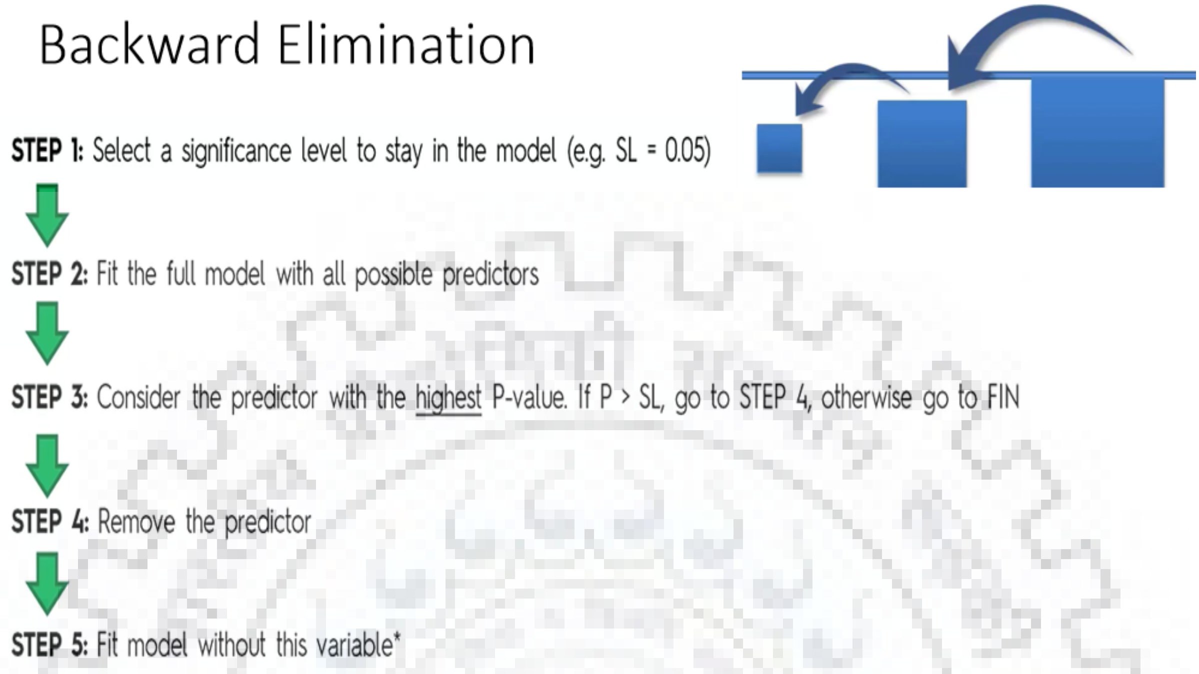


Figure 3.2: Land Use Regression Methodology *Source: Soni (2020)*

3.2 Data

Most of the studies use vector analysis while this work try to perform and obtain result from raster analysis. Raster analysis provides better and intuitive map algebra making it faster and easy to use and update. Raster grid formatting is data model for satellite data and other remote sensing data. The image in raster format is stored in form of matrix consisting of pixels. Each pixel have different value based on the land use of that particular pixel as shown in Figure 3.3. If the pixel is road pavement, traffic characteristics also becomes functional. The count of each type of pixel along their weights can be computed inside any buffer area around the monitoring station. Model can developed based on the linear regression technique with each category of land use as different independent or explanatory variable.

The vehicular emissions are difficult to obtain to higher accuracy levels but with the knowledge of spatial geometry of the area and vehicular flow characteristics a proxy can be created for vehicular emissions. For the dependent variable a training set is required

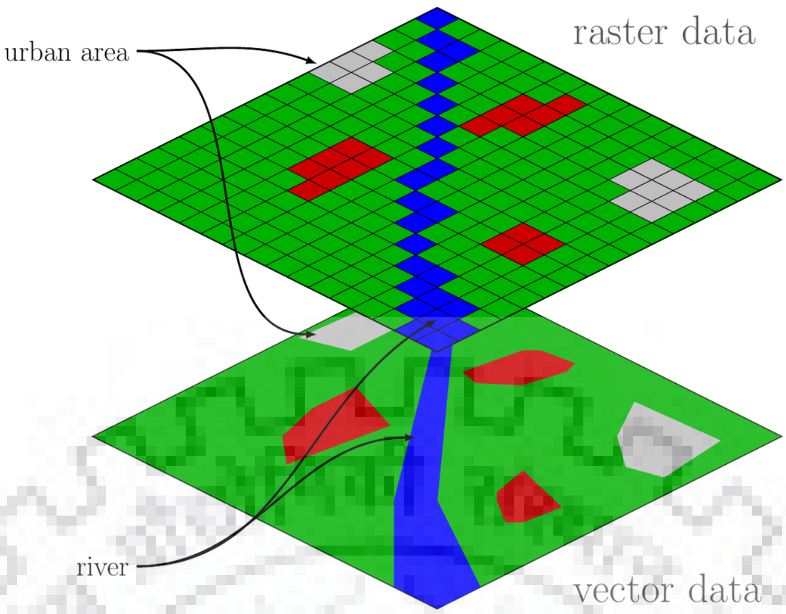


Figure 3.3: Vector to raster image *(Source: eo4geocourses (2018))*

to perform the regression analysis. Therefore mainly 3 types of data are required for the analysis which are :

- Pollution data
- Traffic data
- Land Use data

3.2.1 Pollution data

The pollution data is required as it provides the basic framework for independent variable for the prediction model. It is obtained from the fixed monitoring stations located at various locations in the study region. For this work, a total of 36 monitoring stations are available in National Capital Territory (NCT) of Delhi as shown in Figure 3.4. Out of these there are 24 sites operated by Delhi Pollution Control Committee (DPCC) while 6 sites are overseen by CPCB and rest 6 sites are maintained by India Meteorological Department (IMD). The real time data is available at Central Control Room for Air Quality Management[1].

Concentrations of major pollutants like CO, NH_3, SO_2, NO_x, are available along

[1]https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing

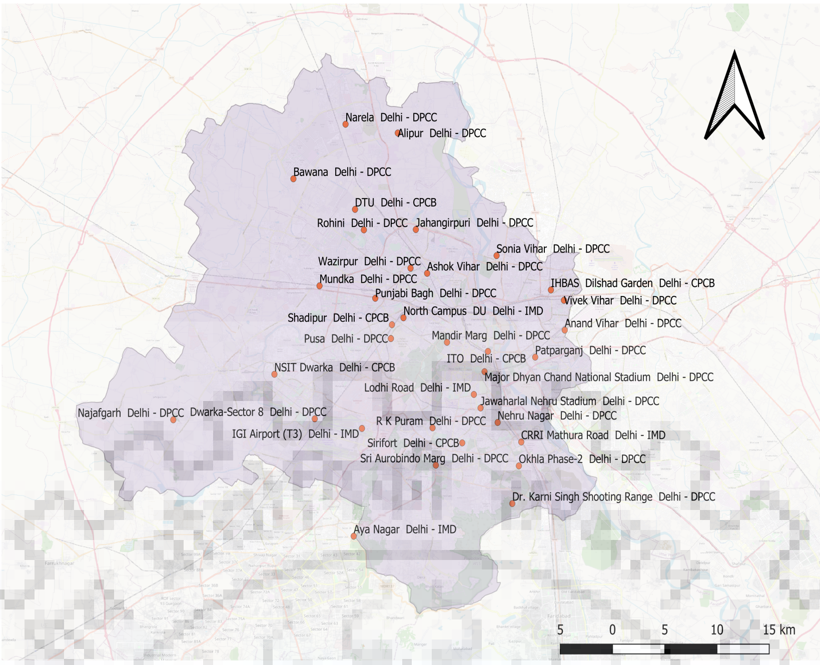


Figure 3.4: Location of different monitoring sites in Delhi

with meteorological data like temperature, wind direction and speed, barometric pressure and relative humidity. For this study information about concentration of $PM_{2.5}$ was collected along with the meteorological data like relative humidity, wind speed, pressure, temperature and solar radiance for 2 week period at 4 hour time interval. This model can be used to predict any pollutant if its historical data is available. Pollution data can be obtained as daily mean as well as hourly mean based on the availability of the traffic data.

3.2.2 Traffic data

The traffic data can be approximated by the hierarchy of the road i.e. arterial, sub-arterial or local road and the congestion on the road i.e. high, medium or low. This flow data on different roads is obtained from HERE Maps Traffic Flow API as shown in Figure 3.5. These two factors can specify about the pollution emission from that region as areas with congestion are major hot-spots for particulate matter pollutants. As the type of road goes up the hierarchy the amount of pollution emission will also rise as more and more number of vehicles would ply on the major roads with more heavy vehicles thus leading to increased pollution emissions. Based on the flow and category of road, a weight can be associated to the road in buffer area. The traffic characteristics of the buffer area

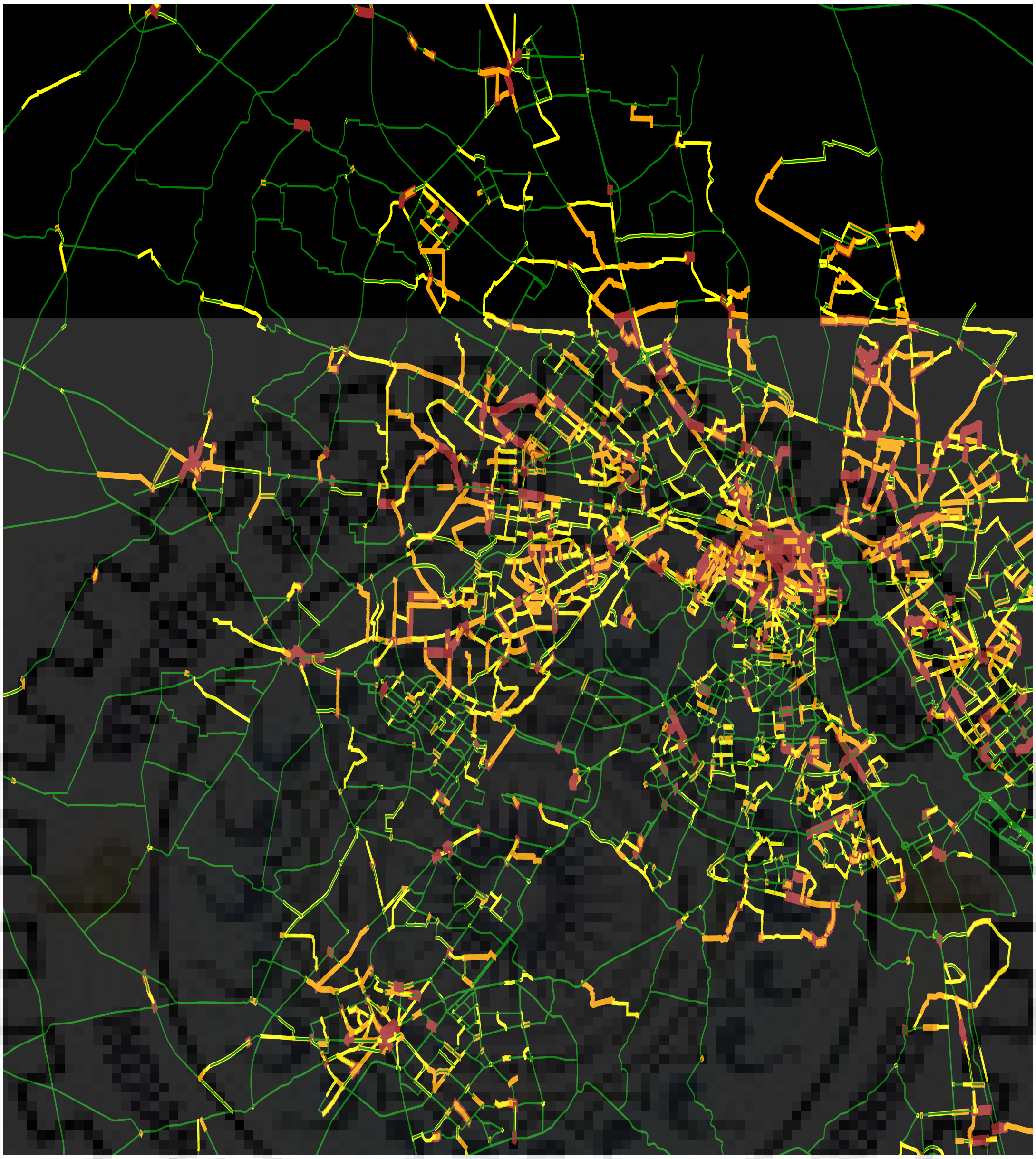


Figure 3.5: Speed Profile of different roads in Delhi

provides temporal variability in the model. If the monitoring site for the pollution is fixed like in this study, real-time data can be useful in predicting the instantaneous pollutant concentration data. It can be very useful for the active commuters to choose their route based on the best option available.

For this study a ratio of average speed to free flow speed was used to categorise the road. The values of this will range from 0 to 1 with 0 being nearly jam condition to 1 being free flow condition (Tsuboi and Yoshikawa, 2020). All the roads were classified like “road_0” to “road_5” as shown in Table 3.1. Traffic conditions were recorded for every 2-hr time bin for the same 2 week periods as pollution data.

3.2.3 Land Use data

The geography and geometry of nearby area affects heavily on dispersion of pollutants in the atmosphere. Thus, it becomes an important factor to consider while predicting air pollution. The area with high density of vegetation and greenery tends to give more surface area for particles to settle and also helps to maintain the natural atmospheric composition of gases. Whereas, the urban area be it commercial or residential tends to increase the amount of air pollution in the nearby area. It is also seen that over the river surface or in general any water body the pollutants concentration is relatively low. Land use data provides the much needed spatial variability in the predicted model. For a better model a higher number of training points are needed for development, thus for a better spatially varied model a higher number of monitoring stations will be required. All the classes of the land use can be used as a possible explanatory variable in the model. The land use data can be obtained directly as services provided by the government agencies like ISRO or one can classify the data from the satellite data available.

3.3 Processing Data

This section gives a brief on the process to obtain the data and the way to process it so it can be used for the analysis. The data obtained from different sources may be in raw form and would require pre-processing or it may be used directly.

3.3.1 Pollution Data

The pollution data is included in the meteorological data obtained. Particulate matter's interaction with other meteorological factors can be seen in Figure 3.6. All of these parameters were obtained from the Central Pollution Control Board's live monitoring stations at 4 hour interval. The data was obtained as xlsx file. The following data was recorded :

- Relative humidity (RH) : Relative humidity is a percentage ratio of the quantity of atmospheric moisture present to the quantity that would be contained if the air was saturated. Because the latter quantity is temperature dependent, relative humidity

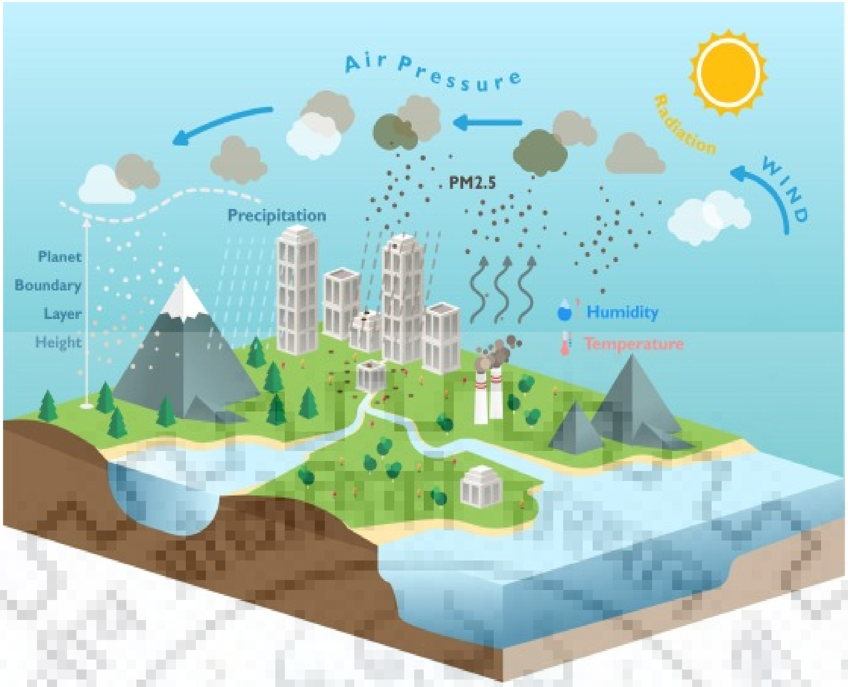


Figure 3.6: Interaction of particulate matter with meteorological factors (*Source:* *Chen et al. (2020)*)

is determined by both moisture content and temperature. The related Temperature and Dew Point for the stated hour are used to calculate the relative humidity.

- Wind Speed (WS) : Wind speed is a basic atmospheric quantity created by air flowing from high pressure to low pressure, mainly due to temperature variations. Wind speed shows a complex trend with particulate pollution.

- Solar Radiation (SR) : The term "solar radiation" refers to the energy emitted by the Sun, a part of which reaches the Ground. This is the principal source of energy for the majority of activities in the atmosphere, oceans, and biosphere.

- Barometric Pressure (BP) : The force or weight of the air at any point is measured by barometric pressure, often known as atmospheric pressure or air pressure.

- Atmospheric Temperature (AT) : The term "atmospheric temperature" refers to the temperature of the Earth's atmosphere at various altitudes. Many things influence it, including incoming sun energy, humidity, and altitude. Temperature fluctuates substantially at various heights relative to the Earth's surface, and this temperature variance distinguishes the four layers of the atmosphere.

- Particulate Matter (PM) : Particulate matter is the aggregate of all solid and liquid particles floating in air, many of which are dangerous. Dust, pollen, soot, smoke, and liquid droplets are among the organic and inorganic particles in this complex combination. The size, nature, and origin of these particles vary widely.

3.3.2 Traffic Data

The traffic data required for this study was obtained from HERE Maps API. The data was obtained in xml file format at every 2-hr time interval for the same 2 weeks period. The data obtained include speed and congestion for the region of interest. Region of interest can be defined by either providing a bounding box pr using proximity or as corridor. The bounding box is limited to maximum of 2 degrees of longitude and latitude. To obtain the data a request is generated and for each request contains information on what data is required. For this study the following request was generated :
https://traffic.ls.hereapi.com/traffic/6.2/flow.xml?apiKey='+HERE_API_KEY+'&bbox=28.39157,%2076.85863;28.87727,%2077.30729&responseattributes=sh,fc&units=metric

This request can be subdivided as shown in the Figure 3.7. The response from the request is obtained in xml format which was converted to ordered dictionary in python and the objects from the results are extracted.

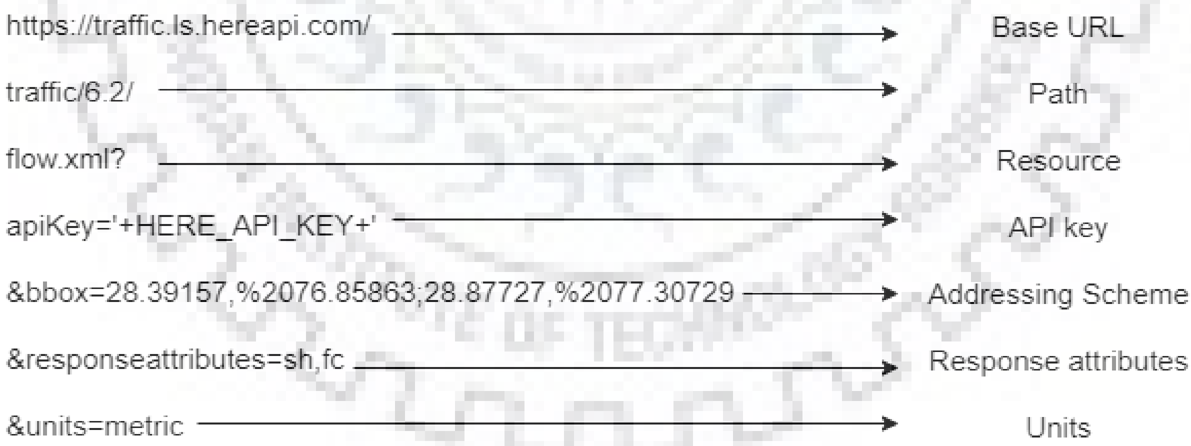


Figure 3.7: Request for Traffic API data

The response to the request contains the following data as shown in fig Figure 3.8:

- "RWS" : It represents a list of roadway items (RW)
- "RW" : This is the overall flow item for a roadway. Each roadway will have a roadway item with accessible traffic flow information

- "FIS" : It represents a list of flow item elements (FI)
- "FI" : FI represents a single flow item
- "TMC" : TMC stands for "Traffic message channel". Information about location in coded format can be sent and received via TMC if the location code table is integrated with the maps service provided.
- "PC" : TMC location code for any point
- "DE" : It represents the text description of the road
- "QD" : QD stands for queuing direction which can be "+" or "-"


```xml
<?xml version='1.0' encoding='UTF-8'?>
<TRAFFICML_REALTIME CREATED_TIMESTAMP="2022-01-18T22:18:56Z" MAP_VERSION="" UNITS="metric" VERSION="3.2" xmlns="http://traffic.nokia.com/trafficml-flow-3.2">
    <RWS EBU_COUNTRY_CODE="5" EXTENDED_COUNTRY_CODE="F2" MAP_VERSION="202104" TABLE_ID="4" TY="TMC" UNITS="metric">
        <RW DE="100 Feet Road" LI="504-00200" PBT="2022-01-18T22:18:56Z" mid="1d3bbd55-b130-47ee-ac04-48eeb36990cf">
            <FIS>
                <FI>
                    <TMC DE="Loni  " LE="0.0335" PC="202" QD="+"/>
                    <SHP FC="4">28.74354,77.28788 28.74352,77.28776 </SHP>
                    <SHP FC="4">28.74352,77.28776 28.7435,77.2876 </SHP>
                    <CF CN="0.71" FF="10.0" JF="0.0" SP="10.0" SU="16.89" TY="TR"/>
                </FI>
                <FI>
                    <TMC DE="Ghaziabad" LE="0.5064" PC="31408" QD="+"/>
                    <SHP FC="4">28.7435,77.2876 28.7434,77.2868 </SHP>
                    <SHP FC="4">28.7434,77.2868 28.74335,77.28642 </SHP>
                    <SHP FC="4">28.74335,77.28642 28.74332,77.28621 </SHP>
                    <SHP FC="4">28.74332,77.28621 28.74328,77.28595 </SHP>
                    <SHP FC="4">28.74328,77.28595 28.74325,77.28577 </SHP>
                    <SHP FC="4">28.74325,77.28577 28.7432,77.28546 </SHP>
                    <SHP FC="4">28.7432,77.28546 28.74317,77.28531 </SHP>
                    <SHP FC="4">28.74317,77.28531 28.74312,77.28507 </SHP>
                    <SHP FC="4">28.74312,77.28507 28.74307,77.28484 </SHP>
                    <SHP FC="4">28.74307,77.28484 28.74302,77.28463 </SHP>
                    <SHP FC="4">28.74302,77.28463 28.74299,77.28448 28.74294,77.2843 </SHP>
                    <SHP FC="4">28.74294,77.2843 28.74288,77.28411 28.74281,77.28392 </SHP>
                    <SHP FC="4">28.74281,77.28392 28.74255,77.28335 </SHP>
                    <SHP FC="4">28.74255,77.28335 28.74232,77.28291 </SHP>
                    <SHP FC="4">28.74232,77.28291 28.74228,77.2828 28.74225,77.28266 </SHP>
                    <CF CN="0.81" FF="20.0" JF="0.37469" SP="19.44" SU="19.44" TY="TR"/>
                </FI>
                <FI>
                    <TMC DE="Lalbagh Colony Road (North)" LE="0.23636" PC="31407" QD="+"/>
                    <SHP FC="4">28.74225,77.28266 28.74202,77.28251 </SHP>
                    <SHP FC="4">28.74202,77.28251 28.74165,77.2823 28.74127,77.28207 </SHP>
                    <SHP FC="4">28.74127,77.28207 28.741,77.28189 </SHP>
                    <SHP FC="4">28.741,77.28189 28.74069,77.28164 </SHP>
                    <SHP FC="4">28.74069,77.28164 28.74042,77.28141 </SHP>
                    <CF CN="0.76" FF="20.0" JF="1.07628" SP="18.39" SU="18.39" TY="TR"/>
                </FI>
                <FI>
                    <TMC DE="Lalbagh Colony Road (South)" LE="0.64559" PC="31406" QD="+"/>
                    <SHP FC="4">28.74042,77.28141 28.73999,77.28105 </SHP>
                    <SHP FC="4">28.73999,77.28105 28.73928,77.28053 </SHP>
                    <SHP FC="4">28.73928,77.28053 28.73911,77.28039 </SHP>
                    <SHP FC="4">28.73911,77.28039 28.73862,77.27999 </SHP>
                    <SHP FC="4">28.73862,77.27999 28.73816,77.27961 </SHP>
                    <SHP FC="4">28.73816,77.27961 28.73779,77.27929 </SHP>
                    <SHP FC="4">28.73779,77.27929 28.73692,77.27844 </SHP>
                    <SHP FC="4">28.73692,77.27844 28.73655,77.2781 </SHP>
                    <SHP FC="4">28.73655,77.2781 28.73576,77.27744 </SHP>
                    <CF CN="0.71" FF="19.0" JF="0.16762" SP="18.76" SU="18.76" TY="TR"/>
                </FI>
```


Figure 3.8: XML file from Traffic API

- "LE" : It is the road segment length in units specified in request url
- "SHP" : SP represents the road shape item containing information about geometry of given road segment
- "FC" : FC represents the functional class of the road ranging from 1 to 5 with 1 being lower road while 5 being higher road in hierarchy of roads
- "CF" :CF stands for Current Flow containing details about speed and Jam Factor information for the given flow item.

- "CN" : CN stands for Confidence Number indicating percentage of real time data used. Data is said to be in real time if the CN value is greater than 0.7
- "FF" : FF represents the free flow speed on the given stretch of the road
- "JF" : JF stands for Jam Factor which represents the quality of travel. JF value ranges from 0 to 10 with 10 being a condition of road closure. As the number increases the quality of the travel will degrade
- "SP" : SP denotes the average speed in units specified in url and capped by speed limit
- "SU" : SU stands for Speed Uncut which also represents average speed in units specified in url but this is not capped by speed limit.

The following xml document was converted to ordered dictionary and each object was retrieved and a dataframe was created specifying all the linestring coordinates and their corresponding attributes. The factor of ratio of average speed to free flow speed was calculated. Based on this factor all the line segments were divided into 6 categories viz. "road_0", "road_1", "road_2", "road_3", "road_4" and "road_5" as shown in Table 3.1. Thus a total of 168 created dataframe were saved in csv format.

Table 3.1: Road factor category corresponding to ratio of average speed to free flow speed

Ratio of average speed with free flow speed	road factor category
<0.05	road_0
0.05-0.275	road_1
0.275-0.5	road_2
0.5-0.725	road_3
0.725-0.95	road_4
>0.95	road_5

All the files were loaded in QGIS and data was clipped based on the buffer area as shown in Figure 3.9. Then the sum of line length in all the buffer zones were calculated and files were saved in csv format with each file name depicting the time and road factor category and each file contain ringID and location with their respective sum of road network.

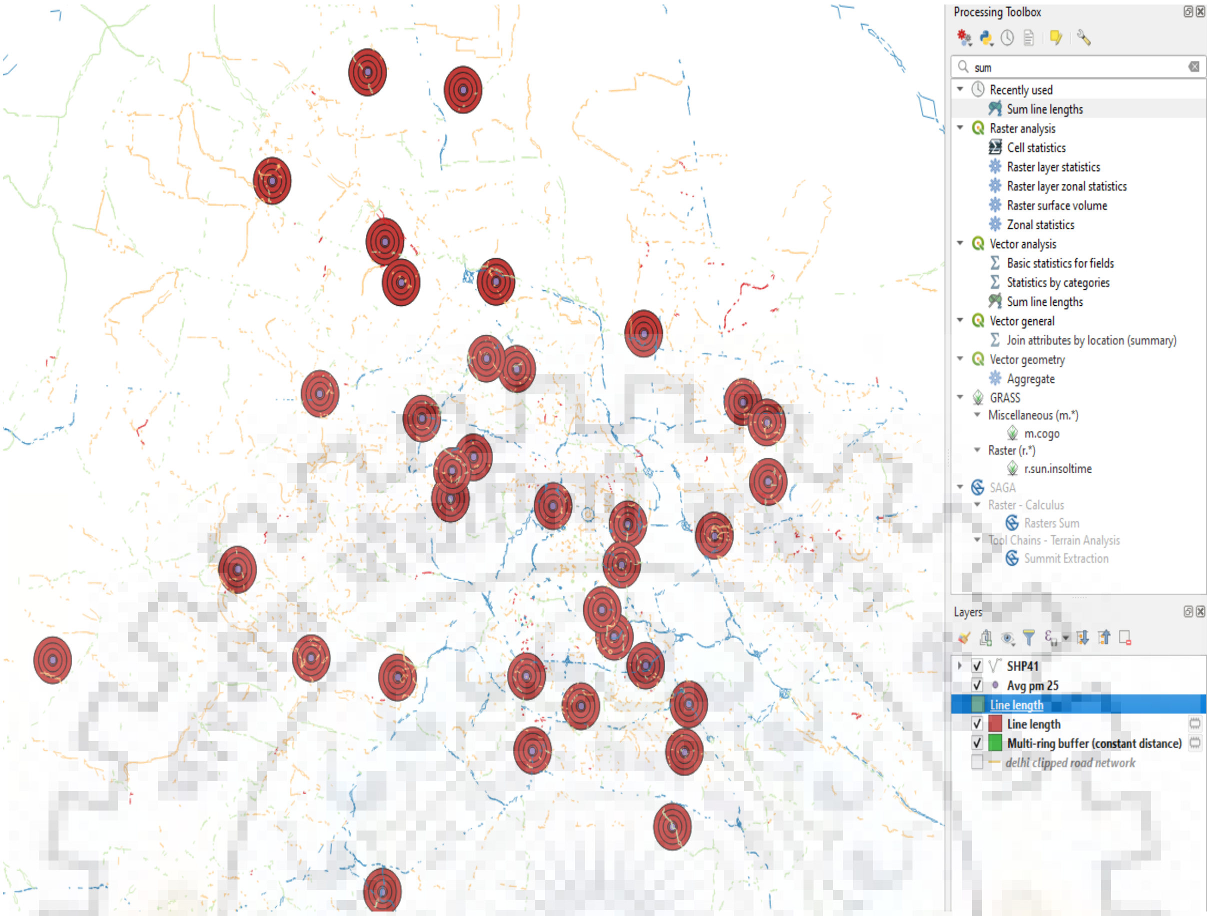


Figure 3.9: Road shapefile and monitoring stations

3.3.3 Land Use Data

For this study the data for Land use was obtained from United States Geological Survey (USGS)[2]. Data from Landsat 8 satellite was used to classify the study area into various categories and sub-categories. The classification was done using Semi-automatic Classification Plugin (SCP) in QGIS. The Semi-Automatic Classification Plugin (SCP) facilitates the supervised learning for the images obtained from remote sensing. Direct images from ASTER, MODIS, GOES, Landsat, Sentinel-1, Sentinel-2, and Sentinel-3 can be obtained using this. SCP provides necessary tools for the pre-processing as well as post-processing of the satellite images.

The images from satellite is obtained in form of tiles or different bands which can is clipped according to study area. For the Landsat 8 there are 6 bands namely Band 2, Band 3, Band 4, Band 5, Band 6 and Band 7 for classification process corresponding

[2]https://earthexplorer.usgs.gov/

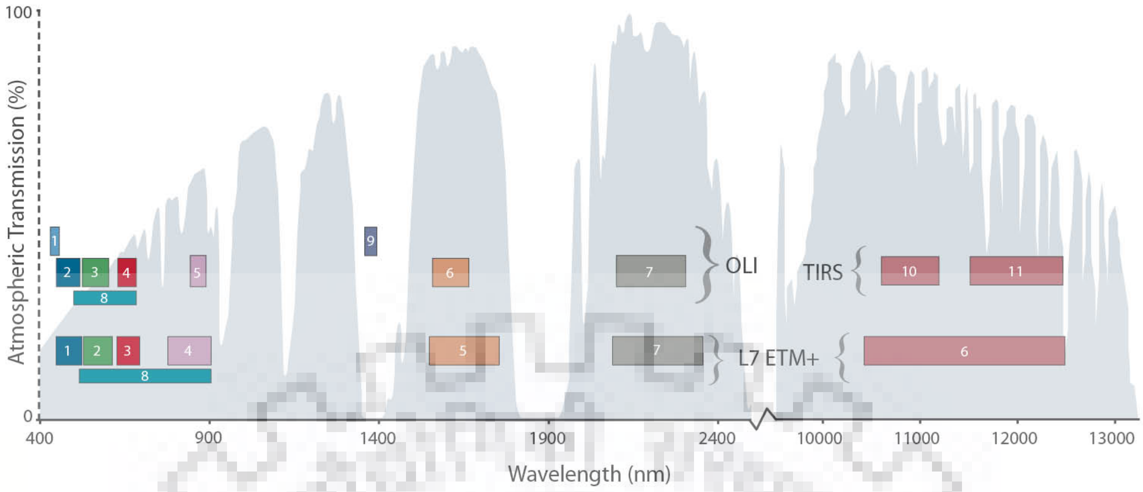


Figure 3.10: Bandset of Landsat 8 (*Source : Brown (2017)*)

to spectral reflectance detected by different sensors on the satellite for Blue, Green, Red, Near Infra Red and Short Wave Infra Red (SWIR) 1 and 2 from Table 3.2. The resolution for these bands is 30m as shown in Figure 3.10. After selecting the bands the metadata information is used to convert the bands into surface reflectance by using the central wavelength for each band-set and additive and multiplicative factor defined for each image.

Table 3.2: Bandset Values of Landsat 8

Landsat 8 OLI and TRIS Bands (μm)			
30m	Coastal/Aerosol	0.435-0.451	Band 1
30m	Blue	0.452-0.512	Band 2
30m	Green	0.533-0.590	Band 3
30m	Red	0.636-0.673	Band 4
30m	NIR	0.851-0.879	Band 5
30m	SWIR1	1.566-1.651	Band 6
100m	TIR1	10.60-11.19	Band 10
100m	TIR2	11.50-12.51	Band 11
30m	SWIR2	2.107-2.294	Band 7
15m	Pan	0.503-0.676	Band 8
30m	Cirrus	1.363-1.384	Band 9

The combination of these bands give different colour composites like True Colour Composite (TCC) which is obtained from using Blue-Green-Red and False Colour Composite (FCC) which is obtained from using Green-Red-NIR bands. Each different colour composite helps in differentiating different objects based on their reflectance properties for e.g. Vegetation is best shown in FCC as dark red areas as shown in Figure 3.11.

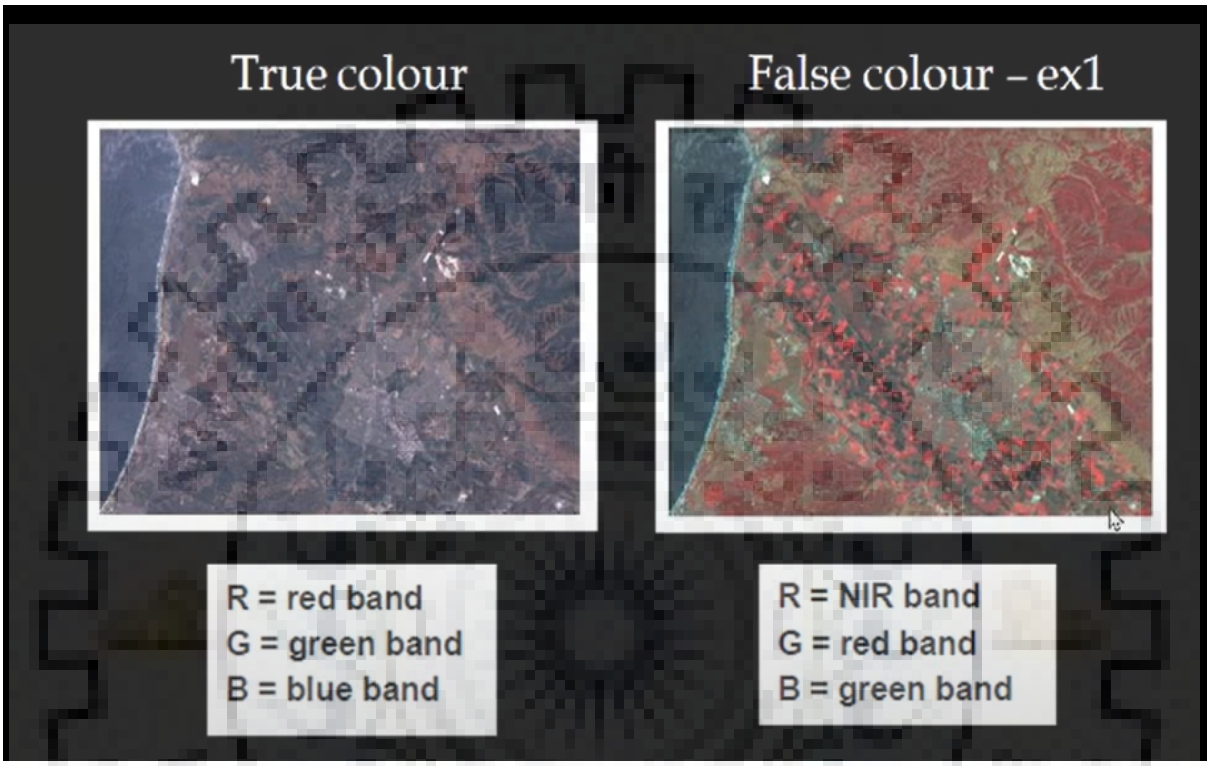


Figure 3.11: Colour Composites (*Source:* *GeoSage (2016)*)

For the classification process first aim is to define the land cover classes followed by creation of ROIs and process of supervised training. Following this, spectral signature is assessed based on which a go ahead is given for the classification. This classification is further refined and an accuracy assessment is done. A training input file is defined and created where the ROIs are created and stored for training the data.

QGIS was used to create a buffer area around the fixed monitoring stations. The whole study region was classified using Semi-automatic Classification Plugin (SCP) into various class and sub classes as mentioned:

- Water bodies - (a) river (b) drain
- Vegetation - (a) forest (b) park
- Built up area - (a) roads (b) builtup

Further the delimited file with details about location of monitoring stations were loaded to QGIS. It will create a layer with point data with each point representing a monitoring station. Then a multi-ring buffer area was created around each monitoring station comprising 4 rings with distance between each concentric ring to be 250m. All the rings were allotted a number from 1 to 4 with innermost ring ID to be 1. Further the raster pixels in each buffer area was converted into a point and the resulting table is used for further analysis. No. of points in the buffer are calculated which comes around 3670 points at an average in each buffer.

3.4 Data Merging

The data obtained from all the 3 data set was merged on the basis of location, time and ringID. Data from traffic flow API included the sum of line length in each buffer zone and location categorised on basis of ringID while the date and time was included in file name. For the meteorological data, all the rings had same value for RH, WS, SR, BP and AT. all these values were obtained on each monitoring station and for the case of land-use data the densities for each type of land-use was calculated as sum of number of points in any given buffer zone. Thus the final data obtained was analysed using correlation analysis and linear regression.

3.5 Validation Data

The data validation was done in 2 ways one being validation from 10% of the dataset and in other way pollutant concentration was predicted using data collected from the field. The data for validation was obtained at 2 different location in Delhi. 10 instances were chosen. Pollution data along with temperature and humidity were recorded using mobile monitors. Buffer was created on the 2 locations and land use was recorded. Other meteorological parameters were collected from nearest monitoring station for that duration. Pollutant concentration was predicted for that time and compared with the obtained data. 2 performance measure functions viz. Normalized Mean Square Error (NMSE) and Fractional Bias (FB) were used for comparing the predicted values obtained by using the result from model and the obtained concentration from the field.

Chapter 4

Results

4.1 Analysis

The analysis is done on three parts firstly to check the quality of classification of land use, secondly as descriptive statistics of explanatory or independent variables and finally to assess the goodness of fit of the model.

4.1.1 Spectral signatures

While classification any pixel value which lies inside the variance limit of the ROIs can be classified otherwise it would remain unclassified as shown in Figure 4.1. On the other hand, increasing the variance limit can result in mis-classification or wrong classification for near equal spectral signatures. For this study the value was chosen to be mean $\pm\ \sigma$. Mean corresponds to mean value of all the ROIs in the given class. In spectral signature it is denoted by dark solid line as shown in Figure 4.1 In the overlapping area, the

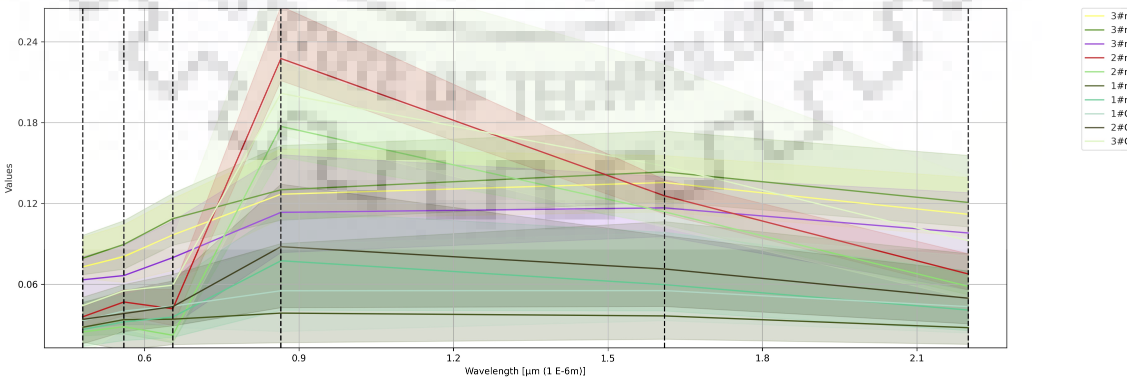


Figure 4.1: Spectral Signatures

classification can be little misguided but the training set with maximum possible coverage across wavelengths give the best results. From the image it is clear the plants show high value of spectral reflectance in near infrared range of spectrum. Due to excessive overlapping in signatures the differentiation among the two becomes very difficult.

Effect from both leaf pigments and physiological structure give plant leaves a characteristic reflectance signature which comprises of low reflectance for red light as well as blue coloured light. Reflectance of green coloured light is in medium range while the highest reflectance is of near infrared region. Thus a sharp peak is observed in vegetation area in near infrared region.

In same way the depth of water, suspended materials in water and roughness of water surface will determine the spectral signature of water body. A major part from incident radiation on water surface is either absorbed or transmitted and only a small part is reflected back, thus very low values are obtained in spectral signatures.

4.1.2 Descriptive statistics

From the Table 4.1 it can be seen there are four sets of variables, first dependent variable of pollutant concentration. In this study for $PM_{2.5}$ is taken as a marker for pollution levels. Secondly, the traffic flow variables are taken into account. These include the variables namely road_1, road_2, road_3, road_4 and road_5, but due to no valid data in road_0 that variable is neglected. These variables represents the effect of different roads categorised based on ratio of average speed and free flow speed. Thirdly, the meteorological variables are used namely Relative Humidity (RH), Solar Radiance (SR), Barometric Pressure (BP), Atmospheric Temperature (AT) and Wind Speed (WS). Finally, the land use variables consisting of density of identified near road points count, vegetation points count , built up points count, bare soil points count and water points count are considered. Any point count will represent the density of that land use as number of points for any land use in any buffer will correspond to density.

It can be seen that $PM_{2.5}$ ranges from near about 17 to 384 μgm i.e. it covers a wide range of 367μgm. Mean pollutant concentration was 106μgm making the air quality to lie in "moderate" bracket according to Indian system while the mode stands at "Satisfactory". Standard deviation can be used to see the deviation from the mean value. Highest deviation can be observed in the value of road category 4 as the difference in sum

Table 4.1: Descriptive statistics of variables

Descriptive statistics for $PM_{2.5}$, road_1 , road_2 , road_3

pm2.5		*road_1*	
Mean	106.0602564	Mean	37.70346371
Standard Error	0.465568519	Standard Error	0.883164453
Median	99.815	Median	0
Standard Deviation	47.95581744	Standard Deviation	90.97022577
Sample Variance	2299.760426	Sample Variance	8275.581976
Kurtosis	1.729655672	Kurtosis	15.37643777
Skewness	1.017249133	Skewness	3.473588458
Range	366.63	Range	963.76
Minimum	17.62	Minimum	0
Maximum	384.25	Maximum	963.76
Sum	1125299.32	Sum	400033.75
Count	10610	Count	10610

road_2		*road_3*	
Mean	158.1576494	Mean	138.1194317
Standard Error	2.132650949	Standard Error	1.867527167
Median	60.665	Median	56.985
Standard Deviation	219.6734002	Standard Deviation	192.3643636
Sample Variance	48256.40274	Sample Variance	37004.0484
Kurtosis	4.137649543	Kurtosis	4.622037978
Skewness	1.89050049	Skewness	1.989698174
Range	1592.23	Range	1409.85
Minimum	0	Minimum	0
Maximum	1592.23	Maximum	1409.85
Sum	1678052.66	Sum	1465447.17
Count	10610	Count	10610

Descriptive statistics for road_4 , road_5 , RH , SR

road_4		*road_5*	
Mean	321.5434788	Mean	13.21019133
Standard Error	3.865576196	Standard Error	0.325674
Median	166.39	Median	0
Standard Deviation	398.1731127	Standard Deviation	33.54600289
Sample Variance	158541.8277	Sample Variance	1125.33431
Kurtosis	2.522913765	Kurtosis	12.57507415
Skewness	1.605135284	Skewness	3.357694816
Range	2490.8	Range	209.23
Minimum	0	Minimum	0
Maximum	2490.8	Maximum	209.23
Sum	3411576.31	Sum	140160.13
Count	10610	Count	10610

RH		*SR*	
Mean	56.13223434	Mean	140.0301624
Standard Error	0.167735801	Standard Error	1.90003537
Median	54.93	Median	30.43
Standard Deviation	16.37723434	Standard Deviation	181.9773692
Sample Variance	268.2138045	Sample Variance	33115.76292
Kurtosis	-0.7820929	Kurtosis	0.051124373
Skewness	0.064776704	Skewness	1.161247139
Range	82.86	Range	781.82
Minimum	11.69	Minimum	2.14
Maximum	94.55	Maximum	783.96
Sum	535108.59	Sum	1284496.68
Count	9533	Count	9173

Descriptive statistics for BP , AT , near road point count , vegetation point count

BP		*AT*	
Mean	971.4930747	Mean	26.77707174
Standard Error	0.559795794	Standard Error	0.053744918
Median	982.59	Median	26.76
Standard Deviation	50.67929983	Standard Deviation	4.865622116
Sample Variance	2568.391431	Sample Variance	23.67427857
Kurtosis	17.55154424	Kurtosis	-0.595676301
Skewness	-4.403733321	Skewness	0.188461221
Range	259.14	Range	28.46
Minimum	735.67	Minimum	13.31
Maximum	994.81	Maximum	41.77
Sum	7962357.24	Sum	219464.88
Count	8196	Count	8196

near road count		*vegetation count*	
Mean	281.0111216	Mean	175.1993402
Standard Error	2.16428773	Standard Error	1.893259057
Median	230	Median	117
Standard Deviation	222.9321421	Standard Deviation	195.0148732
Sample Variance	49698.74	Sample Variance	38030.80077
Kurtosis	0.013329131	Kurtosis	6.589708854
Skewness	0.953564571	Skewness	2.332498839
Range	938	Range	1058
Minimum	2	Minimum	0
Maximum	940	Maximum	1058
Sum	2981528	Sum	1858865
Count	10610	Count	10610

Descriptive statistics for built-up points count , bare soil points count , water points count , WS

built up count		*bare soil count*	
Mean	379.7140434	Mean	113.6504241
Standard Error	2.732172722	Standard Error	1.321903768
Median	350	Median	64
Standard Deviation	281.4270531	Standard Deviation	136.162505
Sample Variance	79201.1862	Sample Variance	18540.22777
Kurtosis	-0.57278498	Kurtosis	10.35893697
Skewness	0.628574674	Skewness	2.685020012
Range	1043	Range	935
Minimum	0	Minimum	0
Maximum	1043	Maximum	935
Sum	4028766	Sum	1205831
Count	10610	Count	10610

water count		*WS*	
Mean	4.047031103	Mean	1.092066506
Standard Error	0.124613418	Standard Error	0.008715466
Median	0	Median	0.89
Standard Deviation	12.83578692	Standard Deviation	0.85095267
Sample Variance	164.7574259	Sample Variance	0.724120446
Kurtosis	56.05182598	Kurtosis	4.049818924
Skewness	6.734628528	Skewness	1.758245084
Range	121	Range	6.4
Minimum	0	Minimum	0.06
Maximum	121	Maximum	6.46
Sum	42939	Sum	10410.67
Count	10610	Count	9533

of total road length with high speed varies greatly spatially. On other hand wind speed as well as other meteorological parameters accounts for the least deviation suggesting all the buffer areas have relatively more consistency with regard to meteorological data.

As there were no valid points in road category with ratio of average speed to free flow speed around 0, this road category was removed from the analysis. For the remaining roads highest mean comes out to be of road category 4 while road category 1 and road category 5 have lower value for mean. It can be due to the fact that none of the roads achieve full free flow capacity while having nearly full jam condition is also difficult. For all the roads the mode was 0 while the standard deviation was maximum for road category 4. From Table 4.1 it can be seen road 4 corresponds to highest values while road 5 results in lowest values.

Temperature was not available for the given period at all times and only 24 stations provided information on the temperature while no water body was detected in 7 of the sites. This may be due to mis-classification or the spectral signature presence was less than the pixel size making it a sub-pixel target.

4.2 Correlation Analysis

From Figure 4.2 it can be found that none of the predictor variable is strongly correlated to the pollutant concentration. Particulate pollution is weakly yet positively related to river density which is explained as air column above river moves along the river thus pollutants are transported from another location. Particulate pollution is also positively correlated with built up area as more hot-spots are developed in highly dense areas. It is negatively related to park and forest density which is due to the fact that leaves and tree canopy provides more surface area for the particulate pollutants to settle as well as they also restore normal gaseous concentration. Negative co relation from temperature. This may be accounted from the fact that higher temperature will result in more movement of wind due to which more dilution happens. For pressure also particulate concentration is negatively correlated. Higher Solar radiation will result in higher temperature and as the temperature is negatively correlated and temperature is positively correlated with solar radiance thus a natural outcome for SR and PM will be negative. For the wind speed

Variables	pm2.5	road_1	road_2	road_3	road_4	road_5	RH	WS	SR	BP	AT	near road count	vegetation count	built up count	bare soil count	water count
pm2.5	1	-0.165	-0.177	-0.065	0.177	-0.007	0.498	-0.307	-0.177	-0.096	-0.305	-0.010	-0.120	0.089	0.007	0.090
road_1	-0.165	1	0.337	0.066	-0.145	0.121	-0.171	0.076	0.029	-0.022	0.118	0.285	0.028	0.192	-0.069	0.054
road_2	-0.177	0.337	1	0.275	-0.113	0.251	-0.290	0.103	0.092	-0.011	0.233	0.417	0.089	0.348	0.078	0.129
road_3	-0.065	0.066	0.275	1	0.111	0.231	-0.153	0.046	0.052	0.033	0.139	0.251	0.165	0.234	0.230	0.160
road_4	0.177	-0.145	-0.113	0.111	1	0.152	0.270	-0.003	-0.147	0.040	-0.247	0.263	0.218	0.144	0.197	0.093
road_5	-0.007	0.121	0.251	0.231	0.152	1	0.028	-0.041	-0.024	0.105	-0.007	0.377	0.139	0.053	0.092	-0.038
RH	0.498	-0.171	-0.290	-0.153	0.270	0.028	1	-0.356	-0.590	0.061	-0.795	-0.020	0.058	-0.025	0.002	-0.005
WS	-0.307	0.076	0.103	0.046	-0.003	-0.041	-0.356	1	0.424	0.140	0.188	0.067	0.028	-0.018	-0.043	-0.099
SR	-0.177	0.029	0.092	0.052	-0.147	-0.024	-0.590	0.424	1	0.077	0.569	-0.034	-0.001	0.013	0.013	-0.021
BP	-0.096	-0.022	-0.011	0.033	0.040	0.105	0.061	0.140	0.077	1	-0.061	0.010	0.096	-0.107	0.038	-0.001
AT	-0.305	0.118	0.233	0.139	-0.247	-0.007	-0.795	0.188	0.569	-0.061	1	0.014	-0.081	0.059	0.022	0.094
near road count	-0.010	0.285	0.417	0.251	0.263	0.377	-0.020	0.067	-0.034	0.010	0.014	1	0.093	0.485	-0.050	0.084
vegetation count	-0.120	0.028	0.089	0.165	0.218	0.139	0.058	0.028	-0.001	0.096	-0.081	0.093	1	-0.026	0.268	-0.018
built up count	0.089	0.192	0.348	0.234	0.144	0.053	-0.025	-0.018	0.013	-0.107	0.059	0.485	-0.026	1	0.176	0.263
bare soil count	0.007	-0.069	0.078	0.230	0.197	0.092	0.002	-0.043	0.013	0.038	0.022	-0.050	0.268	0.176	1	0.084
water count	0.090	0.054	0.129	0.160	0.093	-0.038	-0.005	-0.099	-0.021	-0.001	0.094	0.084	-0.018	0.263	0.084	1

Figure 4.2: Correlation analysis

as normal mean wind speed is less than 3m/s it helps it dispersing the pollutants thus wind speed is showing negative co relation with pollutants concentration. For relative humidity a positive correlation can be seen. As the data collection was done in month of March humidity was less due to which positive correlation is seen. From literature it is known that at higher humidity levels correlation with pollutant concentration reverses the sign. The positive correlation between pollutant concentration and road network density is because the one of the major air pollution source increases as road network density increases.

4.3 Model Development and Evaluation

Model parameters are depicted in Table 4.2 with obtained p-values. The significance level set for this study was 5%. For this model 3 iterations were done by initially including all the variables for the model building. The highest p-value obtained was 0.977 for bare soil points. A R^2 value of 0.37 was obtained. As this p-value is way above the significance level of 0.05 this variable was removed from the model building and again regression analysis was done from which highest p-value obtained was 0.952 for road_5. This variable was also removed and again iteration was done and highest p-value in this model was obtained to be 0.856 for near road variable. In the next iteration variable road_2 was removed as its p-value was 0.187. Thus the final model obtained had highest p-value of nearly 0.06 for road_3 which was kept in the model.

Table 4.2: Model Parameters

Source	Value	Standard error	t	Pr > —t—	Lower bound (95%)	Upper bound (95%)
Intercept	68.167	10.440	6.529	<**0.0001**	47.702	88.633
road_1	-0.033	0.005	-6.911	<**0.0001**	-0.042	-0.024
road_3	0.005	0.002	1.878	0.060	0.000	0.009
road_4	0.009	0.001	7.610	<**0.0001**	0.007	0.012
RH	1.971	0.049	39.935	<**0.0001**	1.875	2.068
WS	-9.371	0.584	-16.042	<**0.0001**	-10.516	-8.226
SR	0.064	0.003	19.733	<**0.0001**	0.058	0.071
BP	-0.103	0.009	-11.810	<**0.0001**	-0.120	-0.086
AT	1.064	0.155	6.882	<**0.0001**	0.761	1.368
vegetation count	-0.034	0.002	-16.098	<**0.0001**	-0.038	-0.030
built up count	0.011	0.002	6.535	<**0.0001**	0.008	0.015
water count	0.155	0.032	4.873	<**0.0001**	0.093	0.218

4.3.1 ANOVA

From the analysis of variance the obtained results are shown in Table 4.3. Adjusted mean squares are a tool to measure how much variation is ebeing explained by the model. Mean square can be seen as variance around the fitted values. Given the p-value of the F statistic computed in the ANOVA table, and given the significance level of 5%, the information brought by the explanatory variables is significantly better than what a basic mean would bring.

Table 4.3: Analysis of Variance

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	11	6783541.226	616685.566	417.358	<**0.0001**
Error	8084	11944866.170	1477.594		
Corrected Total	8095	18728407.395			

4.3.2 Goodness of fit statistics

The goodness of fit statistics obtained from the analysis of result is shown in Table 4.4. A total of 8096 observations were taken for the model development while 100 other observations were used for data validation. All the observations were given equal weights. An adjusted R^2 value of 0.361 i.e. 36% was obtained for the model. The root mean square error of 38.439 was obtained for the training set while mean absolute percentage error (MAPE) is 30.33%.

Table 4.4: Goodness of fit statistics

Statistic	Training set	Validation set
Observations	8096	100
Sum of weights	8096	100
DF	8084	88
R^2	0.362	0.348
Adjusted R^2	0.361	
MSE	1477.594	1342.789
RMSE	38.439	36.644
MAPE	30.333	31.391

4.4 Model Validation

The model was validated by 2 ways. First the validation was done on 100 random observations from the input data set. For the second 10 instances from 2 locations were chosen for which particulate pollution data was collected. Along with particulate pollution, humidity and temperature were also recorded. Buffer area around the 2 locations were made and land use category density of all land use types were found. Similarly the road category sum length was also done and length of road in each category was determined. The obtained model was then applied on this data set to obtain the predicted values. Normalised Mean Square Error (NMSE) and Fractional Bias (FB) was calculated to evaluate statistical performance measures.

In the first way, the goodness of fit statistics can be seen in Table 4.4. It can be observed that for 100 observations the degree of freedom was 88, while the obtained R^2 value was 0.348 or approx 35%. The root mean square error was obtained to be 36 while MAPE was 31.4%.

In the second way for performance measure following were taken into account:

- Normalised Mean Square Error (NMSE) : It can be calculated as :

$$NMSE = \frac{\overline{(C_p - C_o)^2}}{\overline{C_p} \times \overline{C_o}}$$

where C_p is predicted concentration of pollutant while C_o is observed concentration of pollutant.

For the model to be acceptable the value of NMSE should be less than 0.5. The obtained value of NMSE is 0.099.

- Fractional Bias (FB) : It can be calculated as :

$$FB = 2 \times \frac{(\overline{C_p} - \overline{C_o})}{(\overline{C_p} + \overline{C_o})}$$

For the model to be acceptable the value of FB should lie between ± 0.5. The obtained value of FB is -0.13.

Chapter 5

Inferences and Conclusions

5.1 Discussion

The land use regression technique has been use in prediction for pollutants concentration by various studies in the past like by Beelen et al. (2013) and Briggs et al. (1997). For this work the data was collected for one 2-week period in the month of March. Kanaroglou et al. (2005) and Gilbert et al. (2005) also conducted the study for one 2-week period although the number of monitoring stations in both the studies were 100 and 67 respectively while in this study only 36 monitoring stations were taken into account. Various researchers have used different pollutant as dependent variable like NO_2 (Sahsuvaroglu et al., 2006; Ross et al., 2005; Briggs et al., 1997; Brauer et al., 2003), $PM_{2.5}$ (Brauer et al., 2003; HOEK et al., 2001; ROSS et al., 2007; Morgenstern et al., 2006) and NO_x (Stedman et al., 1997; Madsen et al., 2007). The buffer area in these above-mentioned studies varied from 500m to 2000m with concentric ring. These rings were spaced at a distance of 50m to 300m. For this work 4 concentric ringed buffer zones were selected with spacing for concentric rings at 250m, thus creating a total buffer of 1000m radius.This analysis was performed based on the linear regression model using backward elimination method. The results obtained from the modelling were found to be on lower side according to the literature reviewed. The R^2 value of various research ranged from 35% to 80% (Henderson et al., 2007). Hoogh et al. (2013) reported the model performed worst in case of $PM_{2.5}$ with explained variablity being just under 50%. The R^2 value obtained for model developed in this study was 0.36 signifying that only 36% of the variability is explained. Hochadel et al. (2006) reported the explained variability to be 17% for particulate pollution while for NO_2 it was 90%.

5.2 Conclusions

In the given study, the predictor variables of land use i.e. vegetation density, built-up density and water density in any buffer area had significant effect in the final model. Other than land use, meteorological variables also accounted for high number of statistically significant variables with humidity, wind speed, solar radiance, pressure and temperature being important variables.

To include the effect of real time traffic congestion, road were categorised into different categories and effect of each road category depicted the effect of any congestion level in the model.

If all the road length in all categories are summed up and regressed with pollutant concentration correlation is positive thus it can be concluded that if total length of road is increased in buffer area then the pollutants concentration is increased.

From the analysis of variance it can be concluded that model is significant at 5% confidence interval.

MAPE value of the model is 30.33% which is quite high. This may be due to confounding effect among meteorological variables or road categories.

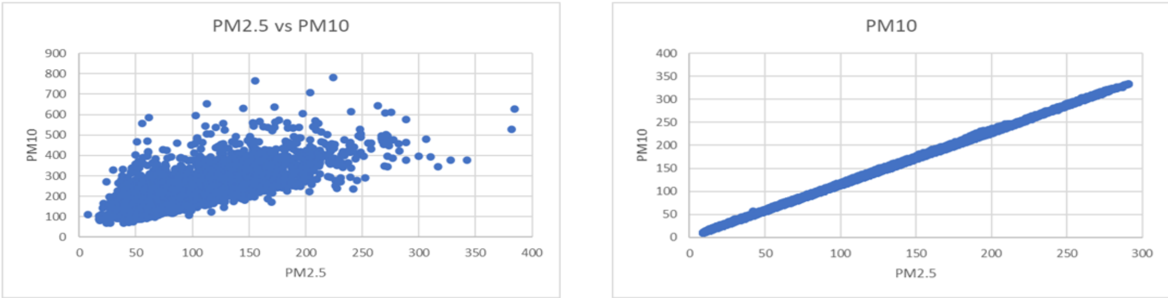
For validation of 100 random sets the explained variability values lies between 35% to 42%.

NMSE and FB validate the model with the actual data obtained in the field.

5.3 Limitations

The data obtained from CPCB showed a R^2 value of 0.67 between PM_{10} and $PM_{2.5}$, while this value should be approximately 0.95 as for the data obtained by mobile monitors as shown in Figure 5.1

Low value for explained variability may be resulted from a poor classification or limited data for training. The obtained spectral signatures resulted in degraded surface reflectance from roads. And this is the reason behind the failure in detecting the road network with a higher accuracy. This can be avoided by increasing the resolution to 1m but at same time it can increase the computational time. Another proposed way can be to use the open street map (OSM) data to determine road network in the buffer area.



(a) Data from CPCB (a) Data from mobile monitors

Figure 5.1: Comparison of $PM_{2.5}$ data

For this study the 4 hour pollutants concentration was used. This can be replaced to smaller time bins for better accuracy.

Another major drawback of the Land use model comes from its area specificity i.e. the resulting model can be applied only to the area with same geographic features.

Another limitation of this study arises due to the size of buffer area. As each predictor variable has a different radii till which it can affect the surroundings, for this study a multi-ring buffer area for 4 rings of 250m was defined around the fixed monitoring site.

The biggest limitation in this land use regression methodology is due to incorrect training. This type of error may creep in due to both human factors as well as incorrect data recorded by the sensors due to any factor like high cloud cover percentage. Human errors can be limited by exercising caution while classification while some unavoidable circumstances arise like the following cases:

- presence of small, sub-pixel targets
- presence of boundaries of discrete land cover classes
- gradual transition between land cover classes (continuum)
- contribution of areas outside the area represented by a pixel

5.4 Future Scope

This model can be redeveloped on a much smaller time interval like one hour basis. This can also help in increasing the training set which may help in improving the goodness of fit for the model.

The model can also be applied on all other live monitoring stations available in the country.

Data from mobile monitors can be used in the study.

References

J. S. Apte, T. W. Kirchstetter, A. H. Reich, S. J. Deshpande, G. Kaushik, A. Chel, J. D. Marshall, and W. W. Nazaroff. Concentrations of fine, ultrafine, and black carbon particles in auto-rickshaws in New Delhi, India. *Atmospheric Environment*, 45(26):4470–4480, 2011. doi:10.1016/j.atmosenv.2011.05.028.

R. Beelen, G. Hoek, D. Vienneau, M. Eeftens, K. Dimakopoulou, X. Pedeli, M.Y. Tsai, N. Künzli, T. Schikowski, A. Marcon, K.T. Eriksen, O.R. Nielsen, E. Stephanou, E. Patelarou, T. Lanki, T.Y. Tuomi, and C. Declercq. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in europe – the ESCAPE project. *Atmospheric Environment*, 72:10–23, 2013. doi:10.1016/j.atmosenv.2013.02.037.

T. Bellander, N. Berglind, P. Gustavsson, T. Jonson, F. Nyberg, and G. Pershagen. Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in stockholm. *Environmental Health Perspectives*, 109(6):633–639, 2001. doi:10.1289/ehp.01109633.

M. Brauer, G. Hoek, P. van Vliet, K. Meliefste, P. Fischer, U. Gehring, J. Heinrich, J. Cyrys, T. Bellander, M. Lewne, and B. Brunekreef. Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems. *Epidemiology*, 14(2):228–239, 2003. ISSN 10443983. URL http://www.jstor.org/stable/3703461.

M. Brauer, G. Freedman, J. Frostad, A. van Donkelaar, R.V. Martin, F. Dentener, R. van Dingenen, K. Estep, H. Amini, J.S. Apte, K. Balakrishnan, L. Barregard, D. Broday, V. Feigin, S. Ghosh, P.K. Hopke, L.D. Knibbs, Y. Kokubo, Y. Liu, S. Ma, L. Morawska, J.L.T. Sangrador, G. Shaddick, H.R. Anderson, T. Vos, M.H. Forouzanfar, R.T. Burnett, and A. Cohen. Ambient air pollution exposure estimation for the global burden of disease 2013. *Environmental Science & Technology*, 50(1):79–88, December 2015. doi:10.1021/acs.est.5b03709.

D. Briggs. The role of gis: Coping with space (and time) in air pollution exposure assessment. *Journal of Toxicology and Environmental Health, Part A*, 68(13-14):1243–1261, 2005. doi:10.1080/15287390590936094.

D.J. Briggs, S. Collins, P. Elliott, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H.V. Reeuvyk, K. Smallbone, and A.V.D. Veen. Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographical Information Science*, 11(7):699–718, 1997. doi:10.1080/136588197242158.

D.J. Briggs, C. Hoogh, J. Gulliver, J. Wills, P. Elliott, S. Kingham, and K. Smallbone. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *Science of The Total Environment*, 253(1-3):151–167, 2000. doi:10.1016/s0048-9697(00)00429-0.

I. Brown. Landsat-8: how not to plan a mission (Part 2: the sensors), 2017. URL https://digital-geography.com/landsat-8-how-not-to-plan-a-mission-part-2-the-sensors/,.

Z. Chen, D. Chen, C. Zhao, M. Kwan, J. Cai, Y. Zhuang, B. Zhao, X. Wang, B. Chen, J. Yang, and R. Li. Influence of meteorological conditions on PM2.5 concentrations across china: A review of methodology and mechanism. *Environment International*, 139:105558, 2020. doi:10.1016/j.envint.2020.105558.

G. Ciccone, F. Forastiere, N. Agabiti, A. Biggeri, L. Bisanti, E. Chellini, G. Corbo, V. Dell'Orco, P. Dalmasso, T.F. Volante, C. Galassi, S. Piffer, E. Renzoni, F. Rusconi, P. Sestini, and G. Viegi. Road traffic and adverse respiratory effects in children. SIDRIA collaborative group. *Occupational and Environmental Medicine*, 55(11):771–778, 1998. doi:10.1136/oem.55.11.771.

J. Cyrys, M. Eeftens, J. Heinrich, C. Ampe, A. Armengaud, R. Beelen, T. Bellander, T. Beregszaszi, M. Birk, M. Cesaroni, and G. Hoek. Variation of NO2 and NOx concentrations between and within 36 european study areas: Results from the ESCAPE study. *Atmospheric Environment*, 62:374–390, December 2012. doi:10.1016/j.atmosenv.2012.07.080. URL https://doi.org/10.1016/j.atmosenv.2012.07.080.

Explained Desk. Explained: What is grap, delhi-ncr's action plan as air pollution increases?, 2020. URL https://indianexpress.com/article/explained/explained-what-is-grap-delhi-ncrs-action-plan-as-air-pollution-increases-6719746/.

A. Dutta and W. Jinsart. Gaseous and particulate matter emissions from road transport: The case of kolkata, india. *Environmental and Climate Technologies*, 25(1):717–735, 2021. doi:10.2478/rtuect-2021-0054.

EDF. Why smog standards are important for our health, 2017. URL https://www.edf.org/health/why-smog-standards-are-important-our-health#:~:text=When%20inhaled%2C%20smog%20irritates%20our,you%20may%20cough%20and%20wheeze.

M. Eeftens, R. Beelen, K. de Hoogh, T. Bellander, G. Cesaroni, M. Cirach, C. Declercq, A. Dėdelė, E. Dons, A. de Nazelle, K. Dimakopoulou, K. Eriksen, G. Falq, P. Fischer, C. Galassi, R. Gražulevičienė, J. Heinrich, B. Hoffmann, M. Jerrett, D. Keidel, M. Korek, T. Lanki, and S. Lindley. Development of land use regression models for PM2.5, PM2.5 absorbance, PM10 and PMcoarse in 20 european study areas: Results of the ESCAPE project. *Environmental Science & Technology*, 46(20):11195–11205, 2012. doi:10.1021/es301948k.

Delhi Envis CPCB. Air pollution in Delhi - An analytical study, 2016. URL http://www.cpcbenvis.nic.in/envis_newsletter/Air%20Quality%20of%20Delhi.pdf.

eo4geocourses. Urban Heat Islands – Basic GIS knowledge vector and raster data, 2018. URL http://www.eo4geo.eu/training/urban-heat-islands-basic-gis-knowledge-vector-and-raster-data/.

R.L. Forstall, R.P. Greene, and J.B. Pick. WHICH ARE THE LARGEST? WHY LISTS OF MAJOR URBAN AREAS VARY SO GREATLY. *Tijdschrift voor economische en sociale geografie*, 100(3):277–297, 2009. doi:10.1111/j.1467-9663.2009.00537.x.

GeoSage. Spectral Discovery for Landsat-8 Imagery, 2016. URL https://www.geosage.com/highview/features_landsat8.html.

N.L. Gilbert, M.S. Goldberg, B Beckerman, J.R. Brook, and M. Jerrett. Assessing spatial variability of ambient nitrogen dioxide in montréal, canada, with a land-use regression model. *J. Air Waste Manag. Assoc.*, 55(8):1059–1063, 2005.

S.K. Guttikunda, K.A. Nishadh, and P. Jawahar. Air pollution knowledge assessments (APnA) for 20 indian cities. *Urban Climate*, 27:124–141, 2019. doi:10.1016/j.uclim.2018.11.005.

S.B. Henderson, B. Beckerman, M. Jerrett, and M. Brauer. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental Science &amp Technology*, 41(7):2422–2428, 2007. doi:10.1021/es0606780.

M. Hochadel, J. Heinrich, U. Gehring, V. Morgenstern, T. Kuhlbusch, U. Link, H.E. Wichmann, and U. Krämer. Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. *Atmospheric Environment*, 40(3):542–553, 2006. doi:10.1016/j.atmosenv.2005.09.067.

G. HOEK, P. FISCHER, P.D. BRANDT, S. GOLDBOHM, and B. BRUNEKREEF. Estimation of long-term average exposure to outdoor air pollution for a cohort study on mortality. *Journal of Exposure Science &amp Environmental Epidemiology*, 11(6):459–469, 2001. doi:10.1038/sj.jea.7500189.

K. Hoogh, M. Wang, M. Adam, C. Badaloni, R. Beelen, M. Birk, G. Cesaroni, M. Cirach, B. Brunekreef, and G. Hoek. Development of land use regression models for particle composition in twenty study areas in europe. *Environmental Science &amp Technology*, 47(11):5778–5786, 2013. doi:10.1021/es400156t.

M. Jerrett, R. T. Burnett, P. Kanaroglou, J. Eyles, N. Finkelstein, C. Giovis, and J.R. Brook. A GIS–environmental justice analysis of particulate air pollution in hamilton, canada. *Environment and Planning A: Economy and Space*, 33(6):955–973, 2001. doi:10.1068/a33137.

M. Jerrett, A. Arain, P. Kanaroglou, B. Beckerman, D. Potoglou, T. Sahsuvaroglu, J. Morrison, and C. Giovis. A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology*, 15(2):185–204, 2004. doi:10.1038/sj.jea.7500388.

P.S. Kanaroglou, M. Jerrett, J. Morrison, B. Beckerman, M. Arain, N.L. Gilbert, and J.R. Brook. Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. *Atmospheric Environment*, 39(13):2399–2409, 2005. doi:10.1016/j.atmosenv.2004.06.049.

T. Kaneda and K. Bietsch. World population data sheet 2015. Technical report, Population Reference Bureau, 2015.

S. Kathuria, P. Puri, S. Nandar, and V. Ramesh. Effects of air pollution on the skin: A review. *Indian Journal of Dermatology, Venereology and Leprology*, 83(4):415, 2017. doi:10.4103/0378-6323.199579.

P. Kumar, A.P. Patton, J.L. Durant, and H.C. Frey. A review of factors impacting exposure to PM2.5, ultrafine particles and black carbon in Asian transport microenvironments. *Atmospheric Environment*, 187:301–316, 2018. doi:10.1016/j.atmosenv.2018.05.046.

A. Larkin, A. Donkelaar, J.A. Geddes, R.V. Martin, and P. Hystad. Relationships between changes in urban characteristics and air quality in east asia from 2000 to 2010. *Environmental Science &amp Technology*, 50(17):9142–9149, 2016. doi:10.1021/acs.est.6b02549.

E. Lebret, D. Briggs, H. van Reeuwijk, P. Fischer, K. Smallbone, H. Harssema, B. Kriz, P. Gorynski, and P. Elliott. Small area variations in ambient NO2 concentrations in four european areas. *Atmospheric Environment*, 34(2):177–185, 2000. doi:10.1016/s1352-2310(99)00292-7.

J.H. Lee, C.F. Wu, G. Hoek, K. de Hoogh, R. Beelen, B. Brunekreef, and C.C. Chan. Land use regression models for estimating individual NOx and NO2 exposures in a metropolis with a high density of traffic roads and population. *Science of The Total Environment*, 472: 1163–1171, 2014. doi:10.1016/j.scitotenv.2013.11.064.

X. Li, Y.J. Feng, and H.Y. Liang. The impact of meteorological factors on PM2.5 variations in hong kong. *IOP Conference Series: Earth and Environmental Science*, 78:012003, 2017. doi:10.1088/1755-1315/78/1/012003.

C. Madsen, K.C. Carlsen, G. Hoek, B. Oftedal, P. Nafstad, K. Meliefste, R. Jacobsen, W. Nystad, K. Carlsen, and B. Brunekreef. Modeling the intra-urban variability of outdoor traffic pollution in oslo, norway—a GA2len project. *Atmospheric Environment*, 41(35):7500–7511, 2007. doi:10.1016/j.atmosenv.2007.05.039.

I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou. Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8, 2020. doi:10.3389/fpubh.2020.00014.

N. Manojkumar, M. Monishraj, and B. Srimuruganandam. Commuter exposure concentrations and inhalation doses in traffic and residential routes of Vellore city, India. *Atmospheric Pollution Research*, 12(1):219–230, 2021. doi:10.1016/j.apr.2020.09.002.

X. Meng, L. Chen, J. Cai, B. Zou, C.F. Wu, Q. Fu, Y. Zhang, Y. Liu, and H. Kan. A land use regression model for estimating the NO2 concentration in shanghai, china. *Environmental Research*, 137:308–315, 2015. doi:10.1016/j.envres.2015.01.003.

C. Mitsakou, J.P. Adamson, A. Doutsi, H. Brunt, S.J. Jones, A.M. Gowers, and K.S. Exley. Assessing the exposure to air pollution during transport in urban areas – Evidence review. *Journal of Transport & Health*, 21:101064, 2021. doi:10.1016/j.jth.2021.101064.

A. Mölter and S. Lindley. Developing land use regression models for environmental science research using the XLUR tool – more than a one-trick pony. *Environmental Modelling &amp Software*, 143:105108, 2021. doi:10.1016/j.envsoft.2021.105108.

V. Morgenstern, A. Zutavern, J. Cyrys, I. Brockow, U. Gehring, S. Koletzko, C.P. Bauer, D. Reinhardt, H.E. Wichmann, and J. Heinrich. Respiratory health and individual estimated exposure to traffic-related air pollutants in a cohort of young children. *Occupational and Environmental Medicine*, 64(1):8–16, 2006. doi:10.1136/oem.2006.028241.

J.A. Mulholland, A.J. Butler, J.G. Wilkinson, A.G. Russell, and P.E. Tolbert. Temporal and spatial distributions of ozone in atlanta: Regulatory and epidemiologic implications. *Journal of the Air & Waste Management Association*, 48(5):418–426, 1998. doi:10.1080/10473289.1998.10463695.

M. E. Nicholls, R. A. Pielke, J. L. Eastman, C. A. Finley, W. A. Lyons, C. J. Tremback, R. L. Walko, and W. R. Cotton. Applications of the RAMS numerical model to dispersion over urban areas. In *Wind Climate in Cities*, pages 703–732. Springer Netherlands, 1995. doi:10.1007/978-94-017-3686-2_34.

F. Nyberg, P. Gustavsson, L. Jarup, T. Bellander, N. Berlind, R. Jakobsson, and G. Pershagen. Urban air pollution and lung cancer in stockholm. *Epidemiology*, 11(5):487–495, 2000. doi:10.1097/00001648-200009000-00002.

B. Polednik and A. Piotrowicz. Pedestrian exposure to traffic-related particles along a city road in Lublin, Poland. *Atmospheric Pollution Research*, 11(4):686–692, 2020. doi:10.1016/j.apr.2019.12.019.

Z. Ross, J.B. English, R. Scalf, R. Gunier, S. Smorodinsky, S. Wall, and M. Jerrett. Nitrogen dioxide prediction in southern california using land use regression modeling: potential for environmental health analyses. *Journal of Exposure Science &amp Environmental Epidemiology*, 16(2):106–114, 2005. doi:10.1038/sj.jea.7500442.

Z. ROSS, M. JERRETT, K. ITO, T. TEMPALSKI, and G. THURSTON. A land use regression for predicting fine particulate matter concentrations in the new york city region. *Atmospheric Environment*, 41(11):2255–2269, 2007. doi:10.1016/j.atmosenv.2006.11.012.

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

T. Sahsuvaroglu, A. Arain, P. Kanaroglou, N. Finkelstein, B. Newbold, M. Jerrett, B. Beckerman, J. Brook, M. Finkelstein, and N.L. Gilbert. A land use regression model for predicting ambient concentrations of nitrogen dioxide in hamilton, ontario, canada. *Journal of the Air & Waste Management Association*, 56(8):1059–1069, 2006. doi:10.1080/10473289.2006.10464542.

Y. Shi, C. Ren, K.K.L. Lau, and E. Ng. Investigating the influence of urban land use and landscape pattern on PM2.5 spatial variation using mobile monitoring and WUDAPT. *Landscape and Urban Planning*, 189:15–26, 2019. doi:10.1016/j.landurbplan.2019.04.004.

R. Sinharay, J. Gong, B. Barratt, P.O. Strickland, S. Ernst, F.J. Kelly, J. Zhang, P. Collins, P. Cullinan, and K.F. Chung. Respiratory and cardiovascular responses to walking down a traffic-polluted road compared with walking in a traffic-free area in participants aged 60 years and older with chronic lung or heart disease and age-matched healthy controls: a randomised, crossover study. *The Lancet*, 391(10118):339–349, 2018. doi:10.1016/s0140-6736(17)32643-0.

Manik Soni. How to do backward elimination in machine learning?, 2020. URL https://maniksonituts.medium.com/how-to-do-backward-elimination-in-machine-learning-ad1690c5bc49.

J.R. Stedman, K.J. Vincent, G.W. Campbell, W.L. Justin, and C.E.H. Downing. New high resolution maps of estimated background ambient NOx and NO2 concentrations in the u.k. *Atmospheric Environment*, 31(21):3591–3602, 1997. doi:10.1016/s1352-2310(97)00159-3.

S. Tiwari, A. Thomas, P. Rao, D.M. Chate, V.K. Soni, S. Singh, S.D. Ghude, D. Singh, and P.K. Hopke. Pollution concentrations in delhi india during winter 2015–16: A case study of an odd-even vehicle strategy. *Atmospheric Pollution Research*, 9(6):1137–1145, 2018. doi:10.1016/j.apr.2018.04.008.

T. Tsuboi and N. Yoshikawa. Traffic flow analysis in ahmedabad (india). *Case Studies on Transport Policy*, 8(1):215–228, March 2020. doi:10.1016/j.cstp.2019.06.001. URL https://doi.org/10.1016/j.cstp.2019.06.001.

USEPA. Naaqs table, 2022. URL https://www.epa.gov/criteria-air-pollutants/naaqs-table.

E. Velasco, A. Retama, E. Segovia, and R. Ramos. Particle exposure and inhaled dose while commuting by public transport in Mexico City. *Atmospheric Environment*, 219:117044, 2019. doi:10.1016/j.atmosenv.2019.117044.

A. Venn. Local road traffic activity and the prevalence, severity, and persistence of wheeze in school children: combined cross sectional and longitudinal study. *Occupational and Environmental Medicine*, 57(3):152–158, 2000. doi:10.1136/oem.57.3.152.

P. Vliet, M. Knape, J. de Hartog, N. Janssen, H. Harssema, and B. Brunekreef. Motor vehicle exhaust and chronic respiratory symptoms in children living near freeways. *Environmental Research*, 74(2):122–132, 1997. doi:10.1006/enrs.1997.3757.

J. Wang and S. Ogawa. Effects of meteorological conditions on PM2.5 concentrations in nagasaki, japan. *International Journal of Environmental Research and Public Health*, 12 (8):9089–9101, 2015. doi:10.3390/ijerph120809089.

S. Wang, C. Fang, L. Sun, Y. Su, X. Chen, C. Zhou, K. Feng, and K. Hubacek. Decarbonizing china's urban agglomerations. *Annals of the American Association of Geographers*, 109(1): 266–285, 2018. doi:10.1080/24694452.2018.1484683.

S. Wang, S. Gao, S. Li, and K. Feng. Strategizing the relation between urbanization and air pollution: Empirical evidence from global countries. *Journal of Cleaner Production*, 243: 118615, 2020. doi:10.1016/j.jclepro.2019.118615.

WHO. 7 million premature deaths annually linked to air pollution, 2014. URL https://www.who.int/news/item/25-03-2014-7-million-premature-deaths-annually-linked-to-air-pollution.

WHO. Ambient air pollution: A global assessment of exposure and burden of disease. Technical report, World Health Organization, 2018.

WHO. Ambient (outdoor) air pollution, 2021. URL https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

P. Wilkinson, P. Elliott, C. Grundy, G. Shaddick, B. Thakrar, and P. Walls. Case-control study of hospital admission with asthma in children aged 5-14 years: relation with road traffic in north west london. *Thorax*, 54(12):1070–1074, 1999. doi:10.1136/thx.54.12.1070.

W. Xu, X. Jin, M. Liu, Z. Ma, Q. Wang, and Y. Zhou. Analysis of spatiotemporal variation of PM2.5 and its relationship to land use in china. *Atmospheric Pollution Research*, 12(9): 101151, 2021. doi:10.1016/j.apr.2021.101151.

D. Yang, D. Lu, J. Xu, C. Ye, J. Zhao, G. Tian, X. Wang, and N. Zhu. Predicting spatio-temporal concentrations of PM2.5 using land use and meteorological data in yangtze river delta, china. *Stochastic Environmental Research and Risk Assessment*, 32(8):2445–2456, 2017. doi:10.1007/s00477-017-1497-6.