TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PASHCHIMANCHAL CAMPUS

A REPORT ON

'Estimation and Prediction of PM2.5 & PM10 Concentration in Kathmandu Valley Using Remote Sensing & Machine learning'

By

Aashish Baral (PAS077BGE002)
Aayush Chand (PAS077BGE003)
Ashim Poudel (PAS077BGE005)
Gaurav Baral (PAS077BGE021)
Indra Paudel (PAS077BGE022)

SUBMITTED TO:

DEPARTMENT OF GEOMATICS ENGINEERING
IOE PASHCHIMANCHAL CAMPUS
TRIBHUVAN UNIVERSITY
POKHARA, NEPAL

UNDER THE SUPERVISION OF:

Er. Netra Bahadur Katuwal

April 18, 2025

1. Abstract

In recent years, public health and the environment have been seriously threatened by air pollution, especially in low and middle-income nations like Nepal, where the ground-level stations are sparse and thus have limited spatial coverage for monitoring air quality at specific hotspot areas. This study uses machine learning algorithms (Linear Regression, Random Forest, and Gradient Boosting) and satellite data to meet the importunate demand for precise PM2.5 and PM10 prediction. As the ground station, we used data from the monitoring stations of the Department of Environment covering the period January 2022 to December 2023 and satellite-derived parameters, including Aerosol Optical Depth (AOD), Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI) and soil moisture were processed, analyzed and used to train the models. The Gradient Boosting algorithm performed consistently well with average R^2 scores of 0.82 and 0.84 for PM2.5 and PM10, respectively, proving efficient in catching the non-linear relationships in the air quality data. Likewise, Random Forest also showed reliable accuracy with an average R^2 value of scores 0.80 and 0.82 for PM2.5 and PM10, respectively. Meanwhile, the performance of linear regression was not satisfactory with R^2 scores of 0.45 and 0.50 for PM2.5 and PM10, respectively, showing its inability to model complex non-linear data distributions. Spatial maps created from model predictions highlight pollution hotspots, giving active insights for targeted interventions. This study demonstrates the potential of integrating satellite data with machine learning techniques to predict air quality accurately. This finding provides valuable insights for developing targeted intervention strategies. Additionally, this approach aids in air quality monitoring in the regions that lack ground-based monitoring stations, ensuring more comprehensive environmental assessments and supporting data-driven policy and informed decision-making in Nepal and similar regions. Future research should incorporate high-resolution datasets and additional meteorological variables to enhance model reliability and scalability.

Keywords: Air quality, Pm2.5, Pm10, Machine learning, Kathmandu valley, Remote sensing

2. Acknowledgement

Throughout the course of this project, we have received immense support, guidance, and encouragement from various individuals and institutions, to whom we are deeply grateful.

First and foremost, we would like to express our sincere appreciation to our respected supervisor, Er. Netra Bahadur Katuwal, for his valuable guidance, insightful feedback, and continuous encouragement, which were crucial in shaping the direction and outcome of our research.

We are also thankful to the Government of Nepal, Ministry of Population and Environment, Department of Environment, for providing us with reliable air quality monitoring data, which formed a key part of our study.

Our special thanks go to the Department of Geomatics Engineering, IOE Pashchimanchal Campus, Tribhuvan University, for providing the platform and academic environment necessary for carrying out this final year bachelor's project. The resources and support from the department have been vital to the successful completion of our work.

Table of Contents

LIST OF ABBREVIATIONS

- **AQI** Air Quality Index
- **AOD** Aerosol Optical Depth
- **CHIRPS** Climate Hazards Group InfraRed Precipitation with Station data
- **CO** Carbon Monoxide
- **GIS** Geographic Information System
- **GEE** Google Earth Engine
- **LST** Land Surface Temperature
- **LU/LC** Land Use/Land Cover
- **MCD** MODIS Combined Dataset
- **MSE** Mean Square Error
- **MAE** Mean Absolute Error
- **MODIS** Moderate Resolution Imaging Spectroradiometer
- **NDVI** Normalized Difference Vegetation Index
- **NO_2** Nitrogen Dioxide
- **PM** Particulate Matter
- **PM10** Particulate Matter with diameter less than 10 micrometers
- **PM2.5** Particulate Matter with diameter less than 2.5 micrometers
- **R^2** Coefficient of Determination
- **RF** Random Forest
- **GBR** Gradient Boosting Regression
- **SRFR** Simple Random Forest Regression
- **SLR** Simple Linear Regression
- **MLR** Multiple Linear Regression
- **SRFR** Multiple Random Forest Regression
- **RMSE** Root Mean Square Error
- **SDGs** Sustainable Development Goals
- **SHAP** Shapley Additive Explanations
- **SLR** Simple Linear Regression
- **SMAP** Soil Moisture Active Passive
- **SO_2** Sulfur Dioxide
- **TROPOMI** Tropospheric Monitoring Instrument
- **TSP** Total Suspended Particles
- **UN** United Nations
- **WHO** World Health Organization

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Background

Air pollution refers to the presence of harmful substances in the air, leading to adverse health effects and causing millions of premature deaths each year as people worldwide are exposed to toxic air pollutants that contribute to various respiratory and cardiovascular diseases. Air pollution has severe consequences, contributing to 4.2 million premature deaths globally in 2016. It impacts both urban and rural areas, with approximately 91% of the burden falling on low and middle-income countries, especially in Southeast Asia and the Western Pacific (WHO, 2021). Air pollution is caused by various pollutants, including particulate matter (PM), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and ozone (O_3) (NSW, 2022). Key pollutants of public health concern, including particulate matter, carbon monoxide, ozone, nitrogen dioxide, and sulphur dioxide, contribute significantly to respiratory and other diseases, making outdoor and indoor air pollution major causes of morbidity and mortality (WHO, 2021). Addressing this problem of air pollution, which is the second highest risk factor for noncommunicable diseases, is key to protecting public health. Particulate matter (PM) refers to inhalable particles composed of sulphate, nitrates, ammonia, sodium chloride, black carbon, mineral dust, or water. Particulate matter (PM) varies in size and is typically classified by aerodynamic diameter, with PM2.5 and PM10 being the most common types in regulatory standards and the most significant for health (WHO, 2021). Particulate matter (PM), also known as atmospheric aerosol, refers to a complex mixture of solid and liquid particles that vary in size and composition and can remain airborne for extended periods (Arvani et al., 2015). Particulate matter (PM) with aerodynamic diameters smaller than 2.5 μm (PM2.5) and 10 μm (PM10) has received significant attention due to its impact on human health and ecosystems, making it a key focus of recent research(Bodor et al., 2022; Faraji Ghasemi et al., 2020; Leão et al., 2023; Yunesian et al., 2019; Zoran et al., 2020a). Elevated PM10 levels are linked to higher hospital admissions for lung and heart diseases, while PM2.5 poses a greater health risk as it penetrates deeper into the respiratory system(Mamić et al., 2023).

According to WHO data, nearly 99% of the global population breathes air that exceeds the organization's guideline limits and contains high pollutant levels, with low- and middle-income countries experiencing the highest exposure (WHO, 2021). In 2015, the Sixty-Eighth World Health Assembly adopted resolution WHA68.8, titled "Health and the Environment: Addressing the Health Impact of Air Pollution," which was supported by 194 Member States (WHO, 2015). This resolution stated the need to redouble efforts to protect populations from the health risks posed by air pollution. In addition, the United Nations (UN) Sustainable Development Goals (SDGs) were designed to address the public health threat posed by air pollution via specific targets to reduce air pollution exposure and the disease burden from household and ambient exposure.

In Nepal, air pollution leads to 42,100 deaths every year, out of which 19% are in under-five children and about 27% in adults above 70 years of age. It reduces the life expectancy

of an average Nepali by 4.1 years. The data on the major causes of death in Nepal also shows that air pollution is a major contributor to the top five causes of death, namely COPD (66%), ischemic heart disease (34%), stroke (37%), Lower respiratory infection (47%) and neonatal deaths (22%). In many developing countries such as Nepal, air pollution is responsible for an estimated two million premature deaths globally each year(Timilsina et al., 2023). Most of the developed countries do not include the Total Suspended Particles (TSP), only including the PM10 under their consideration. According to different studies, the concentration of TSP in the air in Nepal is found to be dominating. Some of the Asian countries have already included TSP in the calculation of the air quality hence, Nepal has also included it in the air quality-related national standard. The limited spatial coverage of monitoring stations and the absence of a dense monitoring network, driven by economic and feasibility constraints, may introduce bias in epidemiological studies(Timilsina et al., 2023).

Air pollution severity is commonly assessed using the Air Quality Index (AQI), a widely recognized standard (Epa & of Air, 2014). The AQI is an essential tool for conveying air quality observations and evaluating the severity of air pollution, with higher AQI values signifying greater health risks. Ground-based air quality monitoring is traditionally used to determine the AQI because it offers accurate surface-level data (Bechle et al., 2013; EPA, 2018). However, the spatial resolution of these monitors is often limited, resulting in their placement mainly in areas identified as potential pollution hotspots(Bechle et al., 2013a; Lo Re et al., 2014). This issue is further compounded by the fact that several countries lack air quality monitoring stations, making the task of monitoring air quality even more difficult (WAQI, 2022). Additionally, the required maintenance costs can hinder their widespread deployment (Lo Re et al., 2014), prompting the exploration of alternative methods. Due to its extensive area coverage, remote sensing technology has emerged as a viable alternative for minimizing spatial distribution uncertainty(Martin, 2008) and supporting large-scale observations. However, air pollution observed by remote sensing is primarily limited to the Earth's upper atmosphere, where measurement sensitivities are high. Despite these limitations, these observations can still reflect variations in the distribution of air pollution across the Earth's surface(Filonchyk et al., 2021a; Sakti et al., 2023a; Z. Zheng et al., 2019a). Several studies have sought to estimate AQI values using remote sensing data or air pollution measurements from surface observations(Mozumder et al., 2013). From a remote sensing point of view, the first Copernicus mission with the main purpose of monitoring the atmosphere and tracking air pollutants, the Sentinel 5P TROPOMI mission, has been widely used. It is the most recent global satellite mission in monitoring air quality and daily measures concentrations of ozone (O3), methane (CH4), formaldehyde (HCHO), carbon monoxide (CO), nitrogen oxide (NO2), sulphur dioxide (SO2), and aerosol—provided as an aerosol index (AI). Particulate matter of a diameter smaller than 2.5 and 10 μm (PM2.5 and PM10) significantly determines air quality. Still, there are no available satellite sensors that allow us to track them remotely with high accuracy, but only using ground stations (Mamić et al., 2023). New modelling approaches incorporating satellite and other data may also be useful. In recent decades, in addition to existing air pollution monitoring networks,

advanced methods of exposure assessment have become available with the use of satellite observations and various modelling tools to support epidemiological studies, as well as health impact and risk assessment according to WHO global air quality guidelines.

Three types of numerical values have been applied in exposure assessments of ambient particles: (1) monitoring observations, (2) satellite remote sensing measurements of aerosol, and (3) air quality model simulations. Routine monitors were widely used to predict air pollution concentrations across an area using geostatistical methods such as Kriging (Jerrett et al., 2005; Mercer et al., 2011) or land use regression (LUR) to incorporate external spatial covariates (Eeftens et al., 2012; Henderson et al., 2007), but such monitors may be sparsely distributed in suburban or rural areas (Xue et al., 2017). Satellite remote sensing can capture integrated column concentrations of gases and aerosols from the Earth's surface to the top of the atmosphere and has been used to assess ground-level air pollution(Martin, 2008).

Given its high spatial coverage, using satellite-derived aerosol optical depth (AOD) data to estimate PM2.5 and PM10 mass concentrations has become a key focus of current research(S. Wu et al., 2024). Multiple studies suggest that PM2.5 and PM10 are among the primary contributors to air pollution. In Nepal, most air quality monitoring stations measure only PM10 and PM2.5 to determine the air quality index. Therefore, this study focuses on predicting the concentration of PM10 and PM2.5 by integrating satellite data with ground station data using multiple machine learning algorithms.

1.2 Statement of Problem

Air quality monitoring in Nepal faces significant challenges due to the limited spatial coverage provided by ground-based sensors. The existing network of air quality monitoring stations is sparse, with many areas lacking any form of coverage. Additionally, the problem is exacerbated by inactive ground-based sensors, which further reduce the reliability and comprehensiveness of air quality data.

The number of monitoring stations is insufficient to capture the spatial variability of air pollutants such as PM2.5 and PM10 across different regions. This scarcity of stations leads to large areas being left unmonitored, making it difficult to obtain an accurate assessment of air quality. This gap in data collection hampers the ability to fully understand and manage the air quality issues that affect public health and the environment.

Given these limitations, there is a pressing need for innovative approaches to enhance air quality monitoring in Nepal. Satellite remote sensing can reduce these uncertainties by providing broader spatial coverage. However, satellite observations are primarily sensitive to pollutants in the Earth's upper atmosphere, which necessitates robust modelling to represent surface-level air quality. Existing models often fail to account for the unique geographical, climatic, and socio-economic characteristics of each location, making a location-specific approach essential for accurate air quality assessment. The integration of remote sensing data with ground-based observations, coupled with advanced machine learning techniques, presents a promising solution. This approach can

address the gaps in spatial coverage, improve data reliability, and enable the detection of a broader range of pollutants

1.3 Objectives

- **Primary Objectives**
  The main objective of this study is to estimate PM2.5 and PM10 concentrations in Kathmandu Valley using Linear Regression, Random Forest, and Gradient Boosting models by integrating remote sensing and ground station data.

- **Secondary Objectives**
  i) To map PM2.5 and PM10 levels hotspots area using remote sensing data.
  ii) To bridge the gap in spatial coverage left by ground stations.
  iii) To increase the precision of air quality monitoring by utilising geospatial technology.
  iv) To validate the models using ground-based PM2.5 & PM10 data, ensuring the accuracy and reliability for forecasting all over Nepal in air quality prediction and management.

1.4 Study Area

The selected study area for this project is the Kathmandu Valley, which spans an area of 933.73 square kilometers. The geographical boundaries of the valley are defined by its northernmost position at 27.818° N, southernmost position at 27.403° N, easternmost position at 85.5657° E, and westernmost position at 85.189° E. The Kathmandu Valley, in the lesser Himalayas of Central Nepal, is bowl-shaped with a central elevation of 1,425 m. It is surrounded by four mountain ranges: Shivapuri (2,732 m), Phulchowki (2,695 m), Nagarjun (2,095 m), and Chandragiri (2,551 m) with the Bagmati River flowing through it. It's made up of three districts Kathmandu, Bhaktapur, and Lalitpur with respective population densities of 5169, 3631, and 1433 people per sq km. Its subtropical, continental, semi-humid climate has an average temperature of 18.3°C and a mean annual rainfall of 1,439.7 mm. Pollution mainly comes from open fires, vehicular emissions, construction dust, and post-earthquake damage, worsened by its low elevation and lack of dispersing winds.

Kathmandu Valley has been chosen as the study area due to its significant concentration of air quality monitoring stations, which is the highest in Nepal. Within this area, there are 7 governmental air quality monitoring stations and 44 purple-air sensors available, providing a substantial amount of ground-based data for model training and validation. This dense network of stations is crucial for developing and validating the models, as it offers a rich dataset for training purposes. While the ultimate goal is to extend the model to cover all of Nepal, starting with the Kathmandu Valley allows for a more manageable and focused approach. The complexity and time-consuming nature of processing remotely sensed data for a large area make it impractical to begin with the entire country.

Additionally, the availability of a relatively large number of ground stations in Kathmandu Valley, though still not as extensive as in some developed cities globally,

supports the justification for this study. The data from these stations will enable robust testing and applicability of the project, ensuring the model's accuracy and reliability. Once the model is successfully developed and validated in Kathmandu Valley, it can be extended to cover the entire country with necessary adjustments to the training parameters.

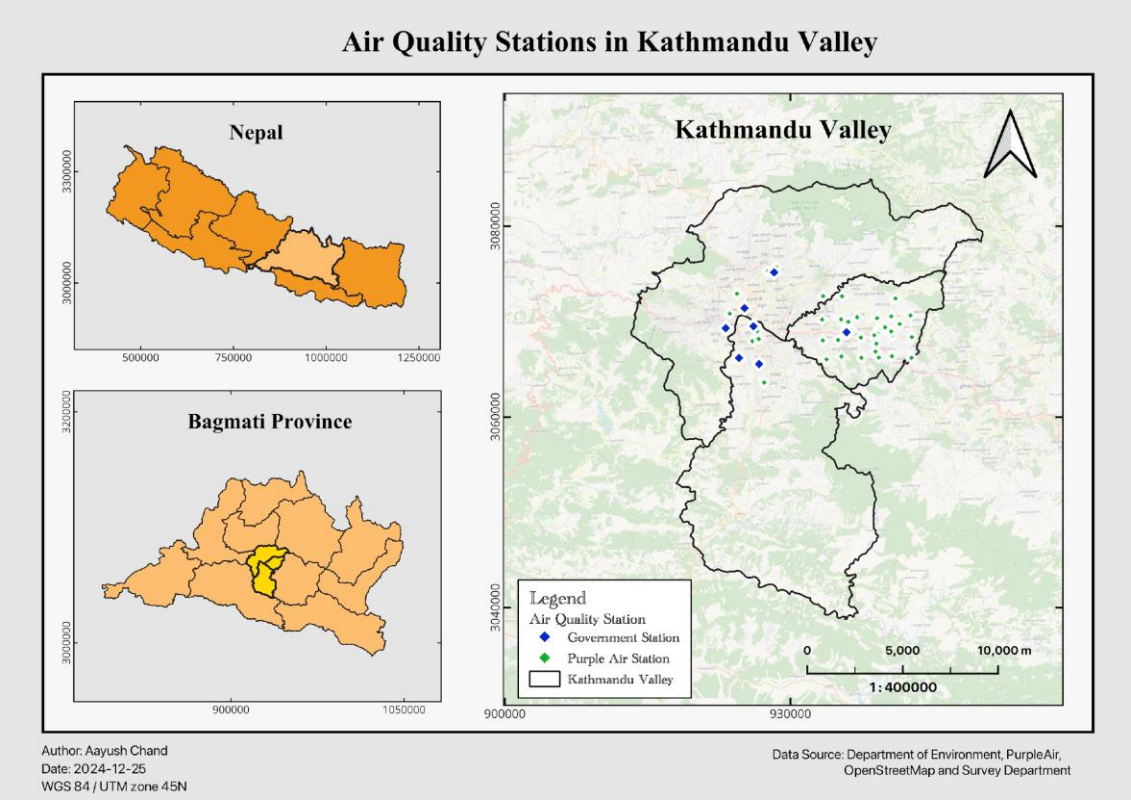


Figure 1.1: Study area

Chapter 2: Literature Review

Satellite remote sensing of trace gases and aerosols for air quality applications has a rich history.(Engel-Cox et al., 2004) Compared qualitative true color images and quantitative aerosol optical depth data from the Moderate Resolution Imaging Spectro-radiometer (MODIS) sensor on the Terra satellite with ground-based particulate matter data from US Environmental Protection Agency (EPA) monitoring networks. They covered the period from 1 April to 30 September 2002. Following were some of the interesting facts about this approach:1) Using both imagery and statistical analysis, satellite data enabled the determination of the regional sources of air pollution events, the general type of pollutant (smoke, haze, dust), the intensity of the events, and their motion. 2) Very high and very low aerosol optical depths were found to be eliminated by the algorithm used to calculate the MODIS aerosol optical depth data. 3) Correlations of MODIS aerosol optical depth with ground-based particulate matter were better in the eastern and Midwest portions of the United States (east of 100°W). Human activity negatively impacts the Earth through pollution that disrupts natural processes and environments, including sea and marine life pollution, global warming, acid rain and smog, thermal pollution, groundwater contamination, soil degradation, radioactive pollution, and even light and noise pollution affecting the natural world(Stahl & Nagy, 2015). There exist multiple types of pollution such as air pollution, sound pollution, light pollution, soil pollution, water pollution, etc. all of which hurt our environment to some extent. Air pollution is more harmful to human life because worldwide 5 million deaths were recorded due to air pollution and related activities. Air pollution has led to increased temperature fluctuations, as well as a rise in respiratory diseases, lung cancer, asthma, and skin-related issues(Zhu et al., 2020; Zoran et al., 2020b). According to the World Health Organization, ambient air pollutants are associated with heart disease, stroke, chronic obstructive pulmonary disease, lung cancer, and acute respiratory infections in children, which are responsible for about 4.2 million premature deaths every year globally. According to the World's Worst Polluted Places by the Blacksmith Institute in 2008, two of the worst pollution problems in the world are urban air quality and indoor air pollution. Air quality maintenance is the key focus of researchers, scientists, and policy-makers for sustainable planning and development of human life in addition to the environment. A study in China found a strong correlation between reduced human mobility and decreased air pollution in 44 Chinese cities, where the Intercity Migration Index (IMI) was used to calculate the results(Bao & Zhang, 2020). The first step in maintaining air quality is to assess the current air quality condition. Air pollution severity is commonly measured using the Air Quality Index (AQI), a widely recognized standard(Epa & of Air, 2014). The air pollution parameter considered is API (Air Pollution Index) or AQI (Air Quality Index), which can be defined as a scheme that transforms the weighted values of individual air pollution-related parameters (e.g., SO2 concentration or SPM) into a single number or set of numbers. The AQI is an essential tool for communicating air quality data and evaluating pollution levels, with higher AQI values signifying greater health risks(Anggraini et al., 2024). In studies on environmental pollution that focus solely on population distribution and pollution levels, notable differences in spatial distribution are observed(Caplin et al., 2019; Rivas et al., 2017;

Trewhela et al., 2019). Aerosol optical depth (AOD) is an air quality indicator observable through satellite remote sensing, representing the columnar aerosol content in the atmosphere. Several studies have established a positive correlation between satellite-derived AOD and surface-level particulate matter(Trewhela et al., 2019; You et al., 2015). A global study found that 69% of the total AOD are within the PBLH (Bourgeois et al., 2018), other studies have shown that temperature plays an important role in capturing AOD and understanding its vertical distribution that improves PM analysis (Liu et al., 2022). AOD-based methods have been widely utilized to estimate PM2.5 concentrations on regional (Y. Zheng et al., 2016), national (Ma et al., 2016) or global scale (van Donkelaar et al., 2010). Ground-based air quality monitoring is traditionally employed to determine AQI (Bechle et al., 2013b) due to its ability to provide precise surface-level data. However, the spatial resolution of such monitors is often limited leading to their placement predominantly in areas considered as potential pollution hotspots. Resolving spatial variability in ambient air pollutants and quantifying contributing factors are critical to human exposure assessment and effective pollution control. Due to its wide area coverage, remote sensing technology has become a promising alternative for reducing uncertainty in spatial distribution(Martin, 2008). Satellite remote sensing can capture integrated column concentrations of gases and aerosols from the Earth's surface to the top of the atmosphere and has been used to evaluate ground-level air pollution(Martin, 2008). While remote sensing primarily measures air pollution in the Earth's upper atmosphere with high sensitivity, these observations can still reflect variations in the distribution of air pollution on the surface(Filonchyk et al., 2021b; Sakti et al., 2023b; Z. Zheng et al., 2019b). The use of remote sensing in air pollution studies began in the 1970s, with the estimation of air pollution levels based on changes in the reflectance of ground objects observed in aerial photographs(Deng & Xiong, 2005). Sentinel-5P's lead instrument, TROPOMI, has a resolution that provides an unprecedented level of detail to be studied. This is high enough to be problematic, and special care is required to retrieve the height of the pollution layers from raw sources, however, the study of tropospheric column densities is less taxing (Efremenko & Kokhanovsky, 2021). NO_2 monitoring using TROPOMI is commonly found in academic studies; for instance, researchers utilized S5P NO_2 data to illustrate the changes in pollution across Europe caused by stay-at-home orders during the COVID-19 pandemic (Vîrghileanu et al., 2020). Several research studies have explored the relationship between Land Use/Land Cover (LU/LC), satellite reflectance, and air pollutants (Ahmad Jabatan Sains et al., n.d.; Deng & Xiong, 2005). Among various factors, vegetation has been identified as a negative contributor to air pollutants, with vegetation indices being used as criteria and indicators in urban air pollution studies (Manawadu and Samarakoon 2005; Hogda et al. 1995). Other indicators and predictor variables include geographic attributes, road length, proximity to roads, vehicle density, bus stop density, land-use data, building density, and population density (Gu et al., 2021). Using 6 years of meteorological and pollutant data, this research offered an ML approach for predicting PM2.5 concentrations from wind (speed and direction) and precipitation levels. The findings of the classification model showed good reliability in classifying low (10 g/m3) against high (>25 g/m3) PM2.5 concentrations, as well as low (10 g/ m3)

versus moderate (10–25 g/m3) PM2.5 concentrations (Kleine Deters et al., 2017). Satellite imagery has been applied for observing air pollutants since the early 1970s through the application of the Geostationary Operational Environmental Satellite (GOES), Advanced Very High-Resolution Radiometer (AVHRR), and Landsat. In addition, the Meteorological Operational satellite (MetOP), Aura, and Sentinel-5 precursor (Sentinel-5P) were among the additional satellite-based datasets that have been commonly applied on behalf of air quality observation since 1978. Regarding the literature, numerous investigators have deliberately observed, examined, and reclaimed air pollutants like Aerosol optical depth, SO2, NO2, CO, PM2.5, PM 10, CH4, and O3 by applying these RS-based satellites. During COVID-19, worldwide air pollutants are also observed to monitor the situation before, during, and after the COVID-19 lockdown. The GEE platform is commonly used to calculate air pollutants using Sentinel-5P and MCD19A2 data for major Indian cities during the period from April 2018 to 2021 (Prakash et al., 2021). In addition to physically based dispersion models and chemical transport models, data-driven modeling techniques have become widely adopted for their ability to capture complex data relationships, lower computational costs, and ease of implementation (James et al., 2021a). Air quality modeling results have been utilized in risk assessment of ambient pollutants (Lelieveld et al., 2015) but rarely in epidemiological studies because of their low accuracy and potential bias. Linear regression (LR) and random forest (RF) are two commonly used data-driven approaches to assess air pollutant variability. LR constructs a linear model using potentially predictive land-use variables to estimate pollution concentrations at various locations and has been applied globally to various pollutants, including in many metropolitan areas (Hoek et al., 2008; Kashima et al., 2009).(Beelen et al., 2013; Wang et al., 2013). However, LR is limited to capturing only linear relationships between the predictor and target variables (James et al., 2021a). The main drawback of RF is its lack of interpretability, as the model is not descriptive (James et al., 2021a). In the LR model, the predictor variables chosen for the final prediction are typically regarded as important, with their coefficients of partial determination, which indicate the percentage of explained variability, serving as the measure of importance (Beelen et al., 2013; Knibbs et al., 2014; Son et al., 2018; Wang et al., 2013). In the RF model, the importance of a predictor variable is measured by permutation importance, which evaluates the change in the model's score when the variable is randomly shuffled (K. Huang et al., 2018; Kamińska, 2018). In both LR and RF regression models, we demonstrate that including the stable variance measure of model performance alongside the mean value is essential for improving evaluation reliability. Additionally, to quantify the contributions of different predictor variables, we propose the use of Shapley Additive Explanations (SHAP) (Lundberg et al., 2017). To predict the air quality index of significant pollutants such as PM2.5, PM10, CO, NO2, SO2, and O3, they employed a variety of classification and regression approaches, including linear regression, SDG regression, and random forest regression. Evaluations were carried out using MSE, MAE, and R-SQUARE, which showed that ANN and SVM worked best for AQI prediction in New Delhi (Srivastava et al., 2019). Of the algorithms linear regression, decision tree regression, SVR, and RFR, the random forest regression algorithm yielded the best accuracy of

0.99985 on the test data with the least mean square error of 0.00013 and the mean absolute error of 0.00373 (Halsana, 2020). Logistic regression was used to assess whether the given data sample of daily weather and environmental conditions in a particular city indicated pollution or not (C R et al., 2018). Land-use regression (LUR) and Land-use land-cover classification (LULC) are two fields of remote sensing study among which statistical analysis techniques play a key role (land-cover being an indication of the Earth's physical characteristics and land-use referring to how these characteristics are used). Before the widespread availability of machine learning techniques in the early 2010s, traditional statistical methods were used for LUR and LULC tasks, often relying on linear or logistic regression models (Hoek et al., 2008). Earlier work applying ML techniques includes the use of Support Vector Machines (SVMs) to detect patterns in remotely sensed images (C. Huang et al., 2002). Linear regression was used as a machine learning algorithm to predict air quality for the next day using sensor data from three specific locations in the Capital City of India, Delhi, and the National Capital Region (NCR). The model's performance was evaluated using four metrics: MAE, MSE, RMSE, and MAPE. This paper focused on AQI prediction using data generated by IoT systems (Kumar et al., 2020). Monitoring air quality through measurement stations is valuable, but predicting air pollution in areas without monitoring stations is crucial for implementing preventive measures. Machine learning, a branch of artificial intelligence, has proven effective in environmental research by enabling accurate air quality prediction and forecasting (Q. Wu & Lin, 2019). As noted by Gorelick et al. (2017), GEE's data catalog contains a repository of publicly available geospatial datasets, including observations from various satellite and aerial imaging systems in optical and non-optical wavelengths environmental variables, weather and climate forecasts and hindcasts, land cover, topographic, and socioeconomic datasets. All of this data is pre-processed into a format that preserves information, enabling efficient access and eliminating many of the challenges associated with data management (Gorelick et al., 2017). A multimodal artificial intelligence architecture was used to estimate NO_2, O_3, and PM_{10} air pollution in the UK and Ireland by integrating Sentinel-5P, Sentinel-2, and local field station data (Rowley & Karakuş, 2023).

Chapter 3: Materials And Methods

3.1 Description of Datasets

To determine the concentration of PM2.5 and PM10 in the Kathmandu Valley, we utilized data from both remote sensing and ground-based observations (Table 1) spanning two years, from January 1, 2022, to December 31, 2023. Our data acquisition strategy focused on obtaining reliable air quality monitoring data from the Government of Nepal, Ministry of Population and Environment, Department of Environment, which operates seven air quality monitoring stations across the Kathmandu Valley. These stations provided vital data on various particulate matter, including PM2.5 and PM10. Historical data from these stations were accessed through the Department of Environment's Air Quality Monitoring website.

The temporal range for data acquisition extended from January 1, 2022, to December 31, 2023. These two years provided sufficient data to capture seasonal variations, long-term trends, and short-term pollution spikes caused by events such as wildfires, dust storms, industrial accidents, crop burning, and traffic congestion.

The satellite-based air pollution data used in this study were sourced from the MCD19A2.061 MODIS Terra and Aqua MAIAC satellites via the Google Earth Engine Catalog. This dataset enabled the extraction of aerosol optical depth (AOD) information within the study area. Leveraging the MCD19A2.061 product, derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) instruments aboard the Terra and Aqua satellites, the study accessed a daily AOD dataset with a spatial resolution of 1 kilometer. Distributed by NASA's Land Processes Distributed Active Archive Center (LP DAAC), MCD19A2.061 utilized the Multi-angle Implementation of Atmospheric Correction (MAIAC) algorithm, offering daily global coverage for analysis of short-term aerosol variations. This product provided additional parameters such as AOD uncertainty, cloud information, and solar/viewing geometry data, enhancing aerosol characterization and atmospheric correction procedures.

In addition to pollution data from satellite imagery, we incorporated other environmental factors crucial to air quality variation. Environmental factors included the Normalized Difference Vegetation Index (NDVI) from MODIS satellite data, with a spatial resolution of 1 km and a temporal resolution of 1 month. Furthermore, we incorporated land surface temperature (LST) data from MODIS MOD11A1V6.1, which provided daily LST and emissivity values with 1 KM spatial resolution. These temperature values were derived from the MOD11_L2 swath product. In regions above 30 degrees latitude, some pixels had multiple observations meeting clear-sky criteria. In such cases, the pixel value represented the average of all qualifying observations. The dataset included both daytime and nighttime surface temperature bands, along with their quality indicator layers (MODIS bands 31 and 32 and six observation layers). Precipitation data were obtained from the CHIRPS dataset of the "UCSB-CHG/CHIRPS/DAILY" image collection, having a spatial resolution of 5.56 kilometers and a daily temporal resolution. Soil moisture data were sourced from the SMAP dataset of the

"NASA/SMAP/SPL4SMGP/007" image collection, having a spatial resolution of 9 kilometers and resampled to 1 kilometer and temporal resolution of 3 hours. The soil moisture data was selected with a 3-hour temporal resolution to account for its dynamic nature, as soil moisture levels fluctuate throughout the day based on sunlight and other environmental factors. This resolution was specifically chosen to align the soil moisture data with the exact timestamps of the aerosol optical depth (AOD) data captured by the MODIS sensor over the study area. By synchronizing the temporal resolutions, we aimed to enhance the accuracy of the soil moisture data, ensuring that it corresponded closely to the conditions present during the AOD measurements. This alignment was critical for improving the reliability of the subsequent analysis and modeling.

Table 3.1: Description of Datasets

Datasets	Data Source	Data Type	Spatial Resolution	Temporal Resolution	Description
Shapefile	Survey Department	Vector	-	-	Boundary of Kathmandu Valley
Ground PM2.5 and PM10	Department of Environment	CSV	-	1 Day	Ground PM2.5 and PM10 concentration of selected stations
AOD	MODIS	Raster	1 Km	1 Day	Aerosol Optical Depth
LST	MODIS	Raster	1 Km	1 Day	Land Surface Temperature
NDVI	MODIS	Raster	1 Km	1 Month	Vegetation Index
Soil Moisture	SMAP	Raster	9 Km (Resampled to 1 Km)	3 Hour	Moisture Content
Precipitation	CHIRPS	Raster	5.56 Km	24 Hour	Precipitation information

3.2 Methodology

The main aim of our study is to establish a machine learning model to estimate and predict the ground-level air pollution in Kathmandu Valley from 2019 to 2022 through heterogeneous data such as satellite and meteorological data. In particular, we focused on the determination of ground-level concentrations of PM2.5 and PM10 through a machine-learning approach, presented in Figure 2.

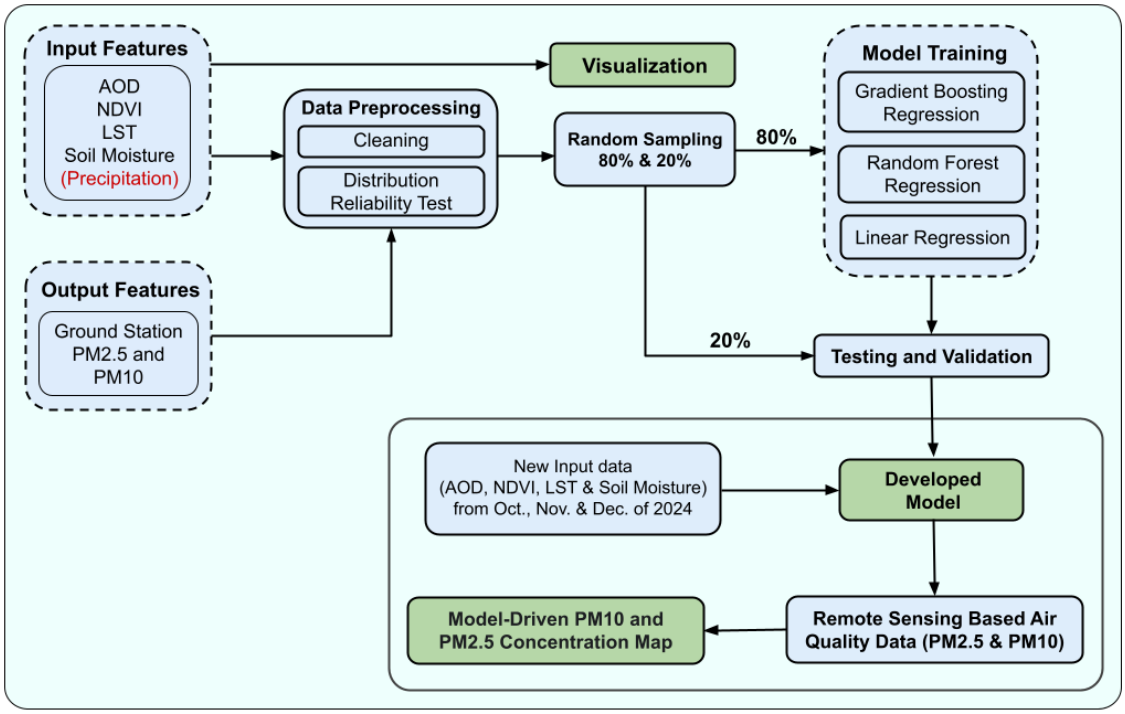


Figure 3.2: Methodological flowchart used for estimation of PM2.5 and PM10

3.2.1 Data Extraction, Cleaning, Processing, and Analysis

In this work, the ground-based PM2.5 and PM10 data were collected as daily averages, ensuring consistency and reliability in temporal resolution for analysis. The AOD data used in this study was sourced from MODIS, Terra's satellite, and similarly had a daily temporal resolution. This alignment of temporal resolution between ground-based and satellite-derived AOD data minimized discrepancies during analysis. Land Surface Temperature (LST) data, another critical variable, was also obtained with a daily temporal resolution, enabling consistent comparisons across datasets.

The NDVI dataset, however, had a temporal resolution of one month, captured precisely at the start of each month. To address this temporal disparity, NDVI values were interpolated by assigning the value of the first day of the month to all days within 15 days before and after this date. This method ensured that NDVI values represented a consistent vegetation index across the month while maintaining computational simplicity. Soil moisture data presented a unique challenge, as it had a higher temporal resolution of three hours, resulting in eight data points per day. To ensure that the soil moisture data aligned with the temporal resolution of the other variables, the value closest to the time when AOD data was extracted was selected. The remaining seven values for each day were excluded from the analysis. For spatial alignment, point values for each remote sensing dataset were extracted for the exact pixel in which the ground station was located, ensuring spatial correspondence between the datasets. Once all datasets were successfully extracted, they were combined by matching ground station locations and observation times. During preprocessing, missing values and duplicates were identified and removed using Python's Pandas library to enhance data quality and consistency. We detect and remove outliers, which are data points that vary outside of the range of three standard

deviations from the mean. To ensure uniformity across different variables, data scaling and normalization techniques were applied. These preprocessing steps collectively enhanced the dataset's suitability for model development.

To better understand the dataset and identify patterns, various visualization techniques were employed. Scatter plots, boxplots, and Kernel Density Estimation (KDE) plots were created to examine data distribution and detect potential issues such as skewness or anomalies. For instance, an important insight emerged from the visualization of precipitation data, which revealed that its values were zero more than 75% of the time. This lack of variability indicated that precipitation would likely not contribute significantly to the model and was, therefore, excluded from further analysis.

Furthermore, a correlation heatmap was plotted to examine the relationships between predictor and target variables. This heatmap revealed valuable insights into variable dependencies. The correlation among predictor variables was found to be weak, meaning that multicollinearity was not a concern. As a result, feature scaling was deemed unnecessary. After careful evaluation, all variables except precipitation were retained for further analysis as they showed potential relevance to the target variables. When analyzing the relationships between independent and dependent variables, AOD emerged as a particularly important factor, exhibiting the strongest positive correlation with PM2.5 and PM10 concentrations. This strong association underscored the importance of AOD as a predictive variable in understanding air quality.

Through careful data extraction, cleaning, and preprocessing, the study established a solid foundation for the subsequent modeling and analysis of PM2.5 and PM10 levels. Each step, from aligning temporal resolutions to identifying and addressing data anomalies, was carefully designed to verify that the datasets were accurate, consistent, and insightful, paving the way for efficient predictive analysis.

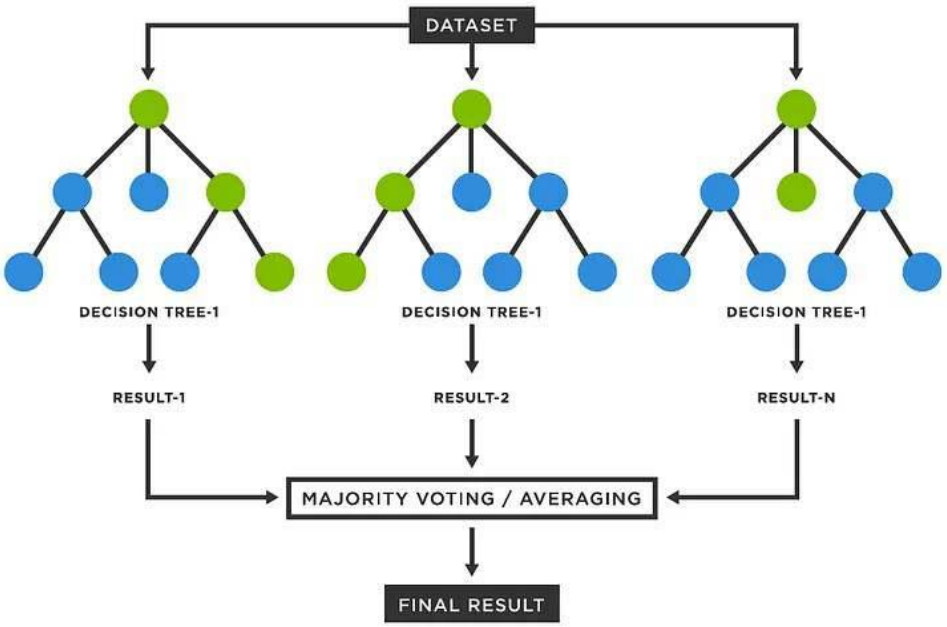
3.2.2 Model Development

This study's primary motivation is to estimate the ground PM2.5 and PM10 values from the remote sensing observations. To achieve this, we leverage 3 algorithms: Linear Regression, Random Forest, and Gradient Boosting models.

Simple Linear Regression (SLR) estimates the relationship between a single independent variable (X) and a dependent variable using a straight line. It's the foundation for more complex models. Multiple Linear Regression (MLR) extends SLR to model the relationship between a dependent variable and multiple independent variables (X_1, X_2, ..., Xn). It captures the combined effect of all predictors on Y.

Equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$

It has similar notation as SLR, but with multiple independent variables (X_1, X_2, ..., Xn) and corresponding slopes (β_1, β_2, ..., β_n).



Simple Random Forest Regression (SRFR) is less common than SLR. SRFR assumes the error term (ε) in the linear regression model follows a normal distribution with zero mean and constant variance (homoscedasticity). It's often used as a theoretical concept or for specific statistical tests.

Multiple Random Forest Regression (MRFR) is not a linear regression model but a powerful ensemble machine learning method. It combines the predictions of multiple decision trees to create a more robust and accurate model, especially for complex relationships between variables. No single equation represents a random forest model. It's built from a collection of decision trees, each making predictions based on splitting rules learned from the data. The final prediction is an average or weighted average of the individual tree predictions.

Gradient Boosting is a powerful machine learning technique widely used for regression and classification problems. It is an ensemble method that builds models sequentially, where each new model corrects the errors of the previous one. By combining the strengths of multiple weak learners, usually decision trees, Gradient Boosting creates a robust model capable of capturing complex patterns in the data. This method is particularly effective when dealing with structured data, as it focuses on reducing bias and variance to improve prediction accuracy.

After completing the preprocessing steps, a total of 987 rows of clean and consistent data remained, ready for model development. This dataset was split into training and testing subsets to evaluate model performance. Specifically, 80% of the data was allocated for training the models, ensuring they could learn patterns and relationships within the dataset, while the remaining 20% was reserved for testing purposes, allowing us to validate the models' predictive capabilities.

Three models were developed for this study: Linear Regression, Random Forest, and Gradient Boosting. Separate models were created to predict PM2.5 and PM10 concentrations. Each model was carefully designed to capture the relationships between the predictor variables and the target variables, leveraging their unique strengths.

To enhance the models' performance, hyperparameters were fine-tuned for each model using a trial-and-error technique. This involved systematically adjusting key parameters, such as the learning rate and the number of estimators for Gradient Boosting, to find the optimal combination that delivered the best results. Fine-tuning ensured that the models were neither underfitted nor overfitted, striking a balance between generalization and precision.

Finally, the fine-tuned models were used to predict PM2.5 and PM10 concentrations using the test datasets. This step allowed us to evaluate the models' accuracy and reliability in predicting air quality parameters. By comparing the predicted values with the actual test data, we could determine how well each model performed and assess their potential for real-world applications.

3.2.3 Testing and Validation

Out of the total cleaned and processed data, 80% of the data was used for training the models, and 20% of the data was used for testing and validation of the models. We applied random sampling to split our data into training and testing sets. This technique ensures that each individual has an equal chance of being chosen, creating representative samples and minimizing bias. Random splitting involves dividing the dataset into a training set, used for model training, and a testing set, used for performance evaluation. This method ensured both sets represented the overall data distribution, providing reliable inferences about the population. Root-mean square error (RMSE), Mean Square Error (MSE), R-squared (R^2), and adjusted R-squared methods were used for accuracy assessment. If y is the predicted value, $\hat{y}$ is the observed value (AQI-S), n is the total of data, and v is the residual, then:

$$RMSE = \sqrt{\frac{\sum (y-\hat{y})^2}{n}} = \sqrt{\frac{\sum v^2}{n}} \qquad \textit{Equation 3.1}$$

The MSE measured the average of the squared differences between predicted and observed values, providing a more severe penalty for larger errors. It was expressed as:

$$MSE = \frac{\sum (y-\hat{y})^2}{n} \qquad \textit{Equation 3.2}$$

One of the key accuracy metrics, R-squared (R^2) was calculated to measure prediction accuracy.

$$R^2 = 1 - \frac{\sum_i (y_i-\hat{y}_i)^2}{(y_i-\underline{y})^2} \qquad \textit{Equation 3.3}$$

where:
$\underline{y}$ = Mean of observed values

The Adjusted R^2 improved upon R^2 by accounting for the number of predictors in the model. It prevented overfitting by adjusting for model complexity. Its formula was:

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)} \qquad \textit{Equation 3.4}$$

where:
R^2 = Adjusted coefficient of determination
N = Total number of observations
K = Number of predictors

The MAE calculated the average of absolute errors between the observed and predicted values. It was less sensitive to outliers compared to RMSE and was defined as:

$$MAE = \frac{\Sigma \quad |y-\hat{y}|^2}{n} \qquad \textit{Equation 3.5}$$

where:
$|y - \hat{y}|$ = Absolute error of each prediction

3.2.4 Visualization of Prediction Results

As part of this study, our developed models were applied to predict PM2.5 and PM10 concentrations across Kathmandu Valley. For the visualization phase, recent input data, including Aerosol Optical Depth (AOD), Soil Moisture, Normalized Difference Vegetation Index (NDVI), and Land Surface Temperature (LST), were collected at 10-day intervals during October, November, and December 2024. Predictions were generated at 935 evenly spaced points, each 1 km part, using both trained models, and the final PM2.5 and PM10 values were obtained by averaging the outputs from both approaches. The predicted values were then mapped using a color scale ranging from white for lower concentrations and red for higher concentrations, effectively capturing spatial pollution variations. These visualizations help to highlight pollution hotspots and temporal trends, offering valuable insights into air quality distribution. This approach demonstrates the capability of the model in identifying high-risk zones and aiding policymakers in designing targeted air quality management strategies.

Chapter 4: Result

4.1 Data Visualization

To represent the spatial and temporal variations of air quality and associated environmental parameters, we developed comprehensive maps for each parameter: Aerosol Optical Depth (AOD), Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), and Soil Moisture. These maps were created for four distinct four-month intervals spanning the years 2022 and 2023, providing a seasonal perspective of air quality dynamics in the Kathmandu Valley.

4.1.1 Aerosol Optical Depth (AOD)

The average AOD was extracted from the MCD19A2.061 MODIS Terra and Aqua MAIAC satellite data using the Google Earth Engine Catalog. These datasets, with a 1 km spatial resolution, were averaged over the specified intervals to visualize the concentration of aerosols across the study area.

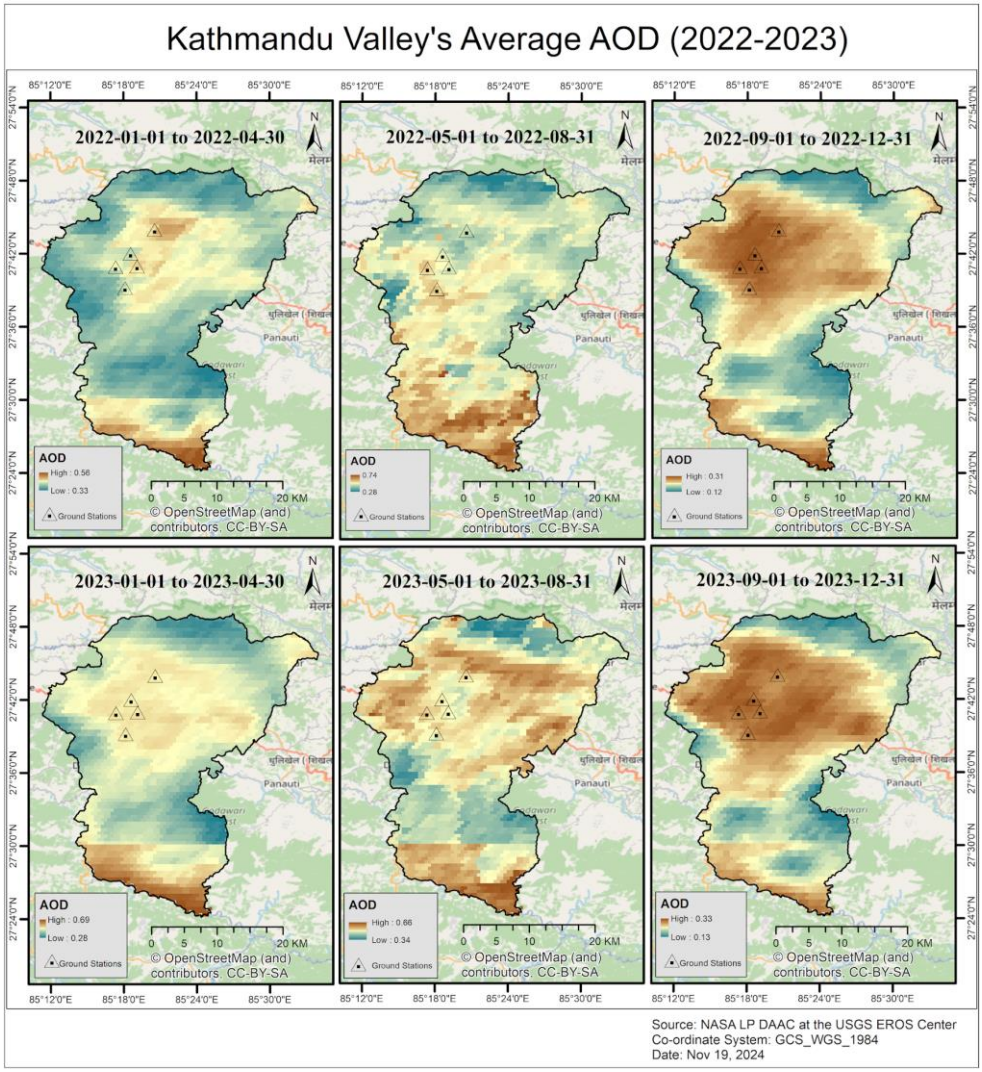


Figure 4.1.1: Average AOD of Kathmandu Valley (2022-2023)

4.1.2 Land Surface Temperature (LST)

LST data were derived from the MODIS MOD11A1 V6.1 dataset, which provides daily temperature and emissivity values with a 1 km spatial resolution. The data were aggregated to represent the average LST for each four-month interval. These maps help correlate temperature variations with aerosol distributions, emphasizing seasonal effects on air quality.

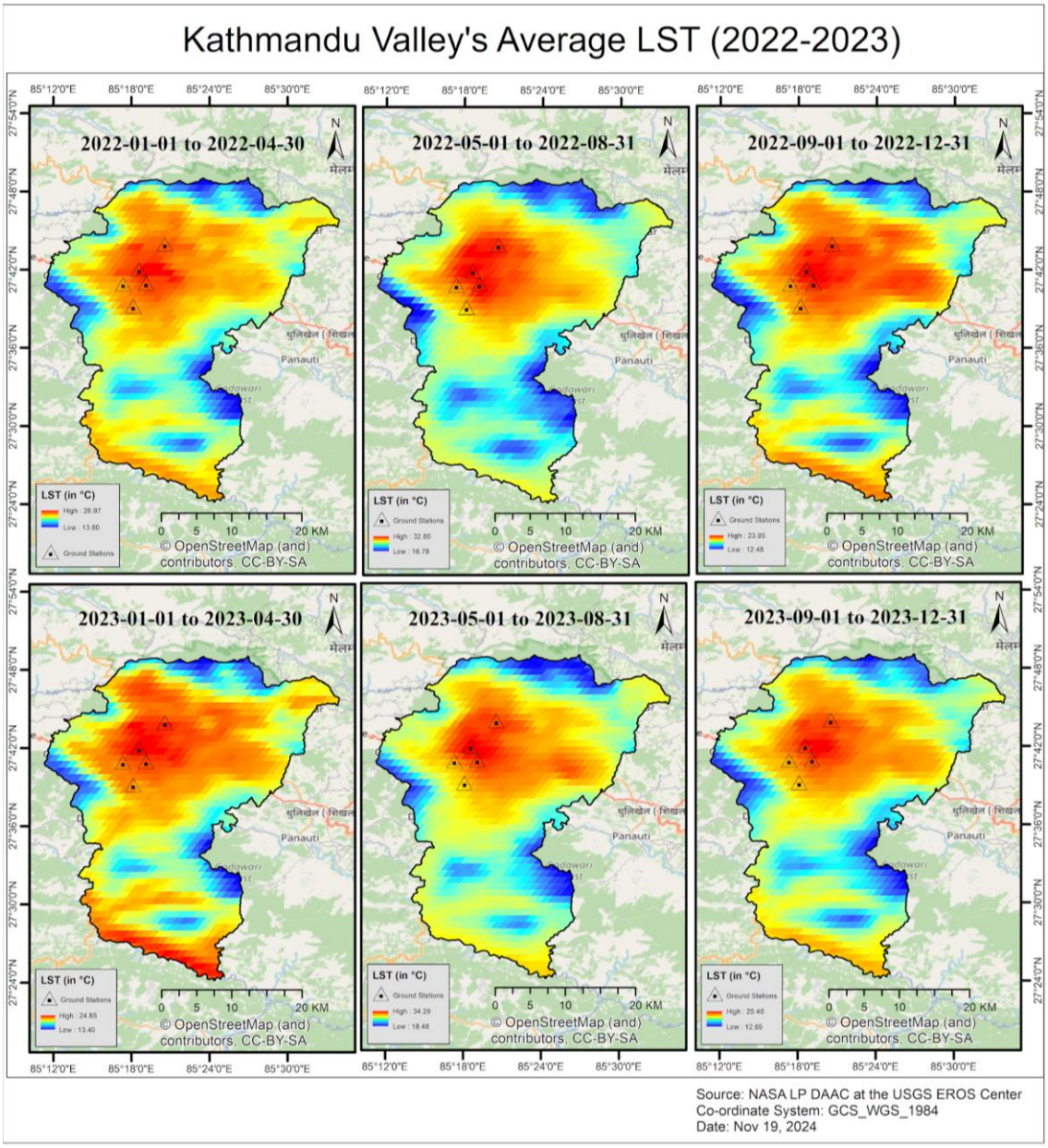


Figure 4.1.2: Average LST of Kathmandu Valley (2022-2023)

4.1.3 Normalized Difference Vegetation Index (NDVI)

NDVI data were sourced from MODIS satellite data at a 1 km spatial resolution and a monthly temporal resolution. The interval-averaged NDVI maps provide insights into vegetation coverage and health, factors that influence air quality by regulating particulate deposition and secondary aerosol formation.

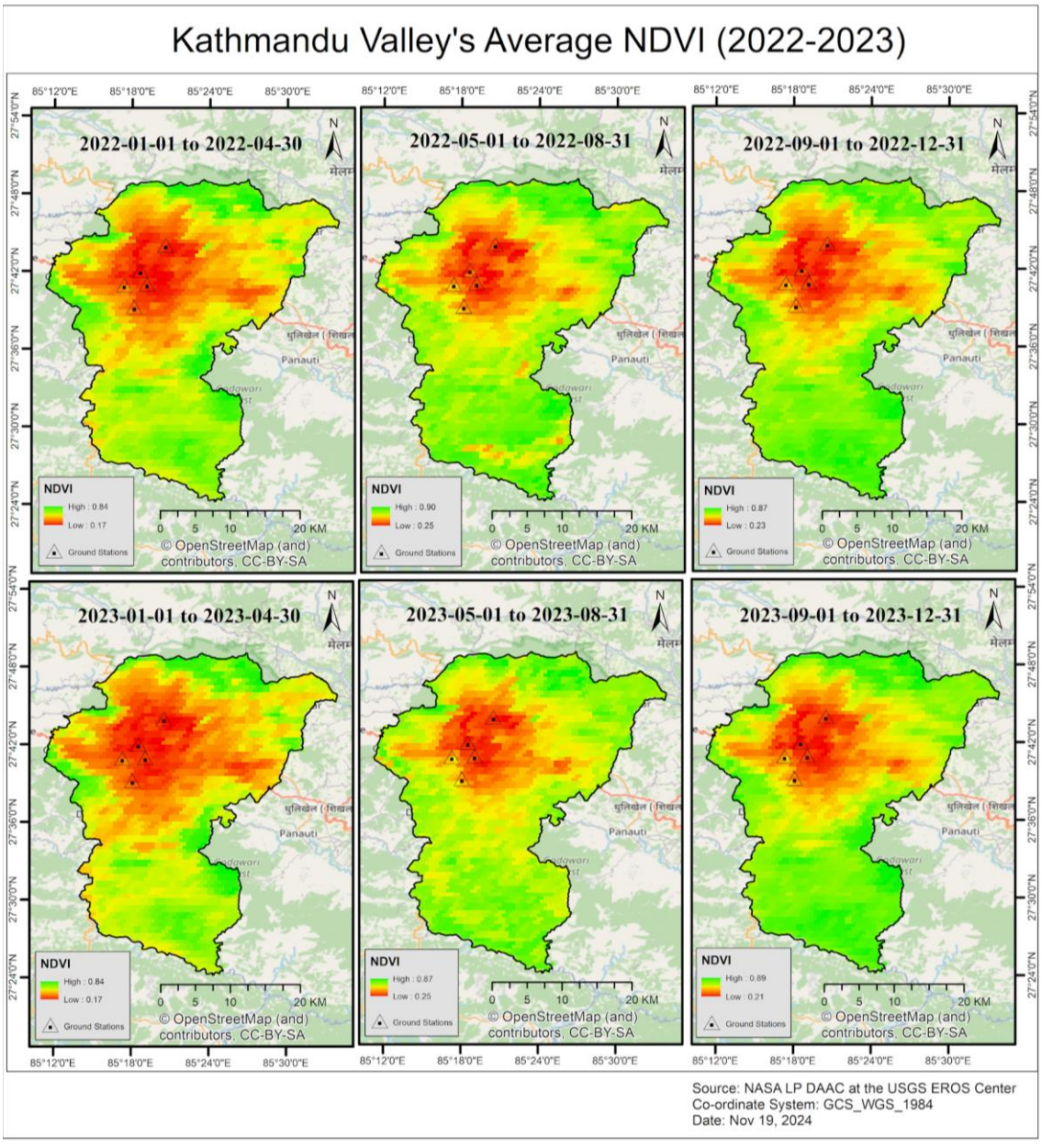


Figure 4.1.3: Average NDVI of Kathmandu Valley (2022-2023)

4.1.4 Soil Moisture

Soil moisture data were obtained from the SMAP dataset ("NASA/SMAP/SPL4SMGP/007"), originally available at an 11 km resolution. These data were resampled to a 1 km spatial resolution and aggregated for four-month intervals. The resulting maps highlight the influence of soil moisture on surface-level air quality by affecting aerosol resuspension and deposition dynamics.

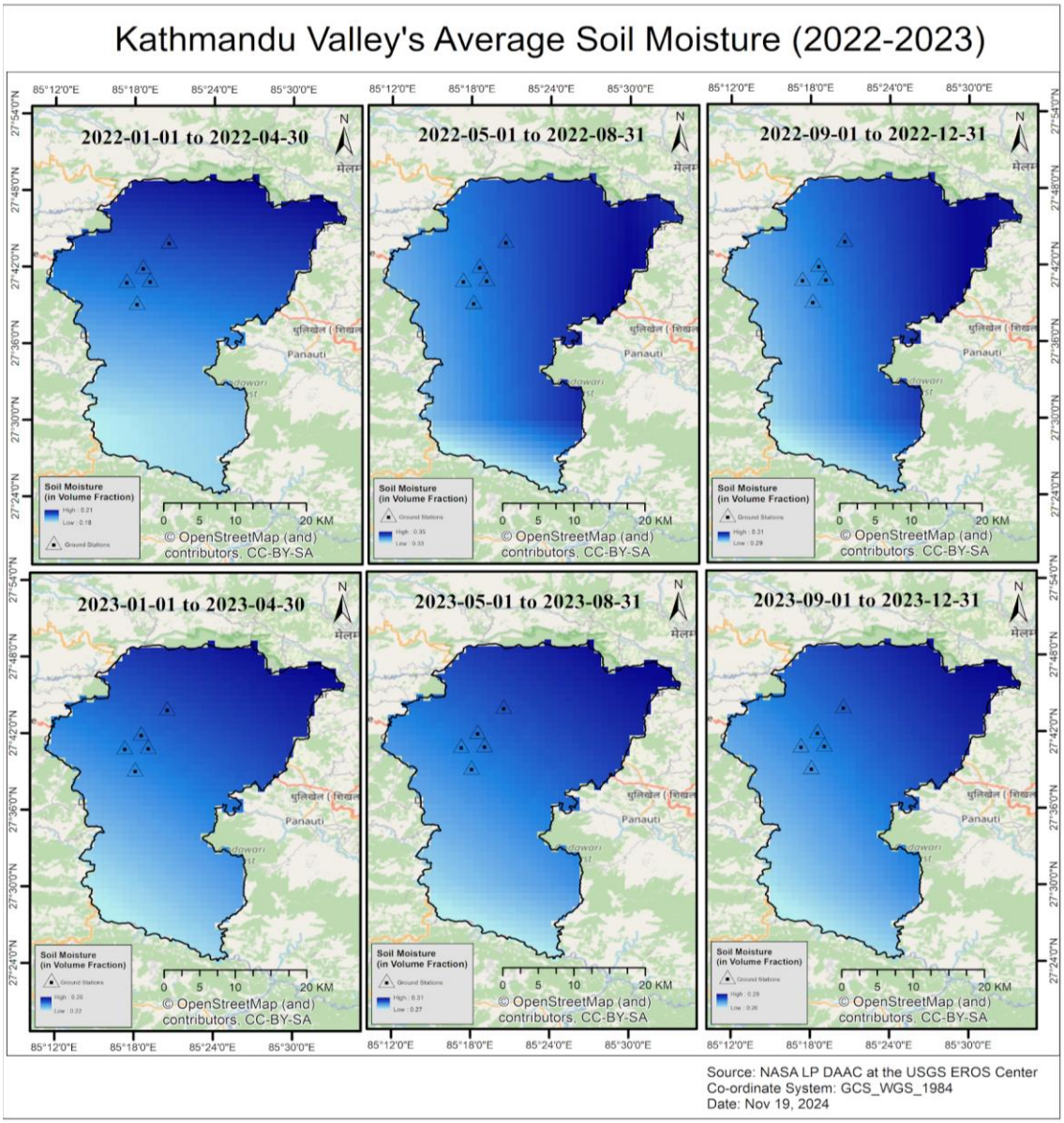


Figure 4.1.4: Average Soil Moisture of Kathmandu Valley (2022-2023)

By integrating these environmental variables, the study provides a holistic understanding of air quality variation in the Kathmandu Valley, aiding in the identification of high-risk zones and contributing factors.

4.2 Statistical Evaluation

Figure 4.5 shows the data distribution of predictor and target variables through the KDE plots. The data distribution of the AOD is positively skewed, with most values concentrated between 0.2 and 0.7, with a long tail expanding up to nearly 2. This indicates that lower AOD values dominate, representing clearer atmospheric conditions most of the time, while higher AOD values are relatively rare. The distribution of LST is close to a normal distribution, with values predominantly between 15°C and 35°C, peaking around 27°C. The distribution of LST shows moderate variability in temperature, with most regions having mild to warm surface temperatures most of the time. The NDVI values are moderately skewed towards lower values, concentrating between 0.2 and 0.6, with fewer values closer to 0.7. This represents sparse variation in most areas, indicating dense vegetation is less common in the study area. The precipitation data values are strongly right-skewed, with more than 75% of values being zero, due to which precipitation data were not utilized for further analysis, considering its less significance in model development. The distribution of soil moisture is close to a normal distribution, with most values concentrated between 0.2 and 0.4, peaking around 0.3, indicating fairly uniform soil moisture levels across the regions with moderate variations. The distribution of PM2.5 is close to a normal distribution with a slight right skew. Most of the values lie between 50 and 150 µg/m³, reflecting moderately high levels of PM2.5, which could indicate significant air pollution in the regions analyzed. The distribution of PM10 is right-skewed, with values concentrated between 50 and 100 µg/m³, with fewer values exceeding 200 µg/m³. This indicates a higher concentration of coarse particulate matter in some areas, but most values suggest moderately polluted air.

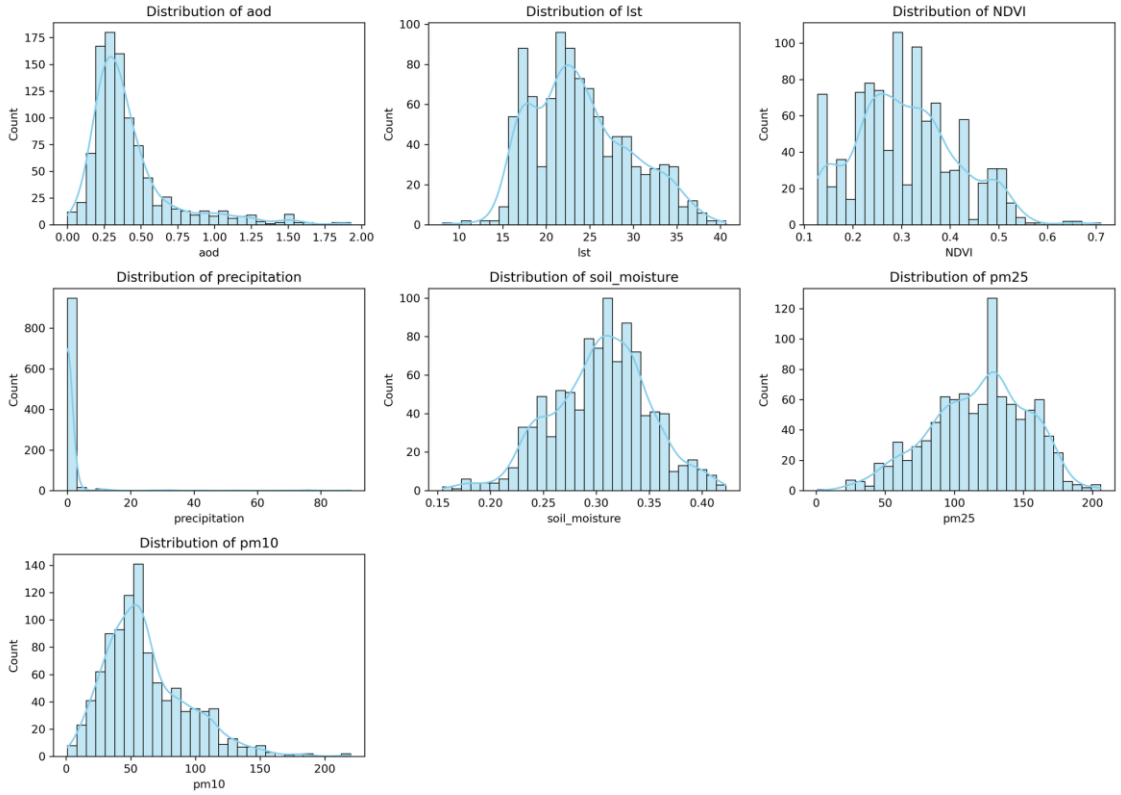


Figure 4.3: Data distribution of predictor and target variables

The boxplots in Figure 4.6 also show the distribution of the environmental variables. The analysis of these boxplots provides insights into the central tendencies, variabilities, and presence of outliers for each parameter. The median AOD value lies close to 0.3, and the interquartile range is between 0.2 and 0.5, suggesting variability in aerosol concentration within this range. A significant number of outliers are present above 0.5, representing occasional high aerosol loads. The distribution of LST shows the median temperature is approximately 25°C, and the IQR spans from 20°C to 30°C. A few outliers exist beyond 40°C, indicating rare occurrences of extreme surface temperatures. The IQR range of NDVI values ranges from approximately 0.2 to 0.4, with a median NDVI around 0.3, pointing to sparse vegetation coverage in the study area. A small number of outliers are above 0.6, indicating small areas with higher vegetation density. The median value of precipitation is 0, and the IQR is tightly clustered at low values, showing consistent precipitation levels. Numerous outliers extending beyond 80 mm suggest occasional heavy rainfall events. The median soil moisture value is close to 0.3, reflecting moderate moisture levels across the study area. The IQR spans from 0.25 to 0.35, indicating consistent soil moisture conditions. A few outliers below 0.15 represent drier regions. The IQR of PM10 ranges between 50 and 100 µg/m³, with a median value of approximately 75 µg/m³, indicating moderate air pollution levels. Outliers above 150 µg/m³ reflect occasional severe air pollution. The median PM2.5 value is around 100 µg/m³, highlighting significant particulate matter pollution. The IQR ranges from 75 to 150 µg/m³. Outliers below 50 µg/m³ represent occasional lower air pollution levels.

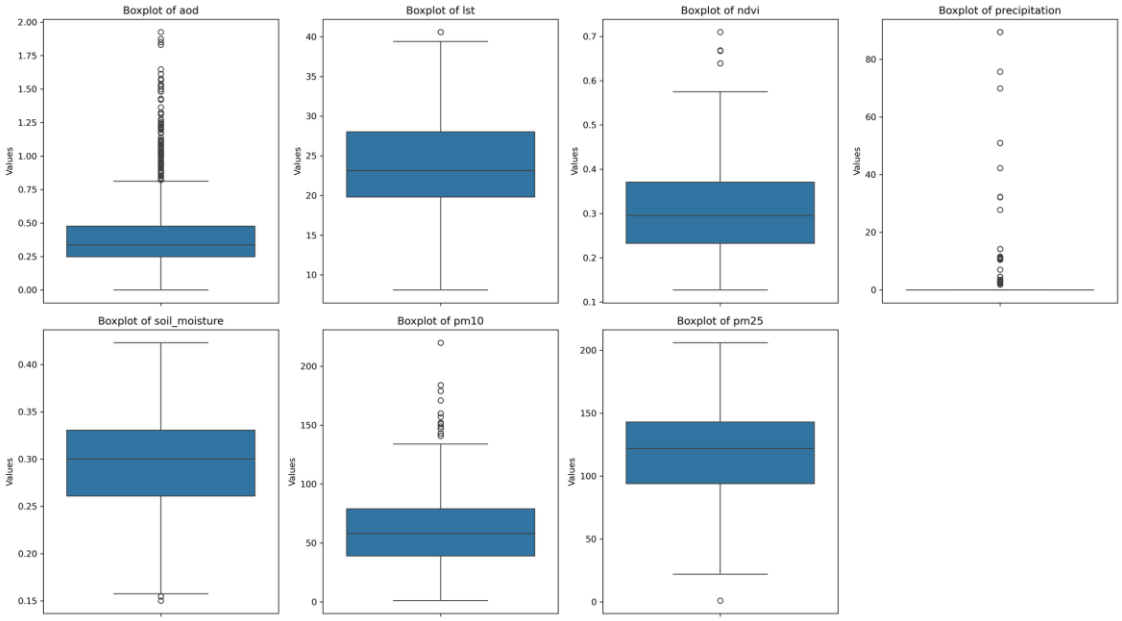


Figure 4.4: Boxplots of predictor and target variables

The correlation heatmaps in Figure 4.7 provide a visual representation of the relationships between variables, measured using Pearson's correlation coefficient. The values range from -1 to 1, where 1 represents a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation between variables. The AOD shows a maximum positive correlation with PM10 (0.53) and PM2.5 (0.49). This suggests that,

as AOD increases, PM concentrations also tend to increase. AOD has a slight positive correlation with LST (0.32), indicating that areas with higher aerosol levels might experience slightly higher surface temperatures. LST shows a moderate positive correlation with NDVI (0.39), indicating that vegetation density is somewhat related to surface temperature. LST shows a weak negative correlation with PM2.5 (-0.08). NDVI shows a weak negative correlation with PM2.5 (-0.22) and PM10 (-0.05), suggesting that areas with denser vegetation may have slightly lower PM levels, possibly due to vegetation acting as a filter to pollutants. Soil moisture exhibits a negative correlation with PM10 (-0.43) and PM2.5 (-0.43), suggesting that higher soil moisture may be associated with lower particulate matter levels, likely due to reduced dust generation. It is weakly negatively correlated with AOD (-0.30), indicating that aerosols might decrease with increasing soil moisture. PM10 and PM2.5 have a strong positive correlation (0.77), indicating that these two variable metrics are closely related and often change together.

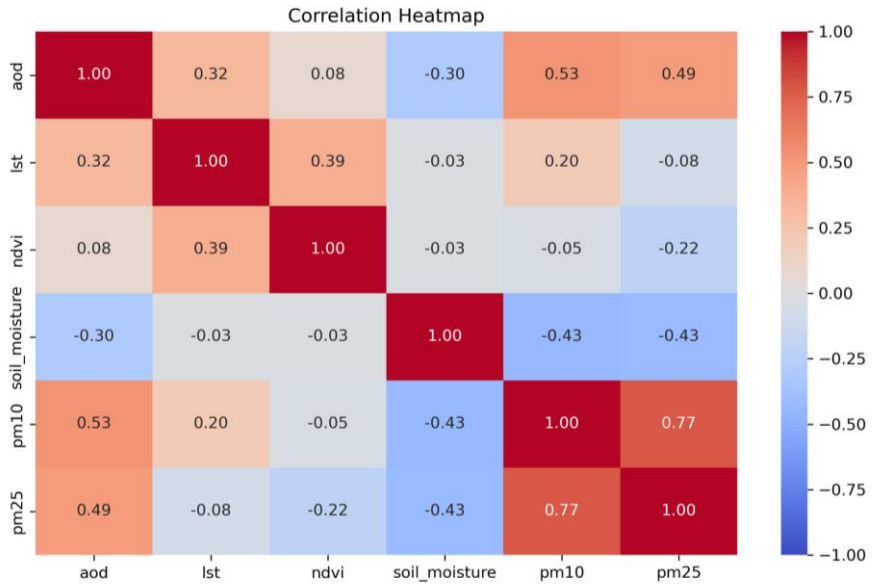


Figure 4.5: Correlation Heatmap

4.3 Linear Regression

Figures 4.8 and 4.9 demonstrate the performance of linear regression models. In Figure 4.8, the scatter plot of predicted versus actual values, with the red line representing the regression line and blue dots representing data points, highlights the inaccuracies of the model. The R^2 score for PM2.5 is 0.38, indicating that the model explains only 38% of the variance in PM2.5 concentrations. This low value suggests that a substantial portion of the variability remains unexplained by the model. Moreover, the error metrics show considerable deviations, with a Root Mean Squared Error (RMSE) of 27.80 and a Mean Absolute Error (MAE) of 23.38, underscoring the model's inability to make accurate predictions. For PM10 predictions, the results are similarly concerning. The scatter plot for PM10 also exhibits weak predictive accuracy, as evidenced by an R^2 score of 0.36.

This score reflects the model's failure to explain 64% of the variance in PM10 concentrations. The RMSE and MAE for PM10 are 23.41 and 19.59, respectively, further indicating significant errors in the model's predictions. The scatter points show poor alignment with the regression line, demonstrating an inconsistent relationship between predicted and actual values.

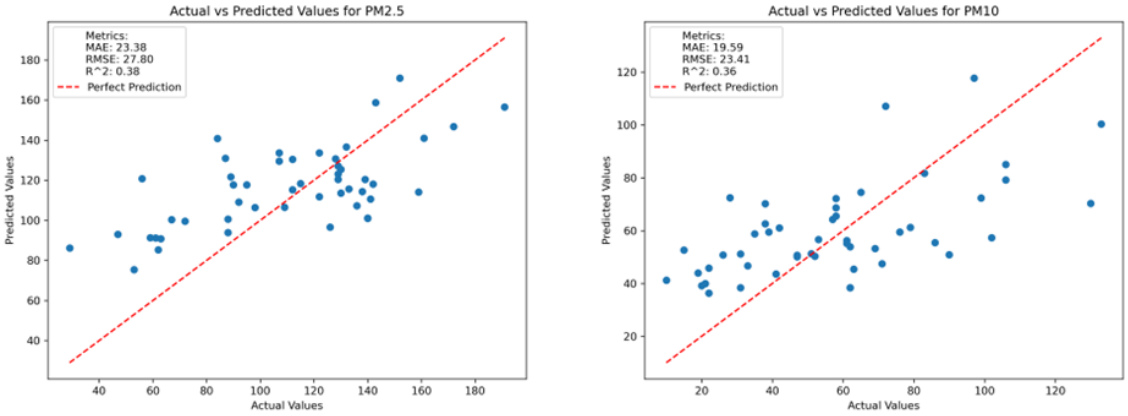


Figure 4.6: Scatter Plot of Actual vs Predicted Values (Linear Regression)

The cross-validation results in Figure 4.9 provide additional evidence of the model's limitations. For PM2.5, the metrics—R^2, MAE, and RMSE—fluctuate drastically across the folds, as seen in the left panel of the figure. The R^2 values range from negative numbers to over 60, reflecting extreme inconsistency. Although MAE and RMSE appear relatively stable compared to R^2, they still exhibit noticeable variations. Similar trends are observed in the cross-validation results for PM10 in the right panel. The R^2 values for PM10 fluctuate significantly, while MAE and RMSE show varying degrees of error across folds. These fluctuations suggest that the model's performance is highly sensitive to the training data, making it unreliable for generalization.

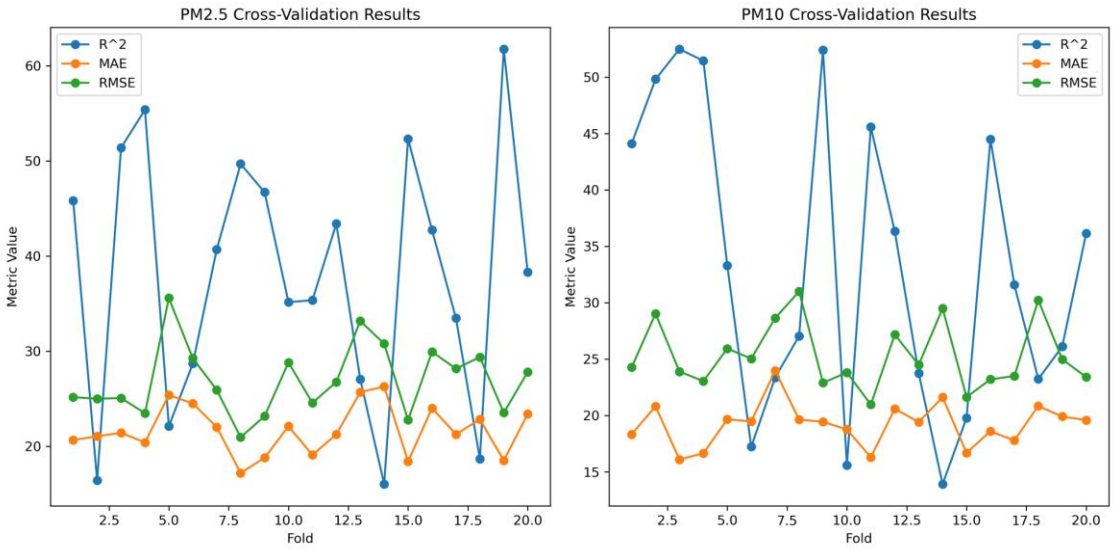


Figure 4.7: Cross-Validation Results (Linear Regression)

In conclusion, the results demonstrate that the linear regression model is unsuitable for predicting PM2.5 and PM10 concentrations using the given remote sensing-based environmental variables in our case. The low R^2 scores, high error metrics, and inconsistent cross-validation performance collectively indicate that the model struggles to establish a meaningful relationship between the predictors and the target variables. These findings suggest the need for more advanced modeling techniques, such as non-linear models or machine learning algorithms, to better capture the complex interactions in the data. Exploring improved data preprocessing, feature selection, and alternative approaches could lead to more accurate and reliable predictions.

4.4 Random Forest Model

The Random Forest model demonstrates strong performance in predicting PM2.5 concentrations. The R^2 score of 0.8 indicates that the model explains 80% of the variance in PM2.5 data, showcasing its ability to capture the complex relationships in the dataset. The error metrics further highlight the model's accuracy, with a Mean Absolute Error (MAE) of 10.98 and a Root Mean Squared Error (RMSE) of 15.16. These metrics suggest that the predictions are both precise and reliable, minimizing deviations from the observed values. For PM10 concentrations, the Random Forest model exhibits similar results. The R^2 score of 0.82 signifies that the model explains 82% of the variance in PM10 data, demonstrating a strong correlation between predicted and actual values. The error metrics are also favorable, with an MAE of 9.31 and an RMSE of 15.05, indicating a high degree of accuracy and consistency in the predictions.

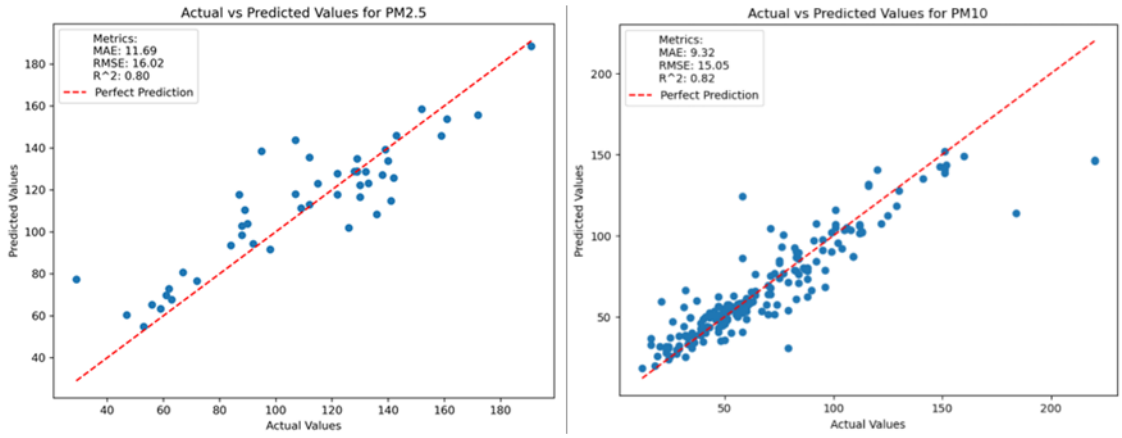


Figure 4.8: Actual vs Predicted Values (Random Forest)

The cross-validation results provide further evidence of the model's reliability. For PM2.5, the metrics are stable across folds, as shown in the left panel of the figure 4.11. The R^2 values remain high, while the MAE and RMSE stay low, underscoring the model's generalization capability. Similarly, for PM10, the right panel of the figure 4.11 reveals consistent performance across different folds. The stability in R^2, MAE, and RMSE values indicates that the model maintains its accuracy and robustness on unseen data.

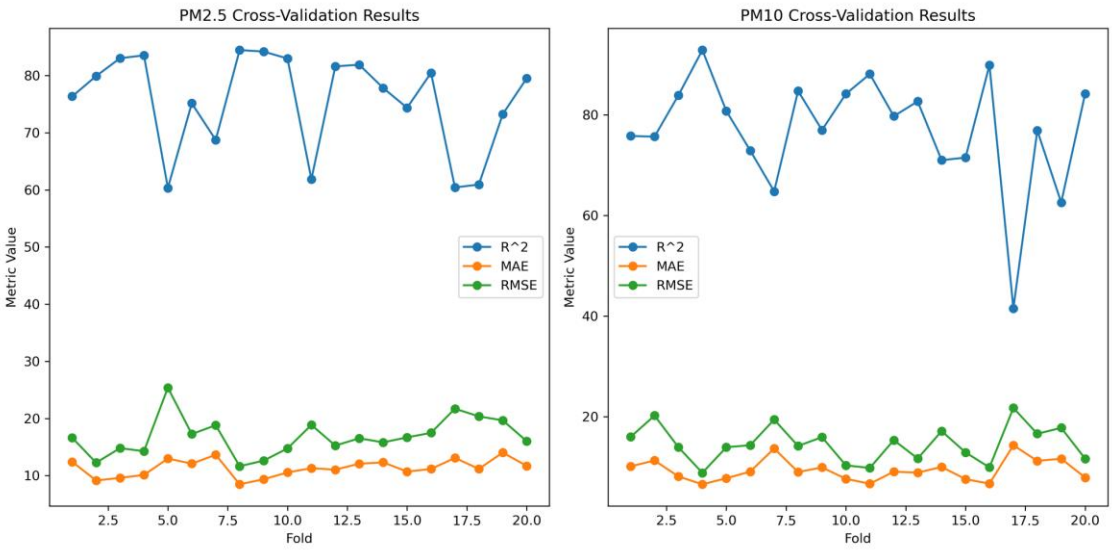


Figure 4.9: Cross-Validation Results (Random Forest)

In summary, the Random Forest model effectively predicts PM2.5 and PM10 concentrations, as evidenced by its high R^2 scores, low error metrics, and consistent cross-validation performance. These results highlight the model's capability to handle the complexities of the dataset and provide reliable predictions, making it a valuable tool for air quality analysis and forecasting.

4.5 Gradient Boosting Model

The Gradient Boosting model exhibits strong predictive performance for both PM2.5 and PM10, achieving metrics that underscore its accuracy and effectiveness. For PM2.5, the model achieves an R^2 of 0.82, signifying that 82% of the variance in the data is explained by the model. Additionally, it records a Mean Absolute Error (MAE) of 10.42, indicating the average absolute difference between predicted and actual values, and a Root Mean Square Error (RMSE) of 14.12, which provides insight into the magnitude of prediction errors. Similarly, for PM10, the model demonstrates even higher accuracy with an R^2 of 0.84, capturing 84% of the variance in the data. The model also achieves a low MAE of 8.16 and an RMSE of 14.08, reflecting its ability to make precise predictions with minimal error. These metrics highlight the robustness of the Gradient Boosting model for air quality analysis.

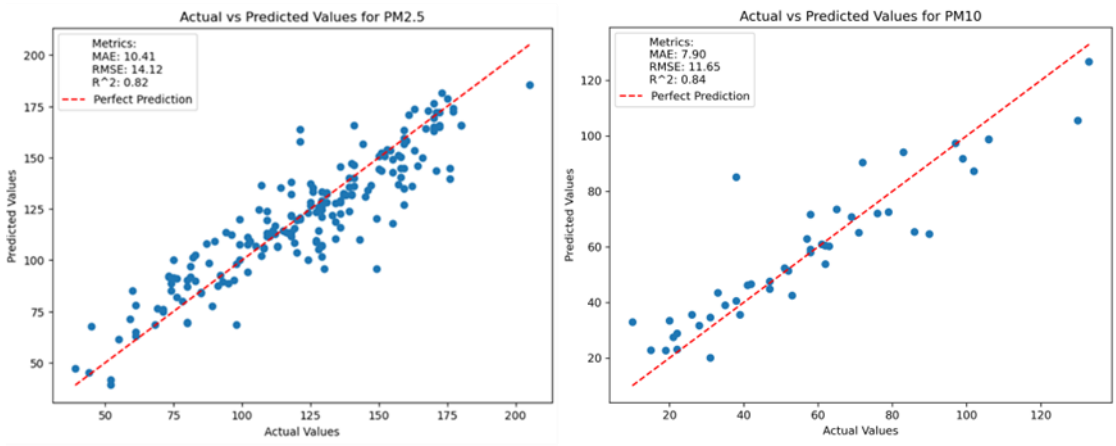


Figure 4.10: Actual vs Predicted values (Gradient Boosting)

The cross-validation results further establish the reliability of the Gradient Boosting model, showing consistent performance across all 20 folds. For PM2.5, the cross-validation results indicate high R^2 values with minimal variability, ensuring that the model generalizes well to unseen data. Both MAE and RMSE remain stable across the folds, indicating that the model avoids overfitting and maintains a balanced tradeoff between bias and variance. For PM10, the cross-validation results are equally impressive, with consistently high R^2 values across folds and minimal fluctuations in MAE and RMSE, emphasizing the robustness and adaptability of the model to varying data distributions. The visual representation of cross-validation metrics supports these observations, with smooth trends and negligible outliers.

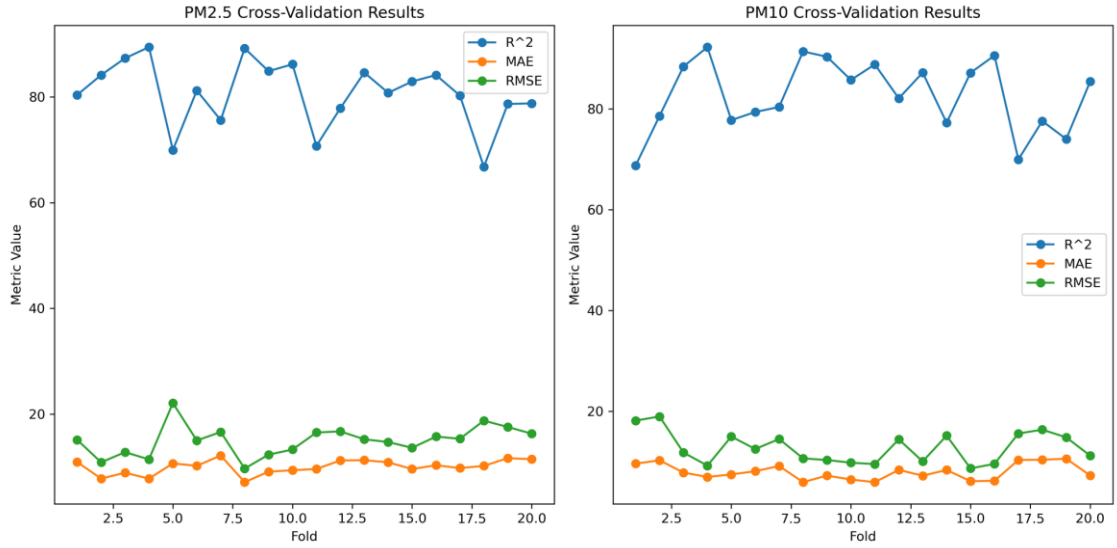


Figure 4.11: Cross-Validation Results (Gradient Boosting)

In conclusion, the Gradient Boosting model demonstrates exceptional predictive accuracy and consistency for both PM2.5 and PM10 concentrations. Its ability to achieve high R^2 values and maintain low errors across folds underscores its suitability for air

quality monitoring and prediction tasks. This makes it a reliable and efficient choice for environmental applications, supporting informed decision-making for air pollution management and public health initiatives.

4.6 Model-Driven PM2.5 and PM10 Concentration Map

To visualize the predicted air quality across the Kathmandu Valley, two maps (figure 4.14) were developed for average PM2.5 and PM10 concentrations during October, November, and December 2024 at 10-day intervals. The maps are based on the predicted values generated by our model, using Aerosol Optical Depth (AOD), Soil Moisture, Normalized Difference Vegetation Index (NDVI), and Land Surface Temperature (LST) as input parameters. These inputs were collected at 10-day intervals for the three months. The predictions were made for 935 equally spaced points at a 1 km spacing across the Kathmandu Valley. The model output was then visualized to create average concentration maps of PM2.5 and PM10. Key insights from these maps include the spatial variation of particulate matter, highlighting areas of higher pollution levels during the study period.

The PM2.5 concentration map (Figure 4.14 left) displays average values for the study period, with the spatial variation showing hotspots of higher concentrations toward the central regions of the valley. The predicted PM2.5 values range from 119.1 µg/m³ to 158.2 µg/m³, with the highest concentrations observed near urbanized and densely populated areas. This pattern indicates significant contributions from vehicular emissions, industrial activities, and urban heating systems in these regions. The PM10 concentration map (Figure 4.14 right) highlights a similar distribution, with concentrations ranging from 64.2 µg/m³ to 108.5 µg/m³. Higher values are evident in the central parts of the valley, closely aligning with PM2.5 trends. The elevated PM10 levels in these areas suggest additional contributions from road dust and construction activities, which are prominent in urban and semi-urban zones.

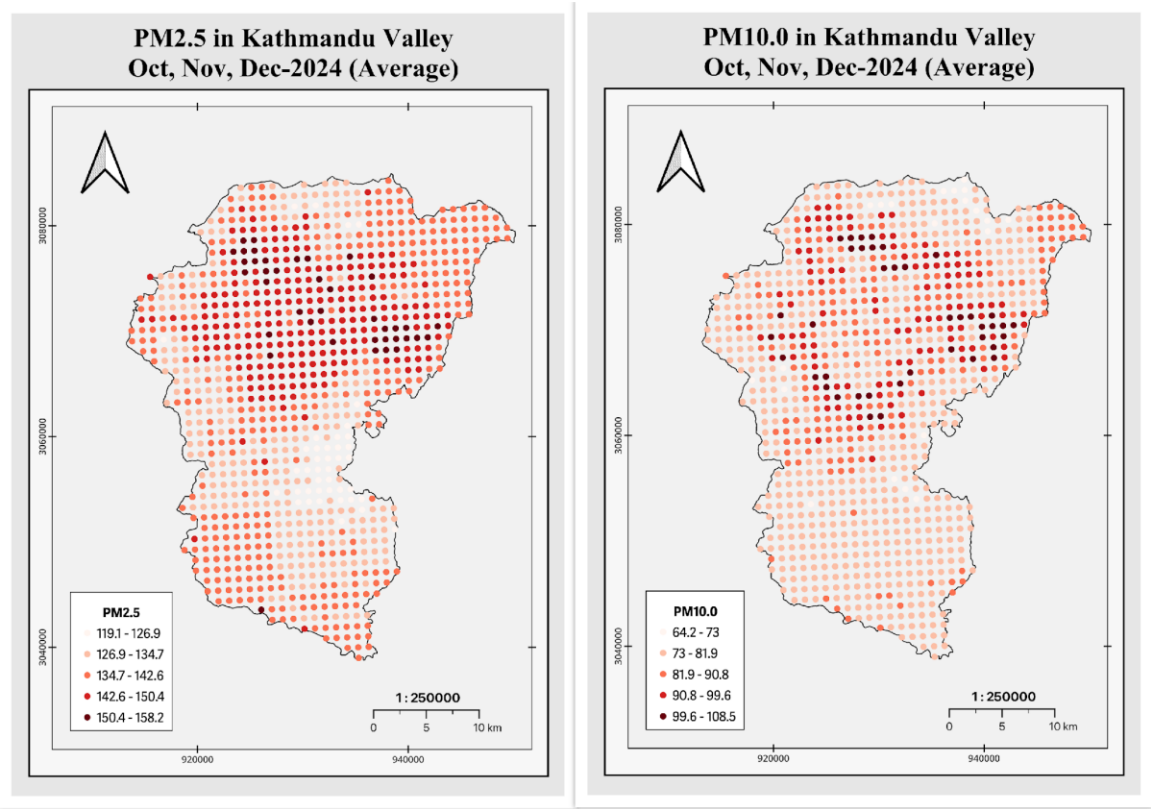


Figure 4.12: PM2.5 and PM10 Concentration Map

These maps clearly demonstrate spatial and temporal trends, with certain regions within the valley exhibiting consistently higher pollutant concentrations. The visualizations emphasize the capability of our model to identify pollution hotspots and temporal variations. These maps offer critical insights into spatial pollution trends and can aid policymakers in targeting interventions to reduce air pollution in high-risk areas.

Chapter 5: Limitations and Discussions

Air pollution poses severe challenges to public health and the environment, particularly in low- and middle-income countries like Nepal, where limited spatial coverage of air quality monitoring stations hinders effective assessment. This study integrates remote sensing data with ground-based observations to estimate PM2.5 and PM10 concentrations in the Kathmandu Valley using machine learning models- Linear Regression, Random Forest, and Gradient Boosting. Data spanning 2022–2023 from ground stations and satellite-derived parameters, including Aerosol Optical Depth (AOD), Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), and soil moisture, were processed and analyzed.

The results reveal that Gradient Boosting outperformed other models, achieving R^2 values of 0.82 and 0.84 for PM2.5 and PM10, respectively, demonstrating its robustness in capturing the non-linear relationships in air quality data. Random Forest also showed high accuracy, with R^2 scores exceeding 0.80. At the same time, Linear Regression proved less effective due to its limitations in modeling complex data distributions, which also resembles the results of other previous studies globally. Spatial maps generated from model predictions highlight pollution hotspots, offering actionable insights for targeted interventions.

While the study underscores the potential of remote sensing and machine learning to address gaps in air quality monitoring, it acknowledges limitations, including coarse data resolution and the exclusion of key meteorological factors. Future research should incorporate higher-resolution datasets and additional variables to improve model reliability and scalability. This work demonstrates a viable approach to enhancing air quality assessments, supporting data-driven policy decisions in Nepal and similar regions globally.

Table 5.1: Metrics for LR, RF, and GBR models

Metrics	Linear Regression		Random Forest		Gradient Boosting	
	PM2.5	PM10	PM2.5	PM10	PM 2.5	PM10
MAE	21.10	18.08	10.99	9.32	10.42	8.16
MSE	624.97	641.24	229.97	226.64	199.55	198.40
RMSE	24.99	25.32	15.17	15.05	14.13	14.09
R2 Score	0.45	0.50	0.80	0.82	0.82	0.84
Adjusted R2 Score	0.44	0.49	0.79	0.82	0.82	0.84

s

Chapter 6: Conclusion

There is a very low availability of air quality stations in Nepal, making it almost impossible to accurately monitor the air quality in most regions with ground-based sensors. The installation of air quality measurement stations in all the areas demands high cost, and it is almost impossible to cover the entire region of the country. One possible solution to the problem of limited spatial coverage of air quality stations is to use remotely sensed raster data to monitor air quality. This study shows the ability of remotely sensed data to predict air quality, particularly the concentration of PM10 and PM2.5, to solve the above-stated problem with ground-based monitoring stations. Like the multiple studies in the global context show the ability of machine learning and remote sensing data to determine air quality, our study also supports this in the context of Nepal.

Since the data distribution of influencing factors is most likely to be nonlinear. This study suggests choosing nonlinear models like bagging and boosting-based regression models, e.g., Random Forest and Gradient Boosting as in this study, which are supposed to be highly efficient in predicting the concentrations of pollutants despite the complex distribution of data. In our case, the gradient boosting model is most reliable, and the random forest model is also efficient for analysis with an accuracy of over 80%. This Study shows that the Aerosol Optical Depth (AOD) and soil moisture show the strongest influence on predicting PM10 and PM2.5 concentrations compared to LST and NDVI. In conclusion, the Integration of remotely sensed data with machine learning models can solve the problem of the limited spatial coverage of air quality stations in Nepal.

References

Ahmad Jabatan Sains, A., Pusat Perkhidmatan Akademik, M., Hashim, M., Md Hashim Pusat Pengajian Sosial, N., dan Persekitaran, P., Nizam Ayof Jabatan Sains, M., & Setyo Budi Jabatan Sains, A. (n.d.). *The Use of Remote Sensing and GIS to Estimate Air Quality Index (AQI) Over Peninsular Malaysia*.

Anggraini, T. S., Irie, H., Sakti, A. D., & Wikantika, K. (2024). Machine learning-based global air quality index development using remote sensing and ground-based stations. *Environmental Advances*, *15*. https://doi.org/10.1016/j.envadv.2023.100456

Arvani, B., Pierce, R. B., Lyapustin, A. I., Wang, Y., Ghermandi, G., & Teggi, S. (2015). High spatial resolution aerosol retrievals used for daily particulate matter monitoring over Po valley, northern Italy. *Atmos. Chem. Phys. Discuss*, *15*, 123–155. https://doi.org/10.5194/acpd-15-123-2015

Bao, R., & Zhang, A. (2020). Does lockdown reduce air pollution? Evidence from 44 cities in northern China. *The Science of the Total Environment*, *731*. https://doi.org/10.1016/J.SCITOTENV.2020.139052

Bechle, M. J., Millet, D. B., & Marshall, J. D. (2013a). Remote sensing of exposure to NO2: Satellite versus ground-based measurement in a large urban area. *Atmospheric Environment*, *69*, 345–353. https://doi.org/10.1016/J.ATMOSENV.2012.11.046

Bechle, M. J., Millet, D. B., & Marshall, J. D. (2013b). Remote sensing of exposure to NO2: Satellite versus ground-based measurement in a large urban area. *Atmospheric Environment*, *69*, 345–353. https://doi.org/10.1016/J.ATMOSENV.2012.11.046

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M. Y., Künzli, N., Schikowski, T., Marcon, A., Eriksen, K. T., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., … de Hoogh, K. (2013). Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmospheric Environment*, *72*, 10–23. https://doi.org/10.1016/J.ATMOSENV.2013.02.037

Bodor, K., Szép, R., & Bodor, Z. (2022). The human health risk assessment of particulate air pollution (PM2.5 and PM10) in Romania. *Toxicology Reports*, *9*, 556–562. https://doi.org/10.1016/J.TOXREP.2022.03.022

Bourgeois, Q., Ekman, A. M. L., Renard, J. B., Krejci, R., Devasthale, A., Bender, F. A. M., Riipinen, I., Berthet, G., & Tackett, J. L. (2018). How much of the global aerosol optical depth is found in the boundary layer and free troposphere? *Atmospheric Chemistry and Physics*, *18*(10), 7709–7720. https://doi.org/10.5194/ACP-18-7709-2018

Caplin, A., Ghandehari, M., Lim, C., Glimcher, P., & Thurston, G. (2019). Advancing environmental exposure assessment science to benefit society. *Nature Communications 2019 10:1*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-09155-4

C R, A., Deshmukh, C. R., D K, N., Gandhi, P., & astu, V. (2018). Detection and Prediction of Air Pollution using Machine Learning Models. *International Journal of Engineering Trends and Technology*, *59*(4), 204–207. https://doi.org/10.14445/22315381/IJETT-V59P238

Deng, R., & Xiong, S. (2005). The coupled model of atmosphere and ground for air pollution remote sensing and its application on guangdong province, China. *International Geoscience and Remote Sensing Symposium (IGARSS)*, *7*, 5133–5136. https://doi.org/10.1109/IGARSS.2005.1526837

Eeftens, M., Beelen, R., De Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., De Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Gražulevičiene, R., Heinrich, J., Hoffmann, B., Jerrett, M., … Hoek, G. (2012). Development of land use regression models for PM2.5, PM 2.5 absorbance, PM10 and PMcoarse in 20 European study areas; Results of the ESCAPE project. *Environmental Science and Technology*, *46*(20), 11195–11205. https://doi.org/10.1021/ES301948K/SUPPL_FILE/ES301948K_SI_001.PDF

Efremenko, D., & Kokhanovsky, A. (2021). Foundations of Atmospheric Remote Sensing. *Foundations of Atmospheric Remote Sensing*. https://doi.org/10.1007/978-3-030-66745-0

Engel-Cox, J. A., Holloman, C. H., Coutant, B. W., & Hoff, R. M. (2004). Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmospheric Environment*, *38*(16), 2495–2509. https://doi.org/10.1016/J.ATMOSENV.2004.01.039

Epa, U., & of Air, O. (2014). Air Quality Index - A Guide to Air Quality and Your Health. Brochure 2014. EPA-456/F-14-002.

Faraji Ghasemi, F., Dobaradaran, S., Saeedi, R., Nabipour, I., Nazmara, S., Ranjbar Vakil Abadi, D., Arfaeinia, H., Ramavandi, B., Spitz, J., Mohammadi, M. J., & Keshtkar, M. (2020). Levels and ecological and health risk assessment of PM2.5-bound heavy metals in the northern part of the Persian Gulf. *Environmental Science and Pollution Research*, *27*(5), 5305–5313. https://doi.org/10.1007/S11356-019-07272-7

Filonchyk, M., Peterson, M., & Hurynovich, V. (2021a). Air pollution in the Gobi Desert region: Analysis of dust-storm events. *Quarterly Journal of the Royal Meteorological Society*, *147*(735), 1097–1111. https://doi.org/10.1002/QJ.3961

Filonchyk, M., Peterson, M., & Hurynovich, V. (2021b). Air pollution in the Gobi Desert region: Analysis of dust-storm events. *Quarterly Journal of the Royal Meteorological Society*, *147*(735), 1097–1111. https://doi.org/10.1002/QJ.3961

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/J.RSE.2017.06.031

Gu, J., Yang, B., Brauer, M., & Zhang, K. M. (2021). Enhancing the Evaluation and Interpretability of Data-Driven Air Quality Models. *Atmospheric Environment*, *246*, 118125. https://doi.org/10.1016/J.ATMOSENV.2020.118125

Halsana, S. (2020). Air Quality Prediction Model using Supervised Machine Learning Algorithms. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 190–201. https://doi.org/10.32628/CSEIT206435

Hanusz, Z., & Tarasińska, J. (2015). Normalization of the Kolmogorov–Smirnov and Shapiro–Wilk tests of normality. *Biometrical Letters*, *52*(2), 85–93. https://doi.org/10.1515/BILE-2015-0008

Henderson, S. B., Beckerman, B., Jerrett, M., & Brauer, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental Science and Technology*, *41*(7), 2422–2428. https://doi.org/10.1021/ES0606780/SUPPL_FILE/ES0606780_SI.PDF

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*(33), 7561–7578. https://doi.org/10.1016/J.ATMOSENV.2008.05.057

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, *23*(4), 725–749. https://doi.org/10.1080/01431160110040323

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D., & Liu, Y. (2018). Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. *Environmental Pollution*, *242*, 675–683. https://doi.org/10.1016/J.ENVPOL.2018.07.016

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021a). *An Introduction to Statistical Learning*. https://doi.org/10.1007/978-1-0716-1418-1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021b). *An Introduction to Statistical Learning*. https://doi.org/10.1007/978-1-0716-1418-1

Jerrett, M., Burnett, R. T., Ma, R., Arden Pope, C., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. E., & Thun, M. J. (2005). Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*, *16*(6), 727–736. https://doi.org/10.1097/01.EDE.0000181630.15826.7D

Kamińska, J. A. (2018). The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław. *Journal of Environmental Management*, *217*, 164–174. https://doi.org/10.1016/J.JENVMAN.2018.03.094

Kashima, S., Yorifuji, T., Tsuda, T., & Doi, H. (2009). Application of land use regression to regulatory air quality data in Japan. *Science of The Total Environment*, *407*(8), 3055–3062. https://doi.org/10.1016/J.SCITOTENV.2008.12.038

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

Knibbs, L. D., Hewson, M. G., Bechle, M. J., Marshall, J. D., & Barnett, A. G. (2014). A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environmental Research*, *135*, 204–211. https://doi.org/10.1016/J.ENVRES.2014.09.011

Kumar, R., Kumar, P., & Kumar, Y. (2020). Time Series Data Prediction using IoT and Machine Learning Technique. *Procedia Computer Science*, *167*, 373–381. https://doi.org/10.1016/J.PROCS.2020.03.240

Leão, M. L. P., Zhang, L., & da Silva Júnior, F. M. R. (2023). Effect of particulate matter (PM2.5 and PM10) on health indicators: climate change scenarios in a Brazilian metropolis. *Environmental Geochemistry and Health*, *45*(5), 2229–2240. https://doi.org/10.1007/S10653-022-01331-8/TABLES/5

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature 2015 525:7569*, *525*(7569), 367–371. https://doi.org/10.1038/nature15371

Liu, B., Ma, X., Ma, Y., Li, H., Jin, S., Fan, R., & Gong, W. (2022). The relationship between atmospheric boundary layer and temperature inversion layer and their aerosol capture capabilities. *Atmospheric Research*, *271*, 106121. https://doi.org/10.1016/J.ATMOSRES.2022.106121

Lo Re, G., Peri, D., & Vassallo, S. D. (2014). Urban Air Quality Monitoring Using Vehicular Sensor Networks. *Advances in Intelligent Systems and Computing*, *260*, 311–323. https://doi.org/10.1007/978-3-319-03992-3_22

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://github.com/slundberg/shap

Mamić, L., Gašparović, M., & Kaplan, G. (2023). Developing PM2.5 and PM10 prediction models on a national and regional scale using open-source remote sensing data. *Environmental Monitoring and Assessment*, *195*(6). https://doi.org/10.1007/s10661-023-11212-x

Manderscheid, L. V. (1965). Significance Levels—0.05, 0.01, or? *American Journal of Agricultural Economics*, *47*(5), 1381–1385. https://doi.org/10.2307/1236396

Martin, R. V. (2008). Satellite remote sensing of surface air quality. *Atmospheric Environment*, *42*(34), 7823–7843. https://doi.org/10.1016/J.ATMOSENV.2008.07.018

Ma, Z., Hu, X., Sayer, A. M., Levy, R., Zhang, Q., Xue, Y., Tong, S., Bi, J., Huang, L., & Liu, Y. (2016). Satellite-Based Spatiotemporal Trends in PM2.5 Concentrations: China, 2004-2013. *Environmental Health Perspectives*, *124*(2), 184–192. https://doi.org/10.1289/EHP.1409481

Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., Avol, E. L., Oron, A. P., Larson, T., Liu, L. J. S., & Kaufman, J. D. (2011). Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric Environment*, *45*(26), 4412–4420. https://doi.org/10.1016/J.ATMOSENV.2011.05.043

Mozumder, C., Reddy, K. V., & Pratap, D. (2013). Air Pollution Modeling from Remotely Sensed Data Using Regression Techniques. *Journal of the Indian Society of Remote Sensing*, *41*(2), 269–277. https://doi.org/10.1007/S12524-012-0235-2/METRICS

Prakash, S., Goswami, M., Khan, Y. D. I., & Nautiyal, S. (2021). Environmental impact of COVID-19 led lockdown: A satellite data-based assessment of air quality in Indian megacities. *Urban Climate*, *38*, 100900. https://doi.org/10.1016/J.UCLIM.2021.100900

Raymaekers, J., & Rousseeuw, P. J. (2021). Transforming variables to central normality. *Machine Learning*, 1–23. https://doi.org/10.1007/S10994-021-05960-5/FIGURES/13

Rivas, I., Kumar, P., & Hagen-Zanker, A. (2017). Exposure to air pollutants during commuting in London: Are there inequalities among different socio-economic groups? *Environment International*, *101*, 143–157. https://doi.org/10.1016/J.ENVINT.2017.01.019

Rowley, A., & Karakuş, O. (2023). Predicting air quality via multimodal AI and satellite imagery. *Remote Sensing of Environment*, *293*, 113609. https://doi.org/10.1016/J.RSE.2023.113609

Sakti, A. D., Anggraini, T. S., Ihsan, K. T. N., Misra, P., Trang, N. T. Q., Pradhan, B., Wenten, I. G., Hadi, P. O., & Wikantika, K. (2023a). Multi-air pollution risk assessment in Southeast Asia region using integrated remote sensing and socio-economic data products. *Science of The Total Environment*, *854*, 158825. https://doi.org/10.1016/J.SCITOTENV.2022.158825

Sakti, A. D., Anggraini, T. S., Ihsan, K. T. N., Misra, P., Trang, N. T. Q., Pradhan, B., Wenten, I. G., Hadi, P. O., & Wikantika, K. (2023b). Multi-air pollution risk assessment in Southeast Asia region using integrated remote sensing and socio-economic data products. *Science of The Total Environment*, *854*, 158825. https://doi.org/10.1016/J.SCITOTENV.2022.158825

Son, Y., Osornio-Vargas, Á. R., O'Neill, M. S., Hystad, P., Texcalac-Sangrador, J. L., Ohman-Strickland, P., Meng, Q., & Schwander, S. (2018). Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. *Science of The Total Environment*, *639*, 40–48. https://doi.org/10.1016/J.SCITOTENV.2018.05.144

Srivastava, C., Singh, S., & Singh, A. P. (2019). Estimation of air pollution in Delhi using machine learning techniques. *2018 International Conference on Computing, Power and Communication Technologies, GUCON 2018*, 304–309. https://doi.org/10.1109/GUCON.2018.8675022

Stahl, S. A., & Nagy, W. E. (2015). The Science of Environmental Pollution, Third Edition. *The Science of Environmental Pollution, Third Edition*, 1–209. https://doi.org/10.1201/9781315226149/SCIENCE-ENVIRONMENTAL-POLLUTION-FRANK-SPELLMAN

Steinskog, D. J., Tjøtheim, D. B., & Kvamstø, N. G. (2007). A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality. *Monthly Weather Review*, *135*(3), 1151–1157. https://doi.org/10.1175/MWR3326.1

Timilsina, S., Gautam, P., & Shrestha, K. L. (2023). Relation between Modis-based Aerosol Optical Depth and Particulate Matter in Kathmandu using Regression Model. *Journal of Environment Sciences*, 1–12. https://doi.org/10.3126/JES.V9I1.56417

Trewhela, B., Huneeus, N., Munizaga, M., Mazzeo, A., Menut, L., Mailler, S., Valari, M., & Ordoñez, C. (2019). Analysis of exposure to fine particulate matter using passive data from public transport. *Atmospheric Environment*, *215*, 116878. https://doi.org/10.1016/J.ATMOSENV.2019.116878

van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., & Villeneuve, P. J. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. *Environmental Health Perspectives*, *118*(6), 847–855. https://doi.org/10.1289/EHP.0901623

Vîrghileanu, M., Săvulescu, I., Mihai, B. A., Nistor, C., & Dobre, R. (2020). Nitrogen Dioxide (NO2) Pollution Monitoring with Sentinel-5P Satellite Imagery over Europe during the Coronavirus Pandemic Outbreak. *Remote Sensing 2020, Vol. 12, Page 3575*, *12*(21), 3575. https://doi.org/10.3390/RS12213575

Wang, R., Henderson, S. B., Sbihi, H., Allen, R. W., & Brauer, M. (2013). Temporal stability of land use regression models for traffic-related air pollution. *Atmospheric Environment*, *64*, 312–319. https://doi.org/10.1016/J.ATMOSENV.2012.09.056

Wu, Q., & Lin, H. (2019). A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Science of The Total Environment*, *683*, 808–821. https://doi.org/10.1016/J.SCITOTENV.2019.05.288

Wu, S., Sun, Y., Bai, R., Jiang, X., Jin, C., & Xue, Y. (2024). Estimation of PM2.5 and PM10 Mass Concentrations in Beijing Using Gaofen-1 Data at 100 m Resolution. *Remote Sensing*, *16*(4). https://doi.org/10.3390/rs16040604

Xue, T., Zheng, Y., Geng, G., Zheng, B., Jiang, X., Zhang, Q., & He, K. (2017). Fusing observational, satellite remote sensing and air quality model simulated data to estimate spatiotemporal variations of PM2.5 exposure in China. *Remote Sensing*, *9*(3). https://doi.org/10.3390/rs9030221

You, W., Zang, Z., Pan, X., Zhang, L., & Chen, D. (2015). Estimating PM2.5 in Xi'an, China using aerosol optical depth: A comparison between the MODIS and MISR retrieval models. *Science of The Total Environment*, *505*, 1156–1165. https://doi.org/10.1016/J.SCITOTENV.2014.11.024

Yunesian, M., Rostami, R., Zarei, A., Fazlzadeh, M., & Janjani, H. (2019). Exposure to high levels of PM2.5 and PM10 in the metropolis of Tehran and the associated health risks during 2016–2017. *Microchemical Journal*, *150*, 104174. https://doi.org/10.1016/J.MICROC.2019.104174

Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys and Tutorials*, *12*(2), 159–170. https://doi.org/10.1109/SURV.2010.021510.00088

Zheng, Y., Zhang, Q., Liu, Y., Geng, G., & He, K. (2016). Estimating ground-level PM2.5 concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements. *Atmospheric Environment*, *124*, 232–242. https://doi.org/10.1016/J.ATMOSENV.2015.06.046

Zheng, Z., Yang, Z., Wu, Z., & Marinello, F. (2019a). Spatial Variation of NO2 and Its Impact Factors in China: An Application of Sentinel-5P Products. *Remote Sensing 2019, Vol. 11, Page 1939*, *11*(16), 1939. https://doi.org/10.3390/RS11161939

Zheng, Z., Yang, Z., Wu, Z., & Marinello, F. (2019b). Spatial Variation of NO2 and Its Impact Factors in China: An Application of Sentinel-5P Products. *Remote Sensing 2019, Vol. 11, Page 1939*, *11*(16), 1939. https://doi.org/10.3390/RS11161939

Zhu, Y., Xie, J., Huang, F., & Cao, L. (2020). Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. *The Science of the Total Environment*, *727*. https://doi.org/10.1016/J.SCITOTENV.2020.138704

Zoran, M. A., Savastru, R. S., Savastru, D. M., & Tautan, M. N. (2020a). Assessing the relationship between surface levels of PM2.5 and PM10 particulate matter impact on COVID-19 in Milan, Italy. *The Science of the Total Environment*, *738*. https://doi.org/10.1016/J.SCITOTENV.2020.139825

Zoran, M. A., Savastru, R. S., Savastru, D. M., & Tautan, M. N. (2020b). Assessing the relationship between surface levels of PM2.5 and PM10 particulate matter impact on COVID-19 in Milan, Italy. *The Science of the Total Environment*, *738*. https://doi.org/10.1016/J.SCITOTENV.2020.139825

APPENDIX

	A	B	C	D	E	F	G	H
1		aod	lst	NDVI	precipitati	soil_moist	pm25	pm10
2	0	1.033	25.53	0.2967	0	0.33523	182	112
3	3	0.173	24.39	0.3706	0	0.33569	98	50
4	4	0.158	24.39	0.3706	0	0.33569	98	50
5	5	0.181	25.27	0.3706	0	0.33422	89	36
6	6	0.284	25.27	0.3706	0	0.33422	89	36
7	10	0.256	17.73	0.2135	0	0.33881	137	63
8	11	0.178	17.73	0.2135	0	0.33881	137	63
9	12	0.265	17.49	0.2135	0	0.33367	145	69
10	13	0.37	17.49	0.2135	0	0.33367	145	69
11	14	0.461	16.51	0.2135	0	0.35992	160	71
12	15	0.48	16.51	0.2135	0	0.35992	160	71
13	16	0.707	13.89	0.2135	0	0.34376	162	77
14	17	0.621	13.89	0.2135	0	0.34376	162	77
15	18	0.367	15.51	0.2135	0	0.33931	161	99
16	19	0.249	15.51	0.2135	0	0.33931	161	99
17	20	0.212	18.17	0.2135	0	0.31708	120	67
18	21	0.099	18.17	0.2135	0	0.31708	120	67
19	22	0.578	17.79	0.2135	0	0.30701	134	117
20	23	0.534	17.79	0.2135	0	0.30701	134	117
21	24	0.224	20.85	0.2215	0	0.32697	136	71
22	25	0.113	20.85	0.2215	0	0.32697	136	71
23	26	0.206	19.95	0.2215	0	0.31518	133	69
24	27	0.247	19.95	0.2215	0	0.31518	133	69
25	28	0.292	20.25	0.2215	0	0.30564	139	82
26	29	0.285	20.25	0.2215	0	0.30564	139	82
27	30	0.3	21.45	0.2215	0	0.30196	146	77
28	31	0.342	21.45	0.2215	0	0.30196	146	77
29	32	0.294	20.57	0.2215	0	0.33094	128	179

Figure: Cleaned and Pre-processed Data Sample