Developing Spatial Simulator Engine For Urban Air Pollution Monitoring-Application on MODIS data

A Project Report Submitted to the
Department of Computer Science of
Ramakrishna Mission Vivekananda Educational and Research Institute, Belur,
in partial fulfilment of the requirements for the degree of
MSc in Computer Science.

Submitted by
DEBAJYOTI MAITY
ID No. B2230023

Supervisor:
Dr. Swalpa Kumar Roy
Department of Computer Science and Engineering
Alipurduar Government Engineering and Management College

Department of Computer Science
Ramakrishna Mission Educational and Research Institute
Belur Math, Howrah 711202, West Bengal, India
April 28, 2024

Developing Spatial Simulator Engine For Urban Air Pollution Monitoring-Application on MODIS data

By

Debajyoti Maity

Declaration by student:

“I hereby declare that the present dissertation is the outcome of my project work under the guidance of Dr. Swalpa Kumar Roy and I have properly acknowledged the sources of materials used in my project report."

(Debajyoti Maity, ID No. B2230023)

A project report in the partial fulfilment of the requirements of the degree of MSc in Computer Science

Examined and approved on

by

Swalpa Kumar Roy(supervisor)
Department of Computer Science and Engineering

Alipurduar Goverment Engineering and Management College

Countersigned by

Registrar
Ramakrishna Mission Vivekananda Educational and Research Institute

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah 711202, West Bengal, India

Acknowledgement

The present project work is submitted in partial fulfilment of the requirements for the degree of Master of Science at Ramakrishna Mission Vivekananda University (RKMVU). I express my deepest gratitude to my supervisor, Prof. Swalpa Kumar Roy of Alipurduar Government Engineering and Management College, for his inestimable support, encouragement, profound knowledge, and largely helpful conversations. I am thankful for his guidance and for providing me with a systematic approach for the completion of my project work. His dedication and hard work have been a great inspiration to me.I would also like to extend my gratitude to Professors Dr. Mainak Thakur, Dr. Mrinmoy Ghorai, and Mr. Subhojit Mandal for their valuable contributions, guidance, and insights that have significantly enriched my academic journey. Additionally, I am immensely grateful to the Vice-Chancellor of Ramakrishna Mission Vivekananda University for his encouragement and unwavering support throughout the course.Last but not least, I owe a debt of gratitude to my parents for their constant support, encouragement, and understanding. Their love and encouragement have been my pillar of strength. I am also thankful to my fellow classmates for their camaraderie and collaborative spirit.

Belur
April 28, 2024

Debajyoti Maity
Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute

Contents

List of Tables

List of Figures

Chapter 1

Abstract

As we come into the 21st century, the air pollution emerges as a vital global concern.Its effects ripple across every aspect of our lives, affecting individual's health, throwing ecosystems off balance. and even the stability of our climate. Recent studies have extensively explored the relationship between Aerosol Optical Depth (AOD) derived from satellite data and the concentrations of fine particulate matter (PM2.5). However, limited attention has been directed towards comprehending the temporal and spatial dynamics of this relationship, especially within the context of the Delhi region in India. Our research aimed to investigate the connection between locally estimated PM2.5 concentrations, sourced from environmental monitoring, and AOD measurements obtained from NASA's Terra and Aqua satellites during the period spanning 2018 to 2022.

Our analysis underscored that the combined utilization of AOD data from both Terra and Aqua satellites significantly extended the coverage of AOD readings compared to employing either satellite independently. Notably, our findings unveiled noteworthy variations in the correlation between AOD measurements and PM2.5 concentrations across diverse geographic locations within the Delhi region (5.2). Remarkably, this correlation exhibited heightened strength during the summer and fall seasons, while displaying comparatively weaker associations during the winter and spring periods.

In my recent project focusing on the assessment of spatial-temporal patterns ofAerosol Optical Depth(AOD) levels within the Delhi region of India, I utilized a high-resolution 1 km Aerosol Optical Depth (AOD) product derived from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm developed for

MODIS satellite data.

Employing day-specific calibrations of AOD information, I successfully predicted Aod value in the Delhi area. To enhance the predictive accuracy of my model, I integrated essential variables related to land use and meteorology.

Additionally, I addressed missing values using the kriging method[2], selectively filling in data points that fell below the 33% observation threshold. This process significantly augmented the completeness of the dataset, ensuring a more comprehensive analysis.

The project methodology involved accounting for the variability in AOD data and addressing nested regions within days to effectively capture spatial variation. This approach aimed to control for day-to-day variability inherent in the AOD-PM2.5 relationship, influenced by dynamic parameters such as particle optical properties, vertical and diurnal concentration profiles, and ground surface reflectance.

The outcomes of this study shed light on the potential utilization of satellite-derived data, specifically the MAIAC data, to assess and understand the spatial variability of PM2.5 levels in the Delhi region. The focus was to explore its application in analyzing and mapping urban air quality dynamics, particularly within areas characterized by high traffic density.

Keywords: Aerosol Optical Depth, Moderate Resolution Imaging Spectroradiometer, Terra, Aqua, PM2.5, Delhi, India

Chapter 2

Introduction

Aerosols present in the atmosphere wield a substantial influence on the Earth's climate system [19], profoundly impacting human health ([22]; [17]; [1]). However, the precise effect on the Earth's climate remains poorly understood, drawing significant interest from atmospheric scientists ([15]). Aerosols in the atmosphere play a pivotal role in altering the climate by absorbing and reflecting solar radiation ([13]). They enter the atmosphere through diverse sources, including anthropogenic and natural origins, along with formation via various atmospheric chemical and physical processes ([20]). Aerosol Optical Depth (AOD) is a crucial atmospheric parameter that quantifies the extent to which aerosol particles disperse and absorb sunlight as it traverses through the Earth's atmosphere. These aerosol particles, encompassing minute solid or liquid particles like dust, smoke, pollen, pollutants, and sea salt, contribute to the optical thickness of the atmosphere. In essence, AOD serves as a quantitative measure of the atmospheric turbidity caused by aerosols. It plays a pivotal role in elucidating atmospheric composition, air quality, climate dynamics, and visibility conditions. By assessing AOD, scientists gain insights into how aerosols influence the transmission of both direct and diffuse solar radiation, subsequently impacting the solar energy reaching the Earth's surface. AOD values are indicative of the concentration of aerosol particles present in the atmosphere. Higher AOD values signify elevated aerosol concentrations, which can have tangible effects such as diminished visibility, alterations in the Earth's energy balance, and even exert cooling influences on the climate. Monitoring AOD is, therefore, instrumental in comprehending the complex interplay between aerosols and the Earth's atmosphere, contributing significantly to atmospheric science and environmental research.

Aerosol Optical Depth (AOD) serves as an alternative index for estimating particulate matter (PM) in the atmosphere ([8]). Exposure to PM poses substantial health risks ([5]), contributing to conditions such as lung cancer, cardiopulmonary mortality, and pulmonary inflammation ([7]). To comprehend the impact of aerosols on climate and human health, continuous monitoring of aerosols through surface in situ measurement techniques and satellite-based column-integrated observations remains an indispensable tool. ([8]).

Chapter 3

Problem Statement

3.1 Why satellite data have missing value?

Satellite data, such as INSAT 3D, often encounter missing values due to various reasons, with one prominent factor being the geostationary nature of observations. Geostationary satellites have limited coverage, leading to data gaps in certain regions or under specific conditions. Understanding the causes of missing values is crucial for devising effective strategies to address them. One significant contributor to these gaps is the presence of cloud cover, hindering the satellite's ability to capture clear, uninterrupted observations. Cloud cover introduces variability in the availability of satellite data, leading to missing values in certain regions and during specific atmospheric conditions. **Sun glint** occurs when sunlight reflects off water surfaces and directly into the satellite sensors. This phenomenon can create bright spots in satellite imagery, making it challenging to obtain accurate measurements in affected areas.

3.2 AOD

The significance of AOD lies in its capacity to quantify the extent to which aerosol particles influence the transmission of solar radiation through the atmosphere. High AOD values indicate an increased concentration of aerosols, reflecting elevated levels of particulate matter in the air. As sunlight interacts with these aerosol particles, it undergoes scattering and absorption, directly impacting visibility and the overall energy balance in the atmosphere.

AOD plays a crucial role in environmental monitoring, providing insights into the spatial and temporal distribution of aerosols and their potential impact on climate, weather patterns, and air quality. In the context of pollution studies, AOD serves as a valuable indicator, enabling researchers to assess the degree of particulate pollution and its implications for human health, ecosystem dynamics, and atmospheric processes. The ability to quantify the optical properties of aerosols through AOD measurement positions it as a key tool in the comprehensive toolkit for understanding and addressing contemporary environmental challenges.

3.3 Why Cloud cover can be a problem for pollutant monitoring using MODIS?

Cloud cover poses a challenge for pollutant monitoring using MODIS (Moderate Resolution Imaging Spectroradiometer) as it obstructs the direct observation of the Earth's surface. In the presence of clouds, AOD data becomes unreliable, impacting the accuracy of pollutant assessments. Addressing this challenge requires developing models capable of predicting missing AOD values under cloud-covered conditions. Developing effective models for imputing missing AOD values under cloud-covered conditions becomes imperative to enhance the precision of pollutant monitoring using MODIS data.

3.4 How the fusion of ground sensor observations can help?

Integrating ground sensor observations from systems like ERA5, MERRA-2, and urban air pollution monitoring stations can enhance the accuracy and reliability of pollutant monitoring. Ground-based data provides additional contextual information, aiding in the imputation of missing AOD values and improving the overall performance of environmental models.

3.5 How Cloud Cover imputations can help the Modelers?

This work aims to address the limitations posed by cloud cover in pollutant monitoring using MODIS data. By developing models capable of predicting missing AOD values under cloud-covered conditions, modelers can obtain more comprehensive and continuous datasets. The imputed AOD values contribute to a more accurate representation of pollutant distribution, supporting robust and reliable environmental modeling efforts. The outcome of this work has broader implications for understanding air quality dynamics and facilitating informed decision-making in pollution management.

Chapter 4

Literature Survey

In this section, we present a concise summary of the latest advancements in deep learning models and machine learning models dedicated to the analysis of sequential data. Additionally, we discuss spatio-temporal models relevant to environmental monitoring and explore techniques for managing missing data.

4.1 Deep Learning Models for Sequential Data

Recurrent Neural Networks (RNNs) have stood out as prominent deep learning models for handling sequential data[6]. Nevertheless, RNNs encounter challenges with the vanishing gradient problem, hindering their ability to capture long-term dependencies within the data. To tackle this limitation, the introduction of Long Short-Term Memory (LSTM) networks [27] and CNN-LSTM networks [23] has been instrumental. These models find application in diverse tasks and time series forecasting .

4.2 Deep Learning Models for Spatio-Temporal data

The realm of spatio-temporal forecasting is frequently cast within the framework of multivariate time series forecasting, where each time series corresponds to a variable at a specific location. While the utilization of deep learning models for spatio-temporal forecasting is not a novel concept, its application has evolved over time. For instance, Zhang et al. (2018)[26] employed a Long Short-Term Memory (LSTM) model to predict daily land surface temperature, and McDermott and Wikle

[14] devised an ensemble quadratic echo state network for forecasting Pacific sea surface temperature.

4.3 Addressing Missing Values in Modelling

Random Forest Regressors[4] have emerged as a potent tool in environmental monitoring, showcasing their prowess in predicting continuous variables essential for understanding and managing environmental processes. These regressors belong to the family of ensemble learning algorithms, leveraging the collective intelligence of multiple decision trees to yield accurate and robust predictions.

In the context of air quality forecasting, Random Forest Regressors[9] excel in modeling the complex relationships between various meteorological parameters and air pollutant concentrations[11]. By harnessing the power of numerous decision trees, these models adeptly capture the intricate patterns in the data, enabling reliable predictions of pollutant levels. This capability proves invaluable for assessing the potential impact of emissions, implementing timely interventions, and safeguarding public health.

Chapter 5

Materials

5.1 Study area

The study area for the analysis of Aerosol Optical Depth (AOD) levels encompasses the Indian Territory, with a specific focus on Delhi Figure 5.2. Delhi, situated in northern India figure 5.1, is positioned between latitudes 28.40° N and 28.88° N, and longitudes 76.83° E and 77.34° E. The Tropic of Cancer intersects the middle of the Indian subcontinent, rendering the majority of the region climatically tropical. This research aims to investigate and understand the variations in AOD levels within the specified geographic coordinates, providing valuable insights into the atmospheric conditions of the specific urban environment of Delhi.

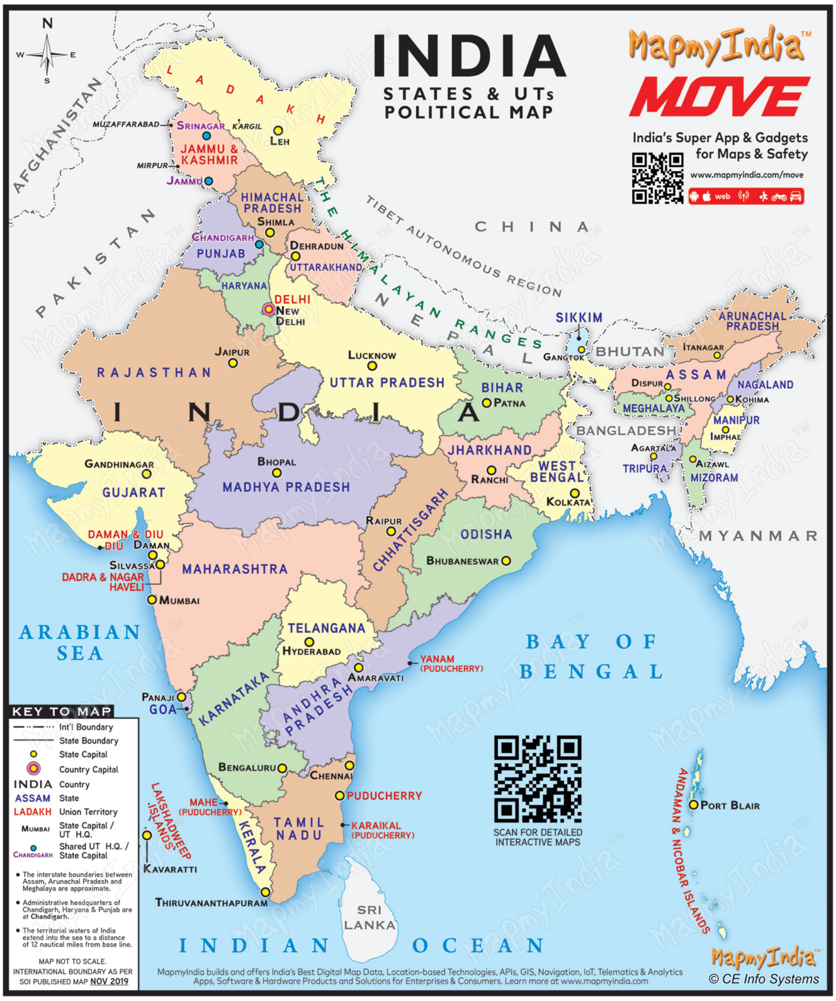


Figure 5.1: India Map

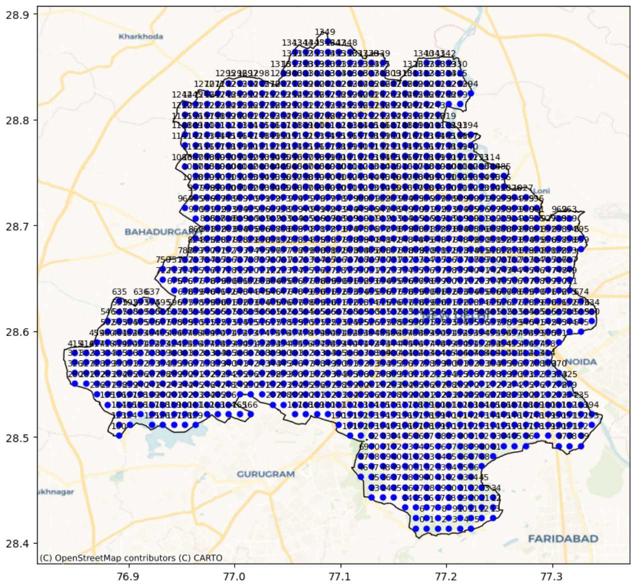


Figure 5.2: Geographic Locations of Delhi Region

5.1.1 Data types and sources

In the present study, mainly three types of data were used. These are MODIS AOD data [1] $PM_{2.5}$ and PM_{10} mass concentration data, and meteorological data(obtained from the ERA-5)[2] and meteorological data ,obtained from Central Pollution Control Board(CPCB)[3] for the development of a prediction model using multiple regression and random forest regression analyses. The characteristics of the data and its sources are explained below.

5.1.2 MODIS data

MODIS(Moderate Resolution Imaging Spectroradiometer) Aqua [4] and Terra [5] satellite data, which provide a wealth of high-resolution and multi-spectral information, this project seeks to create a sophisticated spatial engine capable of accurately detecting and mapping air pollutants on a global scale. Terra and Aqua are two Earth-

[1] https://search.earthdata.nasa.gov/search/granules?portal=idn&p=C2324689816-LPCLOUD&q=aod&sb[0]=76.5%2C27.41702%2C78.22266%2C29.27992&fi=MODIS&tl=1691835349!3!!&lat=19.810554109292468&long=68.58984375000001&zoom=4

[2] https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form

[3] https://www.cpcb.nic.in/

[4] https://en.wikipedia.org/wiki/Aqua_(satellite)

[5] https://terra.nasa.gov/

observing satellites launched as part of NASA's Earth Observing System (EOS) program. They are designed to collect data about various aspects of Earth's atmosphere, land, oceans, and climate.

MODIS products are available in different processing levels (level 1.0—geolocated and calibrated, level 2.0—derived geophysical data products, and level 3.0—gridded time-averaged products) and collections ([10]). In the current study, two different types of MODIS data (MODIS Atmosphere Monthly Global Product and MODIS Atmosphere Daily Global Product) at a wavelength of 470 nm were used. The data used were for 5 years (2018–2022). It was used to study annual AOD variation. The data extracted for the year 2018 were used to analyze the monthly AOD variations. The daily data from January 1, 2018, to december 30, 2022 were extracted for analyzing the diurnal variations as these days recorded the highest AOD values compared to other days of the year.

5.1.3 Meteorological data

This twelve features enriches the analytical scope, enabling a more nuanced exploration of the complex interactions and interdependencies within the meteorological system. The random forest regression analyses leverage these twelve variables to uncover intricate relationships and enhance the overall predictive modeling of atmospheric phenomena. These observations are taken from ERA5 dataset and Central Pollution Control Board, matched with ground level monitoring stations locations.

1. **Temperature (°C):** This crucial parameter provides insights into the thermal state of the atmosphere, influencing weather patterns and climate conditions. Temperature can influence the emissions of certain aerosol precursors. For example, higher temperatures may enhance the release of biogenic volatile organic compounds (VOCs)[6] from vegetation, contributing to the formation of secondary organic aerosols[7]. On the other hand, temperature can also affect anthropogenic emissions[8], such as those from industrial processes and vehicular traffic.

   Source:Central pollution Control Board(CPCB)

[6] https://en.wikipedia.org/wiki/Volatile_organic_compound

[7] https://en.wikipedia.org/wiki/Secondary_organic_aerosol

[8] https://en.wikipedia.org/wiki/Greenhouse_gas_emissions

2. **Relative Humidity (%):** Representing the percentage of moisture in the air, relative humidity is pivotal for understanding atmospheric moisture levels and potential precipitation. Relative humidity influences the hygroscopic growth of aerosol particles. Hygroscopic aerosols [9] are those that can absorb water vapor from the surrounding air, leading to an increase in their size. This growth can affect the scattering and absorption properties of aerosols, influencing AOD. For example, hygroscopic growth may lead to increased scattering of solar radiation.
   Source:Central pollution Control Board(CPCB)

3. **Wind Speed (m/s):** Wind speed, measured in meters per second, characterizes the horizontal movement of air, playing a significant role in weather dynamics.Higher wind speeds can enhance the horizontal and vertical mixing of aerosols, leading to their dispersion over larger areas and potentially reducing local AOD levels. Additionally, strong winds may contribute to the removal of aerosols through dry deposition and can impact the formation and behavior of aerosol particles in the atmosphere, affecting AOD patterns.
   Source:Central pollution Control Board(CPCB)

4. **Precipitation (mm):** Precipitation data, measured in millimeters, offers valuable information about rainfall or snowfall, impacting hydrological processes. Precipitation, especially rain, captures and carries aerosol particles as it falls through the atmosphere, effectively removing them from the air.Precipitation events can contribute to the dispersion and dilution of aerosol concentrations, influencing their spatial distribution.
   Source:Central pollution Control Board(CPCB)

5. **Solar Radiation (MJ/m^2):** Solar radiation quantifies the amount of radiant energy received from the sun, influencing temperature variations and energy balance. Increased solar radiation can enhance atmospheric stability and vertical mixing, potentially leading to the dispersion and dilution of aerosol particles,thereby reducing AOD levels.Additionally, solar radiation can impact the photochemical processes involving particulate matter, influencing the formation and transformation of aerosols, and subsequently affecting AOD.

[9] https://www.sciencedirect.com/topics/earth-and-planetary-sciences/hygroscopic-aerosols

Source: Central Pollution Control Board(CPCB)

6. **Particulate Matter 2.5 (pm2.5):** This feature assesses fine inhalable particles, contributing to the evaluation of air quality.PM2.5 [10] includes particles with a diameter of 2.5 micrometers or smaller, which can remain suspended for longer periods in the atmosphere. The presence of elevated PM2.5 concentrations is associated with increased AOD levels. Fine particles can serve as nuclei for aerosol formation and growth, leading to the accumulation of aerosols in the atmosphere. PM2.5 can also influence the scattering and absorption of sunlight, affecting the radiative balance and subsequently influencing AOD.
   Source:Central Pollution Control Board(CPCB)

7. **K-Index (KINDEX):** The K-Index is a meteorological metric indicating the potential for thunderstorm activity, aiding in the identification of severe weather conditions.While the K-Index itself may not have a direct impact on aerosol optical depth (AOD), it is linked to weather dynamics that can influence AOD levels.or example, the vertical motion in thunderstorms can lead to the uplift of aerosols to higher altitudes or contribute to the scavenging of aerosols through precipitation. The subsequent rainfall associated with thunderstorms can "wash out" aerosols from the air.
   Source: ERA5

8. **Planetary Boundary Layer (PBL):** This atmospheric layer, influenced by the Earth's surface, is vital for understanding the exchange of heat, moisture, and momentum.The PBL is where the Earth's surface strongly interacts with the atmosphere. Changes in the PBL height, temperature, and humidity can influence the dispersion, transport, and vertical distribution of aerosols. For instance, a higher PBL height might allow aerosols to disperse more widely in the atmosphere, affecting AOD levels over a larger vertical extent.
   Source:ERA5

9. **Vertical Wind Speed (windy):** Vertical wind speed complements horizontal wind components, providing a comprehensive view of wind behavior in different directions. Strong vertical winds contribute to better mixing of aerosols throughout the air column, impacting their concentration and distribution.

[10] https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

The interaction between vertical wind speed and planetary boundary layer dynamics plays a key role in shaping the patterns of AOD.
Source:ERA5

10. **Eastward Wind Component (u):** The eastward wind component offers information about horizontal wind flow in the eastward direction. This component influences the transport and dispersion of aerosol particles, affecting the spatial distribution and concentration of aerosol optical depth (AOD). Changes in the eastward wind component can impact the long-range transport of aerosols, contributing to variations in AOD levels across different geographic locations..
Source:ERA5

11. **Northward Wind Component (v):** Representing the northward component of wind, this parameter contributes to the overall wind vector analysis.The magnitude and direction of the northward wind component can impact the geographical distribution and concentration of AOD, affecting air quality and atmospheric conditions.

    Source:ERA5

12. **Wind Speed - Horizontal (windx):** Indicating the horizontal component of wind speed, this feature enhances the understanding of wind patterns.Changes in horizontal wind speed can affect the spatial distribution and concentration of aerosol optical depth (AOD).
Source: ERA5

5.1.4 Data Matching

The study area was divided into the geographic grids with 0.01°× 0.01°spatial resolution. The aforementioned data from 2018 to 2022 were employed and matched spatiotemporally for the long-term seam less daily AOD retrieval in this study. The grids with in-situ AOD monitoring stations were extracted for training and evaluation of the proposed model.

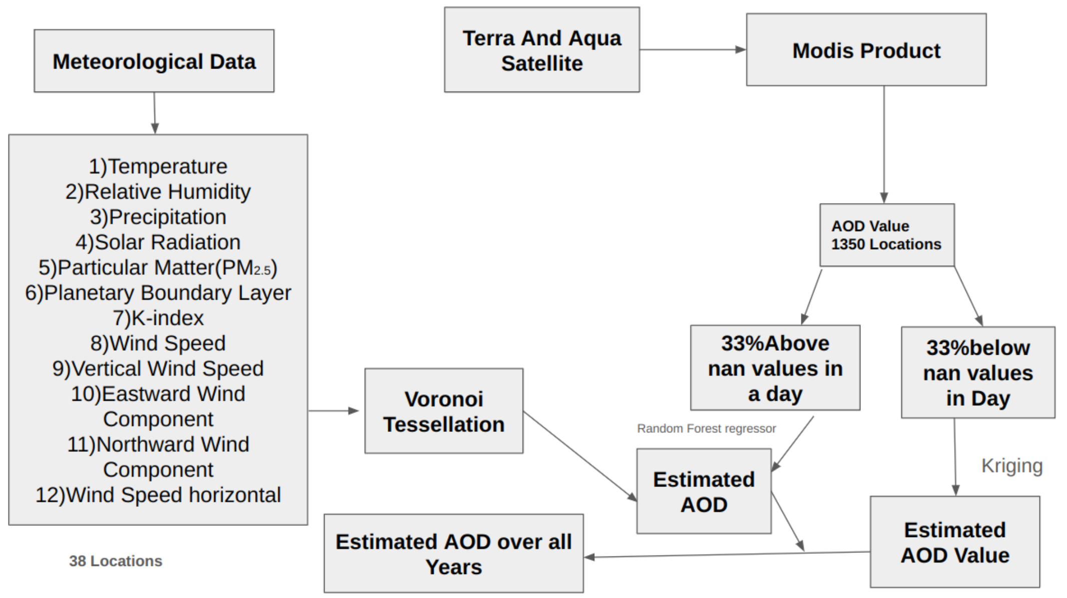


Figure 5.3: Spatial and Temporal Data Matching

Chapter 6

Method

In this study, the methodology for estimating long-term daily Aerosol Optical Depth (AOD) data involves a multi-step process. The first step includes obtaining continuous AOD data from the Moderate Resolution Imaging Spectroradiometer (MODIS). Subsequently, days with missing AOD values are identified, and those with more than 33% missing values are excluded. To address gaps in the remaining dataset, Kriging interpolation is applied, resulting in a continuous and gapless AOD dataset. Ground station data, encompassing meteorological variables and geographic information, is then acquired and seamlessly integrated with the continuous AOD dataset. The integrated dataset serves as input features for the Random Forest algorithm, where AOD is designated as the target variable. The dataset is appropriately split into training and testing sets for model training and evaluation. This comprehensive approach ensures the incorporation of both satellite and ground-based data, enhancing the accuracy and reliability of the long-term daily AOD estimates.

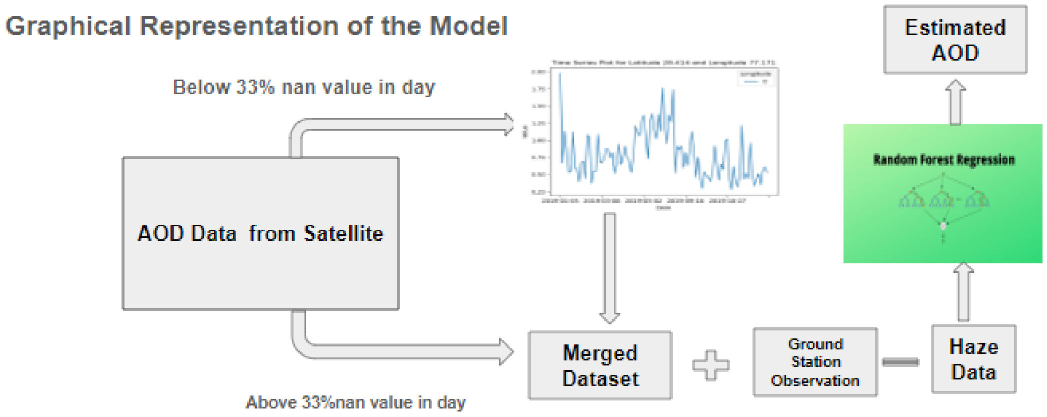

Graphical Representation of the Model
Below 33% nan value in day
AOD Data from Satellite
Above 33%nan value in day
Merged Dataset
Ground Station Observation
Haze Data
Random Forest Regression
Estimated AOD

Figure 6.1: Model Architecture

The study's methodology is visually represented in Figure 6.1 via a flowchart. We utilized MODIS satellite data from both AQUA and Terra for the period spanning 2018 to 2022. The initial step involves acquiring daily gridded MODIS AOD satellite data. These AOD datasets were then extracted, focusing on the Delhi region with a resolution of 1km by 1km. A total of 1350 locations within Delhi were selected for our study.

To access the Delhi shape file, we obtained it from the following link: [1]. From the satellite image we get the aod value of these location using haversine formula

Let the central angle θ between any two points on a sphere be:

$$\theta = \frac{d}{r}$$

where:

d is the distance between the two points along a great circle of the sphere (see spherical distance),

r is the radius of the sphere.

The haversine formula allows the haversine of θ (that is, $\operatorname{hav}(\theta)$) to be computed directly from the latitude (represented by φ) and longitude (represented by λ) of

[1] https://www.societyforplanners.in/2021/01/gis-shapefiles-states-ut.html

the two points:

$$\operatorname{hav}(\theta) = \operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)$$

where:

φ_1, φ_2 are the latitude of point 1 and latitude of point 2,

λ_1, λ_2 are the longitude of point 1 and longitude of point 2.

Finally, the haversine function $\operatorname{hav}(\theta)$, applied above to both the central angle θ and the differences in latitude and longitude, is:

$$\operatorname{hav}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos(\theta)}{2}$$

The haversine function computes half a versine of the angle θ.

To solve for the distance d, apply the archaversine (inverse haversine) to $h = \operatorname{hav}(\theta)$ or use the arcsine (inverse sine) function:

$$d = r \operatorname{archav}(h) = 2r \arcsin\left(\sqrt{h}\right)$$

or more explicitly:

$$\begin{aligned}
d &= 2r \arcsin\left(\operatorname{hav}(\varphi_2 - \varphi_1) + (1 - \operatorname{hav}(\varphi_1 - \varphi_2) - \operatorname{hav}(\varphi_1 + \varphi_2)) \cdot \operatorname{hav}(\lambda_2 - \lambda_1)\right) \\
&= 2r \arcsin\Bigg(\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) \\
&\quad + (1 - \sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) - \sin^2\left(\frac{\varphi_2 + \varphi_1}{2}\right)) \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)\Bigg) \\
&= 2r \arcsin\Bigg(\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) \\
&\quad + \cos(\varphi_1) \cdot \cos(\varphi_2) \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)\Bigg).
\end{aligned}$$

6.1 Kriging

The dataset was partitioned into two categories based on the percentage of days with missing AOD values, distinguishing between those with less than 33% and those with greater than 33% missing values. For instances where the missing values were below 33%, we applied the kriging method to impute the absent AOD values [12] .

In statistics, originally in geostatistics, *Kriging*, also known as Gaussian process regression, is a method of interpolation based on Gaussian process governed by prior covariances. Under suitable assumptions of the prior, kriging gives the best linear unbiased prediction (*BLUP*) at unsampled locations. Interpolating methods based on other criteria such as smoothness (e.g., smoothing spline) may not yield the BLUP. The method is widely used in the domain of spatial analysis and computer experiments. The technique is also known as Wiener–Kolmogorov prediction, after Norbert Wiener and Andrey Kolmogorov.

A value from location x_1 (generic coordinates) is interpreted as a realization $z(x_1)$ of the random variable $Z(x_1)$. In space A, where samples are dispersed, there are N realizations of the random variables $Z(x_1), Z(x_2), \ldots, Z(x_N)$, correlated between themselves.

The set of random variables constitutes a random function, of which only one realization is known – the set $z(x_i)$ of observed AOD data. With only one realization of each random variable, determining any statistical parameter of the individual variables or the function is theoretically impossible. The proposed solution in the geostatistical formalism assumes various degrees of stationarity in the random function to make the inference of some statistical values possible.

For instance, assuming the homogeneity of samples in area A where the AOD variable is distributed, the hypothesis that the first moment is stationary (i.e., all random variables have the same mean) allows estimating the mean by the arithmetic mean of sampled AOD values.

The hypothesis of stationarity related to the second moment is defined such that the correlation between two random variables depends only on the spatial distance between them and is independent of their location. If $\mathbf{h} = x_2 - x_1$ and $|\mathbf{h}| = h$, then:

$$C(Z(x_1), Z(x_2)) = C(Z(x_i), Z(x_i + \mathbf{h})) = C(h),$$

$$\gamma(Z(x_1), Z(x_2)) = \gamma(Z(x_i), Z(x_i + \mathbf{h})) = \gamma(h).$$

For simplicity, we define $C(x_i, x_j) = C(Z(x_i), Z(x_j))$ and $\gamma(x_i, x_j) = \gamma(Z(x_i), Z(x_j))$.

This hypothesis allows inferring two measures – the variogram and the covariogram:

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} (Z(x_i) - Z(x_j))^2,$$

$$C(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} (Z(x_i) - m(h))(Z(x_j) - m(h)),$$

where:

$m(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} (Z(x_i) + Z(x_j))$;

$N(h)$ denotes the set of pairs of observations (i, j) such that $|x_i - x_j| = h$, and $|N(h)|$ is the number of pairs in the set. In this set, (i, j) and (j, i) denote the same element. Generally, an "approximate distance" h is used, implemented using a certain tolerance.

6.2 Random Forest

Random Forest (RF) constitutes a methodology that involves the construction of an extensive ensemble of regression trees[3]. At a conceptual level, regression trees strive to identify the optimal binary split among covariates (features) based on mean-squared error, subsequently partitioning the data accordingly. This recursive process persists until a predefined condition is met, such as having only one observation remaining, precluding further splits. Two pivotal elements of RF encompass: (1) bagging, or bootstrap aggregation, where each tree is trained on a random sample (with replacement) drawn from the original dataset, and (2) the algorithm's consideration of a random subset (denoted as m) of the original p predictors at each node to determine the optimal split. Given that a single decision tree might tend to overfit the data, the incorporation of these two components and the averaging of multiple trees mitigate variance, striking a balance with low bias . Following the imputation of missing AOD values, we incorporated ground station meteorological measurement data as detailed in Section 5.1.3 to enhance the accuracy of the remaining missing AOD value estimations. In this context, Voronoi tessellation[2] was employed to derive predictor variables of the satellite data, enhancing the overall predictive capabilities of the model. This was achieved through the application of the Random Forest Regressors. [21] In this process, we utilize Voronoi tessellation to determine the nearest ground station location for predicting the missing AOD data.[3]

[2] https://en.wikipedia.org/wiki/Voronoi_diagram

[3] https://en.wikipedia.org/wiki/Voronoi_diagram

Chapter 7

Results

Given that every AOD (Aerosol Optical Depth) prediction model ultimately aims to provide an accurate estimation of a continuous value, standard error metrics commonly used for regression tasks, such as mean squared error (MSE), and root mean squared error (RMSE), R-squared value can be applied. However, since AOD values are strongly dependent on various environmental factors, the mean and variance of a given dataset can vary substantially for different measurement locations. A region with consistently high aerosol concentrations, for example, may exhibit AOD values that vary smoothly over time and are therefore much easier to predict. Using a simple difference-based error metric might not adequately consider the differences in prediction difficulty. An overview of the loss metrics is given below:

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

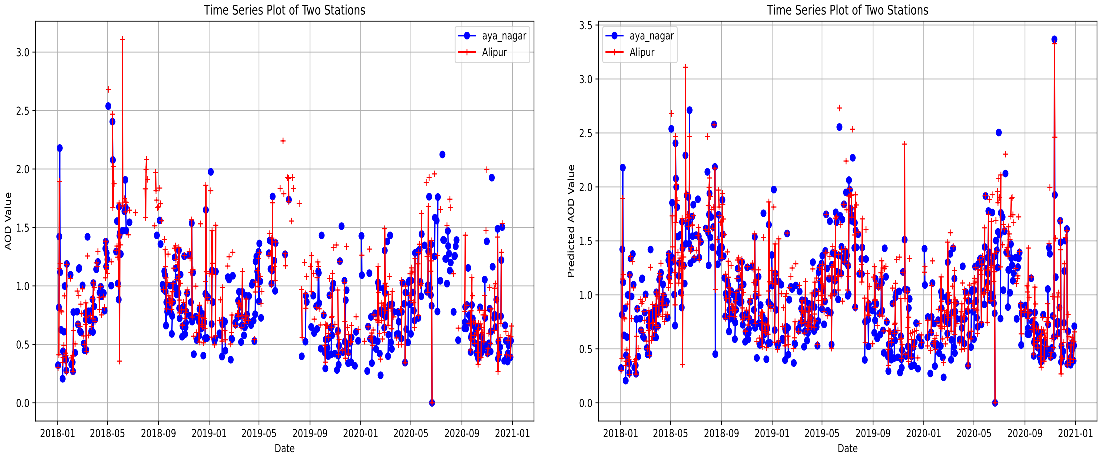
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Here, y_i represents the actual AOD values, $\hat{y}_i$ denotes the predicted AOD values, $\bar{y}$ denotes the mean AOD values, and n is the number of data points.

This approach accounts for the varying difficulty in predicting AOD values across different locations, providing a more nuanced and fair assessment of model performance. The table shows the results for our fitted model (see Table 7.2).

Figure 7.1 (a) depicts the time series of Aerosol Optical Depth (AOD) values



(a)

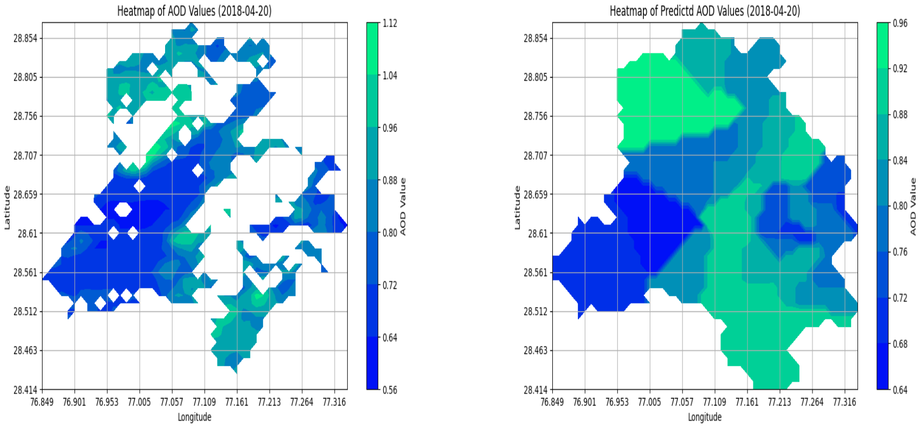
(b)

Figure 7.1: (a) Actual AOD value , (b) Predicted AOD values

for the two stations Table 7.1. It is evident that there are numerous missing values in both stations. Despite the presence of missing data, both stations exhibit a similar pattern, characterized by a peak in AOD values during the spring months (March-May) followed by a decline in the summer and fall. This observed pattern may be attributed to seasonal variations in meteorological conditions, impacting the transport and dispersion of aerosols.

After applying kriging and random forest regression models and fitting them to the data, the resulting time series plot for both stations is presented in Figure 7.1 (b). Observing this plot, we draw the conclusion that the missing values have been effectively imputed by these models.

From Figure 7.2 (a) While the AOD map provides valuable insights into the spatial distribution of pollution across the Delhi region, it is important to acknowledge the presence of several missing data points. These missing values, primarily concentrated in specific areas of the Delhi region, limit the comprehensiveness of the analysis and necessitate the use of imputation techniques for generating a complete image. Despite the availability of a high-resolution AOD map for the Delhi region, the presence of a significant number of missing data points, particularly in certain areas of the Delhi region, poses a challenge in accurately assessing the overall air quality and identifying potential pollution hotspots. In addressing the intricate spatial distribution and potential non-linear relationships inherent in AOD values and environmental factors, we employed a hybrid imputation strategy that com-


Heatmap of AOD Values (2018-04-20)
Latitude
Longitude
AOD Value
Heatmap of Predictd AOD Values (2018-04-20)
Latitude
Longitude
AOD Value

(a)

(b)

Figure 7.2: (a) Actual AOD value of the Delhi region on the date 20/04/2018, (b) Predicted AOD values of the Delhi region on the date 20/04/2018 using random forest regressor.

Table 7.1: Satellite Image Coordinates for Aya Nagar and Alipur

Place	Latitude	Longitude
Aya Nagar	28.414	77.171
Alipur	28.854	77.192

bines Kriging and Random Forest regressors.Kriging served as a robust foundation for capturing spatial dependence among AOD measurements. The integration of Random Forest regressors further enhanced the accuracy of imputation, particularly in regions with limited data, by accounting for potential non-linear relationships. Highlighting the advantages of our hybrid approach. Our hybrid methodology capitalizes on the strengths of both Kriging and Random Forest regressors, resulting in a more nuanced understanding of AOD distribution. By amalgamating the spatial insights from Kriging with the non-linear learning capabilities of Random Forest, our approach achieved heightened accuracy in imputing missing AOD values, thereby enhancing the reliability of our air quality analysis.The superior performance of our hybrid approach, as illustrated in Figure 7.2(b), outshines traditional Kriging or Random Forest imputation methods, particularly in areas exhibiting complex spatial patterns or non-linear relationships between AOD and environmental factors.The validation plot of the Alipur station 7.3 Shows that the model predict AOD accurately.

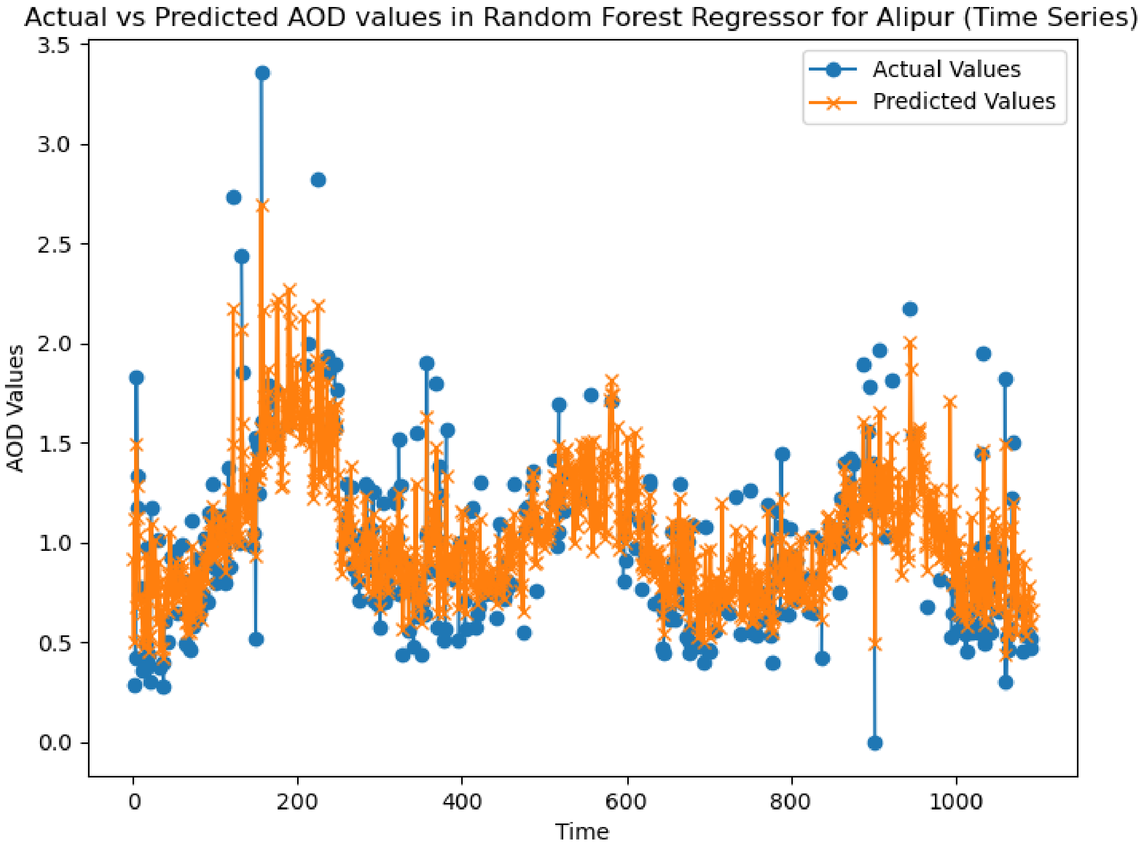


Figure 7.3: Alipur Station Prediction Plot

Table 7.2: Performance Metrics for the Model

Metric	RMSE	MSE	R-squared value
Value	0.06	0.0007	0.79

Chapter 8

Conclusion & Future Work

8.1 Conclusion

The escalating industrialization-related pollution sources globally, particularly in developing nations, necessitate enhanced air quality monitoring, especially for particulate matter (PM10 and PM2.5). However, establishing extensive monitoring stations across diverse locations is impractical due to logistical, temporal, and financial constraints.

This study addresses these challenges by successfully demonstrating the estimation of Aerosol Optical Depth (AOD) levels using MODIS satellite data. Multiple regression models, leveraging both Kriging and Random Forest regressors, were developed to predict PM10 and PM2.5 concentrations based on MODIS-derived AOD data. This innovative approach provides a cost-effective and efficient means of estimating particulate matter concentrations.

The utilization of Kriging and Random Forest regressors in this study underscores the importance of leveraging advanced spatial interpolation techniques and machine learning methodologies. Results from the linear regression models indicated suboptimal performance, revealing the limitations of predicting particulate matter concentration from MODIS-based AOD levels alone. However, the integration of Kriging and Random Forest regressors demonstrated a notable enhancement, particularly after considering meteorological factors. This improvement in the regression coefficients (R2) highlights the efficacy of combining these techniques for more accurate predictions.

The superior performance of the integrated Kriging and Random Forest models over simple regression models emphasizes the significance of accounting for spatial variability and non-linear relationships in accurately predicting AOD. Statistically significant outcomes were observed in most cases, reinforcing the reliability and robustness of the developed models for predicting Aerosol optical depth.

In conclusion, this study presents a novel approach that integrates Kriging and Random Forest regressors for estimating AOD value using MODIS-derived AOD data. The incorporation of meteorological factors, alongside these advanced techniques, significantly enhances predictive accuracy, making this methodology a valuable tool for air quality assessment and public health studies in regions facing challenges in establishing extensive monitoring networks.

8.2 Future Work

8.2.1 Neural Process

There are several avenues to enhance the predictive capabilities of Neural Processes[18] (NPs) for Aerosol Optical Depth (AOD) prediction. First, there is a need to extend the current model to effectively incorporate temporal dependencies in AOD data. This expansion would involve adapting the Neural Processes framework to handle time-series aspects, enabling the model to capture temporal patterns and trends in aerosol distribution over time. Additionally, addressing the spatial variability in AOD is crucial. Enhancements to the model architecture should be explored to capture spatial correlations and variations in aerosol distribution. Incorporating spatial information into Neural Processes would contribute to a more robust prediction of AOD across diverse geographical regions.

8.2.2 STA-GAN:A Spatio-Temporal Attention Generative Adversarial Network for Missing Value Imputation in Satellite Data[25]

Figure 8.1 illustrates the structure of the STA-GAN model [25], comprising the spatio-temporal attention (STA) module and the Generative Adversarial Network (GAN) module. Initially, the STA module learns short-term temporal dependence and dynamic spatial dependence in satellite data using GAT, resulting in the production of the short-term temporal dependence representation matrix F and the dynamic spatial dependence representation matrix S. Subsequently, the generator and

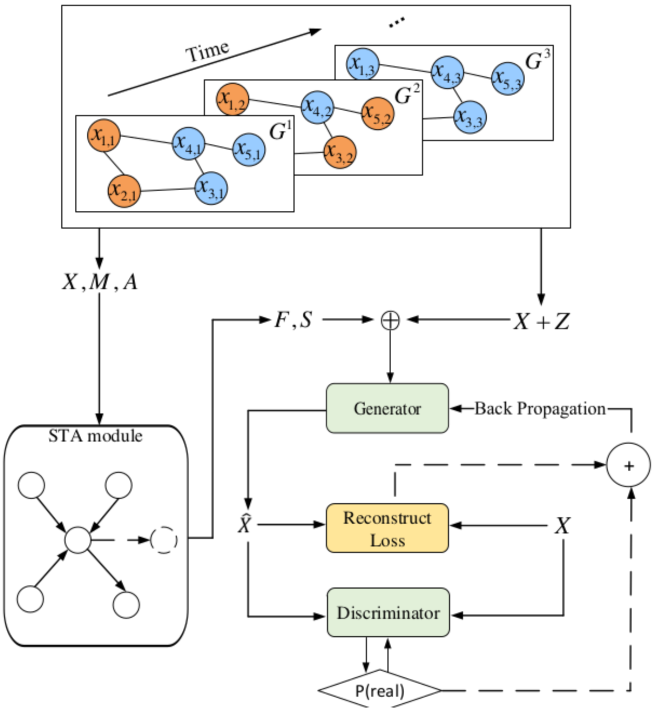


Figure 8.1: Illustration of the STA-GAN model structure.

discriminator of the GAN are trained by incorporating the learned spatio-temporal dependence features. Finally, the missing data is filled using the generated data from the GAN module.

8.2.3 Vision Trasformer

Aerosol Optical Depth (AOD) prediction, the Temporo-Spatial Vision Transformer (TSViT) [24] serves as a robust foundation, paving the way for several avenues of future exploration. One promising avenue involves the integration of multi-sensor fusion, where TSViT could benefit from incorporating data from ground-based monitoring systems, ERA5, MERRA-2, and other relevant datasets. The exploration of fusion strategies aims to harness the synergies among these diverse sources, enhancing the model's contextual understanding and further elevating its predictive accuracy.

A key focus in the future work will be on enhancing TSViT's adaptability to dynamic environmental contexts. This entails the integration of real-time meteorological data and environmental indices to capture nuanced temporal variations in atmospheric conditions, providing a more responsive and accurate representation of AOD levels. Moreover, TSViT will be tailored for urban air quality monitoring, addressing the distinctive challenges posed by urban environments. This includes a thorough investigation into the interplay between satellite-derived AOD data and ground-based sensor observations, offering detailed insights into local air quality dynamics.

Fine-grained spatial analysis of AOD will be a priority, extending TSViT's spatial capabilities to discern intricate patterns in aerosol distribution. Strategies involving refined patching mechanisms and overlapping patches will empower the model to capture fine-scale spatial details, contributing to a more comprehensive understanding of aerosol dispersion.

Transfer learning strategies will be explored for AOD prediction across diverse geographical regions. The goal is to transfer knowledge from well-modeled regions to areas with limited satellite coverage, fostering adaptability to varying environmental conditions and improving prediction accuracy.

8.2.4 SERT:A Transfomer Based Model for Spatio-Temporal Sensor Data with Missing Values for Environmental Monitoring[16]

we intend to build upon the foundations laid by our proposed SERT (Spatio-temporal Encoder Representations from Transformers) and SST-ANN (Sparse Spatio-Temporal Artificial Neural Network) models for multivariate spatio-temporal forecasting. Our primary focus will be on advancing and refining these models to address specific challenges and enhance their capabilities.

One key direction for future exploration involves improving the interpretability of both SERT and SST-ANN. We aim to develop techniques that provide deeper insights into the decision-making processes of these models, making their predictions more transparent and understandable for end-users and stakeholders.

Another crucial aspect to be addressed in future work is the continued enhancement of the models' ability to handle missing data. We plan to investigate and implement advanced strategies, particularly in scenarios with increased sparsity or dynamic changes, to further improve the robustness and reliability of our models in the face of incomplete data.

Bibliography

[1] Antonis Analitis, Klea Katsouyanni, Konstantina Dimakopoulou, Evangelia Samoli, Aristidis K Nikoloulopoulos, Yannis Petasakis, Giota Touloumi, Joel Schwartz, Hugh Ross Anderson, Koldo Cambra, et al. Short-term effects of ambient particles on cardiovascular and respiratory mortality. *Epidemiology*, 17(2):230–233, 2006.

[2] Shrutilipi Bhattacharjee, Pabitra Mitra, and Soumya K. Ghosh. Spatial interpolation to predict missing attributes in gis using semantic kriging. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4771–4780, 2014.

[3] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[4] Cole Brokamp, Roman Jandarov, Monir Hossain, and Patrick Ryan. Predicting daily urban fine particulate matter concentrations using a random forest model. *Environmental science & technology*, 52(7):4173–4179, 2018.

[5] Robert D Brook, Sanjay Rajagopalan, C Arden Pope III, Jeffrey R Brook, Aruni Bhatnagar, Ana V Diez-Roux, Fernando Holguin, Yuling Hong, Russell V Luepker, Murray A Mittleman, et al. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the american heart association. *Circulation*, 121(21):2331–2378, 2010.

[6] Mohamed Eltahan and Karim Moharm. Atmospheric aerosol prediction over egypt with lstm-rnn using nasa's merra-2. In *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 93–98. IEEE, 2020.

[7] Amit K Gorai, Francis Tuluri, and Paul B Tchounwou. A gis based approach for assessing the association between air pollution and asthma in new york state, usa. *International journal of environmental research and public health*, 11(5):4845–4869, 2014.

[8] Pawan Gupta, Maudood N Khan, Arlindo da Silva, and Falguni Patadia. Modis aerosol optical depth observations over urban areas in pakistan: quantity and quality of the data for air quality monitoring. *Atmospheric pollution research*, 4(1):43–52, 2013.

[9] Jana Handschuh, Thilo Erbertseder, and Frank Baier. Systematic evaluation of four satellite aod datasets for estimating pm2. 5 using a random forest approach. *Remote Sensing*, 15(8):2064, 2023.

[10] Michael D King, W Paul Menzel, Yoram J Kaufman, Didier Tanré, Bo-Cai Gao, Steven Platnick, Steven A Ackerman, Lorraine A Remer, Robert Pincus, and Paul A Hubanks. Cloud and aerosol properties, precipitable water, and profiles of temperature and water vapor from modis. *IEEE Transactions on Geoscience and Remote Sensing*, 41(2):442–458, 2003.

[11] Rajesh Kumar, Sachin D Ghude, Mrinal Biswas, Chinmay Jena, Stefano Alessandrini, Sreyashi Debnath, Santosh Kulkarni, Simone Sperati, Vijay K Soni, Ravi S Nanjundiah, et al. Enhancing accuracy of air quality and temperature forecasts during paddy crop residue burning season in delhi via chemical data assimilation. *Journal of Geophysical Research: Atmospheres*, 125(17):e2020JD033019, 2020.

[12] Long Li, Runhe Shi, Lu Zhang, Jie Zhang, and Wei Gao. The data fusion of aerosol optical thickness using universal kriging and stepwise regression in east china. In *Remote Sensing and Modeling of Ecosystems for Sustainability XI*, volume 9221, pages 219–229. SPIE, 2014.

[13] Natalie Mahowald. Aerosol indirect effect on biogeochemical cycles and climate. *Science*, 334(6057):794–796, 2011.

[14] Patrick L McDermott and Christopher K Wikle. An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat*, 6(1):315–330, 2017.

[15] Juan Pedro Mellado, Chiel C van Heerwaarden, and Jade Rachele Garcia. Near-surface effects of free atmosphere stratification in free convection. *Boundary-Layer Meteorology*, 159:69–95, 2016.

[16] Amin Shoari Nejad, Rocío Alaiz-Rodríguez, Gerard D McCarthy, Brian Kelleher, Anthony Grey, and Andrew Parnell. Sert: A transfomer based model for spatio-temporal sensor data with missing values for environmental monitoring. *arXiv preprint arXiv:2306.03042*, 2023.

[17] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california counties: results from calfine. *Environmental health perspectives*, 114(1):29–33, 2006.

[18] Shenghao Qin, Jiacheng Zhu, Jimmy Qin, Wenshuo Wang, and Ding Zhao. Recurrent attentive neural process for sequential data. *arXiv preprint arXiv:1910.09323*, 2019.

[19] VCPJ Ramanathan, Paul J Crutzen, JT Kiehl, and Daniel Rosenfeld. Aerosols, climate, and the hydrological cycle. *science*, 294(5549):2119–2124, 2001.

[20] John H Seinfeld and Spyros N Pandis. *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons, 2016.

[21] Yanchuan Shao, Zongwei Ma, Jianghao Wang, and Jun Bi. Estimating daily ground-level pm2. 5 in china with random-forest-based spatiotemporal kriging. *Science of The Total Environment*, 740:139761, 2020.

[22] Rod Simpson, Gail Williams, Anna Petroeschevsky, Trudi Best, Geoff Morgan, Lyn Denison, Andrea Hinwood, Gerard Neville, and Anne Neller. The short-term effects of air pollution on daily mortality in four australian cities. *Australian and New Zealand journal of public health*, 29(3):205–212, 2005.

[23] Yuxuan Su, Junyu Li, Lilong Liu, Xi Guo, Liangke Huang, and Mingyun Hu. Application of cnn-lstm algorithm for pm2. 5 concentration forecasting in the beijing-tianjin-hebei metropolitan area. *Atmosphere*, 14(9):1392, 2023.

[24] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10418–10428, 2023.

[25] Shuyu Wang, Wengen Li, Siyun Hou, Jihong Guan, and Jiamin Yao. Sta-gan: A spatio-temporal attention generative adversarial network for missing value imputation in satellite data. *Remote Sensing*, 15(1):88, 2022.

[26] Jianfeng Zhang, Yan Zhu, Xiaoping Zhang, Ming Ye, and Jinzhong Yang. Developing a long short-term memory (lstm) based model for predicting water table depth in agricultural areas. *Journal of hydrology*, 561:918–929, 2018.

[27] Qi Zhang, Yang Han, Victor OK Li, and Jacqueline CK Lam. Deep-air: A hybrid cnn-lstm framework for fine-grained air pollution estimation and forecast in metropolitan cities. *IEEE Access*, 10:55818–55841, 2022.