$PM_{2.5}$ Estimation using Machine Learning Models and Satellite Data: A Literature Review

Mitra Unik[1], Imas Sukaesih Sitanggang[2], Lailan Syaufina[3], I Nengah Surati Jaya[4]
Department of Computer Science, Institut Pertanian Bogor, IPB Bogor, Indonesia[1, 2]
Department of Silviculture, Institut Pertanian Bogor, IPB Bogor, Indonesia[3]
Department of Forest Management, Institut Pertanian Bogor, IPB Bogor, Indonesia[4]

***Abstract*—Most researchers are beginning to appreciate the use of remote sensing satellites to assess $PM_{2.5}$ levels and use machine learning algorithms to automate the collection, make sense of remote sensing data, and extract previously unseen data patterns. This study reviews delicate particulate matter ($PM_{2.5}$) predictions from satellite aerosol optical depth (AOD) and machine learning. Specifically, we review the characteristics and gap-filling methods of satellite-based AOD products, sources and components of $PM_{2.5}$, observable AOD products, data mining, and the application of machine learning algorithms in publications of the past two years. The study also included functional considerations and recommendations in covariate selection, addressing the spatiotemporal heterogeneity of the $PM_{2.5}$ -AOD relationship, and the use of cross-validation, to aid in determining the final model. A total of 79 articles were included out of 112 retrieved records consisting of articles published in 2022 totaling 43 articles, as of 2023 (until February) totaling 19 articles, and other years totaling 18 articles. Finally, the latest method works well for monthly $PM_{2.5}$ estimates, while daily $PM_{2.5}$ and hourly $PM_{2.5}$ can also be achieved. This is due to the increased availability and computing power of large datasets and increased awareness of the potential benefits of predictors working together to achieve higher estimation accuracy. Some key findings are also presented in the conclusion section of this article.**

Keywords*—*AOD; machine learning; $PM_{2.5;}$ remote sensing; pollutant

I. Introduction

Interest in the study of $PM_{2.5}$ (particulate matter aerodynamic diameter ≤ 2.5 μm/m^3) concentration estimates from various outdoor and indoor particle sources has increased dramatically recently, as evidenced by the number of academic journals that have published articles on it. Studies identified the impact of $PM_{2.5}$ contamination on humans as the initial problem of various adverse effects on the health of fetal growth during pregnancy to early death [1]. Direct and long-term exposure can significantly impact climate change, visibility degradation, ecosystem disruption, and social, ecological, and economic impacts [2]–[4].

$PM_{2.5}$ monitoring is a critical need for public health, especially in densely populated areas, where exposure to airborne particles poses significant health risks [5]. Ground station monitoring is the most direct and accurate method of $PM_{2.5}$ monitoring. However, it is impossible to fully identify the spatial distribution and obtain historical measurements of $PM_{2.5}$ concentrations across the region. Most researchers are beginning to appreciate the use of remote sensing satellites to assess $PM_{2.5}$ levels. Estimating $PM_{2.5}$ concentrations using Aerosol optical depth (AOD) as a remote sensing satellite derivative can be used to fill the gap of spatial and temporal data gaps left by ground stations [6]. Various remote sensing satellite sensors, such as Moderate Resolution Imaging Spectrometer (MODIS) [7], [8] The Visible-infrared Imaging Radiometer Suite (VIIRS) [9], [10], the Advanced Himawari Imager [11], [12] the Advanced Geosynchronous Radiation Image (AGRI) [13], [14] have been applied to estimate $PM_{2.5}$ concentrations.

Models for predicting $PM_{2.5}$ concentrations can be useful for filling data gaps from existing monitoring networks. Air pollutant concentration prediction methods can generally be classified into three categories: numerical, statistical, and artificial intelligence (AI) models. Numerical models simulate the physical and chemical changes and transport processes of atmospheric pollutants by specifying and solving complex differential equations. Recent representative numerical models include Community Multiscale Air Quality (CMAQ) and Weather Research and Forecasting coupled with Chemistry (WRF-Chem). The accuracy of these models relies heavily on detailed emission data from pollutant sources, which often need to be made more precise and available. In addition, the complex modeling process requires more time and computing power [15]. Therefore, it is necessary to develop a faster and more accurate model to improve the prediction of air pollutants.

Statistical models have not involved complex physical changes, chemical reactions, and transportation processes. Statistical models rely entirely on data-driven mining of internal relationships to historical data. Therefore, the computational effort is significantly lower compared to numerical models. It is easy to implement classical statistical models such as autoregressive integrated moving average (ARIMA) [16] and autoregressive moving average (ARMA). However, these models are suitable for small data sets and univariate time series models. In addition, these models are based on linear assumptions that require strict stationarity of the data. Therefore, capturing nonlinear relationships in the data is inherently complex. These limitations greatly restrict the performance and applicability of classical statistical models in air pollution forecasting.

In contrast, adopting machine learning models in remote sensing is considered the optimal solution for predicting $PM_{2.5}$ concentration time series due to its advantages of flexible nonlinear regression capabilities and classification features based on large data sets with complex data relationships between many variables. [17], [18]. The initial study that utilized a Neural Network (NN) to tackle the intricate correlation between AOD-PM$_{2.5}$ [19]. Since the 1990s, Machine Learning algorithms have been used to automate the collection, understand remote sensing data, and extract previously unseen data patterns [20], [21].

Machine learning capabilities make it possible to non-parametrically examine the relationship between predictors of pollutant concentrations and measured pollutant concentrations [22], [23]. A number of research investigations have indicated that machine learning [17], [24], [25] such as deep learning [26], Random Forest [27] , and deep ensemble models [28], have a remarkable ability to estimate PM$_{2.5}$ concentrations at various temporal and spatial scales. Several models have been developed to predict indoor [29] and atmospheric PM$_{2.5}$ concentrations based on data obtained from air quality monitoring stations, such as meteorological variables from weather stations such as air temperature (T), relative humidity (RH), wind speed (WS), wind direction, Precipitation (PRE). Land variables, such as NDVI. Variables related to population) such as population density, road network density, height, and the number of buildings, and others, including data on PM$_{2.5}$, carbon monoxide (CO), ozon (O$_3$), nitrogen oxides (NO), nitrogen dioxide (NO2), and sulfur dioxide (SO$_2$) [10], [11], [14], [26], [30].

The success of PM$_{2.5}$ concentration estimation studies using machine learning and satellite remote sensing data depends on the quantity and quality of the researcher's domain knowledge, regional knowledge, and time spent. This review article aims to summarize the literature on the use of machine learning and satellite remote sensing in estimating large-scale and long-term PM$_{2.5}$ concentrations. This literature review includes articles from 2022 to 2023 related to this crucial topic. However, some articles that can provide insights into various remote sensing technologies on PM$_{2.5}$, air pollution, and other specific studies were also added without being limited by the year of publication period. Specific search terms and study selection are illustrated in the second section to summarize the current state of development in estimating PM$_{2.5}$ concentrations. The third section investigates factors affecting PM$_{2.5}$ concentrations, levels, and model measurements. The following section is a personal presentation on using machine learning models.

II. Literature Search and Selection

In line with the multidisciplinary research topics, several other disciplines, ranging from computer science, forestry, remote sensing, atmosphere, and disaster, which intersect with the main topic without being limited by the year of publication period to provide additional insight, are included. Four general stages of literature search and determination were conducted, such as:

- Identification: The initial set was conducted by identifying keywords to search for articles relevant to the topic of this literature review from electronic databases Web of Science, Google Scholar, and sources of Elsevier, ScienceDirect, and Springer, both from National and International journals. Based on the topic raised, this study needs to summarize (1) literature from indoor PM$_{2.5}$ concentration research, (2) specific indoor PM$_{2.5}$ sources (cooking, cigarettes, vacuum cleaners, and more.), and (3) monitoring via landline networks. Application of keywords as follows: "Estimating PM$_{2.5}$", "machine learning," "satellite remote sensing," "PM$_{2.5}$", and "outdoor." Findings of article titles corresponding to the research topic were then stored and evaluated. The search continued by checking for other articles cited or quoted in this set and removing double-identified documents. Due to the core topic of this literature review, we focused on published articles from January 2022 to February 2023.

- Screening: The articles found were screened by labeling them as relevant or not to this study after checking the abstracts. These potential papers were then carefully reviewed to ensure their eligibility as references in this literature review.

- Eligibility: Eligibility was determined by reading the main findings, use of data, results, and discussion. The authors considered journal articles and books published by reputable publishers as high-quality research and included them in summary. The authors used "Scimago Journal & Country Rank" to check the rankings of the included articles.

- Inclusion: The research then lists literature articles that correspond to the main topic.

Our initial search yielded 112 articles. After passing the initial screening to eligibility assessment, this study used 61 primary and 18 other articles.

is a summary flowchart of the following literature search and selection statistics:

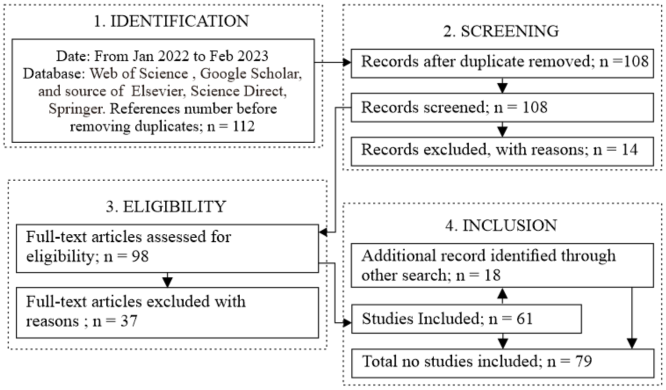
1. IDENTIFICATION
Date: From Jan 2022 to Feb 2023
Database: Web of Science , Google Scholar, and source of Elsevier, Science Direct, Springer. References number before removing duplicates; n = 112
2. SCREENING
Records after duplicate removed; n =108
Records screened; n = 108
Records excluded, with reasons; n = 14
3. ELIGIBILITY
Full-text articles assessed for eligibility; n = 98
Full-text articles excluded with reasons ; n = 37
4. INCLUSION
Additional record identified through other search; n = 18
Studies Included; n = 61
Total no studies included; n = 79

Fig. 1. Literature search and selection flow chart.

illustrates the number of references in the literature review. The collected articles from 2022 totaled 43 articles, 2023 (up to February) totaled 19 articles, and the other years totaled 18 articles.

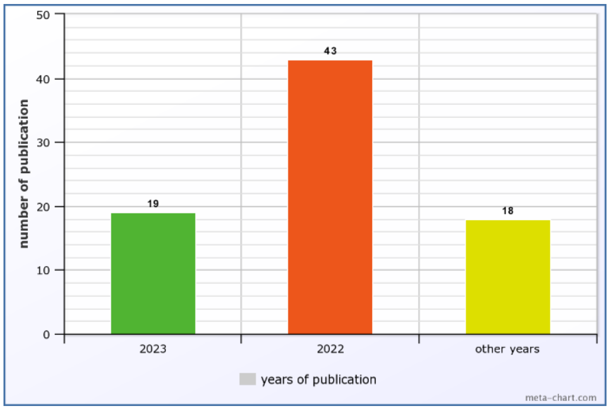

Fig. 2. Number of literature review references.

III. Remote Sensing Technique

A. Moderate Resolution Imaging Spectroradiometer(MODIS)

MODIS is an instrument on the Aqua and Terra satellites capable of detecting small changes in surface reflectance due to changes in $PM_{2.5}$ concentrations. Reflectance changes estimate $PM_{2.5}$ concentrations through a statistical approach that does not require calibration or data collection from ground-level locations. In addition, this method is more resistant to noise than other methods [31]. The MODIS instrument captures data from 36 spectral bands with wavelengths ranging from 0.4 to 14.385 µm and observes the entire Earth's surface every one to two days. The Terra satellite follows a north-south orbit and crosses the equator in the morning, while the Aqua satellite travels in the opposite direction and flies over the equator in the afternoon. Various sources provide various MODIS data products:

- MODIS level 1 data, geolocation, cloud mask and atmospheric products: http://ladsweb.nascom.nasa.gov/
- MODIS ground products: https://lpdaac.usgs.gov/.
- MODIS cryosphere products:: http://nsidc.org/daac/modis/index.html.
- MODIS ocean color and sea surface temperature products: *http://oceancolor.gsfc.nasa.gov/*.

MODIS images have spatial resolutions of 250 m, 500 m, and 1km. The range of wavelengths between 0.47 and 2.12 µm in various channels is utilized to determine aerosol properties, specifically AOD, in order to estimate $PM_{2.5}$ [7], [8]. A research study based on theoretical analysis of data gathered by a multiangle imaging spectrometer aboard the Terra satellite in the US has shown that the range of particle sizes appropriate for AOD retrieval, which closely corresponds to the particle size range of PM2.5, falls between 0.1-2 nm in the visible and near-infrared wavelength bands.

B. Himawari-8

The Japan Meteorological Agency (JMA) operates Himawari-8, a geosynchronous weather satellite. The satellite was launched on 7 October 2014, and is stationed at 140.7 degrees east longitude, providing uninterrupted observations over the Asia-Pacific region, which includes Southeast Asia, Australia, Japan and the Western Pacific. Himawari-8 carries a suite of advanced instruments to observe the Earth's atmosphere and weather systems. These instruments include the Advanced Himawari Imager (AHI), which provides high-resolution images of the Earth's surface and clouds, and the Himawari Cast data collection system, which receives data from other weather satellites and ground-based weather stations [32]. Himawari-8 can be used to measure Aerosol Optical Depth (AOD) to investigate the diurnal variation of air pollution with high temporal resolution. [12]. Recently, some studies have started to estimate hourly ground-level $PM_{2.5}$ in real-time from Himawari-8 AOD products [33]–[35].

C. Sentinel 5-P

The Sentinel-5 Precursor Satellite (Sentinel-5P) was launched on October 13, 2017, carrying the following TROPOspheric Monitoring Instrument (TROPOMI) to generate global high-coverage total/tropospheric vertical columns of precursors (e.g., NO_2) for $PM_{2.5}$ and PM10. TROPOMI has a legacy to the Ozone Monitoring Instrument (OMI) as well as the Scanning Imaging Absorption spectroMeter for Atmospheric CartograpHY (SCIAMACHY) TROPOMI is a single instrument from the Sentinel-5P spacecraft covering wavelengths from ultraviolet (UV) to ShortWave InfraRed (SWIR). This hyperspectral spectrometer is designed to provide routine observations of key atmospheric constituents including ozone, NO_2 , SO_2 , CO, CH_4 , CH_2O and aerosol properties at high spatial resolution using passive remote sensing methods. [36]. The typical pixel size (near nadir) is defined as 7 × 3,5 km^2 for all spectral bands except UV1 (7 × 28 km^2) and SWIR band (7 × 7 km^2). In terms of accuracy, the evaluation results show that the quality of the TROPOMI atmospheric product meets the requirements in $PM_{2.5}$ pollutant estimation [37], [38].

IV. Predictors Used for Estimation of $PM_{2.5}$ Concentrations

A. Sources of $PM_{2.5}$

There are several types of outdoor $PM_{2.5}$ sources originating from the combustion of fossil materials, such as automotive vehicle exhaust emissions, coal, and biomass combustion [39], industrial activities, soil dust, secondary sulfates, secondary nitrates, as well as through release into the volcanic atmosphere [40]. The sources and concentrations of $PM_{2.5}$ can vary significantly between locations due to the different characteristics of climatic conditions, emission sources, and distribution patterns [41]. Black carbon, aryl hydrocarbons, polycyclic aromatic hydrocarbons, volatile organic hydrocarbons, biological materials, heavy metals, minerals, inorganic ions, and organic compounds are the primary constituents of $PM_{2.5}$, which account for around 79-85% of the entire mass when considered together [42].

B. Explanatory Variables of $PM_{2.5}$

Two characteristics of variables used in $PM_{2.5}$ research are dependent and independent variables. The dependent variable contains $PM_{2.5}$ values (µg/m3) obtained through air quality measurements using ground stations. On the other hand, the independent can contain co-pollutant, meteorological, and anthropic information that can significantly improve the model's accuracy. Regarding this critical difference, independent data is essential information to help estimate PM$_{2.5}$, including AOD.

1) Aerosol optical depth: (AOD is a quantitative measure of the reduction of light by aerosol particles in the Earth's atmosphere. AOD describes how much light from the sun is reduced or blocked by aerosol particles in the atmosphere. The higher the AOD value, the more significant the attenuation of light caused by aerosol particles. AOD can be measured using devices such as spectrometers or photometers [43]. Thus, AOD is an essential predictor of PM$_{2.5}$, according to the close relationship with AOD.

The starting point for knowing satellite AOD about surface PM$_{2.5}$ is through Equation (1) [44], which shows the dependence of PM$_{2.5}$ and cloud-free AOD relationships on various factors:

$$\text{AOD} = \text{PM}_{2.5} \times H \times f(RH) \times \frac{3\text{Q}_{ext,dry}}{4\rho\,\text{r}_{\text{eff}}} = PM2.5\ x\ H\ x\ S \qquad (1)$$

where H is the boundary layer height (BLH), f (RH) is the ratio of the ambient and dry extinction coefficient to the relative humidity (RH), ρ is the aerosol mass density (g m^{-3}), Q$_{ext,dry}$ is the Mie extinction efficiency, and r$_{eff}$ is the effective radius of the particle. S is the specific extinction efficiency (m^2 g^{-1}) of the aerosol at ambient RH. This equation assumes the aerosol is homogeneously distributed throughout the BLH.

The relationship between PM2.5 and AOD could take the form of a multivariate function that is linked to numerous meteorological and spatial factors that influence it. [45], [46]. AOD is a variable that includes changes in PM$_{2.5}$, resulting from a comprehensive combination of emissions, chemical reactions, and others. However, there are still three main differences between AOD and PM$_{2.5}$ data:

- AOD is a unitless value that reflects the total light blackout effect of the aerosol in the column, while PM$_{2.5}$ is the mass concentration at the soil surface.
- In the presence of moisture, water-soluble particles will become more prominent through the water absorption process, thus affecting the light-extinguishing ability of the aerosol.
- PM$_{2.5}$ is only part of aerosols with a diameter equal to or less than $_{2.5}$ g/m^3, but this does not apply to all aerosols.

Some AOD products that can be found:

a) AERONET AOD: The Aerosol Robotic Network (AERONET) is a global aerosol monitoring network widely recognized as the benchmark for evaluating satellite source AID products. It provides long-term AOD ground measurements with low drift (0.01-0.02) and high time resolution (15 minutes). AOD measurements in the 550nm band are not available through AERONET. However, to estimate AOD in this band, the Angstrom exponent is typically used to interpolate AOD values between 440nm and 675nm. AERONET AOD is currently categorised into three quality levels: L1.0, L1.5, and L2.0, which represent unfiltered data, filtered and quality-controlled data, and quality-assured data, respectively. Version 3 of the database is currently under development, which will feature more stringent quality control measures, particularly for cirrus cloud pollution [47].

b) DT AOD: The development of the Dark Target (DT) Algorithm is enabled to obtain AOD values in high vegetation cover, dark soil, and low sea surface albedo environments at 10 Km or 3 Km spatial resolution. DT selects dark pixels with atmospheric reflectance (TOA) of 0.01-0.25 in the 2.12 μm channel to retrieve AOD. DT provides fine (low, medium, high) and coarse three-surface aerosol models. Furthermore, it selects from these three models according to the season and geographical conditions [45]. Collection 6 DT AOD (C6 DT AOD) was established in early 2014 and has completed updates to calibration, cloud mask, and land/ocean symbols. Subsequently, the C6-based C6.1 DT AOD was released to address the continuous changes in surface reflectivity caused by the rapid growth of tall buildings worldwide [48]. Specifically, in a pixel network covering an area of 10km×10km, consisting of ≥ 50% coastal pixels or ≥ 20% water pixels, C6.1 DT will reduce the capture quality to zero and modify it with surface reflections in certain AOD areas [48]. The observed rise in value confirms a 2.17% increase in the correlation coefficient between C6.1 DT and AERONET AOD in certain urban locations, suggesting that C6.1 DT AOD provides a more accurate representation of the urban situation. The unfiltered product for AOD C6.1 DT, lacking quality detection, is denoted as "Image_-Optical_Depth_Land_And_Ocean," whereas the filtered product with Quality Assurance (QA) greater than 1 (for ocean) and QA equal to 3 (for land) is denoted as "Optical_Depth_Land_And_Ocean."

c) DB AOD: The Deep Blue (DB) algorithm is designed to capture AOD at 10 km spatial resolution for environments with high surface albedo in deserts, drylands, and cities. It can overcome the defects of the DT algorithm on shiny surfaces. Contrary to DT, DB first picks up aerosols at 1Km resolution, and then combines the 10Km pixels. The Collection 6 DB AOD (C6 DB AOD) product is named "Enhanced Deep Blue" to differentiate it from C5 and extend to other global layers beyond snow and ice C6.1 DB has the following improvements over C6: (1) reduction of artefacts from heterogeneous terrain, (2) improved elevation terrain surface models, and (3) updated seasonal or regional aerosol models and better smoke detection [48]. The product without quality detection in AOD C6.1 DB is named "Deep_Blue_Aerosol_Optical_Depth_550_Land" and filtered by QA=2 and QA=3 is named "Deep_Blue_Aerosol_Optical_Depth_550_Land_Best_Estimate".

d) MAIAC AOD: MAIAC AOD refers to the atmospheric aerosol optical depth (AOD) product generated by the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm. The MAIAC algorithm is a sophisticated technique for atmospheric correction of satellite imagery, which allows for the retrieval of high-quality AOD data. MAIAC AOD is derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument on board the Terra and Aqua satellites, which are operated by the National Aeronautics and Space Administration (NASA). MAIAC AOD provides high-resolution AOD data with spatial and temporal coverage, making it a valuable tool for studying air pollution and its impacts on human health and the environment.. Therefore, MAIAC has a high level of quality (L2) [49], [50]. In addition, the 1 km resolution is another important feature of MAIAC AOD. MAIAC has uncertainties under extreme conditions, indicating that MAIAC cannot obtain AOD accurately at high altitudes (>4.2 km). Lyapustin [51] showed that MAIAC could not get the AOD accurately at altitudes (>4.2 km). Tao [49] showed that daily bias varies dramatically in areas where airborne transportation and dusting occur. Otherwise, Lyapustin [51] compared MAIAC products with different surface cover types and found significantly different detection precision. The miscalculation of regression coefficients of surface reflectance at different wavelengths caused MAIAC to be systematically overestimated due to particle scattering properties in northwest China's desert areas.

e) Other AOD products: In addition to the aforementioned AOD products, several radiometers offer satellite-based AOD products, such as the Climate Change Initiative (CCI) products from the European Space Agency (ESA), which include AATSR Dual View (AATSR-ADV), AATSR Swansea University (AATSR-SU), AATSROxford-RAL Retrieval of Aerosol and Cloud (AATSR-ORAC), and AATSR-ENSEMBLE (AATSR-EN), as well as AOD products from other sensors, including the Advanced Very High-Resolution Radiometer (AVHRR), Multi-angle Imaging Spectroradiometer (MISR), Sea-viewing Wide Field-of-view Sensor (SeaWiFS), Visible Infrared Imaging Radiometer (VIIRS), and Polarization and Directionality of the Earth's Reflectance (POLDER) [52] and Advanced Himawari Imager (AHI). Some AOD products are not widely used for $PM_{2.5}$ estimation due to low time resolution, poor overall accuracy, or limited application range.

2) Co-pollutant and meteorological variabels: When hydrocarbons (HC) and nitrogen oxides (NOx) react in sunlight, they produce the secondary pollutants ozone (O3) and secondary organic carbon (SOC). The photochemical reaction of gaseous precursors of primary organic carbon (POC) results in the formation of SOC [53]. Meteorological factors affect the dispersion and transport of fine particles. Commonly used meteorological variables are relative humidity (RH), temperature (TEMP), u/v wind, surface pressure (SP), and wind direction (WD) [54]. Additional studies based on observation have demonstrated that the correlation between $PM_{2.5}$ and AOD is influenced by the Planetary Boundary Layer Height (PBLH). When the PBLH is greater, the AOD is also higher; however, the $PM_{2.5}$ concentration is lower. [55]. RH changes aerosol particles, affecting AOD by increasing humidity, hygroscopicity (the ability of a substance to take up water molecules from its surroundings, either by absorption or adsorption), and aerosol particles. Furthermore, evaporation has a strong positive correlation with temperature ($R > 0.6$) and a strong negative correlation with relative humidity ($R < -0.6$).

The survey results summarize the various source variables and individual chemical constituents of the data set used for the $PM_{2.5}$ study. The chemical sources were divided into the categories of natural and anthropogenic-biogenic [42] :

- Natural Sources
  - Biomass (Potassium (K))
  - Sea spray aerosols (Sodium (Na))
  - Coal burning (Aluminium (Al), Selenium (Se), Cobalt (Co), Arsenic (As))
  - Soil and road dust (Aluminium (Al), Silicon (Si), Calcium (Ca))
  - Volcanic dust particles and wild land fire particles (Potassium (K), Zinc (Zn), Lead (Pb))
- Anthropogenic-biogenic sources
  - Diesel, petrol and coal combustion (Elemental carbon (EC), Sulfates (SO4)
  - Heavy industry—high temperature combustion (Iron (Fe), Zinc (Zn), Copper (Cu), Lead (Pb), Nitrates (NO3)
  - Fertilizer and animal husbandry (Ammonium (NH4)
  - Oil burning (Vanadium (V), Nickel (Ni), Manganese (Mn), Iron (Fe), Organic carbon (OC)

3) Anthropic variables: According to existing research on $PM_{2.5}$ forecasting, road and rail density, population density, and proportion of land use (agricultural land and forest land) as human influencing factors of $PM_{2.5}$ [56], [57]. Land use variables have always been the conventional choice in $PM_{2.5}$ driving research, representing the degree of landscape modification by humans and as a proxy for local emissions and background air pollution levels. Land use variables approximate air pollutant emissions, often at the kilometer or sub-kilometer scale. Land use can be (1) type of land use coverage, (2) distance to the nearest highway, (3) distance to the coastline, (4) elevation, and (5) NDVI (normalized vegetation difference index). (6) The distribution of PM2.5 is influenced by elevation due to the difficulty of reaching PM2.5 at a higher elevation above the earth's surface from sea level [58].

Land use variables are potential sources of $PM_{2.5}$ and are the areas of most significant concern. Existing studies on the dependence of land use variables on AOD or $PM_{2.5}$ show significant differences between lower and higher areas. Grassland, shrubs, water bodies, and artificial surfaces positively depend on AOD (maximum partial dependence of about 0.63) and are insignificant on $PM_{2.5}$ [43]. Since land cover properties can be assumed to change gradually, missing values at the temporal scale are then replaced through linear interpolation between adjacent values [59].

Nighttime (Nigh Light (NLT) population density variables, such as road network density, height, and number of buildings, are used to identify the degree of population agglomeration and urbanization in the scale of urban industrial development [30]. For example, coal, forest fires and vehicle emissions can be a major source of haze, as the larger composition of released fine soot particles affects the higher AOD and $PM_{2.5}$ measurements in MODIS. Another study showed that NLT has an increased MSE: about 30 n plots with partial dependence on $PM_{2.5}$ generally increase slowly as NLT increases [60]. This suggests that high NLT represents densely populated areas and still operating factories. However, the impact of emissions in a short period is not very influential on high $PM_{2.5}$. Studies by [61] have shown that population density is significantly positively correlated with AOD, with $PM_{2.5}$ concentrations increasing sharply near population density = 6 (people/KM) then increasing slowly. This pattern shows the contribution of population density to $PM_{2.5}$ concentrations, which can rise above pollutant limits due to high human activity.

C. Analysis of Variables Affecting $PM_{2.5}$

Understanding the variables that trigger $PM_{2.5}$ is essential. The study by Su [56], adopted spatial autocorrelation analysis to explain the spatial correlation of $PM_{2.5}$ in the study area and period. This study uses spatial cluster and outlier methods to analyze the distribution and spatial-temporal variation of the $PM_{2.5}$ surface. Meanwhile, the Random Forest algorithm was used to analyze the influence variables on $PM_{2.5}$. The relationship between $PM_{2.5}$ concentration and the explanatory variables was well modeled, and the explanation level of the drivers to $PM_{2.5}$ was more than 0.9. Temperature, rainfall, and wind speed are the main driving forces of $PM_{2.5}$ emissions. The impact of forest fires is also slowly influencing the driving force of $PM_{2.5}$ concentration [61]. Another study related to the importance of explanatory variables in explaining $PM_{2.5}$ variations, using ensemble models (deep learning (DL), Random Forest Distribution (DRF), and Gradient Boosting Machines (GBM)) by explaining $PM_{2.5}$ variations such as wind speed, inversion strength, and aerosol optical depth (AOD) to be the most influential in DRF and GBM models. For the deep learning algorithm, wind direction emerged as the most influential, followed by the land cover variable [62].

D. Missing Values

The relationship between AOD and $PM_{2.5}$ varies considerably across regions, seasons, and time periods. Hence, studies that employ a single machine learning technique to estimate $PM_{2.5}$ concentrations over a vast area require some enhancements in spatial distribution. Additionally, the accuracy of machine learning methods for $PM_{2.5}$ estimation is linked to the training sample used. Since satellites cannot detect atmospheric aerosols below the clouds, it has a gap of missing values in the spatial distribution.

The advantages of atmospheric model data are fully utilized to obtain comprehensive coverage results. Therefore, the map produced by the interpolation analysis of $PM_{2.5}$ concentration distribution using measured values from each monitoring station can be evidence of the validity of the $PM_{2.5}$ concentration prediction technique. To obtain values for unknown spatial data, a spatial interpolation approach can be used. Various researchers have referred to standard spatial interpolation methods such as Trend Surface (TS) interpolation, Collaborative Kriging (CK), Inverse Distance Weighted (IDW) interpolation, Ordinary Kriging (OK) interpolation [56], and radial basis function.

1) The OK; interpolation method assumes that the spatial correlation of surface changes can be explained based on the distance or direction between sampling points, and it adjusts a mathematical function at all points to determine the value of each outlet by considering a certain number of nearby points or a certain radius. Calculated through Eq. (2) as follows:

$$z_v^*(x) = \sum_{i=1}^{n} \lambda_j Z(x_i) \tag{2}$$

Where *Z(x)* is the measurement of position *i*, λi is the unknown weight of the measurement value at a position *i*, is the predicted position, and n is the number of measurements. Wong [56] used the OK method to generate continuous air pollutant concentrations and meteorological factors covering Taiwan.

2) The IDW: interpolation method calculates pixel values by linearly combining a series of sample points, with the goal of minimizing the distance between the mapped variable and the sample locations. Calculated through Equation (3) as follows:

$$z = \left[\sum_{i=1}^{n} \frac{z_i}{d_i^k}\right] / \left[\sum_{i=1}^{n} \frac{1}{d_i^k}\right] \tag{3}$$

This formula represents the calculation of the inverse distance weighting (IDW) method, which is a spatial interpolation technique used to estimate values at unsampled locations based on the values of neighboring sampled locations. In the formula, " z " represents the estimated value at the unsampled location, "n" is the number of neighboring sampled locations, " z_i " is the value at each neighboring sampled location, "d_i " is the distance from the unsampled location to each neighboring sampled location, and " k " is a power parameter that determines the influence of the distance on the estimated value. The formula calculates the weighted average of the neighboring sampled values based on their distances to the unsampled location, where the weights are determined by the inverse of the distances raised to the power of " k ", and divides the sum of the weighted values by the sum of the weights to obtain the estimated value. Chae [63] used the IDW method to interpolate the missing values uniformly and generate grid-shaped data in the Convolutional Neural Network (ICNN) Interpolation prediction model in South Korea from January 1, 2018, to December 31, 2019, with $PM_{2.5}$ and PM_{10} measurements [63].

3) The TS: method involves applying statistical techniques to create continuous mathematical surfaces by matching them to known spatial points to examine patterns of change in regional and local geological variables. It is calculated through Equation (3) as follows:

$$z = \beta_1 + \beta_2 + \beta_{13y} + \beta_{4x^2} + \beta_{5xy} + \beta_{6y^2} + \cdots \tag{4}$$

Where Z is the address variable, x and y are the coordinates of the observation point.

The CK method refers to kriging interpolation, which is a geostatistical method based on variogram theory and structural analysis. It is considered an unbiased and optimal estimation method for regional variables [64], [65]. Liu [66], employed the CK Method to generate simulation maps depicting the spatial distribution of $PM_{2.5}$ mass concentration on Changsha's Third Ring Road. Furthermore, an additional interpolation analysis map was generated using the measured values from each monitoring station, to serve as a reference for the map generated using predicted values. The aim is to validate the $PM_{2.5}$ concentration prediction method, which uses the CK method. This method uses one or more secondary variables to interpolate the primary variable of interest. The method assumes that the correlation between these variables can improve the accuracy of the primary predictor [66]. Usually, some measurement points correspond to a normal distribution. To estimate each unknown point, the estimator is expressed as a linear combination of the valid sample values. In other words, a linear combination of valid sample values is used as an estimator for each unknown point to be estimated:

$$\hat{Z}(S_i) = \sum_{j-1}^{n} \lambda_j \, Z(S_j) \tag{5}$$

where $\hat{Z}(S_i)$ is the estimated value of the variable at location S_i , $Z(S_j)$ is the observed value of the variable at location S_j, λ_j is the weight assigned to the observed value at location S_j, and n is the total number of observed values used in the estimation.

One way to guarantee that the model provides unbiased estimates is by:

$$\sum_{j-1}^{n} \lambda_j = 1 \tag{6}$$

The value of $\hat{Z}(S_i)$ can be determined while ensuring that the kriging variance is kept at a minimum.

V. Applied Machine Learning Models

Advanced machine learning models have been applied to $PM_{2.5}$ forecasting by developing methods that reflect transport and formation characteristics in suitable algorithms. Compared to classical statistical models and generalized additive models that have been used to calculate empirical models of $PM_{2.5}$[67], machine learning has become a popular method for developing satellite-based AOD-$PM_{2.5}$ models due to its advantages in selecting and using many independent factors that can affect the dependent variable to be estimated [62], [68].

The feed-forward neural network [69] and Recurrent Neural Network (RNN) [41] are some of the fundamental algorithms to simulate the temporal variation of PM2.5 concentration by describing the stratigraphic characteristics of the predicted area. Observation data from monitoring stations in the forecast area and surrounding areas are utilized to develop Convolutional Neural Network (CNN) and Graph Neural Network (GNN) models that directly capture transportation characteristics [41], [70]. These models can effectively represent the spatial correlation between the forecast area and the downwind emission source.

Hybrid models that combine CNN and GNN with the temporal property of LSTM, such as CNN-LSTM and GNN-LSTM, can reflect the temporal variation of the forecast area and the transmission of the wind direction area. Theoretically, the convolutional LSTM (ConvLSTM) network structure makes it an ideal algorithm for combining transportation and formation features; however, these features cannot be accurately predicted after 12 hours [71]. The ensemble technique of Deep Neural Network (DNN) [69], RNN, CNN algorithmic models for real-time estimation of $PM_{2.5}$ is considered capable of reducing the average bias and improving the accuracy index of models that are substantially limited by the uncertainties in the input data of anthropogenic emissions and meteorological fields, as well as the inherent limitations of each model [65].

The paper by Wong [56] uses four types of machine learning algorithms GBM, eXtreme gradient boosting (XGBoost), LightGBM, and CatBoost, after influential variables are identified through interpolation models. The results of the study by comparing the ensemble mixed spatial model and LUR showed that the forecast performance increased from 0.514 to 0.895 (from 0.478 to 0.879) during the day and from 0.523 to 0.878 at night [56].

Note that both the LightGBM model [72] and the eXtrem Gradient Boosting (XGBoost) model [72], [73] are decision tree-based Gradient Boosting frameworks. The XGBoost objective function equation is as follows:

$$Ob^{(t)} = \sum_{j=1}^{T} \left[G_j W_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + yT \tag{7}$$

where:

- Ob^(t) is the objective function at iteration t
- T is the total number of leaf nodes in the tree
- $G_j$and H_j are the cumulative sum of the first-order and second-order partial derivatives of the samples contained in leaf node j, respectively
- λ and γ are constants
- W_j is the score value of the j-th leaf node
- $w_j$is the weight of the j-th leaf node

LightGBM uses the same gain formula $G_j W_j + \frac{1}{2}(H_j + \lambda) w_j^2$as as XGBoost, however, it employs a histogram-based algorithm, as well as techniques like leaf-wise growth with depth restrictions and Gradient-based One-Side Sampling (GOSS) to accelerate the training process. These approaches enable LightGBM to attain better prediction accuracy and lower memory consumption.

The Random Forest (RF) regression algorithm produced a good fit in detecting the relationship between $PM_{2.5}$ and its drivers [61], [74], [75]. Liu [12] utilized RF as a gap filler on the Himawari-8 AOD, using MERRA-2 to estimate hourly $PM_{2.5}$ concentrations, respectively. The results of this random forest study indicate that a set of input variables are used at each node to grow the tree. The algorithm (random forest) used resulted in gap-filling capability with AOD MERRA-2 can provide reliable spatial and temporal $PM_{2.5}$ predictions and significantly reduce errors in $PM_{2.5}$ estimation [12]. he $PM_{2.5}$ concentration estimation model (night) was also conducted by Ma [77] by integrating Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) radiance, moon phase angle, and meteorological data in the Beijing Tianjin-Hebei region. The study developed a NightPMES model using random forests and compared its cross-validation results with those of MLR and DNN models. The NightPMES model achieved an R2 of 0.82, and an RMSE and MAE of 16.67 and 10.20, respectively. In addition, the NightPMES model performed better than most previous models [76].

The general framework for estimating $PM_{2.5}$ concentrations in RF is as follows:

$$f(x) = \sum_{z=1}^{z} C_z I(x \in R_z)$$

$$\hat{c}_z = \text{mean}(y_i | x \in R_z) \tag{8}$$

The formula involves the regression tree function, $f(x)$ where the output value is the estimated $PM_{2.5}$ value. The sample (*xi*, *yi*) is taken from the Z region (R1, R2, ..., Rz), and there are N samples in total. The best estimate of the output mean for the data set is denoted as ĉz. The RF division strategy is expressed as follows:

$$z_1(m,n) = \{X|X_j \le N\} \& Z_2(m,n) = \{\{X|X_j > N\}$$

$$\min_{m,n}\left[\min_{m,n} \sum_{x_i \in R_1(m,n)} (y_i - C_1)^2 + \min_{m,n} \sum_{x_i \in R_2(m,n)} (y_i - C_2)^2\right] \tag{9}$$

$$\hat{c}_z = \text{mean}(y_i | x_i \in R_1(m,n)) \& \hat{c}_2 = \text{mean}(y_i | x_i \in R_2(m,n))$$

where, m is the splitting variable, n is the split point. Diagrammatic representation of it is given in Fig. 3.

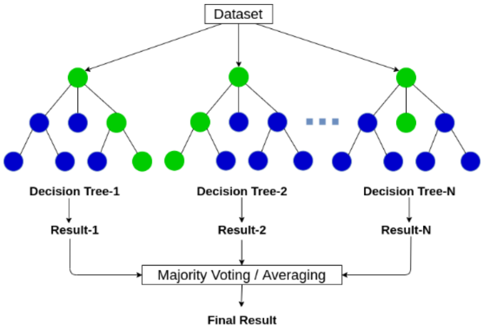

Fig. 3. Random forest based prediction process.

The study by Aguilera [62], used three base learners available within the H2O framework for machine learning: deep learning, random forest (RF), and Gradient Boosting. Each model was trained individually on all response ($PM_{2.5}$) and independent variables, with the optimal parameters of each machine learning algorithm selected by performing a grid search which was then stacked to find the optimal combination of the set of prediction algorithms (H2O's Stacked Ensemble method). Of the three machine learning algorithms, the optimal combination of three base learners (RF, deep learning, and gradient boosting) achieved excellent prediction performance (R2 of 0.78 and RMSE of 3.51 µg m-3) [62].

Feng conducted a study in the Beijing-Tianjin-Hebei region, where they developed an integrated model using RF and LightGBM after wavelet decomposition of PM2.5 observations. The study showed a high degree of consistency between the estimated and actual values. The cross-validation using time-based R2, RMSE, and MAE showed good model performance, with respective values of 0.91, 11.60, and 7.34 [77]. A later study by Falah [73] explored the use of RF and XGBoost models based on the fusion of multiple satellite-borne remote sensing aerosol products retrieved from two platforms (Aqua and Aura), two sensors (MODIS and OMI), and three retrieval algorithms (MAIAC, DB, and OM AE RUV). This study developed thirteen different performance models for each algorithm based on the input data sources MODIS/MAIAC (AOD, aerosol type), MODIS/DB (Angstrom exponent), and OMI (UV Aerosols Index). The UVAI OMI is used to classify aerosols into three categories: scattering aerosols, UVAI < 0.25; mixed-type aerosols, 0.25 ⩽ UVAI < 0.25; and absorbing aerosols, 0.25 ⩽ UVAI. Similarly, MODIS/DB AE is used to classify aerosols into three size fractions: coarse, e.g., dust, AE < 0.7; mixed mode, 0.7 ⩽ AE < 1.3; and delicate, e.g., smoke, 1.3 ⩽ AE. Overall, both RF and XGBoost models showed good performance, with variance (RF; R2 0.753 and NRMSE 0.884 - XGBoost R2 0.741 and NRMSE 0. 874) explained by high cross-validation and low normalized root mean square error even for the base model (MAIAC AOD: AOD, CWV, PBLH, SP), with both models showing much better overall weighting performance when the model input data is subdivided into categories representing different aerosol types/properties [72].

Mahmud [54] conducted a study that used six supervised machine learning algorithms for regression and classification to predict PM2.5 values from 2015 to 2019 in the North Paso region. The variables were analyzed by six different machine learning algorithms using various evaluation metrics. The study showed that the ML model successfully detected the effects of other variables on PM2.5, made accurate predictions, and identified areas of potential risk. The random forest algorithm showed the best performance among all machine learning models with 92% accuracy[54]. This technique has several advantages over other machine learning methods, such as shorter computation time, ease of handling high-dimensional data, strong fault tolerance, and parallel processing, making it suitable even for very high-dimensional data.

Support Vector Machines (SVMs) are flexible and powerful techniques for supervised machine learning, which are used for classification, pattern recognition, and functional regression problems. SVMs find an N-dimensional hyperplane with large margins to classify data into specific groups or labels [78]. A hyperplane divides the class into two, and the margin is used to divide the hyperplane. The predicted value, close to the best margin, is sampled to one of the classes. The predicted output includes one of the high-dimensional spaces as class 1 or 0, which concludes the prediction as traffic or less traffic area.

Recent research indicates that artificial neural networks are effective in both classification and regression tasks. One approach to predict areas with high levels of air pollution is by utilizing support vector machines (SVM), which aim to identify an N-dimensional hyperplane that maximizes the separation gap (margin) for the training data points. The optimal hyperplane is located at the center of the margin, and the data points located close to this hyperplane are known as support vectors. SVMs use kernel functions, such as linear, radial basis function, polynomial, Fisher, and Bayesian, to bridge linearity to non-linearity. In Masood's work [25], kernel functions were found to be crucial in this process. This study employed both linear and polynomial kernel functions. A visual representation of the SVM approach is shown in Fig. 4.

$$Linear\ Kernel = k(X_i, X_j) = x_j^T x_j \tag{10}$$

$$Polynomial\ Kernel = k(X_i, X_j) = (1 + x_j^T x_j)^p \tag{11}$$

Where X_i and X_j are independent random vectors, and p is a polynomial kernel order.

The efficacy of the SVM kernels depends on the calibration of controlling parameters such as kernel width (σ), regularization parameter (C), and gamma parameter (γ). In this research, two kernels (linear and polynomial) were employed for modeling. The SVM_lin model had NSE, RMSE, IA, R2, and R values of 0.938, 22.733, 0.983, 0.938, and 0.968, respectively, for the training phase, and 0.896, 29.634, 0.970, 0.923, and 0.961 for the testing phase. For the SVM_poly model, the NSE, RMSE, IA, R2, and R values for the training phase were 0.934, 23.334, 0.982, 0.935, and 0.967, respectively, and for the testing phase, they were 0.893, 30.071, 0.939, 0.840, and 0.916. Overall, the results of the SVM_lin and SVM_poly models were satisfactory for both the training and testing phases, indicating their ability to accurately predict PM2.5 concentrations. [25].

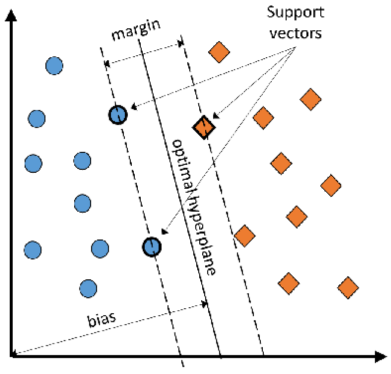

Fig. 4. This diagram illustrates the ideal hyperplane that effectively separates the data points, with the support vectors located near it.

A. Model Validations and Predictions for the Model

Statistical indicators such as coefficient of determination (R2, the higher, the better), correlation coefficient (R), Mean percentage error (MPE, the lower, the better), root means squared prediction error (RMSE, the lower, the better), index of agreement (IA), Mean Absolute Percentage Error, (MAPE the lower, the better), mean absolute error (MAE) and Nash-Sutcliffe efficiency index (NSE), are evaluation metrics typically used for model evaluation. The mathematical expressions of these metrics are given as follows:

1) *Determination coefficient (R^2)*

$$R^2 = \frac{\sum_{i=1}^{N}(I_0 - \bar{I_0})\,(I_0 - \bar{I_0})}{\sum_{i=1}^{N}(I_0 - \bar{I_0})^2\ \sum_{i=1}^{N}(I_p - \bar{I_p})^2} \tag{12}$$

2) *Correlation coefficient (R)*

$$R = \frac{N\ \sum I_o I_p - (\sum I_o)\ (\sum I_p)}{\sqrt{N(\sum I_0^2) - (\sum I_0)^2}\ \sqrt{N(\sum I_p^2) - (\sum I_p)^2}} \tag{13}$$

3) *Root mean square error (RMSE)*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(I_p - I_0)^2}{N}} \tag{14}$$

4) *Index of Agreement (IA)*

$$IA = \left(\frac{\sum_{i=1}^{N}(\,|I_0 - I_p|)^2}{\sum_{i=1}^{N}(\,|I_p - \bar{I_0}| + |I_0 - \bar{I_0}|)^2}\right) \tag{15}$$

5) *Mean Absolute Percentage Error (MAPE)*

$$M = \frac{1}{n}\sum_{t-1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{16}$$

6) *Mean Absolute Error (MAE)*

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{true} - y_{predict}\right| \tag{17}$$

7) *Nash–Sutcliffe efficiency index (NSE)*

$$NSE = 1 - \left(\frac{\sum_{i=1}^{N}(\,I_0 - I_p)^2}{\sum_{i=1}^{N}(\,I_0 - \bar{I_0})^2}\right) \tag{18}$$

VI. Conclusions

Although ground stations are considered precise for measuring $PM_{2.5}$ concentrations, obtaining values that reflect the overall situation is challenging due to their uneven presence. As a result, many researchers are turning to satellite remote sensing to fill in the gaps in spatial and temporal data left by ground stations. AOD, a satellite remote sensing derivative, has been utilized to calculate $PM_{2.5}$ concentrations due to its relationship-based nature. However, satellite products have limitations such as being unable to detect atmospheric aerosols below clouds and the variation of AOD and $PM_{2.5}$ relationship between regions, time, and seasons. To address this, researchers commonly use spatial interpolation methods like OK, IDW, TS, and CK to obtain missing spatial data.

In recent years, machine learning has become popular for predicting unknown values at spatial and temporal scales, and to establish the relationship between $PM_{2.5}$ concentrations and AOD values at each grid. Many new methods, ensembles, and refinements of existing methods have been applied to PM2.5

estimation based on the derived satellite data. Large datasets, increased computing power, and awareness of the potential benefits of predictors working together have contributed to achieving higher estimation accuracy at different scales and spatial-temporal resolutions.

To support our explanations, we reviewed relevant papers, including classic papers, in this study. Key findings include:

- The distribution of ground-level PM2.5 observatories is generally uneven, which makes PM2.5 estimates less reliable in areas with fewer stations compared to those with more stations. This also raises concerns about the effectiveness of the commonly used cross-validation approach based on ground stations, as the observational data used for training and validation are concentrated in areas with more stations.
- Several factors determine the accuracy of PM2.5 estimates, including the specific conditions of the study area, the resolution of the source data, the use of predictors in a particular model, and the details of the methods used to estimate $PM_{2.5}$ concentrations.
- The low consistency between independent and dependent variables in the same atmospheric environment can affect the estimation results. Therefore, it is essential to have different data as predictors to increase confidence in the results obtained. However, data availability is a challenge. In some research areas, the potential for predicting the temporal variation of $PM_{2.5}$ based on satellite AOD needs to be further explored in the future.
- The use of classical methods to estimate $PM_{2.5}$ concentrations is known to be not as accurate as those obtained using new methods. On the other hand, to estimate $PM_{2.5}$ concentrations, there are more and more models and ensemble methods that can be implemented for various conditions.
- A simple and fast method is a spatial interpolation. Several improvements and integrations to various methods have been made to obtain more accurate results. However, the accuracy of these methods is relatively low compared to more complex methods (machine learning).
- Currently, research on $PM_{2.5}$ concentration estimation models is starting to lead to the use of deep learning models. In some recent studies, especially those published in 2023, deep learning models dominate many studies.

From the importance of variables and correlations between variables in different articles, the following conclusions can be drawn:

- Meteorological variables are a class of predictors that make an essential contribution to $PM_{2.5}$ after AOD;
- The contribution of land use variables has a low correlation with meteorological variables;
- The importance of population-related variables depends on the economic development of the study area.

REFERENCES

[1] W. J. Chen et al., “Susceptible windows of exposure to fine particulate matter and fetal growth trajectories in the Spanish INMA (INfancia y Medio Ambiente) birth cohort,” Environ. Res., vol. 216, Jan. 2023, doi: 10.1016/j.envres.2022.114628.

[2] J. Tan-soo and S. K. Pattanayak, “Seeking natural capital projects : Forest fires , haze , and early-life exposure in Indonesia,” in PNAS, 2019, pp. 1–7. doi: 10.1073/pnas.1802876116.

[3] T. Schikowski and H. Altuğ, “The role of air pollution in cognitive impairment and decline,” Neurochem. Int., vol. 136, no. February, p. 104708, 2020, doi: 10.1016/j.neuint.2020.104708.

[4] L. Yang, C. Li, and X. Tang, “The Impact of PM2.5 on the Host Defense of Respiratory System,” Front. Cell Dev. Biol., vol. 8, no. March, pp. 1–9, 2020, doi: 10.3389/fcell.2020.00091.

[5] S. Yin, T. Li, X. Cheng, and J. Wu, “Remote sensing estimation of surface PM2.5 concentrations using a deep learning model improved by data augmentation and a particle size constraint,” Atmos. Environ., vol. 287, Oct. 2022, doi: 10.1016/j.atmosenv.2022.119282.

[6] Z. Ma et al., “A review of statistical methods used for developing large-scale and long-term PM2.5 models from satellite data,” Remote Sens. Environ., vol. 269, p. 112827, 2022, doi: https://doi.org/10.1016/j.rse.2021.112827.

[7] D. K. and N. T. M. T. and P. V. H. Pham Phan Hong Danhand Le, “Estimating PM2.5 Mass Concentration from MODIS AOD Products in Ho Chi Minh City, Vietnam,” in ICSCEA 2021, 2023, pp. 579–588.

[8] P. and J. A. M. and A. J. S. and S. A. and V. G. K. Scaria Haritha P.and Avanthika, “Relational Study of PM2.5 Surface Concentration with MODIS Level 3 AOD Data Over India,” in Recent Advances in Civil Engineering, 2023, pp. 99–113.

[9] S. Gündoğdu, G. Tuna Tuygun, Z. Li, J. Wei, and T. Elbir, “Estimating daily PM2.5 concentrations using an extreme gradient boosting model based on VIIRS aerosol products over southeastern Europe,” Air Qual. Atmos. Heal., vol. 15, no. 12, pp. 2185–2198, 2022, doi: 10.1007/s11869-022-01245-5.

[10] N. Erkin, M. Simayi, X. Ablat, P. Yahefu, and B. Maimaiti, “Predicting spatiotemporal variations of PM2.5 concentrations during spring festival for county-level cities in China using VIIRS-DNB data,” Atmos. Environ., vol. 294, p. 119484, 2023, doi: https://doi.org/10.1016/j.atmosenv.2022.119484.

[11] Z. Song, B. Chen, and J. Huang, “Combining Himawari-8 AOD and deep forest model to obtain city-level distribution of PM2.5 in China,” Environ. Pollut., vol. 297, p. 118826, 2022, doi: https://doi.org/10.1016/j.envpol.2022.118826.

[12] Z. Liu, Q. Xiao, and R. Li, “Full Coverage Hourly PM2.5 Concentrations’ Estimation Using Himawari-8 and MERRA-2 AODs in China,” Int. J. Environ. Res. Public Health, vol. 20, no. 2, Jan. 2023, doi: 10.3390/ijerph20021490.

[13] Z. Song et al., “High temporal and spatial resolution PM2.5 dataset acquisition and pollution assessment based on FY-4A TOAR data and deep forest model in China,” Atmos. Res., vol. 274, p. 106199, 2022, doi: https://doi.org/10.1016/j.atmosres.2022.106199.

[14] B. N. Vu, J. Bi, W. Wang, A. Huff, S. Kondragunta, and Y. Liu, “Application of geostationary satellite and high-resolution meteorology data in estimating hourly PM2.5 levels during the Camp Fire episode in California,” Remote Sens. Environ., vol. 271, p. 112890, 2022, doi: https://doi.org/10.1016/j.rse.2022.112890.

[15] L. Wang, Y. Zhang, K. Wang, B. Zheng, Q. Zhang, and W. Wei, “Application of Weather Research and Forecasting Model with Chemistry (WRF/Chem) over northern China: Sensitivity study, comparative evaluation, and policy implications,” Atmos. Environ., vol. 124, pp. 337–350, 2016, doi: https://doi.org/10.1016/j.atmosenv.2014.12.052.

[16] G. E. Kulkarni, A. A. Muley, N. K. Deshmukh, and P. U. Bhalchandra, “Autoregressive integrated moving average time series model for forecasting air pollution in Nanded city, Maharashtra, India,” Model.

Earth Syst. Environ., vol. 4, no. 4, pp. 1435–1444, 2018, doi: 10.1007/s40808-018-0493-2.

[17] [17] H. Karimian, Y. Li, Y. Chen, and Z. Wang, "Evaluation of different machine learning approaches and aerosol optical depth in PM2.5 prediction," Environ. Res., vol. 216, Jan. 2023, doi: 10.1016/j.envres.2022.114465.

[18] M. Unik and Sri Nadriati, "Overview: Random Forest Algorithm for PM2.5 Estimation Based on Remote Sensing," J. CoSciTech (Computer Sci. Inf. Technol., vol. 3, no. 3, pp. 422–430, Dec. 2022, doi: 10.37859/coscitech.v3i3.4380.

[19] P. Gupta and S. A. Christopher, "Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach," J. Geophys. Res. Atmos., vol. 114, no. 20, pp. 1–14, 2009, doi: 10.1029/2008JD011497.

[20] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," Geosci. Front., vol. 7, no. 1, pp. 3–10, 2016, doi: 10.1016/j.gsf.2015.07.003.

[21] K. J. Bergen, P. A. Johnson, M. V. De Hoop, and G. C. Beroza, "Machine learning for data-driven discovery in solid Earth geoscience," Science (80-.)., vol. 363, no. 6433, 2019, doi: 10.1126/science.aau0323.

[22] Q. Di et al., "An ensemble-based model of PM(2.5) concentration across the contiguous United States with high spatiotemporal resolution.," Environ. Int., vol. 130, p. 104909, Sep. 2019, doi: 10.1016/j.envint.2019.104909.

[23] Q. Di et al., "Assessing NO(2) Concentration and Model Uncertainty with High Spatiotemporal Resolution across the Contiguous United States Using Ensemble Model Averaging.," Environ. Sci. Technol., vol. 54, no. 3, pp. 1372–1384, Feb. 2020, doi: 10.1021/acs.est.9b03358.

[24] P. Zhang, L. Yang, W. Ma, N. Wang, F. Wen, and Q. Liu, "Spatiotemporal estimation of the PM2.5 concentration and human health risks combining the three-dimensional landscape pattern index and machine learning methods to optimize land use regression modeling in Shaanxi, China," Environ. Res., vol. 208, p. 112759, 2022, doi: https://doi.org/10.1016/j.envres.2022.112759.

[25] A. Masood and K. Ahmad, "Data-driven predictive modeling of PM2.5 concentrations using machine learning and deep learning techniques: a case study of Delhi, India," Environ. Monit. Assess., vol. 195, no. 1, Jan. 2023, doi: 10.1007/s10661-022-10603-w.

[26] H. Feizi, M. T. Sattari, R. Prasad, and H. Apaydin, "Comparative analysis of deep and machine learning approaches for daily carbon monoxide pollutant concentration estimation," Int. J. Environ. Sci. Technol., vol. 20, no. 2, pp. 1753–1768, 2023, doi: 10.1007/s13762-022-04702-x.

[27] X. Li, L. Li, L. Chen, T. Zhang, J. Xiao, and L. Chen, "Random Forest Estimation and Trend Analysis of PM2.5 Concentration over the Huaihai Economic Zone, China (2000–2020)," Sustain., vol. 14, no. 14, Jul. 2022, doi: 10.3390/su14148520.

[28] W. Yu, S. Li, T. Ye, R. Xu, J. Song, and Y. Guo, "Deep Ensemble Machine Learning Framework for the Estimation of PM2:5 Concentrations," Environ. Health Perspect., vol. 130, no. 3, Mar. 2022, doi: 10.1289/EHP9752.

[29] Z. Li, Z. Di, M. Chang, J. Zheng, T. Tanaka, and K. Kuroi, "Study on the influencing factors on indoor PM2.5 of office buildings in beijing based on statistical and machine learning methods," J. Build. Eng., vol. 66, p. 105240, 2023, doi: https://doi.org/10.1016/j.jobe.2022.105240.

[30] M. D. Yazdi et al., "Predicting fine particulate matter (PM2.5) in the greater london area: An ensemble approach using machine learning methods," Remote Sens., vol. 12, no. 6, 2020, doi: 10.3390/rs12060914.

[31] M. Liu, X. Liu, L. Wu, X. Zou, T. Jiang, and B. Zhao, "A modified spatiotemporal fusion algorithm using phenological information for predicting reflectance of paddy rice in southern china," Remote Sens., vol. 10, no. 5, 2018, doi: 10.3390/rs10050772.

[32] F. Yang et al., "Preliminary investigation of a new AHI aerosol optical depth (AOD) retrieval algorithm and evaluation with multiple source AOD measurements in China," Remote Sens., vol. 10, no. 5, 2018, doi: 10.3390/rs10050748.

[33] J. Wei et al., "Himawari-8-derived diurnal variations in ground-level PM2.5 pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM)," Atmos. Chem. Phys., vol. 21, no. 10, pp. 7863–7880, 2021, doi: 10.5194/acp-21-7863-2021.

[34] J. Sun, J. Gong, and J. Zhou, "Estimating hourly PM2.5 concentrations in Beijing with satellite aerosol optical depth and a random forest approach," Sci. Total Environ., vol. 762, p. 144502, 2021, doi: https://doi.org/10.1016/j.scitotenv.2020.144502.

[35] L. Zang, F. Mao, J. Guo, W. Gong, W. Wang, and Z. Pan, "Estimating hourly PM1 concentrations from Himawari-8 aerosol optical depth in China," Environ. Pollut., vol. 241, pp. 654–663, 2018, doi: https://doi.org/10.1016/j.envpol.2018.05.100.

[36] J. P. Veefkind et al., "TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications," Remote Sens. Environ., vol. 120, pp. 70–83, 2012, doi: https://doi.org/10.1016/j.rse.2011.09.027.

[37] D. Griffin et al., "High-Resolution Mapping of Nitrogen Dioxide With TROPOMI: First Results and Validation Over the Canadian Oil Sands," Geophys. Res. Lett., vol. 46, no. 2, pp. 1049–1060, Jan. 2019, doi: https://doi.org/10.1029/2018GL081095.

[38] K. Garane et al., "TROPOMI/S5P total ozone column data: global ground-based validation and consistency with other satellite missions," Atmos. Meas. Tech., vol. 12, no. 10, pp. 5263–5287, 2019, doi: 10.5194/amt-12-5263-2019.

[39] Y. Pang et al., "In-vitro human lung cell injuries induced by urban PM2.5 during a severe air pollution episode: Variations associated with particle components," Ecotoxicol. Environ. Saf., vol. 206, Dec. 2020, doi: 10.1016/j.ecoenv.2020.111406.

[40] L. Liu et al., "Chemical composition, oxidative potential and identifying the sources of outdoor PM2.5 after the improvement of air quality in Beijing," Environ. Geochem. Health, 2022, doi: 10.1007/s10653-022-01275-z.

[41] Q. Zhang, Y. Han, V. O. K. Li, and J. C. K. Lam, "Deep-AIR: A Hybrid CNN-LSTM Framework for Fine-Grained Air Pollution Estimation and Forecast in Metropolitan Cities," IEEE Access, vol. 10, pp. 55818–55841, 2022, doi: 10.1109/ACCESS.2022.3174853.

[42] P. Thangavel, D. Park, and Y. C. Lee, "Recent Insights into Particulate Matter (PM2.5)-Mediated Toxicity in Humans: An Overview," International Journal of Environmental Research and Public Health, vol. 19, no. 12. MDPI, Jun. 01, 2022. doi: 10.3390/ijerph19127511.

[43] R. Zhang et al., "A nonparametric approach to filling gaps in satellite-retrieved aerosol optical depth for estimating ambient PM2.5 levels," Environ. Pollut., vol. 243, pp. 998–1007, 2018, doi: https://doi.org/10.1016/j.envpol.2018.09.052.

[44] R. B. A. Koelemeijer, C. D. Homan, and J. Matthijsen, "Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe," Atmos. Environ., vol. 40, no. 27, pp. 5304–5315, 2006, doi: https://doi.org/10.1016/j.atmosenv.2006.04.044.

[45] A. Mhawish et al., "Estimation of High-Resolution PM2.5over the Indo-Gangetic Plain by Fusion of Satellite Data, Meteorology, and Land Use Variables," Environ. Sci. Technol., vol. 54, no. 13, pp. 7891–7900, 2020, doi: 10.1021/acs.est.0c01769.

[46] J. Zhong et al., "Robust prediction of hourly PM2.5 from meteorological data using LightGBM," Natl. Sci. Rev., vol. 8, no. 10, 2021, doi: 10.1093/nsr/nwaa307.

[47] D. M. Giles et al., "Advancements in the Aerosol Robotic Network (AERONET) Version 3 database - Automated near-real-time quality control algorithm with improved cloud screening for Sun photometer aerosol optical depth (AOD) measurements," Atmos. Meas. Tech., vol. 12, no. 1, pp. 169–209, 2019, doi: 10.5194/amt-12-169-2019.

[48] J. Wei, Z. Li, Y. Peng, and L. Sun, "MODIS Collection 6.1 aerosol optical depth products over land and ocean: validation and comparison," Atmos. Environ., vol. 201, pp. 428–440, 2019, doi: https://doi.org/10.1016/j.atmosenv.2018.12.004.

[49] M. Tao et al., "Performance of MODIS high-resolution MAIAC aerosol algorithm in China: Characterization and limitation," Atmos. Environ.,

vol. 213, pp. 159–169, 2019, doi: https://doi.org/10.1016/j.atmosenv.2019.06.004.

[50] H. Bagheri, “A machine learning-based framework for high resolution mapping of PM2.5 in Tehran, Iran, using MAIAC AOD data,” Adv. Sp. Res., vol. 69, no. 9, pp. 3333–3349, 2022, doi: https://doi.org/10.1016/j.asr.2022.02.032.

[51] A. Lyapustin, Y. Wang, S. Korkin, and D. Huang, “MODIS Collection 6 MAIAC algorithm,” Atmos. Meas. Tech., vol. 11, no. 10, pp. 5741–5765, 2018, doi: 10.5194/amt-11-5741-2018.

[52] J. Wei, Y. Peng, R. Mahmood, L. Sun, and J. Guo, “Intercomparison in spatial distributions and temporal trends derived from multi-source satellite aerosol products,” Atmos. Chem. Phys., vol. 19, no. 10, pp. 7183–7207, 2019, doi: 10.5194/acp-19-7183-2019.

[53] Z. Wang et al., “The seasonal variation, characteristics and secondary generation of PM2.5 in Xi’an, China, especially during pollution events,” Environ. Res., vol. 212, p. 113388, 2022, doi: https://doi.org/10.1016/j.envres.2022.113388.

[54] S. Mahmud, T. B. I. Ridi, M. S. Miah, F. Sarower, and S. Elahee, “Implementing Machine Learning Algorithms to Predict Particulate Matter (PM2.5): A Case Study in the Paso del Norte Region,” Atmosphere (Basel)., vol. 13, no. 12, Dec. 2022, doi: 10.3390/atmos13122100.

[55] S. Lu et al., “Impact of thermal structure of planetary boundary layer on aerosol pollution over urban regions in Northeast China,” Atmos. Pollut. Res., vol. 14, no. 2, p. 101665, 2023, doi: https://doi.org/10.1016/j.apr.2023.101665.

[56] P. Y. Wong, H. J. Su, S. C. C. Lung, and C. Da Wu, “An ensemble mixed spatial model in estimating long-term and diurnal variations of PM2.5 in Taiwan,” Sci. Total Environ., vol. 866, no. 1, p. 161336, 2023, doi: 10.1016/j.scitotenv.2022.161336.

[57] N. Liu, B. Zou, S. Li, H. Zhang, and K. Qin, “Prediction of PM2.5 concentrations at unsampled points using multiscale geographically and temporally weighted regression,” Environ. Pollut., vol. 284, p. 117116, 2021, doi: https://doi.org/10.1016/j.envpol.2021.117116.

[58] L. Yang, H. Xu, and Z. Jin, “Estimating ground-level PM2.5 over a coastal region of China using satellite AOD and a combined model,” J. Clean. Prod., vol. 227, pp. 472–482, 2019, doi: 10.1016/j.jclepro.2019.04.231.

[59] Z. Chen et al., “Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism,” Environment International, vol. 139. Elsevier Ltd, Jun. 01, 2020. doi: 10.1016/j.envint.2020.105558.

[60] Y. Liu, G. Cao, N. Zhao, K. Mulligan, and X. Ye, “Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach,” Environ. Pollut., vol. 235, pp. 272–282, 2018, doi: 10.1016/j.envpol.2017.12.070.

[61] Z. Su, L. Lin, Y. Chen, and H. Hu, “Understanding the distribution and drivers of PM2.5 concentrations in the Yangtze River Delta from 2015 to 2020 using Random Forest Regression,” Environ. Monit. Assess., vol. 194, no. 4, Apr. 2022, doi: 10.1007/s10661-022-09934-5.

[62] R. Aguilera et al., “A novel ensemble-based statistical approach to estimate daily wildfire-specific PM2.5 in California (2006–2020),” Environ. Int., vol. 171, Jan. 2023, doi: 10.1016/j.envint.2022.107719.

[63] S. Chae, J. Shin, S. Kwon, S. Lee, S. Kang, and D. Lee, “PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network,” Sci. Rep., vol. 11, no. 1, pp. 1–9, 2021, doi: 10.1038/s41598-021-91253-9.

[64] K. Zhao, X. Ma, H. Zhang, and Z. Dong, “Performance zoning method of asphalt pavement in cold regions based on climate Indexes: A case study of Inner Mongolia, China,” Constr. Build. Mater., vol. 361, p. 129650, 2022, doi: https://doi.org/10.1016/j.conbuildmat.2022.129650.

[65] Y. S. Koo et al., “A Development of PM2.5 Forecasting System in South Korea Using Chemical Transport Modeling and Machine Learning,” Asia-Pacific J. Atmos. Sci., 2023, doi: 10.1007/s13143-023-00314-8.

[66] J. Liu, B. Zheng, and J. Fan, “Long Short-Term Memory Network and Ordinary Kriging Method for Prediction of PM2.5 Concentration BT - Proceedings of the 2022 International Conference on Green Building, Civil Engineering and Smart City,” 2023, pp. 1158–1169.

[67] X. Yang et al., “Spatiotemporal estimates of daily PM2.5 concentrations based on 1-km resolution MAIAC AOD in the Beijing–Tianjin–Hebei, China,” Environ. Challenges, vol. 8, no. March, p. 100548, 2022, doi: 10.1016/j.envc.2022.100548.

[68] X. Xu, C. Zhang, and Y. Liang, “Review of satellite-driven statistical models PM2.5 concentration estimation with comprehensive information,” Atmos. Environ., vol. 256, no. February, p. 118302, 2021, doi: 10.1016/j.atmosenv.2021.118302.

[69] J. B. Lee et al., “Development of a deep neural network for predicting 6 h average PM2.5 concentrations up to 2 subsequent days using various training data,” Geosci. Model Dev., vol. 15, no. 9, pp. 3797–3813, 2022, doi: 10.5194/gmd-15-3797-2022.

[70] A. Gilik, A. S. Ogrenci, and A. Ozmen, “Air quality prediction using CNN+LSTM-based hybrid deep learning architecture,” Environ. Sci. Pollut. Res., vol. 29, no. 8, pp. 11920–11938, 2022, doi: 10.1007/s11356-021-16227-w.

[71] C. Wen et al., “A novel spatiotemporal convolutional long short-term neural network for air pollution prediction,” Sci. Total Environ., vol. 654, pp. 1091–1099, 2019, doi: https://doi.org/10.1016/j.scitotenv.2018.11.086.

[72] S. Falah, F. Kizel, T. Banerjee, and D. M. Broday, “Accounting for the aerosol type and additional satellite-borne aerosol products improves the prediction of PM2.5 concentrations,” Environ. Pollut., vol. 320, no. January, p. 121119, 2023, doi: 10.1016/j.envpol.2023.121119.

[73] J. Wang et al., “A full-coverage estimation of PM2.5 concentrations using a hybrid XGBoost-WD model and WRF-simulated meteorological fields in the Yangtze River Delta Urban Agglomeration, China,” Environ. Res., vol. 203, p. 111799, 2022, doi: https://doi.org/10.1016/j.envres.2021.111799.

[74] W. Zhou, X. Wu, S. Ding, X. Ji, and W. Pan, “Predictions and mitigation strategies of PM(2.5) concentration in the Yangtze River Delta of China based on a novel nonlinear seasonal grey model.,” Environ. Pollut., vol. 276, p. 116614, May 2021, doi: 10.1016/j.envpol.2021.116614.

[75] X. Y. Jin et al., “Machine learning driven by environmental covariates to estimate high-resolution PM2.5 in data-poor regions,” PeerJ, vol. 10, pp. 1–21, 2022, doi: 10.7717/peerj.13203.

[76] Y. Ma, W. Zhang, L. Zhang, X. Gu, and T. Yu, “Estimation of Ground-Level PM2.5 Concentration at Night in Beijing-Tianjin-Hebei Region with NPP/VIIRS Day/Night Band,” Remote Sensing, vol. 15, no. 3. 2023. doi: 10.3390/rs15030825.

[77] Y. Feng, S. Fan, K. Xia, and L. Wang, “Estimation of Regional Ground-Level PM2.5 Concentrations Directly from Satellite Top-of-Atmosphere Reflectance Using A Hybrid Learning Model,” Remote Sens., vol. 14, no. 11, 2022, doi: 10.3390/rs14112714.

[78] M. M. Hameed, M. K. AlOmar, A. A. A. Al-Saadi, and M. A. AlSaadi, “Inflow forecasting using regularized extreme learning machine: Haditha reservoir chosen as case study,” Stoch. Environ. Res. Risk Assess., vol. 36, no. 12, pp. 4201–4221, 2022, doi: 10.1007/s00477-022-02254-7.