An Estimation of Daily PM2.5 Concentration in Thailand Using Satellite Data at 1-kilometer Resolution

Suhaimee Buya *, Sasiporn Usanavasin *, Gokon Hideomi, Jessada Karnjana

Posted Date: 26 May 2023

doi: 10.20944/preprints202305.1833.v1

Keywords: PM2.5 estimation; Satellite data; Aerosol optical depth; Machine learning; Random Forest; Thailand

Article

An Estimation of Daily PM2.5 Concentration in Thailand Using Satellite Data at 1-Kilometer Resolution

Suhaimee Buya [1,2,*], Sasiporn Usanavasin [1,*], Gokon Hideomi [2] and Jessada Karnjana [3]

[1] School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand; suhaimee.buy@dome.tu.ac.th; sasiporn.us@siit.tu.ac.th

[2] School of Knowledge Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan; gokon@jaist.ac.jp

[3] NECTEC, National Science and Technology Development Agency, Pathum Thani, Thailand; jessada.karnjana@nectec.or.th

* Correspondence: suhaimee.buy@dome.tu.ac.th; sasiporn.us@siit.tu.ac.th

Abstract: This study addresses the limited coverage of regulatory monitoring for particulate matter 2.5 microns or less in diameter (PM2.5) in Thailand due to the lack of ground station data by developing a model to estimate daily PM2.5 concentrations in small regions of Thailand using satellite data at a 1-kilometer resolution. The study employs multiple linear regression and three machine learning models and finds that the random forest model performs the best for PM2.5 estimation over the period of 2011-2020. The model incorporates several factors such as Aerosol Optical Depth (AOD), Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), Elevation (EV), Week of the year (WOY), and year and applies them to the entire region of Thailand without relying on monitoring station data. Model performance is evaluated using the coefficient of determination (R^2) and root mean square error (RMSE), and the results indicate high accuracy for training (R^2: 0.95, RMSE: 5.58 μg/m^3), validation (R^2: 0.78, RMSE: 11.18 μg/m^3), and testing (R^2: 0.71, RMSE: 8.79 μg/m^3) data. These PM2.5 data can be used to analyze the short- and long-term effects of PM2.5 on population health and inform government policy decisions and effective mitigation strategies.

Keywords: PM2.5 estimation; satellite data; aerosol optical depth; machine learning; random forest; Thailand

1. Introduction

According to the World Health Organization (WHO), ambient air pollution causes approximately 7 million premature deaths globally, with particulate matter, ozone, nitrogen dioxide, sulfur dioxide, and other contaminants being some of the leading pollutants [1]. The most dangerous among them is PM2.5, which is particulate matter with an aerodynamic diameter of less than 2.5 μm. These particles can easily enter the lungs and become trapped in the lung's parenchyma, leading to inflammation and oxidative stress [2]. This can cause severe cardiovascular and respiratory diseases and even lung cancer. PM2.5 plays a critical role in air pollution, and environmental health and its impact on human health are of great concern [3].

PM2.5 has been associated with increased mortality and morbidity in several studies [4–6]. However, the coverage of ground-level PM2.5 monitoring sites is limited, which makes it challenging to capture the spatial variability of PM2.5 for exposure and epidemiological research. Researchers have increasingly used satellite-derived atmospheric aerosol optical depth (AOD) to address this challenge as a proxy for ground-level PM2.5 [7–11]. AOD measures the aerosol in the atmosphere and can serve as a proxy for surface PM2.5 [12]. Additionally, other factor variables, including meteorological factors, land use and cover, and time variables, are often included to improve the accuracy of the modeling. These variables can explain seasonal variations and long-term trends in PM2.5 levels and indicate potential PM2.5 sources and areas of concern [14,15]. However, the

 doi:10.20944/preprints202305.1833.v1

importance of these factors varies among studies, and some analyses have found that satellite-derived AODs do not improve model performance [15]. Therefore, the association between satellite data and PM2.5 in different locations must be considered.

Previous studies on the estimation of PM2.5 using satellite data have employed a variety of models, but most have chosen only one [16]. The five studies [14–18] were done to compare model performance comprehensively with the Random Forest (RF) model showing a high coefficient of determination (R^2) in three studies, and the eXtreme Gradient Boosting (XGBoost) model showing a high R^2 in two studies. However, it should be noted that the RF model performed similarly to the XGBoost model. Among the other Machine Learning (ML) models, Multiple Linear Regression (MLR) had the lowest accuracy. Despite this, MLR is still widely used for its simplicity and practicality. Estimating PM2.5 concentrations is challenging due to the numerous variables that can affect it. ML has become popular for solving complex problems because it can find and use multiple independent factors that impact the predicted variable [19].

Earlier research on estimating PM2.5 levels in Thailand using satellite data has been limited due to a scarcity of data from both ground stations and satellites. Two previous studies conducted in Thailand's Chiangmai and central regions estimated PM2.5 using MLR models with AOD (10 kilometers (km)), resulting in R^2 values of 0.77 and 0.49 when considering monitoring station meteorological parameters and 0.22 and 0.11 when not considering them [21,22]. However, these meteorological parameters do not cover small areas such as 1 km, 3 km, and 10 km, limiting the accuracy of PM2.5 estimation. A review article on predicting ground PM2.5 concentration using satellite AOD found that MLR had the lowest R^2 accuracy compared to other models [16]. The low R^2 values suggest further examination into including covariates such as meteorological factors, land use, cover, and season variables in MLR models [23].

In this study, we aim to develop a method for estimating PM2.5 concentrations throughout Thailand using satellite data with a 1 km pixel resolution. Our approach seeks to overcome the limitation of ground-level PM2.5 monitoring by not relying on monitoring station factor variables. Instead, we begin with AOD as a base factor and then add other variables to improve accuracy in estimating PM2.5 levels in Thailand. Specifically, we have selected Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), and Elevation (EV) data to represent land use and cover, as well as year and week of the year (WOY) as time factors. All factor variables are applied at a 1 km pixel resolution throughout Thailand without the need for monitoring station data, which can be costly and not cover all areas of the country. We will use MLR as the standard regression model and other ML models such as RF, XGBoost, and Support Vector Machines (SVM) to compare their performance. The final model with the highest accuracy will be selected to estimate PM2.5 levels in Thailand.

Our study will serve as a reference for future satellite-based PM2.5 estimation studies and will aid in exposure assessment in health studies of the Thai population. Using satellite data to estimate PM2.5 concentrations at a high spatial resolution, our study can provide a more comprehensive understanding of the distribution of PM2.5 in Thailand, which can help inform policy and public health efforts to reduce exposure to harmful air pollutants.

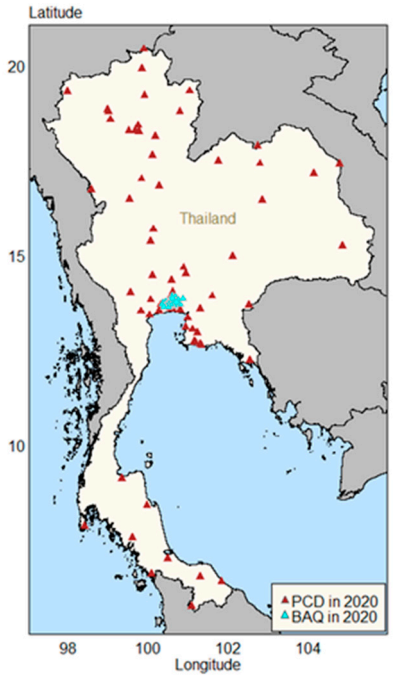
2. Materials and Methods

2.1. PM2.5 data and area of study

Thailand is a Southeast Asian country that borders the Andaman Sea and the Gulf of Thailand, with an approximate population of 70 million people and an area of 513,120 square kilometers. The Pollution Control Department (PCD) is a legally recognized government agency in Thailand that collects data on air pollution parameters from meteorological stations throughout the country. Bangkok's Air Quality and Noise Management Division (BAQ) also operates ground stations for monitoring PM2.5 in Bangkok. The PCD and BAQ measure PM2.5 data using the same standard, the beta-ray attenuation method, which follows the United States Environmental Protection Agency

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1833.v1

(USEPA) reference method. Figure 1 presents PM2.5 data and the number of stations from PCD and stations for BAQ from 2011 to 2020.



Year	Number of stations	
	PCD	BAQ
2011	3	-
2012	5	-
2013	7	-
2014	10	-
2015	12	4
2016	19	4
2017	26	4
2018	53	4
2019	65	14
2020	68	49

Figure 1. The map of PM2.5 stations and the number of stations.

2.2. Satellite data

This study employed remote sensing data obtained from the MODIS satellite products, specifically AOD, LST, NDVI, and EV, which were all retrieved from the National Aeronautics and Space Administration (NASA) Earth Observing System Data and Information System (EOSDIS) offered by the Distributed Active Archive Center (DAAC). AOD data were processed from the MCD19A2 product of both Terra and Aqua satellites, which included "Aerosol Optical Depth at 045 Microns" [24]. The daily AOD data had a spatial resolution of 1 km per pixel and was collected at 10:30 a.m. and 1:30 p.m. local standard time. LST data was collected from Terra's MOD11A1 product [25] and Aqua's MYD11A1 product [26], and their measurements were combined with increasing the sample size. Daily average LST values were calculated by taking the arithmetic mean of the two satellite measurements or using only one satellite's data. The study utilized NDVI data from MOD13A1, with a temporal resolution of 16 days and a spatial resolution of 500 meters, which was beneficial in monitoring vegetation conditions, depicting land cover changes, and providing insights for modeling global biogeochemical and hydrologic processes and regional climates. Additionally, EV data from "Land Digital Elevation Model (MODDEM1KM)—Land/sea mask and digital elevation model" with a spatial resolution of 1 km was used.

2.3. Data analysis

For this study, we found that satellite data and PM2.5 readings were consistent when the sky was clear. To match the daily PM2.5 concentrations for each station from 2011 to 2020, we selected the average satellite data within a 5 km radius. We established a link between PM2.5 outcomes and factors such as AOD, LST, NDVI, EV, WOY, and year by using daily average PM2.5 data. Four models were developed to predict daily PM2.5: MLR, RF, XGBoost, and SVM. We evaluated the model's accuracy using R^2 and root mean square errors (RMSE). A higher R^2 and lower RMSE indicate better-estimating performance. The data handling and analysis were conducted using the R programming language.

 doi:10.20944/preprints202305.1833.v1

2.3.1. Multiple Linear Regression (MLR)

The MLR statistical model is a commonly used method for identifying the relationship between a continuous response variable and one or more predictor variables, which can be continuous or categorical. MLR is a parametric model that assumes a normal distribution, constant variance, and a linear relationship between the response and predictor variables. This study uses a log-linear regression model, and the MLR model can be represented as:

$$\log(PM2.5) = \beta_0 + \beta_1 AOD + \beta_2 LST + \beta_3 NDVI + \beta_4 EV + \beta_5 WOY + \beta_6 Year \quad (1)$$

where β_0 is the intercept, $\beta_{(1-6)}$ is the coefficient of determinant.

2.3.2. Random Forest (RF)

RF is a method for creating an ensemble of decision trees. The RF algorithm builds each tree using a bootstrap sample of the data, and each tree node is split based on the best of a subset of randomly selected predictors [27]. The predictions of each tree are then combined to produce an ensemble prediction of the target variable. The model also calculates the "importance" of each predictor by measuring how much prediction error increases when the data for that variable is permuted. In contrast, the data for the other variables remain unchanged [28]. This study uses the R package "randomForest" [29].

2.3.3. eXtreme Gradient Boosting (XGBoost)

XGBoost is a gradient-boosting technique that improves performance and speed using a tree-based ensemble ML algorithm [30]. Gradient boosting is a method where the loss function is minimized by sequentially adding weak learners through gradient descent optimization. The gradient boosting approach has three key components: a loss function, a weak learner, and an additive model. The loss function measures how well the model predicts the data. Even though a weak learner may not classify things accurately, it is still better than guessing randomly. The additive model is a method of adding decision trees one at a time and iteratively. This study uses the R package "xgboost" [31].

2.3.4. Support Vector Machines (SVM)

SVM is a supervised learning model for regression concerns in ML [32]. SVM builds a set of hyperplanes in a high-dimensional space using a nonlinear transformation based on the following function [33]

$$f(x)=wx+b \quad (2)$$

where x is the input predictors' vector (6 variables), w is the weight vector of x, and b is the error, which defines the hyperplane's distance from the original. SVM is based on decreasing the gap between the expected and actual output values. It reduces prediction errors. This study uses the R package "e1071" [34].

2.3.5. Model assessment

The rows of the PCD dataset were randomly shuffled and divided into a training dataset (80%) and a validation dataset (20%) to ensure that model performance comparisons could be made. A consistent random state was used for this purpose. Table 1 presents the structure of the PCD and BAQ data. The distribution of the training and validation datasets were similar; however, the testing dataset was different as it only included BAQ data collected in Bangkok provinces.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1833.v1

Table 1. The data structure of datasets.

Variables	Types	PCD (n=34,748)		BAQ (n=7,339)
		Training (n =27,798)	Validation (n=6,950)	Testing
Stations	Nominal	68 stations	68 stations	49 stations
Date	Date	2,778 days	1,865 days	734 days
Month	Nominal	12 months	12 months	12 months
Year	Discrete	10 years	10 years	6 years
WOY	Nominal	53 weeks	53 weeks	53 weeks
PM2.5 (μg/m³)	Continuous	μ: 32.2, *s*: 23.7, *IQR*: 26	μ: 32.4, *s*: 23.8, *IQR*: 26	μ: 30.1, *s*: 16.2, *IQR*: 21
AOD	Continuous	μ: 0.5, *s*: 0.3, *IQR*: 0.4	μ: 0.5, *s*: 0.3, *IQR*: 0.4	μ: 0.5, *s*: 0.3, *IQR*: 0.4
LST (°C)	Continuous	μ: 33.3, *s*: 4.5, *IQR*: 6	μ: 33.4, *s*: 4.5, *IQR*: 6	μ: 36.1, *s*: 3.8, *IQR*: 4.3
NDVI	Continuous	μ: 0.1, *s*: 0.2, *IQR*: 0.3	μ: 0.1, *s*: 0.2, *IQR*: 0.3	μ: -0.1, *s*: 0.1, *IQR*: 0.2
EV (m)	Continuous	μ: 144.6, *s*: 198.9, *IQR*: 265.3	μ: 142.4, *s*: 197.3, *IQR*: 265.3	μ: 6.8, *s*: 1.6, *IQR*: 2.9

n: Rows; μ: Mean; s: Standard deviation; IQR: Interquartile range; m: Meter.

After training the model, the model's performance was evaluated by indicators such as R^2 and RMSE, shown in the following formulas:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}} \quad (4)$$

where y_i is the observations of PM2.5, $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the mean of the observations of PM2.5, and n is the total sample count.

3. Results

3.1. Data descriptive statistics

Figure 2 presents a scatterplot matrix of the variables, with the first row and column displaying positive skew histograms of the PM2.5 distribution. Each scatterplot matrix includes the correlation coefficient (R) values, with the top row showing the relationship between each predictor variable and PM2.5. The first column displays the R values for all determinants with PM2.5. Positive R correlations between PM2.5 and AOD, LST, and EV indicate that these variables increase along with PM2.5 (R = 0.51, 0.20, and 0.13, respectively), while negative R correlations between WOY (R = -0.27), NDVI (R = -0.19), and year (R = -0.05) and PM2.5 suggest that as these variables increase, PM2.5 will decrease. AOD has the highest positive association, and lower PM2.5 levels are observed during WOY 20-40 in Thailand's rainy season, indicating a negative correlation. Dry seasons with increased LST show higher PM2.5 levels, while higher NDVI levels decrease PM2.5. Finally, EV and Year have lower correlation values with PM2.5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1833.v1

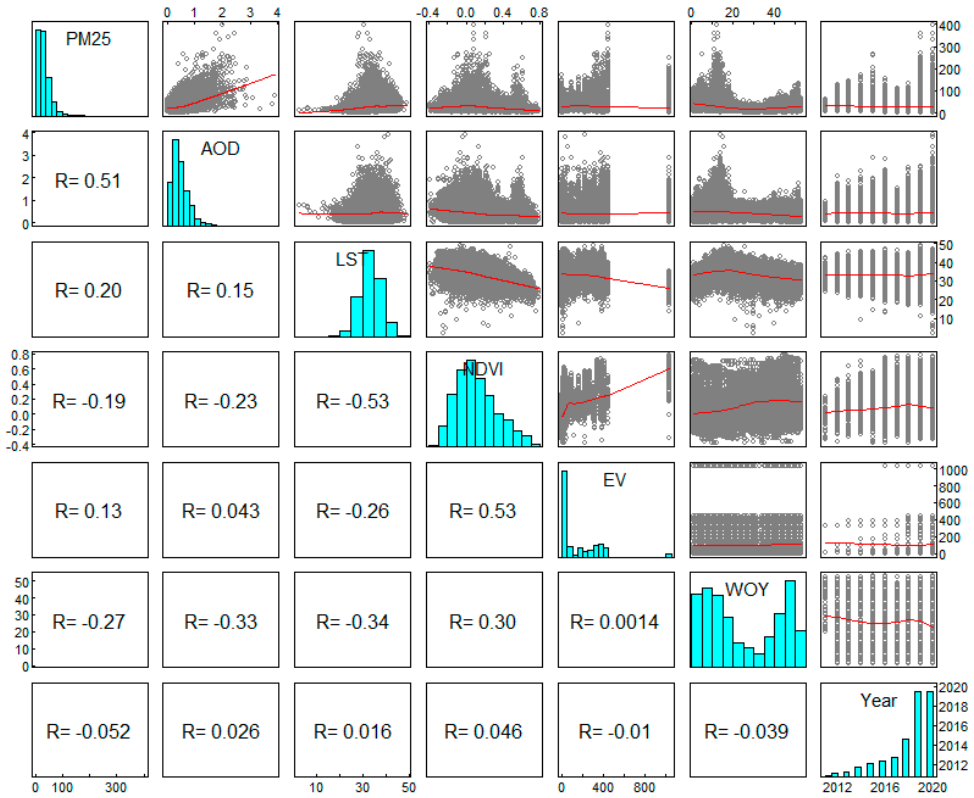


Figure 2. The scatterplot matrix of variables.

3.2. Modeling results

Table 2 presents the estimated performance of each model for the three datasets. The results indicate that the RF model, which includes AOD, LST, NDVI, EV, WOY, and year, is the most effective in predicting PM2.5 across all datasets. The R^2 values for the training, validation, and testing datasets were 0.95, 0.78, and 0.71, respectively, with RMSE values of 5.58 μg/m^3, 11.18 μg/m^3, and 8.79 μg/m^3, respectively (as seen in the scatter plot in Figure S5). In terms of model performance, XGBoost and SVM were similar. However, the LR model had the worst performance.

Although the final RF model has a higher R^2 accuracy in the validation dataset than the testing dataset, the testing dataset has a lower RMSE than the validation dataset. This means the RF model can estimate PM2.5 in the validation dataset more accurately than in the testing dataset. However, the difference between the actual and estimated PM2.5 in the testing dataset is closer than in the validation dataset due to the lower RMSE. This discrepancy could be attributed to the fact that the testing dataset only covers Bangkok provinces and thus has more data from these areas. In contrast, the validation dataset covers all areas of Thailand.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1833.v1

Table 2. The performance of models for estimation of PM2.5.

Models	R^2 (RMSE (μg/m³))		
	Training	Validation	Testing
LR			
AOD	0.18 (21.48)	0.19 (21.26)	0.04 (16.79)
AOD+LST	0.21 (21.25)	0.22 (21.04)	0.01 (17.15)
AOD+LST+NDVI	0.22 (21.26)	0.22 (21.19)	0.01 (17.27)
AOD+LST+NDVI+EV	0.25 (20.49)	0.25 (20.38)	0.01 (17.35)
AOD+LST+NDVI+EV+WOY	0.51 (18.42)	0.51 (17.94)	0.35 (14.07)
AOD+LST+NDVI+EV+WOY+Year	0.51 (18.28)	0.52 (17.83)	0.35 (13.78)
RF			
AOD	0.79 (11.39)	0.16 (23.08)	0.02 (20.52)
AOD+LST	0.86 (10.12)	0.25 (20.88)	0.04 (18.59)
AOD+LST+NDVI	0.90 (8.82)	0.44 (17.87)	0.10 (16.03)
AOD+LST+NDVI+EV	0.89 (8.82)	0.60 (15.17)	0.15 (15.05)
AOD+LST+NDVI+EV+WOY	0.92 (7.23)	0.74 (12.35)	0.60 (10.47)
AOD+LST+NDVI+EV+WOY +Year	**0.95 (5.58)**	**0.78 (11.18)**	**0.71 (8.79)**
XGBoost			
AOD	0.31 (19.77)	0.27 (20.27)	0.04 (17.45)
AOD+LST	0.34 (19.34)	0.30 (19.85)	0.05 (17.63)
AOD+LST+NDVI	0.40 (18.39)	0.38 (18.71)	0.08 (15.90)
AOD+LST+NDVI+EV	0.49 (16.94)	0.47 (17.34)	0.12 (15.23)
AOD+LST+NDVI+EV+WOY	0.61 (14.93)	0.60 (15.14)	0.43 (12.40)
AOD+LST+NDVI+EV+WOY+Year	0.62 (14.74)	0.60 (15.00)	0.45 (12.12)
SVM			
AOD	0.28 (20.59)	0.28 (20.66)	0.04 (17.15)
AOD+LST	0.31 (20.08)	0.31 (20.16)	0.05 (16.91)
AOD+LST+NDVI	0.39 (18.83)	0.38 (18.93)	0.09 (15.68)
AOD+LST+NDVI+EV	0.47 (17.60)	0.46 (17.79)	0.14 (15.65)
AOD+LST+NDVI+EV+WOY	0.59 (15.64)	0.60 (15.44)	0.51 (11.51)
AOD+LST+NDVI+EV+WOY+Year	0.61 (15.32)	0.62 (15.17)	0.52 (11.63)

3.3. Estimation of daily PM2.5

RF approaches were used to estimate daily PM2.5 concentrations in Thailand, and it was found that the model that included AOD, LST, NDVI, EV, WOY, and year had the best performance. The RF results also show two alternative measurements of each predictor variable's relative contribution in Figure 3. The %IncMSE is a percentage increase in mean square error, equivalent to accuracy-based importance. The IncNodePurity, calculated similarly to Gini-based importance, is based on reducing the sum of squared errors whenever a variable is split. Without WOY, AOD, EV, year, LST, and NDVI as predictors, the %IncMSE was 72.4%, 59.3%, 50.7%, 43.2%, 32.4%, and 31.5%, respectively. The important variables for IncNodePurity were WOY, AOD, EV, NDVI, LST, and year, respectively. These two measurements were calculated using different methods due to their strong association with ground-level PM2.5. Additionally, all the factors were needed to estimate PM2.5 levels in Thailand, where WOY, AOD, and EV were the three most essential variables in the two measurements.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1833.v1

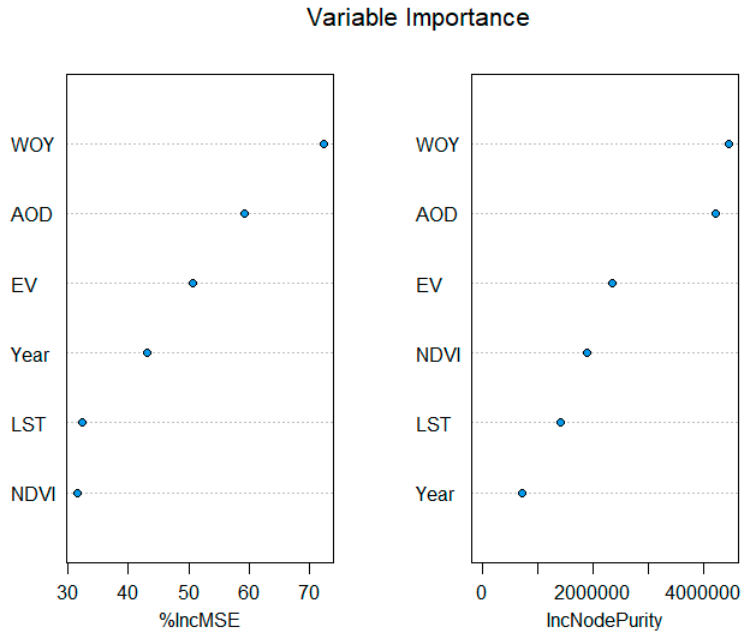


Figure 3. The importance variables for estimation of PM2.5.

Figure 4 presents the PM2.5 time series plot and estimation for the training, validation, and testing data. The three plots exhibit a consistent pattern in the observed and estimated PM2.5 concentrations, with the highest concentrations observed during weeks 45 to 53 (November to December) and 1 to 10 (January to March). The difference between the measured and estimated PM2.5 concentrations in the testing dataset was slight in 2015 and 2016 but remained consistent in 2017 and 2020.

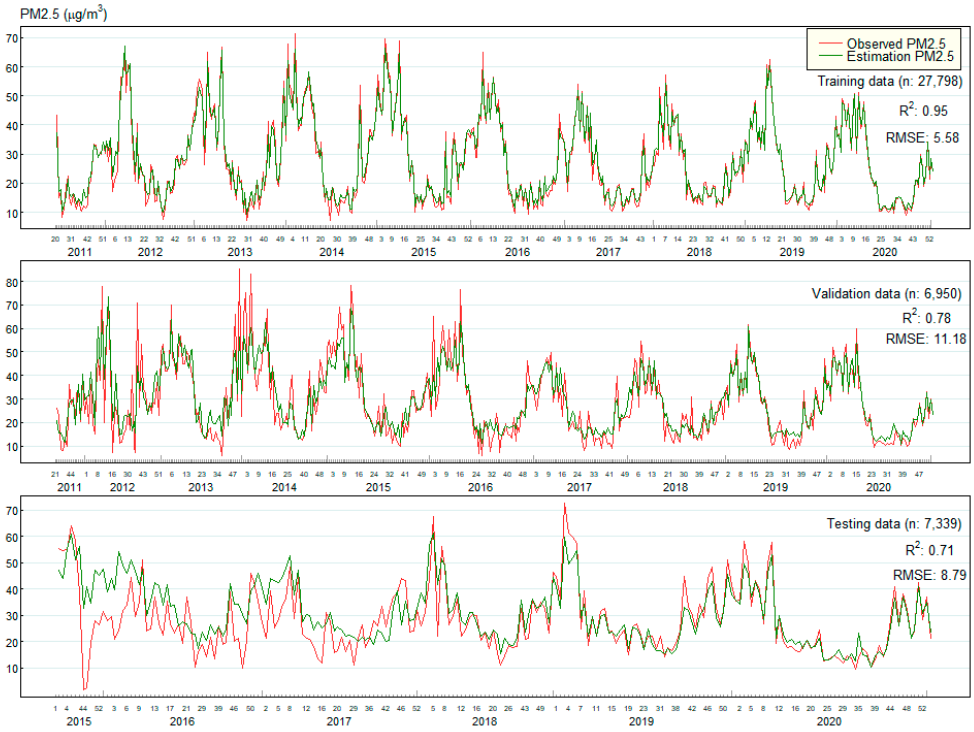
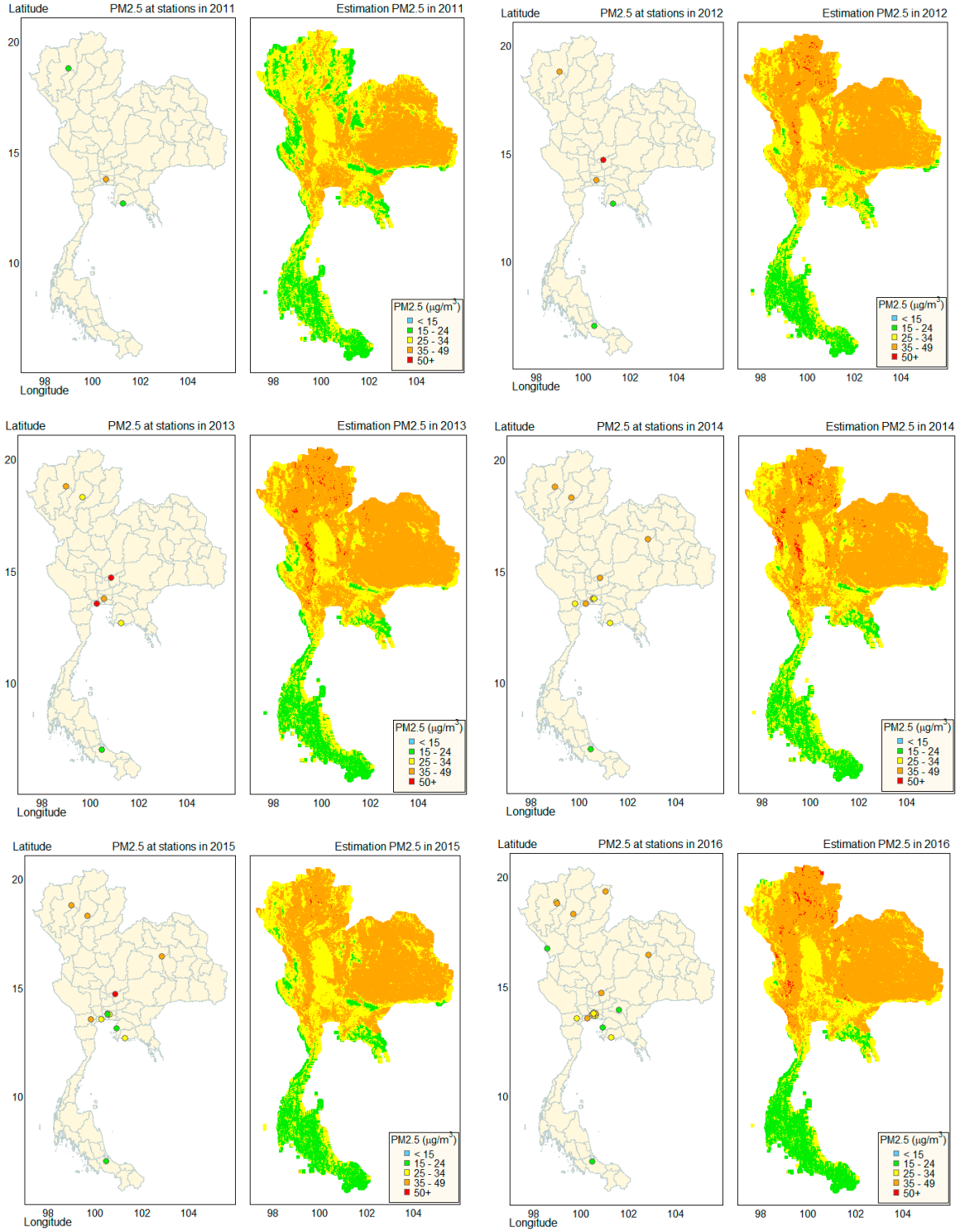


Figure 4. Time series plot of PM2.5 observed and estimation of PM2.5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1833.v1

Figure 5 presents the estimation of PM2.5 concentrations from 2011 to 2020 at a 1 km resolution using the RF model. The values of PM2.5 at stations and the estimated PM2.5 are comparable. Northern Thailand exhibited the highest PM2.5 concentrations, while Southern Thailand showed the lowest levels. Except for the southern part of Thailand, most of the region's PM2.5 levels exceeded the WHO 24-hour standard of 15 μg/m³ but remained below Thailand's national standard limit of 50 μg/m³ overall.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 May 2023 doi:10.20944/preprints202305.1833.v1

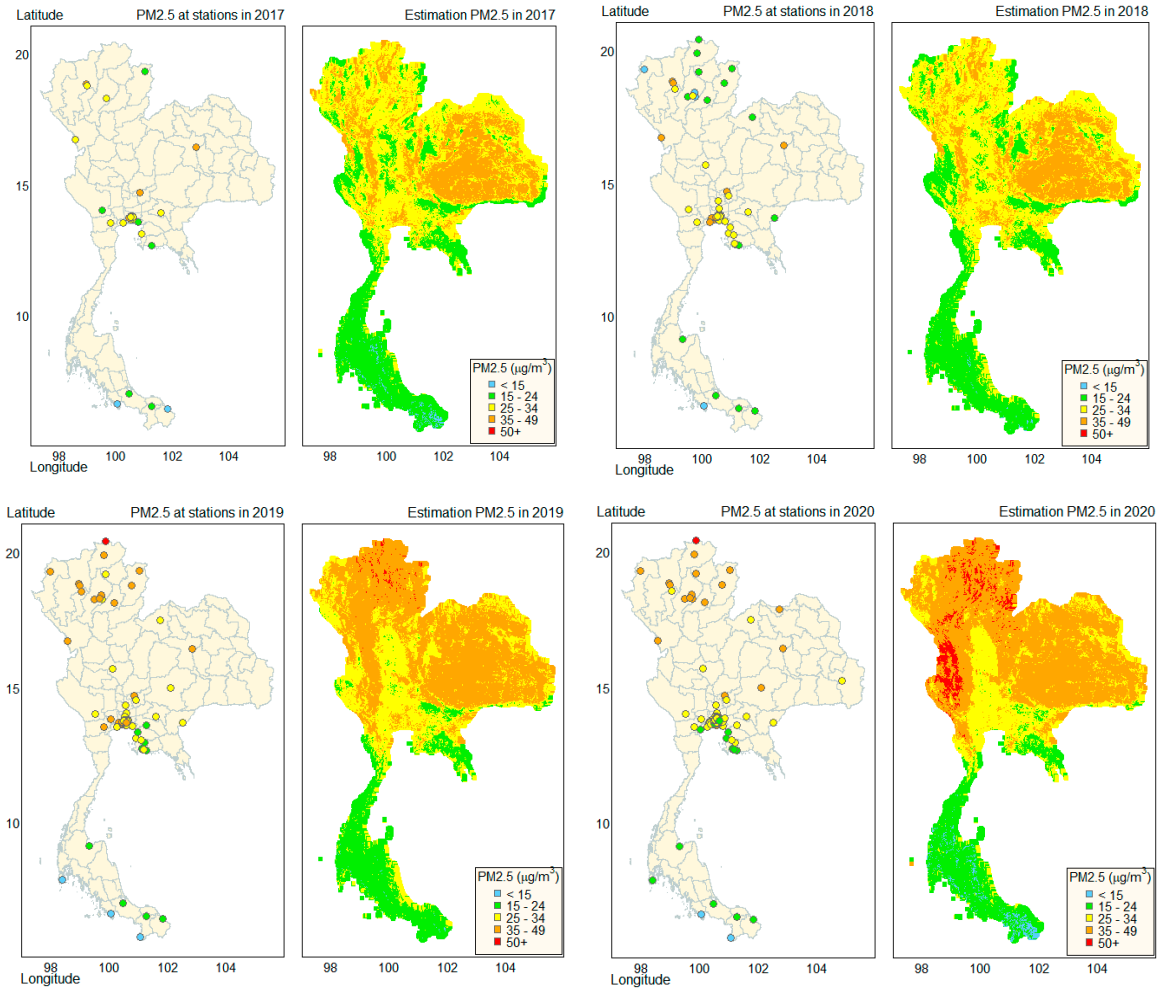


Figure 5. Estimation of PM2.5 in Thailand 2011-2020 in each pixel has a 1 km resolution.

4. Discussion

We proposed using satellite data with a 1 km resolution to predict daily PM2.5 concentrations in Thailand and identified the best model to achieve this. The results of this model estimation can be utilized as standards for simulating PM2.5 in other areas with a similar mix of pollution sources and a need for more monitoring to understand the particle's spatiotemporal distribution. Investigating the spatiotemporal variations of PM2.5 at small scales was made possible by estimating PM2.5 in 1 km grid cells. These PM2.5 values are intended to aid epidemiological research and assist individuals in making informed decisions about air pollution.

In our trials, RF outperformed LR, XGBoost, and SVM models. Our findings align with previous PM2.5 estimating studies from other countries, with an R^2 of 0.95 (RMSE of 5.58 µg/m^3) for training data, 0.78 (RMSE of 11.18 µg/m^3) for validation data, and 0.71 (RMSE of 8.79 µg/m^3) for testing data. For example, the predicted PM2.5 in Greater London using RF, GBM, and KNN, with RF providing the best estimation with an R^2 of 0.83 and RMSE of 4.28 µg/m^3 [20]. In another study, using remote sensing data and AOD, eight approaches were used to anticipate monthly PM2.5 in British Columbia, and RF was found to be the most reliable ML method, with an R^2 of 0.49 (RMSE of 2.67 µg/m^3) [18]. The predicted daily PM2.5 at a 1 km grid for 2013–2015 in Italy using RF with an R^2 of 0.80 (RMSE = 7.05 µg/m^3) [35]. The computed 1 km-resolution PM2.5 concentrations in China using RF, with an R^2 of 0.98 (RMSE = 6.40 µg/m^3) for model fitting and an R^2 of 0.81 (RMSE = 17.91 µg/m^3) for model validation [19]. Another Chinese study used RF to predict daily PM2.5 from 2005 to 2016, with an R^2 of 0.77 (RMSE of 22 µg/m^3) [17]. These studies demonstrate that estimating PM2.5 from satellite data using the RF model with an R^2 of 0.49–0.83 (RMSE = 2.67–22 µg/m^3) in the validation data is acceptable. On the other hand, the LR model performed poorly in this study. This may be due to the

 doi:10.20944/preprints202305.1833.v1

positively skewed and non-normally distributed nature of PM2.5 data, which may not be well suited for LR models [36–38].

The study found that the RF model, utilizing AOD, LST, NDVI, EV, WOY, and year as predictors, produced the best results for estimating daily PM2.5 concentrations in Thailand. The strength of the RF model lies in its ability to avoid overfitting data by utilizing the strength of individual trees in the forest and their correlation. However, the results of our study differ from those of other studies, where other models, such as XGBoost, have been found to outperform RF [17]. This may be due to how these decision tree-based models take in and process training data. Our findings suggest that decision tree-based models are recommended for estimating PM2.5 using satellite data.

The results indicate that WOY, AOD, and EV are significant factors in determining PM2.5 concentrations, as shown by the two measurements of the RF model. This is consistent with previous studies, which found AOD and EV to contribute to PM2.5 modeling significantly [18]. Daily PM2.5 concentrations often exhibit a favorable skewed distribution similar to AOD. Similar to the research conducted in China, the bivariate correlation analysis revealed that independent variables such as AOD strongly associate with PM2.5 [19]. Our results also show that the estimated PM2.5 concentrations align well with the observed values at monitoring stations, with similar patterns in the time-series plots for observed and estimated PM2.5. However, there was some discrepancy between observed and estimated PM2.5 concentrations in 2015-2016. This may be due to the less varied geographical distribution of pollutants in the PM2.5 sample taken before 2017, as suggested by research from the United Kingdom [20].

The PM2.5 assessment indicates that northern Thailand experiences higher levels of PM2.5 than other regions, particularly during the dry seasons of WOY 1–10 (January–March) and WOY 45–53 (November–December). This is attributed to extensive agricultural fields and open-air biomass burning in northern Thailand and neighboring countries [22]. These activities contribute to the elevated PM2.5 levels and also have a significant impact on climate change. Except for the southern region, most areas in Thailand surpass the WHOs 24-hour standard of 15 μg/m^3 for PM2.5 levels, although they remain within the national limit of 50 μg/m^3. The high PM2.5 levels can negatively impact population health, including respiratory and cardiovascular diseases. Our model's PM2.5 data can be used to identify links between PM2.5 levels and specific geographic areas, such as provinces, districts, and sub-districts.

Although satellite data can provide higher coverage than ground monitoring stations for PM2.5 data, it often has lower temporal coverage due to lousy observation conditions such as clouds and fog. We used average satellite data within a 5 km radius of the stations to decrease missing values. In our analysis, we used 42,009 (or 33.6%) data points out of 124,846 valid data points. According to evaluate MODIS collection 6 AOD retrievals against ground sunphotometer observations over East Asia cloud cover or high surface reflectance can cause an average of 40% to 70% of satellite retrievals to go unrecovered [39]. Furthermore, Thailand's overcast or foggy weather can invalidate the satellite retrieval technique by reducing the sampling frequency of accessible satellite data. This issue has also been identified in a study conducted in China [9]. As a result, new monitoring methods with wider spatial coverage and fewer weather limitations should be developed. These strengths can be used as benchmarks when estimating ground-level PM2.5 or other air pollution metrics in Thailand or other countries using remote sensing.

5. Conclusions

This study proposed an efficient method for estimating daily PM2.5 concentrations in Thailand using satellite data with a pixel resolution of 1 km. The RF model was the most effective compared to LR, XGBoost, and SVM models. The use of AOD, LST, NDVI, EV, WOY, and year as predictor variables improved the model's performance, resulting in R^2 values of 0.95 (RMSE of 5.58 μg/m^3) for the training dataset, 0.78 (RMSE of 11.18 μg/m^3) for the validation dataset, and 0.71 (RMSE of 8.79 μg/m^3) for the testing dataset. The results from 2011 to 2020 were consistent with PM2.5 values obtained from monitoring stations. Using satellite data in this study allowed for examining air quality

at various regional and temporal scales. The developed models and projections can aid regulatory operations and future epidemiological research in Thailand.

Author Contributions: S.B., Conceptualization, Formal analysis, Writing - original draft. S.U., Supervision, Writing - review & editing. G.H., Writing - review & editing. J.K., Writing - review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This study was encouraged by the Sirindhorn International Institute of Technology (SIIT), Thammasat University Research Fund and Japan Advanced Institute of Science and Technology (JAIST), and the research fund of Thailand's National Electronics and Computer Technology Centre (NECTEC).

Data Availability Statement: PM2.5 data from PCD (http://air4thai.pcd.go.th/webV2/history/, accessed on 18 May 2023) and BAQ (https://bangkokairquality.com/bma/report?lang=en, accessed on 18 May 2023). The satellite data can be assessed at (https://ladsweb.modaps.eosdis.nasa.gov/search/, accessed on 18 May 2023).

Acknowledgments: The Pollution Control Department and Bangkok's Air Quality and Noise Management Division provided the PM2.5 data, which the authors are thankful for. We appreciate Professor Don McNeil's wise counsel.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO. WHO | Air pollution. *World Health Organization*. Published online 2019.
2. Dockery DW. Health Effects of Particulate Air Pollution. *Ann Epidemiol*. 2009;19(4). doi:10.1016/j.annepidem.2009.01.018
3. WHO. Ambient (outdoor) air pollution. World Health Organisation Geneva.
4. Chung Y, Dominici F, Wang Y, Coull BA, Bell ML. Associations between long-term exposure to chemical constituents of fine particulate matter (PM2. 5) and mortality in Medicare enrollees in the eastern United States. *Environ Health Perspect*. 2015;123(5):467-474.
5. Lu F, Xu D, Cheng Y, et al. Systematic review and meta-analysis of the adverse health effects of ambient PM<inf>2.5</inf> and PM<inf>10</inf> pollution in the Chinese population. *Environ Res*. 2015;136:196-204. doi:10.1016/j.envres.2014.06.029
6. Bae S, Kwon HJ. Current state of research on the risk of morbidity and mortality associated with air pollution in korea. *Yonsei Med J*. 2019;60(3). doi:10.3349/ymj.2019.60.3.243
7. Carmona JM, Gupta P, Lozano-García DF, Vanoye AY, Hernández-Paniagua IY, Mendoza A. Evaluation of modis aerosol optical depth and surface data using an ensemble modeling approach to assess pm<inf>2.5</inf> temporal and spatial distributions. *Remote Sens (Basel)*. 2021;13(16). doi:10.3390/rs13163102
8. Maheshwarkar P, Sunder Raman R. Population exposure across central India to PM2. 5 derived using remotely sensed products in a three-stage statistical model. *Sci Rep*. 2021;11(1):544.
9. Xu X, Zhang C. Estimation of ground-level PM<inf>2.5</inf>concentration using MODIS AOD and corrected regression model over Beijing, China. *PLoS One*. 2020;15(10 October). doi:10.1371/journal.pone.0240430
10. Yang Q, Yuan Q, Yue L, Li T, Shen H, Zhang L. The relationships between PM2.5 and aerosol optical depth (AOD) in mainland China: About and behind the spatio-temporal variations. *Environmental Pollution*. 2019;248:526-535. doi:10.1016/j.envpol.2019.02.071
11. Zeydan Ö, Wang Y. Using MODIS derived aerosol optical depth to estimate ground-level PM<inf>2.5</inf> concentrations over Turkey. *Atmos Pollut Res*. 2019;10(5):1565-1576. doi:10.1016/j.apr.2019.05.005
12. Pavolonis M, Sieglaff J. *GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document for Volcanic Ash (Detection and Height)*. University of Wisconsin--Madison; 2010.
13. Engel-Cox JA, Holloman CH, Coutant BW, Hoff RM. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmos Environ*. 2004;38(16). doi:10.1016/j.atmosenv.2004.01.039
14. Zhang X, Chu Y, Wang Y, Zhang K. Predicting daily PM<inf>2.5</inf> concentrations in Texas using high-resolution satellite aerosol optical depth. *Science of the Total Environment*. 2018;631-632:904-911. doi:10.1016/j.scitotenv.2018.02.255
15. Joharestani MZ, Cao C, Ni X, Bashir B, Talebiesfandarani S. PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere (Basel)*. 2019;10(7). doi:10.3390/atmos10070373
16. Chu Y, Liu Y, Li X, et al. A review on predicting ground PM<inf>2.5</inf> concentration using satellite aerosol optical depth. *Atmosphere (Basel)*. 2016;7(10). doi:10.3390/atmos7100129

17. Xiao Q, Chang HH, Geng G, Liu Y. An Ensemble Machine-Learning Model to Predict Historical PM<inf>2.5</inf> Concentrations in China from Satellite Data. *Environ Sci Technol*. 2018;52(22):13260-13269. doi:10.1021/acs.est.8b02917
18. Xu Y, Ho HC, Wong MS, et al. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environmental Pollution*. 2018;242. doi:10.1016/j.envpol.2018.08.029
19. Wei J, Huang W, Li Z, et al. Estimating 1-km-resolution PM<inf>2.5</inf> concentrations across China using the space-time random forest approach. *Remote Sens Environ*. 2019;231. doi:10.1016/j.rse.2019.111221
20. Danesh Yazdi M, Kuang Z, Dimakopoulou K, et al. Predicting fine particulate matter (PM2. 5) in the greater london area: An ensemble approach using machine learning methods. *Remote Sens (Basel)*. 2020;12(6):914.
21. Kanabkaew T. Prediction of hourly particulate matter concentrations in Chiangmai, Thailand using MODIS aerosol optical depth and ground-based meteorological data. *EnvironmentAsia*. 2013;6(2).
22. Phuengsamran P, Lalitaporn P. Estimating Particulate Matter Concentrations in Central Thailand Using Satellite Data. *Thai Environmental Engineering Journal*. 2021;35(3):1-11.
23. Kloog I, Koutrakis P, Coull BA, Lee HJ, Schwartz J. Assessing temporally and spatially resolved PM<inf>2.5</inf> exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos Environ*. 2011;45(35):6267-6275. doi:10.1016/j.atmosenv.2011.08.066
24. Lyapustin A, Wang Y. MCD19A2 MODIS/Terra+ aqua land aerosol optical depth daily L2G global 1km SIN grid V006 [data set]. *NASA EOSDIS land processes DAAC*. Published online 2018.
25. Wan Z, Hook S, Hulley G. MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006. 2015, Distributed by NASA EOSDIS Land Processes DAAC. Published online 2015.
26. Wan Z, Hook S, Hulley G. MYD11A1 MODIS/Aqua land surface temperature/emissivity daily L3 global 1km SIN Grid V006. NASA EOSDIS LP DAAC. Published online 2015.
27. Breiman L. Random forests. *Mach Learn*. 2001;45(1). doi:10.1023/A:1010933404324
28. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(3).
29. Breiman L, Cutler A, Liaw A, Wiener M. Package 'randomForest' - Breiman and Cutler's Random Forests for Classification and Regression. *CRAN Repository*. Published online 2018.
30. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-August-2016. ; 2016. doi:10.1145/2939672.2939785
31. Chen T, He T, Benesty M, et al. *Package "Xgboost."*; 2019.
32. Sain SR, Vapnik VN. The Nature of Statistical Learning Theory. *Technometrics*. 1996;38(4). doi:10.2307/1271324
33. Zhao D, Qi L. Prediction of Maximum Power of PV System based on SVR Algorithm. *Journal of Jilin Institute of Chemical Technology*. 2015;32(06):89-94.
34. Meyer D. Support vector machines: the interface to libsvm in package e1071. *… Systems and their …*. 2014;1. doi:10.1007/978-0-387-77242-4
35. Stafoggia M, Bellander T, Bucci S, et al. Estimation of daily PM<inf>10</inf> and PM<inf>2.5</inf> concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environ Int*. 2019;124:170-179. doi:10.1016/j.envint.2019.01.016
36. Boulesteix AL, Schmid M. Machine learning versus statistical modeling. *Biometrical Journal*. 2014;56(4). doi:10.1002/bimj.201300226
37. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods*. 2018;15(4). doi:10.1038/nmeth.4642
38. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13. doi:10.1016/j.csbj.2014.11.005
39. Xiao Q, Zhang H, Choi M, et al. Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia. *Atmos Chem Phys*. 2016;16(3). doi:10.5194/acp-16-1255-2016.