Thesis for the Degree of Master of Engineering

Prediction of $PM_{2.5}$ concentration based on machine learning using MODIS AOD and LDAPS data

by

Minji Ryu

Division of Earth Environmental System Science

(Major of Spatial Information Engineering)

The Graduate School

Pukyong National University

February, 2023

Prediction of PM$_{2.5}$ concentration based on machine learning using MODIS AOD and LDAPS data

(MODIS AOD와 LDAPS 자료를 활용한 머신러닝 기반 PM$_{2.5}$ 농도 예측)

Advisor: Prof. Jinsoo Kim

by
Minji Ryu

A thesis submitted in partial fulfillment of the requirements
for the degree of
Master of Engineering
in Division of Earth Environmental System Science (Major of Spatial Information Engineering), The Graduate School,
Pukyong National University

February, 2023

Prediction of $PM_{2.5}$ concentration based on machine learning using MODIS AOD and LDAPS data

A dissertation

by

Minji Ryu

Approved by:

(Chairman) Prof. Kyung-Soo Han

ywLee

(Member) Prof. Yangwon Lee

(Member) Prof. Jinsoo Kim

February 17, 2023

CONTENTS

LIST OF FIGURES

LIST OF TABLES

MODIS AOD와 LDAPS 자료를 활용한 머신 러닝 기반 $PM_{2.5}$ 농도 예측

류 민 지

부경대학교대학원지구환경시스템과학부공간정보시스템공학전공

요 약

미세먼지(particle matter less than 2.5μg/㎥, $PM_{2.5}$)는 주요 대기오염물질로 호흡기나 심혈관 질환 등의 원인이 되며 건강에 위협을 주는 물질이다. 인간의 건강을 위해 미세먼지를 모니터링하고 예방하는 것이 중요하며 이를 위해서는 미세먼지를 예측하는 것이 필요하다. 본 연구에서는 $PM_{2.5}$의 정확한 예측에 필요한 머신 러닝 모델을 구축하였다. 예측을 위한 독립변수로 2017년부터 2019년까지의 local data assimilation and prediction system (LDAPS) 모델링 데이터의 기상인자 15가지(NCPCP, UGRD, VGRD, TMP, TMIN, TMAX, RH, MAXGUST, 50MINU, 50MINV, 50MAXU, 50MAXV, STMP, HPBL, PRES)와 화학인자 4가지(CO, O_3, SO_2, NO_2), aerosol optical depth (AOD) 2가지(470㎚, 550㎚) 파장대의 데이터를 선정하였다. 종속변수는 서울시 40개 AQMS 지점에서 관측된 $PM_{2.5}$ 값으로 하였다. 예측에 사용된 머신러닝 모델은 앙상블 기반의 알고리즘인 RF, GBM, XGB를 이용하였으며 예측 정확도 지표는 R^2, Bias, RMSE, MAE를 적용하였다. 예측 결과, XGB가 R^2=0.89, Bias = -0.143, RMSE = 4.719, MAE = 3.502로 예측에 적합한 모델임을 확인하였다. 그러나, RF와 GBM 모델 또한 정확도 면에서 큰 차이를 보이지 않아 세 모델 모두 좋은 성능임을 확인하였다. 변수 중요도 평가를 통해 모델의 훈련에 기여한 인자를 파악하였으며, O_3와 SO_2,, NCPCP 등이 높은 기여도를 나타냈다. 또한, 결측 값으로 인한 AOD의 영향력을 평가하기 위해 AOD를 제외한 모델의 성능을 비교 분석하였으며, AOD가 포함되었을 때, 약 5%의 정확도 향상의 결과를 나타냈다. 계절별 모델 성능의 분석도 실시하였으며, 봄과 겨울에 높은 정확도를 보였으며, 여름과 가을에 낮은 정확도를 나타냈다. 본 연구는 서울시의 $PM_{2.5}$ 예측을 위해 LDPAS 기상인자, MODIS AOD, 지상관측 자료인 화학인자 등을 활용하고자 하였으며, 안정적인 성능을 보이는 앙상블 기반의 RF, GBM, XGB알고리즘을 이용하여 예측을 수행하였다. 본 연구의 결과를 통해 실시간으로 변하는 $PM_{2.5}$ 농도의 모니터링과 관리 및 대책 수립에 도움이 될 것으로 기대된다.

1. Introduction

1.1. Background

Air pollutants are divided into gaseous substances and particulate matter, and $PM_{2.5}$ (Particle Matter) and PM_{10} included in particulate matter are air pollutants with aerodynamic diameters of 2.5μg/m^3 and 10μg/m^3 or less, respectively (Han and Kim, 2015). $PM_{2.5}$ is a chemical reaction to automobile exhaust, fossil fuel combustion, and air pollutants emitted from factory manufacturing processes such as sulfur oxide, nitrogen oxide, and ammonia, which are produced as secondary substances, and is closely related to weather changes such as temperature, wind speed, air pressure, and humidity (Kim and Jang, 2021).

Along with the recent discussion of health risks caused by air pollution, the World Health Organization (WHO) designates $PM_{2.5}$ emitted from diesel cars as a class 1 carcinogen (Hwang et al, 2018). $PM_{2.5}$ is also associated with the development of respiratory and cardiovascular diseases as it can penetrate the alveoli when inhaled by humans due to its very small particle size (Choi, 2018a; Yoo et al., 2020; Azari et al., 2021). In order to prepare for the risks caused by $PM_{2.5}$, monitoring and forecasting activities to find vulnerabilities in advance are essential (Shin and Kim, 2015). Therefore, the concentration of particulate matter in the air must be continuously monitored, and a model that accurately predicts high concentrations of PM is needed (Chae et al., 2021).

In Korea, increased economic activity, concentration of population in large cities, and increased number of vehicles are causing serious air quality degradation and related problems in small and medium-sized cities and metropolitan areas (Park et al, 2017a). In particular, the air quality in Seoul, which is located in the metropolitan area, deteriorated compared to other regions due to the concentration of population and facilities (Kim and Yeo, 2013).

According to the Seoul Metropolitan Government's air pollution statistics, the average concentration of $PM_{2.5}$ in Seoul from 2015 to 2020 was about 24$\mu g/m^3$, far exceeding the current annual average environmental standard for $PM_{2.5}$ in Korea of 15$\mu g/m^3$, and management for this is necessary.

Recently, research on reliable measurement technology and control technology and services in terms of management of PM has been continuously conducted, and research and development for rapid response are being conducted (Kim, 2022). Therefore, in order to continuously reduce and manage $PM_{2.5}$ concentrations in Seoul, it is necessary to analyze trends in $PM_{2.5}$ concentrations and study concentration predictions considering various factors.

1.2. Literature review

PM is a substance that is complexly affected by various factors, and its effect on PM concentration varies depending on weather factors and seasons, so understanding the effect of these environmental factors in PM management is essential (Choi et al, 2018b). Song and Park (2022) analyzed the occurrence pattern of $PM_{2.5}$ in consideration of land use type, temperature, and wind speed factors in Changwon National Industrial Complex, and derived a result that the lower the temperature and wind speed, the higher the concentration of $PM_{2.5}$. Park and Shin (2017) conducted a study in consideration of factors such as seasonal wind direction factors to analyze the influencing factors of $PM_{2.5}$ in Korea, and as a result of the analysis, it was reported that the west wind direction had an effect on $PM_{2.5}$. According to a study that predicts fine dust by applying various independent variables, Seo and Yom (2019) performed prediction using weather factors such as temperature, precipitation, and wind speed for predicting PM. Chen et al., (2018a) used data such as aerosol optical depth (AOD), temperature, air pressure, wind speed, humidity weather factors, and land cover as factors for predicting $PM_{2.5}$ concentrations across Beijing, China. Choi et al. (2022) performed $PM_{2.5}$ concentration prediction using carbon monoxide (CO), sulfer oxides (SO_2), nitrogen oxides (NO_2), ozone (O_3), and PM_{10} as independent variables for $PM_{2.5}$ concentration prediction. The study of Donkelaar et al., (2006) predicted $PM_{2.5}$ on the ground using AODs observed on several satellites and reported that temporal variations in AODs are the most influential parameters of the relationship between satellite and $PM_{2.5}$ ground measurements. Zhang and Kondragunta (2021)'s study attempted to predict the concentration of $PM_{2.5}$ through AOD by applying the regression relationship between

$PM_{2.5}$ and AOD. As described above, meteorology, chemical factors, and AOD were selected as essential independent variables for $PM_{2.5}$ concentration prediction in various studies.

Looking at the $PM_{2.5}$ concentration prediction study using machine learning, Kim (2020) predicted environmental and meteorological variables as independent variables for $PM_{2.5}$ concentration in Seoul, and Lee and Lee (2020) proposed a random forest (RF) method that used the number of bootstraps adjusted by preprocessing ground observation data in time series. In a study using a deep learning-based model, Lin et al. (2020) used Classification and Regression Trace (CART), support vector machine, gradient boosting machine (GBM), long short-term memory (LSTM), and Recurrent neural network as $PM_{2.5}$ concentration prediction models, and derived high accuracy results. In the study of Gao and Li (2021), a graph-based LSTM model was proposed for $PM_{2.5}$ concentration prediction in Gansu Province, China. Shogrkhodaei et al., (2021) performed $PM_{2.5}$ spatio-temporal modeling using three machine learning algorithms: RF, AdaBoost, and Stochastic gradient descent, and reported that the modeling accuracy of the RF algorithm is the best result. Recently, most of the $PM_{2.5}$ prediction research cases have used artificial intelligence algorithms, and among them, many studies using ensemble-based algorithms such as RF, GBM, and XGB, which show stable performance in classification or regression models, or deep learning have been actively conducted.

1.3. Objectives

For $PM_{2.5}$ concentration prediction, this study aims to use RF, a model that is basically used in many studies, GBM, which has high accuracy in prediction among machine learning algorithms, and XGB, which has fast learning speed and excellent performance, as methodologies. As independent variables, weather, satellite AOD, and ground observation data such as wind direction, wind speed, and temperature were used, and a total of 24 independent variables were applied to the model to find a model with excellent prediction performance and compare and analyze the performance of each model. In addition, since the concentration of PM is affected by various emission sources and weather conditions, factors that affect the concentration distribution of fine dust due to time-space differences should be identified (Jeong, 2017). Based on this, it is intended to grasp and analyze the influence of factors that contributed to the learning of the model through the evaluation of the importance of each factor. Finally, based on the seasonal characteristics that high concentrations of $PM_{2.5}$ in Korea appear mainly in winter and early spring, we tried to conduct a seasonal analysis by model by evaluating whether the model used in this study performed predictions well (Son et al., 2020). The flow chart of the study is shown in Fig. 1.

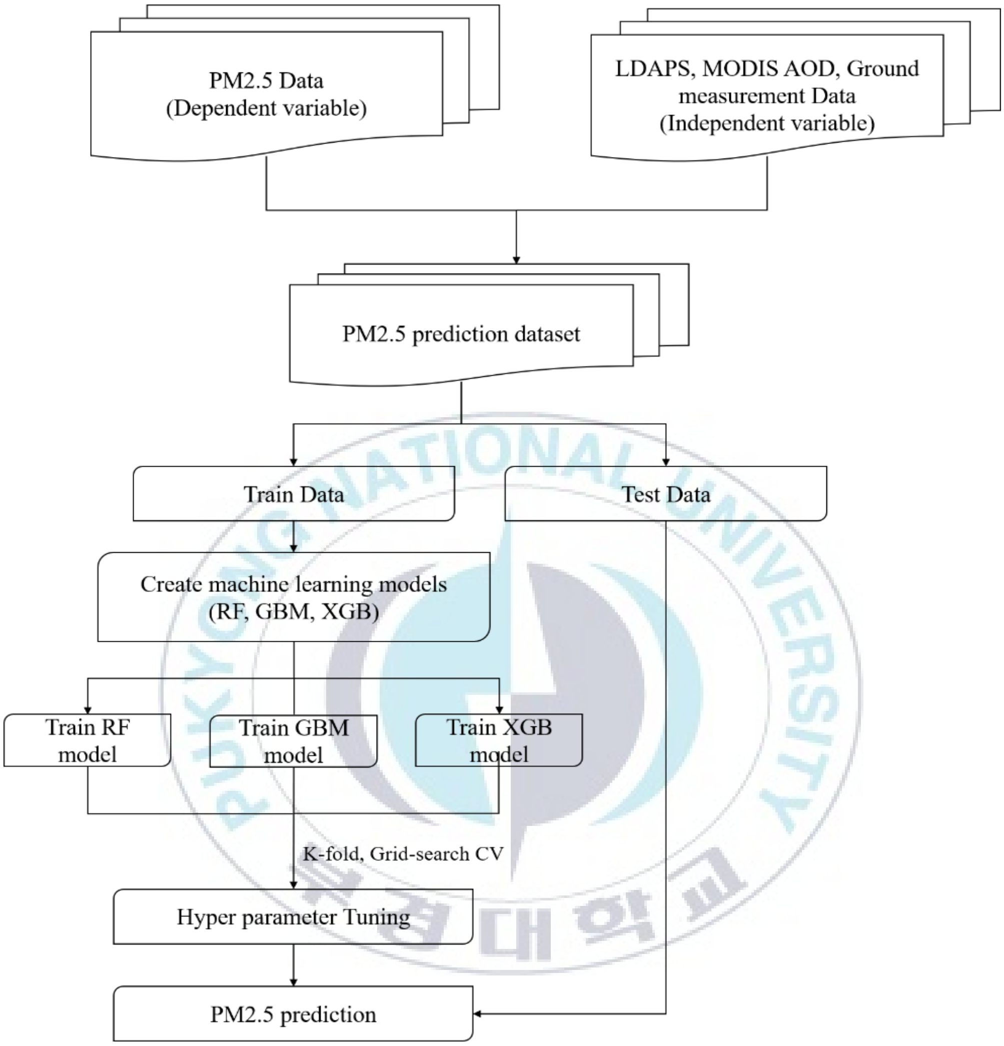

Fig. 1. Flow chart of this study

2. Methodology

2.1. Study Area

Seoul is the capital of Korea and consists of 25 autonomous districts and 423 administrative districts (Lee et al., 2017). In this study, Seoul, where high-rise buildings are concentrated and high vehicle density frequently generates high concentrations of fine dust, was selected as the study target area (Son and Kim, 2021). In addition, the fine dust concentration in Seoul, which has been very important among local governments in terms of population and economic size, is quite high among major cities around the world, so it was selected as a research target area for $PM_{2.5}$ prediction (Hwang, 2018).

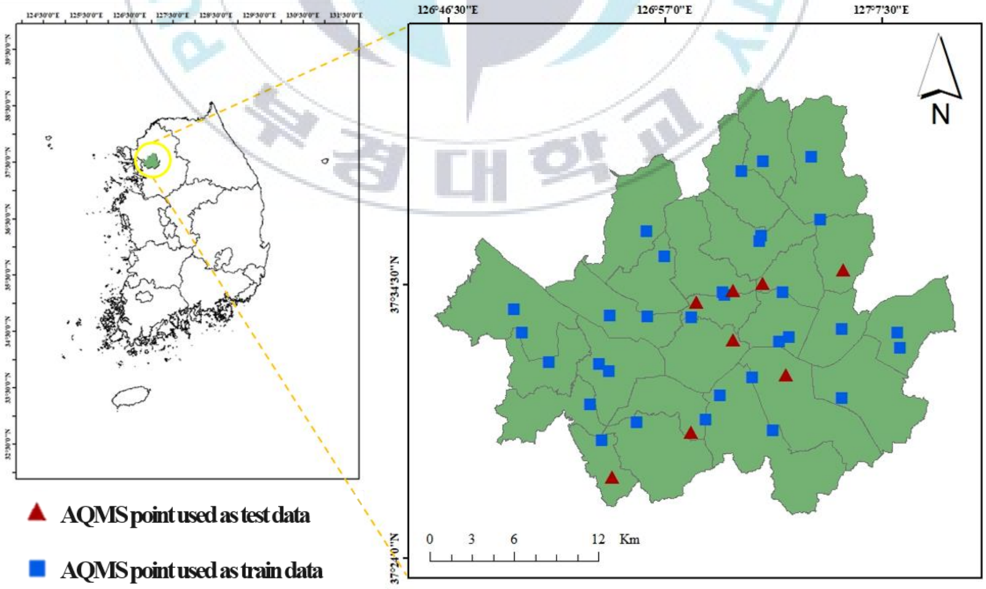

Fig. 2. Study area (a) Seoul location (b) The point where 40 Seoul AQMS points are divided by the train test

The study period is from January 1, 2017 to December 31, 2019, and the total number of air quality monitoring stations (AQMS) located in Seoul during the period 2017 to 2019 is 40, the largest number of AQMS compared to the area. For training and verification of the machine learning model used in this study, train and test data were classified for each AQMS. Among 40 AQMS in Seoul, the train and test stations were split in a ratio of 8:2, and the data of 32 stations were used as the train data set and the data from 8 stations were used as the test data set. The total number of train data sets was 11,550 and the total number of test data sets was 3,750. (Fig. 2).

2.2. Data

A total of 25 data collected for the study were collected as dependent variables $PM_{2.5}$ and independent variable date and AQMS longitude coordinate data, local data estimation and prediction system (LDAPS) meteorological factors, moderation resolution imaging spectrometer (MODIS) AOD, and chemical factors (Table 1). 15 factors such as precipitation, wind-related factors, temperature, humidity, air pressure, and PBHL (Planetary boundary layer) were selected as the LDAPS meteorological factors. MODIS AOD collected data of 470 nm and 550 nm, which are wavelength bands including aerosols. In addition, the observed values of chemical factors CO, SO_2, NO_2, O_3 observed as PM in AQMS and the location data and date, which are spatio-temporal data, were used as factors. The original data of LDAPS and MODIS AOD, excluding ground observation factors that provide numerical data, were constructed as raster format files in tiff format. The two data constructed the final data set by joining the corresponding value of the pixel where the Republic of Korea is located to the ground observation factor based on the AQMS location data. After matching all input data to AQMS points, all data on the day when any one of the 25 factors included missing values were removed, and the final data was constructed with a total of 15,300.

Table 1. Data used for predictors of PM$_{2.5}$

Data	Name	Details	Source	Time Resolution
Air pollutants	PM$_{2.5}$	Particulate matter with a diameter of 2.5 μm or less	K-eco	Hourly
	CO	carbon monoxide		Hourly
	SO$_2$	sulfur dioxide		Hourly
	NO$_2$	nitrogen dioxide		Hourly
	O$_3$	Ozone		Hourly
Time series	Date	Date of Year	-	-
Location	Lat	latitude coordinates AQMS		-
	Lon	Longitude coordinates AQMS		-
Satellite	AOD_470㎚	AOD 470 nm wavelength band	MAIAC MODIS	Daily
	AOD_550㎚	AOD 550 nm wavelength band		Daily
Meteorological (LDAPS)	NCPCP	Large-scale precipitaion	KMA	Hourly
	UGRD	U-component of wind		Hourly
	VGRD	V-component of wind		Hourly
	TMP	Temperature (1.5m)		Hourly
	TMIN	Minimum temperature (1.5m)		Hourly
	TMAX	Maximum temperature (1.5m)		Hourly
	RH	Relative humidity		Hourly
	MAXGUST	Maximum wind speed		Hourly
	50MINU	50m-wind u-component(min)		Hourly
	50MINV	50m-wind v-component(min)		Hourly
	50MAXU	50m-wind u-component(max)		Hourly
	50MAXV	50m-wind v-component(max)		Hourly
	STMP	Surface temperature		Hourly
	HPBL	Planetary boundary layer height		Hourly
	PRES	Surface pressure		Hourly

2.2.1. LDAPS Data

Weather-related factors are essential factors that are reflected in $PM_{2.5}$ prediction, and local forecast model LDAPS data belonging to short- and medium-term prediction among the numerical models provided by Korea Meteorological Administration (KMA) were collected. LDAPS is a numerical prediction model operated by the Korea Meteorological Administration, with a total of 8 predictions per day, including 36 hours for 00, 06, 12, 18 UTC, and 3 hours for initial data generation in 03, 09, 15, and 21 UTC (Byon et al., 2021). LDAPS provides three types of data: isobaric plane, model plane, and single plane, and the data format is provided in the grip2 format suggested by the world meteorological organization. In this study, single-sided data were used, and a total of 78 variables were included for atmosphere, ground, and soil, and 15 weather-related factors were selected to be used for predicting fine dust (Yu et al., 2016). An example of the LDAPS data that performed the preprocessing process is shown in Fig. 3.

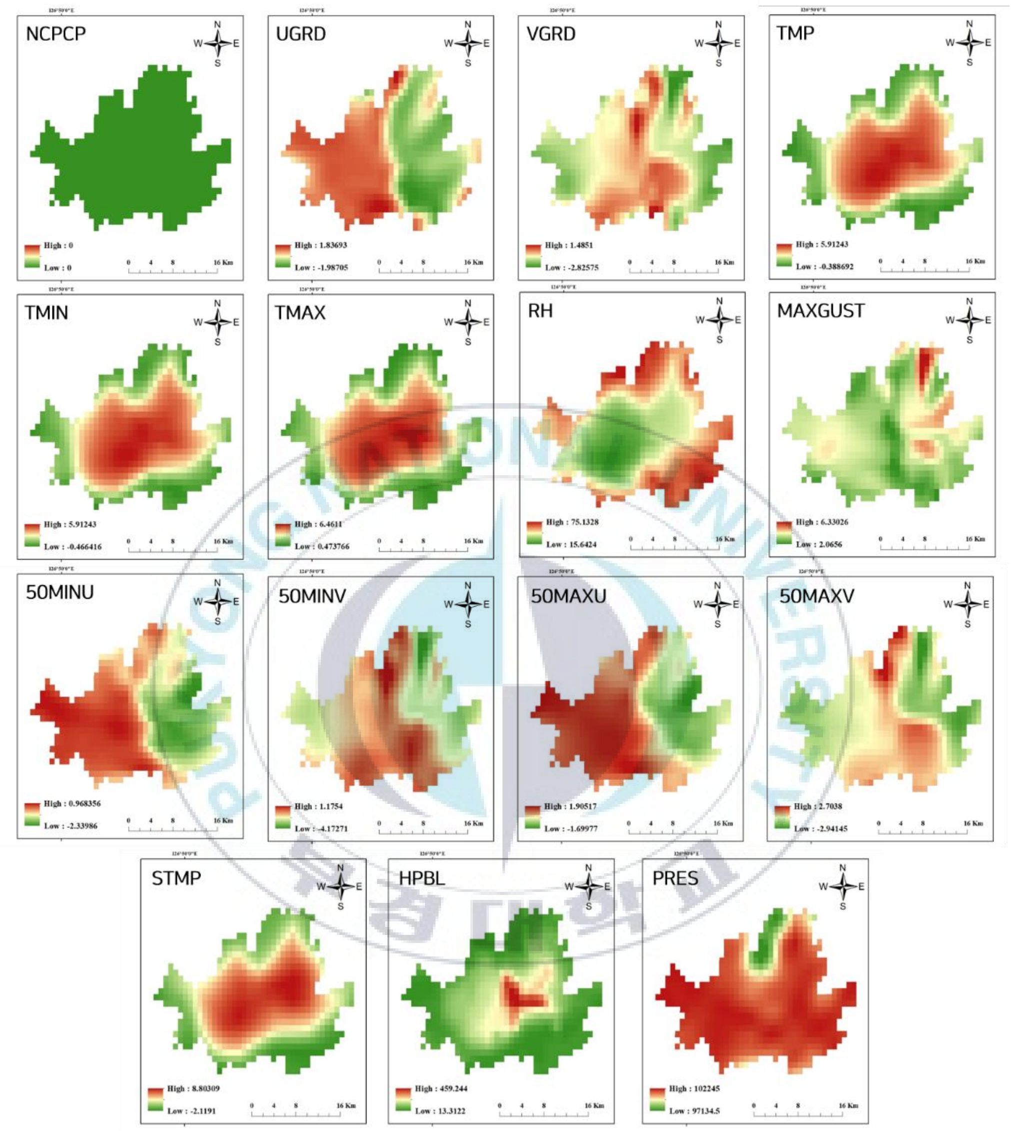

Fig. 3. LDAPS data converted to raster format (February 26, 2018)

2.2.2. MODIS AOD

AOD is a numerical value of solar radiation attenuation by aerosols in the atmosphere (Chen et al., 2018b; Park et al., 2021a). Most of the fine dust prediction studies performed predictions, including AOD, and there were also many studies analyzing the relationship between AOD and fine dust (Kim et al., 2016; Guo et al., 2017; Xie et al., 2015). Therefore, it was judged to be an important variable in predicting PM, and AOD was used as an independent variable. The AOD used in this study is provided by MODIS sensors in Terra and Aqua combined Multi-angle Implementation of Atomospheric Correction (MAIAC) satellites. MODIS has three types of spatial resolution and 36 channels, providing various outputs as high-resolution data, and AOD data necessary for corrective calculation among MODIS data are produced at a spatial resolution of 1 km x 1 km at Level 2 (Park et al., 2017b). An example of MODIS AOD data after preprocessing is shown in Fig. 4.

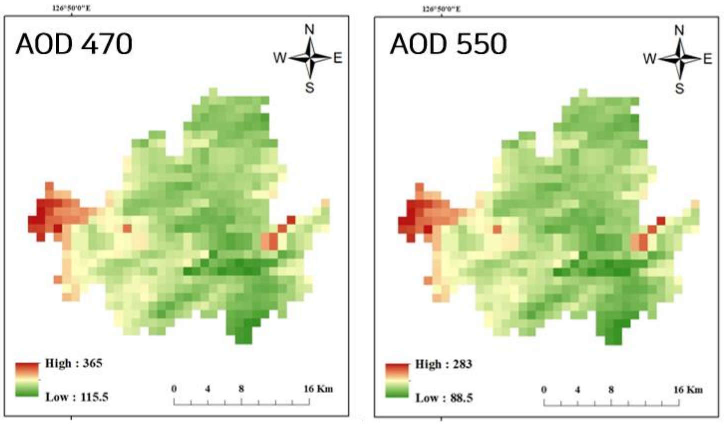

Fig. 4. AOD data converted to raster format (February 26, 2018)

2.2.3. Ground measurement data

In the Air Quality Conservation Act of Korea, a total of 64 types of air pollutants are specified, including PM_{10}, $PM_{2.5}$, CO, SO_2, NO_2, and O_3. Six types of air pollutant data, including PM_{10} and $PM_{2.5}$, were collected from AQMS measured values provided by Air Korea of Korea Environment Corporation (K-eco). There are total of five types of AQMS measurement network: urban atmosphere, national background concentration, suburban atmosphere, roadside atmosphere, and port measurement network. In this study, four measurement network data were used, excluding the port measurement network.

In addition, since phenomena such as PM have both time dependence and spatial dependence, analysis using appropriate tools to consider both spatiotemporal patterns were required, and DOY data and 40 longitude coordinate data of Seoul AQMS were collected (Hwang et al., 2022).

3. Methodology

3.1. RF

RF is an ensemble technique that combines bagging with decision tree (DT) and is a model in which bagging and variable selection methods are the main operating principles (Jang and Park, 2020). It is a method of forming various DTs with bootstrap samples extracted from training data, and it is a concept that improves predictive power by adding randomness in the process of forming multiple DTs to form correlated trees and synthesizing their classification or regression results (Jeong and Jin, 2020). The advantage of RF is that bootstrap samples extracted randomly from the entire data are used for each decision tree analysis, so they are not significantly affected by noise or outlier, and the higher the number of trees, the less overfitting problem occurs according to the law of algebra, resulting in stable results (Breiman, 2001; Kim and Park, 2019).

3.2. GBM

GBM is an algorithm proposed by Friedman (2001) and is based on the boosting principle of Ensemble learning, using gradient descent approach to build a model in the negative sense of the partial derivative of the loss function with respect to the prediction set, and then perform predictions based on it and obtain initial residuals (Ribero and Coelho, 2020)

GBM has three components: loss function, weak learner, and additive model, of which weak learner improves the error rate of existing weak learner and increases accuracy according to itation, which is used to form prediction and strong learner and applies additional weak learner or decision tree (Park, 2022).

GBM minimizes the difference between predicted and actual values by adjusting the weights each time a model is generated, and unlike bagging-based methods, it takes more time to learn the model (Jang et al., 2020).

3.3. XGB

XGB is a type of GBM, designed by Chen and Guestrin (2016). XGB is a method of finding the best tree by reducing the error value using multiple CART. A tree is randomly generated as many times as a set number of times and calculations are repeated, and a model is generated by combining trees with high scores when finally calculated (Sung et al., 2020). XGB was widely recognized in several machine learning and data mining problems, and more than half of 2015's winning tasks on the machine learning competition site 'Kaggle' used XGB, which runs 10 times faster than traditional popular techniques and uses parallel and distributed computing methods to speed up model exploration (Chen and Guestrin, 2016). It aims to solve the problem of overfitting in linear or tree-based models and improve the stability and training speed of large datasets, and XGB, based on such fast and efficient advantages, is used in the prediction model (Ha et al., 2017).

3.4. Model construction and validation

In the RF model construction process, k-fold cross validation (K-fold CV) was applied to find hyperparameters to prevent overfitting and improve model accuracy. The k-fold CV is a method of splitting the data into k pieces, creating k models, training them in k-1 splits, and evaluating them in the remaining splits. The Grid-Search CV is a method in which the user directly inputs the hyperparameter values of the model in the form of a list, and proceeds with the process of finding the optimal parameter values while measuring and evaluating the prediction performance for each number of cases for the values. 5-fold CV and Grid-Search CV were applied, and the final RF model was built by adjusting the n_estimators value of RF. GBM performed the parameter optimization process by applying the 5-fold CV method and the Grid-search CV method. GBM selected hyperparameters of n_estimator, learning_rate, and max_depth. Based on the advantage of fast processing speed, XGBoost went through a parameter optimization process by applying 10-fold CV and Grid-Search CV. XGB has a wide variety of parameter types, so the range of adjustment is very wide. In this study, the final model was built by adjusting subsample, max_depth, colsample_bytree, learning_rate, nthread, n_estimators, and min_child_weight. Table 2 shows the hyperparameter values of each model selected through the parameter optimization process.

Table 2. Hyperparameters selected for model construction

Model	Parameter	Value
RF	n_estimator	300
GBM	n_estimator	100
	learning_rate	0.1
	max_depth	5
XGB	n_estimator	500
	learning_rate	0.07
	max_depth	7
	subsample	0.7
	colsample_bytree	0.7
	nthread	4
	min_child_weight	4

4. Results

4.1. Model performance

In order to evaluate the predictive performance of the three finally built models, the accuracy was calculated by selecting the r-squared score (R²), root mean square errors (RMSE), mean absolute errors (MAE), and bias as indicators. The formula of the evaluation index is as follows.

$$R^2 = 1 - \frac{\sum(\widehat{Y_i} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \tag{1}$$

$$Bias = \frac{1}{n}\sum_{n}(\hat{y} - y) \tag{2}$$

$$RMSE = \sqrt{\frac{\sum(\hat{y} - y)^2}{n}} \tag{3}$$

$$MAE = \frac{1}{n}\sum|\hat{y} - y| \tag{4}$$

Equation (1) is an expression of R², and the higher the value, the better the performance. Equation (2) is an expression of bias, which indicates how far the predicted value is from the actual observed value. If the bias is high, it can be interpreted as meaning that the difference between the predicted value and the observed value is large. Equations (3) and (4) are formulas for RMSE and MAE, and the smaller the value, the better the performance. The training accuracy of RF is $R^2 = 0.981$, Bias = 0.035, RMSE = 2.289μg/m^3, MAE = 1.627μg/m^3, and for GBM $R^2 = 1$, Bias = 0.000, RMSE = 0.327μg/m^3, MAE = 0.248μg/m^3, XGB confirmed that $R^2 = 0.992$, Bias = -0.001, RMSE = 1.337μg/m^3, MAE = 0.996μg/m^3. The prediction accuracy was $R^2 = 0.881$, Bias = 0.421, RMSE = 5.006μg/m^3, MAE = 3.67μg/m^3 for RF, and $R^2 = 0.888$, Bias = -0.436, RMSE = 4.779μg/m^3, MAE = 3.489μg/m^3 for GBM, and XGB confirmed that $R^2 = 0.89$, Bias = -0.143, RMSE = 4.719μg/m^{3}and MAE = 3.502μg/m^3.

In the case of training accuracy, by evaluating how well the model learned, GBM's R^2 showed a value close to 1, and it was confirmed that the model learned best. GBM, XGB and RF showed the highest accuracy. Table 3 is a table summarizing the accuracy of each model.

Table 3. Model performance results (R^2, Bias, RMSE, MAE)

	Train				Test			
model	**R^2**	**Bias**	**RMSE (㎍/㎥)**	**MAE (㎍/㎥)**	**R^2**	**Bias**	**RMSE (㎍/㎥)**	**MAE (㎍/㎥)**
RF	0.981	0.035	2.289	1.627	0.881	0.421	5.006	3.670
GBM	1.000	0.000	0.327	0.248	0.888	-0.436	4.779	3.489
XGB	0.992	-0.001	1.337	0.996	0.890	-0.143	4.719	3.502

In the case of prediction accuracy, it is the result of predicting the concentration of $PM_{2.5}$ using the test data set and evaluating the accuracy. XGB showed the highest prediction accuracy among the three models, showing an accuracy of $R^2 = 0.89$. Next to XGB, GBM and RF showed similar accuracy, with $R^2 = 0.888$ and 0.881, respectively (Fig. 5).

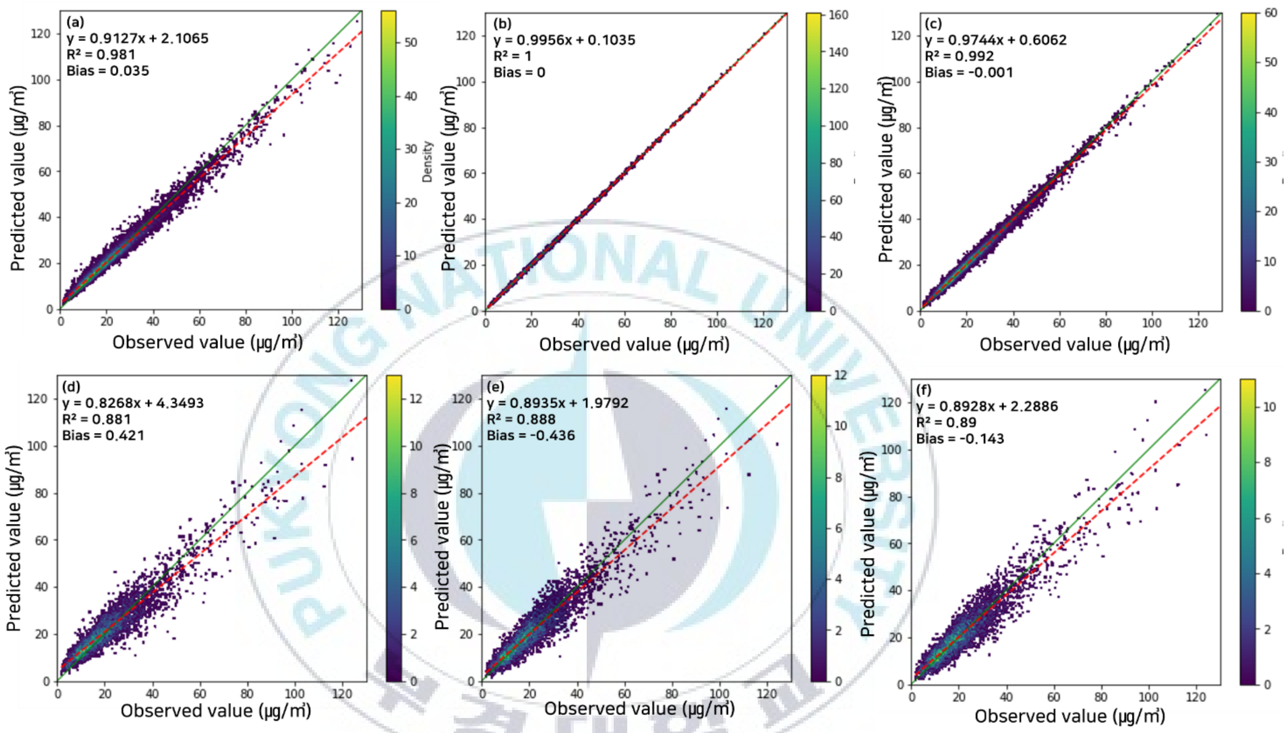

Fig. 5. Scatterplot by model performance

(a) RF train (b) GBM train (c) XGB train (d) RF test (e) GBM test (f) XGB test

4.2. Feature Importance

In order to analyze the contribution to model learning of 24 independent variables used for $PM_{2.5}$ prediction, variable importance was evaluated. The three models used in this study basically have a built-in feature_importances function in Python, so it is possible to evaluate the importance of variables in each model. Therefore, variable importance analysis was performed using the corresponding function (Fig. 6). All models showed similar results, and it was confirmed that SO,, CO, AOD, wind-related factors (50MINU, 50MINV, MAXGUST) were the upper contribution factors, and O,, NO,, Date, and temperature-related factors (TMP, TMAX) were the lower factors. Although there are some differences between models, similar results were shown in common.

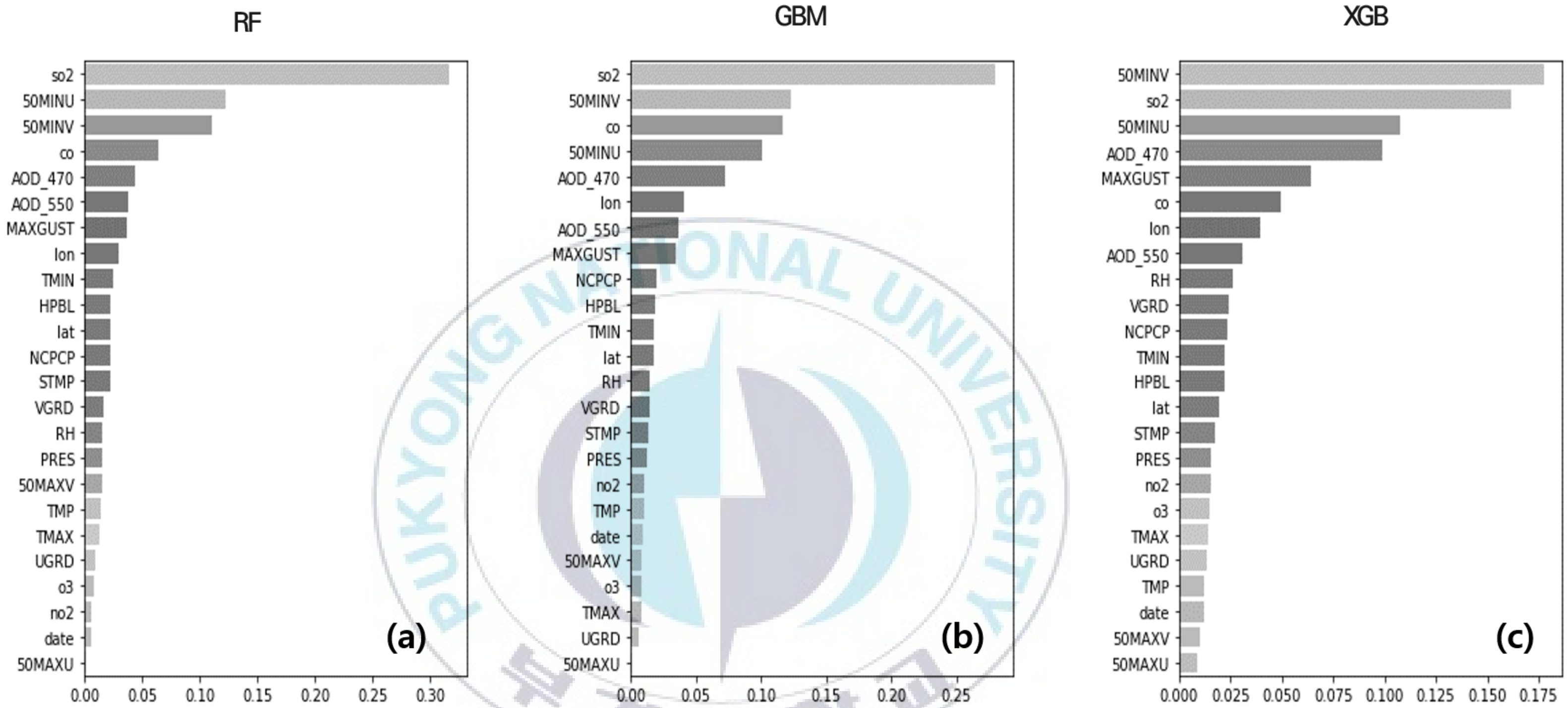

Fig. 6. Feature importance results (a) RF (b) GBM (c) XGB

4.3. Comparison with and without AOD

AOD is a key variable that can predict changes in $PM_{2.5}$ concentration, but one of the limitations of AOD is that there are a large number of missing values due to cloud cover, snow, or reflection of water (stafoggia et al., 2019). This study also uses AOD data and includes a large amount of missing values, so it is necessary to evaluate the performance of the model according to the presence or absence of AOD. The model was performed except for the AOD 470nm and 550 nm data in the existing research data set. RF results excluding AOD, R^2 = 0.836, Bias = -0.069, RMSE = 5.915μg/m³, MAE = 4.132μg/m³, which decreased the accuracy by about 4% compared to when the R^2 standard AOD was included. GBM had R^2 = 0.841, Bias = -1.308, RMSE = 5.844μg/m³, and MAE = 4.236μg/m³, which was about 5% less accurate than when the R^2-based AOD was included. XGB showed R^2 = 0.867, Bias = -0.653, RMSE = 5.238μg/m³, and MAE = 3.842μg/m³, showing the least difference among the three models with an accuracy decrease of about 2%. It can be seen that the performance of the three models decreased when AOD was excluded (Fig. 7; Table 4).

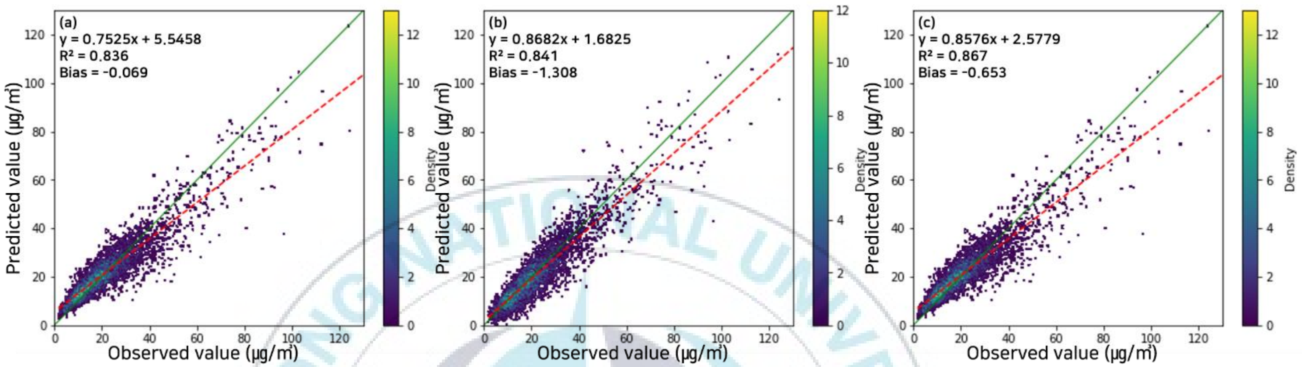

Fig. 7. Scatterplot of results without AOD

Table 4. Comparison of model performance results with and without AOD (R², Bias, RMSE, MAE)

	With AOD				Without AOD			
Model	**R²**	**Bias**	**RMSE (㎍/㎥)**	**MAE (㎍/㎥)**	**R²**	**Bias**	**RMSE (㎍/㎥)**	**MAE (㎍/㎥)**
RF	0.881	0.421	5.006	3.670	0.836	-0.069	5.915	4.132
GBM	0.888	-0.436	4.779	3.489	0.841	-1.308	5.844	4.236
XGB	0.890	-0.143	4.719	3.502	0.867	-0.653	5.238	3.842

4.4. Model Performance Analysis by Time Series

Seasonal analysis was performed for time series analysis of the model used in this study. Fig. 8 is a figure showing the actual observed value and the predicted value of each model as a box plot to check the distribution and outliers. The line drawn in the center of the box in the figure represents the median (50% percentile), the third quartile above the median, that is, 75% of the total data, and the first quartile represents 25% of the total data. A line drawn above the box (upper whisker) represents the maximum value, and a line drawn below the box (lower whisker) represents the minimum value. Points distributed above the maximum value represent outliers, and the white dot in the center of the box represents the average value. Looking at the overall distribution of the graph, the concentration in spring is the highest, and the concentration distribution in summer and fall is generally low. On the other hand, it seems that there are many outliers in spring, but it is judged that relatively many outliers are expressed due to the wide range of concentration distribution in spring.

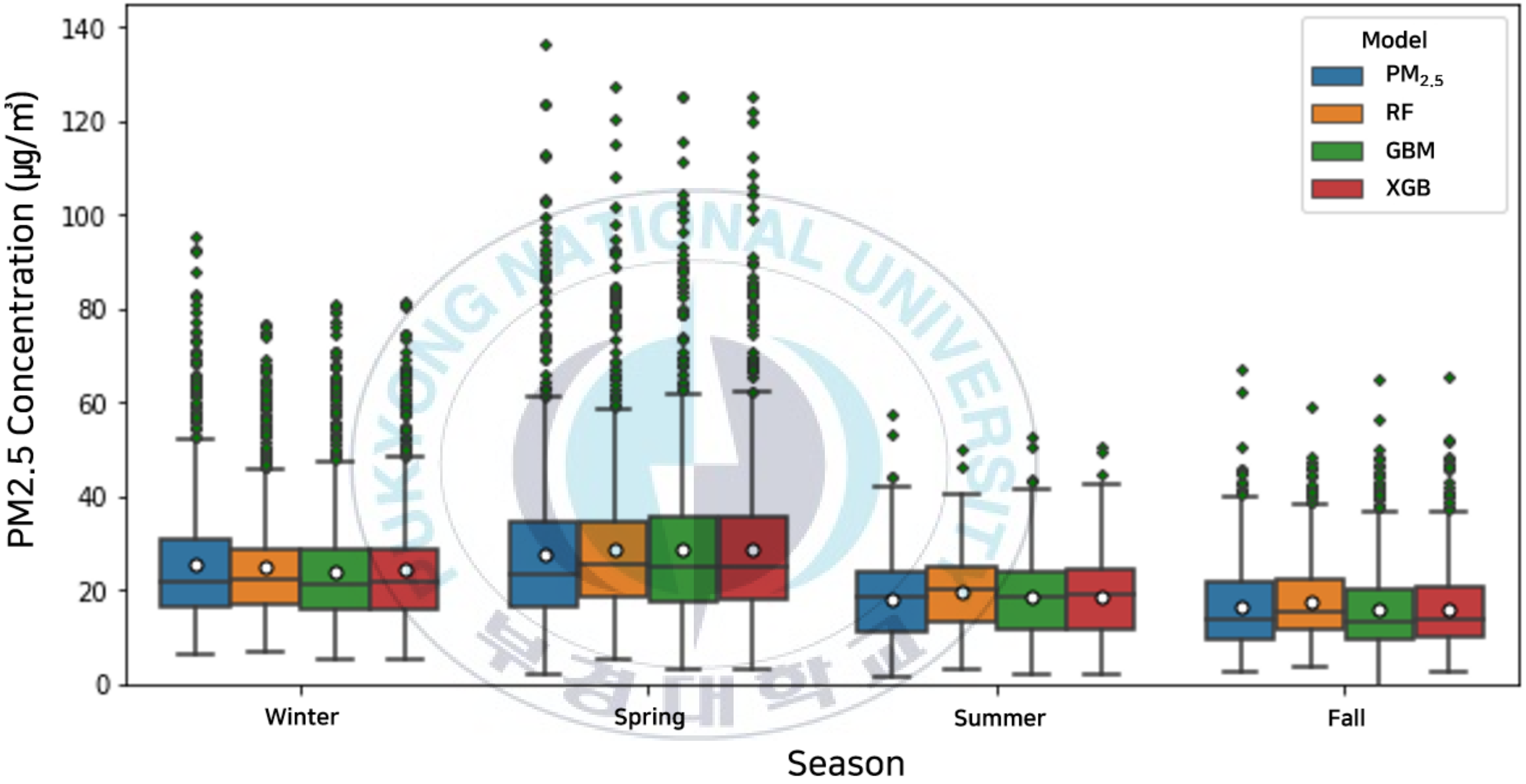

Fig. 8. Seasonal $PM_{2.5}$ concentration comparison results in boxplot

Fig. 9, Table 5 shows the seasonal performance results of the model. Looking at the graph, spring generally shows a high concentration, and the model performance of spring shows the best results when evaluated by R^2. The performance of the spring RF model was R^2 = 0.897, Bias = 0.745, RMSE = 5.872μg/m^3, and MAE = 4.194μg/m^3. GBM showed R^2 = 0.911, Bias = 0.598, RMSE = 5.350μg/m^3, and MAE = 3.795μg/m^3. In the case of XGB, R^2 = 0.913, Bias = 0.864, RMSE = 5.335μg/m^3, MAE = 3.830μg/m^3. In the case of spring, the performance of XGB was the best. In the case of the summer RF model, R^2 = 0.790, Bias = 1.415, RMSE = 4.321μg/m^3, and MAE = 3.426μg/m^3. GBM showed R^2 = 0.791, Bias = 0.101, RMSE = 4.102μg/m^3, and MAE = 3.104μg/m^3. XGB showed that R^2 = 0.795, Bias = 0.569, RMSE = 4.119μg/m^3, MAE = 3.108μg/m^3. Summer also showed the best performance of XGB. The model performance in fall is R^2 = 0.774, Bias = 0.996, RMSE = 4.365μg/m^3, MAE = 3.298μg/m^3 for RF, R^2 = 0.764, Bias = -0.590, RMSE = 4.429μg/m^3, MAE = 3.298μg/m^3 for GBM. It was found to be 3.301μg/m^3. XGB showed the results of R^2 = 0.766, Bias = -0.366, RMSE = 4.373μg/m^3, and MAE = 3.294μg/m^3. Fall showed high performance in RF unlike other seasons. The winter RF model showed an accuracy of R^2 = 0.876, Bias = -0.720, RMSE = 4.959μg/m^3, and MAE = 3.635μg/m^3. GBM showed R^2 = 0.886, Bias = -1.380, RMSE = 4.787μg/m^3, and MAE = 3.301μg/m^3. XGB showed R^2 = 0.889, Bias = -1.070, RMSE = 4.655μg/m^3, and MAE = 3.549μg/m^3. In the winter model, XGB's performance was the highest. In summary, it was found that the performance of XGB was the best in all seasons except fall.

Seasonal comparative analysis results, in spring, the peak shows a value of about 100 ~ 120μg/m³, and the range between the maximum and minimum values is very wide compared to other seasons. Therefore, it is judged that the values of RMSE and MAE, which are calculated as averages by obtaining errors, are relatively high. On the other hand, summer and fall showed low concentrations, and the performance of the model also showed low results. Among them, the prediction performance in fall showed values of $R^2 = 0.76 \sim 0.77$, showing the lowest model performance among the four seasons. However, it is judged that the RMSE and MAE values of summer and fall were low because the RMSE and MAE values of spring were high due to the generally low concentration distribution. In addition, among the three models, the performance of RF was lower than that of GBM and XGB, and the performance of GBM and XGB was similar.

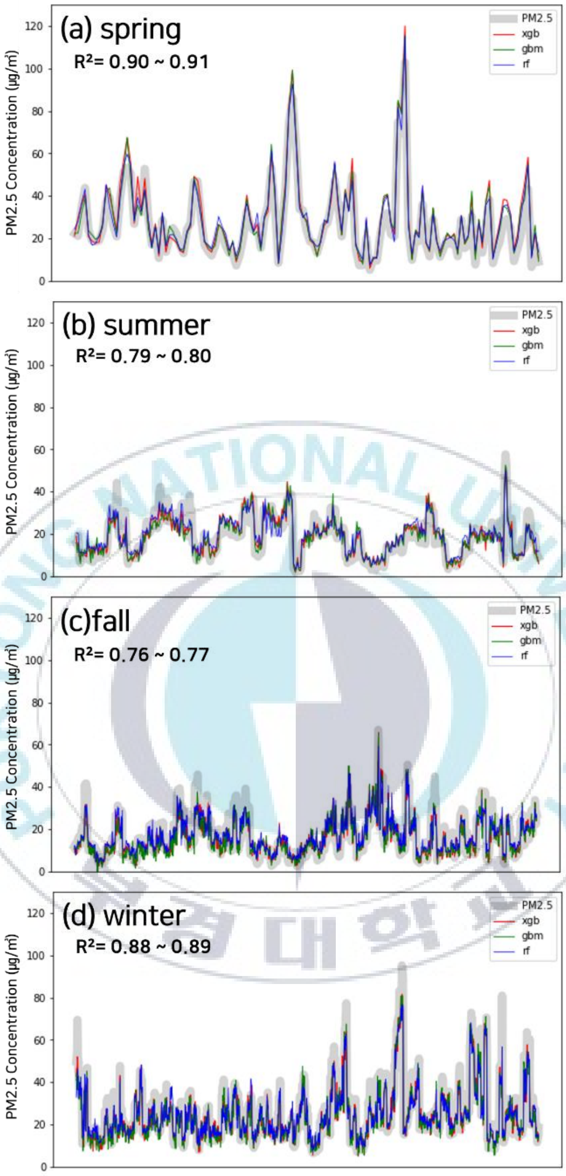

Fig. 9. Seasonal Model Results Trend Analysis

Table 5. Comparison of seasonal model performance (R^2, Bias, RMSE, MAE)

Season	**Model**	**R^2**	**Bias**	**RMSE (㎍/㎥)**	**MAE (㎍/㎥)**
Spring	**RF**	0.897	0.745	5.872	4.194
	GBM	0.911	0.598	5.350	3.795
	XGB	0.913	0.864	5.335	3.830
Summer	**RF**	0.790	1.415	4.321	3.426
	GBM	0.791	0.101	4.102	3.104
	XGB	0.795	0.569	4.119	3.108
Fall	**RF**	0.774	0.996	4.365	3.298
	GBM	0.764	-0.590	4.429	3.301
	XGB	0.766	-0.366	4.373	3.294
Winter	**RF**	0.876	-0.720	4.959	3.635
	GBM	0.886	-1.380	4.787	3.533
	XGB	0.889	-1.070	4.655	3.549

5. Discussion

In this study, RF (R^2 = 0.881, bias = 0.421, RMSE = 5.006$\mu g/m^3$, MAE = 3.67$\mu g/m^3$), GBM (R^2 = 0.888, bias = - 0.436, RMSE = 4.779$\mu g/m^3$, MAE = 3.489$\mu g/m^3$) and XGB (R^2 = 0.89, bias = -0.143, RMSE = 4.719$\mu g/m^3$, MAE = 3.502$\mu g/m^3$). All models produced results with prediction accuracy of R^2 = 0.88 ~ 0.89. However, as a result of training accuracy, GBM showed a tendency of overfitting, and in terms of bias, the bias value of the RF model showed a higher value compared to the two models, indicating a tendency of underfitting. It is believed that this is because max_depth, a basic element, was not optimized in the process of optimizing the RF model, and 5-fold CV was applied to RF and GBM due to the problem of model processing speed. Compared to the above two models, XGB has the advantage of fast execution time and strong against overfitting regulation. Therefore, it is judged that XGB is more suitable for $PM_{2.5}$ prediction model than RF and GBM. However, all three models seem to show high accuracy, and the study of Sihag et al., (2019) used models such as RF for $PM_{2.5}$ prediction. It was confirmed that the RF with the best performance among all models was R^2 = 0.691, MAE = 30.776$\mu g/m^3$, and RMSE = 44.695$\mu g/m^3$. Luo et al., (2020) built a GBM model for $PM_{2.5}$ prediction and reported that the prediction results were R^2 = 0.85, MAE = 3.56$\mu g/m^3$, and RMSE = 10.02$\mu g/m^3$. Peng et al., (2022) evaluated training, verification, and performance with an XGB model, etc., to predict $PM_{2.5}$ concentration in Hunan Province, central China, and the XGB prediction accuracy after parameter optimization process showed a result of R^2 of 0.761. Compared with the models in this study, all three models showed high prediction accuracy, and the prediction performance of the models is judged to be excellent. Among the results of this study, when comparing the performance of RF and GBM, GBM showed better performance but similar performance. Yazdi et al., (2020) used algorithms such as RF, GBM, and KNN to predict $PM_{2.5}$, and as a result, when 10-fold CV was applied, the RF model $R^2 = 0.83$ and the GBM $R^2 = 0.826$ reported similar performance. A study by Pu and Yoo (2021) predicted $PM_{2.5}$ using DNN, RF, GBM, etc., with $R^2 = 0.81$ for RF and $R^2 = 0.85$ for GBM, GBM's performance produced better prediction results than RF. Since the RF and GBM models in this study showed excellent performance even when 5-fold was applied, both the RF and GBM models built in this study are judged to have good performance. In addition, when comparing RF and XGB in this study, XGB showed significantly better results than RF. In the study of Joharestani et al. (2019), a similar method using RF and XGB was implemented to predict $PM_{2.5}$, and when RF and XGB were compared, XGB had $R^2 = 0.81$, MAE = 9.93 μg/m³, RMSE = 13.58 μg/m³, which was selected as the model with the best performance. As a result of comparing the GBM and XGB accuracy of this study, the two models showed similar performance, but the accuracy of XGB was higher with a slight difference. Park et al., (2021b) used GBM, XGB, and LightGBM, which are boosting-based ensemble models, to predict PM_{10}, and reported that the R^2 of GBM was 0.829 and the R^2 of XGB was 0.839. In the case of the three models, 5-fold or 10-fold CV was applied, and the performance is judged to be very good when compared to the results of similar studies. In addition, Just et al. (2018)'s study corrected AOD measurement errors using RF, GBM, and XGB for PM_{10} modeling, and XGB was selected as the optimal model. A study by Ribero and Coelho (2020) used RF, GBM, and XGB to predict agricultural prices and reported that XGB performed the best. It is impossible to compare the performance due to the different accuracy indicators, but compared to RF and GBM, XGB has the best performance, and it can be seen that XGB is a suitable model for the prediction model.

As a result of evaluating the importance of variables, SO2, CO, AOD, and wind-related factors (50MINU, 50MINV, MAXGUST) showed high contributions, and O3, NO2, Date, and wind-related factors (50MAXU, 50MAXV) showed low contributions. According to Yeo and Kim (2020), when the concentration of sulfur dioxide gas in Seoul increases, the average annual concentration of $PM_{2.5}$ more than doubled from the total $PM_{2.5}$ annual average concentration, confirming that the correlation between sulfur dioxide and $PM_{2.5}$ concentration is high. In the case of CO, as a result of analyzing the correlation between $PM_{2.5}$ and CO as Pearson's correlation in the study of Park and Ha (2008), it was reported that $PM_{2.5}$ and CO were significant correlations in the $p<0.01$ reliability interval and showed high correlation with a correlation coefficient of 0.520. On the other hand, wind-related factors show high contribution, but they also belong to sub-factors, so it is judged that discrimination is necessary. In addition, the results of factor importance in this study are based on impurities, which tend to be somewhat biased, so other methods need to be mixed and used. And since this result is the importance calculated from the train dataset, variables that are not actually important in the test dataset can be calculated as the most important variables in the learning process, so it is judged that improvement and further research are needed.

In order to confirm the effect of missing values of AOD on the model performance, the performance of the model with and without AOD was compared and analyzed. As a result, all three models showed a decrease in accuracy of 2 to 5%. A study by Chen et al. (2021) compared a model with and without AOD in the daily $PM_{2.5}$ concentration measurement in Guangdong Province, China by RF, and as a result, the R^2 of the model with AOD was 0.8 ~ 0.83, indicating an R^2 of 0.78 to 0.82 for the non-included model. Compared to this study, it was said that the accuracy was lowered, but there was no significant difference. Some studies have shown that the correlation between AOD and $PM_{2.5}$ shows a high correlation, but there are also studies that show a low correlation according to seasonal fluctuations (Guo et al., 2017; Li et al., 2015). As a result of comparison, AOD seems to have an effect on the performance of the model, and it is thought to show a significant correlation with $PM_{2.5}$ through additional analysis of variable importance.

Finally, the results of analyzing the model performance by season showed a trend of high concentration in spring and low concentration in summer and fall, and the performance of the model was also the highest in spring, and similarly low in summer and fall. It was confirmed that the value of AOD, a major factor, increased in spring and summer. In spring, yellow dust is affected through the northwest wind direction and the AOD value increases. In summer, pollutant accumulation due to relatively low wind speed and stable atmospheric condition, generation of secondary aerosol due to increased solar radiation, It is assumed that the AOD value increased because hygroscopic aerosols grew due to the increase in relative humidity (Lee et al., 2010; Lee et al., 2007). Therefore, it is estimated that the high AOD value was effectively reflected in spring when the actual observed concentration was high, and in the case of summer, the actual $PM_{2.5}$ concentration was low, but the high value of AOD was reflected and the accuracy was judged to be low. In addition, in the case of fall, the concentration trend is higher than that of summer, but the performance of the model is judged to be low because the predictive power for the peak value is low.

6. Conclusion

This study used three ensemble-based algorithms, RF, GBM, and XGB, to predict $PM_{2.5}$ concentration in Seoul. Also, as independent variables, 15 meteorological factors of LDAPS and 470nm of MODIS AOD. The 550 nm wavelength band and four chemical factors observed on the ground were used. As a result of prediction, all three models showed good performance with high prediction accuracy of R^2= 0.88 ~ 0.89. Among the three models, XGB's performance is judged to be the best, but RF and GBM also showed similar performances. In addition, as a result of comparative analysis of model performance according to the presence or absence of MODIS AOD, it was shown that the accuracy of the model improved as AOD was used as a factor. Therefore, it is judged that AOD should be an essential independent variable. In addition, in the results of time series analysis, it is determined that XGB, which has the best performance, is a suitable model for predicting $PM_{2.5}$ concentration. RF, GBM, and XGB used in this study are used in research in various fields, and recently, a number of studies using hybrid models combining algorithms have been conducted (Lin et al., 2022; Zhang et al., 2018). Therefore, additional studies that can increase the prediction accuracy should be conducted.

The $PM_{2.5}$ concentration prediction results of this study showed sufficiently good predictive power even at the present time when various algorithms are being developed, so it is expected to contribute to increasing the stability and accuracy of prediction for monitoring.

References

Azhari, A., Halim, N. D. A., Mohtar, A. A. A., Aiyub, K., Latif, M. T., & Ketzel, M. (2021). Evaluation and prediction of PM10 and PM2. 5 from road source emissions in Kuala Lumpur City Centre. Sustainability, 13(10), 5402.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Byon, J. Y., Hong, S. O., Park, Y. S., & Kim, Y. H. (2021). Evaluation of the Urban Heat Island Intensity in Seoul Predicted from KMA Local Analysis and Prediction System. The Journal of The Korean Earth Science Society, 42(2), 135-148.

Chae, S., Shin, J., Kwon, S., Lee, S., Kang, S., & Lee, D. (2021). PM10 and PM2. 5 real-time prediction models using an interpolated convolutional neural network. Scientific Reports, 11(1), 1-9.

Chen, G., Li, S., Knibbs, L. D., Hamm, N. A., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M. J., & Guo, Y. (2018). A machine learning method to estimate PM2. 5 concentrations across China with remote sensing, meteorological and land use information. Science of the Total Environment, 636, 52-60.

Chen, G., Li, Y., Zhou, Y., Shi, C., Guo, Y., & Liu, Y. (2021). The comparison of AOD-based and non-AOD prediction models for daily PM2. 5 estimation in Guangdong province, China with poor AOD coverage. Environmental Research, 195, 110735.

Chen, G., Wang, Y., Li, S., Cao, W., Ren, H., Knibbs, L. D., Abramson, M. J., & Guo, Y. (2018). Spatiotemporal patterns of PM10 concentrations over China during 2005–2016: A satellite-based estimation using the random forests approach. Environmental pollution, 242, 605-613.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, Choi, I., Lee, W., Eun, B., Heo, J., Chang, K., & Oh, J. (2022). A Study on Prediction of PM2.5 Concentration Using DNN. Journal of Environmental Impact Assessment, 31(2), 83-94.

Choi, S. I., An, J., & Jo, Y. M. (2018). Review of Analysis Principle of Fine Dust. Korean Industrial Chemistry News, 21(2), 16-23.

Choi, T. Y., Moon, H. G., Kang, D. I., & Cha, J. G. (2018). Analysis of the Seasonal Concentration Differences of Particulate Matter According to Land Cover of Seoul – Focusing on Forest and Urbanized Area –. Journal of Environmental Impact Assessment, 27(6), 635-646.

Danesh Yazdi, M., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., Lyapustin, A., Katsouyanni, K., & Schwartz, J. (2020). Predicting fine particulate matter (PM2. 5) in the greater London area: an ensemble approach using machine learning methods. Remote sensing, 12(6), 914.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Gao, X., & Li, W. (2021). A graph-based LSTM model for PM2. 5 forecasting. Atmospheric Pollution Research, 12(9), 101150.

Guo, J., Xia, F., Zhang, Y., Liu, H., Li, J., Lou, M., He, J., Yan, Y., Wang, F., & Min, M. (2017). Impact of diurnal variability and meteorological factors on the PM2. 5-AOD relationship: Implications for PM2. 5 remote sensing. Environmental pollution, 221, 94-104.

Ha, J. E., Shin, H. C., & Lee, Z. K. (2017). Korean Text Classification Using Randomforest and XGBoost Focusing on Seoul Metropolitan Civil Complaint Data. The Korea Journal of BigData, 2(2), 95-104.

Han, S. H., & Kim, Y. P. (2015). Long-term Trends of the Concentrations of Mass and Chemical Composition in PM2.5 over Seoul. Journal of Korean Society for Atmospheric Environment, 31(2), 143-156.

Hwang, I. C. (2018). Particulate Matter Management Policy of Seoul: Achievements and Limitations. The Korea Association for Policy Studies, 27(2), 27-51.

Hwang, K. l., Han, B. H., Kwark, J. I., & Park, S. C. (2018). A Study on Decreasing Effects of Ultra-fine Particles (PM2.5) by Structures in a Roadside Buffer Green - A Buffer Green in Songpa-gu, Seoul -. Journal of the Korean Institute of Landscape Architecture, 46(4), 61-75.

Hwang, S., Kim, T. H., Kim, M., & Choi, J. (2022). A Study on the Time Series Characteristics of High-concentration Fine Dust Generation by Local Indicator of Temporal Burstiness. Journal of the Korean Geographical Society, 57(1), 97-108.

Jang, D., & Park, M. (2020). A Study on the Art Price Prediction Model Using the Random Forests. Journal of Applied Reliability, 20(1), 34-42.

Jang, Y., Jang, I., & Choe, Y. (2020). Prediction of Soil Moisture with Open Source Weather Data and Machine Learning Algorithms. Korean Journal of Agricultural and Forest Meteorology, 22(1), 1-12.

Jeong, J. C. (2017). Spatial Information Application Case for Appropriate Location Assessment of PM10 Observation Network in Seoul City. Journal of Cadastre & Land InformatiX, 47(2), 175-184.

Jeong, S. H., & Jin, C. (2020). A Study on the Office Rent Estimation by the Machine Learning Methods -Focusing on the Use of Random Forests, Artificial Neural Networks, Support Vector Machines. Journal of the Korea Real Estate Analysts Association, 26(2), 23-53.

Just, A. C., Carli, M. M. D., Shtein, A., Dorman, M., Lyapustin, A., & Kloog, I. (2018). Correcting measurement error in satellite aerosol optical depth with machine learning for modeling PM2. 5 in the Northeastern USA. Remote sensing, 10(5), 803.

Kim, H. (2020). The Prediction of PM2.5 in Seoul through XGBoost Ensemble. Journal of The Korean Data Analysis Society, 22(4), 1661-1671.

Kim, H. Y., & Moon, T. H. (2021). Machine learning-based Fine Dust Prediction Model using Meteorological data and Fine Dust data. Journal of the Korean Association of Geographic Information Studies, 24(1), 92-111.

Kim, K., Lee, D., Lee, K. Y., Lee, K. H., & Noh, Y. (2016). Estimation of surface-level PM 2.5 concentration based on MODIS aerosol optical depth over Jeju, Korea. Korean Journal of Remote Sensing, 32(5), 413-421.

Kim, M., & Park, M. (2019). An Analysis of the Characteristics of College Students According to First-Time Participation in Private Tutoring Using A Random Fores. CNU Journal of Educational Studies, 40(1), 1-33.

Kim, T. Y. (2022). Design of Fine Dust Monitoring System based on the Internet of Things. The Journal of Korea Institute of Information, Electronics, and Communication Technology, 15(1), 14-26.

Kim, Y., & Chang, K. (2021). Comparison and analysis of prediction performance of fine particulate matter(PM2.5) based on deep learning algorithm. Journal of Convergence for Information Technology, 11(3), 7-13.

Kim, Y., & Yeo, M. (2013). The Trend of the Concentrations of the Criteria Pollutants over Seoul. Journal of Korean Society for Atmospheric Environment, 29(4), 369-377.

Lee, D., & Lee, S. (2020). Hourly Prediction of Particulate Matter (PM2.5) Concentration Using Time Series Data and Random Forest. Korea Information Processing Society Transactions on Software and Data Engineering, 9(4), 129-136.

Lee, K. l., Ryu, J., Jeon, S. W., Jung, H. C., & Kang, J. Y. (2017). Analysis of the Effect of Heat Island on the Administrative District Unit in Seoul Using LANDSAT Image. Korean Journal of Remote Sensing, 33(5), 821-834.

Lee, S., Yoon, S. C., & Ghim, Y. S. (2007). Comparison of Aerosol Optical Properties from Different Aerosol Chemical Compositions in Seoul and Gosan. Proceedings of the Korea Air Pollution Research Association Conference, 116-119.

Lee, S. B., Kang, C. H., Jung, D. S., Ko, H. J., Kim, H. B., Oh, Y. S., & Kang, H. L. (2010). Composition and pollution characteristics of TSP, PM2.5 atmospheric aerosols at Gosan site, Jeju Island. ANALYTICAL SCIENCE & TECHNOLOGY, 23(4), 372-382.

Li, J., Carlson, B. E., & Lacis, A. A. (2015). How well do satellite AOD observations represent the spatial and temporal variability of PM2. 5 concentration for the United States? Atmospheric Environment, 102, 260-273.

Lim, J. M. (2019). An Estimation Model of Fine Dust Concentration Using Meteorological Environment Data and Machine Learning. Journal of Information Technology Services, 18(1), 173-186.

Lin, L., Chen, C. Y., Yang, H. Y., Xu, Z., & Fang, S. H. (2020). Dynamic system approach for improved PM 2.5 prediction in Taiwan. IEEE Access, 8, 210910-210921.

Lin, L., Liang, Y., Liu, L., Zhang, Y., Xie, D., Yin, F., & Ashraf, T. (2022). Estimating PM2. 5 Concentrations Using the Machine Learning RF-XGBoost Model in Guanzhong Urban Agglomeration, China. Remote sensing, 14(20), 5239.

Luo, Z., Huang, F., & Liu, H. (2020). PM2. 5 concentration estimation using convolutional neural network and gradient boosting machine. Journal of Environmental Sciences, 98, 85-93.

Park, D. U., & Ha, K. C. (2008). Characteristics of PM10, PM2. 5, CO2 and CO monitored in interiors and platforms of subway train in Seoul, Korea. Environment International, 34(5), 629-634.

Park, J. K., Choi, Y. J., & Jung, W. S. (2017). Understanding on Regional Characteristics of Particular Matter in Seoul. Journal of Environmental Science International, 26(1), 55-65.

Park, J. Y., Kwon, T. Y., & Lee, J. Y. (2017). Estimation of surface visibility using MODIS AOD. Korean Journal of Remote Sensing, 33(2), 171-187.

Park, S., Kim, M., & Im, J. (2021). Estimation of Ground-level PM10 and PM2.5 Concentrations Using Boosting-based Machine Learning from Satellite and Numerical Weather Prediction Data. Korean Journal of Remote Sensing, 37(2), 321-335.

Park, S., & Shin, H. (2017). Analysis of the Factors Influencing PM2.5 in Korea : Focusing on Seasonal Factors. Journal of Environmental Policy and Administration, 25(1), 227-248.

Park, S., Son, S., Bae, J., Lee, D., Kim, J-J., & Kim, J-S. (2021). Robust Spatiotemporal Estimation of PM Concentrations Using Boosting-Based Ensemble Models. Sustainability, 13(24), 13782.

Park, S. Y., Baek, J., Park, S., & Hur, J. (2022). Implementation of a Short-term Wind Power Output Forecasting Model based on Gradient Boosting Machine(GBM) Algorithms. Proceedings of the korean institute of electrical Engineers Conference, 305-306.

Peng, J., Han, H., Yi, Y., Huang, H., & L. Xie. (2022). Machine learning and deep learning modeling and simulation for predicting PM2. 5 concentrations. Chemosphere, 308, 136353.

Pu, Q., & Yoo, E. Y. (2021). Ground PM2. 5 prediction using imputed MAIAC AOD with uncertainty quantification. Environmental pollution, 274, 116574.

Ribeiro, M. H. D. M., & Coelho, L. d. S. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Applied Soft Computing, 86, 105837.

Seo, Y. M., & Yom, J. H. (2019). Comparison of LSTM-based Fine Dust Concentration Prediction Method using Meteorology Data. Korea Society of Surveying, Geodesy, Photogrammetry, and Cartography, 117-120.

Shin, D. H., & Kim, Y. M. (2015). The Utilization of Big Data's Disaster Management in Korea. The Journal of the Korea Contents Association, 15(2), 377-392.

Sihag, P., Kumar, V., Afghan, F. R., Pandhiani, S. M., & Keshavarzi, A. (2019). Predictive modeling of PM2. 5 using soft computing techniques: case study—Faridabad, Haryana, India. Air Quality, Atmosphere & Health, 12(12), 1511-1520.

Son, K., Bae, M., You, S., Kim, E., Kang, Y.-H., Bae, C., Kim, Y., Kim, H.-C., Kim, B.-U., & Kim, S. (2020). Meteorological and Emission Influences on PM2.5 Concentration in South Korea during the Seasonal Management: A Case of December 2019 to March 2020. Journal of Korean Society for Atmospheric Environment, 36(4), 442-463.

Son, S., & Kim, J. (2021). Vulnerability Assessment for Fine Particulate Matter (PM2.5) in the Schools of the Seoul Metropolitan Area, Korea: Part I – Predicting Daily PM2.5 Concentrations. Korean Journal of Remote Sensing, 37(6), 1881-1890.

Song, B. G., & Park, K. H. (2022). Analysis of PM2.5 Pattern Considering Land Use Types and Meteorological Factors - Focused on Changwon National Industrial Complex -. Journal of the Korean Association of Geographic Information Studies, 25(2), 1-17.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., Hoogh, K. D., De'Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., & Renzi, M. (2019). Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environment International, 124, 170-179.

Sung, S. H., Kim, S., & Ryu, M. H. (2020). A Comparative Study on the Performance of Machine Learning Models for the Prediction of Fine Dust: ocusing on Domestic and verseas Factors. Innovation studies, 15(4), 339-357.

Van Donkelaar, A., Martin, R. V., & Park, R. J. (2006). Estimating ground-level PM2. 5 using aerosol optical depth determined from satellite remote sensing. Journal of Geophysical Research: Atmospheres, 111(D21).

Xie, Y., Wang, Y., Zhang, K., Dong, W., Lv, B., & Bai, Y. (2015). Daily estimation of ground-level PM2. 5 concentrations over Beijing using 3 km resolution MODIS AOD. Environmental science & technology, 49(20), 12280-12288.

Yeo, M. & Kim. M. (2020). Long-term Trend of Sulfur Dioxide Concentration by District in Korea. Journal of Korean Society for Atmospheric Environment, 36(6), 742-756.

Yoo, H. G., Hong, J. W., Hong, J., Sung, S., Yoon, E. J., Park, J. H., & Lee, J. H. (2020). Impact of Meteorological Conditions on the PM2.5 and PM10 concentrations in Seoul. Journal of Climate Change Research, 11(5-2), 521-528.

Yu, M., Lee, Y., & Yi, J. (2016). Flood inflow forecasting on HantanRiver reservoir by using forecasted rainfall. Korea Water Resources Association, 49(4), 327-333.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere, 10(7), 373.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and XGboost. IEEE Access, 6, 21020-21031.

Zhang, H., & Kondragunta, S. (2021). Daily and hourly surface PM2. 5 estimation from satellite AOD. Earth and Space Science, 8(3), e2020EA001599.