Korean Journal of Remote Sensing, Vol.37, No.2, 2021, pp.321~335
https://doi.org/10.7780/kjrs.2021.37.2.11

ISSN 1225-6161 (Print)
ISSN 2287-9307 (Online)

Article

부스팅 기반 기계학습기법을 이용한 지상 미세먼지 농도 산출

박서희[1] · 김미애[2] · 임정호[3]†

Estimation of Ground-level PM_{10} and $PM_{2.5}$ Concentrations Using Boosting-based Machine Learning from Satellite and Numerical Weather Prediction Data

Seohui Park[1] · Miae Kim[2] · Jungho Im[3]†

Abstract: Particulate matter (PM_{10} and $PM_{2.5}$ with a diameter less than 10 and 2.5 μm, respectively) can be absorbed by the human body and adversely affect human health. Although most of the PM monitoring are based on ground-based observations, they are limited to point-based measurement sites, which leads to uncertainty in PM estimation for regions without observation sites. It is possible to overcome their spatial limitation by using satellite data. In this study, we developed machine learning-based retrieval algorithm for ground-level PM_{10} and $PM_{2.5}$ concentrations using aerosol parameters from Geostationary Ocean Color Imager (GOCI) satellite and various meteorological parameters from a numerical weather prediction model during January to December of 2019. Gradient Boosted Regression Trees (GBRT) and Light Gradient Boosting Machine (LightGBM) were used to estimate PM concentrations. The model performances were examined for two types of feature sets–all input parameters (Feature set 1) and a subset of input parameters without meteorological and land-cover parameters (Feature set 2). Both models showed higher accuracy (about 10 % higher in R^2) by using the Feature set 1 than the Feature set 2. The GBRT model using Feature set 1 was chosen as the final model for further analysis (PM_{10}: $R^2 = 0.82$, nRMSE = 34.9 %, $PM_{2.5}$: $R^2 = 0.75$, nRMSE = 35.6 %). The spatial distribution of the seasonal and annual-averaged PM concentrations was similar with *in-situ* observations, except for the northeastern part of China with bright surface reflectance. Their spatial distribution and seasonal changes were well matched with *in-situ* measurements.

Key Words: Particulate Matter, PM_{10}, $PM_{2.5}$, Machine learning, AOD, GOCI

Received April 2, 2021; Revised April 16, 2021; Accepted April 23, 2021; Published online April 26, 2021

[1] 울산과학기술원 도시환경공학부 석박통합과정생 (Combined Master and PhD Student, Department of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology)

[2] 울산과학기술원 도시환경공학부 연구조교수 (Research Assistance Professor, Department of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology)

[3] 울산과학기술원 도시환경공학부 정교수 (Professor, Department of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology)

† Corresponding Author: Jungho Im (ersgis@unist.ac.kr)

요약: 미세먼지 (PM_{10}) 및 초미세먼지 ($PM_{2.5}$)는 인체에 흡수 가능하여 호흡기 질환 및 심장 질환과 같이 인체 건강에 악영향을 미치며, 심각할 경우 조기 사망에 영향을 줄 수 있다. 전 세계적으로 현장관측기반의 모니터링을 수행하고 있지만 미 관측지역에 대한 대기질 분포의 공간적인 한계점이 존재하여 보다 광범위한 지역에 대한 지속적이고 정확한 모니터링이 필요한 상황이다. 위성기반 에어로졸 정보를 사용함으로써 이러한 현장 관측자료의 한계점을 극복할 수 있다. 따라서 본 연구에서는 다양한 위성 및 모델자료를 활용하여 2019년도에 대해 한 시간 단위의 지상 PM_{10} 및 $PM_{2.5}$ 농도를 추정하였다. GOCI 위성의 관측영역을 포함하는 동아시아 지역에 대해 트리 기반 앙상블 방법을 사용하는 Boosting 기법인 GBRTs (Gradient Boosted Regression Trees)와 LightGBM (Light Gradient Boosting Machine)을 활용하여 모델을 구축하였다. 또한, 기상변수 및 토지피복변수의 사용유무에 따른 모델의 성능을 비교하기 위해 두 가지 festure set으로 나누어 테스트하였다. 두 기법 모두 주요 변수인 AOD (Aerosol Optical Depth), SSA (Single Scattering Albedo), DEM (Digital Eelevation Model), DOY (Day of Year), HOD (Hour of Day)와 기상변수 및 토지피복변수를 함께 사용한 Feature set 1을 사용하였을 때 높은 정확도를 보였다. Feature set 1에 대해 GBRT 모델이 LightGBM에 비해서약 10%의 정확도 향상을 보였다. 가장 정확도가 높았던 기상 및 지표면 변수를 포함한 Feature set1을 사용한 GBRT기반 모델을 최종모델로 선정하였으며 (PM_{10}: R^2 = 0.82 nRMSE = 34.9%, $PM_{2.5}$: R^2 = 0.75 nRMSE = 35.6%), 계절별 및 연평균 PM_{10} 및 $PM_{2.5}$ 농도에 대한 공간적인 분포를 확인해본 결과, 현장관측자료와 비슷한 공간 분포를 보였으며, 국가별 농도 분포와 계절에 따른 시계열 농도 패턴을 잘 모의하였다.

1. 서론

대기오염물질은 엽면에 부착되어 식물 생장 감소 및 고사를 유발하며, 인체에 장시간 노출되면 호흡기 및 심장 질환 등 인체 건강에 악영향을 미친다. 특히, 크기가 작은 미세먼지(입자크기가 10 μm 이하인 Particulate Matter (PM_{10}))와 초미세먼지(입자크기가 2.5 μm 이하인 Particulate Matter ($PM_{2.5}$))는 인체에 흡수되어 다양한 질병을 유발하는 원인 중 하나로, 심각할 경우 조기 사망에 이를 수 있다(Pope III *et al*., 2009; Jerrett *et al*., 2017). 대기오염물질에 대한 위험성이 대두되면서 정확한 대기오염물질 농도에 대한 모니터링이 필요한 상황이다. 전 세계적으로 현장관측기반의 대기질 모니터링을 수행하고 있지만 미관측 지역에 대한 대기질 분포의 공간적인 불확실성이 여전히 존재한다. 위성은 넓은 지역에 대한 관측값을 제공함으로써 이러한 한계점을 극복할 수 있다. 위성으로부터 제공되는 대기중 에어로졸 광학 깊이(AOD; Aerosol Optical Depth)는 대기중 에어로졸에 의한 복사 에너지 값의 감소 정도를 나타내는 값으로 기존의 많은 연구에서 지상 대기오염물질 농도를 추정하기 위해 활용하고 있다(Park *et al*., 2019; She *et al*., 2020; Lee *et al*., 2020; Guo *et al*., 2021).

지상 PM_{10} 및 $PM_{2.5}$ 농도 추정 연구는 수년간 변화하고 있으며, 보다 정확한 농도 추정을 위해 다양한 입력변수를 사용한 여러 기법들을 적용하고 있다. 초기 연구에서는 주로 위성 기반 AOD를 기반한 PM 농도를 추정하였으며, 단순 통계기법위주의 선형모델을 적용하여 지상농도를 추정하였다(Liu *et al*., 2005; Gupta and Christopher, 2009a). 최근에는 위성 기반 AOD와 기상학적인 변수 및 공간적인 변수를 함께 사용하여 PM 농도를 추정하는 방식으로 변화하고 있다. 많은 기존연구에서는 위성기반 AOD만을 사용했을 때보다 기상 변수 및 공간적인 변수를 추가적으로 사용하였을 때 PM_{10} 및 $PM_{2.5}$ 농도 추정에 더 높은 정확도를 보였다(Kloog *et al*., 2015; Ghotbi *et al*., 2016; Lee *et al*., 2016; Soni *et al*., 2018). Kloog *et al*. (2015)이 개발한 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델의 경우, MAIAC AOD만 단독으로 사용하였을 때보다 AOD와 기상변수, 토지피복변수를 모두 사용하였을 때 R^2가 0.77에서 0.84로 정확도가 증가하였다.

사용된 모델의 기법적인 변화도 지상 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델의 정확도 향상에 많은 기여를 했다(Gupta and Christopher, 2009a; Gupta and Christopher, 2009b). 초기연구에서 많이 사용되었던 단일 및 다중선형회귀분석의 경우 AOD와 지상 PM_{10} 및 $PM_{2.5}$ 농도 사이의 비선형성이 고려되지 않아 정확도 향상에 한계점이 존재하였다. 최근 활발히 사용되는 기계학습 및 딥러닝 기법의 경우 이러한 비 선형성을 고려할 수 있어 모델의 정확도 향상을 위한 연구사례들이 늘고 있다(Park *et al*.,

2019; Gui *et al.*, 2020; Stirnberg *et al.*, 2021; Zhang *et al.*, 2021). 중국 전역에 대한 $PM_{2.5}$ 농도 추정 연구의 경우 Boosting 기법을 활용한 연구(Gui *et al.*, 2020; Zhang *et al.*, 2021)가 다중선형회귀분석을 사용한 기존 연구(Ma *et al.*, 2014; Geng *et al.*, 2015)에 비해 높은 정확도를 보였다.

향상된 기법들을 적용한 기존연구에서도 MODIS (Moderate Resolution Imaging Spectroradiometer)와 같은 극궤도 위성기반 AOD를 주로 사용하고 있어, 한 시간 단위의 현장자료를 일일 평균하여 이용하고 있다. 그러나 지상 PM_{10} 및 $PM_{2.5}$ 농도는 시간에 따라 변동성이 크기 때문에 보다 정확한 모니터링을 수행하기 위해서는 시간 단위의 추정이 필수적이다. 에어로졸 정보를 하루에 한 번만 제공하는 극궤도위성과 달리 정지궤도위성은 하루 내 시간별 정보 제공이 가능하다. 본 연구에서는 정지궤도 위성인 GOCI기반 AOD를 사용하여 한 시간 간격의 PM 농도 추정을 수행하였으며, 다양한 위성기반 및 모델 기반 자료와 현장관측자료를 활용하여 기계학습 기반의 지상 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델 개발 및 농도의 공간적인 분포 표출을 목표로 하였다. 최신 Boosting 기반의 대표적인 기법인 GBRTs와 GBRTs 모델의 단점을 보완한 LightGBM를 사용하여 지상 PM_{10} 및 $PM_{2.5}$ 농도를 추정하였다.

2. 연구자료

본 연구의 연구 지역은 GOCI (Geostationary Ocean Color Imager) 관측 지역으로 한반도를 중심으로 한 중국 동부 지역과 일본을 포함한 육상 지역 (Fig. 1)이며, 연구 기간은 2019년 1월부터 12월까지이다. 이 지역은 지난 수 십년간 급속한 경제발전과 산업화로 인해 인구밀도가 높은 대도시가 집중되어 있어 심각한 대기오염 문제를 겪고 있다. 또한, 연구지역 주변에 고비 사막 등 자연적인 에어로졸 발원지들로부터 불어오는 황사의

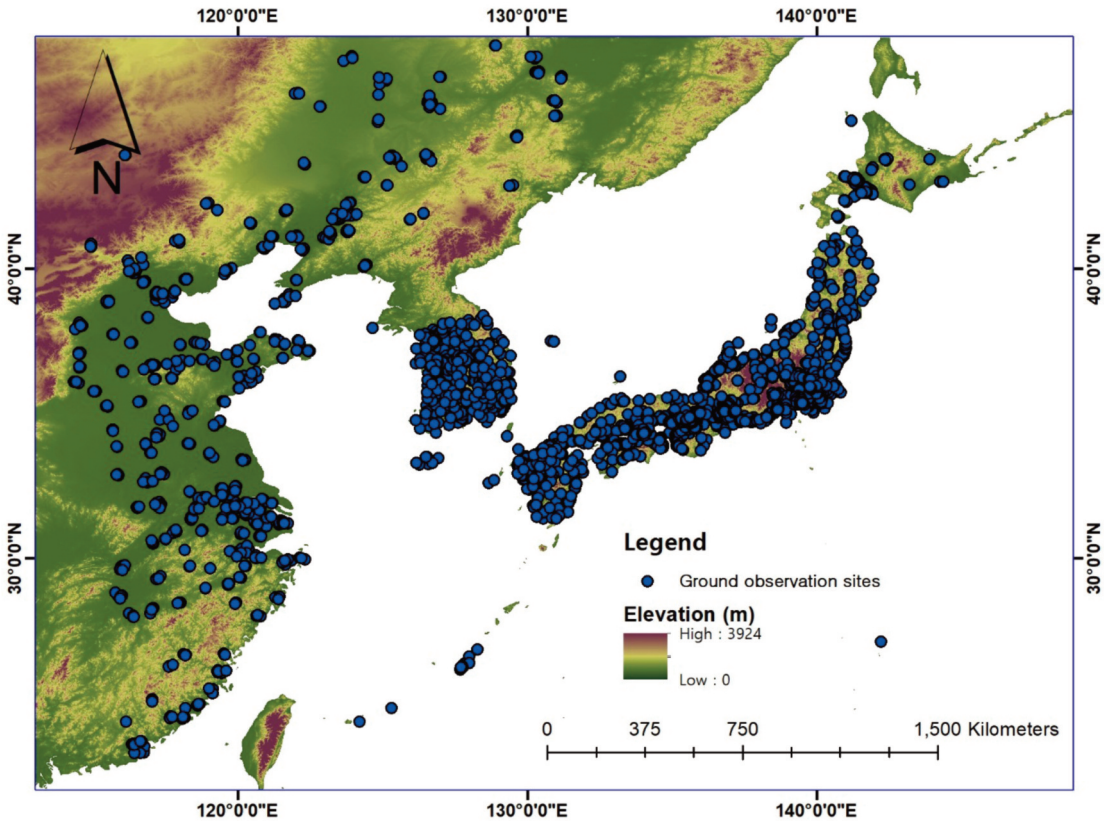

Fig. 1. Study area with ground particulate matter (PM) observation sites (blue dots). The background image is elevation for study area.

영향을 직접적으로 받고 있다. 본 연구에서는 다양한 위성-GOCI, MODIS, SRTM (Shuttle Radar Topography Mission), GPM (Global Precipitation Measurement)-을 기반한 자료와 UM-RDAPS (Unified Model-Regional Data Assimilation and Prediction System)의 기상자료 및 기타 보조자료를 사용하여 기계학습기반 지상 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델을 개발하였다.

1) PM 지상농도 현장관측 자료

본 연구에서는 남한, 중국 동부, 일본 지역을 포함하는 연구 지역내에 존재하는 총 3278개의 관측소에서 측정된 지상 PM_{10} 및 $PM_{2.5}$ 현장관측자료를 타깃변수로 활용하였다. 남한의 경우 AirKorea 웹사이트 (https://www.airkorea.or.kr/index)에서 제공하는 확정자료를 다운받아 사용하였으며, 중국의 경우 BMEMC (Beijing Municipal Environmental Protection Monitoring Center)에서 제공하는 자료를 API (Application Programming Interface)를 통해 다운받아 사용하였다. 일본도 중국과 마찬가지로 API를 통해 NIES (National Institute for Environmental Studies, Japan)에서 제공하는 자료를 다운받아 사용하였다. 현장관측자료는 서로 다른 로컬 시간을 가지고 있기 때문에 UTC 시간에 맞추었다. Fig. 1에는 연구지역과 함께 PM 지상농도 현장관측소 위치를 나타내고 있다(Fig. 1).

2) 기계학습 모델 입력 변수

Table 1는 본 연구에 사용된 모델 입력 변수를 보여준다. 본 연구에서는 GOCI 기반 AOD와 SSA (Single Scattering Albedo, 단일 산란 알베도) 산출물 자료를 입력 변수로 사용하였다. AOD는 대기 중 에어로졸에 의한 태양 복사의 감소를 수치화한 값이며(Lim *et al.*, 2016; Jin, 2018), SSA는 대기 중 에어로졸이 태양빛을 반사시키는 정도를 나타내는 값이다. 일반적으로 AOD가 낮을수록 낮은 SSA 값을 보이며, 에어로졸의 산란 및 흡수 특성을 구별하는데 사용된다. 기존의 지상 PM 농도 추정 연구에서는 이러한 에어로졸 정보를 주요 입력 변수로 활용해왔다(Soni *et al.*, 2018; Park *et al.*, 2019).

지상 PM 농도는 기상 등의 다양한 외부 요인에 의해서 영향을 받는다. 따라서 본 연구에서는 우리나라에 특화된 수치 모델인 RDAPS기반 기온(Air Temperature, Temp), 강수량(Precipitation, Precip), 상대습도(Relative Humidity, RH), 경계층고도(Planetary Boundary Layer Height, PBLH), 최대 풍속(Maximum wind speed, MaxWS)을 입력 변수로 사용하였다. 시간에 따른 에어로졸 이류의 영향을 확인하기 위해서 1, 3, 5, 7일 동안 최대 풍속을 누적하여 만든 stack1_maxWS, stack3_maxWS, stack5_maxWS, stack7_maxWS을 추가적으로 고려해주었다. 또한 강수량에 의한 대기 중 에어로졸 제거를 고려하기 위해서 GPM의 강수량 자료를 입력 변수로 이용하였다.

Table 1. Two input feature sets used in the machine learning models in this study

	Input parameters	Source
Feature set 1	Aerosol Optical Depth	GOCI
	Single Scatttering Albedo	
	Normalized Difference Vegetation Index	MODIS
	Urban Ratio	
	Digital Elevation Model	SRTM
	Temperature	UM-RDAPS
	Precipitation	
	Relative Humidity	
	Planetary Boundary Layer Height	
	Maximum Wind Speed	
	1, 3, 5, 7-day Stacked Maximum Wind Speed	
	Population Density	Ancillary data
	Road Density	
	Day Of Year	
	Hour Of Day	
Feature set 2	Aerosol Optical Depth	GOCI
	Single Scatttering Albedo	
	Digital Elevation Model	SRTM
	Day Of Year	Ancillary data
	Hour Of Day	

에어로졸은 지표면으로부터의 영향을 받기 때문에 다양한 표면 정보를 입력 변수로 사용하였다. 식생의 활력 정도를 나타내는 지수인 NDVI (Normalized Difference Vegetation Index)를 사용한다. 또한, MODIS의 500 m 공간 해상도의 연간 토지 피복도 자료(MCD12Q1)로부터 중심픽셀과 인접한 픽셀(13×13 윈도우) 내의 도시 지역의 비율을 계산하여 입력 변수로 활용하였다. 추가적으로 SEDAC (Socioeconomic Data and Applications Center)

에서 제공하는 인구 밀도 자료와 GRIP (Global Roads Inventory Project)에서 제공하는 도로밀도 자료를 이용하였다. 토지피복자료(도시비율, 인구 밀도, 도로밀도)는 에어로졸의 발생원으로써 지상 PM 농도와 큰 상관관계가 있어 지상 PM 농도를 추정하는데 유용한 정보가 될 수 있다 (Kloog *et al*., 2015).

추가 보조 변수로는 DOY (Day Of the Year)와 HOD (Hour Of Day) 등의 시간 변수 및 DEM (Digital Elevation Model)의 고도 정보를 추가 변수로 이용하였다. 본 연구에서는 DEM을 제외한 토지피복변수와 기상변수의 사용 유무에 따라 모델 성능 차이를 확인하고자 Table 1에서와 같이 두 가지의 Feature set을 구분하여 모델링을 수행하였다.

3. 연구방법

1) 기계학습 모델 훈련자료 구축 방법

기계학습 모델의 훈련자료를 구축하기 위해서 다양한 공간 해상도를 가지는 입력 변수들을 시공간적으로 서로 매칭시켰다. 서로 다른 공간 해상도를 가지는 위성 및 모델기반 입력 자료들을 GOCI 산출물의 공간 해상도(6 km × 6 km) 에 맞춰 bilinear 보간법을 이용해서 리샘플링 하였다. 타깃으로 이용되는 지상 PM_{10} 및 $PM_{2.5}$ 농도 현장 관측자료는 GOCI 관측 시간인 09:00-16:00 KST (Korea Standard Time) (하루 8시간)동안의 자료를 사용하였다. 또한, 하나의 픽셀에 2개 이상의 관측소가 존재할 경우, 거리가중평균을 통해 픽셀 당 하나의 타겟 값이 나오도록 계산해주었다. 현장 관측자료의 경우 고농도 샘플이 부족하고 저농도 샘플이 많은데 이러한 데이터셋의 불균형은 모델의 과소추정을 야기할 수 있다. 따라서 본 연구에서는 이를 방지하기 위해 고농도 샘플에 대한 오버샘플링(oversampling)과 저농도 샘플에 대한 서브샘플링(subsampling) 방법을 적용하여 데이터셋의 균형을 맞춰주었다.

오버샘플링의 경우 관측소 주변으로 농도 값이 크게 변하지 않는다는 가정하에 고농도 샘플이 존재하는 픽셀 주변으로 5%의 값의 변화량을 주어 샘플 수를 늘려주었다. 서브샘플링은 샘플 수가 많은 저농도 샘플을 랜덤하게 선택하는 방식을 적용하였다. 연구지역 내 지상 PM_{10} 및 $PM_{2.5}$ 농도는 범위가 크고 양으로 치우친 분포를 갖기 때문에 로그변환 후 모델링에 사용되었다. 2019년에 대해 최종적으로 총 448,779개의 샘플을 구축하였으며, 구축된 최종 샘플들은 계절별로 랜덤하게 8:2로 나누어 계절기반 계층적 샘플링(stratified sampling)을 수행하였고, 80%의 자료는 모델의 훈련자료로써 이를 이용하여 각 모델은 교차 검증 기반으로 하이퍼파라미터 최적화가 수행되며, 나머지 20%의 자료는 구축된 예측모델의 성능을 평가하는 자료로 이용된다.

2) 기계학습 모델

기계학습기법 적용 시 나타나는 가장 큰 문제점은 과적합(overfitting) 문제이다. Boosting기법은 트리(Tree)기반으로 데이터셋을 학습하지만 각각의 트리가 서로 독립적인 bagging 기법과는 달리 순차적으로 오차가 큰 부분에 가중치를 부여하는 방식으로 트리를 학습하여 과적합에 취약한 기법 중 하나이다. Boosting 모델들의 다양한 하이퍼파라미터들을 최적화 시킴으로써 과적합 문제를 예방할 수 있다. 이러한 Boosting 기반의 대표적인 기계학습 기법은 최신 Gradient boosting 앙상블 기법인 GBRTs이 있다. 하지만 GBRTs는 모델 학습하는데 있어 상당한 계산 시간이 소요되어 이를 보완하고자 여러 알고리즘이 개발되었고 LightGBM 알고리즘은 그 중 대표적으로 학습시간이 상당히 단축된 모델이다. 따라서 본 연구에서는 두 가지 GBRTs와 LightGBM을 적용하여 지상 PM_{10} 및 $PM_{2.5}$ 농도추정 모델을 개발하였다.

(1) Gradient Boosted Regression Trees

본 연구에서는 GBRTs 모델을 사용하여 PM 지상 농도를 추정하였다. Gradient boosting 은 경사하강법(gradient descent) 알고리즘으로, 경사하강법이란 미분이 가능한 손실함수(loss 혹은 cost function) 를 최소화하기 위해서 모델의 변수들을 변경시킴으로써 손실함수의 경사도가 감소하는 방향으로 수행된다(Friedman, 2002; Pedregosa *et al*., 2011). 극소(local minimum) 문제를 피하고 전역적인 해(global optimum)를 찾기 위해서 훈련자료의 부분샘플(subsample)을 사용하는 방식으로 모델이 학습된다. 트리기반의 기계학습 기법인 GBRTs는 구축된 훈련자료를 기반으로 결정트리들(decision trees)을 생성한다. 트리는 결정 노드(node)를 따라서 부분샘플의 오차를

Table 2. Hyperparameters of the GBRT model tested for model optimization in this study. Each element in square bracket for each hyperparameter was examined, and the final hyperparameters are shown by superscript letters

Number of trees (n_estimators)	Maximum depth of a tree (max_depth)	Minimum number of samples in a node (min_samples_split)	Minimum samples in a terminal node or leaf (min_samples_leaf)
[100, 500, 1000$^{PM10,PM2.5}$]	[4, 6$^{PM2.5, 8PM10}$]	[10$^{PM10, PM2.5}$, 50, 100]	[10$^{PM10, PM2.5}$, 50, 100]

최소화하는 방향으로 훈련자료를 나눈다. 모델의 결과에 대한 신뢰도를 높이기 위해서 결정 트리들이 합쳐짐으로써(즉, 앙상블; ensemble) 최종 예측 값을 산출하게 된다. 트리들은 점진적으로 이전의 앙상블 트리 모델에 더해지며 다음의 새로운 트리는 경사하강법을 통해서 이전 모델의 오차에 대해 피팅이되는 방식으로 모델이 훈련된다. 이러한 방법은 기존의 앙상블 트리 방법(e.g., Random Forest)과는 달리 트리들이 체계적으로 구축이 되며 더 적은 모델 반복 (interation)이 요구된다는 장점이 있다.

GBRTs 모델은 Python 3 sikit-learn 환경에서 수행되었고 sikit-learn에서 제공하는 기계학습 library (Gradient Boosting Regressor)를 이용하여 수행하였다(Pedregosa *et al.*, 2011). 모델의 구조는 모델에 사용되는 트리의 개수, 트리의 깊이, 학습율 등의 하이퍼파라미터(hyperparameters)에 의해서 결정된다. Table 2는 본 연구에서 테스트된 하이퍼파라미터 조합을 보여준다. 하이퍼파라미터들은 교차 검증을 기반한 grid search 방법을 이용하여 최적화된다. Grid search 방법은 모델에 사용되는 여러 종류의 하이퍼파라미터들 사이의 여러 조합들을 하나씩 테스트하며 모델의 성능 결과를 기반으로 최적의 변수 조합을 찾는 방법이다. 트리의 개수(n_estimator)은 훈련과정에서 생성되는 트리의 개수를 결정한다. GBRTs 모델은 많은 수의 트리를 이용하더라도 과적합 문제에 비교적 견고한 모델이지만 교차 검증을 통해서 학습률에 따라 최적화를 시켜 과적합 문제를 예방할 필요가 있다. 트리의 최대 깊이(max_depth) 변수는 한 트리가 얼마큼 깊게 나눠지는지를 결정한다. 깊이가 깊어질 수록 특정 샘플에 집중해서 학습이 되는 경향이 있기 때문에 최적화가 필요한 부분이다. 최소샘플분할(min_samples_split) 변수는 모델이 훈련자료를 분할할 때 결정 노드에서 필요되어지는 최소한의 샘플을 결정한다. 이 변수는 값이 높을수록 특정 샘플에 학습되는 과적합을 예방할 수 있으며 한편 너무 큰 값은 과소적합될 수 있기 때문에 최적화가 필요하다. 최소샘플리프(min_samples_leaf) 변수는 최종노드나 리프에서 필요 되어지는 최소한의 샘플수를 결정한다.

(2) Light Gradient Boosting Model

GBRT 알고리즘의 단점은 자료의 갯수가 많아지면 계산 시간이 크게 증가하여 모델 훈련의 효율성이 떨어진다. 이러한 문제점을 해결하기 위해서 Ke *et al.* 2017은 GOSS (Gradient-based One-Side Sampling) 방법을 도입했다. 이는 작은 경사도를 가지는 샘플들은 작은 에러를 가짐으로써 잘 훈련이 된 반면 큰 경사도를 가지는 샘플들은 덜 훈련된다는 사실을 기반으로해서 GOSS 방법은 작은 경사도를 가지는 샘플들에 대해서는 랜덤 샘플링을 통해 부분 추출을 하고 큰 경사도를 가지는 샘플들은 전부를 다음 훈련으로 가져가는 방식으로 큰 경사도를 가지는 샘플에 좀 더 집중을 하는 방법이다. LightGBM은 GOSS를 기반으로하여 보다 모델의 복잡성을 줄이고 훈련을 효율적으로 수행한다.

LightGBM 모델은 Python 3 sikit-learn 환경에서 수행되었고 제공되는 패키지인 lightgbm의 LGBMRegressor를 이용하여 수행하였다. Table 3는 본 연구에서 테스트된 LightGBM의 하이퍼파라미터 조합을 보여준다. num_leaves는 리프의 갯수로 트리 모델의 복잡성을 결

Table 3. Hyperparameters of the LightGBM model tested for model optimization in this study. Each element in square bracket for each hyperparameter was examined, and the final hyperparameters are shown by superscript letters

Boosting types (boosting_type)	Number of trees (n_estimators)	Maximum depth of a tree (max_depth)	Number of leaves (num_leaves)	Minimum data in a leaf (min_data_in_leaf)
['goss'$^{PM10, PM2.5}$, 'dart']	[100,500, 1000$^{PM10, PM25}$]	[4P^{M25},6,8 ^{PM10}]	[8^{PM25}, 32^{PM10},128]	[20^{PM25}, 50^{PM10},100, 250]

정하는 주요 변수이다. Leaf-wise 트리인 LightGBM은 자칫 특정 리프에 대해서 너무 깊이 훈련이될 수 있기 때문에 대개 num_leaves 변수를 2^{max_depth} 보다 더 작게 설정해준다. min_data_in_leaf는 리프에 필요한 최소한의 샘플수로 과적합을 조절하는데 사용되는 변수이다. 이 변수의 최적의 값은 훈련 샘플의 수와 num_leaves 변수에 영향을 받는다. 이 값을 높게 지정하면 지나치게 깊은 트리로 자라는 것을 피할 수 있지만 한편 과소적합이 발생할 수 있기 때문에 최적화시킬 필요가 있다. num_iterations 변수는 수행되는 Boosting 횟수로 트리의 갯수를 의미한다. 이 변수와 학습율(learning_rate) 변수는 수로 영향을 주며 대개 트리 갯수를 감소시킬 때 학습율을 증가시킨다.

3. 결과 및 토론

1) 모델 성능 결과

2019년에 대하여 두 기계학습 기법을 기반으로 개발된 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델의 성능을 평가하였으며, 한 시간 단위의 자료를 활용하여 모델링하였기 때문에 시간별 검증을 수행하였다. Fig. 2와 3는 GBRT기법 기반으로 각각 Feature set 1과 2에 대한 지상 PM_{10} 및 $PM_{2.5}$ 농도 추정 값과 현장관측자료를 비교한 결과이다. PM_{10} 및 $PM_{2.5}$ 둘 다 기상 변수를 포함하는 모든 변수를 활용한 Feature set 1을 사용하였을 때 높은 정확도를 보였다(PM_{10}: R^2 = 0.83, nRMSE = 32.7%, $PM_{2.5}$: R^2 = 0.80, nRMSE = 31.8%). 특히, PM_{10} 결과에서 저농도 사례 (200 $\mu g/m^3$ 이하)에 대한 과대추정이 줄고 고농도 사례들의 모의 성능이 향상하면서 전체적인 R^2 정확도가 약 10% 정도 증가하였다(Fig. 2과 3). LightGBM기법을 적용하여 Feature set 1과 2에 대해 모델링한 결과는 Fig. 4와 5에 각각 나타내었다. GBRT 모델 결과와 마찬가지로 기상 변수들이 포함되지 않은 Feature set 2보다 Feature set 1에 대해서 더 높은 정확도를 보였으며(PM_{10}: R^2 = 0.79, nRMSE = 36.6%, $PM_{2.5}$: R^2 = 0.65, nRMSE = 42.6%), PM_{10} 결과에서 큰 정확도 향상이 나타났다.

두 기법 간의 결과를 비교하였을 때, LightGBM기반 모델결과보다 GBRT기반 모델결과가 더 높은 성능을 보였으며, 특히, $PM_{2.5}$ 추정 결과에서 10 % 이상 높은 성능을 보였다(Fig. 2). 하지만 LightGBM의 하이퍼파라미터 튜닝은 GBRT에 비해서 최대 45배 빠르게 수행되었다 (예, Feature set 2를 이용한 $PM_{2.5}$에 대한 LightGBM 러닝 타임 약 28초 소요, 반면 GBRT 약 1276초 소요. 모델 구동 시스템 사양은 Intel(R) Xeon(R) 160GB memory 의 CPU E5-2650 v3 @2.30GHz). Zhang *et al.* (2021)는 중국 전역에 대하여 GBDT (gradient boosting decision tree) 기법기반 일평균 $PM_{2.5}$ 농도추정 모델을 개발하여 R^2 = 0.81, RMSE =

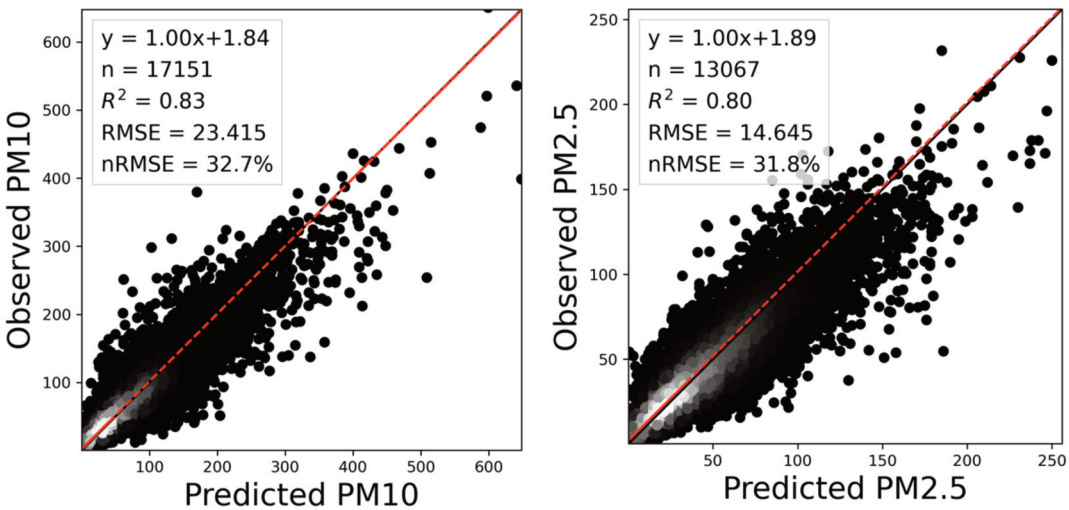

Fig. 2. Scatter plots using the GBRT model results for independent test data with feature set 1 (All variables) for PM_{10} (left) and $PM_{2.5}$ (right). Points are represented as the number of observations estimated by a Gaussian kernel density estimation. The brighter the point, the higher density of the samples.

11.57 μg/m³의 정확도를 보였으며, 본 연구에서 개발된 시간 단위의 $PM_{2.5}$ 결과(R^2 = 0.80, RMSE = 14.65 μg/m³)와 유사한 성능을 보였다. XGBoost (Extreme gradient boosting models) (Chen *et al.*, 2019; Qadeer and Jeon, 2019; Gui *et al.*, 2020), GW-GBM (Geographically-Weighted Gradient Boosting Machine) (Zhan *et al.*, 2017), and GBDT (Zhang *et al.*, 2021) 등 다양한 Boosting기법 기반의 기존 연구들은 각기 다른 연구 지역 및 연구 기간, 산출 시간해상도를 갖기 때문에 직접적인 비교는 어렵지만 본 연구에서 제시한 GBRT 모델의 정확도가 기존 연구들과 비슷하거나 더 나은 결과를 보였다. 게다가, 본 연구에서 제시한 모델은 기존 연구와 다르게 시간별 지상 PM_{10} 및 $PM_{2.5}$ 농도 제공이 가능하다는 장점이 있다. 따라서 본 연구에서는 모델 성능이 가장 좋았던 Feature set 1을 사용하여 GBRT 기법을 적용한 지상 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델을 최종 모델로 선정하여 추가적인 모델 분석을 수행하였다.

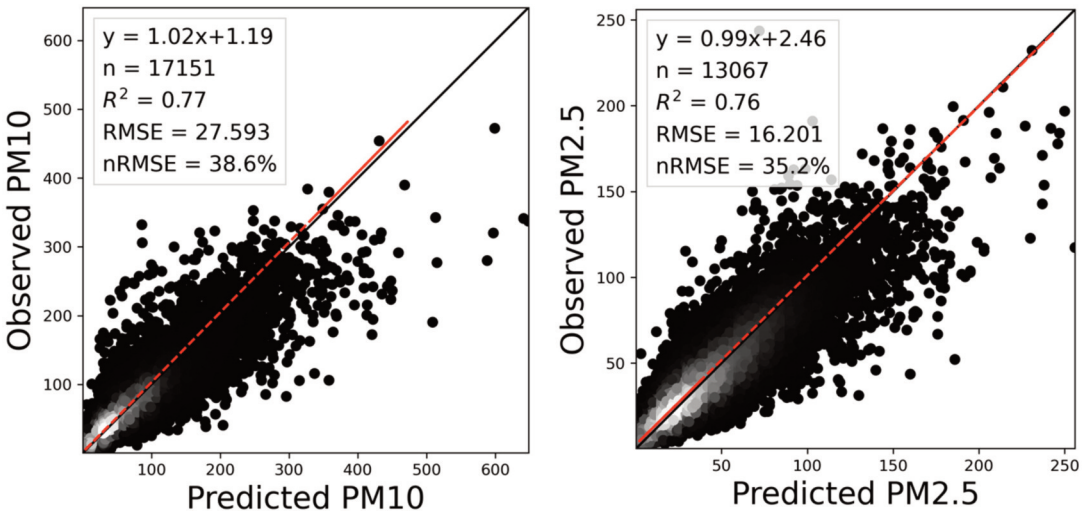

Fig. 3. Scatter plots using the GBRT model results for independent test data with feature set 2 (5 selected variables) for PM_{10} (left) and $PM_{2.5}$ (right). Points are represented as the number of observations estimated by a Gaussian kernel density estimation. The brighter the point, the higher density of the samples.

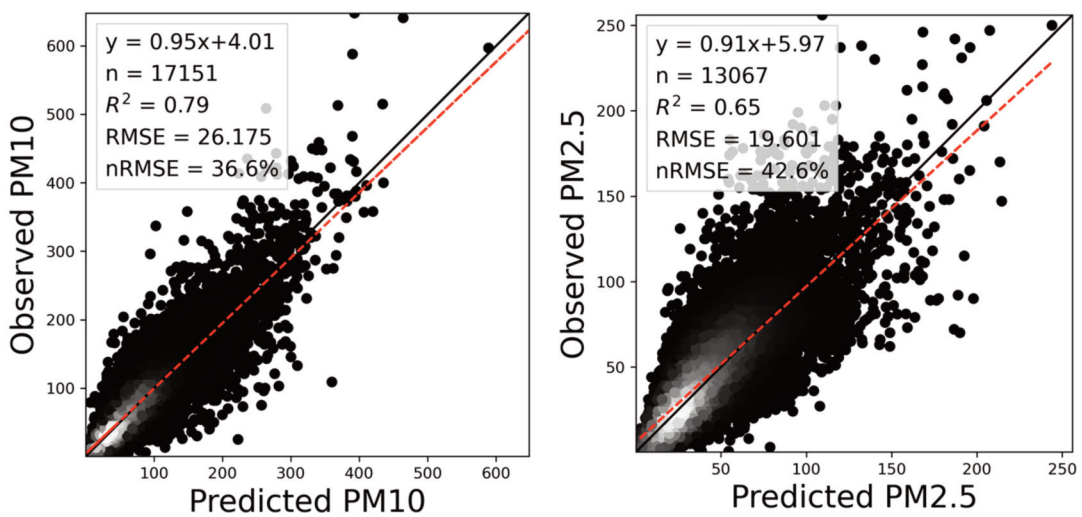

Fig. 4. Scatter plots using the LightGBM model for independent test data with feature set 1 (All variables) for PM_{10} (left) and $PM_{2.5}$ (right). Points are represented as the number of observations estimated by a Gaussian kernel density estimation. The brighter the point, the higher density of the samples.

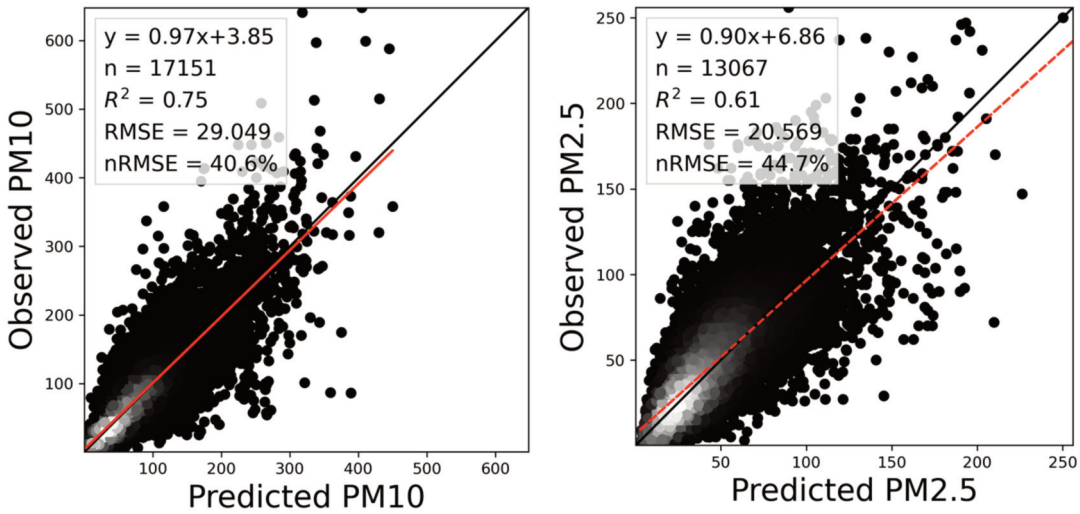

Fig. 5. Scatter plots using the LightGBM model for independent test data with feature set 2 (5 selected variables) for PM_{10} (left) and $PM_{2.5}$ (right). Points are represented as the number of observations estimated by a Gaussian kernel density estimation. The brighter the point, the higher density of the samples.

2) 모델 분석 결과

(1) 변수 중요도

사용된 기계학습 기법은 변수 중요도 정보를 제공하여 모델링에 중요하게 사용된 변수들에 대한 분석이 가능하다. Fig. 6은 최종모델로 선정된 GBRT기반 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델에 사용된 입력변수들의 변수 중요도(Permutation importance)이다. 이는 하나의 변수 값이 랜덤하게 섞인 값으로 대체될 때 모델 정확도의 감소를 측정한 값이다. 모델의 정확도가 크게 감소했다는 것은 해당 변수가 모델에 큰 기여를 했다는 것을 의미한다. 본 연구에서는 R^2 값을 기반으로 모델의 정확도를 평가하고 변수 중요도를 산출하였다. 총 10회 동안 무작위로 섞어서 변수중요도 값을 반복하여 뽑은 다음 최종 변수 중요도를 산정하였고 이는 표준편차를 보여주는 에러바로 표현되었다.

PM_{10}과 $PM_{2.5}$ 에 대해서 공통적으로 DOY, AOD, Temp, NDVI가 PM을 추정하는데 있어 가장 높은 기여도를 보였다. 이는 위성기반 AOD를 사용한 PM 농도 추정의 기존 연구와 비슷한 기여도 패턴을 보였으며, 대체적으로 위성기반 AOD와 기상 변수들의 모델 기여도가 높았다(Xu *et al.*, 2018; Park *et al.*, 2019). 지상 기반 PM 농도는 계절에 따라 농도 변화가 확연하게 구분되기 때문에 시계열 정보를 나타내는 DOY가 가장 높은 변수 중요도를 보였다. AOD는 PM 패턴에 대한 가장 좋은 예측 변수이기 때문에 GOCI기반 AOD 변수가 PM 추정에 큰 기여를 한 것으로 확인된다. 한편 AOD는 전체의 태양 복사 에너지의 감소를 나타내지만 SSA는 전체 감소에 대한 산란의 영향을 나타낸다(Jethva *et al.*, 2014). 대개 SSA는 도시 및 산업 지역에서 높은 값이 산출되며 따라서 이는 $PM_{2.5}$와 더 높은 관련성을 가진다(Jethva *et al.*, 2014). 마찬가지로 본 연구에서 SSA 변수는 PM_{10} 보다 $PM_{2.5}$에 대해서 모델 구축에서 높은 기여를 하였다(Fig. 6). 이는 작은 에어로졸 입자가 큰 입자에 비해서 태양빛을 더 많이 반사하는 경향이 있기 때문이다. 기온의 경우 온도가 높아지면 대류가 활발해서 PM_{10} 및 $PM_{2.5}$ 농도가 낮아지는 음의 상관관계를 갖으며, 본 연구의 연구지역은 사계절이 뚜렷해 온도가 높은 여름철 낮은 PM_{10} 및 $PM_{2.5}$ 농도를 보인다. 또한, 최대 풍속이 상대적으로 중요한 변수로 확인되었다. 기존 연구에 따르면 풍속이 강한 경우 활발한 대류를 통해 대기오염물질이 정체되는 것을 방지함으로써 풍속과 PM_{10} 및 $PM_{2.5}$ 농도 사이에 높은 음의 상관관계를 가진다(Kukkonen *et al.*, 2005). 본 연구에서는 누적 최대풍속 자료가 모델 구축에 높은 기여를 하였으며, 이는 하루 이상 강하거나 약한 풍속이 지속되었을 때 PM_{10} 및 $PM_{2.5}$ 농도 산정에 영향을 끼친다는 것을 알 수 있다. 또한, 에어로졸의 인위적인 발원지와 관련되는 인구밀도 및 도로밀도 변수

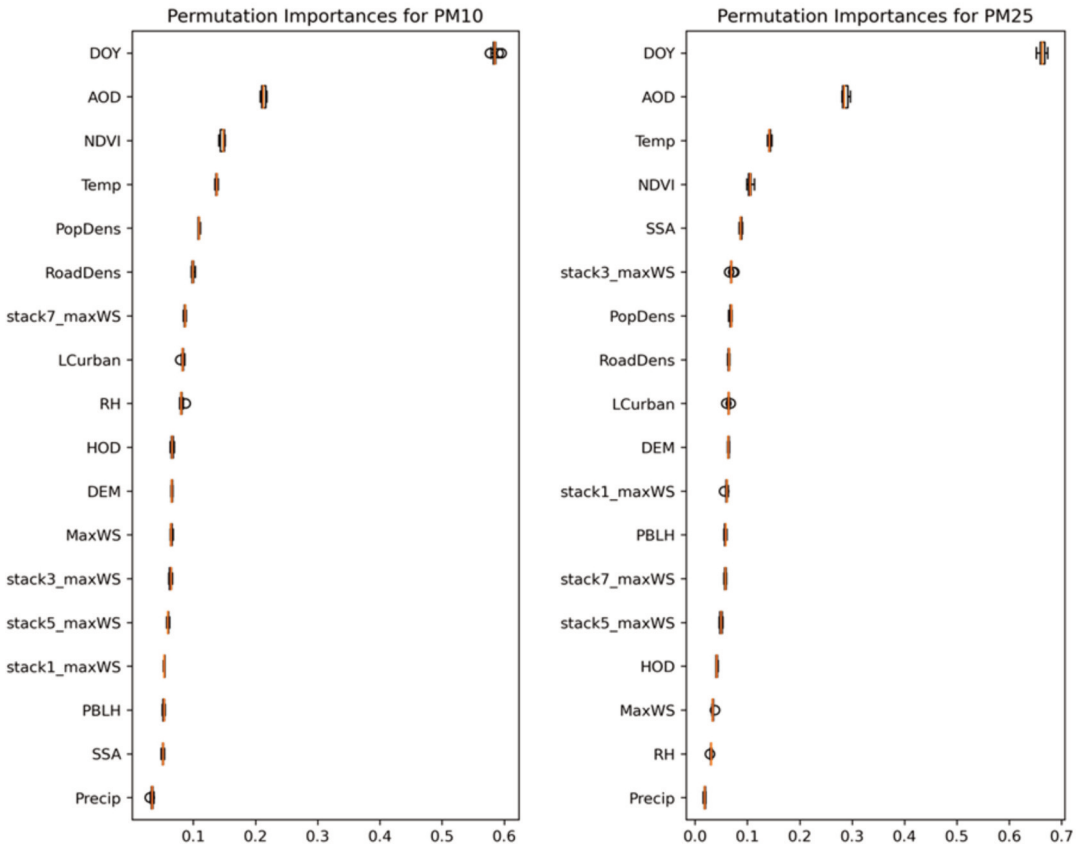

Fig. 6. Relative variable importance of the GBRT models for PM_{10} (left) and $PM_{2.5}$ (right).

가 PM_{10} 및 $PM_{2.5}$ 추정에 꽤 높은 정도의 중요도를 가지는 것으로 확인되었다.

3) 공간분포 분석

Fig. 7과 8은 GBRT 모델 기반으로 추정된 PM_{10} 및 $PM_{2.5}$의 계절별 분포를 각각 나타내고 있다. PM_{10}의 경우 봄철과 겨울철에서 다른 계절에 비해 높은 값을 보인다(Fig. 7). 이는 봄철에는 중국 동북쪽 몽골 지역으로부터의 황사에 의해 PM 농도 값이 높아지는 것으로 확인된다(Rodrıguez *et al.*, 2001; Wang *et al.*, 2006; Krasnov *et al.*, 2016). 반면 큰 에어로졸 입자인 황사는 PM_{10} 농도에 영향을 주기 때문에 작은 에어로졸 입자 농도를 나타내는 $PM_{2.5}$에 비해서 봄철에 PM_{10}이 전체적으로 높은 값을 보인다(Fig. 8). $PM_{2.5}$의 경우 인위적 에어로졸 배출과 높은 연관성을 가짐으로써, 화석연료의 사용이 늘어가는 겨울철에 상대적으로 더 높은 농도 값이 관측된다. 여름철의 경우 온도가 높아지면서 PBLH가 높아지고 대류의 순환이 활발해지면서 지상 PM_{10} 및 $PM_{2.5}$ 농도가 감소하는 패턴을 보였다(Du *et al.*, 2013; Wang *et al.*, 2006; Lv *et al.*, 2020). 본 연구결과에서도 이와 같은 공간적인 패턴이 잘 모의되었다(Fig. 7과 8).

Fig. 9는 추정된 PM_{10} 및 $PM_{2.5}$ 농도의 연평균 분포와 각 물질별 현장관측자료의 공간적인 분포를 나타내었다. PM_{10} 및 $PM_{2.5}$ 모두 중국동부에서 높은 농도 분포를 보였으며, 공장이 밀집 되어있는 베이징 등 대도시 지역에서 높은 분포를 나타냈다. 또한, 일본과 남한에서 중국과 비교하여 낮은 농도 분포를 갖는 지역별 농도 분포 패턴이 잘 모의되었다. 그러나 본 연구를 통해 추정된 PM_{10} 및 $PM_{2.5}$ 농도 값은 실제 현장관측자료에 비해 전반적으로 과대 추정하는 경향을 보였다(Fig. 9). 중국 동북쪽 또한, 실제 관측 값에 비해 과대 추정되는 한계점을 보였다. PM_{10}의 경우, 사막지역을 포함한 몽골지역에서도 높은 농도를 보였으나 현장관측자료와 잘 매칭되지 않는 한계점을 보였다. 이는 건조한 사막지역의

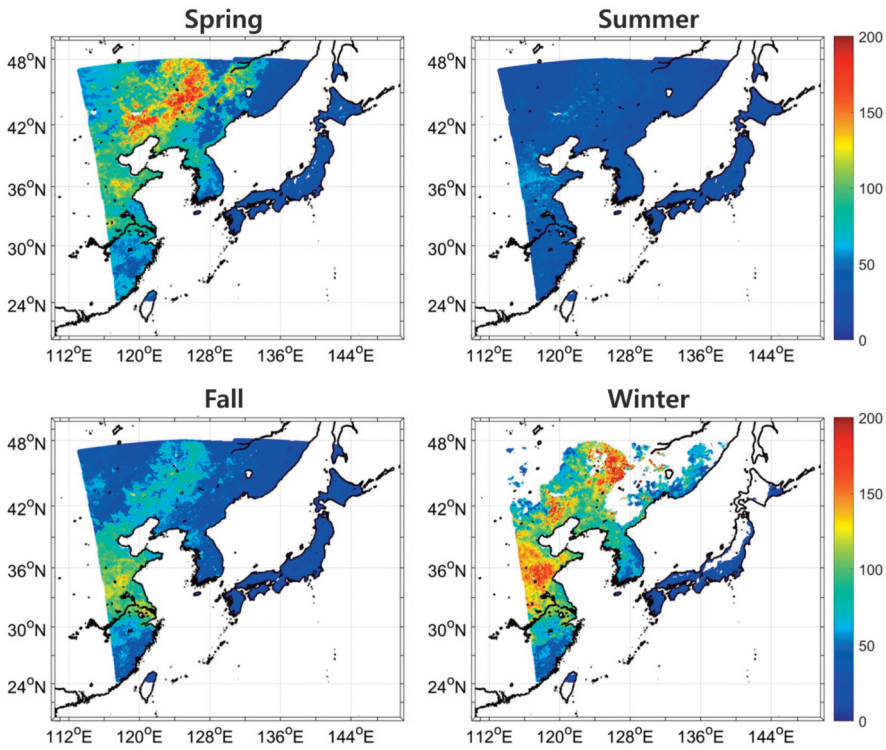

Fig. 7. Seasonal distribution of PM_{10} concentrations estimated by the GBRT model.

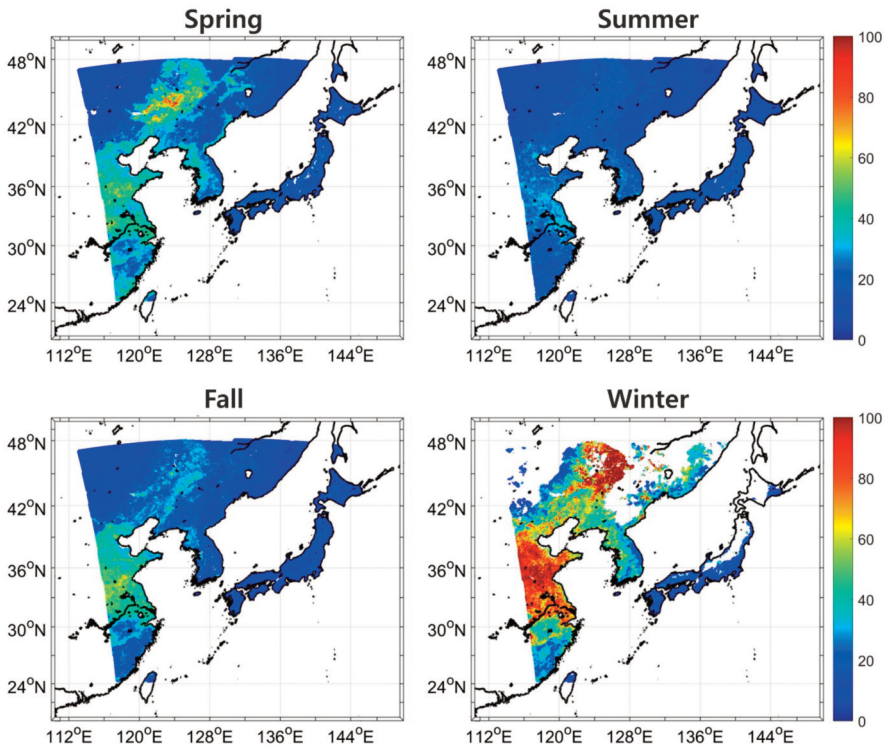

Fig. 8. Seasonal distribution of $PM_{2.5}$ concentrations estimated by the GBRT model.

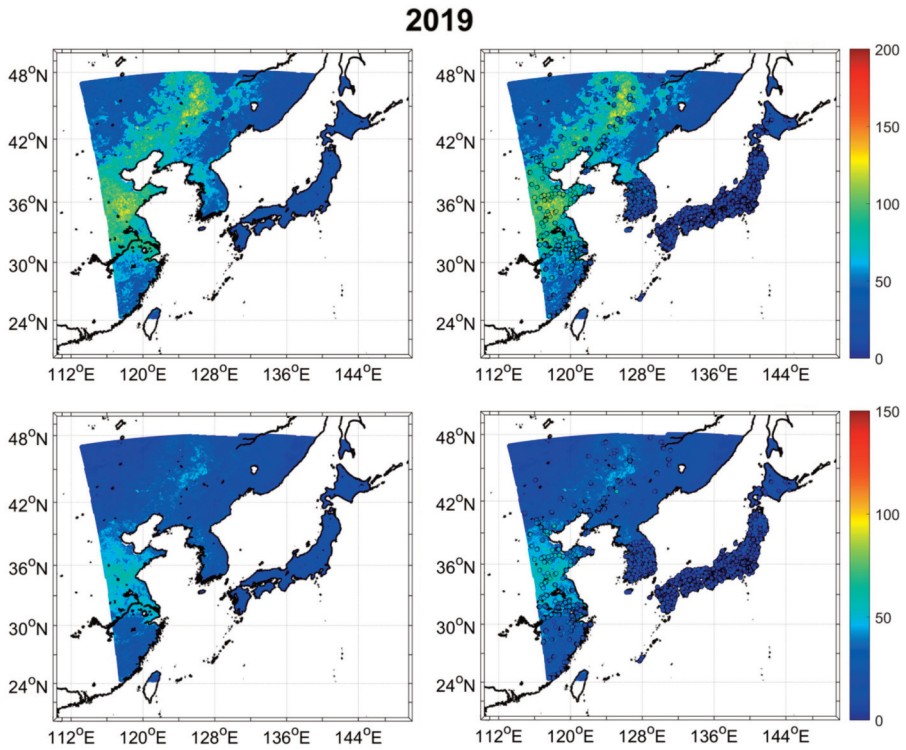

Fig. 9. Annual distribution of PM_{10} (upper) and $PM_{2.5}$ (bottom) estimation with observed PM_{10} and $PM_{2.5}$ concentrations, respectively.

경우 지표면의 영향에 의해 정확한 에어로졸 산출이 어렵기 때문으로 판단된다 (Sorek-Hamer *et al.*, 2015). 지표면 반사도가 높은 지역에서 대개 AOD 산출이 과대 추정되는 경향이 있는데, AOD는 PM 지상농도 추정에 매우 중요한 변수로 활용되기 때문에 중국 동북쪽에서 높은 PM 농도가 산출된 것으로 보인다.

4. 결론

본 연구에서는 2019년에 대하여 정지궤도 위성인 GOCI 위성 기반 AOD 산출물과 다중위성기반 자료와 수치모델로부터 다양한 기상 변수들 및 기타보조자료를 융합 활용하여 Boosting 기법기반 기계학습을 적용하여 지상 PM_{10} 및 $PM_{2.5}$ 농도를 추정하였다. 한국의 AirKorea, 중국 BMEMC 및 일본의 NIES에서 제공하는 검증된 PM 지상농도 현장관측 자료를 이용하여 모델의 기준자료로 활용했다. 두 가지 Boosting 기법인 GBRT와 LightGBM기법을 각각 활용하여 모델을 구축하였으며, 기상변수 및 토지피복변수의 모델 기여도를 확인하기 위해 두 가지의 Feature set을 구분하여 모델의 성능을 평가하였다.

LightGBM이 GBRT에 비해 모델 구동시간이 빠르다는 장점이 있지만, GBRT기반 지상 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델의 성능이 전반적으로 높게 나타났다. PM_{10}과 $PM_{2.5}$ 농도를 추정하는데 공통적으로 AOD, SSA, DEM, DOY, HOD가 중요 변수로 확인되었다. PM_{10}과 $PM_{2.5}$ 모두 시계열 정보를 포함한 DOY와 에어로졸 정보를 가장 잘 반영하고 있는 AOD 변수가 모델 구축에 가장 높은 기여를 하였다. 하지만 공통 변수들만 사용하였을 때보다 기상변수 및 토지피복변수를 함께 사용하였을 때 특히 PM_{10} 농도 추정에 대해서 대략 10%까지 정확도가 향상되었다. 변수중요도에서도 온도, 최대풍속 등 기상변수와 토지피복변수(식생지수, 인구밀도, 도로밀도)가 상위 10개안에 포함되어 모델 구축에 중요한 역할을 하는 것으로 나타났다. 최종모델로 선정된 Feature set 1 (기상 및 토지피복변수 등 포함 모든 변수)을 사용한 GBRT기반으로 추정된 지상 PM_{10} 및 $PM_{2.5}$ 농도의 계절

별 분포와 연평균 분포 또한 현장관측자료와 일치하는 양상을 보였다. 다만, 사막지역의 경우 지표면의 영향에 의해 정확한 에어로졸 산출이 어려워 과대추정되는 한계점을 보였다. 본 연구에서 개발된 지상 PM_{10} 및 $PM_{2.5}$ 농도 추정 모델은 넓은 지역에 대한 시간별 농도 분포 제공이 가능하여 대기질 모니터링에 큰 기여를 할 것으로 사료된다.

중국지역을 제외한 남한, 일본에 대해서는 다소 과대추정하는 경향을 보여 향후 SHAP (SHapley Additive exPlanations) value와 같은 상세한 변수분석기법을 통해 지역별 또는 시기별로 어떠한 변수가 중요하게 활용되고 있는지 추가적인 분석이 가능하다. 본 연구에서는 2019년 한 해에 대해서만 모델링을 하였지만, 연구 기간을 추가하여 장기간 지상 PM_{10} 및 $PM_{2.5}$ 농도의 시공간적 변화를 확인해볼 수 있다. 하지만 장기간 자료를 사용할 경우 시간에 따른 PM 농도의 장기 변화 등의 추가적인 고려가 필요하다. 또한, 2020년 2월에 발사된 정지궤도 환경위성(GK-2B; GEO-KOMPSAT-2B)에 탑재된 GEMS (Geostationary Environment Monitoring Spectrometer)에서 제공 예정인 파장대별 AOD와 컬럼별 정보를 활용한다면 모델의 정확도 향상에 기여할 것으로 기대된다. GOCI기반 산출물과 비교하여 더 넓은 지역에 대한 정보 제공이 가능하며, 최대 하루 10회 관측하기 때문에 지상 PM_{10} 및 $PM_{2.5}$ 농도의 연속적인 시계열 흐름 파악이 가능할 것으로 기대된다.

사사

본 논문은 환경부의 재원으로 국립환경과학원의 지원을 받아 수행하였고(NIER-2020-04-02-086), 과학 기술정보통신부 및 정보통신기획평가원의 대학ICT연 구센터지원사업의 연구결과로 수행되었음(IITP-2021-2018-0-01424).

References

Chen, Z.-Y., T.-H. Zhang, R. Zhang, Z.-M. Zhu, J. Yang, P.-Y. Chen, C.-Q. Ou, and Y. Guo, 2019. Extreme gradient boosting model to estimate $PM_{2.5}$ concentrations with missing-filled satellite data in China, *Atmospheric Environment*, 202: 180-189.

Du, C., S. Liu, X. Yu, X. Li, C. Chen, Y. Peng, Y. Dong, Z. Dong, and F. Wang, 2013. Urban boundary layer height characteristics and relationship with particulate matter mass concentrations in Xi'an, central China, *Aerosol and Air Quality Research*, 13(5): 1598-1607.

Friedman, J.H., 2002. Stochastic gradient boosting, *Computational Statistics & Data Analysis*, 38(4): 367-378.

Geng, G., Q. Zhang, R.V. Martin, A. van Donkelaar, H. Huo, H. Che, J. Lin, and K. He, 2015. Estimating long-term $PM_{2.5}$ concentrations in China using satellite-based aerosol optical depth and a chemical transport model, *Remote sensing of Environment*, 166: 262-270.

Ghotbi, S., S. Sotoudeheian, and M. Arhami, 2016. Estimating urban ground-level PM_{10} using MODIS 3 km AOD product and meteorological parameters from WRF model, *Atmospheric Environment*, 141: 333-346.

Gui, K., H. Che, Z. Zeng, Y. Wang, S. Zhai, Z. Wang, M. Luo, L. Zhang, T. Liao, and H. Zhao, 2020. Construction of a virtual $PM_{2.5}$ observation network in China based on high-density surface meteorological observations using the Extreme Gradient Boosting model, *Environment International*, 141: 105801.

Guo, B., D. Zhang, L. Pei, Y. Su, X. Wang, Y. Bian, D. Zhang, W. Yao, Z. Zhou, and L. Guo, 2021. Estimating $PM_{2.5}$ concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017, *Science of The Total Environment*, 778: 146288.

Gupta, P. and S.A. Christopher, 2009a. Particulate matter air quality assessment using integrated surface,

satellite, and meteorological products: Multiple regression approach, *Journal of Geophysical Research: Atmospheres*, 114(D14205): 1-13.

Gupta, P. and S.A. Christopher, 2009b. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach, *Journal of Geophysical Research: Atmospheres*, 114(D20205): 1-14.

Jerrett, M., M.C. Turner, B.S. Beckerman, C.A. Pope III, A. Van Donkelaar, R.V. Martin, M. Serre, D. Crouse, S.M. Gapstur, and D. Krewski, 2017. Comparing the health effects of ambient particulate matter estimated using ground-based versus remote sensing exposure estimates, *Environmental Health Perspectives*, 125(4): 552-559.

Jethva, H., O. Torres, and C. Ahn, 201., Global assessment of OMI aerosol single-scattering albedo using ground-based AERONET inversion, *Journal of Geophysical Research: Atmospheres*, 119(14): 9020-9040.

Jin, K., 2018. LEO and GEO satellite programs for space-borne measurement of aerosol, *Current Industrial and Technological Trends in Aerospace*, 16(1): 53-62.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, 2017, Lightgbm: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, 30: 3146-3154.

Kloog, I., M. Sorek-Hamer, A. Lyapustin, B. Coull, Y. Wang, A.C. Just, J. Schwartz, and D.M. Broday, 2015. Estimating daily $PM_{2.5}$ and PM_{10} across the complex geo-climate region of Israel using MAIAC satellite-based AOD data, *Atmospheric environment*, 122: 409-416.

Krasnov, H., I. Katra, and M. Friger, 2016. Increase in dust storm related PM_{10} concentrations: A time series analysis of 2001-2015, *Environmental Pollution*, 213: 36-42.

Kukkonen, J., M. Pohjola, R.S. Sokhi, L. Luhana, N. Kitwiroon, L. Fragkou, M. Rantamäki, E. Berge, V. Ødegaard, and L.H. Slørdal, 2005. Analysis and evaluation of selected local-scale PM_{10} air pollution episodes in four European cities: Helsinki, London, Milan and Oslo, *Atmospheric Environment*, 39(15): 2759-2773.

Lee, H.J., R.B. Chatfield, and A.W. Strawa, 2016. Enhancing the applicability of satellite remote sensing for $PM_{2.5}$ estimation using MODIS deep blue AOD and land use regression in California, United States, *Environmental Science & Technology*, 50(12): 6546-6555.

Lee, P.S.-H., J.C. Park, and J.-Y. Seo, 2020. Estimation of ambient PM_{10} and $PM_{2.5}$ concentrations in Seoul, South Korea, using empirical models based on MODIS and Landsat 8 OLI imagery, *Korean Journal of Agricultural Science*, 47(1): 59-66 (in Korean with English abstract).

Lim, H., M. Choi, M. Kim, J. Kim, and P. Chan, 2016. Retrieval and validation of aerosol optical properties using Japanese next generation meteorological satellite, Himawari-8, *Korean Journal of Remote Sensing*, 32(6): 681-691 (in Korean with English abstract).

Liu, Y., J.A. Sarnat, V. Kilaru, D.J. Jacob, and P. Koutrakis, 2005. Estimating ground-level $PM_{2.5}$ in the eastern United States using satellite remote sensing, *Environmental Science & Technology*, 39(9): 3269-3278.

Lv, Z., W. Wei, S. Cheng, X. Han, and X. Wang, 2020, Meteorological characteristics within boundary layer and its influence on $PM_{2.5}$ pollution in six cities of North China based on WRF-Chem, *Atmospheric Environment*, 228: 117417.

Ma, Z., X. Hu, L. Huang, J. Bi, and Y. Liu, 2014. Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing, *Environmental Science & Technology*, 48(13): 7436-7444.

Park, S., M. Shin, J. Im, C.-K. Song, M. Choi, J. Kim, S. Lee, R. Park, J. Kim, and D.-W. Lee, 2019.

Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea, *Atmospheric Chemistry and Physics*, 19(2): 1097-1113.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, 2011. Scikit-learn: Machine learning in Python, *The Journal of Machine Learning Research*, 12: 2825-2830.

Pope III, C.A., M. Ezzati, and D.W. Dockery, 2009, Fine-particulate air pollution and life expectancy in the United States, *New England Journal of Medicine*, 360(4): 376-386.

Qadeer, K. and M. Jeon, 2019. Prediction of PM_{10} Concentration in South Korea Using Gradient Tree Boosting Models, *Proc. of 2019 the 3rd International Conference on Vision*, Image and Signal Processing, Vancouver, CA, Aug. 26-28, pp. 1-6.

Rodrıguez, S., X. Querol, A. Alastuey, G. Kallos, and O. Kakaliagou, 2001. Saharan dust contributions to PM_{10} and TSP levels in Southern and Eastern Spain, *Atmospheric Environment*, 35(14): 2433-2447.

She, Q., M. Choi, J.H. Belle, Q. Xiao, J. Bi, K. Huang, X. Meng, G. Geng, J. Kim, and K. He, 2020. Satellite-based estimation of hourly $PM_{2.5}$ levels during heavy winter pollution episodes in the Yangtze River Delta, China, *Chemosphere*, 239: 124678.

Soni, M., S. Payra, and S. Verma, 2018. Particulate matter estimation over a semi arid region Jaipur, India using satellite AOD and meteorological parameters, *Atmospheric Pollution Research*, 9(5): 949-958.

Sorek-Hamer, M., I. Kloog, P. Koutrakis, A.W. Strawa, R. Chatfield, A. Cohen, W.L. Ridgway, and D.M. Broday, 2015. Assessment of $PM_{2.5}$ concentrations over bright surfaces using MODIS satellite observations, *Remote Sensing of Environment*, 163: 180-185.

Stirnberg, R., J. Cermak, S. Kotthaus, M. Haeffelin, H. Andersen, J. Fuchs, M. Kim, J.-E. Petit, and O. Favez, 2021. Meteorology-driven variability of air pollution (PM_1) revealed with explainable machine learning, *Atmospheric Chemistry and Physics*, 21(5): 3919-3948.

Wang, S., W. Yuan, and K. Shang, 2006. The impacts of different kinds of dust events on PM_{10} pollution in northern China, *Atmospheric Environment*, 40(40): 7975-7982.

Xu, Y., H.C. Ho, M.S. Wong, C. Deng, Y. Shi, T.-C. Chan, and A. Knudby, 2018. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level $PM_{2.5}$, *Environmental Pollution*, 242: 1417-1426.

Zhan, Y., Y. Luo, X. Deng, H. Chen, M.L. Grieneisen, X. Shen, L. Zhu, and M. Zhang, 2017. Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm, *Atmospheric Environment*, 155: 129-139.

Zhang, T., W. He, H. Zheng, Y. Cui, H. Song, and S. Fu, 2021. Satellite-based ground $PM_{2.5}$ estimation using a gradient boosting decision tree, *Chemosphere*, 268: 128801.