Check for updates

Journal of Korean Society for Atmospheric Environment
Vol. 38, No. 1, February 2022, pp. 57-73
https://doi.org/10.5572/KOSAE.2022.38.1.57
p-ISSN 1598-7132, e-ISSN 2383-5346

논 문

시계열 머신러닝 기법을 이용한 컬럼에어로솔 예측 및 평가

Evaluation and Prediction of Column Aerosol by Using the Time Series Machine Learning Technique

김영일[1),3)], **이권호**[2),3),*], **이규태**[2),3)]

[1)]강릉원주대학교 공간정보협동과정, [2)]강릉원주대학교 대기환경과학과,
[3)]강릉원주대학교 복사-위성 연구소

접수일 2021년 12월 15일
수정일 2021년 12월 27일
채택일 2021년 12월 28일

Received 15 December 2021
Revised 27 December 2021
Accepted 28 December 2021

***Corresponding author**
Tel : +82-(0)33-640-2319
E-mail : kwonho.lee@gmail.com

Yeong-Il Kim[1),3)], **Kwon-Ho Lee**[2),3),*], **Kyu-Tae Lee**[2),3)]

[1)]*Spatial Information Cooperative Program, Gangneung-Wonju National University, Gangneung, Republic of Korea*
[2)]*Department of Atmospheric & Environmental Sciences, Gangneung-Wonju National University, Gangneung, Republic of Korea*
[3)]*Research Institute for Radiation-Satellite, Gangneung-Wonju National University, Gangneung, Republic of Korea*

Abstract Column aerosol observation has the advantage of obtaining microphysical information of aerosols present in the vertical atmospheric column at a given point. In this study, in order to predict column aerosol loads in the local atmosphere, time series machine learning technique was applied by using Aerosol Optical Depth (AOD) and Ångström Exponent (AE) data acquired from the selected ground-based Sun-sky radiometer observation network. For the determination of the best time series machine learning model, the independent optional properties including three regression models (glmnet, lm, and spark), four training period (1–4 years), and five regularization parameters (0–0.08) were tested in time series modelling. The results showed that spark model with the 1 year training period and regularization constant of 0.08 has the highest accuracy (RMSE=0.302, bias=0.213) for the AOD prediction. In the case of AE prediction, the highest accuracy (RMSE=0.356, bias=0.238) was obtained by the glmnet model with 1 year training period and the regularization constant of 0.08. In addition, machine learning clustering results shows that urban/industrial aerosol types occurred at a rate of 63.7% in Korea. The methodology and results of this study can be used for short-term aerosol prediction models and remote sensing data.

Key words: Aerosol, Machine learning, Aerosol optical depth, Classification, Air quality

1. 서 론

대기 중 다양한 형태로 존재하는 에어로솔(Aerosol) 입자는 대기 중에서 태양광을 흡수하거나 산란하므로, 시정거리 감소와 기후변화에 큰 영향을 미치는 것으로 알려져 있다(IPCC, 2013; Hansen *et al.*, 1997; Charlson *et al.*, 1992). 일반적으로, 에어로솔의 주요 성분은 주요 공업 지역이나 대도시 지역에서 발생하는 인위적 기원 오염성 입자가 큰 부분을 차지하고 있고, 황사나 산불 같은 자연 배출원에 의한 영향도 받는다(Lee, 2012). 특히, 동북아시아의 급속한 산업화로 인하여 대규모 오염물질이 배출되고 있으며, 대기 중으로 배출된 오염물질은 국가 간으로 이동하여 지역 대기 환경에 심각한 영향을 미치고 있다(Park *et al.*, 2013).

대기 에어로솔을 측정하는 방법은 주로 지점 관측(point measurement)이 사용되어 왔으며, 지점 관측의 대표적인 사례는 대기오염 관측망이나 대기질 관

측 및 연직 대기의 정보를 포함하는 컬럼 대기를 관측하기 위한 전 세계 Sun-sky radiometer 관측 네트워크인 Aerosol Robotic Network (AERONET)가 있다. 지상에서 에어로솔을 관측하는 기기 중 다파장 분광계 (multi-wavelength spectroradiometer)의 일종인 Sun-sky radiometer는 대기 컬럼 중 에어로솔의 광학 특성값을 연속 관측할 수 있으며 하루 1~2회만 관측할 수 있는 위성원격탐사에 비해 시간적 변화를 파악하기에 매우 용이하다 (Holben *et al.*, 2001). 이와 관련된 최근의 연구 사례 중, Park (2020) 연구에서는 국내 대기질 관측망 자료를 이용하여 COVID-19 기간 동안의 대구 지역 내 PM 농도 감소의 원인은 교통량의 급감 및 배출 저감 정책과 외부 유입량 감소에 따른 결과임을 밝혔다. 또한, AERONET 관측 자료를 이용하여 컬럼에어로솔의 시간적 변동성과 황사 현상에 분석이 이루어졌으며 (Kim *et al.*, 2008), 에어로솔 광학 두께 (Aerosol optical depth; AOD)와 Ångström Exponent (AE)는 여름에는 높고, 겨울에는 낮은 계절별 변화가 뚜렷하게 나타남이 보고되었다 (Lee *et al.*, 2016, 2007). 그리고 장기간 동안 관측된 AERONET 데이터 베이스를 이용한 통계분석에 관한 연구결과, 한반도 대기 중 분포하는 미세먼지 입자 크기가 과거에 비해 작아지고 있고 (Noh, 2021), 주로 비흡수성 AOD가 증가하는 추세가 나타나고 있음이 보고되었다 (Joo *et al.*, 2020). Lee (2018) 연구에서는 천리안 위성의 AOD 산출물 검증을 위해 지상 AERONET AOD 자료와 비교 분석한 결과, 최대 상관계수 (R) 값이 0.905로 유사성을 나타냈으나, 지역적인 한계가 있음을 보고하였다.

최근에는 다양한 머신러닝 기법을 적용하여 대기환경 자료를 예측하거나 분석하는 연구가 활발히 수행되고 있다. 예를 들면, Son *et al.* (2020)은 PM_{10}, $PM_{2.5}$ 농도 자료와 기상인자 (기온, 습도 등) 자료를 기반으로 다중 선형 회귀 기법 (Multiple Linear Regression Method), Support Vector Machine (SVM) 및 Random Forest (RF) 기법을 적용하여 PM_{10}, $PM_{2.5}$ 농도를 예측한 결과, 결정계수 (R^2)값이 0.772~0.929의 범위를 가지는 것으로 나타났다. 보다 최근에는 서울지역에서 관측된 PM 농도 자료를 하이브리드 머신러닝 모델로 예측한 결과, 기존의 모델보다 개선된 성능을 보였다 (Yang *et al.*, 2020). 또한, 그리스의 월평균 기온과 월 총 강수량의 시계열 자료를 기반으로 신경망 및 SVM 기법을 이용하여 기온과 강수량을 예측한 결과는 기온은 Root Mean Square Error (RMSE)값이 1.31로 높은 정확도를 보였지만, 강수량은 RMSE값이 47로 낮은 정확도를 보였다 (Papacharalampous *et al.*, 2018).

이와 같이 지상에서 지점 관측 기반의 PM 및 기상자료를 머신러닝 기법에 적용한 사례가 보고되고 있으나, 현재까지 머신러닝 기법을 적용하여 컬럼에어로솔의 예측을 수행한 연구 사례는 부족한 실정이다. 따라서, 본 연구는 대기 중 입자상 물질의 총 부하량인 컬럼에어로솔을 예측하고 평가하기 위하여 시계열 머신러닝 기법을 적용하였다. 이를 위하여, 첫째, 시계열 머신러닝을 위한 모델 수행 요소인 학습 기간 (Training Period), 회귀 모델 (Regression Model), 정규화 상수 (Regularization parameter)에 대한 조건별 모델링 결과를 비교 및 평가하여 가장 정확도가 높은 예측 모델을 결정하였다. 둘째, 최적의 모델링 기법을 이용한 컬럼 에어로솔의 예측값과 실제 관측한 관측값으로부터 모델링 결과에 대한 정확도 비교 및 검증을 수행하였다. 마지막으로 실제 관측한 AOD, AE, 단산란알베도 (Single Scattering Albedo, SSA), 입자크기분포함수 (Particle Size Distribution, PSD) 자료를 기반으로 클러스터링 기법에 적용하여 에어로솔의 유형을 분류하여 한반도에서 관측된 컬럼에어로솔의 특성을 파악하였다.

2. 자료 및 방법

2.1 관측 자료

연구 대상 지역은 대한민국에 위치한 총 28개의 AERONET 관측 지점 중 최소 3년 이상의 유효

관측 자료가 존재하는 6개의 관측 지점인 Anmyon (36.539°N, 126.330°E, 47 m agl.), Gangneung_WNU (37.771°N, 128.867°E, 60 m agl.), Gosan_SNU (33.300°N, 126.206°E, 72 m agl.), Gwangju_GIST (35.228°N, 126.843°E, 52 m agl.), Seoul_SNU (37.458°N, 126.951°E, 116 m agl.), Yonsei_University (37.564°N, 126.935°E, 97 m agl.) 지점을 포함한다(그림 1). 각 관측 지점에서는 Cimel사의 Sun-sky radiometer (모델명 CE-318)를 이용하여 대기 컬럼에어로솔이 관측되었으며, 본 연구에서 사용된 시계열 머신러닝 기법에 적용한 입력 자료는 AOD_500 nm, AE_440-675 nm 자료이고, 클러스터링 기법에는 AOD, AE, SSA, PSD 자료를 사용하였다. Sun-sky radiometer 관측 자료는 구름이나 다른 외부 요인에 의해 값이 비정상적인 값을 모두 제거하여 산출한 Level 2.0자료(download date: 2021.10.02.)를 AERONET 데이터 베이스(https://aeronet.gsfc.nasa.gov/)로부터 획득하였다. 자세한 관측 지점별 기간과 목록에 대한 상세 설명은 표 1 과 같다.

본 연구에서 사용된 관측 장비인 CE-318 Sun-sky radiometer는 8개의 중심 파장 대역(340 nm, 380 nm, 440 nm, 500 nm, 675 nm, 870 nm, 939 nm 및 1020 nm)에서 직사광 및 산란광을 측정하는 다중 파장 태양 분광광도계이다(Model: CE-318, CIMEL Electronique., France, https://www.cimel.fr/). CE-318

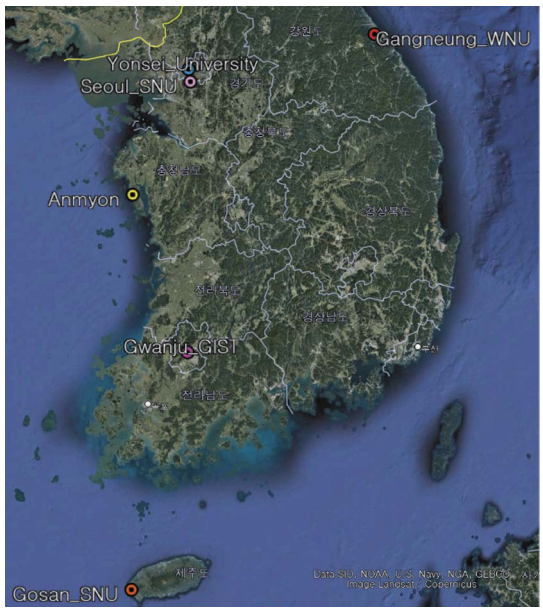

Fig. 1. Geographic locations of six Sun-sky radiometer measurement sites used in this study (Anmyon (36.539°N, 126.330°E, 47 m agl.), Gangneung_WNU (37.771°N, 128.867°E, 60 m agl.), Gosan_SNU (33.300°N, 126.206°E, 72 m agl.), Gwangju_GIST (35.228°N, 126.843°E, 52 m agl.), Seoul_SNU (37.458°N, 126.951°E, 116 m agl.), Yonsei_University (37.564°N, 126.935°E, 97 m agl.)).

Table 1. List of AERONET datasets used in this study.

Site	Period	Parameter (Level 2.0, daily average)	Number of data
Anmyon (36.539°N, 126.330°E)	2014.01.01~2020.12.31	Aerosol Optical Depth (AOD), Ångström Exponent (AE), Single Scattering Albedo (SSA), Particle Size Distribution (PSD) (VolCon-F, EffRad-F, VolMedianRad-F, StdDev-F, VolCon-C, EffRad-C, VolMedianRad-C, StdDev-C)	1510
Gangneung_WNU (37.771°N, 128.867°E)	2015.01.01~2018.12.31		1123
Gwangju_GIST (35.228°N, 126.843°E)	2011.01.01~2014.12.31		721
Gosan_SNU (33.300°N, 126.206°E)	2012.01.01~2016.12.31		609
Seoul_SNU (37.458°N, 126.951°E)	2016.01.01~2020.12.31		1237
Yonsei_University (37.564°N, 126.935°E)	2012.01.01~2020.12.31		2390

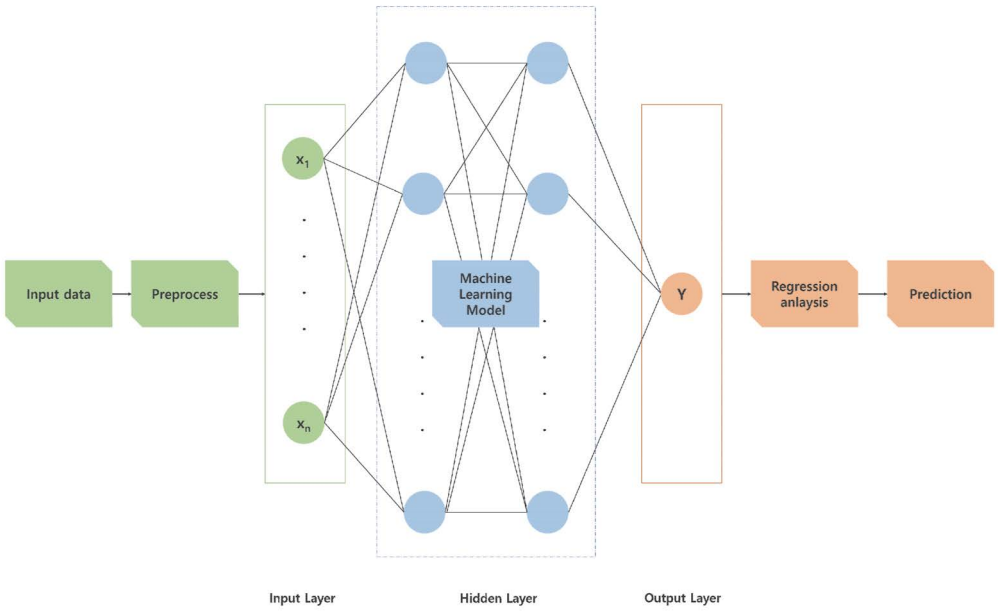

Fig. 2. Conceptual diagram of time series machine learning technique used in this study.

Sun-sky radiometer는 대기 광경로(Airmass)값이 7 미만의 낮 동안 직사광 또는 산란광을 여러 번 측정하도록 프로그래밍 되어 있다. 직사광 측정은 7개 파장 대역에서 이루어지며 레일리 산란(Rayleigh Scattering), 오존 및 기타 미량 기체의 흡수에 의한 투과도를 보정한 후, 식 (1, 2)와 같이 Beer-Bouguer 법칙을 사용하여 AOD를 산출한다(Giles *et al.*, 2019).

$$V(\lambda) = V_0(\lambda) \cdot d^2 \cdot \exp[-\tau(\lambda)_{Total} \cdot m] \quad (1)$$

$$\tau(\lambda)_{Aerosol} = \tau(\lambda)_{Total} - \tau(\lambda)_{Rayleigh} - \tau(\lambda)_{gas} \quad (2)$$

식 (1)에서 $V(\lambda)$는 파장(λ)에 의존하는 계측기의 측정된 전압이며, $V_0(\lambda)$는 파장에 의존하는 보정 계수이다. d는 지구-태양 거리에 대한 평균 비율(Michalsky, 1988), $\tau(\lambda)_{Total}$은 총 광학 두께, m은 대기의 상대 광경로(Kasas and Young, 1989)이다. 식 (2)에서 $\tau(\lambda)_{Aerosol}$, $\tau(\lambda)_{Rayleigh}$, $\tau(\lambda)_{gas}$는 각각 AOD, 레일리 산란에 의한 광학 두께, 그리고 가스성 물질 흡수에 의한 광학 두께이다. 또한 이 장비는 태양의 주 평면(태양 방위각과 동일한 고정 방위각으로 산란각을 변경하여 관측)을 따라 4개 파장 대역(440 nm, 675 nm, 870 nm 및 1020 nm)의 산란광을 측정함으로써, 에어로솔 크기 분포(입자 크기 범위 0.1~15 μm), 산란 위상 함수 및 굴절률을 역산하여 산출한다(Dubovik and King, 2000).

2.2 시계열 머신러닝

연구 대상 지역의 컬럼에어로솔 예측을 하기 위하여 머신러닝 기법 중 하나인 시계열 머신러닝을 사용하였다. 본 연구에서 시계열 머신러닝 분석은 R의 timetk 패키지를 이용하여 수행하였다. R은 무료로 개방된 통계패키지로서 통계 모형, 통계 계산 및 통계 그래픽 관련 라이브러리가 2021년 12월 24일 현재 18,653개가 존재한다(Jang, 2020). timetk 패키지는 과거 데이터의 추세를 분석하고, 분석한 추세를 바탕으로 미래의 값까지 예측할 수 있는 패키지이다. 그림 2는 본 연구에서 사용한 시계열 머신러닝 기법의 자료처리 과정을 설명하는 것이다. 먼저, 수집된 자료(input data)를 전처리(preprocess) 한 후, 신경망 내부에서 사용자가 지정한 회귀 기법을 이용하여 종속변

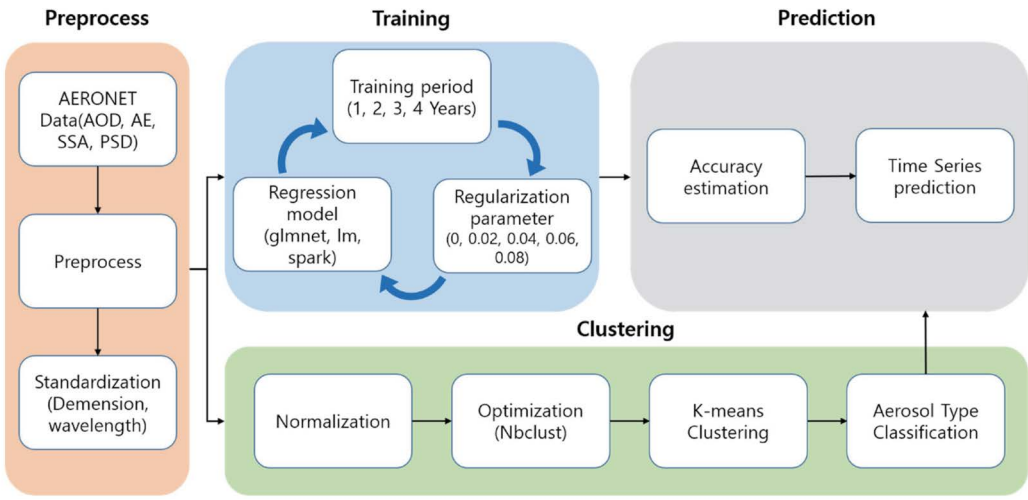

Fig. 3. Flowchart explaining the time series forecast of column aerosol measurement data in each modules.

수(AOD, AE)와 독립변수(시간)의 관계를 나타내는 함수식과 각 자료를 가장 잘 설명하는 해석 모델을 찾아, 함수식과 해석 모델을 토대로 다변량 입력값(x)에 따른 미래의 결과값(y)을 예측한다(Dancho *et al.*, 2021).

시계열 머신러닝을 위한 모델 수행 요소는 학습 기간, 회귀 모델, 정규화 상수 총 3가지이다. 따라서, 본 연구에서는 학습 기간이 모델에 미치는 영향을 파악하기 위하여 4가지 기간별(1, 2, 3, 4년) 조건을 지정하였다. 회귀 모델별 차이는 timetk 패키지의 엔진 모델에서 사용되고 있는 3가지 회귀 모델(glmnet, lm, spark)을 독립적으로 수행한 결과를 비교하였으며, 각 모델에 대한 상세 설명은 다음과 같다. 첫 번째, generalized linear model (glmnet)은 Penalized Maximum Likelihood (PML)를 통해 추정하는 일반화 선형 모델이다(Friedman *et al.*, 2010). 두 번째, linear regression model (lm)은 단순 선형 회귀 모델이다. 마지막으로 spark 모델은 분산 머신러닝 기법을 이용한 모델이며, 2011년 버클리대학의 AMP Lab에서 개발하였다(Venkataraman *et al.*, 2016). 마지막으로, 정규화상수는 0부터 0.08까지 0.02 간격으로 총 5개의 정규화상수를 지정한 후 시계열 머신러닝 기법을 적용한 결과를 비교하였다. 이렇게 제시된 입력 자료 및 조건별 시계열 머신러닝 기법을 이용하여 예측한 컬럼에어로솔값과 실제 관측값을 비교하여, 모델의 정확도를 비교 및 검증하였다.

3. 결과 및 토의

3.1 시계열 머신러닝 결과

그림 1에서 제시된 국내 6개 Sun-sky radiometer 관측 지점에서 관측된 월평균 컬럼에어로솔 자료를 이용한 시계열 그래프는 그림 4와 같다. 각 지점별 평균 AOD는 0.368±0.298 (Anmyon), 0.295±0.223 (Gangneung_WNU), 0.426±0.326 (Gwangju_GIST), 0.378±0.273 (Gosan_SNU), 0.399±0.327 (Seoul_SNU), 0.448±0.376 (Yonsei_University)으로서, 각 지점별 평균 AOD의 최대값과 최소값은 Yonsei_University와 Gangneung_WNU 지점이다. 또한, 각 지점별 평균 AE는 1.204±0.253 (Anmyon), 1.279±0.244 (Gangneung_WNU), 1.309±0.262 (Gwangju_GIST), 1.148±0.272 (Gosan_SNU), 1.256±0.262 (Seoul_SNU), 1.280±0.260 (Yonsei_University)이다. 각 지점

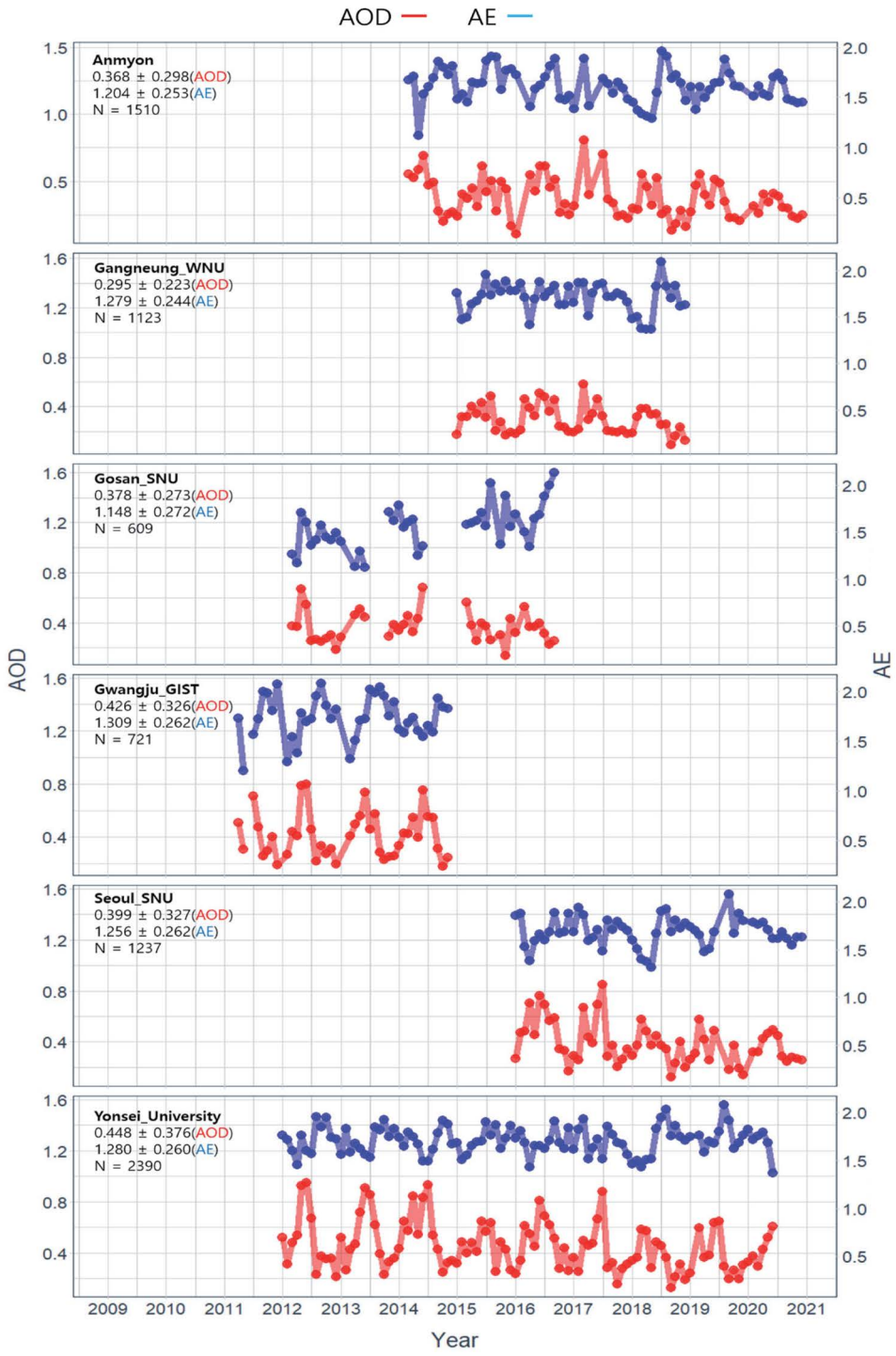

Fig. 4. Time series of the monthly mean AOD at 500 nm and AE at 440~675 nm observed at 6 selected sites.

Table 2. RMSE and bias derived from different input factors in the machine learning processes.

	Training period				Regression model			Regularization parameter				
	1yr	2yr	3yr	4yr	glmnet	lm	spark	0	0.02	0.04	0.06	0.08
RMSE	0.437	0.517	0.376	0.350	0.437	0.334	0.334	0.517	0.513	0.507	0.503	0.500
Bias	0.327	0.351	0.290	0.259	0.227	0.257	0.257	0.351	0.348	0.345	0.343	0.341

별 평균 AE의 최대값과 최소값은 Gwangju_GIST와 Gosan_SNU 지점이다. 관측 기간 중 유효 데이터의 개수(Ndata)는 Yonsei_University 지점이 2390개로 가장 많았고, Gosan_SNU 지점이 609개로 가장 적었다. 그리고 시계열 변화 경향은 6개 관측 지점 모두 대체로 여름에 AOD와 AE값이 높아지고, 겨울에 낮아지는 계절별 변화를 확인할 수 있고, 그림 4에서 보여진 계절별 변화는 Lee *et al.* (2007) 연구와 유사한 결과를 보인다. 이러한 계절별 변동성의 원인은 여름철 대기 중 입자상 물질의 부하량 증가 또는 대기 중 상대습도의 증가로 인한 입자의 흡습성장으로 인한 광산란 효과의 증가가 주요 원인으로 알려져 있다.

그림 5~7은 6개 관측 지점 중 유효 관측 자료가 2390개로 가장 많은 Yonsei_University의 자료를 대표 사례로 각 입력 요소 변화에 따른 분석 결과를 나타낸다. 먼저 그림 5는 학습 기간별 모델 예측 결과를 비교한 사례로서, 회귀 모델은 glmnet, 정규화 상수는 0으로 고정하고, 학습 기간을 1~4년 기간으로 각각 설정하여 Training (Blue area)하여, 학습 기간의 바로 다음 연도를 Test (Yellow area)기간으로 설정하였다. 이 사례의 경우, 학습 기간이 4년일 때 RMSE = 0.350, bias = 0.259로 4개의 학습 기간 중 가장 높은 정확도를 보였다. 반면 학습 기간이 2년일 때 RMSE = 0.517, bias = 0.351로 가장 정확도가 낮았다. 따라서, 이 경우에는 학습 기간이 4년인 경우에 최적의 예측 결과를 얻을 수 있게 되는 조건임을 확인할 수 있다.

그림 6은 회귀 모델별 결과를 비교하기 위하여, 서울 지점의 관측 자료를 기반으로 3가지 다른 회귀 모델(glmnet, lm, spark)을 이용하고, 학습 기간은 1년, 정규화 상수는 0으로 고정하여 각각 Training하여 비교한 사례이다. 이 사례에서는 lm, spark 모델을 사용한 예측 테스트 결과는 모두 RMSE = 0.334, bias = 0.257로 glmnet 모델을 사용한 결과값(RMSE = 0.437, bias = 0.227)보다 정확도가 높게 나타났다.

그림 7은 정규화 상수가 예측 결과에 미치는 정확도 분석결과를 비교한 사례이다. 이 사례는 정규화 상수 값을 0부터 0.08까지 0.02 간격으로 총 5개의 정규화 상수값을 적용하고, 학습 기간은 2년, 회귀 모델은 glmnet로 고정하여 Training 하였다. 이 사례의 경우, 정규화 상수값이 클수록 정확도가 더 높아짐을 알 수 있다. 정규화 상수가 0.08일 때 RMSE = 0.500, bias = 0.341로 가장 정확도가 높은 결과를 보였고, 정규화 상수가 0일 때 RMSE = 0.517, bias = 0.351 정확도가 가장 낮았다.

3.2 모델별 정확도 비교 및 검증

그림 8에서는 이전 장에서 제시된 사례별 분석방법을 토대로, 6개의 관측 지점별 입력 요소를 시계열 머신러닝 기법에 적용한 결과 중 가장 높은 정확도와 가장 낮은 정확도를 나타낸 사례를 비교하였다. 먼저 전체 관측 자료에 대한 AOD값의 예측과 관측값을 비교한 경우는 그림 8(a), (c)의 결과와 같다. 가장 높은 정확도를 보이는 그림 8(a)는 모델의 학습 기간이 1년, 정규화 상수가 0.08일 때 RMSE = 0.282, bias = 0.199로 전체 관측값의 표준편차(σ = 0.319)보다 값이 작아 유효한 결과로 판단되었다. 그러나, 그림 8(c)와 같이 학습 기간이 4년, 정규화 상수가 0일 때의 예측 결과는 RMSE = 0.338, bias = 0.238로 관측값의 표준편차(σ = 0.319)보다 RMSE 값이 높게 나타나 유효하지 않은 결과로 판단할 수 있으며, 예측 정확도 값도 가장 낮았다.

그림 8(b), (d)는 AE의 예측값과 관측값을 비교한

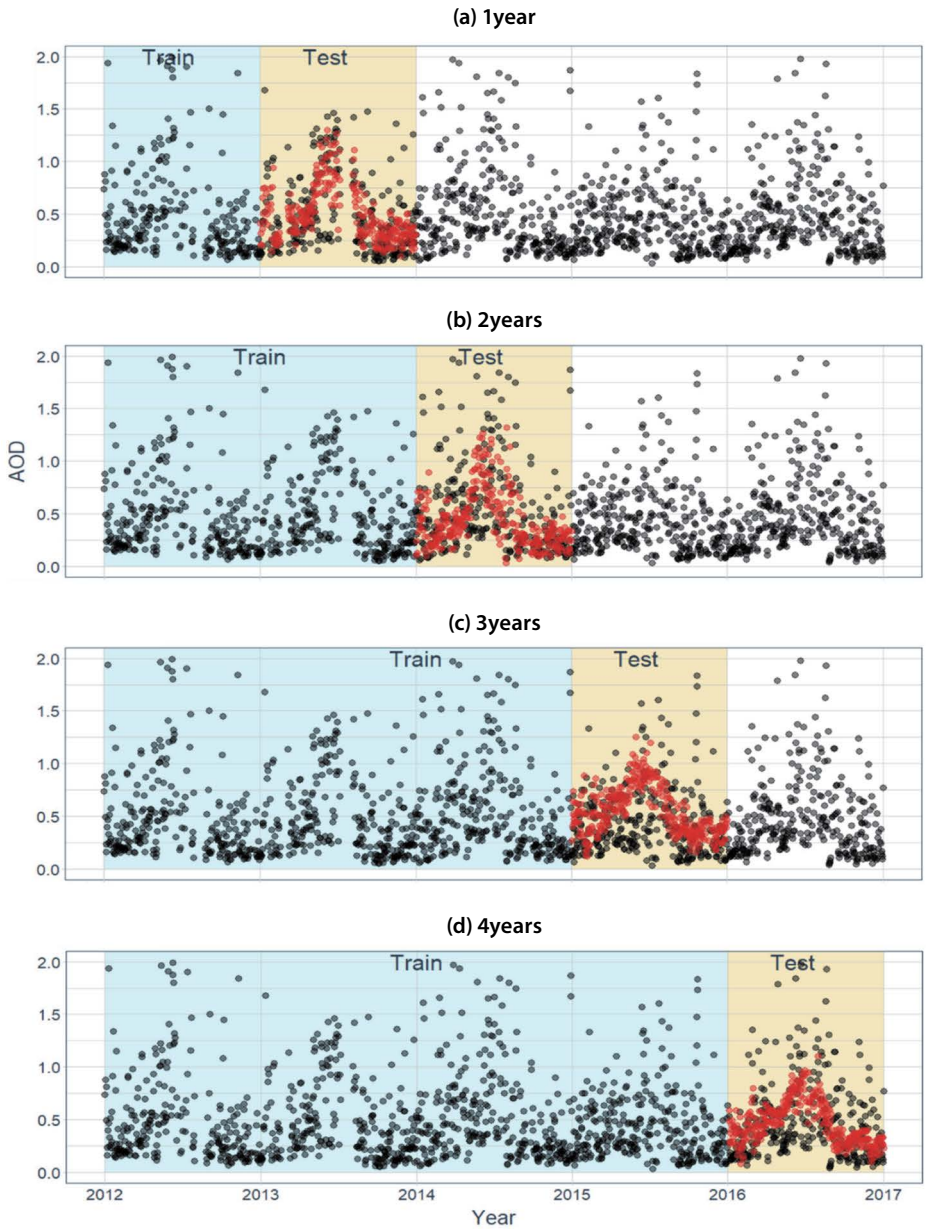

Fig. 5. Comparison of machine learning prediction results by four different training periods (a) training period of 1year, (b) training period of 2years, (c) training period of 3years, (d) training period of 4years. Symbols and colors represent total observations (black dots), learning period (blue regions), and test periods (yellow regions), and test results (red dots), respectively.

결과이다. AOD의 분석결과와 마찬가지로 학습 기간이 1년, 정규화 상수가 0.08일 때 가장 높은 예측 정확도를 보였으며, 이때 RMSE = 0.246, bias = 0.189로 관측값의 표준편차 ($\sigma = 0.262$)보다 값이 작아 유효한 결과로 나타났다. 반대로 학습 기간이 2년, 정규화 상수가 0일 때 RMSE = 0.490, bias = 0.301로 두 값 모두

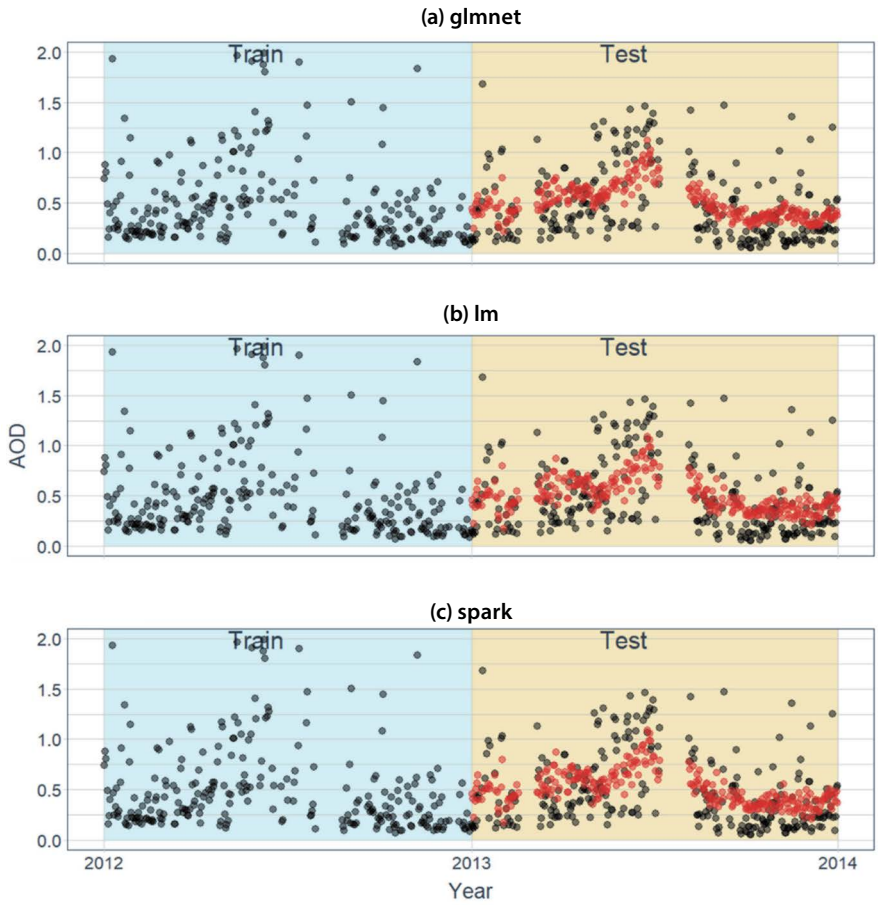

Fig. 6. Comparison of machine learning prediction results by three different regression models (a) glmnet, (b) lm, (c) spark. Symbols and colors represent total observations (black dots), training period (blue regions), and test periods (yellow regions), and test results (red dots), respectively.

관측값의 표준편차(σ=0.262)보다 큰 값을 가지므로 유효하지 않은 결과이며, 정확도가 제일 낮았다. 이러한 결과를 토대로 AOD와 AE 모두 큰 값의 정규화 상수 값으로 정규화를 진행하여 이상치를 최소화할 때 정확도가 높고, 관측값은 전년과 비슷한 추세를 보이는 것을 의미하며 본 연구에 적용된 시계열 머신러닝 기법은 장기간보다 단기간 예측에 더 효율적일 것으로 판단된다. 또한, 예측값과 관측값의 비교를 통해서 예측값의 과대 및 과소 추정 여부까지 확인할 수 있다. 예를 들어 그림 8(a)는 y=x (검은색 실선)보다 x축 방향으로 더 많이 분포하기 때문에 상대적으로 예측값이 관측값보다 과대추정되는 것을 알 수 있다. 그림 8(b)는 그림 8(a)와 반대로 검은색 실선보다 y축 방향으로 더 많이 분포하기 때문에 상대적으로 예측값이 관측값보다 과소추정되는 것을 알 수 있다.

그림 9는 각 입력 요소를 시계열 머신러닝 기법에 적용하여 나온 결과를 세 가지의 입력 요소별로 RMSE와 bias를 나타낸 그림이다. 학습 기간의 경우 AOD와 AE 모두 학습 기간이 1년일 때 RMSE와 bias가 낮아 제일 높은 정확도를 보였다. AOD는 학습 기간이 길어질수록 정확도가 다소 떨어지는 추세를 보였지만, AE의 경우 학습 기간이 2년일 때 가장 정확

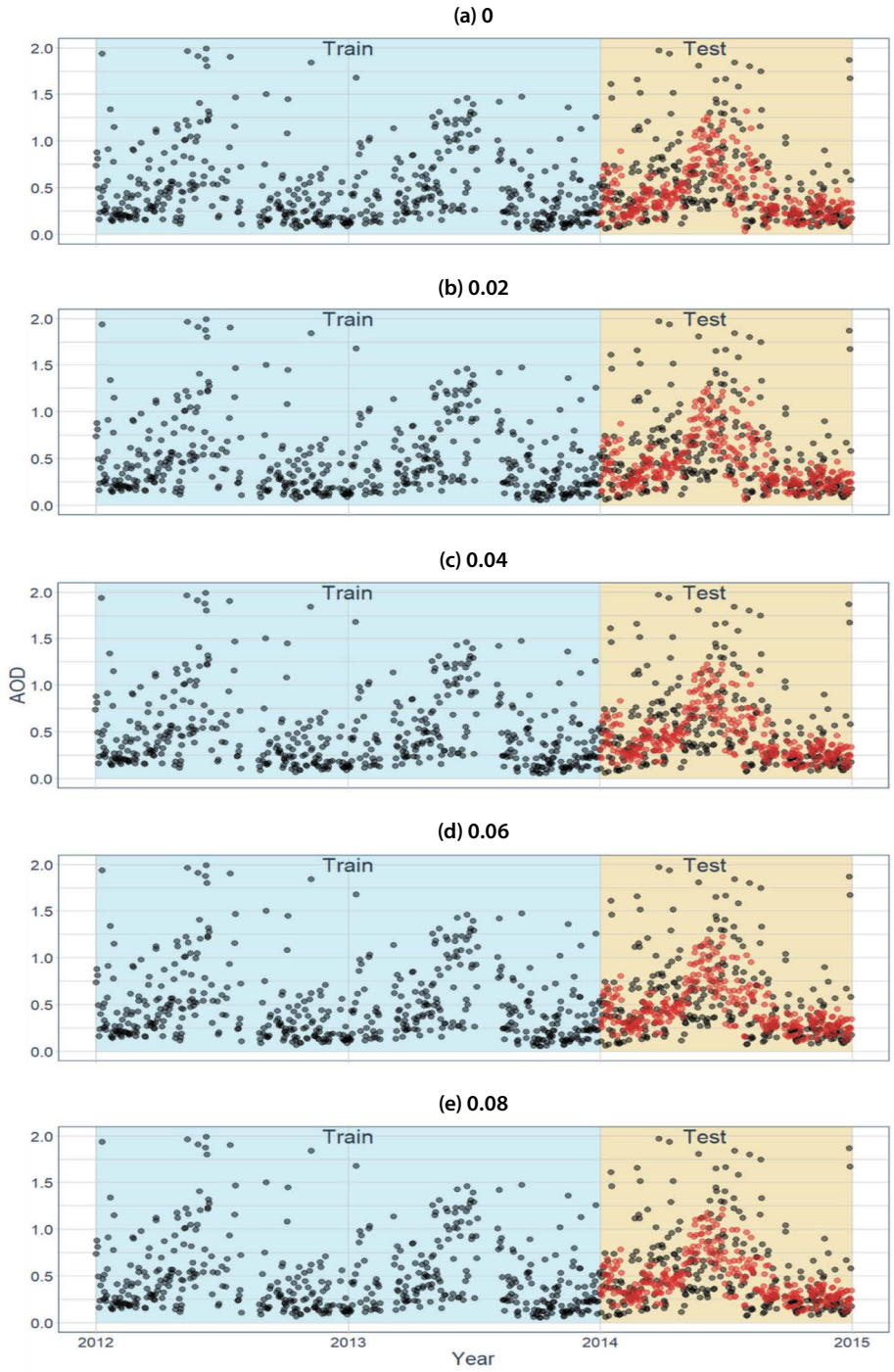

Fig. 7. Comparison of machine learning prediction results by five different regularization parameters (a) regularization parameter of 0, (b) regularization parameter of 0.02, (c) regularization parameter of 0.04, (d) regularization parameter of 0.06 and (e) regularization parameter of 0.08. Symbols and colors represent total observations (black dots), training period (blue regions), and test periods (yellow regions), and test results (red dots), respectively.

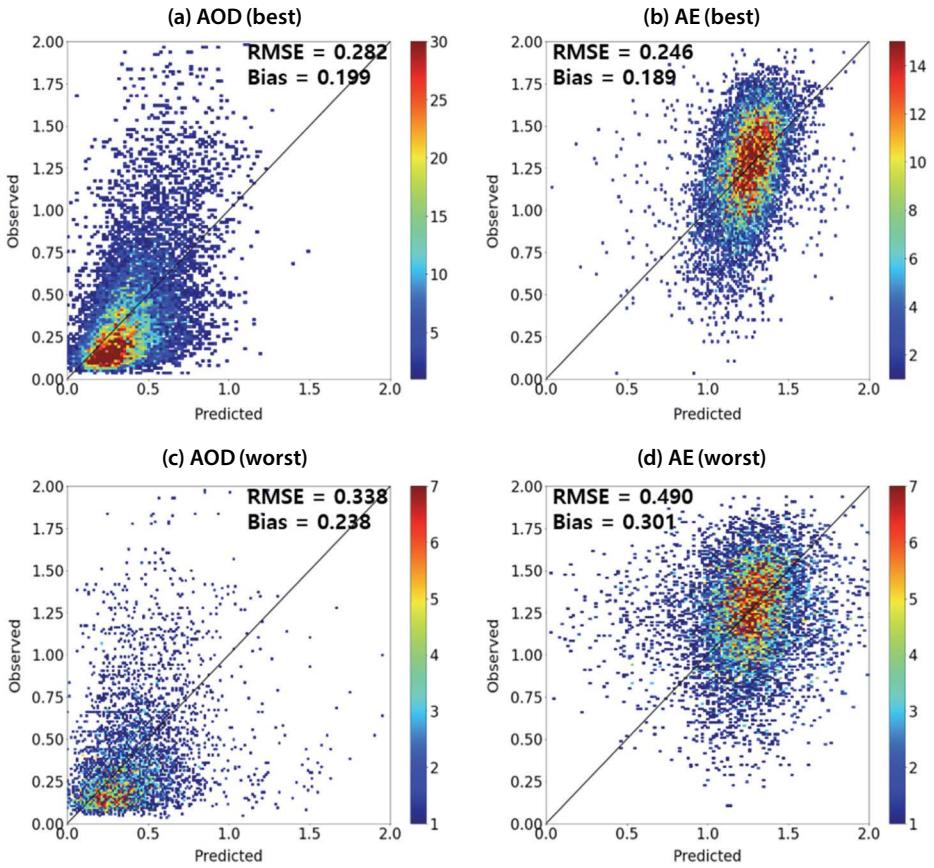

Fig. 8. Scatterplots of model predicted and observed data for (a) AOD (best): training period of 1 year, regularization parameter of 0.08, (b) AE (best): training period of 1 year, regularization parameter of 0.08 and (c) AOD (worst): training period of 4 years, regularization parameter of 0, (d) AE (worst): training period of 2 years, regularization parameter of 0.

도가 낮고 기간이 길어질수록 정확도가 조금씩 높아진다. 회귀 모델에 따른 비교 결과, AOD는 spark 모델일 때 가장 높은 정확도를 보였으며, AE는 glmnet 모델일 때 가장 높은 정확도를 보였다. 반면에 AOD와 AE 모두 lm 모델일 때 정확도가 가장 낮았다. 정규화 상수는 AOD와 AE 모두 정규화 상수가 클수록 정확도가 높아져서 0.08일 때 가장 높은 정확도를 보이지만, 정확도가 가장 낮은 0일 때와 큰 차이를 가지진 않았다.

3. 3 머신러닝 클러스터링 결과

국내 대기 중 컬럼에어로솔 타입을 분류하기 위해 AOD, AE, SSA 그리고 PSD 자료를 기반으로 머신러닝 클러스터링 기법을 적용하였다. SSA는 입자의 광흡수 비율을 나타내는 변수이며, 입자의 성분별 특징에 의하여 영향을 받게 된다. SSA는 AERONET에서 기하조건별 관측을 통하여 수치적으로 계산한 값이다(오차범위 < 0.01). 그러나, AOD가 관측되더라도 SSA를 산출하기 위하여 필요한 유효 데이터의 수가 부족한 경우에는 SSA 산출이 불가능하므로(구름, 태양천정각 > 50°, AOD440nm < 0.4), 일반적으로 SSA의 유효 데이터 개수는 AOD 데이터에 비하여 적다(Holben *et al.*, 2006). 이와 같은 이유로 클러스터링을 진행할 때 SSA를 포함한 자료와 포함하지 않는 자료

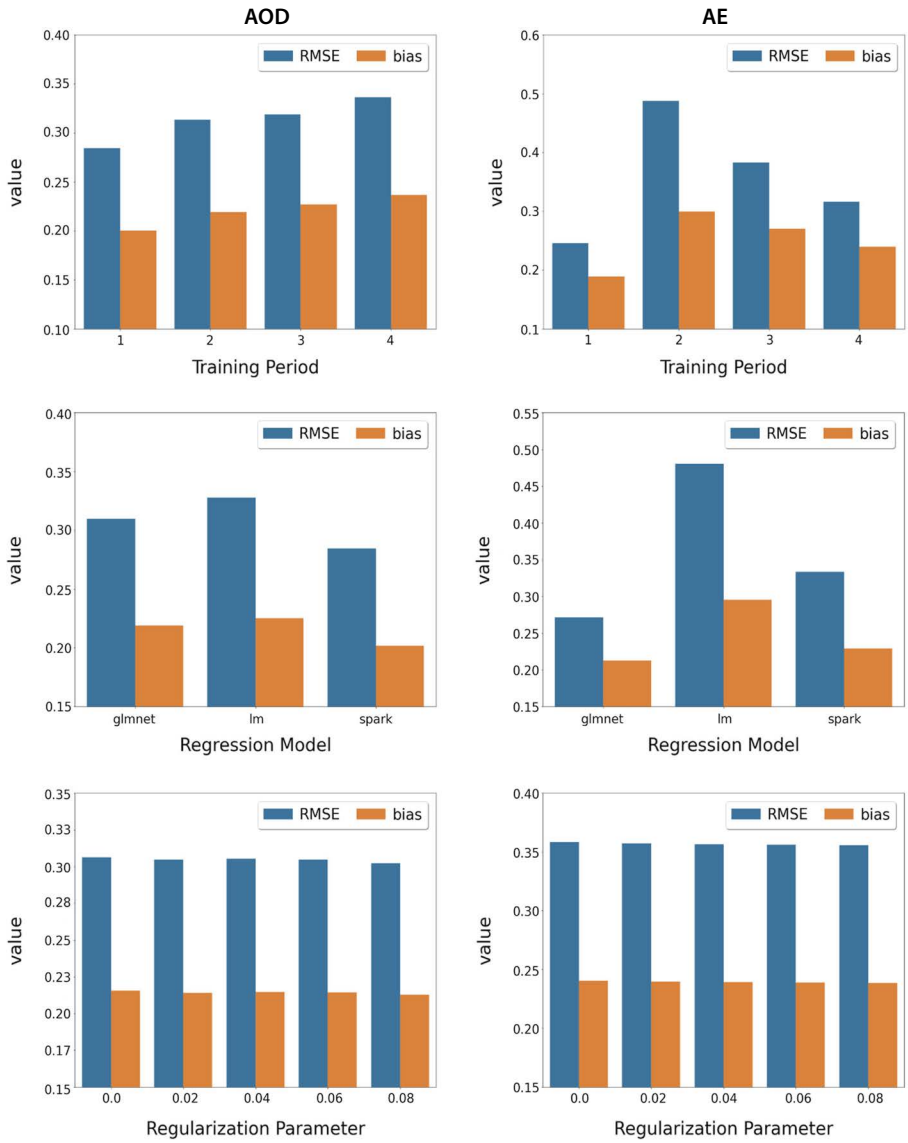

Fig. 9. Comparisons of RMSE and bias values regarding model performance conditions by training periods, regression models, regularization parameter.

로 나누어 수행되었다.

본 연구에서 사용된 클러스터링 기법은 K-Means 클러스터링으로, 군집 분류에 앞서 결과물의 클러스터 숫자를 지정해야 한다. 클러스터링에 필요한 최적의 클러스터 숫자를 결정하기 위하여 R의 NbClust 패키지 (Charrad *et al.*, 2014)가 사용되었으며, NbClust 패키지는 입력값으로 사용된 데이터로부터 30가지의 클러스터링 테스트를 수행하여 각 클러스터링 방법을 통해 생성된 결과로 최적의 클러스터 수를 선정한다. 따라서 30개의 테스트 결과 중 가장 많은 스코어를 획득한 클러스터 개수를 선택하여 K-Means 클러스터링을 수행하였다. 그림 10은 NbClust 패키지를

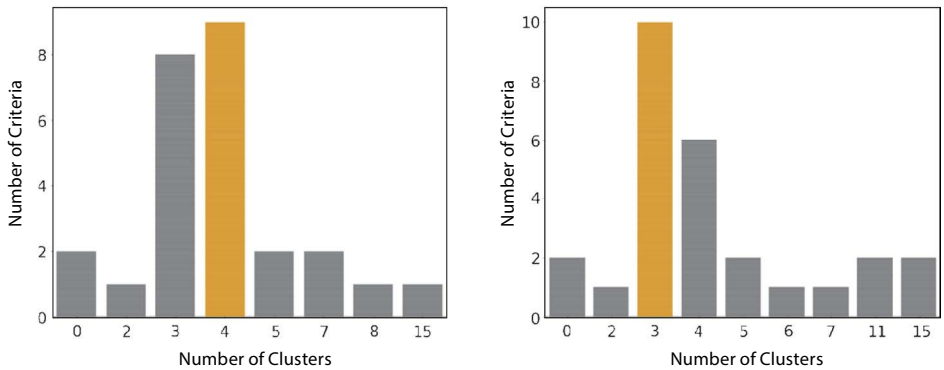

Fig. 10. The optimal number of clusters in (left) with SSA and (right) without SSA determined from the NbClust package.

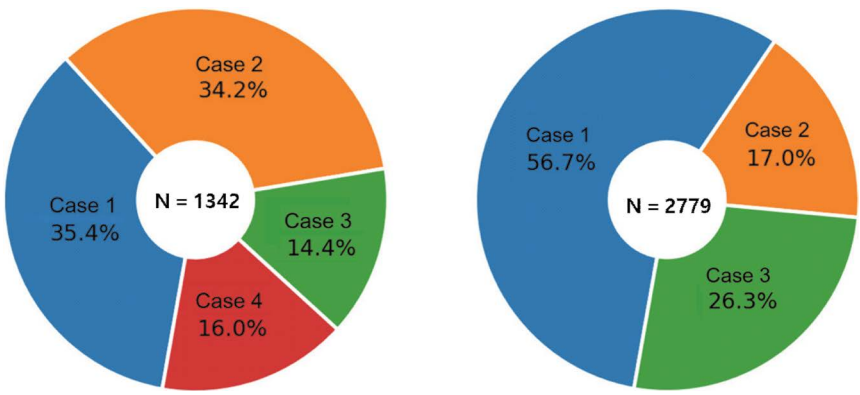

Fig. 11. Machine learning clustering results using data with and without SSA.

이용한 최적 클러스터링 테스트 결과이다. SSA를 포함한 자료는 30가지 계산식 중 9개 계산식에 선정을 받아 총 4개의 Case로 분류하였고, 데이터 개수는 1342개이다. SSA를 포함하지 않는 총 데이터 개수는 2779로 SSA를 포함하였을 때보다 약 2배가 많은 수치이며, 최적 클러스터링 테스트 결과 30가지 계산식 중 10개 계산식에 선정을 받아 총 3개의 Case로 분류하였다. 그림 11은 최적 Case 수를 적용하여 K-Means 클러스터링을 수행한 결과이다. SSA를 포함하였을 때는 Case 1이 35.4%로 가장 높은 비율을 차지했고, Case 3이 14.4%로 가장 낮은 비율을 차지하였다. SSA를 포함하지 않았을 때는 Case 1이 56.7%로 가장 높은 비율을 차지했으며, Case 2가 17.0%로 가장 낮은 비율을 차지했다.

그림 12는 SSA를 포함하였을 때 각 파장별, SSA 분포와 각 Case별 크기 분포를 나타낸 그림이다. 파장별, Case별 SSA 결과는, Case 1, 4는 도시/산업(Urban/Industrial, UI) 에어로솔 유형, Case 2는 탄소성(Carbonceous; Cb) 에어로솔 유형, Case 3은 황사/먼지(Desert dust; DD) 에어로솔 유형과 유사하게 나타났다. Case별 크기 분포 또한, 파장별, Case별 SSA 수치와 같은 결과가 나왔다. 따라서, SSA를 포함하였을 때 우리나라는 UI 에어로솔 유형이 51.4%로 가장 빈번하게 발생하였고, DD 유형이 14.4%로 가장 적게 발생하였다.

그림 13은 SSA를 포함하지 않았을 때 파장별, Case

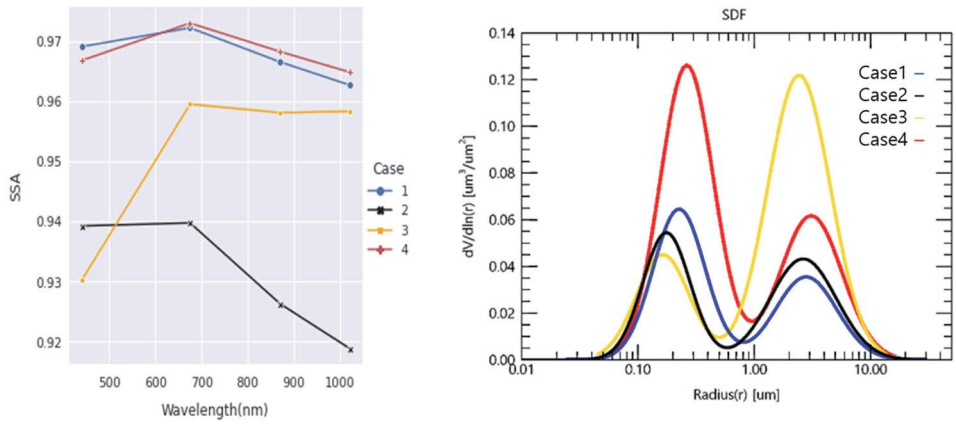

Fig. 12. Classified SSA for each wavelength and size distribution.

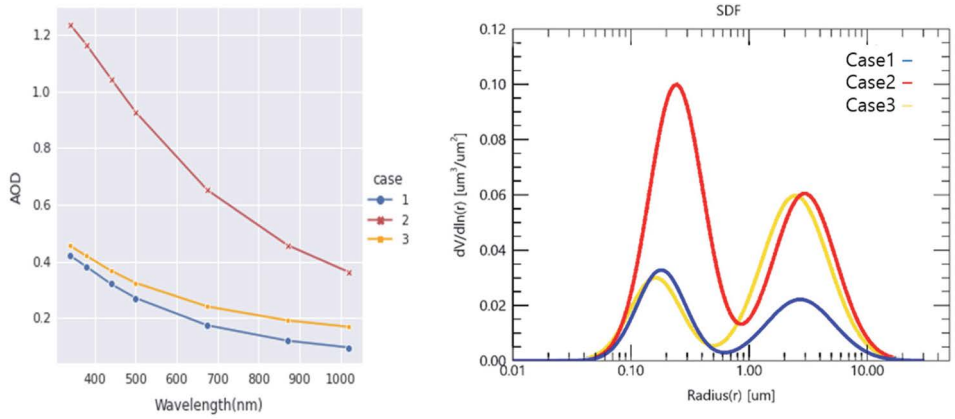

Fig. 13. Classified AOD for each wavelength and size distribution.

별 AOD 분포와 각 Case별로 크기 분포를 나타낸 그림이다. Case 1, 2 유형은 UI 에어로솔 유형, Case 3 유형은 DD 유형과 유사하다. 따라서 SSA를 포함하지 않았을 때 우리나라는 UI 에어로솔 유형이 63.7%로 가장 빈번하게 발생한다. Case 1, 2는 같은 UI 에어로솔 유형이지만 크기 분포에서 AOD 절대값의 차이로 인해 큰 차이를 보이고 있다. 또한, SSA를 포함하였을 때와는 달리 Cb 유형이 산출되지 않았으며, 이로 인해 Cb 유형과 같은 강한 광흡수성 물질에 대한 특성을 고려해야 하는 경우에는 SSA를 포함하여 클러스터링을 수행해야 한다는 것을 의미한다.

4. 결 론

컬럼에어로솔은 대기 중 입자상 물질의 총 부하량을 나타낸다. 또한, 컬럼에어로솔은 대기 중 에어로솔의 유형과 특성을 알 수 있기 때문에 대기 에어로솔 연구에 있어 굉장히 중요하다. 본 연구에서는 전국의 6개 지점(Anmyon, Gangneung_WNU, Gosan_SNU, Gwangju_GIST, Seoul_SNU, Yonsei_University)의 AERONET Sun-sky radiometer 컬럼에어로솔 자료를 이용하였다. 컬럼에어로솔 자료를 머신러닝 기법(시계열, 클러스터링)에 적용하였고, 대기 중 입자상 물

질의 총 부하량인 컬럼에어로솔을 예측하고 평가하여 다음과 같은 결론을 도출하였다.

첫째, 시계열 머신러닝 기법 중 세 가지 입력 요소인 학습 기간, 회귀 모델, 정규화 상수를 Test한 결과 AOD의 경우 학습 기간이 1년, 회귀 모델은 spark 모델, 정규화 상수는 0.08일 때 가장 정확한 결과를 보였고, AE의 경우 학습 기간이 1년, 회귀 모델은 glmnet 모델, 정규화 상수는 0.08일 때 가장 정확한 결과를 보였다.

둘째, 각 변수에 대한 최적의 입력 요소의 정확도를 비교한 결과, AOD의 경우 학습 기간이 1년일 때 RMSE = 0.285, bias = 0.200이고, 회귀 모델이 spark 모델일 때 RMSE = 0.284, bias = 0.202이다. 정규화 상수가 0.08일 때 RMSE = 0.302, bias = 0.213로 유효한 결과임을 입증하였다. AE의 경우 학습 기간이 1년일 때 RMSE = 0.246, bias = 0.189이고, 회귀 모델이 glmnet 모델일 때, RMSE = 0.271, bias = 0.212이다. 정규화 상수가 0.08일 때 RMSE = 0.356, bias = 0.238로 AOD와 마찬가지로 유효한 결과임을 입증하였다. 이러한 결과는, 시계열 머신러닝 기법은 컬럼에어로솔에 대해서 장기간 예측보다는 단기간 예측에 우수함을 의미한다.

셋째, SSA를 포함한 자료를 이용하여 머신러닝 클러스터링을 수행한 결과, Cb 유형과 유사한 Case가 나왔지만, SSA를 포함하지 않았을 때는 Cb 유형과 유사한 Case가 발견되지 않았다. 이는 Cb 유형과 같은 강한 광흡수성 물질의 거동을 분석할 때는 SSA값의 사용 유무가 모델링 결과에 크게 작용하는 것을 의미한다. 또한, 한반도에서 관측된 컬럼에어로솔의 클러스터링 분석결과는 총 3가지 Case로 분류되었다. 그중 Case 1, 2는 UI 에어로솔 유형과 유사하며 전체에서 약 63.7% 비율을 차지하였다. Case 3은 DD 에어로솔 유형과 유사하며 약 26.3% 비율을 차지하였다. 본 연구에서 사용된 관측 지점별 총 데이터 개수는 상이하지만 관측 기간이 충분히 길고(지점별 관측 기간이 최소 4년 이상), 각 관측 지점의 지리적 위치가 한반도 전역에 분포하고 있으므로 한반도 컬럼 대기의 특성을 대표할 수 있을 것으로 판단된다. 따라서 한반도에서는 황사나 미세먼지보다는 자동차 배기가스나 공장의 매연으로 인한 UI 유형이 보다 빈번하게 발생함을 알 수 있었다.

본 연구의 방법론과 결과는 단기간 에어로솔 및 미세먼지 예측 모델의 검증과 원격탐사 자료를 이용하여 공간적인 분석에 활용할 수 있을 것이라 판단된다.

감사의 글

이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2019R1I1A3A01062804). 본 연구에서 사용된 AERONET 관측 지점(Anmyon, Gangneung_WNU, Gosan_SNU, Gwangju_GIST, Seoul_SNU, Yonsei_University) 사이트를 구축, 유지 관리 및 자료 제공해 주신 연구자분들께 감사드립니다.

References

Charlson, R.-J., Schwartz, S.-E., Hales, J.-M., Cess, R.-D., Coakley Jr, J.-A., Hansen, J.-E., Hofmann, D.-J. (1992) Climate Forcing by Anthropogenic Aerosols, Science, 255(5043), 423-430. https:// doi.org/10.1126/science.255.5043.423

Charrad, M., Chazzali, N., Boiteau, V., Niknafs, A. (2014) NbClust: An R Package for Determining the Relevant Number of Clusters in A Data Set, Journal of Statistical Software, 61(6), 1-36. https://doi.org/10.18637/jss.v061.i06

Dancho, M., Vaughan, D., Dancho, M.-M. (2021) timetk: A Tool Kit for Working with Time Series in R, R package 'timetk' version 2.6.2. https://cran.r-project.org/web/packages/timetk/index.html

Dubovik, O., King, M.-D. (2000) A Flexible Inversion Algorithm for Retrieval of Aerosol Optical Properties from Sun and Sky Radiance Measurements, Journal of Geophysical Research, 105, 20673-20696. https://doi.org/10.1029/2000JD900282

Friedman, J., Hastie, T., Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models Via Coordinate Descent, Journal of Statistical Software, 33(1), 1-4, https://doi.org/10.1163/ej.9789004178922.i-328.7

Giles, D.-M., Sinyuk, A., Sorokin, M.-G., Schafer, J.-S., Smirnov, A., Slutsker, I., Eck, T.-F., Holben, B.-N., Lewis, J.-R., Campbell, J.-R., Welton, E.-J., Korkin, S.-V., Lyapustin, A.-I. (2019) Advancements in the Aerosol Robotic Network (AERONET) Version 3 Database-automated Near-real-time Quality Control Algorithm with Improved Cloud Screening for Sun Photometer Aerosol Optical Depth (AOD) Measurements, Atmospheric Measurement Techniques, 12(1), 169-209. http://doi.org/10.5194/amt-12-169-2019

Hansen, J., Sato, M., Ruedy, R. (1997) Radiative Forcing and Climate Response, Journal of Geophysical Research: Atmospheres, 102(D6), 6831-6864. https://doi.org/10.1029/96JD03436

Holben, B.-N., Tanré, D., Smirnov, A., Eck, T.-F., Slutsker, I., Abuhassan, N., Newcomb, W.-W., Schafer, J.-S., Chatenet, B., Lavenu, F., Kaufman, Y.-J., Castle, J.-V., Setzer, A., Markham, B., Clark, D., Frouin, R., Halthore, R., Karneli, A., O'Neill, N.T., Pietras, C., Pinker, R.-T., Voss, K., Zibordi, G. (2001) An Emerging Ground-based Aerosol Climatology: Aerosol Optical Depth from AERONET, Journal of Geophysical Research: Atmospheres, 106(D11), 12067-1209. https://doi.org/10.1029/2001JD900014

Holben, B., Eck, T., Slutsker, I., Smirnov, A., Sinyuk, A., Schafer, J., Giles, D., Dubovik, O. (2006) AERONET's Version 2.0 Quality Assurance Criteria, Proceedings of SPIE Asia-Pacific Remote Sensing, 6408, 64080Q-64080Q14

International Panel on Climate Change (IPCC) (2013) Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. https://www.ipcc.ch/site/assets/uploads/2017/09/WG1AR5_Frontmatter_FINAL.pdf

Jang, D.-H. (2020) Applications of R Package for Statistical Engineering, The Korean Journal of Applied Statistics, 33(1), 87-105, (in Korean with English abstract). https://doi.org/10.5351/KJAS.2020.33.1.087

Joo, S., Dehkhoda, N., Noh, Y. (2020) A Study on the Characteristic and AOD Variation according to Aerosol Types Using AERONET Sunphotometer Data in Korea, Korean Journal of Remote Sensing, 36(2-1), 93-101, (in Korean with English abstract). https://doi.org/10.7780/kjrs.2020.36.2.1.1

Kasten, F., Young, A.-T. (1989) Revised Optical Air Mass Tables and Approximation Formula, Applied Optics, 28(22), 4735-4738. https://doi.org/10.1364/AO.28.004735

Kim, B.-G., Kim, Y.-J., Eun, S.-H. (2008) An Analysis of Aerosol Optical Properties Around Korea using AERONET, Journal of Korean Society for Atmospheric Environment, 24(6), 629-640, (in Korean with English abstract). https://doi.org/10.5572/KOSAE.2008.24.6.629

Lee, K.-H., Kim, Y.-J., Hoyningen-Huene, W., Burrow, J.-P. (2007) Spatio-Temporal Variability of Satellite-Derived Aerosol Optical Thickness over Northeast Asia in 2004, Atmospheric Environment, 41(19), 3959-3973. https://doi.org/10.1016/j.atmosenv.2007.01.048

Lee, K.-H. (2012) Impact of Northeast Asian Biomass Burning Activities on Regional Atmospheric Environment, Journal of the Korean Association of Geographic Information Studies, 15(1), 184-196, (in Korean with English abstract). https://doi.org/10.11108/kagis.2012.15.1.184

Lee, K.-H., Lee, K.-T. (2016) Aerosol Optical Thickness Measurements from the Microtops-II Multi-wavelength Radiometer, Journal of Korean Society for Atmospheric Environment, 32(1), 57-66, (in Korean with English abstract). https://doi.org/10.5572/KOSAE.2016.32.1.057

Lee, K.-H. (2018) Validation of COMS/MI Aerosol Optical Depth Products Using Aerosol Robotic Network (AERONET) Observations Over East Asia, Korean Journal of Remote Sensing, 34(3), 507-517, (in Korean with English abstract). https://doi.org/10.7780/kjrs.2018.34.3.6

Michalsky, J.-J. (1988) The Astronomical Almanac's Algorithm for Approximate Solar Position (1950-2050), Solar Energy, 40(3), 227-235. https://doi.org/10.1016/0038-092X(88)90045-X

Noh, Y. (2021) A Long-term Trend of Annual Average Aerosol Optical Properties in the Korean Peninsula using AERONET Sun/Sky Radiometer Data, Journal of Korean Society for Atmospheric Environment, 37(3), 456-465, (in Korean with English abstract). https://doi.org/10.5572/KOSAE.2021.37.3.456

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D. (2018) Univariate Time Series Forecasting of Temperature and Precipitation with a Focus on Machine Learning Algorithms: A Multiple-case Study from Greece, Water Resources Management, 32(15), 5207-5239. https://doi.org/10.1007/s11269-018-2155-6

Park, S.-S., Kim, S.-J., Gong, B.-J., Cho, S.-Y., Kim, J.-C., Lee, S.-J. (2013) Investigation on a Haze Episode of Fine Particulate Matter using Semi-Continuous Chemical Composition Data, Journal of Korean Society for Atmospheric Environment, 29(5), 642-655, (in Korean with English abstract). https://doi.org/10.5572/KOSAE.2013.29.5.642

Park, S. (2020) COVID-19 (Coronavirus Disease 2019) Outbreaks and Their Relationship with Atmospheric Concentrations of PM_{10} and $PM_{2.5}$: A Case Study for Daegu Metropolitan City, Journal of the Korean Geographical Society, 55(5), 453-465, (in Korean with English abstract). https://doi.org/10.22776/kgs.2020.55.5.453

Son, S., Kim, J. (2020) Evaluation and Predicting PM_{10} Concentration Using Multiple Linear Regression and Machine Learning, Korean Journal of Remote Sensing, 36(6-3), 1711-1720, (in Korean with English abstract). https://doi.org/10.7780/kjrs.2020.36.6.3.7

Venkataraman, S., Yang, Z., Liu, D., Liang, E., Falaki, H., Meng, X., Xin, R., Ghodsi, A., Franklin, M., Stoica, I., Zaharia, M. (2016) Sparkr: Scaling R Programs with Spark. Proceedings of the 2016 International Conference on Management of Data, 1099-1104.

Yang, G., Lee, H., Lee, G. (2020) A Hybrid Deep Learning Model to Forecast Particulate Matter Concentration Levels in Seoul, South Korea, Atmosphere, 11(4), 348. https://doi.org/10.3390/atmos11040348

Authors Information

김영일 (강릉원주대학교 공간정보협동과정 석사과정)
(kyi3619@gmail.com)

이권호 (강릉원주대학교 대기환경과학과 교수)
(kwonho.lee@gmail.com)

이규태 (강릉원주대학교 대기환경과학과 교수)
(ktlee@gwnu.ac.kr)