

Research Paper

Deep Neural Network를 활용한 초미세먼지 농도 예측에 관한 연구

최인호 · 이원영 · 은범진 · 허정숙 · 장광현 · 오종민

경희대학교 환경학 및 환경공학과

A Study on Prediction of PM_{2.5} Concentration Using DNN

Inho Choi · Wonyoung Lee · Beomjin Eun · Jeongsook Heo ·
Kwang-Hyeon Chang · Jongmin Oh

Department of Environmental Science and Engineering, Kyung Hee University

요약: 본 연구는 국가측정망(에어코리아)에서 제공하는 2017년, 2019년 및 2020년도 대기질확정 데이터를 이용하여 Deep Neural Network(DNN) 모델을 학습하고, 2016년과 2018년도 데이터를 이용하여 학습된 모델을 평가·검증하였다. 피어슨 상관계수 0.2를 기준으로 SO₂, CO, NO₂, PM₁₀ 항목을 독립변수로 하여 초기 모델링을 진행하였고, 예측의 정확도를 높이기 위한 방법으로 시계열적 요소를 반영한 월별 모델링(개선모델)을 진행하여 초기모델과 비교·분석하였다. 분석에 사용한 지표는 RMSE(Root mean square error) 방법으로 오차를 계산하였으며, 예측 결과 초기모델의 RMSE값은 5.78로 국가측정망의 예측이동 평균모델의 결과(10.77)와 비교하여 초기모델에서 약 46% 오차가 감소하였다. 또한, 개선모델의 경우, 초기모델 대비 11월 모델을 제외한 모든 월별모델에서 정확도 향상이 있었다. 따라서, 본 연구에서는 DNN 모델링이 PM_{2.5} 농도 예측에 효과적인 방법임을 제안할 수 있었으며, 향후 추가적인 독립변수 선정 및 시계열 요소를 고려한 방법으로 모델의 정확도 개선 가능성을 확인할 수 있었다.

주요어: 초미세먼지, 대기오염물질, 기계학습, 예측모델, DNN모델

Abstract: In this study, DNN-based models were learned using air quality determination data for 2017, 2019, and 2020 provided by the National Measurement Network (Air Korea), and this models evaluated using data from 2016 and 2018. Based on Pearson correlation coefficient 0.2, four items (SO₂, CO, NO₂, PM₁₀) were initially modeled as independent variables. In order to improve the accuracy of prediction, monthly independent modeling was carried out. The error was calculated by RMSE (Root Mean Square Error) method, and the initial model of RMSE was 5.78, which was

First Author: Inho Choi, E-mail: 2016101262@khu.ac.kr, ORCID : 0000-0002-0480-9697

Corresponding Author: Jongmin Oh, E-mail: jmoh@khu.ac.kr, ORCID : 0000-0002-1104-5867

Co-Authors: Wonyoung Lee, E-mail: 2017101164@khu.ac.kr, ORCID : 0000-0001-9532-0469

Beomjin Eun, E-mail: bumjin0814@khu.ac.kr, ORCID : 0000-0003-2715-5283

Jeongsook Heo, E-mail: jsheo1986@khu.ac.kr, ORCID : 0000-0003-2913-7716

Kwanghyeon Chang, E-mail: chang38@khu.ac.kr, ORCID : 0000-0002-7952-4047

Received: 11 February, 2022. Revised: 18 April, 2022. Accepted: 20 April, 2022.

about 46% better than the national moving average model result (10.77). In addition, the performance improvement of the independent monthly model was observed in months other than November compared to the initial model. Therefore, this study confirms that DNN modeling was effective in predicting PM_{2.5} concentrations based on air pollutants concentrations, and that the learning performance of the model could be improved by selecting additional independent variables.

Keywords: PM_{2.5}, air pollutants, machine learning, prediction model, DNN model

I. 서론

1. 개요

산업화와 도시화에 의한 대기오염과 그 영향에 관한 관심이 집중되고 있으며, 특히 미세먼지(PM₁₀)와 초미세먼지(PM_{2.5})에 대한 관심도가 높아지고 있다. 초미세먼지(PM_{2.5})는 쉽게 분쇄되고 물리·화학적인 영향에 의해 성분과 크기가 변화하는 특성이 있으므로 질량농도만 고려하여 초미세먼지의 특성을 파악하기는 어렵다(Choi et al., 2020; Choe et al., 2015). 따라서, 미세먼지 방지 및 관리 대책 수립을 위해서는 입자상 오염물질의 기원 및 배출 기여도를 정확히 추정하는 것이 중요하다(Hwang et al., 2007). 해당 연구에서 사용된 DNN모델은 기존에 학습되었던 모델에 변수를 추가하는 것과 시시각각 발생하는 기상 데이터를 이용한 모델 추가 학습이 가능하다는 장점이 있다. 이러한 장점을 살린다면 기존에 사용하던 모델링기법에 비해 변수를 추가하는 방식이 간편해지고 이로 인해 농도 예측 정확성도 높아질 수 있을 것으로 기대된다. 본 연구는 여러 대기오염물질 농도를 기본 변수로 하여 PM_{2.5} 농도를 보다 정확하게 예측하는 모델 개발을 목적으로 하며, 정확한 모델 개발을 위해 최근 3년간 국가측정망(한국환경공단 에어코리아)의 600개소, 9,196,385개의 대기질 데이터를 이용하였다. 또한, 일산화탄소(CO), 이산화질소(NO₂), 오존(O₃), PM₁₀과 PM_{2.5}의 변수 간 통계적 유의성을 분석하고 각 오염물질의 월별 특성을 파악하였다. 또한, 피어슨 상관계수 0.2를 기준으로 학습에 사용할 독립 변수를 선정하여 Deep Neural Network 모델을 개발하고 RMSE를 지표로 하여 평가·검증하였다.

2. 선행연구

원동준 등은 2018년부터 2020년까지 2년간 반월 시화국가산업단지의 시간별 PM_{2.5} 농도를 인공지능 기반의 Random Forest, XGBoost, LightGBM, Deep neural network과 Voting 모델을 통해 예측하고, RMSE를 기준으로 비교분석을 진행하였다. 분석 결과 에어코리아 예측 모델 대비 약 56.82% 향상된 결과를 기록하였다.

차진욱 등은 2014년 1월부터 2017년 6월까지 서울시의 미세먼지(PM₁₀) 수치를 기계학습 알고리즘인 Artificial Neural Network (ANN)알고리즘과 K-Nearest Neighbor (K-NN) 알고리즘을 상호 응용하여 예측하였다. ANN 알고리즘을 통한 예측 값을 KNN 알고리즘으로 분류하였고 ANN과 KNN을 사용한 단일 예측 모델의 정확도 62.27%, 58.41%보다 좋아진 83.4%를 기록하였다.

성상하 등은 2010년 1월부터 2019년 11월까지 국내 기상 데이터를 엑스지부스트(XGBoost), 랜덤포레스트(Random Forest), 서포트벡터머신(SVM), 인공신경망(ANN)의 알고리즘을 적용하여 분석하고, 제시된 모형을 통해 미세먼지 수치 예측을 진행했다. 분석결과 교통량, 화력발전 등 국내 관련 변수가 추가됨에 따라 기계학습 모형의 예측 오차율이 가장 낮을 때 0.065로 높은 예측 정확도를 보였다.

손상훈 등은 2017년 1월부터 12월까지 1년 동안 서울시 소재 39개소 대기오염측정망에서 관측된 농도 자료와 기상인자를 multiple linear regression (MLR), support vector machine (SVM), 그리고 random forest (RF) 모델을 통해 서울시 PM₁₀ 농도 예측을 진행하였다. MLS, SVM, 그리고 RF 모델에 의해 예측된 PM₁₀ 농도 간 결정계수(R²)는 각각 0.260, 0.772, 그

리고 0.793이었으며, RF모델에 의한 결과값은 0.793으로 가장 높은 성능을 나타내었다.

II. 이론적 배경 및 선행연구

1. 미세먼지의 생성 및 구성요소

미세먼지는 직경에 따라 PM₁₀과 PM_{2.5} 등으로 구분하며, 특히 2.5 μ m 이하의 직경을 갖는 PM_{2.5}는 2차 대기오염 물질로 여러 대기오염물질을 기원으로 하므로 대기질 판단 인자로서 대표성이 크다. 우리나라는 1978년 SO₂를 시작으로 대기질 관측자료가 공개되고 있으며, 국가관측망에서 측정 및 공개하는 SO₂, CO,

NO₂, PM₁₀, O₃ 중 O₃를 제외한 물질은 PM_{2.5}의 전구 물질로 작용하는 것으로 알려져 있다(Ghim 2013). 황산화물(SO_x), 질소산화물(NO_x), 암모니아(NH₃), 휘발성 유기화합물(VOCs) 등의 전구물질이 대기 중의 특정 조건에서 반응하여 PM_{2.5}가 2차 생성되므로, 이러한 전구물질과의 상관관계를 통해 PM_{2.5}의 농도를 예측 가능할 것으로 가정하였다.

2. 시간 변화에 따른 대기오염물질과 PM_{2.5} 상관관계 추이

시간적 변화에 따른 PM_{2.5} 생성에 대한 대기오염물질의 연관성을 파악하기 위해 PM_{2.5}와 SO₂, CO, O₃,

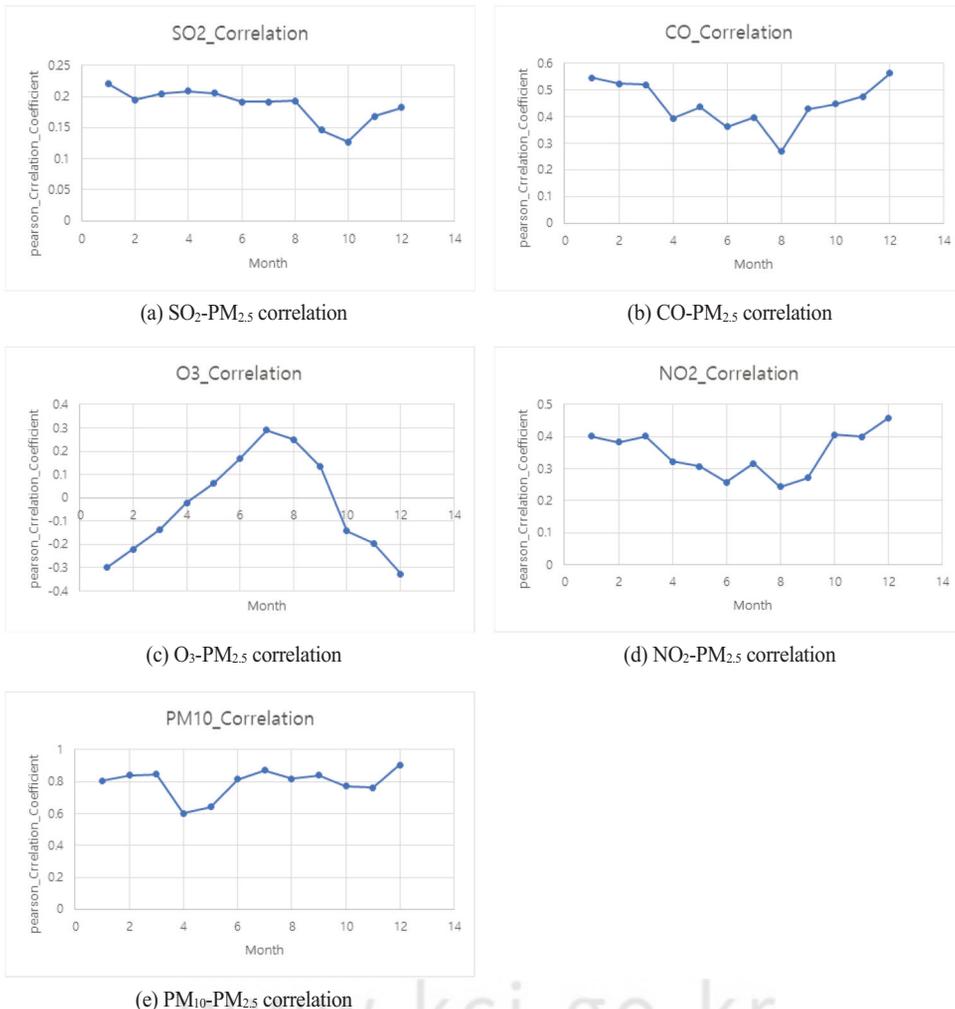


Figure 1. Correlation analysis between each air pollutant and PM_{2.5}

NO₂, PM₁₀의 상관관계를 분석하였다. Figure 1은 2017년, 2019년 및 2020년도 국가측정망 9,196,385 개 데이터셋을 바탕으로 PM_{2.5}와 대기오염물질들의 월별 피어슨 상관계수를 분석한 결과이다. SO₂는 황이 연소할 때에 발생하는 기체로 황과 산소의 화합물이며, 주로 석유, 석탄 속에 들어있는 유황화합물의 연소로 인해 생성되어 대기로 배출된다(Kim, 2019). 또한 SO_x 중 약 98% 비율을 차지하며, 미세먼지의 주요 생성물질로써 PM_{2.5} 총 질량의 약 20~35%를 차지하는 것으로 연구된 바 있다(Pathak et al, 2009; Zhang et al, 2005). Figure 1에서 장마철의 강우 세정효과에 의해 일시적으로 상관관계가 감소하는 경향을 보인다. 주로 화석연료에서 배출되는 CO는 PM_{2.5}의 상관관계가 4월과 8월 감소하였다 다시 증가하는 경향을 나타내었다. 이는 4월의 황사로 인한 PM_{2.5}의 일시적 상승과 8월의 강우에 의한 세정효과 때문에 상관성이 감소한 것으로 해석된다(Yoon 2014).

여름철 강한 일사량에 의해 발생한 고농도 O₃이 대기중 산화력을 증가시켜 PM_{2.5}의 2차 생성을 촉진하여 양의 상관관계를 갖지만, 겨울철 O₃의 NO_x 적정효과에 의한 음의 상관관계를 갖는다는 결과가 연구된 바 있다(Lee et al. 2020). O₃의 경우 PM_{2.5}의 전구물질은 아니지만, 여름철과 겨울철 PM_{2.5}의 생성에 간접적인 영향이 있음을 확인할 수 있었다. NO_x는 대기중 산화를 통해 2차 생성되는 PM_{2.5} 전구물질로, 일사량이 증가할수록 고농도 NO₂의 발생 확률이 증가하며, 계절별 농도변화는 크지 않았다(Lee et al. 1999). 따라서 여름철 상관관계의 감소는 강우 세정효과에 의한 PM_{2.5} 농도 변화에 따른 결과로 추정된다. PM₁₀은 PM_{2.5}에 가장 직접적인 연관성을 갖는 변수로, PM_{2.5} 농도와 높은 연관성을 보이나, 4월의 경우 황사의 영향으로 PM₁₀의 분율이 상승하여 상관관계가 낮은 경향을 나타내는 것으로 해석된다(Jeon 2010). 추가적으로 시간의 변동에 따른 온도, 습도, 풍향, 풍속, 일사량, 자외선량 및 기압 등 다양한 변인이 존재하고 대부분 대기오염물질에서 뚜렷한 계절성이 나타나므로, 시계열 분석 과정이 PM_{2.5} 예측모델 개발 과정에서 고려되어야 할 사항이라 사료된다.

III. 연구 방법

1. 데이터 수집 및 활용

본 연구에서는 초미세먼지 예측을 위한 DNN 모델 제작 및 데이터 분석을 위해 2017, 2019 및 2020년도 국가측정망으로부터 얻은 대한민국 전역의 대기질 측정데이터를 활용하였다. 데이터셋은 Table 1과 같이 측정지역, 측정망, 측정소 코드, 측정소명, 측정일시, SO₂, CO, O₃, NO₂, PM₁₀, PM_{2.5}, 주소 등 총 12개 항목으로 구성된다. 본 연구는 모델링의 대상 범위를 대한민국 전역으로 선정하였고, 지역정보와 측정 일시를 제외한 6가지 대기오염물질(SO₂, CO, O₃, NO₂, PM₁₀, PM_{2.5}) 농도 데이터를 학습에 사용하였다.

Table1. The format of raw data sets

Classification	Feature	Purpose
Location information	Location	Unused
	Local Network	
	Measuring Station Code	
	Measuring Station Name	
	Measuring Station Adress	
Time	Measurement Date and Time	
Air Pollutants	SO ₂	Independent variables
	CO	
	O ₃	
	NO ₂	
	PM ₁₀	Dependent Variables
	PM _{2.5}	

2. 데이터의 전처리

DNN Modeling에서 결측치가 존재하거나 이상치가 존재하는 데이터셋을 학습에 사용하는 것은 모델의 정확도를 저해하는 요인이 될 수 있다. 따라서 2017, 2019 및 2020년도 데이터를 병합 후, 6가지 대기오염물질 항목에서 결측치가 존재하는 데이터셋을 제거하였다. 또한, 이상치를 제거하기 위한 방법으로 IQR(Interquartile range) 방법을 사용하였다. IQR 방법은 이상치를 제거하기 위해 보편적으로 사용되는 통계적 방법으로 사분위수의 상위 75% 지점 값(Q1)

Table 2. Pearson's correlation coefficient of air pollutants

	SO ₂	CO	O ₃	NO ₂	PM ₁₀	PM _{2.5}
SO ₂	1.000000	0.280628	-0.000749	0.298532	0.243470	0.230525
CO	0.280628	1.000000	-0.284623	0.586022	0.452062	0.534664
O ₃	-0.000749	-0.284623	1.000000	-0.471995	0.035559	-0.055721
NO ₂	0.298532	0.586022	-0.471995	1.000000	0.394386	0.430499
PM ₁₀	0.243470	0.452062	0.035559	0.394386	1.000000	0.814862
PM _{2.5}	0.230525	0.534664	-0.055721	0.430499	0.814862	1.000000

과 하위 25% 지점 값(Q3) 차이를 IQR로 하고, 하위 25% 지점(Q1) $-1.5IQR$ 부터 상위 75%지점(Q3) $+1.5IQR$ 범위 밖의 값을 제거하는 방법이다. DNN 모델은 반복학습을 통해 오차를 최소화하는 가중치 값을 찾는 것을 목적으로 하며, 결과는 가중치를 포함한 매개변수들의 곱과 합이 비선형분류 목적의 활성화 함수를 거쳐 표현된다. 이 과정에서 이상치가 포함된 데이터셋 학습은 가중치 설정에 방해요인이 되어 높은 오차의 원인이 되거나, 반복학습에 의한 과적합 현상의 원인이 될 수 있으므로 데이터셋에서 제거하여 학습의 효율성을 증가시키는 방안을 채택하였다. 학습에 사용된 최종 데이터셋은 7,856,008 개이다.

3. 데이터의 상관관계 분석

DNN 모델의 효율성을 증진시키기 위해서는 학습 종속변수와 상관성이 높은 학습데이터 선정과정이 필요하다. 따라서 본 연구는 종속변수인 PM_{2.5}와 독립변수인 SO₂, CO, O₃, NO₂, PM₁₀과의 상관관계 분석을 진행하였다. 종속변수와 독립변수의 관계를 규명하는 과정은 선행연구 사례에 기반한 정성적인 방법과 정량적인 방법을 사용할 수 있다. 본 연구에서는 선형 회귀분석을 통한 피어슨 상관계수 R값을 지표로 하였으며, R은 다음식으로 계산된다.

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

여기서, y_i 는 변수 y 의 i 번째 측정값, \bar{y} 는 모든 y 값의 평균값, x_i 는 변수 x 의 i 번째 측정값을, \bar{x} 는 모든 x 값의 평균값을 의미한다. 상관계수 R값은 $-1 \sim 1$ 값

으로 나타나며, 절대값이 작을수록 낮은 상관관계를 갖는다. 값의 부호는 독립변수와 종속변수 간 상관관계 방향성을 의미하며, 통상적으로 절대값이 0.2 이상 상관계수에서 유의미한 상관관계로, 절대값 0.4 이상 상관계수에서 강한 상관관계로 해석한다(Cha et al, 2018).

Table 2는 PM_{2.5}와 각 대기오염물질 간의 Pearson correlation 분석결과를 제시한 것이다. 종속변수 PM_{2.5}에 대한 SO₂의 상관계수는 0.2 이상으로 유의미한 상관관계를 보였으며, CO, NO₂, PM₁₀은 각각 0.4 이상으로 높은 상관관계를 나타내었다. 단, O₃의 경우는 상관계수가 -0.055721 로 통계적으로 무의미하므로 변수로 채택하지 않았다.

4. 모델링 방법

본 연구의 PM_{2.5} 모델링에 사용된 DNN 기법은 다층구조의 은닉층(hidden layer)을 포함하고, 은닉층에 포함된 퍼셉트론 사이에 연결된 가중치 w 값을 학습과정을 통해 오차를 최소화하는 방향으로 최적화시켜 기대치에 근사한 출력력을 얻는 방법이다. 퍼셉트론은 인공 신경망의 일종으로 다수의 입력으로부터 하나의 결과를 출력하는 알고리즘이다. Figure 2에 본 연구의 예측모델 개발 과정을 제시하였다. 개발 과정은 2017, 2019 및 2020년 대기질 데이터를 PM_{2.5}와의 상관관계를 기준으로 하여 일괄적으로 학습시키는 초기모델을 구성하고, 정확성을 평가하기 위해 국가측정망 예측이동평균 모델 RMSE 값(10.77)과 비교하였다. 이후 시간적, 계절적 특성에 민감한 대기오염물질의 특성을 반영하기 위해 월별로 모델(Improved model)을 개발하여 시계열 요인 반영에 대한 모델 정확성 향상 여부를 평가하였다.

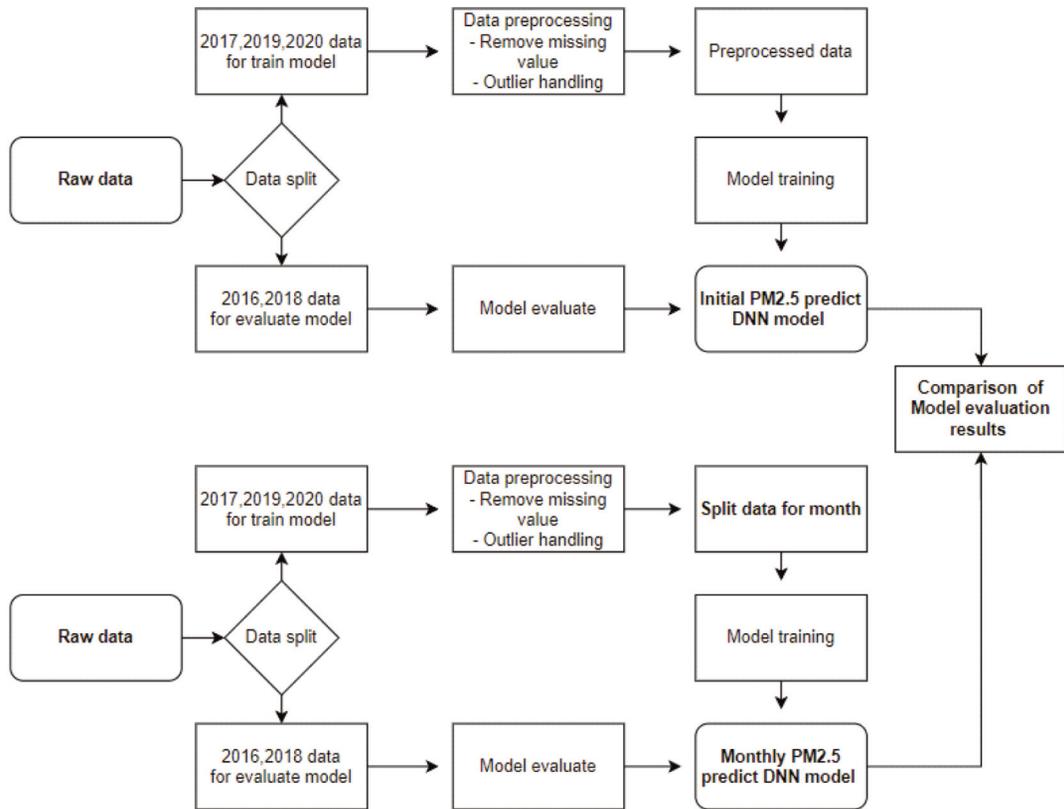


Figure 2. The flowchart of modeling method in this study

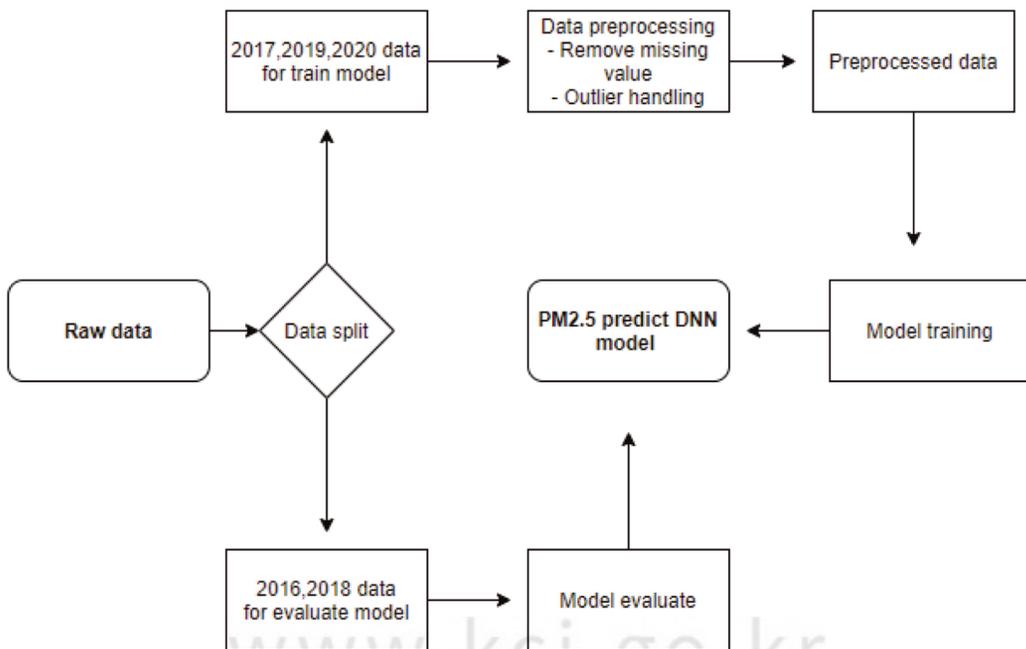


Figure 3. The flowchart of Initial model

1) 초기모델

초기모델 학습은 상관관계 분석 결과에 따라 SO₂, CO, NO₂, PM₁₀ 4개 column으로 구성된 데이터셋을 사용하였으며, 초기모델에서 시계열 특성은 고려하지 않고 학습을 진행하였다. Figure 3은 초기모델 개발의 흐름도이다.

(1) 초기모델 체계에서 DNN 구조

초기모델은 입력층과 출력층 사이 다수의 은닉층으로 구성되어 있으며, DNN을 이용한 regression 설계에서 은닉층의 활성화 함수는 보편적으로 Relu 함수를 사용하지만 본 연구에서는 반복 횟수 증가에 따라 tanh가 우세한 결과를 보여 tanh 함수를 채택하였다.

Relu 함수는 Rectified Linear Unit 함수로, 정의역이 0보다 작을때는 0을, 0 이상일때는 y 값을 반환해 주는 함수이며 tanh 함수는 하이퍼볼릭 탄젠트 함수로 두 함수 모두 은닉층 사이의 비선형 관계 형성을 위해 사용된다. Figure 4는 초기모델의 구조를 나타내며 1개 층의 입력층과 2개 층의 은닉층, 1개 층 출력층으로 구성된다. 입력층의 퍼셉트론 개수는 모델 학습에 필요한 독립변수의 개수와 동일하게 4개이며, 은닉층의 모든 퍼셉트론과 연결된 가중치 w값이 Adam(Adaptive Moment Estimation)방식을 사용하여 오차를 줄여나가는 방향으로 최적화된다. Table

Table 3. Architecture summary of Initial model

Model: "sequential_1"		
Layer (type)	Output Shape	Param #
dense_5 (Dense)	(None, 4)	20
dense_6 (Dense)	(None, 8)	40
dense_7 (Dense)	(None, 8)	72
dense_8 (Dense)	(None, 4)	36
dense_9 (Dense)	(None, 1)	5

Total params: 173
 Trainable params: 17
 Non-trainable params: 0

3은 모델에 사용된 각 층의 퍼셉트론 정보를 제시한 것이다.

(2) 초기모델의 학습환경

Table 4는 초기모델 학습 및 훈련에 사용된 조건들을 의미한다. Optimizer로는 Adam을 사용하여 Loss 값을 최소화하는 가중치를 찾는 방향으로 모델을 학습시켰으며 손실함수는 MSE(Mean Squared Error)을 사용하였다. 총 30회 반복 학습을 통해 모델을 학습시켰으며, Metrics로는 RMSE를 사용하였다.

Table 4. The initial model learnig conditions

Model Learning Conditions				
Optimizer	Epoach	Metrics	Batch_size	Loss function
Adam	30	RMSE	4096	MSE

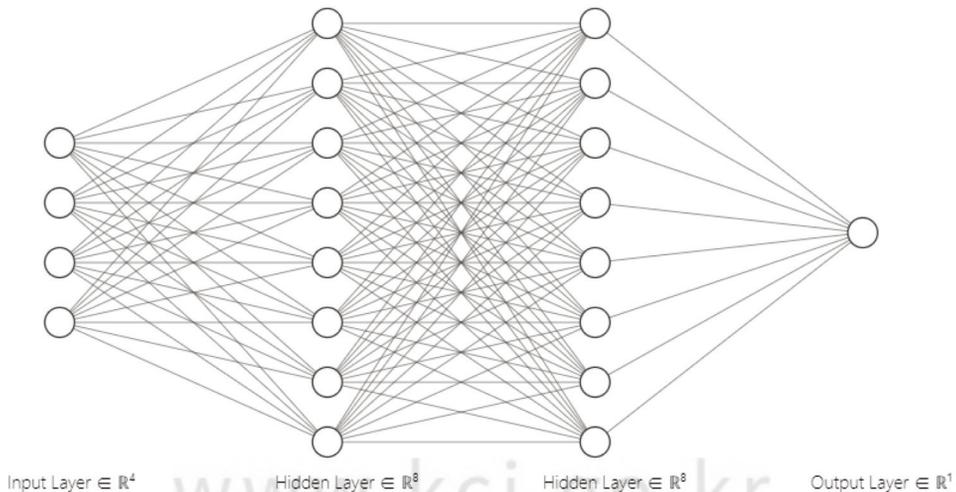


Figure 4. The structure of Initial model

2) 초기모델의 개선

초기모델의 경우 시계열 요소를 고려하지 않고 학습을 진행하였으나 대부분 대기오염 물질은 뚜렷한 계절성이 있으므로 시계열 요소를 고려하는 방법으로 정확성 향상이 가능할 것으로 판단하였다. 이러한 이유로 One-hot encoding 방법을 통해 data의 측정시기를 월 단위로 구별한 모델링 방안을 고려하였으나, 12개 열이 확장되며 용량이 과하게 늘어나고, 학습시간 지연을 야기하는 등 문제가 발견되어 데이터를 월 단위로 세분화하여 12개 월별 모델을 제작하고, 이를 초기모델과 비교하는 방안을 채택하였다. Table 5는 PM_{2.5}에 대한 월별 대기오염물질의 피어슨 상관계수로 0.2를 기준으로 하여 학습에 활용할 대기오염 물질을 선별하였다.

Table 5. The correlation of monthly air pollutants

Model	SO ₂	CO	O ₃	NO ₂	PM ₁₀
Mon1	0.273	0.579	-0.299	0.425	0.823
Mon2	0.196	0.528	-0.187	0.374	0.840
Mon3	0.249	0.602	-0.084	0.422	0.913
Mon4	0.243	0.408	-0.025	0.349	0.631
Mon5	0.263	0.459	0.109	0.371	0.653
Mon6	0.200	0.359	0.174	0.274	0.814
Mon7	0.218	0.394	0.278	0.346	0.881
Mon8	0.189	0.260	0.252	0.228	0.828
Mon9	0.176	0.441	0.130	0.319	0.858
Mon10	0.125	0.456	-0.128	0.405	0.765
Mon11	0.178	0.485	-0.199	0.422	0.770
Mon12	0.191	0.566	-0.329	0.459	0.906

개선모델(Improved model) 개발 과정에서 초기모델과 상이한 조건은 학습에 참여시키는 독립변수 항목과 각 모델 학습에 사용되는 데이터셋의 개수이다. 이는 PM_{2.5}에 대한 월별 대기오염물질의 상관관계 변화를 통해 결정하였다. 학습 참여 기준은 초기모델 판단 기준과 같은 상관관계수 0.2 이상으로 하였으며, 종속변수와의 상관관계가 낮은 항목은 학습에서 배제하는 방법을 사용하였다.

Table 6은 월별로 학습에 참여하는 독립변수를 제시한 것으로 3~5개 독립변수 항목을 모델 학습에 사

Table 6. The training independent variables of each model

Model	Train data set Configuration				
Mon1	SO ₂	CO	O ₃	NO ₂	PM ₁₀
Mon2	CO	NO ₂	PM ₁₀		
Mon3	SO ₂	CO	NO ₂	PM ₁₀	
Mon4	SO ₂	CO	NO ₂	PM ₁₀	
Mon5	SO ₂	CO	NO ₂	PM ₁₀	
Mon6	SO ₂	CO	NO ₂	PM ₁₀	
Mon7	SO ₂	CO	O ₃	NO ₂	PM ₁₀
Mon8	CO	O ₃	NO ₂	PM ₁₀	
Mon9	CO	NO ₂	PM ₁₀		
Mon10	CO	NO ₂	PM ₁₀		
Mon11	CO	NO ₂	PM ₁₀		
Mon12	CO	O ₃	NO ₂	PM ₁₀	

용하였다. 개선모델의 구조와 학습환경의 경우, 입력층 퍼셉트론 개수와 학습에 참여한 독립변수 항목 및 반복 횟수 외에는 초기모델과 동일하게 구성하였다.

IV. 결과 및 고찰

1. 초기모델 평가

Figure 5는 초기모델의 학습곡선이다. 총 30회 학습을 진행하였으며 20% 무작위로 Validation된 RMSE값과 MAE값은 각각 4.36과 5.78이다. RMSE는 평균 제곱근 편차로 정확성 평가에 보편적으로 사용하는 지표 중 하나이며, MAE는 평균 절대오차로 쌍을 이루는 관측치 간 오차 측정값을 의미한다. 예측모델 평가에 보편적으로 사용되는 RMSE 값을 기준으로 국가측정망 24시간 예측이동평균모델의

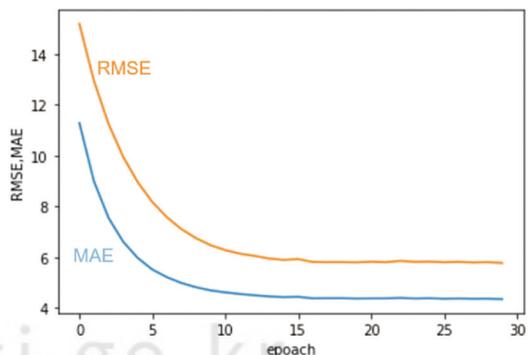


Figure 5. The Learning curve of Initial model

RMSE값인 10.77에 비하여 약 46% 정확성이 개선되었다.

2. 모델 평가결과 비교

학습에 사용하지 않은 2016년과 2018년도 데이터를 12개 월별로 분류하여 각 데이터셋으로 초기모델과 12개의 개선모델을 평가하였다. Figure 6은 RMSE를 지표로 하여 모델의 평가 결과를 나타낸 것이다. 결과적으로 두가지 종류의 모델 모두 7~9월 시기 실측치와 근사한 결과가 도출되는 유사한 경향성을 갖고 있으나, 대체적으로 개선 모델에서 약 4% 정확한 예측 결과가 계산되었다. DNN 모델링에서 모델 정확성을 결정하는 중요한 요인으로는 학습에 사용되는 데이터의 양과 데이터의 품질이 있으며 데이터의 품질은 예측하고자 하는 종속변수와의 상관성이 높을수록, 데이터의 양은 많을수록 모델의 정확성이 증가한다. Figure 6과 같이 월별 데이터를 구분하여 학습시킨 개선모델의 정확성이 대부분 월별 평가에서 향상되었으나, 11월 모델의 경우 오히려 정확성이 저하된 경향을 보인다. 이러한 이유로는 DNN 모델의 특성상 정

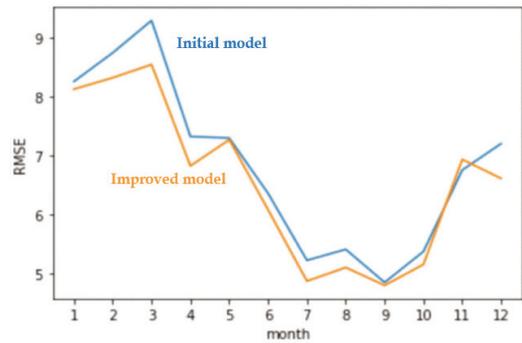


Figure 6. Comparison of evaluation results of Initial model and Improved model

확한 원인규명이 어렵지만, 초기모델 대비 주요한 변동사항인 학습데이터 양의 감소가 주요한 원인이라고 추론되며 향후 추가적인 데이터의 누적 학습으로 해결할 수 있을 것이라 판단된다.

3. 개선된 모델의 예측치와 실측치 비교

2016년과 2018년 데이터로부터 각각 200개 데이터셋을 무작위로 추출하여 개선 모델에 적용한 결과는 Figure 7에 제시하였으며, 실측치와 예측치를 Plot

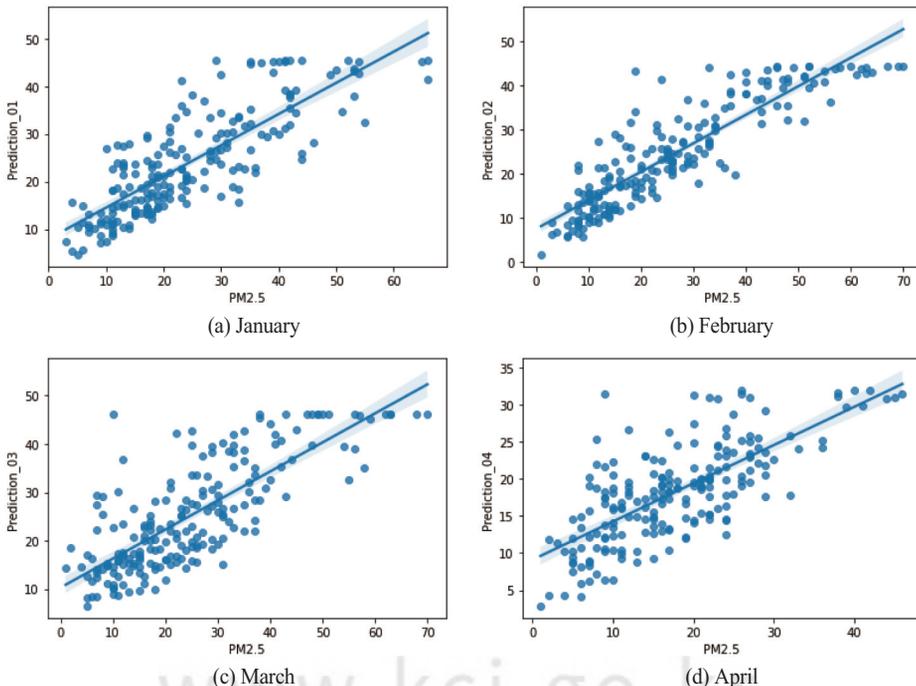


Figure 7. Comparison of Forecasting Model and Observed Concentration based on $PM_{2.5}$

chart 형태로 나타내었다.

1월 모델부터 12월 모델까지의 정확도는 Figure 6의 Improved model의 RMSE 수치와 비례하며, 대

체적으로 기온이 높은 7월 ~ 9월의 예측 결과가 실측치와 밀접하게 분포한다.

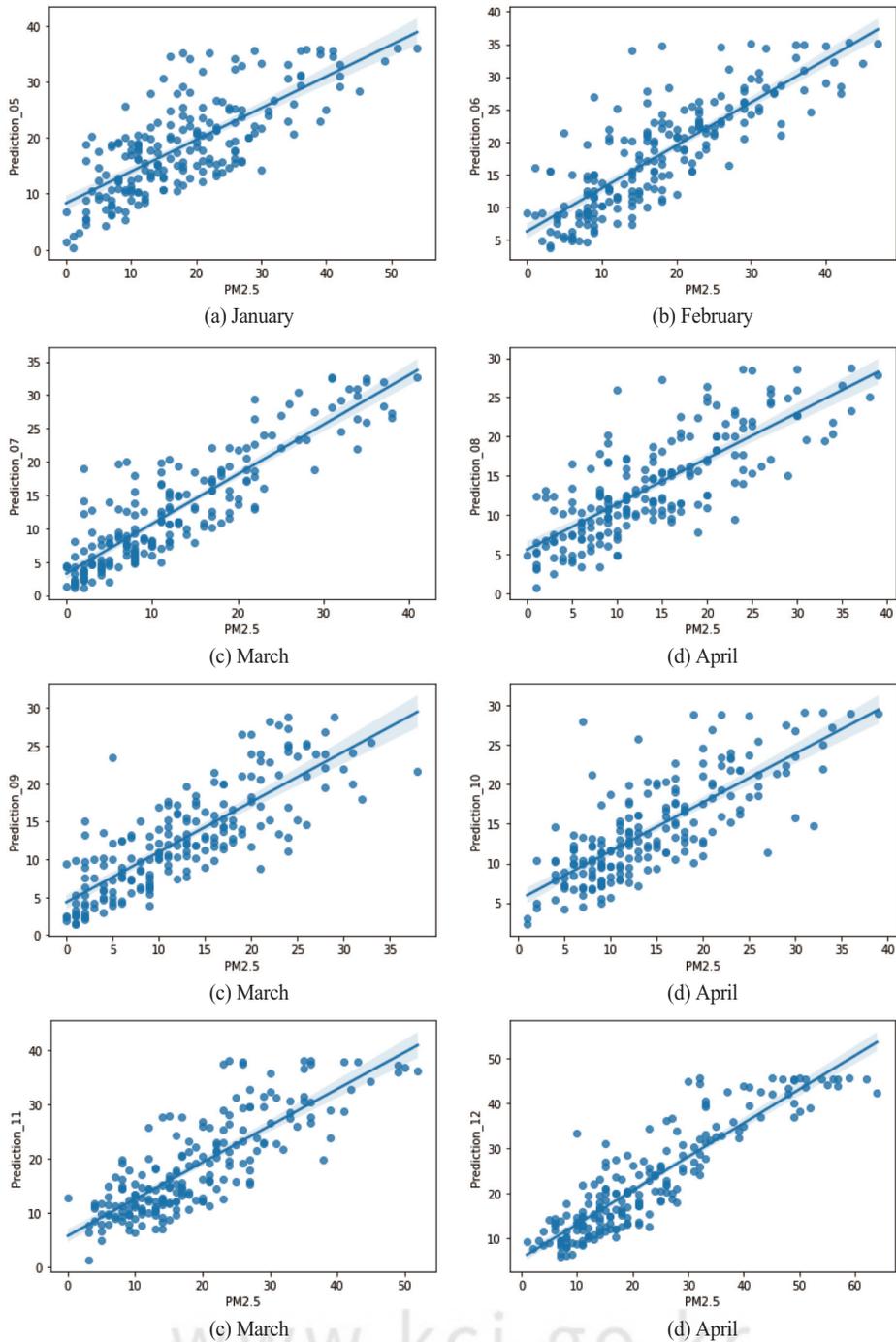


Figure 7. Continued

V. 결론

본 연구에서는 국가측정망 대기질 데이터를 이용하여 CO, NO₂, O₃, PM₁₀과 PM_{2.5}의 변수 간 통계적 유의성을 분석하고 각 오염물질의 월별 특성을 분석한 결과, PM_{2.5}의 전구물질인 대기오염물질이 PM_{2.5}의 생성에 중요한 변수임을 확인하였다. 시간적 요인을 고려하지 않고 학습 데이터의 양에 집중한 초기모델을 개발하였고, 국가측정망 이동평균 예측모델의 RMSE 값 10.77과 비교하여 약 46% 향상된 5.78 RMSE 결과로 DNN 모델링 방식의 타당성을 검증하였다. 대기오염물질의 시간적 및 계절적 특성을 반영하여 월별로 모델을 개발하는 방법을 선택하였다. 초기모델과 개선모델의 객관적 평가를 위해 학습에 사용하지 않은 2016년과 2018년 데이터셋으로 모델의 평가를 진행한 결과, 11월을 제외한 달에서 평균 4% 정도의 정확도 개선이 있었으나, 그 정도가 경미하고 오히려 11월에서는 정확도가 감소하는 결과가 도출되었다. 이는 PM_{2.5} 예측에 있어 시계열 요소 반영의 긍정적인 영향이 존재하나 아직 확보된 데이터의 수가 부족하여 정확도 향상 효과가 크지 않다고 평가된다. 향후 초미세먼지 예측에 있어 충분한 데이터가 축적되고 RNN(Recurrent Neural Network)과 같은 시계열 요인 반영이 용이한 모델링 기법을 적용한다면 더욱 정확한 초미세 먼지 농도 예측이 가능할 것으로 기대된다.

References

- Cha J, Kim J. 2018. Development of Data Mining Algorithm for Implementation of Fine Dust Numerical Prediction Model. *Journal of the Korea Institute of Information and Communication Engineering*, 22(4), 595-601. <https://doi.org/10.6109/JKIICE.2018.22.4.595>
- Kim DS. 2016. Characteristics in Atmospheric Chemistry between NO, NO₂ and O₃ at an Urban Site during MAPS (Megacity Air Pollution Study)-Seoul, Korea, *Journal of Korean Society for Atmospheric Environment* 32(4): 422-434. <http://dx.doi.org/10.5572/KOSAE.2016.32.4.422>
- Won DJ. 2021. Prediction of Fine Dust in Gyeonggi-do Industrial Complex using Machine Learning Methods, *Journal of KIISE* 48(7): 764-773.
- Im DY. 2021. Temporal Analyses of PM Data, Estimation of the Past Unmonitored PM_{2.5} Data, and Assessment of the COVID-19 Effect at the Background Areas in Korea. *Journal of Korean Society for Atmospheric Environment* 37(4): 670-690. <https://doi.org/10.5572/KOSAE.2021.37.4.670>
- Lee HK. 2020. A Study on the Seasonal Correlation between O₃ and PM_{2.5} in Seoul in 2017. *Journal of Korean Society for Atmospheric Environment* 36(4): 533-542. <https://doi.org/10.5572/KOSAE.2020.36.4.533>
- Shin HW. 2021. Changes of Air Pollutants Concentrations Associated with the Rates of Rainfall and Its Duration during Summertime in Korea. *Journal of Korean Society for Atmospheric Environment* 2014(10): 175.
- Lee HW. 1999. The Effect of Meteorological Factors on Variation and Temporal and Spatial Characteristics of NO₂ Concentration in Pusan Area. *Journal of Environmental Science International* 8(4): 465-471.
- Hwang SM, Shin DS, Lee BM, Kim HH, Cho HG, Moon GJ, Jeong SH. 2007. A Study on the Distribution Characteristics of Air Pollutants at PAMS in Seoul Metropolitan Area, *Proceeding of 43th meeting of Korean Society for Atmospheric Environment in 2007*, pp. 1396-1399.
- Lim IH. 2013. A Study on the hourly Characteristics

- of Air Pollution in the Gwangyang Bay. 環境研究論文集 13: 29-39.
- Jeon BI. 2010. Characteristics of Spacio-Temporal Variation for PM₁₀ and PM_{2.5} Concentration in Busan, Journal of Environmental Science International 19(8): 1013-1023. <https://doi.org/10.5322/JES.2010.19.8.1013>
- Kim JS, Bais AL, Kang SH, Lee J, Park K. 2011. A semi-continuous measurement of gaseous ammonia and particulate ammonium concentrations in PM_{2.5} in the ambient atmosphere. J. Atmos. Chem. 68(3): 251-263. doi:10.1007/s10874-012-9220-y
- Lee BK, Lee DS, Kim MG. 2001. Rapid time variations in chemical composition of precipitation in South Korea. Water, Air, and Soil Pollution 130(1-4): 427-432. doi:10.1023/A:1013845620642.
- Sung SH, Kim SJ, Ryu MH. 2020. A Comparative Study on the Performance of Machine Learning Models for the Prediction of Fine Dust: Focusing on Domestic and Overseas Factors. Korea Innovation Studies 15(4): 339-357.
- Son SH, Kim JS. 2020. Evaluation and Predicting PM₁₀ Concentration Using Multiple Linear Regression and Machine Learning. Korean Journal of Remote Sensing 36(6-3): 1711-1720.
- Yoon SY. 2014. Effect of Precipitation Cleaning on Air Pollutants in Summer and Winter. Korean Meteorological Society, Proceedings of the Autumn Meeting of KMS, 2014: 146-147.
- Ghim YS. 2013. Regional Trends in Short-Term High Concentrations of Criteria Pollutants from National Air Monitoring Stations. Journal of Korean Society for Atmospheric Environment 29(5): 545-552.