

## ORIGINAL ARTICLE

# 관측자료와 CMAQ 예측자료를 활용한 부산지역 격자단위 대기질 예측 개선을 위한 머신러닝 기법 평가 연구

윤하은 · 도우곤\* · 박정욱

부산광역시 보건환경연구원 대기환경연구부

## Machine Learning for Improving Grid-Level CMAQ Air Quality Predictions in Busan Using Observations and CMAQ Predictions

Haeun Yun, Woogon Do\*, Jeongok Park

Busan Metropolitan City Institute of Health and Environment, Busan 46616, Korea

### Abstract

Monitoring data alone are not sufficient to understand the spatial distribution of air quality across Busan; this study develops machine learning-based nonlinear regression models to improve local air quality predictions. Hourly CMAQ (Community multiscale air quality model) predictions and observations of PM<sub>10</sub> and PM<sub>2.5</sub> from 2021 to 2024 were analyzed. Grid cells were grouped using cluster analysis, and a paired monitoring station was selected for each grid using correlation analysis. Subsequently, CMAQ predictions and observations from paired monitoring stations were used as input data to train four machine learning models: Random forest, Support vector machine, Gradient boosting, and XGBoost (Extreme gradient boosting). Model performance was evaluated using the correlation coefficient, root mean square error (RMSE), and mean absolute error (MAE). Consequently, XGBoost demonstrated superior performance with a relatively fast training speed. When applied to PM<sub>10</sub> and PM<sub>2.5</sub> for the first half of 2025, the XGBoost-based correction significantly improved prediction accuracy in the mid-range concentration levels (10  $\mu\text{g}/\text{m}^3 \sim 40 \mu\text{g}/\text{m}^3$  for PM<sub>10</sub> and 5  $\mu\text{g}/\text{m}^3 \sim 25 \mu\text{g}/\text{m}^3$  for PM<sub>2.5</sub>). In particular, PM<sub>10</sub> accuracy improved by about 2.7 to 2.9 times in the 20  $\mu\text{g}/\text{m}^3 \sim 30 \mu\text{g}/\text{m}^3$  range, and PM<sub>2.5</sub> accuracy improved by about 1.7 to 2.4 times in the 10  $\mu\text{g}/\text{m}^3 \sim 20 \mu\text{g}/\text{m}^3$  range compared to the conventional CMAQ predictions. However, performance decreased at low and high concentrations owing to the limited training data.

**Key words :** CMAQ, Machine learning, Cluster analysis, Correlation analysis, Regression

### 1. 서 론

미세먼지(particulate matter, PM)는 대기 중에 부유하거나 침적하는 입자상 물질로, 장·단기 노출 시 호흡기계 및 심혈관계 질환의 악화, 조기 사망 위험 증가 등 다양한 건강영향과 관련되는 것으로 알려져 있다. 이에 따라 미세먼지 농도를 정확하게 예측하고, 고농도

발생 가능성에 선제적으로 대응하는 것은 시민 건강 보호와 대기환경 정책의 효과적 이행을 위해 매우 중요하다. 세계보건기구(WHO) 또한 PM<sub>10</sub>과 PM<sub>2.5</sub>를 포함한 대기오염물질의 건강위해성을 제시하며, 보다 엄격한 대기질 관리와 예측정보 활용의 필요성을 강조한 바 있다(WHO, 2021).

Received 30 April, 2026; Revised 21 May, 2026;

Accepted 29 May, 2026

\*Corresponding author : Woogon Do, Busan Metropolitan City Institute of Health and Environment, Busan 46616, Korea  
Phone : +82-51-309-2762  
E-mail : dou777@korea.kr

© The Korean Environmental Sciences Society. All rights reserved.  
© This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

부산광역시 보건환경연구원은 부산지역 27개 도시 대기측정소를 운영하며 시간 단위 대기질을 상시 모니터링하고 있다(Busan Metropolitan City Health and Environmental Information System, 2025). 그러나 측정소 기반 자료는 관측 지점의 농도만을 제공하므로, 복잡한 해안지형과 도시구조를 갖는 부산 전역의 공간적 농도 분포를 상세하게 파악하는 데에는 한계가 있다. 이러한 한계를 보완하기 위한 방법으로 광화학 수치모델을 활용한 고해상도 대기질 예측이 널리 사용되고 있으며, 그중 CMAQ (community multiscale air quality model)은 기상장과 배출량 자료를 바탕으로 대기오염물질의 배출, 수송, 화학반응 및 제거 과정을 모의하는 대표적인 3차원 대기질 모델이다(Byun and Schere, 2006; United States Environmental Protection Agency (U.S. EPA), 2025). CMAQ은 지역 및 도시 규모의 대기질 평가, 배출원 영향 분석, 정책 효과 진단 등에 폭넓게 활용되어 왔다(Burns et al., 2021).

그러나 CMAQ을 포함한 수치모델은 배출량 자료의 불확실성, 기상장 재현 오차, 화학반응 과정의 단순화, 복잡한 국지 배출 및 지형 효과의 미반영 등으로 인해 실제 관측값과 차이를 보일 수 있다. 특히 도시 내 국지적 배출특성과 미세한 공간변동이 큰 지역에서는 예측 오차가 더욱 두드러질 수 있으며, 이에 따라 관측자료를 활용하여 수치모델 결과를 보정하거나 예측 성능을 향상시키기 위한 연구가 지속적으로 수행되어 왔다. Hogrefe et al.(2009), Yu et al.(2018), Jang et al.(2022) 등은 관측자료와 통계적 또는 수치적 후처리 기법을 결합하여 대기질 예측의 정확도를 개선할 수 있음을 보고하였다. 또한 최근에는 머신러닝을 활용한 자료융합, 오차보정, 단기예측 고도화 연구가 활발히 이루어지고 있으며, 이는 대기질 연구의 중요한 발전 방향 중 하나로 평가되고 있다.

기존의 대기질 예측 정확도 개선 연구들은 다수의 기상 및 환경 변수를 입력자료로 사용하거나, 예측 정확도 향상을 위한 알고리즘이 비교적 복잡하여 실제 대기질 예보 업무에 적용하기에는 한계가 있었다. 본 연구에서는 입력자료를 CMAQ 모델의 예측치와 상시 운영되는 도시대기측정소의 관측치로 제한하여 데이터 수집 및 전처리 부담을 최소화하였으며, 4년간의 장기 축적 데이터를 학습시킨 머신러닝 모델 구축을 통해 입력자료의 축소에 따른 한계점을 보완하였다. 또한 실시간으로

생성되는 CMAQ 데이터를 즉각적으로 보정할 수 있는 모델을 구축하여 실제 대기질 예보 시스템에서 요구하는 신속성과 정확성을 동시에 확보하고자 하였다.

머신러닝은 자료로부터 비선형 관계와 복합 패턴을 학습하여 새로운 입력에 대한 예측을 수행하는 데이터 기반 기법으로, 복잡한 회귀 문제에 효과적으로 적용될 수 있다. 대기질 분야에서는 기상자료, 배출량, 위성자료, 수치모델 결과 및 지상관측자료를 결합하여 PM<sub>10</sub>, PM<sub>2.5</sub>, 오존 등의 예측 성능을 향상시키려는 연구가 다수 수행되었으며, 특히 Random forest, Support vector machine, Gradient boosting, XGBoost (Extreme gradient boosting)와 같은 알고리즘은 높은 예측력과 실용성으로 주목받고 있다(Kim et al., 2022; Agbehadji and Obagbuwa, 2024). 최근 연구들은 머신러닝 기반 접근법이 전통적 수치모델의 체계적 편의를 보완하고, 관측망이 제한된 지역에서도 공간적으로 연속적인 대기질 정보를 제공하는 데 유용함을 보여주고 있다(Kim et al., 2022).

이러한 배경에서 본 연구의 목적은 CMAQ 모델이 산출한 부산지역 1 km × 1 km 해상도 862개 격자의 시간 단위 예측값에 관측자료 기반 머신러닝 보정기법을 적용하여 PM<sub>10</sub> 및 PM<sub>2.5</sub> 예측 정확도를 개선하는 데 있다. 이를 위해 2021년부터 2024년까지의 격자별 CMAQ 예측자료를 이용하여 시계열 변동 특성에 따른 군집분석을 수행하고, 동일 군집 내 도시대기측정소 관측값과의 상관분석을 통해 각 격자와 가장 높은 관련성을 갖는 짝측정소를 선정하였다. 이후 각 격자의 CMAQ 예측값과 짝측정소의 관측값을 입력자료로 활용하여 Random forest, Support vector machine, Gradient boosting, XGBoost 기반의 비선형 회귀모델을 구축하고, 상관계수, 평균제곱근오차(RMSE), 평균절대오차(MAE)를 이용하여 예측 성능을 정량적으로 비교하였다. 마지막으로 최종 선정된 모델을 2025년 상반기 CMAQ 예측값에 적용하여 기존 예측결과와 비교함으로써 머신러닝 기반 보정의 적용 가능성과 개선 효과를 평가하였다.

## 2. 자료 및 방법

### 2.1. CMAQ 모델링

대기질 진단평가시스템은 CMAQ 모델을 기반으로 부산지역 대기질을 실시간으로 모의·평가하기 위해

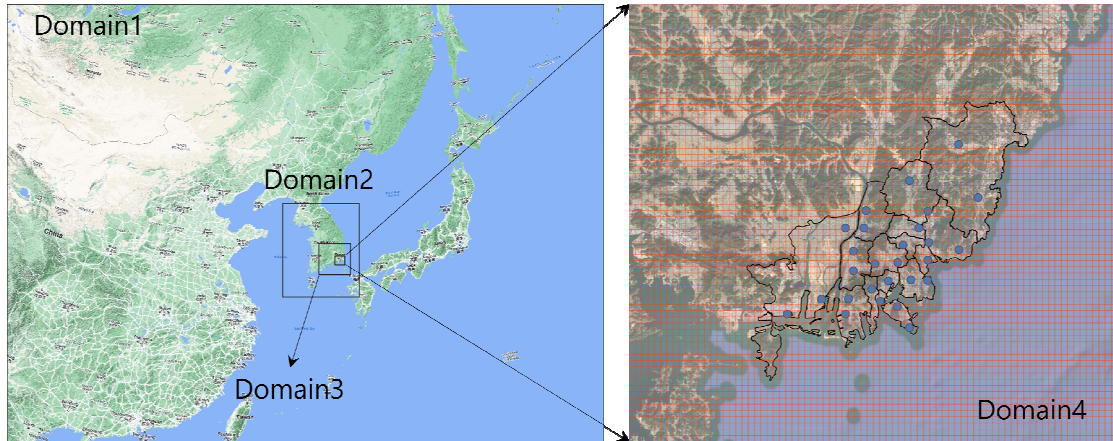


Fig. 1. Modeling domain for air quality diagnosis and evaluation system. Left part shows locations of domain1 to domain3, right part shows geographical features of domain4 and locations of air quality monitoring stations.

Table 1. Grid configurations of modeling system

|                  | Model domain | Grids information                                                                                              |
|------------------|--------------|----------------------------------------------------------------------------------------------------------------|
| Horizontal grids | Domain1      | 174 × 128 × 27 km                                                                                              |
|                  | Domain2      | 67 × 82 × 9 km                                                                                                 |
|                  | Domain3      | 83 × 83 × 3 km                                                                                                 |
|                  | Domain4      | 78 × 70 × 1 km                                                                                                 |
| Vertical levels  | sigma levels | 1.000, 0.995, 0.990, 0.985, 0.970, 0.950, 0.930, 0.910, 0.880, 0.840, 0.800, 0.740, 0.700, 0.600, 0.450, 0.000 |

구축되었으며, 2017년에 최초 도입되었다(Busan Metropolitan City Institute of Health and Environment, 2025). 모델링 영역은 동아시아 27 km 격자 영역에서 한반도 9 km 격자 영역, 영남권 3 km 격자 영역, 최종적으로 부산권의 1 km 격자 영역으로 4단계의 nesting 도메인으로 구성된다(Fig. 1, Table 1).

진단평가시스템의 기상입력자료는 National Centers for Environmental Prediction (NCEP)의 global forecast system (GFS) 전구 데이터를 기상모델 weather research forecast (WRF)에 입력하여 생성하고 있다. GFS는 NCEP에서 운영하고 있는 전구 기상 예보 수치모델이며 하루에 4번 6시간 간격으로 실행되고 384시간의 예보자료를 포함하고 있다(NCAR Research data archive, 2025). WRF 모델은 2005년부터 미국 NOAA 산하 기관인 NCEP의 현업 모델로 사용되고 있고 세계적으로 널리 보급되어 많은 연구에

활용되고 있는 기상모델이다(WRF Model users' page, 2025). 배출량 자료를 생성하기 위해서 미국 EPA에서 제공하는 sparse matrix operator kernel emissions (SMOKE) 모델을 적용하였다. SMOKE는 미국의 Environmental Modeling Center (EMC)에서 개발된 것으로 모델링에 필요한 배출량을 고효율로 계산할 수 있도록 배출량을 matrix 구조체로 생성하는 배출량 모델링 시스템이다(Community Modeling and Analysis System, 2025). SMOKE 모델에 사용하는 동아시아 지역의 배출량자료로 일본 국립환경연구소에서 개발된 아시아지역 대기오염물질배출량, Regional Emission inventory in Asia (REAS) 2015년 자료를 이용하였다. REAS 배출량은 Ohara et al.(2007)에 의해 최초로 개발되었으며 SO<sub>2</sub>, NO<sub>x</sub>, CO, NMVOC 등 총 10개 물질에 대하여 아시아지역 전체에 대하여 0.25도 해상도로 배출량을 공개하고 있다(Ohara et al., 2007; Regional Emission inventory in Asia data

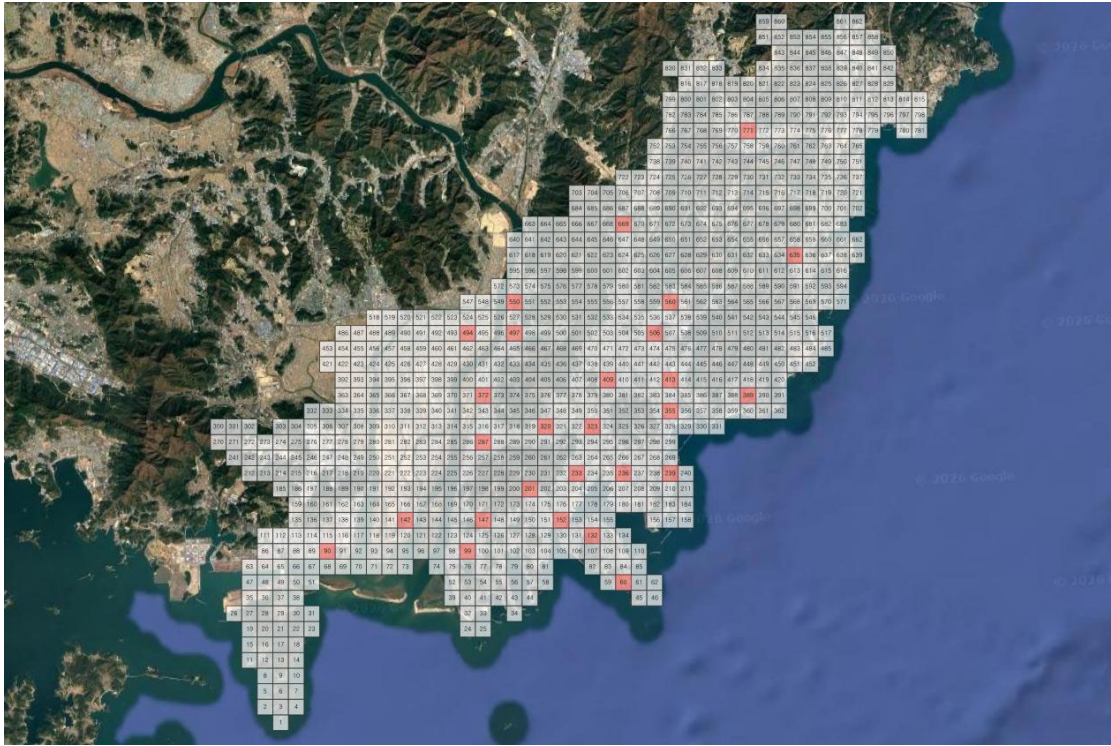


Fig. 2. Spatial distribution of 1 km × 1 km grids in Busan (monitoring stations shown in red).

download site, 2025). 국내배출량 자료는 환경부 국가미세먼지정보센터의 대기정책지원시스템(clean air policy support system, CAPSS)에서 공개되는 2020년 배출원별, 1 km 격자별 배출량을 사용하였다. 생성된 기상 및 배출량 입력자료들은 최종적으로 광화학 수치모델인 CMAQ에 입력된다. CMAQ은 미국 EPA가 정한 규제모형 중에서 가장 많이 이용되는 3차원 광화학 오일러리안 대기질 모델이며, 대기 중 오염농도, 건성침적, 습성침적 등 여러 가지 물리적 과정과 대기 중에서 발생하는 광화학 반응 등 상세한 물리·화학 반응 모듈을 포함하고 있다(Byun and Schere, 2006; U.S. EPA, 2025).

## 2.2. 머신러닝 학습데이터 구축 및 전처리

본 연구에서는 2021년 1월 1일부터 2024년 12월 31일까지의 PM<sub>10</sub> 및 PM<sub>2.5</sub> CMAQ 예측값과 부산지역 도시대기측정소 관측값을 사용하였다.

CMAQ 예측값은 부산시 전역을 1 km × 1 km 해상

도로 구분한 862개 격자의 1시간 단위 시계열 농도 자료를 사용하였으며, 격자 분포는 Fig. 2에 제시하였다. CMAQ 예측자료는 모델결과가 생성되는 시점에서 총 95시간의 예측결과를 생성한다. 본 연구에서는 이러한 모델 예측결과를 예측 당일(Day0), 1일 후(Day1), 2일 후(Day2)로 구분하고 분석에 사용하였다. 다만 시스템 점검 등으로 인해 모델이 수행되지 않은 일자는 분석에서 제외하였다.

관측자료는 부산지역 도시대기측정소 27개소에서 생산된 1시간 단위 PM<sub>10</sub> 및 PM<sub>2.5</sub> 농도 자료를 사용하였으며, 측정소 현황은 Table 2에 제시하였다. 황사의 영향을 최소화하기 위하여 기상청 황사 관측일을 기준으로 황사 발생일과 발생 전후 1일의 자료를 제외하였다. 또한 대기환경측정소의 점검 등으로 인해 발생한 결측치는 제거하였으며, 최종적으로 CMAQ 예측값과 관측값이 모두 존재하는 시간대의 자료만 학습 및 검증에 사용하였다. 그 결과 유효 자료 수는 Day0 33,240건, Day1 33,216건, Day2 33,192건으로 나타났다.



**Table 2.** Air quality monitoring stations in Busan

| No | Grid number | Monitoring station  | No | Grid number | Monitoring station    |
|----|-------------|---------------------|----|-------------|-----------------------|
| 1  | 60          | Taejongdae (TJ)     | 15 | 355         | Yongho-dong (YH)      |
| 2  | 90          | Noksan-dong (NS)    | 16 | 372         | Hakjang-dong (HJ)     |
| 3  | 99          | Jangnim-dong (JR)   | 17 | 389         | Gaegeum-dong (GG)     |
| 4  | 132         | Cheonghak-dong (CH) | 18 | 409         | Jeonpo-dong (JP)      |
| 5  | 142         | Myeongji-dong (MZ)  | 19 | 413         | Jaesong-dong (JS)     |
| 6  | 147         | Dangni-dong (DR)    | 20 | 494         | Daejeo-dong (DJ)      |
| 7  | 152         | Gwangbok-dong (GB)  | 21 | 497         | Deokcheon-dong (DC)   |
| 8  | 201         | Daesin-dong (DS)    | 22 | 506         | Myeongjang-dong (MJ)  |
| 9  | 233         | Sujeong-dong (SJ)   | 23 | 550         | Hwamyong-dong (HM)    |
| 10 | 236         | Gwangan-dong (GA)   | 24 | 560         | Hoedong-dong (HD)     |
| 11 | 239         | Deokpo-dong (DP)    | 25 | 635         | Gijang-eup (GJ)       |
| 12 | 267         | Jwa-dong (JW)       | 26 | 669         | Cheongryong-dong (CR) |
| 13 | 320         | Yeonsan-dong (YS)   | 27 | 771         | Yongsu-ri (YR)        |
| 14 | 323         | Daeyeon-dong (DY)   |    |             |                       |

### 2.3. 군집분석

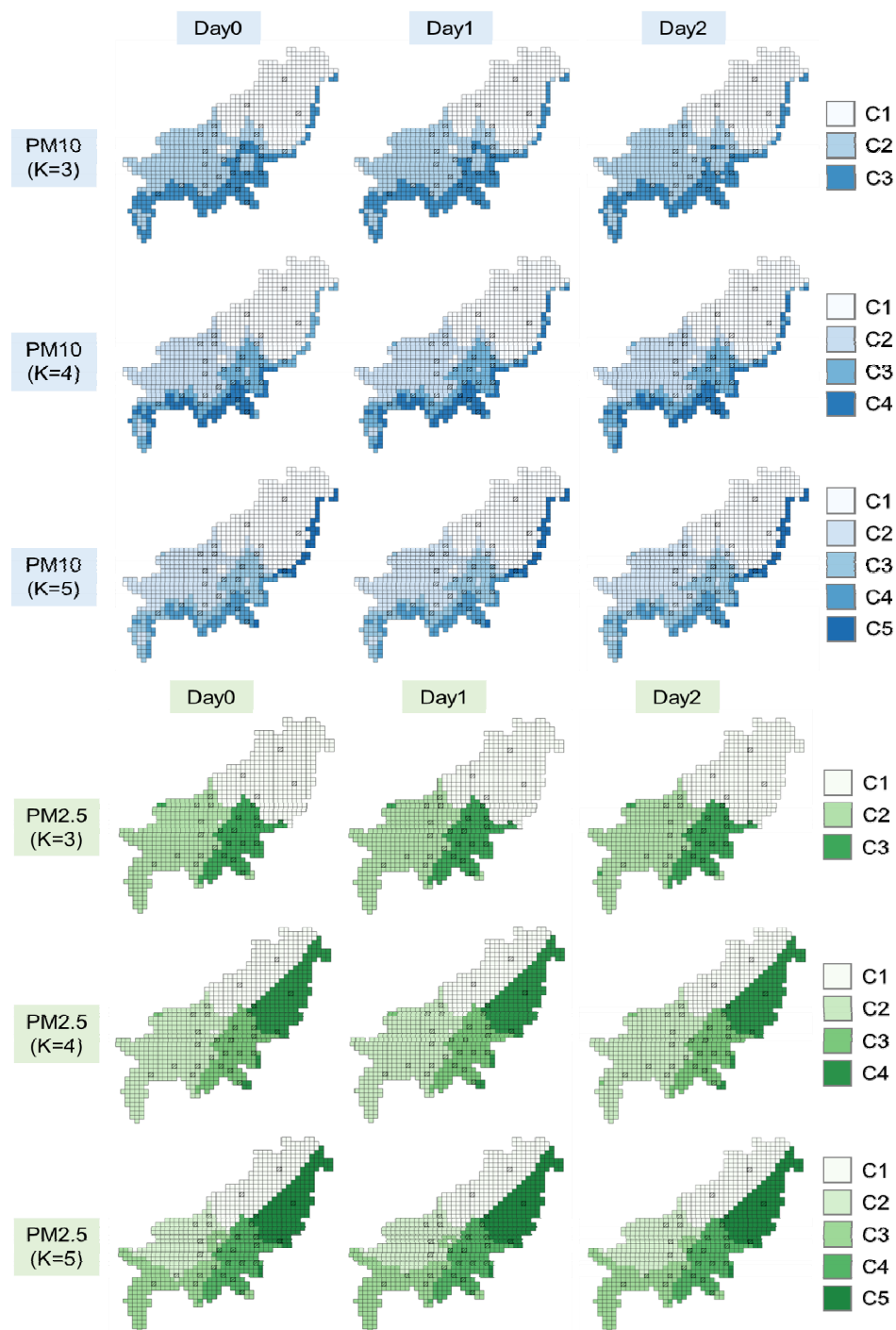
군집분석은 유사한 특성을 갖는 자료를 동일 집단으로 분류하는 통계적 기법이다. 본 연구에서는 부산 지역 1 km × 1 km 격자별 PM<sub>10</sub> 및 PM<sub>2.5</sub> CMAQ 예측자료의 시계열 변동 특성을 바탕으로, 농도 변동성이 유사한 격자들을 군집화하여 각 군집의 공간적 특성을 분석하고자 하였다. 자료 간 유사성은 유클리드 거리(Euclidean distance)를 기준으로 산정하였다.

군집분석 방법으로는 비계층적 군집분석인 K-means 알고리즘을 적용하였다. K-means는 연구자가 군집 수를 사전에 설정한 후 자료를 반복적으로 재배치하여 군집 내 분산을 최소화하는 방법으로, 대용량 자료에 대한 적용이 용이하고 분석 속도가 빠른 장점이 있다(MacQueen, 1967; Lloyd, 1982). 본 연구에서는 군집 수 증가에 따른 군집 내 제곱합(within-cluster sum of squares, WCSS)의 감소 경향을 비교하고, 감소폭이 완만해지는 지점을 기준으로 최적 군집 수를 결정하는 elbow method를 적용하였다. 또한 절대 농도 수준의 차이에 따른 영향을 최소화하고 시계열 변동 패턴 중심의 군집화를 수행하기 위하여 분석 전 자료를 표준화하였다. 군집분석 결과의 공간 분포는 QGIS 3.6.1을 이용하여 시각화하였다.

### 2.4. 머신러닝

격자별 CMAQ 예측값의 정확도를 개선하기 위하여 머신러닝 기반의 비선형 회귀모델을 구축하고, 이를 이용해 보정된 CMAQ 예측값을 산출하였다. 본 연구에서는 다양한 머신러닝 기법을 적용하여 예측 성능을 비교하고, 그 중 보정 효과가 가장 우수한 기법을 최종 모델로 선정하였다. 모델의 학습에는 2021년부터 2023년까지의 자료를, 성능 평가에는 2024년 자료를 사용하였다.

수치모델에서 발생하는 오차의 복잡한 비선형성을 개선하기 위해서 네 가지의 머신러닝 기법을 선정하여 성능을 비교 분석하였다. 먼저, 대표적인 앙상블 학습 방법으로 예측 안정성이 검증된 Random forest (RF)를 선정하였다. RF는 다수의 의사결정나무를 생성한 뒤 그 예측값을 평균하여 최종 결과를 산출하는 앙상블 기법으로, 변수 선택의 임의성을 통해 나무 간 상관성을 낮추고 과적합을 줄일 수 있는 장점이 있다(Ryu and Han, 2022). 또한, 이상데이터에 둔감하여 일반화 성능이 뛰어난 Support vector machine (SVM)을 통해 오차의 일반적인 경향성을 파악하고자 하였다. SVM은 회귀 문제에서 입력자료와 목표변수 간의 비선형 관계를 효과적으로 학습할 수 있는 기법으로, 고차원 자료에



**Fig. 3.** Cluster distributions of  $PM_{10}$  and  $PM_{2.5}$  for Day0, Day1, Day2 predictions with different numbers of clusters ( $K = 3, 4$ , and  $5$ ). Grids containing air quality monitoring stations are indicated by hatching.

서도 안정적인 예측 성능을 나타내는 것으로 알려져 있다(Drucker et al., 1996; Smola and Schölkopf, 2004). 이와 대조적으로, 학습을 반복하며 미세 잔차를 단계적으로 제거하는 Gradient boosting (GB)을 적용하여 오차 개선의 정밀도를 높였다. GB는 이전 단계의 예측 오차를 순차적으로 보완하는 방식으로 학습이 이루어지며, 반복적으로 약한 학습기를 결합하여 예측 성능을 향상시킬 수 있다고 알려져 있다(Friedman, 2001). 마지막으로 4년 동안의 방대한 데이터를 효율적으로 학습하기 위해 학습속도가 빠르면서도 과적합 방지에 효과적인 XGBoost (XGB)를 선정하였다. XGB는 GB를 기반으로 계산 효율성과 과적합 제어 기능을 강화한 알고리즘으로, 병렬 연산이 가능하고 규제항을 포함하여 학습의 안정성을 높일 수 있다(Sim et al., 2022; Fatima et al., 2023).

각 기법은 R v.4.4.1 환경에서 구현하였다. RF는 randomForest 패키지 내 randomForest 함수를 사용하였으며 ntree는 300으로 설정하였다. SVM은 e1071 패키지 내 svm 함수를 사용하였으며 type은 “eps-regression”, kernel은 “radial”로 설정하여 분석하였다. GB는 gbm 패키지 내 gbm 함수를 사용하였으며 Out-of-bag (OOB) 방법을 사용하여 학습하였다. XGB는 XGboost 패키지 내 xgb.train 함수를 사용하였으며 objective는 “reg:squarederror”로 설정하였다.

### 3. 결과 및 고찰

#### 3.1. 격자별 유사성 분류

Fig. 3은 군집 수(K)를 3, 4, 5로 설정하였을 때  $PM_{10}$ 과  $PM_{2.5}$ 의 군집분석 결과를 나타낸다. 빗금으로 표시된 격자는 도시대기측정소가 위치한 격자이다.  $PM_{10}$ 의 경우 K=3에서 각 군집의 공간적 구분이 비교적 명확하고 군집별 격자 수 또한 상대적으로 균형 있게 분포하는 것으로 나타났다. 반면 K가 4와 5로 증가할수록 남부 및 해안 인접 지역의 군집이 점차 세분화되면서 군집 간 격자 수의 불균형이 커지는 경향을 보였다. 또한 elbow method에서도 K=3 부근에서 WCSS의 감소폭이 상대적으로 크게 나타나,  $PM_{10}$ 의 적정 군집 수는 3개로 판단하였다.

$PM_{2.5}$ 의 경우에도 예측일에 관계없이 전반적으로 유사한 공간 분포가 나타났으나,  $PM_{10}$ 과는 달리 K가 증가함에 따라 지역적 특성이 보다 뚜렷하게 구분되는 양상

을 보였다. 특히 K가 3에서 4로 증가할 때 동부산 지역으로 분류된 군집이 해안 인접 지역과 내륙 인접 지역으로 세분화되었으며, K=5에서는 전체적으로 각 군집의 공간 분포와 격자 수가 보다 고르게 나타났다. 한편  $PM_{2.5}$ 는 elbow method에서  $PM_{10}$ 과 같이 뚜렷한 WCSS 감소 지점을 확인하기 어려웠으므로, 본 연구에서는 군집 간 격자 수의 균형과 공간적 해석 가능성을 종합적으로 고려하여 적정 군집 수를 5개로 결정하였다.

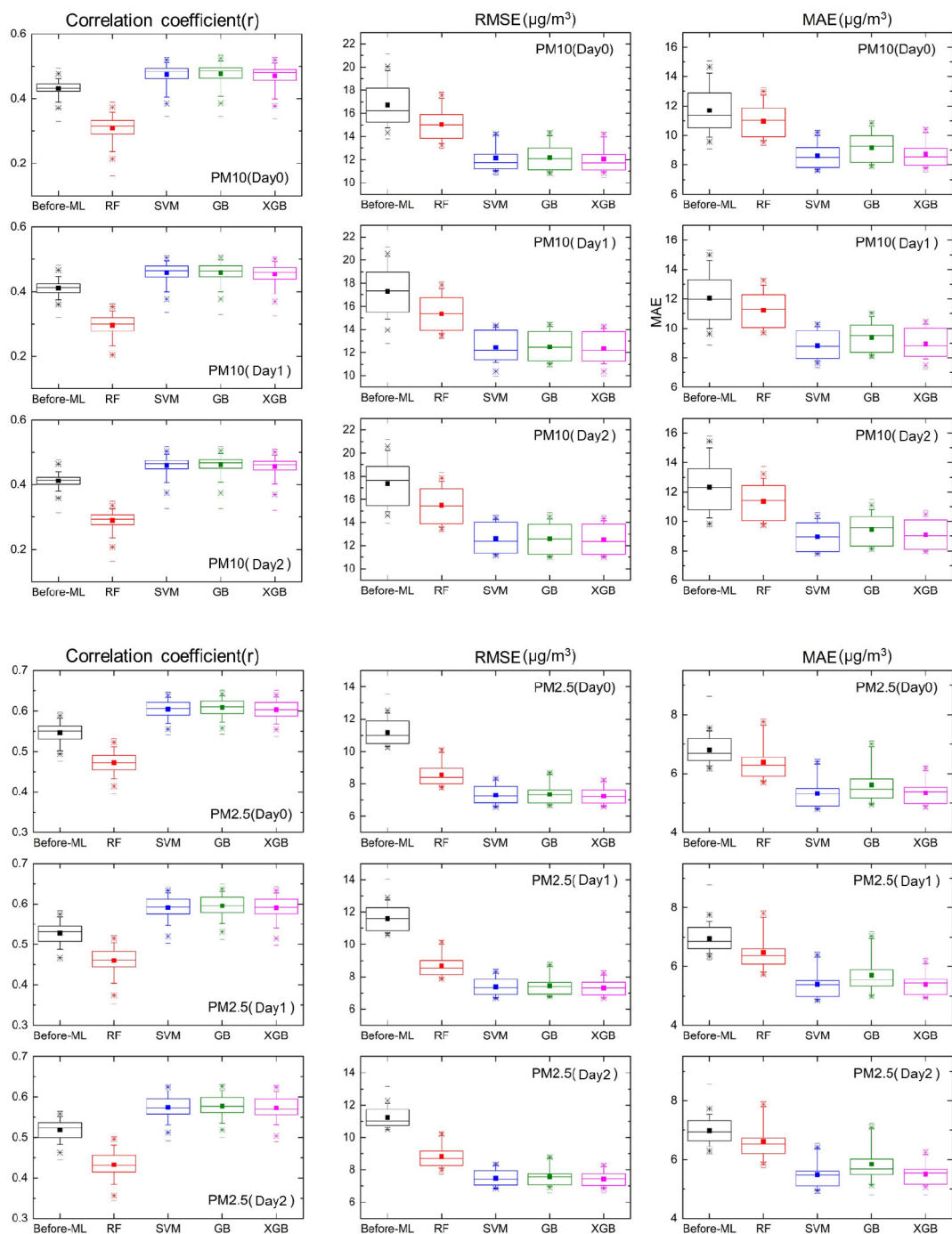
#### 3.2. 예측값과 관측값의 상관분석 및 짝측정소 선정 결과

동일 군집 내 도시대기측정소의 관측값과 격자별 CMAQ 예측값 간 상관분석 결과는 Table 3에 제시하였다.  $PM_{10}$ 과  $PM_{2.5}$  모두 예측시점이 최근일수록 상관계수(r)는 증가하고 RMSE와 MAE는 감소하여, 예측일이 가까울수록 CMAQ의 예측 정확도가 높게 나타났다. 또한  $PM_{2.5}$ 는 모든 예측일에서  $PM_{10}$ 보다 높은 상관계수와 낮은 RMSE 및 MAE를 보여, 전반적으로 더 우수한 예측 성능을 나타냈다. 이는  $PM_{2.5}$ 가 지역 배출과 2차 생성 과정의 영향을 크게 받는 경향이 있는 반면,  $PM_{10}$ 은 황사와 같은 장거리 이동성 먼지 및 조대입자의 영향을 상대적으로 더 크게 받을 수 있기 때문으로 해석된다(Kim et al., 2009; Kumar et al., 2022).

각 격자의 CMAQ 예측값과 동일 군집으로 분류된 측정소의 관측값 간 상관계수를 비교한 결과, 동일 군집 내에서는 전반적으로 유사한 변동 특성이 나타나 측정소 간 상관계수 차이가 크지 않았다. 이에 따라 각 격자별로 가장 높은 상관계수를 보이는 측정소를 최종 짝측정소로 선정하였으며, 그 결과를 Fig. 4에 제시하였다.  $PM_{10}$ 은 짝측정소 분포가 비교적 세분되어 나타난 반면,  $PM_{2.5}$ 는 동일 측정소가 넓은 연속 격자를 대표하는 경향이 뚜렷하여 보다 명확한 공간적 권역성을 보였다. 또한 모든 격자에서 상관관계는 통계적으로 유의하게 나타나( $p<0.001$ ), 본 연구의 짝측정소 선정 방법이 격자별 대표 관측자료를 구성하는 데 적절한 것으로 판단된다.

#### 3.3. 머신러닝 기반 비선형 회귀모델 성능 비교

Fig. 5는  $PM_{10}$  및  $PM_{2.5}$ 에 대하여 머신러닝 기법을 적용한 후 상관계수(r), RMSE, MAE의 분포를 비교한 결과를 나타낸다.  $PM_{10}$ 의 경우 기존 CMAQ 예측값



**Fig. 5.** Box plots of the correlation coefficient, RMSE, and MAE for  $PM_{10}$  and  $PM_{2.5}$  after applying machine learning models (RF, SVM, GB, and XGB). The box shows the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentiles; dots indicate the mean, asterisks(\*) the 1<sup>st</sup> and 99<sup>th</sup> percentiles, and whiskers the minimum and maximum values.



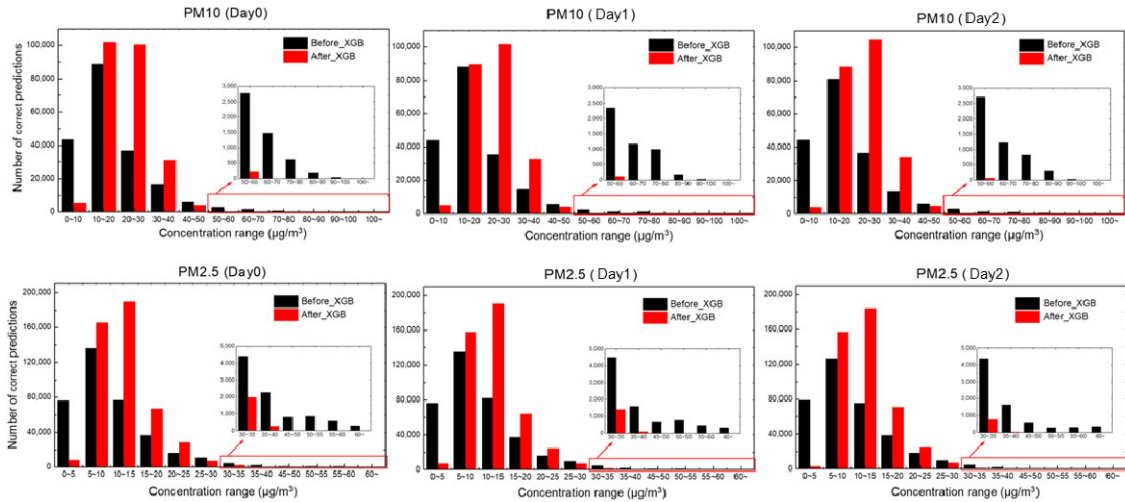


Fig. 6. Comparison of the number of correct predictions for PM<sub>10</sub> and PM<sub>2.5</sub> before and after XGBoost-based nonlinear regression correction across concentration ranges.

(Before-ML)은 상관계수가 낮고 RMSE와 MAE의 분포 범위가 넓어, 관측값과의 일치도가 상대적으로 낮고 예측 오차의 변동성도 큰 것으로 나타났다. RF를 적용한 보정 결과는 RMSE와 MAE가 다소 감소하였으나 상관계수는 오히려 낮아져, 전반적인 예측 성능 개선 효과는 제한적인 것으로 판단되었다. GB는 일부 예측일에서 오차가 감소하였으나 개선 정도가 일관되지 않았던 반면, SVM과 XGB는 모든 예측일에서 상관계수는 증가하고 RMSE와 MAE는 감소하는 경향을 보여 비교적 안정적인 성능 개선 효과를 나타내었다.

PM<sub>2.5</sub>에서도 PM<sub>10</sub>과 유사한 경향이 나타났다. Before-ML에 비해 RF의 개선 효과는 제한적이었고, GB는 일부 지표에서 개선이 나타났으나 일관성은 다소 낮았다. 반면 SVM과 XGB는 예측일 전반에서 상관계수 향상과 오차 감소가 뚜렷하게 나타나, 두 기법이 다른 기법에 비해 우수한 성능을 보이는 것으로 확인되었다. SVM과 XGB의 예측 성능은 전반적으로 유사한 수준이었으나, 본 연구에서는 반복 학습의 효율성과 계산 속도를 고려하여 XGB를 최종 머신러닝 기법으로 선정하였다.

### 3.4. (초)미세먼지 예측 성능 평가

본 연구에서는 XGBoost (XGB) 기법으로 최종 학습한 비선형 회귀모델에 2025년 1월부터 6월까지의

PM<sub>10</sub> 및 PM<sub>2.5</sub> 기존 CMAQ 예측값(Before\_XGB)과 관측값을 입력하여 보정된 CMAQ 예측값(After\_XGB)을 산출하였다. 결정치는 해당 측정소가 위치한 권역의 동일 시간대 평균값으로 대체하였다. 예측 성능 평가는 CMAQ 예측값과 관측값의 차가 1  $\mu\text{g}/\text{m}^3$  이하일 경우 “예측성공”, 1  $\mu\text{g}/\text{m}^3$ 를 초과할 경우 “예측실패”로 구분하여 수행하였으며, 그 결과를 Fig. 6에 제시하였다.

PM<sub>10</sub>의 경우 모든 예측일에서 10  $\mu\text{g}/\text{m}^3$ ~40  $\mu\text{g}/\text{m}^3$  구간의 예측성공 횟수가 보정 후 증가하는 경향을 보였다. 특히 20  $\mu\text{g}/\text{m}^3$ ~30  $\mu\text{g}/\text{m}^3$  구간에서는 기존 CMAQ 예측값 대비 약 2.7배~2.9배 향상된 예측 성능이 나타나, 중간 농도 구간에서 XGB 기반 보정 효과가 가장 크게 나타났다. PM<sub>2.5</sub> 역시 모든 예측일에서 5  $\mu\text{g}/\text{m}^3$ ~25  $\mu\text{g}/\text{m}^3$  구간의 예측성공 횟수가 증가하였으며, 특히 10  $\mu\text{g}/\text{m}^3$ ~20  $\mu\text{g}/\text{m}^3$  구간에서 기존 CMAQ 예측값 대비 약 1.7배~2.4배 향상된 예측 성능을 보였다. 반면 PM<sub>10</sub>과 PM<sub>2.5</sub> 모두 저농도 및 고농도 구간에서는 기존 CMAQ 예측값 대비 예측성공 횟수가 감소하거나 개선 효과가 제한적으로 나타났다.

이러한 경향은 학습자료의 농도 분포 특성과 관련된 것으로 판단된다. PM<sub>10</sub>의 경우 CMAQ 예측값의 약 82%가 30  $\mu\text{g}/\text{m}^3$  이하에 분포하고, 관측값 또한 10  $\mu\text{g}/\text{m}^3$ ~30  $\mu\text{g}/\text{m}^3$  구간이 전체의 58%를 차지하여 중간 농도 범위에 자료가 집중되어 있다. PM<sub>2.5</sub> 역시 CMAQ 예측값의 약 79%, 관측값의 약 77%가 20  $\mu\text{g}/\text{m}^3$  이하

에 분포하는 것으로 나타났다. 이는 자료가 많이 분포한 중간 농도 구간에서는 모델이 충분히 학습되어 보정 성능이 향상된 반면, 상대적으로 사례 수가 적은 저농도 및 고농도 구간에서는 학습이 충분하지 못했기 때문으로 해석된다.

또한 고농도 구간의 예측 성능을 개선하기 위하여 고농도 사례에 가중치를 부여한 추가 학습을 수행하였으나, 고농도 사례의 적중률 개선은 제한적이었고 오히려 자료 비중이 높은 저농도 구간의 적중률이 감소하는 결과가 나타났다. 이에 따라 가중치 기반 학습 결과는 최종 모델에 반영하지 않았다. 따라서 향후에는 고농도 사례의 추가 확보, 자료 불균형 완화를 위한 재표본화 또는 임의 생성 기법의 적용 등을 통해 극값 구간의 예측 정확도를 보완할 필요가 있을 것으로 판단된다.

#### 4. 결론

본 연구는 부산지역 1 km × 1 km 격자별 CMAQ 예측 정확도를 개선하고자, 2021년부터 2024년까지의 CMAQ 예측자료와 도시대기측정소 관측자료를 활용한 머신러닝 기반 보정모델을 구축하고 그 성능을 평가하였다. 주요 결과는 다음과 같다.

첫째, 격자별 CMAQ 예측자료의 시계열 변동 특성을 바탕으로 군집분석을 수행하고, 동일 군집 내 도시대기측정소 관측값과의 상관분석을 통해 각 격자에 대한 짝측정소를 선정하였다. 그 결과 PM<sub>10</sub>은 3개, PM<sub>2.5</sub>는 5개의 군집이 적정 군집 수로 도출되었으며, 이를 통해 부산지역의 공간적 대기질 변동 특성을 반영한 대표 관측자료를 구축할 수 있었다. 특히 PM<sub>2.5</sub>는 PM<sub>10</sub>보다 공간적 권역성이 보다 명확하게 나타나, 짝측정소 선정 결과 역시 보다 일관된 공간 분포를 보였다.

둘째, Random forest (RF), Support vector machine (SVM), Gradient boosting (GB), XGBoost (XGB)의 4개 머신러닝 기법을 비교한 결과, SVM과 XGB가 비교적 우수한 성능을 보였으며, 이 중 XGB가 예측 성능과 계산 효율성 측면에서 최종 모델로 선정되었다. XGB 기반 보정 결과, PM<sub>10</sub>은 20  $\mu\text{g}/\text{m}^3 \sim 30 \mu\text{g}/\text{m}^3$  구간에서 기존 CMAQ 예측값 대비 약 2.7배~2.9배, PM<sub>2.5</sub>는 10  $\mu\text{g}/\text{m}^3 \sim 20 \mu\text{g}/\text{m}^3$  구간에서 약 1.7배~2.4배의 예측성공 향상이 나타났다. 이는 머신러닝 기법이 수치모델 자체의 구조를 변경하지 않고도 예측 오차를 효과적으로 보정할 수 있음을 보여준다.

셋째, XGB 기반 보정 효과는 저농도 및 고농도 구간보다 중간 농도 구간에서 뚜렷하게 나타났다. 이는 학습자료의 대부분이 중간 농도 범위에 집중되어 있어 해당 구간에 대한 학습은 충분히 이루어진 반면, 상대적으로 사례 수가 적은 저농도 및 고농도 구간에서는 학습이 충분하지 못했기 때문으로 판단된다. 또한 고농도 사례에 가중치를 부여한 추가 학습을 수행하였으나 개선 효과는 제한적으로 나타나, 단순 가중치 조정만으로는 자료 불균형 문제를 충분히 해결하기 어려운 것으로 확인되었다.

종합하면, 본 연구에서 제안한 머신러닝 기반 보정모델은 부산 지역의 고해상도 격자 예측 정보 생산에 있어 CMAQ 예측값의 정확도를 향상하고, 관측망의 공간적 한계를 보완할 수 있는 유용한 방법으로 판단된다. 향후 저농도 및 고농도 구간의 예측 정확도 개선을 위하여 기상인자 등 추가 설명변수를 포함한 입력자료 활용, 저농도 및 고농도 사례의 추가 확보, 재표본화 또는 데이터 증강 기법 적용 등 극값 구간의 예측 신뢰도를 높이기 위한 후속 연구를 통해 본 모델의 실제 운영 시스템의 적용 가능성을 높일 수 있을 것으로 보인다. 이러한 개선이 이루어진다면 예보관들은 실시간 CMAQ 예측치를 입력하는 간단하고 직관적인 방법만으로도 향상된 예측 성능을 확보할 수 있을 것이다. 결과적으로 본 연구는 부산시 대기질 예보의 신속성과 정확성을 향상함으로써 지자체 단위 맞춤형 대기질 예보 시스템의 고도화와 과학적 정책지원 체계 구축을 위한 기초자료로 활용될 수 있을 것으로 기대된다.

#### REFERENCES

- Agbehadji, I. E., Obagbuwa, I. C., 2024, Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction, *Atmosphere*, 15, 1352.
- Burns, D. A., Bhatt, G., Linker, L. C., Bash, J. O., Capel, P. D., Shenk, G. W., 2021, Atmospheric nitrogen deposition in the Chesapeake Bay watershed: A History of change, *Atmos. Environ.*, 251, 118277.
- Busan Metropolitan City Health and Environmental Information System, 2025, <https://heis.busan.go.kr/>, accessed November 2025.
- Busan Metropolitan City Institute of Health and Environment, 2025, The annual report of Busan metropolitan city institute of health and environment, Busan, Republic of Korea.

- Byun, D., Schere, K. L., 2006, Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system, *Appl. Mech. Rev.*, 59, 51-77.
- Community Modeling and Analysis System, Sparse matrix operator kernel emissions modeling system, 2025, <https://www.cmascenter.org/smoke/>, accessed November 2025.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., Vapnik, V., 1996, Support vector regression machines, *Adv. Neural Inf. Process. Syst.*, 9.
- Fatima, S., Hussain, A., Amir, S. B., Awan, M. G. Z., Ahmed, S. H., Aslam, S. M. H., 2023, XGBoost and random forest algorithms: An In depth analysis, *Pak. J. Sci. Res.*, 3, 26-31.
- Friedman, J. H., 2001, Greedy function approximation: A Gradient boosting machine, *Ann. Stat.*, 29, 1189-1232.
- Hogrefe, C., Lynn, B., Goldberg, R., Rosenzweig, C., Zalewsky, E., Hao, W., Doraiswamy, P., Civerolo, K., Ku, J. Y., Sistla, G., Kinney, P. L., 2009, A Combined model-observation approach to estimate historic gridded fields of PM<sub>2.5</sub> mass and species concentrations, *Atmos. Environ.*, 43, 2561-2570.
- Jang, E., Kim, M., Do, W., Park, G., Yoo, E., 2022, Real-time estimation of PM<sub>2.5</sub> concentrations at high spatial resolution in Busan by fusing observational data with chemical transport model outputs, *Atmos. Pollut. Res.*, 13, 101277.
- Kim, B. Y., Lim, Y. K., Cha, J. W., 2022, Short-term prediction of particulate matter (PM<sub>10</sub> and PM<sub>2.5</sub>) in Seoul, South Korea using tree-based machine learning algorithms, *Atmos. Pollut. Res.*, 13, 101547.
- Kim, Y. J., Woo, J. H., Ma, Y. I., Kim, S., Nam, J. S., Sung, H., Choi, K. C., Seo, J., Kim, J. S., Kang, C. H., Lee, G., Ro, C. U., Chang, D., Sunwoo, Y., 2009, Chemical characteristics of long-range transport aerosol at background sites in Korea, *Atmos. Environ.*, 43, 5556-5566.
- Kumar, N., Johnson, J., Yarwood, G., Woo, J. H., Kim, Y., Park, R. J., Jeong, J. I., Kang, S., Chun, S., Knipping, E., 2022, Contributions of domestic sources to PM<sub>2.5</sub> in South Korea, *Atmos. Environ.*, 287, 119273.
- Lloyd, S. P., 1982, Least squares quantization in PCM, *IEEE Trans. Inf. Theory*, 28, 129-137.
- MacQueen, J., 1967, Some methods for classification and analysis of multivariate observations, in: *Proc. 5th Berkeley Symp. on Math. Stat. Probab.*, 1, 281-297.
- NCAR Research data archive, 2025, NCEP GFS 0.25 degree global forecast auxiliary grids historical archive, <https://rda.ucar.edu/datasets/ds084.3/#description>, accessed November 2025.
- Ohara, T., Akimoto, H., Kurokawa, J., Horii, N., Yamaji, K., Yan, X., Hayasaka, T., 2007, An Asian emission inventory of anthropogenic emission sources for the period 1980-2020, *Atmos. Chem. Phys.*, 7, 4419-4444.
- Regional Emission inventory in Asia (REAS) data download site, 2021, <https://www.nies.go.jp/REAS/index.html>, accessed November 2025.
- Ryu, L., Han, K., 2022, Machine learning vs. statistical model for prediction modelling: Application in medical imaging research, *J. Korean Soc. Radiol.*, 83, 1219-1228.
- Sim, D., Lee, J. Y., Jang, J., Lee, M., 2022, Prediction of chloride concentration in groundwater on Jeju Island using XGBoost regression machine learning, *J. Geol. Soc. Korea*, 58, 243-255.
- Smola, A. J., Schölkopf, B., 2004, A Tutorial on support vector regression, *Stat. Comput.*, 14, 199-222.
- U.S. EPA, 2025, CMAQ: The community multiscale air quality modeling system, <https://www.epa.gov/cmaq>, accessed November 2025.
- World Health Organization, 2021, WHO global air quality guidelines: Particulate matter (PM<sub>2.5</sub> and PM<sub>10</sub>), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, Geneva, Switzerland.
- WRF Model users' page, 2025, <https://www2.mmm.ucar.edu/wrf/users/>, accessed November 2025.
- Yu, H., Russell, A., Mulholland, J., Odman, T., Hu, Y., Chang, H. H., Kumar, N., 2018, Cross-comparison and evaluation of air pollution field estimation methods, *Atmos. Environ.*, 179, 49-60.
- 
- Researcher. Haeun Yun  
Busan Metropolitan City Institute of Health and Environment  
hey2616@korea.kr
  - Researcher. Woogon Do  
Busan Metropolitan City Institute of Health and Environment  
dou777@korea.kr
  - Researcher. Jeongok Park  
Busan Metropolitan City Institute of Health and Environment  
pjok@korea.kr