Check for updates

Journal of Korean Society for Atmospheric Environment
Vol. 40, No. 5, October 2024, pp. 558-571
https://doi.org/10.5572/KOSAE.2024.40.5.558
p-ISSN 1598-7132, e-ISSN 2383-5346

논 문

서울지역 고농도 오존 상황전파를 위한 분류모델 활용방안 제시

Proposed Utilization of Classification Models for High Ozone Alert Communication in Seoul

이진효, 사창훈, 윤태호[1], 최용석[1], 이현정[2], 구자용[3],*

서울특별시 기후환경본부 대기정책과, [1]서울특별시 보건환경연구원 대기환경연구부, [2]제주대학교 환경공학과, [3]서울시립대학교 환경공학부

접수일 2024년 8월 10일
수정일 2024년 9월 12일
채택일 2024년 9월 30일

Received 10 August 2024
Revised 12 September 2024
Accepted 30 September 2024

***Corresponding author**
Tel : +82-(0)2-6490-2866
E-mail : jykoo@uos.ac.kr

Jinhyo Lee, Changhun Sa, Taeho Yoon[1], Yongsuk Choi[1], Hyunjung Lee[2], Jayong Koo[3],*

Air Quality Policy Division, Seoul Metropolitan Government, Seoul, Republic of Korea
[1]Atmospheric Research Department, Seoul Metropolitan Government Research Institute of Public Health and Environment, Seoul, Republic of Korea
[2]Department of Environmental Engineering, Jeju National University, Jeju, Republic of Korea
[3]Department of Environmental Engineering, University of Seoul, Seoul, Republic of Korea

Abstract A machine learning-based classification model was applied to identify the main influencing factors affecting O_3 advisory (triggered when hourly average O_3 concentrations exceed 0.12 ppm), using existing 25 urban air quality monitoring networks data from Seoul and meteorological data from Seoul automatic weather station (Jongno-gu). From May to September 2023, data were collected and analyzed. The dataset comprised 19 variables, including urban air quality metrics (such as O_3, $PM_{2.5}$, PM_{10}, NO_X) and meteorological parameters (such as wind speed, temperature, relative humidity, rain probability, and cloud cover), recorded on an hourly basis. Using this data, two classification models were developed: the first model (analysis model, ANM) employed decision tree and random forest algorithms to identify the main influencing factors affecting high O_3 concentration events. The second model (prediction model, PRM) was designed to predict the likelihood of O_3 advisory for the following day. Through the application of ANM, the main influencing factors affecting high O_3 concentration were identified, with $PM_{2.5}$, PM_{10}, and temperature emerging as significant variables affecting O_3 advisory. And both decision tree and random forest models have demonstrated strong classification performance. These results indicate that the models effectively classified the data into category 0 (no O_3 advisory) and category 1 (O_3 advisory). Additionally, a second classification model (PRM) was developed to predict the likelihood of O_3 advisory in Seoul for the following day. This model utilized seven independent variables: temperature, relative humidity, rain probability, cloud cover, and forecasted air quality levels ($PM_{2.5}$, PM_{10}, O_3). Overall, these findings suggest that PRM is a viable tool for predicting next-day O_3 advisory. In this study, the application of the proposed classification model methodology based on real-time air quality and meteorological data for a given region is expected to quantitatively explain the performance of PRM and be usefully utilized in reducing O_3 exposure for sensitive and vulnerable populations.

Key words: O_3, O_3 advisory, Classification model, Decision tree, Random forest

1. 서 론

오존은 산소분자(O_2)에 산소원자(O)가 결합한 산소의 동소체로, 무색 기체이며 강한 산화력과 반응성을 가진 물질이다. 오존은 존재하는 위치(성층권, 대류권)에 따라 우리에게 미치는 영향이 달라지는데, 지상에서 약 15~50 km 상공의 성층권에서 생성되는 성층권 오존과 지표면에서 NO_2의 광화학 반응에 의해

생성되는 대류권 오존(지표 오존)으로 구분할 수 있다(Daniel, 1999). 지구상의 오존 중 90%가량은 성층권에 분포하고 있으며, 이미 잘 알려져 있는 바와 같이 성층권 오존은 자외선을 흡수하여 생명체의 보호막 역할을 하고 있다. 반면 약 10% 정도는 지상으로부터 10 km 상공 사이인 대류권에 분포하고 있는데, 성층권에 위치한 오존과 달리 대기오염물질로서의 대류권 오존은 대기 중에 배출된 질소산화물(NO_X)과 휘발성유기화합물(VOCs) 등 오존 전구물질이 자외선과 광화학 반응을 일으켜 생성된 대표적인 2차 오염물질이다. 대류권 오존은 산화력이 매우 강한 오염물질로서 인체 감각 및 호흡기 계통에 영향을 주며, 특히 고농도 노출 시, 만성호흡질환, 기침, 폐활량 감소 등을 유발한다(Lim *et al.*, 2019; Cohen *et al.*, 2017; Turner *et al.*, 2016). 이로 인해 경제적 질병 부담도 증가하고 있는데, 한 연구 결과에 따르면 오존으로 인한 비사고 사망 비용은 2018년 62억 2,800만 원에서 이듬해 76억 4,400만 원으로 증가하였다(Kim *et al.*, 2024). 또한 오존은 작물 생장에도 피해를 주며(Feng *et al.*, 2019; Lapina *et al.*, 2016), 지구온난화에도 기여하기 때문에(Unger *et al.*, 2010) 배출 저감(emission reduction)과 노출 저감(exposure reduction) 등 좀 더 체계적인 오존 관리방안이 필요하다.

서울지역은 고농도 오존에 따른 피해를 줄이고, 신속한 상황전파를 위해 1995년부터 4개 권역(동북권, 서북권, 서남권, 동남권)을 대상으로 오존경보제를 실시하였다. 2011년부터는 발령권역을 5개 권역(도심권, 동북권, 서북권, 서남권, 동남권)으로 조정하여 25개 도시대기측정소 중 1시간 평균농도가 1개소라도 발령기준을 만족하면 해당 권역에 발령이 이루어지고 있다. 하지만 기온상승, VOCs 등 전구물질 배출 증가에 따라 오존 농도는 지속적으로 증가추세를 보이고 있으며(SIHE, 2022), 점차 국내 오존 주의보 첫 발령일이 빨라지고, 발령횟수(일수) 및 주의보 지속시간 또한 길어지고 있어 단순히 오존, 오존 전구물질 및 2차 생성물질 등의 농도변화에 대한 감시만으로는 오존 관리를 위한 적극적인 대처로 보기 어렵다. 동시에 오존오염에 대한 시민들의 관심이 증가되고 있어(Ilaria *et al.*, 2024), 시민들의 알권리 충족에 부합될 수 있는 형태의 이해하기 쉽고, 고농도 오존으로부터 시민들의 건강을 보호하기 위한 신속한 대응책 마련이 무엇보다 중요해졌다. 또한 오존은 2차 생성 대기오염물질로, 배출량뿐만 아니라 기온, 습도, 일사량 등 다양한 기상 조건에도 크게 영향을 받는 등 복잡한 화학반응과 생성 메커니즘으로 인해 오존농도 분석 및 이해하는 데 어려움이 따르며, 고농도 발생과 관련된 원인 규명에서도 많은 시간이 소요된다. 특히, 최근에는 통계적 방법을 이용해서 전국 또는 지역 규모의 오존 예측모형을 활용하고 있는 국외와 달리(Kazemiparkouhi *et al.*, 2020; de Hoogh *et al.*, 2018), 국내 오존의 예측모형에 관한 연구는 주로 물리화학 모형에 기반해서 진행된 것으로(Bae *et al.*, 2018; Kang *et al.*, 2016; Kim, 2011) 특정 과정의 정확한 묘사에 집중하여 모델의 구축과 실행이 매우 복잡하다. 또한 연구기간이나 장소, 지역 등을 감안했을 때, 예측 속도와 실시간 대응 능력이 다소 떨어져 장기적인 예측이나 지역적 제한 등 모델 적용에 있어서 한계가 있다.

따라서 본 연구에서는 배출 저감과 노출 저감으로 구분되는 오존 관리방안 중, 노출 저감에 초점을 맞추어 대량의 데이터 분석을 통한 패턴 인식 등 통계적 방법을 활용한 연구를 수행하였다. 우선 기존에 단순한 오존 농도 감시에서 벗어나 이미 오랜 기간 축적되어 온 방대한 도시대기측정망 대기질자료(서울시), 기상자료(기상청) 등을 활용하여 고농도 오존 발생 시 즉, 오존 주의보(시간당 평균 0.12 ppm 이상) 발령 시, 발령 여부에 영향을 주는 주요 영향인자를 도출하였다. 특히, 오존에 대한 시민들의 관심 증가로 인해 고농도 오존 발생 영향요인 등을 직관적으로 쉽게 이해할 수 있는 형태로 표현하기 위해서 머신러닝(machine learning, 기계학습) 기반의 분류모델을 적용하였다. 아울러 최근에는 오존 노출 저감을 위한 신속한 상황전파 등 대응책 마련의 필요성이 강조되기 때문에, 시민들의 오존 노출을 최소화하기 위한 대책 중 하나로서 분류모델과 주요 영향인자 예보 결과(값, 등급)를

Fig. 1. 25 AQMNs and Seoul AWS (topography basemap of Seoul).

활용하여 사전에 고농도 오존 발생을 예측하고자 하였다.

2. 연구 방법

2.1 자료수집

본 연구에서는 2023년 5~9월 기간(실제 서울지역 오존 주의보 발령일 기준) 서울시 내 25개 도시대기측정망(AQMNs, air quality monitoring networks), 서울(종로구)기상관측소(AWS, automatic weather station), 기상청(www.weather.go.kr)에서 생산되는 도시대기질 및 기상자료 시간데이터를 수집, 분석하여 첫 번째 분류모델(이하 '분석모델(analysis model, ANM)')을 구축하고, 이를 통해 고농도 오존 발생에 영향을 주는 주요 인자들을 우선 도출하였다. 이후 ANM에서 도출된 주요 영향인자 중 당일 오후(17시 기준) 예보 결과(값, 등급)로 제공되는 인자들을 입력변수로 하는 두 번째 분류모델(이하 '예측모델(prediction model, PRM)')을 개발함으로써 다음 날(0~23시) 오존 주의보가 발령될 가능성을 분석하고 전망하였다. 연구대상지점은 그림 1과 같으며, 데이터는 종속변수로 O_3 1항목, 독립변수로는 $PM_{2.5}$, PM_{10}, NO_X, 풍속, 기온, 상대습도, 강수량, 강수 유무, 일조시간, 일사량, 전운량(10분위) 및 다음 날 예보값(기온, 상대습도, 강수확률, 전운량), 대기질 예보등급($PM_{2.5}$_rate, PM_{10}_rate, O_3_rate) 18항목 등 총 19항목으로 구성하였다.

2.2 분석 방법

본 연구에서는 R-3.6.1 프로그램을 이용하여 2가지 목적(고농도 오존 발생 주요 영향요인 선정, 다음 날 오존 주의보 발령 예측)을 달성하기 위한 분류모델을 구축하였다. 분류모델은 변수 중요도 산정, 영향요인 분석 등 다양한 기능으로 인해 많이 활용되고 있으며, 최근에는 기계학습 기반의 조류예보제 적용(Muttil and Chau, 2006), 시계열 데이터를 활용한 대기오염물질 농도 예측(Snezhana *et al.*, 2019) 등 환경 분야뿐만 아니라 대량의 문서를 효율적으로 관리하고 검색하기 위한 자동문서분류(Ruiz and Padmini, 2002), 다양한 사람의 생체데이터를 실시간으로 수집하고, 여기서 획득한 생체신호 및 패턴정보를 이용한 감정분류(Murugappan *et al.*, 2013) 등 다양한 분야에서 높은 정

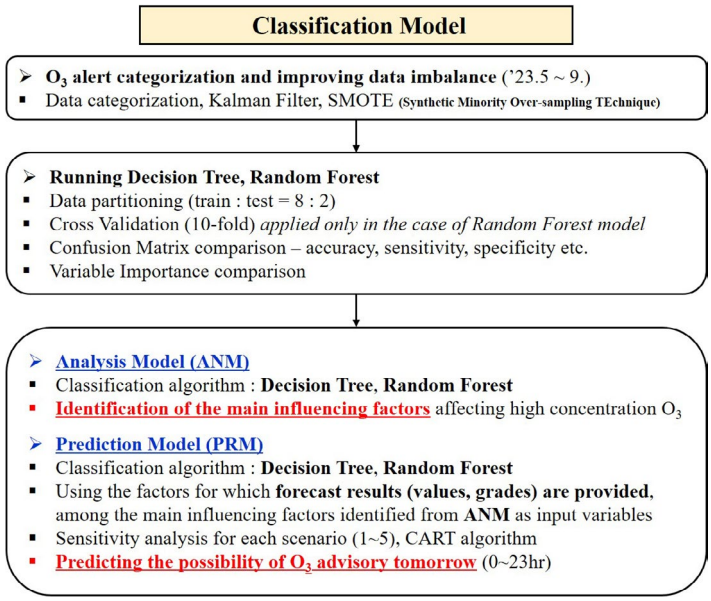

Fig. 2. Flow chart for the development of classification model.

확성을 갖는 머신러닝(machine learning) 기반의 분류모델이 폭넓게 활용되고 있다. 그림 2와 같이 종속변수 범주화[범주 0(주의보 미발령), 범주 1(주의보 발령)]를 실시하였고, 데이터 결측값을 대체하기 위한 칼만 필터(Kalman filter)를 적용하였다. 칼만 필터는 유효 데이터 개수를 확보하기 위하여 항목별 결측값들에 대해 이전 시간 측정치를 바탕으로 상태 예측(state prediction)과 측정 업데이트(measurement update)를 통해 결측값을 추정하는 방법이다(Greg and Gary, 1995). 또한 2개 분류모델(ANM, PRM)의 훈련(학습)이 효과적으로 진행될 수 있도록 SMOTE (Synthetic Minority Over-sampling TEchnique) 기법을 적용하여 데이터 불균형을 개선하였다. 여기서 SMOTE 기법은 비율이 낮은 범주(minor class)의 샘플을 만들어내는 방법으로써, 낮은 비율의 범주 샘플과 K 최근접 이웃(K nearest-neighbor, default K=5)을 활용하여 새로운 샘플을 만드는 알고리즘이다(Chawla *et al.*, 2002). 이를 통해 ANM 모델에서는 SMOTE 전처리 후 주의보 '미발령' 및 '발령' 케이스 개수는 모두 83,968개로(SMOTE 전처리 이전, 주의보 '미발령' 및 '발령' 케이스 개수 각각 83,968개, 332개), PRM 모델에서도 SMOTE 전처리 후의 주의보 '미발령' 및 '발령' 케이스 개수는 각각 83,393, 83,332개로 데이터 균형을 이루었다(SMOTE 전처리 이전, 주의보 '미발령' 및 '발령' 케이스 개수 각각 83,393개, 332개). 그리고 전체 데이터를 학습 데이터(train data) 80%, 테스트 데이터(test data) 20% 비율로 분할하고, 특히 랜덤포레스트 알고리즘의 경우, 최적의 분류모델 구축을 위해 대표적 교차검증 방법 중 하나인 10-fold 교차검증을 적용하였다. 이는 10-fold 교차검증이 전체 데이터를 크기가 동일한 10개의 하부집합(subset)으로 나누고, 10번째 하부집합을 검증용(validation) 자료로, 나머지 9개 하부집합을 훈련용(training) 자료로 사용, 이를 10번(round 1~10) 반복측정하고 각각의 반복측정 결과를 분류모델에 적용하는 평가 방법이기 때문이다. 이처럼 데이터 전처리 후, 의사결정나무(Decision Tree), 랜덤포레스트(Random Forest) 등 2가지 분류모델 알고리즘을 적용하여 변수 중요도를 산정하고, 분류모델 성능평가를 통해 서울지역 고농도 오존 발생에 따른 주요 영향인자 도출을 위한 분석모델(ANM)을 구축하였다.

Table 1. $PM_{2.5}$, PM_{10}, O_3, Cloud cover forecast rate and probability scoring table.

Item	Classification		Forecast rate			
			Good	Moderate	Unhealthy	Very unhealthy
$PM_{2.5}$ ($\mu g/m^3$)	Concentration range		0~15	16~35	36~75	Over 75
	Scoring (1~10)	Scenario 1	1	3	7	10
		Scenario 2	1	3	8	10
		Scenario 3	1	3	8	10
		Scenario 4	1	4	8	10
		Scenario 5	1	3	7	10
PM_{10} ($\mu g/m^3$)	Concentration range		0~30	31~80	81~150	Over 150
	Scoring (1~10)		1	4	8	10
O_3 (ppm)	Concentration range		0~0.0300	0.0301~0.0900	0.0901~0.1500	Over 0.1500
	Scoring (1~10)	Scenario 1	1	5	8	10
		Scenario 2	1	5	8	10
		Scenario 3	1	5	9	10
		Scenario 4	1	5	8	10
		Scenario 5	1	4	8	10
Item	Classification		Forecast probability			
			Clear	Partly cloudy	Mostly cloudy	Cloudy
Cloud cover	Scoring (1~10)		1	4	7	10

예측모델(PRM)은 당일 오후(17시 기준) 기상청에서 발표된 기온, 상대습도, 강수확률, 전운량 4항목의 예보값과 분석모델(ANM)에서 선정된 주요 영향인자 중 환경부에서 예보된 대기질 항목의 예보등급($PM_{2.5}$_rate, PM_{10}_rate, O_3_rate)을 독립변수로 하여 다음 날 오존 주의보 발령 가능성을 예측 및 전망하는 모델이다. 특히 예측모델(PRM)에서는 표 1과 같이 $PM_{2.5}$, PM_{10}, O_3, 전운량 등 4항목 예보 결과에 대해서 시나리오별로 산술 계산하여 점수화(1~10점)한 후, 이 중 CART 알고리즘 기반의(Brian *et al.*, 2014) 의사결정나무 모델을 이용한 민감도 분석을 통해 최적의 시나리오를 적용함으로써 오존 주의보 발령 전망의 정확성을 높이고자 하였다.

일반적으로 분류모델 예측 결과는 분류되는 범주로 나타나기 때문에 분류모델의 성능평가는 분류결과표(confusion matrix)를 주로 사용한다. 따라서 본 연구에서도 분류모델 성능평가를 위해 표 2와 같이 분류결과표를 사용하였다. 분류결과표에서 TP (true positive)는 실측값과 예측값 모두 true인 빈도, TN (true negative)은 실측값과 예측값 모두 false인 빈도, FP (false positive)는 실측값은 false이나 true로 예측한 빈도, FN (false negative)은 실측값은 true이나 false로 예측한 빈도를 나타낸다(Kim, 2018) 성능평가 지표로는 식 (1)~(3)과 같이 모델의 전체적인 예측성능을 나타내는 정확도(accuracy), 오존 주의보 미발령 범주 0에 대한 모델의 예측성능을 나타내는 민감도(sensitivity), 오존 주의보 발령 범주 1에 대한 모델의 예측성능을

Table 2. Confusion matrix of decision tree model.

		Observation	
		No advisory (0)	Advisory (1)
Prediction	No advisory (0)	TP	FP
	Advisory (1)	FN	TN

나타내는 특이도(specificity) 등을 사용하였으며, 지표 값이 클수록 예측성능이 좋은 모형이다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

2.2.1 의사결정나무(Decision Tree)

의사결정나무는 분류기법 중의 하나로써, 의사결정 규칙을 도표화하여 목표대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 계량적 분석 방법이다(Choi and Seo, 1999). 일반적으로 의사결정나무 모델에서는 에러율(error rate)이 가장 낮을 때의 복잡성 매개변수(cp, complexity parameter)를 이용하여 가지치기를 실시하여 변수들 중 가장 설명력이 있는 변수에서 최초로 분리가 일어난다. 따라서 분석 결과가 나무구조로 표현되어 결과를 쉽게 이해하고 설명할 수 있으며, 의사결정을 하는 데 직접적으로 사용할 수 있기 때문에 데이터마이닝(data mining) 적용 시 매우 많이 사용되고 있다. 또한 두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지 쉽게 알 수 있으며, 비모수적 모형에도 적용 가능하기 때문에 선형성(linearity), 정규성(normality), 등분산성(equal variance) 등의 가정이 필요하지 않은 장점을 갖고 있다. 반면에 데이터 특성이 특정 변수에 수직 또는 수평적으로 구분되지 못할 때에는 분류율이 떨어지고, 나무가 복잡해지는 문제가 발생한다. 또한 변수들이 비슷한 수준의 정보력을 갖는 경우, 약간의 차이에 의해서도 다른 변수가 선택되는 나무구조로 바뀌거나, 새로운 자료의 예측에는 불안정(unstabile)하는 등 최적의 해를 보장하지 못하는 단점이 있다.

2.2.2 랜덤포레스트(Random Forest)

랜덤포레스트는 복수의 의사결정나무를 형성하여 새로운 데이터들을 각각의 나무에 동시에 통과시켜, 각각의 나무로부터 도출된 분류 결과에 대해서 투표를 실시하여 가장 많이 득표한 결과를 최종 분류결과로 선택하는 모델이다. 따라서 기존의 의사결정나무에서 가지치기로는 충분히 해결하기 어려운 과적합(overfitting) 문제를 해결하는 데 적합한 좀 더 일반화된 나무모형이라 할 수 있다(Kim *et al.*, 2019). 주요 특징으로는 배깅(bagging), 무작위 입력변수 선택, OOB(out of bags) 오차율, 변수 중요도(variable importance)가 있다. 특히 OOB 오차율은 랜덤포레스트 모델의 성능을 검증하는 데 활용되며, 동시에 OOB 데이터를 통해 예측한 값에 대한 입력변수의 중요도, 즉 변수 중요도를 평가할 수도 있다(Louppe, 2014; Breiman, 2001).

3. 결과 및 고찰

3.1 2023년 서울지역 오존 주의보 발령 현황

2023년 오존 연평균 농도는 0.0310 ppm, 5~9월 기간 오존 주의보 발령횟수(일수)는 45회(14일)로, 2019년(오존 연평균 농도 0.025 ppm) 29회(11일), 2020년(0.025 ppm) 30회(12일), 2021년(0.028 ppm) 32회(11일), 2022년(0.029 ppm) 42회(11일)와 비교했을 때 매년 연평균 오존 농도 및 오존 주의보 발령횟수(일수)가 증가하고 있다. 특히, 향후에도 기온상승, VOCs와 같은 오존생성 전구물질 배출 증가 등으로 인해 조기 발령 및 고농도 오존 지속시간이 증가될 것으로 전망되고 있어 효율적인 오존 대응 정책 추진이 무엇보다 중요해졌다.

3.2 고농도 오존 발생에 따른 주요 영향인자 선정 모델 구축 - 분석모델(ANM)

본 연구에서는 $PM_{2.5}$, PM_{10}, NO_X, 풍속, 기온, 상대습도, 강수량, 강수 유무, 일조시간, 일사량, 전운량(10분위) 등 총 11개 항목을 독립변수로 하고, 종속변수인 25개 도시대기측정망 오존을 범주화하였다. SMOTE 기법을 통한 데이터 불균형 개선작업 후, 의사결정나무, 랜덤포레스트 알고리즘을 적용하여 서울지역 고농

Table 3. O_3 advisory current status and regions (2023).

Year	Advisory number	Advisory day	Region	O_3 advisory					
				May	June	July	August	September	Total
2023	45	14	Central	1	2	2	3	0	8
			Northwestern	1	2	2	5	1	11
			Northeastern	2	1	1	3	0	7
			Southwestern	3	3	2	2	1	11
			Southeastern	0	1	4	2	1	8
			Total	7	9	11	15	3	45
	Advisory region		▶ Central (Jongno, Jung, Yongsan) ▶ Northwestern (Mapo, Seodaemun, Eunpyeong) ▶ Northeastern (Seongdong, Gwangjin, Dongdaemun, Jungnang, Seongbuk, Gangbuk, Dobong, Nowon) ▶ Southwestern (Yangcheong, Gangseo, Guro, Yeongdeungpo, Geumcheon, Dongjak, Gwanak) ▶ Southeastern (Seocho, Gangnam, Songpa, Gangdong)						

Table 4. Confusion matrix and main evaluation index of decision tree model (1).

		Observation		Evaluation index				
		No advisory (0)	Advisory (1)	Accuracy	Sensitivity	Specificity	Cohen's kappa	P-value
Prediction	No advisory (0)	15,955	148	0.9676	0.9444	0.9911	0.9352	<2.2e-16
	Advisory (1)	939	16,485					
Sum		16,894	16,633					

도 오존 발생(오존 주의보 발령)에 영향을 주는 주요 인자들을 도출하였다.

3.2.1 의사결정나무 모델 결과(1)

본 연구에서 의사결정나무 모델의 분류성능을 나타내는 분류성능표 및 주요 평가지표 산정 결과는 표 4를 통해 확인할 수 있다. 실제 오존 주의보 발령 개수 16,633개 중 발령 예측은 16,485개, 미발령 예측은 148개로, 실제 오존 주의보 미발령 개수 16,894개 중 발령 예측은 939개, 미발령 예측은 15,955개로, 이를 통해 정확도 0.9676, 민감도 0.9444, 특이도 0.9911로 나타났다. 동시에 두 개의 범주형 자료 일치도(agreement)를 나타내는 코헨의 카파(kappa, 1에 가까울수록 일치도가 높음) 또한 0.9352로, 전반적으로 범주 0 (오존 주의보 미발령), 범주 1 (오존 주의보 발령)의 분류가 적절하게 이루어졌음을 확인할 수 있었다.

의사결정나무 모델을 통한 고농도 오존에 대한 영향인자를 도출한 결과, 그림 3과 같이 $PM_{2.5}$, PM_{10}, 기온이 오존 주의보 발령 여부에 중요한 변수로 작용함을 판단할 수 있었다. 일반적으로 기온 25°C 이상, 상대습도 75% 이하, 구름이 없는 쾌청한 날씨, 낮은 풍속, 높은 일사량 등의 조건 시 고농도 오존이 발생할 가능성이 크다고 알려져 있다(ME, 2016). 하지만 본 연구에서는 5~9월이라는 특정 기간의 시간데이터를 이용하여 분류모델을 적용하였기 때문에 해당 기간 내 풍속, 상대습도, 일사량 등은 상대적으로 유사하였다. 따라서 실제로 오존 주의보 발령 여부에는 중요한 영향인자로 선정되지 않았으며, 오히려 그림 3, 표 5와 같이 (초)미세먼지 농도에 따라 발령 여부가 결정되는 결과를 확인할 수 있었다.

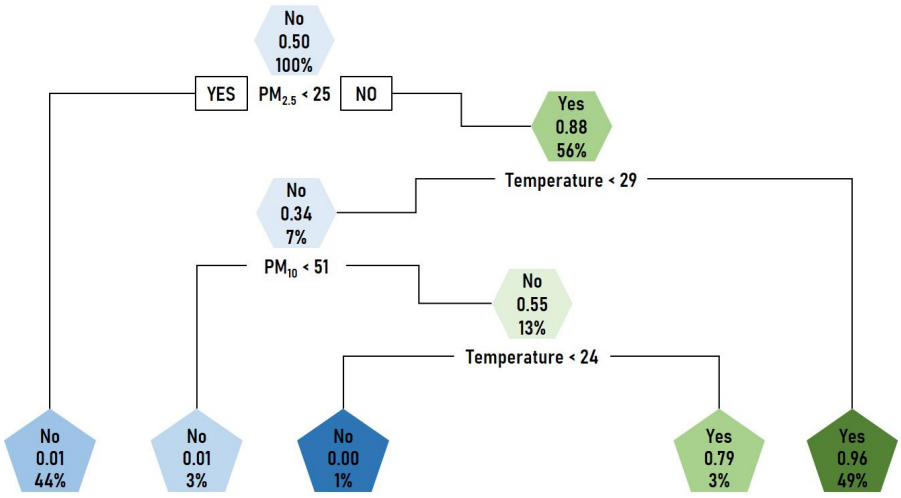

Fig. 3. Decision tree (1) for high concentration O_3 events in Seoul.

Table 5. Cases affecting high concentration O_3 events in Seoul.

Case	$PM_{2.5}$ (μg/m^3)	Temperature (℃)	PM_{10} (μg/m^3)
1	≥25	≥29	–
2	≥25	24＜Temperature＜29	≥51

전체 오존 주의보 발령 사례 중 49%를 차지하는 Case 1의 경우 즉, $PM_{2.5}$ 농도, 기온이 각각 25 μg/m^3 이상, 29℃ 이상이 되었을 때 오존 농도는 주의보 발령 기준을 충족하였으며, 또한 기온은 Case 1보다 다소 낮은 24~29℃이지만 $PM_{2.5}$ 농도 25 μg/m^3 이상, PM_{10} 농도 51 μg/m^3 이상 등 미세먼지 농도가 높은 Case 2 (3%)인 경우에도 오존 주의보 발령은 이루어졌다. 이처럼 서울지역 여름철 기간 동안 오존 주의보 발령은 대부분 Case 1, 2와 같이 (초)미세먼지 농도가 특정 범위 만족 여부에 따라 이루어질 가능성이 높을 것으로 판단된다. 따라서 여름철 하절기에 오존 주의보 발령에 따른 신속한 상황전파를 위해서는 기온, 강수 유무 등 기상조건과 함께 (초)미세먼지 농도 현황 등을 세심하게 살펴봐야 할 것이다.

3.2.2 랜덤포레스트 모델 결과(1)

랜덤포레스트 모델은 의사결정나무 모델의 학습과정에서 과적합을 방지하기 위해 여러 개의 의사결정나무를 활용한 머신러닝 모델로서, 특히 많은 양의 데이터 셋에서도 잘 작동하며 높은 성능을 내기 때문에 대부분의 분류모델에 사용되고 있다. 본 연구에서도 랜덤포레스트 모델 결과와 기존 의사결정나무 모델에서 도출된 고농도 오존에 대한 주요 영향인자들을 비교·분석하였다.

랜덤포레스트 모델에서 OOB (out of bags) 오차율이 가장 낮은 상태가 되는 나무개수(ntree)를 결정하기 위하여 우선 초기값을 600으로 지정하여(default 500) 예측성능이 가장 좋은 랜덤포레스트 모델을 도출하였으며, 그 결과, 나무개수가 100 이후부터는 OOB 오차율이 0.19%로 거의 일정하게 안정된 값을 보였다. 이는 변수 중요도 산정을 위한 랜덤포레스트 모델은 충분히 신뢰할 수 있는 분류모형임을 의미하며, 따라서 본 연구에서는 나무개수를 100으로 하는 등 정확도가 가장 높고, OOB 오차율은 가장 낮게 되도록 표 6과 같은 초매개변수로 랜덤포레스트 모델을 구축하여 분석을 진행하였다.

랜덤포레스트 모델의 분류성능을 보여주는 분류성능표 및 주요 평가지표 산정 결과는 표 7과 같다. 표 7을 보면, 정확도 0.9985, 민감도 0.9971, 특이도 0.9999, 코헨의 카파(kappa) 0.9970으로, 전반적으로 범주 0,

범주 1의 분류가 적절하게 이루어졌으며, 특히 의사결정나무 모델 결과에 비해 상당히 좋은 분류성능을 보이는 결과를 보였다. 따라서 본 연구와 같이 많은 독립변수들 중 분류모델을 통해서 종속변수 변화에 영향을 주는 주요 인자들은 무엇인지, 또는 종속변수 예측모델 구축을 위한 독립변수를 무엇으로 할지 결정하고자 할 때에는 데이터 양, 분석시간, 요구되는 정확도 등을 고려하여 의사결정나무 모델보다 일반적으로 정확성이 높은 랜덤포레스트 모델을 우선 활용하여 영향인자 및 독립변수를 도출하는 것이 타당하다고 판단된다.

랜덤포레스트 모델을 통한 고농도 오존에 대한 영향인자를 도출한 결과를 도트(dot) 형식으로 플로팅(plotting)한 변수의 중요도는 그림 4와 같다. 랜덤포레스트의 변수 중요도를 측정하는 지표에는 평균정확도 감소(MDA, Mean Decrease Accuracy)와 평균불순도

Table 6. Main hyper-parameters for random forest model (1).

Hyper-parameter	trainControl	ntree	mtry	nodesize	importance
Value	method = cv number = 10	100	6	5	TRUE

Table 7. Confusion matrix and main evaluation index of random forest model (1).

		Observation		Evaluation index				
		No advisory (0)	Advisory (1)	Accuracy	Sensitivity	Specificity	Cohen's kappa	P-value
Prediction	No advisory (0)	16,845	2	0.9985	0.9971	0.9999	0.9970	< 2.2e-16
	Advisory (1)	49	16,631					
Sum		16,894	16,633					

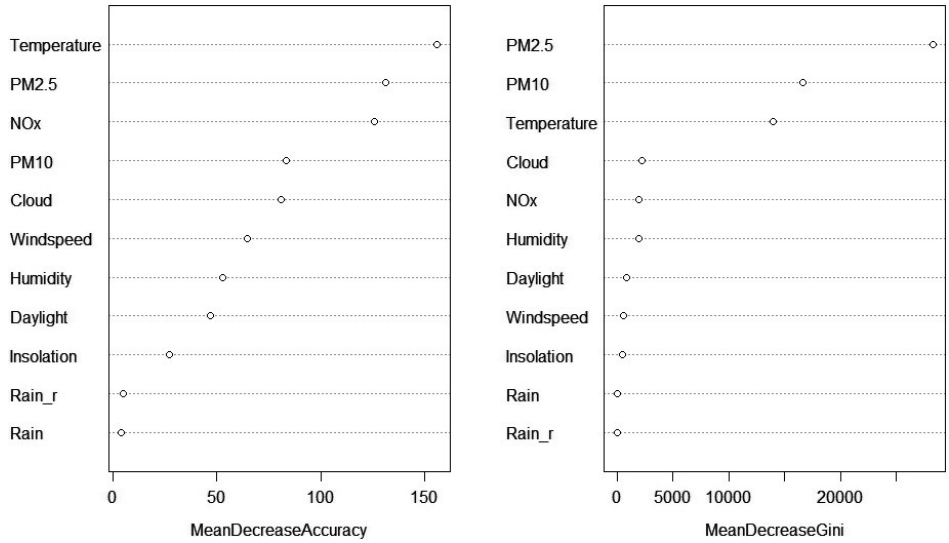

Fig. 4. Rank of factors affecting high concentration O_3 events in Seoul.

감소(MDG, Mean Decrease Gini (in node impurity))가 있으며, 두 지표 모두 값이 커질수록 변수의 중요도가 높아지게 된다. 즉, 그림 4에서 그래프 상단에 위치할수록 고농도 오존에 미치는 중요도가 크다고 할 수 있다. 따라서 전반적으로 기온, (초)미세먼지, 질소산화물이 주로 그래프 상단에 위치하는 등 랜덤포레스트 모델 결과를 통해 의사결정나무 모델과 유사한 주요 영향인자가 도출되었다.

3.3 오존 주의보 발령 예측모델 구축 – 예측모델(PRM)

본 연구에서는 오존 생성과 연관성이 높은 기온, 상대습도, 강수확률, 전운량과 대기질 3개 항목의 예보등급($PM_{2.5}$_rate, PM_{10}_rate, O_3_rate) 등 총 7개 항목을 독립변수로 하여 다음 날 서울지역 오존 주의보 발령 가능성을 예측하는 예측모델(PRM)을 구축하였다. 특히, 표 1과 같이 시나리오 1~5까지 점수화를 달리 조정하면서 의사결정나무 모델을 통해 민감도 분석을 실시하였으며, 그 결과 시나리오 1 조건으로 점수화한 결과 민감도(sensitivity)와 정확도(accuracy)가 가장 양호하였다.

3.3.1 의사결정나무 모델 결과(2)

다음 날 오존 주의보 발령 가능성을 예측하기 위하여 의사결정나무 모델을 실행한 결과(표 8), 정확도 0.8991, 민감도 0.9233, 특이도 0.8747, 코헨의 카파(kappa) 0.7891로, 전반적으로 분류성능이 양호하게 나타나, 예측모델(PRM)을 통해 다음 날 오존 주의보 발령 가능성을 예측할 수 있을 것으로 판단된다. 또한 그림 5에서 보듯이, 의사결정나무 모델을 통한 다음 날 오존 주의보 발령에 대한 주요 영향인자로는 오존 및 초미세먼지 예보등급, 상대습도, 기온이 오존 주의보 발령 여부에 중요한 변수로 작용했음을 확인할 수 있었다.

전체 오존 주의보 발령 사례 중 25%를 차지하는 Case 1의 경우 당일 17시 기준 O_3, $PM_{2.5}$ 예보등급이 각각 "나쁨" 이상일 때 다음 날 오존 주의보 발령 가능성이 높았다. 또한 표 9에서 보듯이 $PM_{2.5}$ 예보등급이 Case 1보다 낮은 "보통" 이하이지만 Case 2, Case 3과 같이 상대습도, 기온, PM_{10} 예보등급, 강수확률 조건 등에 따라 다음 날 오존 주의보 발령 가능성을 확인할 수 있었다. 따라서 앞서 언급하였듯이, 여름철 하절기에 오존 주의보 발령 가능성을 보다 정확히 예측하고, 신속 대응을 위한 상황전파를 위해서는 단순히 오존 예보등급만을 가지고 판단하기보다는 오존 예보등급뿐만 아니라 기존에 제공되고 있는 (초)미세먼지 예보등급 또한 함께 고려해야 할 것이다.

Table 8. Confusion matrix and main evaluation index of decision tree model (2).

		Observation		Evaluation index				
		No advisory (0)	Advisory (1)	Accuracy	Sensitivity	Specificity	Cohen's kappa	P-value
Prediction	No advisory (0)	15,438	2,083	0.8991	0.9233	0.8747	0.7891	< 2.2e-16
	Advisory (1)	1,283	14,541					
Sum		16,721	16,624					

Table 9. Cases affecting O_3 advisory prediction model of 25 AQMNs.

Case	O_3 (score)	$PM_{2.5}$ (score)	Humidity (%)	Temp. (°C)	PM_{10} (score)	Rain prob. (%)
1	≥5	≥5				
2	≥5	2≤$PM_{2.5}$<5	≥70	≥22	<2.5	
3	≥5	2≤$PM_{2.5}$<5	≥70	≥22	≥2.5	<10

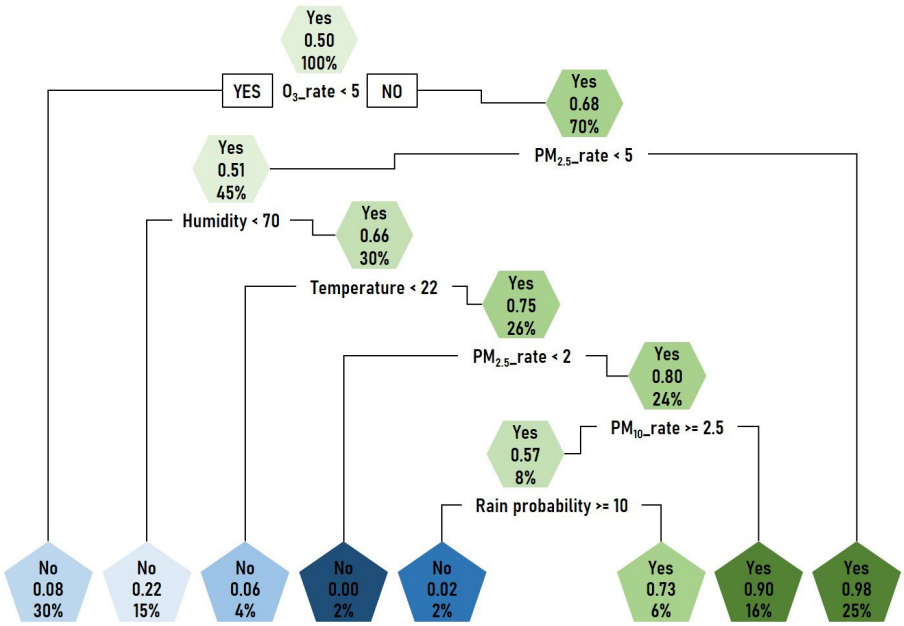

Fig. 5. Decision tree (2) for O_3 advisory prediction model of 25 AQMNs.

Table 10. Main hyper-parameters for random forest model (2).

Hyper-parameter	trainControl	ntree	mtry	nodesize	importance
Value	method = cv number = 10	100	7	5	TRUE

Table 11. Confusion matrix and main evaluation index of random forest model (2).

		Observation		Evaluation index				
		No advisory (0)	Advisory (1)	Accuracy	Sensitivity	Specificity	Cohen's kappa	P-value
Prediction	No advisory (0)	16,339	137	0.9844	0.9772	0.9918	0.9689	< 2.2e-16
	Advisory (1)	382	16,487					
Sum		16,721	16,624					

3. 3. 2 랜덤포레스트 모델 결과(2)

다음 날 오존 주의보 발령 가능성을 예측하기 위한 랜덤포레스트 모델 구축 결과, 나무개수 100 이후부터 OOB 오차율 1.56%로 거의 일정한 값을 보이는 등 표 10과 같은 초매개변수로 랜덤포레스트 모델을 구축하여 분석을 진행하였다.

랜덤포레스트 모델 실행 결과, 분류성능표 및 주요 평가지표 산정 결과는 표 11과 같다. 정확도 0.9844, 민감도 0.9772, 특이도 0.9918, 코헨의 카파(kappa) 0.9689로, 분류성능이 우수하였으며, 의사결정나무 모델 대비 정확성이 약 9.5% 향상되는 등 분류성능이 우수하여 다음 날 오존 주의보 발령 가능성을 예측할 수

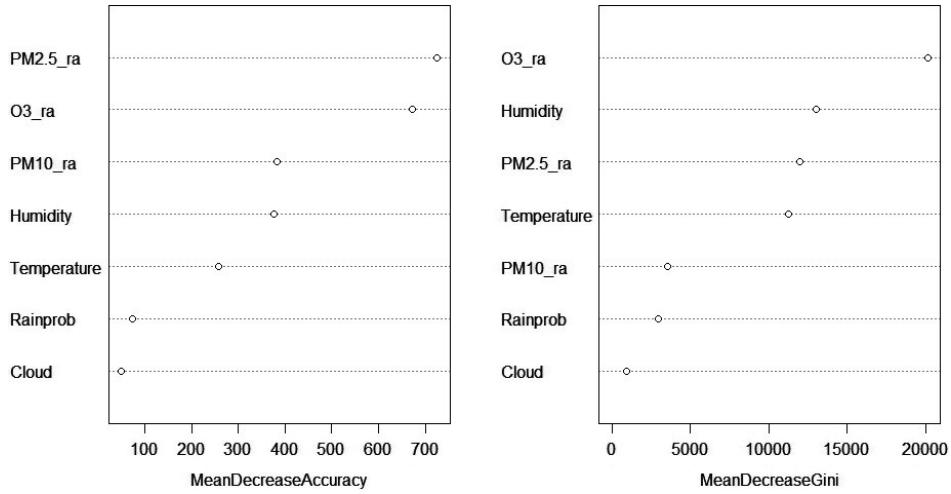

Fig. 6. Rank of factors affecting O_3 advisory prediction model of 25 AQMNs.

있을 것으로 판단된다.

랜덤포레스트 알고리즘을 적용한 예측모델(PRM)을 통해서 다음 날 오존 주의보 발령에 미치는 영향인자 즉, 변수 중요도는 그림 6과 같다. 앞서 언급하였듯이 그림 6에서 그래프 상단에 위치할수록 해당 인자는 다음 날 오존 주의보 발령에 미치는 영향 정도가 크다고 할 수 있다. 따라서 전반적으로 의사결정나무 모델 결과처럼 당일 오후(17시)의 오존 예보등급(O_3_rate)과 (초)미세먼지 예보등급($PM_{2.5}$_rate, PM_{10}_rate)이 다음 날 오존 주의보 발령 여부에 큰 영향을 주는 것으로 판단된다.

4. 결 론

본 연구에서는 서울지역 25개 도시대기측정망, 서울기상관측소를 대상으로 2023년 5~9월 기간의 총 19항목 시간데이터를 이용하여 최근에 영향요인 중요도 산정, 예보제 적용, 농도 예측 등 다양한 분야에서 높은 정확성을 갖는 머신러닝 기반의 분류모델을 구축하였다. 이를 통해 고농도 오존 발생 영향요인 도출 및 다음 날 오존 주의보 발령 가능성 예측을 위한 기초자료를 제공하고자 하였다.

의사결정나무, 랜덤포레스트 알고리즘을 적용한 분류모델을 통해 기온, (초)미세먼지 등 신뢰성 있는 고농도 오존 발생에 영향을 주는 주요 인자를 도출하였으며, 기상조건과 함께 대기질 예보등급을 활용하여 오존 주의보 발령 가능성도 사전에 예측 가능함을 확인하였다. 물론 지역적·권역별 특성, 다양한 배출원 구조 및 해당 지역 기상조건 등에 따라 도출되는 영향인자들이 달라질 수 있다. 하지만 방법론 활용 차원에서 해당 지역의 실시간 대기질·기상데이터 기반의 분류모델을 구축·적용함으로써 높은 정확성과 분석시간 단축 등 효율적인 오존관리 일환으로 고농도 오존 발생 시, 해당 지역에 맞는 과학적인 원인 분석 수행이 가능하고, 사전에 오존 주의보 발령 여부를 예측하여 신속하게 관련 상황을 전파할 수 있을 것으로 기대된다. 향후에는 개인식별정보, 건강의료정보 등 다양한 데이터와 연계하여 오존 취약지역을 도출하고, 오존 취약지역을 중심으로 분류모델을 활용한 고농도 오존 발생 예측 및 신속한 상황전파가 이루어진다면 오존 민감군, 취약군 대상으로 오존 노출 저감 효과를 제고하는 등 다양한 오존 관리정책에 핵심적인 기초자료로서 활용될 수 있을 것으로 기대된다.

References

Bae, C.-H., Kim, B.-U., Kim, H.-C., Kim, S.-T. (2018) Quantitative assessment on contributions of foreign NO_X and VOC emission to ozone concentrations over Gwangyang Bay with CMAQ-HDDM simulations, Journal of Korean Society for Atmospheric Environment, 34(5), 708-726, (in Korean with English abstract). https://doi.org/10.5572/KOSAE.2018.34.5.708

Breiman, L. (2001) Random Forests, Machine Learning, 45, 5-32.

Brian, M., Mark, F., Pei, Y.L., Deborah, M. (2014) Use of CHAID Decision Trees to Formulate Pathways for the Early Detection of Metabolic in Young Adults, Computational and Mathematical Methods in Medicine, 2014, 242717. https://doi.org/10.1155/2014/242717

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321-357. https://www.jair.org/index.php/jair/article/view/10302/24590

Choi, J.H., Seo, D.S. (1999) Decision Trees and Its Application, Statistical Analysis Resarch, 4(1), 61-83.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R. (2017) Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015, The Lancet, 389(10082), 1907-1918. https://doi.org/10.1016/S0140-6736(17)30505-6

Daniel, J.J. (1999) Introduction to Atmospheric Chemistry, Princeton University Press, New Jersey, 85-107.

de Hoogh, K., Chen, J., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Klompmaker, J., Martin, R.V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Brunekreef, B., Hoek, G. (2018) Spatial $PM_{2.5}$, NO_2, O_3 and BC models for Western Europe - Evaluation of spatiotemporal stability, Environment International, 120, 81-92. https://doi.org/10.1016/j.envint.2018.07.036

Feng, Z., Kobayashi, K., Li, P., Xu, Y., Tang, H., Guo, A., Paoletti, E., Calatayud, V. (2019) Impacts of current ozone pollution on wheat yield in China as estimated with observed ozone, meteorology and day of flowering, Atmospheric Environment, 217, 116945. https://doi.org/10.1016/j.atmosenv.2019.116945

Greg, W., Gary, B. (1995) An Instruction to the Kalman Filter, University of North Carolina, North Carolina. https://dl.acm.org/doi/book/10.5555/897831

Ilaria, S., Giuseppe, S., Olivia, C., Sara, M., Anna, A.A., Patrizia, S., Liliana, C., Giovanni, V., Sandra, B. (2024) Air Pollution and Climate Change: A Pilot Study to Investigate Citizens' Perception, Environments, 11(9), 190. https://doi.org/10.3390/environments11090190

Kang, Y.-H., Oh, I.-B., Jeong, J.-H., Bang, J.-H., Kim, Y.-K., Kim, S.-T., Kim, E.-H., Hong, J.-H., Lee, D.-G. (2016) Comparison of CMAQ Ozone Simulation with Two Chemical Mechanisms (SAPRC99 and CB05) in the Seoul Metropolitan Region, Journal of Environmental Science International, 25(1), 85-97, (in Korean with English abstract). https://doi.org/10.5322/JESI.2016.25.1.85

Kazemiparkouhi, F., Eum, K.D., Wang, B., Manjourides, J., Suh, H.H. (2020) Long-term ozone exposures and cause-specific mortality in a US Medicare cohort, Journal of Exposure Science & Environmental Epidemiology, 30(4), 650-658. https://doi.org/10.1038/s41370-019-0135-4

Kim, G.C. (2018) Data Analysis Semi-Professional ADsP: Complete Guide in One Book, Hwangsogeoleum Academy, (in Korean).

Kim, J.N., Chung, S.Y., Oh, I.H., Kim, J.H., Jung, E.J., Ahn, Y.J. (2024) Estimation of Burden of Disease due to Climate Change in the Republic of Korea, Public Health Weekly Report, 17(26), 1119-1142. https://doi.org/10.56786/PHWR.2024.17.26.1

Kim, M.K., Yoon, C.G., Rhee, H.P., Hwang, S.J., Lee, S.W. (2019) A Study on Predicting TDI (Trophic Diatom Index) in tributaries of Han river basin using Correlation-based Feature Selection technique and Random Forest algorithm, Journal of Korean Society on Water Environment, 35(5), 432-438. https://doi.org/10.1155/2014/242717

Kim, S.-T. (2011) Ozone simulations over the Seoul Metropolitan Area for a 2007 June episode, part V: Application of CMAQ-HDDM to predict ozone response to emission change, Journal of Environmental Science International, 27(6), 772-790, (in Korean with English abstract). https://doi.org/10.5572/KOSAE.2011.27.6.772

Lapina, K., Henze, D.K., Milford, J.B., Travis, K. (2016) Impacts of foreign, domestic, and state-level emissions on ozone-induced vegetation loss in the United States, Environmental Science & Technology, 50(2), 806-813. https://doi.org/10.1021/acs.est.5b04887

Lim, C.C., Hayes, R.B., Ahn, J., Shao, Y., Silverman, D.T., Jones, R.R., Garcia, C., Bell, M.L., Thurston, G.D. (2019) Long-term

exposure to ozone and cause-specific mortality risk in the United States, American Journal of Respiratory and Critical Care Medicine, 200(8), 1022-1031. https://doi.org/10.1164/rccm.201806-1161OC

Louppe, G. (2014) Understanding Random Forests, University of Liege.

Ministry of Environment (ME) (2016) Understanding and Preparing for Ozone, (in Korean).

Murugappan, M., Subbulakshmi, M., Bong, S.Z. (2013) Frequency band analysis of electrocardiogram (ECG) signal for human emotional state classification using discrete wavelet transform (DWT), Journal of Physical Therapy Science, 25(7), 753-759. https://doi.org/10.1589/jpts.25.753

Muttil, N., Chau, K.W. (2006) Neural network and genetic programming for modelling coastal algal blooms, International Journal of Environment and Pollution, 28(3), 223-238. https://doi.org/10.1504/IJEP.2006.011208

Ruiz, M.E., Padmini, S. (2002) Hierarchical Text Categorization Using Neural Networks, Information Retrieval, 5(10), 87-118. https://doi.org/10.1023/A:1012782908347

Seoul Metropolitan Government Research Institute of Public Health and Environment (SIHE) (2022) 2022 SEOUL AIR QUALITY REPORT, (in Korean).

Snezhana, G.G., Desislava, S.V., Maya, P.S., Atanas, V.I., Iliycho, P.I. (2019) Regression trees modeling of time series for air pollution analysis and forecasting, Neural Computing and Applications, 31, 9023-9039. https://doi.org/10.1007/s00521-019-04432-1

Turner, M.C., Jerrett, M., Pope III, C.A., Krewski, D., Gapstur, S.M., Diver, W.R., Beckerman, B.S., Marshall, J.D., Su, J., Crouse, D.L., Burnett, R.T. (2016) Long-term ozone exposure and mortality in a large prospective study, American Journal of Respiratory and Critical Care Medicine, 193(10), 1134-1142. https://doi.org/10.1164/rccm.201508-1633OC

Unger, N., Bond, T.C., Wang, J.S., Koch, D.M., Menon, S., Shindell, D.T., Bauer, S. (2010) Attribution of climate forcing to economic sectors, Proceedings of the National Academy of Sciences, 107(8), 3382-3387. https://doi.org/10.1073/pnas.0906548107

Authors Information

이진효 (서울특별시 기후환경본부 대기정책과 주무관)
(co90mp@seoul.go.kr)

사창훈 (서울특별시 기후환경본부 대기정책과 과장)
(sach5436@seoul.go.kr)

윤태호 (서울특별시 보건환경연구원 대기환경연구부 팀장)
(fuco99@seoul.go.kr)

최용석 (서울특별시 보건환경연구원 대기환경연구부 부장)
(hozer87@seoul.go.kr)

이현정 (제주대학교 환경공학과 교수)
(hyunjunglee@jejunu.ac.kr)

구자용 (서울시립대학교 환경공학부 교수)
(jykoo@uos.ac.kr)