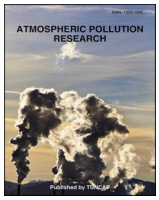
Atmospheric Pollution Research 16 (2025) 102498

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: www.elsevier.com/locate/apr



Impact of land use characteristics on air pollutant concentrations considering the spatial range of influence

Lee Gunwon [a], Han Yuhan [b], Geunhan Kim [c,*]

[a] Department of Architecture, Korea University, Seoul, 02841, Republic of Korea
[b] Department of Geoinformatics, University of Seoul, Seoul, 02504, Republic of Korea
[c] Department of Environmental Planning, Korea Environment Institute, Sejong, 30147, Republic of Korea

ARTICLE INFO

Keywords:
Urban air pollution
Land cover
Multiple regression analysis
XGBoost

ABSTRACT

Prediction models ranging from statistical probability to machine learning techniques have been employed to improve and manage urban air quality. However, the number of air quality monitoring stations (AQMS) for the collection of air quality information is limited. This study established a model that explains the relationship between six air pollutants–SO_2, CO, O_3, NO_2, PM_{10}, and $PM_{2.5}$–measured by approximately 443 AQMS in South Korea and factors, such as the vegetation index, topography, and land cover elements. The model analyzed the impact of land cover changes on air pollutant concentrations and derived scenarios predicting changes in the air quality due to land use changes. Despite the relatively small sample size of approximately 360 AQMS, multiple regression analysis demonstrated higher explanatory power compared with Xtreme Gradient Boosting, a representative machine learning technique. The optimal spatial range for explaining air pollutant concentrations varied for each air pollutant. The highest R^2 in the multiple regression analysis was 0.34 at a distance of 12,000 m for SO_2; 0.27 at 11,000 m for CO; 0.50 at 6000 m for O_3; 0.70 at 18,000 m for NO_2; 0.49 at 18,000 m for PM_{10}; and 0.48 at 11,000 m for $PM_{2.5}$. Certain land cover characteristics were found to significantly affect air quality, whereas small-scale restoration had a minimal impact on air quality improvement, and large-scale development substantially increased pollutant concentrations. This study provides essential information for urban planning and policymaking aimed at improving urban air quality.

1. Introduction

Continuous urbanization has led to the development of industries within cities and a proportional increase in population, which increases automobile traffic. Urbanization has caused rapid changes in land use within cities, leading to environmental issues, such as air quality deterioration (Balew and Korme, 2020). Urban air pollution significantly negatively affects the quality of life and health of urban residents (Faiz, 1993; Akimoto, 2003). Large cities with high population densities, such as Seoul, face crucial air pollution problems (Wang et al., 2004). The World Health Organization (WHO) estimates that exposure to air pollution results in seven million premature deaths annually and the loss of millions of healthy life years (World Health Organization, 2021).

Urban air pollution is driven by several factors, including the use of fossil fuels (Johnston et al., 2011), the scale of the city and its associated production activities (Capello and Camagni, 2000), and the conversion of mountainous areas for urban development (Lee et al., 2016). These activities are primarily driven by socioeconomic activities aimed at addressing basic production and consumption needs, as well as the essential requirements for food, clothing, and shelter of urban residents (Irga et al., 2015; Yang et al., 2017). These activities vary depending on the urban spatial structure and land-use patterns. Therefore, to improve and manage urban air quality, the land use, land cover, and spatial structure of the city must be considered (Chen et al., 2022).

Several studies have attempted to identify urban elements that directly or indirectly cause air pollution in cities. Among the various

Abbreviations: (AQMS), air quality monitoring stations; (WHO), World Health Organization; (ESA), European Space Agency; (NDBI), Normalized Difference Built-up Index; (NDVI), Normalized Difference Vegetation Index; (DEM), digital elevation model; (NSDI), National Spatial Data Infrastructure Portal; (EGIS), Environmental Geographic Information Service; (MLR), Multiple Linear Regression; (SHAP), Shapley Additive Explanations; (OLS), Ordinary Least Squares; (XGBoost), Xtreme Gradient Boosting; (XAI), explainable artificial intelligence; (RMSE), Root Mean Square Error; (MAE), Mean Absolute Error.

* Corresponding author.
E-mail addresses: rhyme2997@korea.ac.kr (L. Gunwon), dbgks25@uos.ac.kr (H. Yuhan), ghkim@kei.re.kr (G. Kim).

https://doi.org/10.1016/j.apr.2025.102498
Received 20 October 2024; Received in revised form 5 March 2025; Accepted 5 March 2025
Available online 6 March 2025

urban elements, urban transportation systems and spatial structures (Baldauf et al., 2013), population distribution and density (Van Der Waals, 2000), and terrain shape (Hanna et al., 1982; Roberts et al., 1994) are associated with urban air pollution. Other studies have focused on the relationship between land use, which is one of the most representative urban components, and air pollution. Patterns and changes in land use affect air pollutant dispersion and air quality (Huang et al., 2013; Wei and Ye, 2014; Zahari et al., 2016; Huang and Du, 2018). Rapid changes in land use can lead to sudden increases in air pollution (Du et al., 2010; Tao et al., 2015; Hien et al., 2020). Furthermore, the composition and concentration of air pollutants vary depending on land use and the spatial ranges over which air pollutants disperse (Nagar et al., 2017; Hsu et al., 2018; Harrison, 2020; Yu and Park, 2021). Additionally, as wind (affected by urban patterns and building heights) influences these air pollutants, their concentrations vary depending on the layout of the city and the height of buildings (Blocken et al., 2016; Tuckett-Jones and Reade, 2017).

To address these issues, previous studies have employed prediction models ranging from statistical probability techniques to machine learning. Commonly used statistical probability techniques include correlation analysis and multiple linear regression models, whereas machine learning algorithms include ANN (Wang et al., 2019; Park et al., 2020; Chen et al., 2021), KNN (Bozdağ et al., 2020; Tella and Balogun, 2021), SVM (Yang et al., 2018; Su et al., 2020; Mogollón-Sotelo et al., 2021; Zhang et al., 2021), Random Forest (RF; Yuchi et al., 2019; Shao et al., 2020; Ma et al., 2021a,b), Ensemble (Lim et al., 2019; Van Roode et al., 2019; Adams et al., 2020; Huang et al., 2022), and Xtreme Gradient Boosting (XGBoost; Hu et al., 2017; AlThuwaynee et al., 2021; Zhao et al., 2021). Machine learning algorithms are powerful tools for modeling complex relationships and interactions within data. They have been proven effective in predicting various scenarios more accurately and uncovering patterns that traditional research methods may overlook (Ma et al., 2024).

However, a vast amount of data is required for performing machine learning or deep learning analysis and the installation of air quality monitoring stations (AQMS) for the collection of air quality information is inherently limited. Therefore, studies should examine whether the unconditional application of machine learning or deep learning techniques is appropriate in cases where the number of available samples is limited (Smith et al., 2013; Luan et al., 2020; Rajput et al., 2023) or whether regression analysis is more a suitable alternative (Bonilla-Bedoya et al., 2021).

This study aimed to differentiate itself from previous studies by focusing on the following aspects: South Korea has 443 AQMS nationwide, which were used for the analysis in this study. Based on the limited AQMS data, we examined the suitability of the model and reviewed the spatial extent required to adequately explain air quality. By examining spatial ranges to the AQMSs from near (1000 m) to distant (20,000 m), we aimed to identify the land use and meteorological characteristics that affect air quality within specific ranges and examine their explanatory power differences. Additionally, we verified how land use and vegetation indices within these ranges impacted specific air quality measures. Finally, we aimed to determine whether the established model could predict changes in air pollutant concentrations resulting from land-cover changes. This study aimed to provide crucial foundational information for urban planning and policymaking aimed at improving urban air quality by analyzing the impact level of independent variables related to land cover and land use on air pollutant concentrations.

2. Materials and methods

2.1. Study area

This study focused on South Korea (Fig. 1), covering a total area of approximately 118,118.94 km^2 (Kim and Kim, 2022). The temporal scope of this study was based on 2019 data. This study was limited to the year 2019 to ensure that the analysis reflects typical air quality conditions, excluding the unusual disruptions caused by the COVID-19 pandemic, which significantly reduced air pollution levels at the beginning of 2020. Numerous studies have reported a substantial reduction in air pollutant concentrations during the early stages of the pandemic due to decreased anthropogenic activity (Chossière et al.,

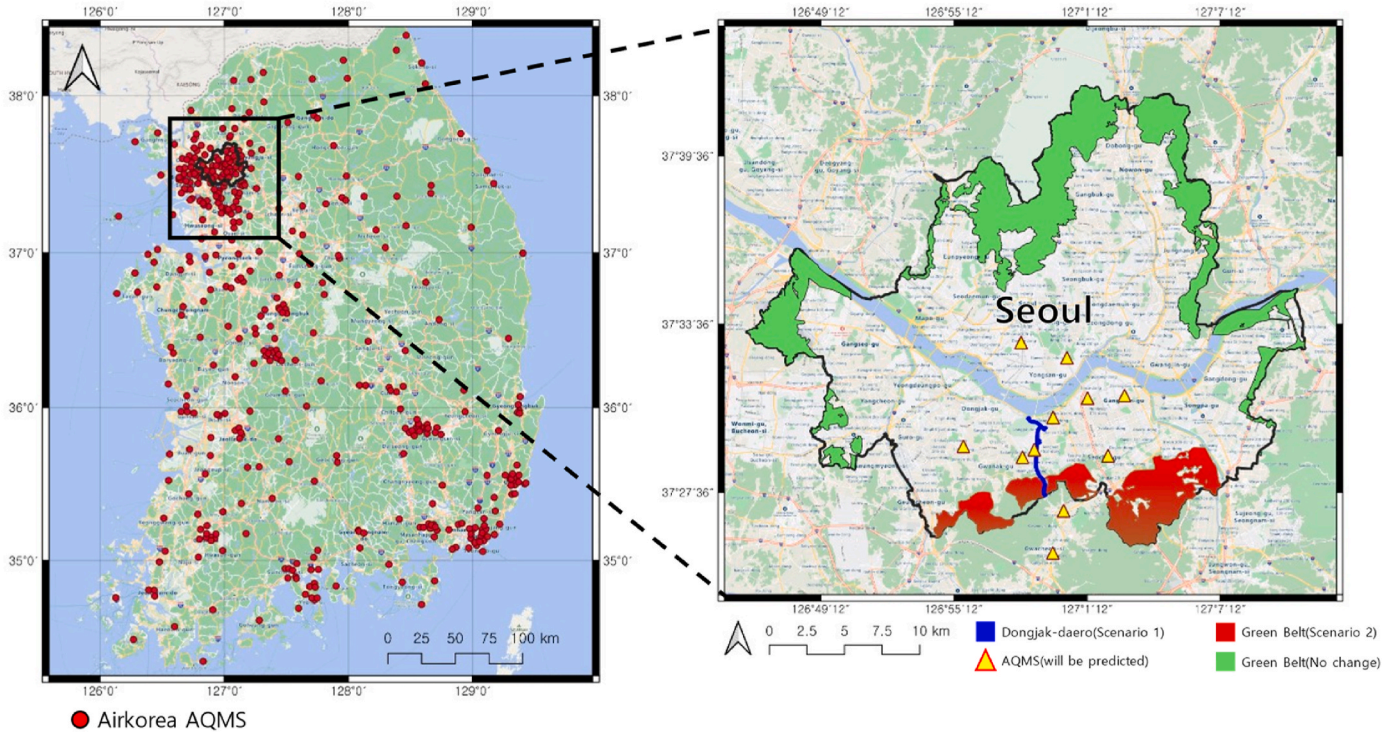


Fig. 1. Study area Korea and Areas Affected by Land-Use Change Scenarios: Dongjak-daero and Surrounding Greenbelt.

2021; Dutheil et al., 2020). Therefore, using data from 2020 onward could introduce biases in the analysis, as the changes in air quality during this period were driven by temporary behavioral and policy shifts rather than long-term land-use changes.

Dongjak-daero serves as a representative case of urban transformation in South Korea, where rapid urbanization has led to increased traffic congestion, reduced green spaces, and significant air quality challenges. To address these issues, an underground road and drainage tunnel have been planned to alleviate congestion and flooding while restoring surface-level green spaces, effectively converting the area into a park (Chosun Biz, 2023). Additionally, the potential impact of the removal of greenbelt development restrictions in Seoul on neighboring areas must be considered. This study analyzed the implications of these urban development initiatives, particularly their effects on air quality and the balance between urban expansion and environmental sustainability, offering insights applicable to cities facing similar challenges.

Fig. 1 illustrates the study area, highlighting the regions affected by major land-use change scenarios, including Dongjak-daero (Scenario 1) and the potential greenbelt removal (Scenario 2). The left panel provides a national-scale view of South Korea, where red dots indicate air quality monitoring stations distributed across the country. The right panel presents a zoomed-in map of Seoul, detailing the specific areas of interest for this study. The blue line represents the planned Dongjak-daero underground road and drainage tunnel project (Scenario 1). The red-shaded areas represent regions where greenbelt restrictions are assumed to be lifted, allowing for potential urban expansion (Scenario 2). The green-shaded areas indicate existing greenbelt zones that remain protected. And the yellow triangles mark the locations of air quality monitoring stations that will be used in Scenario 2.

2.2. Data

Table 1 summarizes the air quality monitoring data used as the dependent variables for analyzing air quality concentrations. The data on vegetation indices [Normalized Difference Built-up Index (NDBI) and Normalized Difference Vegetation Index (NDVI)], topography (Elevation, Slope), and Level-3 land cover were used as independent variables.

2.2.1. Dependent variables

Air pollutant concentration data (CO, NO_2, O_3, SO_2, PM_{10}, and $PM_{2.5}$) from January to December 2019, obtained from AirKorea, were used as dependent variables. After preprocessing, the annual average values were calculated. AirKorea provided AQMS data for South Korea on an hourly basis, which was collected as point data for analysis. A total of 443 AQMSs (Air Quality Monitoring Stations) distributed nationwide were initially collected for analysis. After preprocessing, 351 to 365 AQMSs were used for each type of air quality measure. Some monitoring stations were excluded to ensure data reliability and consistency, as stations with significant NoData values were removed during the preprocessing stage.

Table 1
Air quality monitoring data used in air quality concentration analysis.

Data			Spatial Resolution	Source (Year)
Dependent Variable	Air Quality Concentration	CO (ppm) NO_2 (ppm) O_3 (ppm) SO_2 (ppm) PM_{10} (μg/m³) $PM_{2.5}$ (μg/m³)	Point	AirKorea (2019)
Independent Variable	Vegetation Index	NDBI NDVI	10 × 10 m	Sentinel-2 (2019)
	Topography	Elevation Slope		Digital Terrain Model (2019)
	Level-3 land cover area	Urbanized area Agricultural area Forest area Grassland area Wetland area Bare land area Water area		Level-3 land cover map (2019)

2.2.2. Independent variables

The independent variables included vegetation indices (Fig. 2(a) and 2(b)), topography (Fig. 2(c) and (d)), and level-3 land cover areas (Fig. 2 (e)). The vegetation indices included NDBI and NDVI derived from Sentinel-2 satellite images.

In this study, Sentinel-2A/MSI L1C imagery, captured by the European Space Agency (ESA) on May 23, 2019, was used as raw data to calculate the vegetation indices. The cloud-free imagery was preprocessed using the Semi-Automatic Classification tool. The indices were calculated using Sentinel-2 Band 4 (RED), Band 8 (NIR), and Band 11 (SWIR) as follows:

$$NDBI = \left(\frac{SWIR - NIR}{SWIR + NIR} \right) \text{ and} \quad (1)$$

$$NDVI = \left(\frac{NIR - Red}{NIR + Red} \right). \quad (2)$$

Where SWIR: Short-Wave Infrared, NIR: Near-Infrared, Red: Sentinel Red Band.

Topographical characteristics, including elevation and slope, were also used in this study. Elevation data was derived from digital elevation model (DEM) datasets obtained from the National Spatial Data Infrastructure Portal [NSDI, available online: NSDI (nsdi.go.kr) (accessed on December 6, 2023)], and the slope was calculated using DEM data.

Additionally, a Level-3 land cover map (10 m resolution), provided by the Environmental Geographic Information Service [available online: EGIS (me.go.kr) (accessed on December 6, 2023)] was used. The Level-3 land cover map of South Korea classifies land features, such as residential, commercial, industrial areas, and green spaces, into 41 categories following standardized criteria. This map was created at a 1:5000 scale and has been updated annually by the Ministry of Environment since 2019 (Mun and Kil, 2024). The AQMSs used in this study are mainly located in urban areas, leading to an uneven distribution of land cover types within the buffer zones. Fig. 2(e) shows the Level-3 land cover classification used in this study.

2.3. Methods

2.3.1. Research procedure

This study analyzed the relationship between air pollutant concentrations (CO, NO_2, O_3, SO_2, PM_{10}, and $PM_{2.5}$) and variables, such as vegetation indices, topography, and land use. Zonal statistics were performed within 20 buffer zones (1000–20,000-m at 1000-m intervals) around the AQMSs. Independent variables with a VIF greater than 10 were excluded due to multicollinearity (Fig. 3).

Numerous studies have examined the relationship between air pollutant concentrations and surrounding land use/cover by analyzing spatial ranges extending up to 1–20 km to capture broader spatial impacts (Yang and Jiang, 2021; Ma et al., 2024). In this study, we used 1 km increments for our analyses, creating a balance between fine-scale sensitivity and computational feasibility. Although it is theoretically possible to employ finer increments (e.g., 100 m), such high-resolution analyses would require substantially more computing resources and longer processing times. Consequently, a 1 km scale was deemed appropriate for balancing fine-scale sensitivity with computational

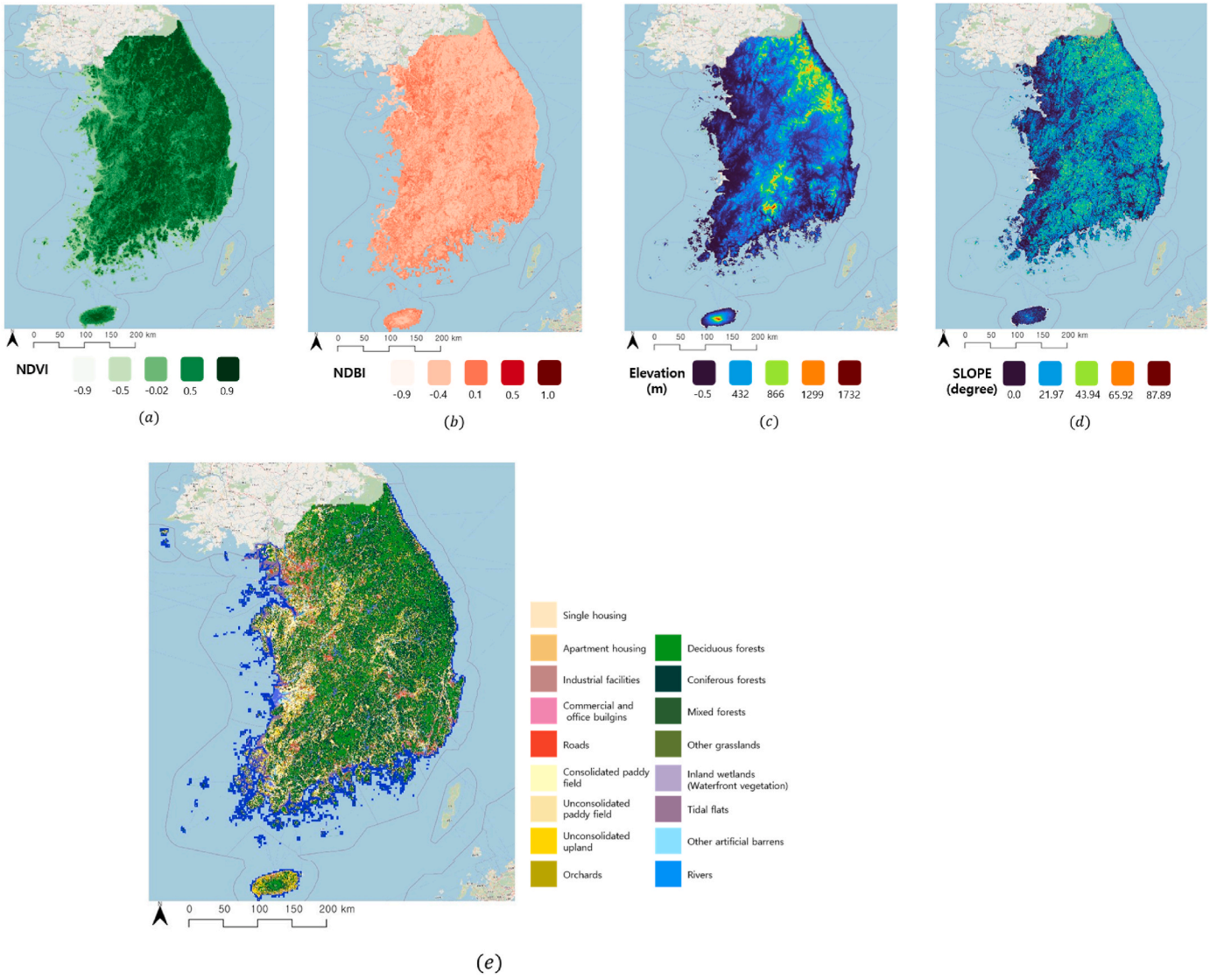


Fig. 2. Variables used in the study: (a) NDBI, (b) NDVI, (c) Elevation, (d) Slope, and (e) Level-3 land cover.

efficiency, a choice that aligns with several existing studies in the field (Liu et al., 2021; Phillips et al., 2021).

Multiple Linear Regression (MLR) and XGBoost models were developed and the best model for each pollutant was selected. The coefficients from the MLR model were examined to identify significant correlations between land cover characteristics and pollutant concentrations. Additionally, Shapley Additive Explanations (SHAP) values were used to interpret the contribution and importance of each feature in the models.

The optimal buffer radius and key influencing factors were identified for each pollutant. Two scenarios were then simulated: converting Dongjak-daero into a green space and urbanizing the surrounding greenbelt, with the predicted air quality changes listed in Tables 2 and 3.

2.3.2. Multiple linear regression

MLR is a statistical method used to analyze the relationship between a dependent variable and multiple independent variables. MLR fits the data into a linear equation to determine the contribution of each independent variable to the dependent variable and, ultimately, identifies the best-fit line that minimizes the difference between the observed and predicted values (ordinary least squares; OLS) (Gulati et al., 2023), defined as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon, \tag{3}$$

where β_0: intercept, β_i: coefficients for each independent variable x_i, x_i: independent variable ϵ: error term.

Separate models were built for SO_2, CO, O_3, NO_2, PM_{10}, and $PM_{2.5}$ using vegetation indices, topography, and land-cover features as independent variables.

2.3.3. XGBoost

XGBoost is a scalable machine learning system for tree boosting that improves on Gradient Boosting by addressing challenges, such as overfitting and learning speed (Chen and Guestrin, 2016). XGBoost was used to handle tabular data due to its efficiency and ability to prevent overfitting through several methods. First, it uses a regularized loss function that reduces model complexity and prevents overfitting, as follows (Dong et al., 2022):

$$\mathscr{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{4}$$

where $\mathscr{L}(\phi)$: total objective function, $l(\hat{y}_i, y_i)$: loss function, $\Omega(f)$: regularization term to control model complexity, γ and λ: the regularization parameters(L1,L2), T: number of leaves, w: leaf weights.

Second, shrinkage reduces the influence of earlier trees, functioning similarly to a learning rate adjustment. Finally, column subsampling

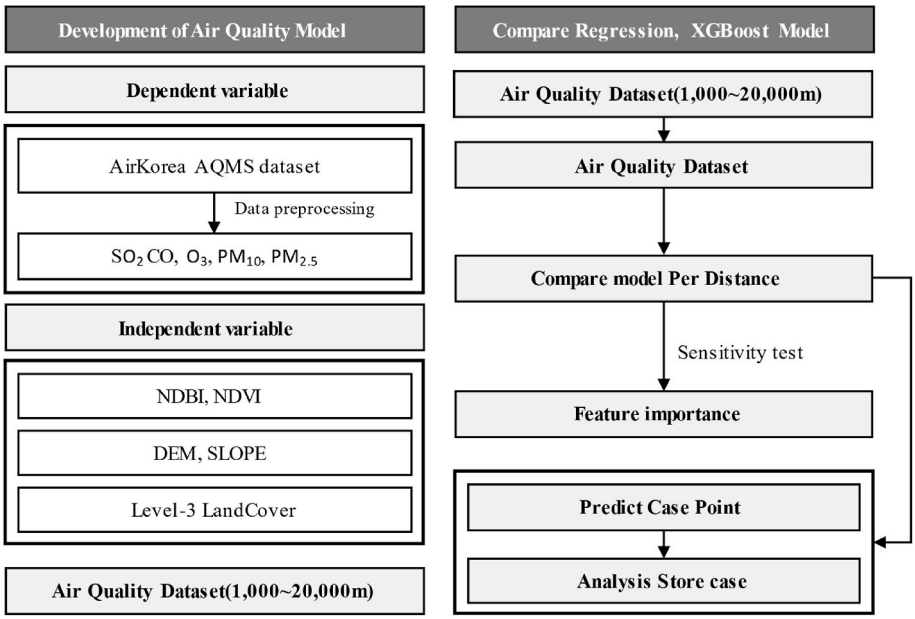


Fig. 3. Flowchart explaining the methods used in this study.

selects a subset of features for training, further reducing overfitting risk (Dong et al., 2022).

Another key reason for selecting XGBoost was its ability to handle multicollinearity effectively, which is a common issue in air quality datasets where several input variables can be highly related. The tree-based structure of the model inherently reduces the impact of multicollinearity by prioritizing the most informative features during the splitting process. Additionally, regularization methods (L1 and L2) further help mitigate the effects of multicollinearity by shrinking the coefficients of less relevant features to zero, thereby improving model stability and performance.

To address potential overfitting in the XGBoost model, hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation (Fig. 4). The parameter grid included n_estimators, learning_rate, reg_alpha, reg_lambda, max_depth, and subsample. Additionally, the dataset was split into 80% training and 20% testing data to ensure unbiased model validation. StandardScaler was applied for scaling and variables with high multicollinearity (VIF >10) were removed before training. The hyperparameter tuning results are shown in Fig. 4.

2.3.4. Scenario assumptions

In this study, two land-use change scenarios were designed to evaluate the impact of land-use alterations on air quality. The assumptions behind each scenario were established based on real-world urban planning discussions and policies to ensure practical relevance (Fig. 5).

Scenario 1: Road Greening (Dongjak-daero Conversion to Green Space)

The first scenario assumed that the surface of Dongjak-daero would be converted into green space after constructing the underground road and drainage tunnel. This assumption was based on recent urban planning initiatives aimed at reducing urban heat islands, improving air quality, and restoring green spaces in metropolitan areas.

The types of vegetation selected for conversion (deciduous forests, coniferous forests, and mixed forests) were chosen to reflect typical urban greening projects due to their air pollution mitigation potential (Nowak et al., 2006). The selection also accounted for vegetation maturity, as mature vegetation can more effectively capture air pollutants and provide cooling effects. It was assumed that traffic previously using Dongjak-daero would not significantly increase on neighboring roads, allowing the analysis to focus solely on the direct environmental benefits of converting the road surface into green space.

Scenario 2: Greenbelt Development (Urbanization of Adjacent Green Areas)

The second scenario assumed that greenbelt areas surrounding Dongjak-daero would be converted into urban developments, such as residential, commercial, and industrial facilities. This assumption was based on ongoing policy discussions in South Korea regarding the relaxation of greenbelt regulations to address housing shortages in urban areas. The types of developments assumed in this scenario were chosen to reflect typical patterns observed in urban expansion projects. It was further assumed that such developments would produce increased emissions from vehicle traffic and industrial activities, contributing to higher air pollutant concentrations. To ensure consistency, it was assumed that existing industrial emission controls would remain constant during the analysis period.

3. Results

3.1. MLR vs. XGBoost

We calculated the mean values of the landscape and topography characteristics, along with the total land cover area, within buffer zones ranging from 1000- to 20,000-m at 1000-m intervals for the following air pollutants: SO_2, CO, O_3, NO_2, PM_{10}, and $PM_{2.5}$. Independent variables with a VIF >10 were excluded from model training. During this process, NDVI and certain land cover variables were removed due to high multicollinearity, as they did not significantly improve model performance.

MLR consistently showed higher explanatory power for predicting air pollutant concentrations compared with XGBoost. Conversely, XGBoost demonstrated overfitting, as seen in certain cases. For instance, in the 1000 m model for CO, the R^2 value for the training model was 0.1652, whereas that of the test model decreased to 0.0107 despite L1 and L2 regularization.

Fig. 6 illustrates the R^2 values of MLR (blue line) and XGBoost (green line) across buffer zones from 1000 to 20,000 m for each dependent variable. The red box highlights the buffer zone where MLR achieved its highest R^2 value for prediction, indicating optimal model performance at each distance.

Tables S.1–6 present the explanatory power (R^2) of multiple linear regression and XGBoost for each dependent variable across buffer distances. The optimal buffer distance for SO_2 was determined to be

Table 2
Predicted air quality changes due to land use changes (Greening) in Dongjak-daero.

		Hangang-daero	Dosan-daero	Seocho-gu	Yongsan-gu	Gangnam-gu	Gwanak-gu	Dongjak-daero	Gangnam-daero	Gwacheon-dong	Byeoryang-dong	Dongjak-gu
SO_2	Original	0.0043	0.0038	0.0038	0.0033	0.0047	0.0043	0.0056	0.0039	0.0033	0.0036	0.0033
	Pred No Change	0.0042	0.0040	0.0040	0.0041	0.0039	0.0040	0.0039	0.0037	0.0036	0.0034	0.0039
	Deciduous Forest	0.0042	0.0040	0.0040	0.0041	0.0039	0.0040	0.0039	0.0037	0.0036	0.0034	0.0039
	Coniferous Forest	0.0042	0.0040	0.0040	0.0041	0.0039	0.0040	0.0039	0.0037	0.0036	0.0034	0.0039
	Mixed Forest	–	–	–	–	–	–	–	–	–	–	–
CO	Original	0.5793	0.7905	0.3749	0.4941	0.4678	0.4509	0.5995	0.6461	0.5942	0.6176	0.4718
	Pred No Change	0.5589	0.5282	0.5357	0.5489	0.5411	0.5362	0.5282	0.5285	0.5331	0.5523	0.5280
	Deciduous Forest	0.5592	0.5420	0.5360	0.5491	0.5414	0.5364	0.5285	0.5288	0.5334	0.5525	0.5282
	Coniferous Forest	0.5589	0.5417	0.5357	0.5489	0.5411	0.5362	0.5282	0.5286	0.5331	0.5523	0.5280
	Mixed Forest	–	–	–	–	–	–	–	–	–	–	–
O_3	Original	0.0170	0.0195	0.0270	0.0229	0.0220	0.0258	0.0170	0.0161	0.0244	0.0232	0.0241
	Pred No Change	0.0234	0.0216	0.0190	0.0225	0.0197	0.0229	0.0216	0.0213	0.0252	0.0273	0.0222
	Deciduous Forest	0.0234	0.0203	0.0190	0.0225	0.0197	0.0228	0.0216	0.0213	0.0251	0.0273	0.0221
	Coniferous Forest	0.0234	0.0203	0.0190	0.0225	0.0197	0.0228	0.0216	0.0213	0.0251	0.0273	0.0222
	Mixed Forest	0.0234	0.0203	0.0190	0.0225	0.0197	0.0228	0.0216	0.0213	0.0251	0.0273	0.0221
NO_2	Original	0.0397	0.0303	0.0300	0.0318	0.0275	0.0305	0.0470	0.0469	0.0256	0.0294	0.0302
	Pred No Change	0.0342	0.0342	0.0347	0.0341	0.0327	0.0335	0.0342	0.0329	0.0328	0.0320	0.0342
	Deciduous Forest	–	–	–	–	–	–	–	–	–	–	–
	Coniferous Forest	0.0342	0.0339	0.0347	0.0341	0.0327	0.0335	0.0342	0.0329	0.0328	0.0320	0.0342
	Mixed Forest	–	–	–	–	–	–	–	–	–	–	–
PM_{10}	Original	48.5521	45.9760	43.3098	33.8829	39.8950	48.6406	46.6591	46.2964	42.6036	46.7572	43.5449
	Pred No Change	45.0509	44.4488	44.6999	44.7248	44.2351	44.6850	44.4488	43.8219	43.9249	43.5393	44.4312
	Deciduous Forest	–	–	–	–	–	–	–	–	–	–	–
	Coniferous Forest	45.0436	44.4967	44.6924	44.7174	44.2273	44.6791	44.4436	43.8170	43.9177	43.5320	44.4237
	Mixed Forest	–	–	–	–	–	–	–	–	–	–	–
$PM_{2.5}$	Original	27.5614	25.8506	25.5377	23.7683	24.7205	27.5339	24.8568	25.5783	22.2822	22.1938	26.4567
	Pred No Change	26.1165	24.2000	24.7671	25.7238	25.2860	24.8759	24.2000	24.2616	23.9556	23.8665	24.0840
	Deciduous Forest	26.1240	25.1916	24.7759	25.7313	25.2939	24.8850	24.2079	24.2733	23.9631	23.8741	24.0915
	Coniferous Forest	26.1161	25.1837	24.7680	25.7233	25.2860	24.8771	24.2000	24.2654	23.9552	23.8661	24.0836
	Mixed Forest	–	–	–	–	–	–	–	–	–	–	–

12,000 m, yielding the highest explanatory power ($R^2 = 0.3402$). The differences in R^2 values among the 11,000-m ($R^2 = 0.34$), 12,000-m ($R^2 = 0.3402$), and 13,000-m ($R^2 = 0.34$) distances are minimal. For other pollutants, the optimal buffer distances varied: CO peaked at 11,000 m ($R^2 = 0.2733$), O_3 showed the highest R^2 at 6000 m ($R^2 = 0.5024$), NO_2 reached its maximum explanatory power at 18,000 m ($R^2 = 0.7037$), PM_{10} had its highest R^2 at 18,000 m ($R^2 = 0.4898$), and $PM_{2.5}$ peaked at 11,000 m ($R^2 = 0.4822$). These optimal buffer distances were chosen to develop models that effectively capture air pollutant variations across different spatial scales.

3.2. Correlations between land cover characteristics and air pollutants

In this study, multiple linear regression was employed to examine the relationships between air pollutant concentrations and various land cover characteristics. The results revealed that certain land cover types significantly affect pollutant levels, consistent with the expected spatial distribution of emission sources and dispersion processes. The results for each pollutant, as summarized in Tables S.7–12, indicate distinct relationships between specific land cover types and air pollutant levels.

For SO_2, industrial facilities and tidal flats exhibited significant positive correlations with SO_2 concentrations. Conversely, orchards and deciduous forests were significantly negatively correlated with SO_2. However, variables, such as elevation and single housing did not reach statistical significance, indicating that the current data cannot reliably support their effects on SO_2 levels.

Similarly, the CO model showed positive relationships with single housing, deciduous forests, and other artificial barren areas, whereas the river variable was significantly negatively correlated. For O_3, significant negative predictors included apartment housing, deciduous forests, and

Table 3
Predicted air quality changes due to land use changes (urbanization) in the greenbelt.

		Hangang-daero	Dosan-daero	Seocho-gu	Yongsan-gu	Gangnam-gu	Gwanak-gu	Dongjak-daero	Gangnam-daero	Gwacheon-dong	Byeoryang-dong	Dongjak-gu
SO_2	Original	0.0043	0.0038	0.0038	0.0033	0.0047	0.0043	0.0056	0.0039	0.0033	0.0036	0.0033
	Pred No Change	0.0042	0.0040	0.0040	0.0041	0.0039	0.0040	0.0039	0.0037	0.0036	0.0034	0.0039
	Single housing	0.0043	0.0044	0.0045	0.0045	0.0043	0.0042	0.0044	0.0042	0.0041	0.0039	0.0044
	Apartment Housing	–	–	–	–	–	–	–	–	–	–	–
	Industrial facilities	0.0050	0.0061	0.0062	0.0058	0.0059	0.0051	0.0062	0.0060	0.0059	0.0057	0.0061
CO	Original	0.5792	0.7905	0.3749	0.4941	0.4678	0.4509	0.5995	0.6461	0.5942	0.6176	0.4718
	Pred No Change	0.5589	0.5282	0.5357	0.5489	0.5411	0.5362	0.5282	0.5285	0.5331	0.5523	0.5280
	Single housing	0.5810	0.6247	0.6233	0.5940	0.6118	0.5746	0.6148	0.6133	0.6215	0.6376	0.6099
	Apartment Housing	–	–	–	–	–	–	–	–	–	–	–
	Industrial facilities	0.5581	0.5371	0.5309	0.5465	0.5360	0.5351	0.5235	0.5239	0.5282	0.5476	0.5234
O_3	Original	0.0170	0.0195	0.0270	0.0229	0.0220	0.0258	0.0170	0.0161	0.0244	0.0232	0.0241
	Pred No Change	0.0234	0.0216	0.0190	0.0225	0.0197	0.0229	0.0216	0.0213	0.0252	0.0273	0.0222
	Single housing	0.0234	0.0207	0.0198	0.0225	0.0204	0.0238	0.0229	0.0243	0.0275	0.0290	0.0235
	Apartment Housing	0.0234	0.0151	0.0083	0.0225	0.0113	0.0088	0.0047	−0.0155	−0.0052	0.0042	0.0039
	Industrial facilities	0.0234	0.0201	0.0185	0.0225	0.0194	0.0221	0.0208	0.0198	0.0238	0.0262	0.0213
NO_2	Original	0.0397	0.0303	0.0300	0.0318	0.0275	0.0305	0.0470	0.0469	0.0256	0.0294	0.0302
	Pred No Change	0.0342	0.0342	0.0347	0.0341	0.0327	0.0335	0.0342	0.0329	0.0328	0.0320	0.0342
	Single housing	–	–	–	–	–	–	–	–	–	–	–
	Apartment Housing	0.0493	0.0491	0.0498	0.0493	0.0479	0.0486	0.0494	0.0480	0.0479	0.0471	0.0493
	Industrial facilities	0.0406	0.0403	0.0410	0.0405	0.0391	0.0398	0.0406	0.0393	0.0392	0.0384	0.0405
PM_{10}	Original	48.5521	45.9760	43.3098	33.8829	39.8950	48.6406	46.6591	46.2964	42.6036	46.7572	43.5449
	Pred No Change	45.0509	44.4488	44.6999	44.7248	44.2351	44.6850	44.4488	43.8219	43.9249	43.5393	44.4312
	Single housing	–	–	–	–	–	–	–	–	–	–	–
	Apartment Housing	50.2786	49.7317	49.9274	49.9524	49.4623	49.9140	49.6786	49.0520	49.1526	48.7670	49.6587
	Industrial facilities	48.2692	47.7223	47.9180	47.9430	47.4529	47.9046	47.6692	47.0426	47.1432	46.7576	47.6493
$PM_{2.5}$	Original	27.5614	25.8505	25.5377	23.7683	24.7205	27.5338	24.8568	25.5783	22.2822	22.1937	26.4567
	Pred No Change	26.1165	24.2000	24.7671	25.7237	25.2859	24.8759	24.2000	24.2616	23.9556	23.8665	24.0840
	Single housing	27.4480	30.1654	30.0212	28.4307	29.5354	27.1849	29.3990	29.3539	29.2625	28.9907	29.0015
	Apartment Housing	–	–	–	–	–	–	–	–	–	–	–
	Industrial facilities	26.0577	24.8614	24.4303	25.5554	24.9484	24.7930	23.8679	23.9427	23.6112	23.5382	23.7649

rivers, with other variables failing to achieve significance. In the case of NO_2, strong positive associations were found for apartment housing and industrial facilities, whereas tidal flats also showed a significant negative effect. Moreover, the PM_{10} and $PM_{2.5}$ models revealed that housing and agricultural areas are generally associated with higher particulate matter concentrations, whereas forested areas contribute to their reduction, as indicated by significant negative coefficients.

3.3. Air quality restoration cases

In this study, two distinct scenarios were analyzed to observe how land-use changes impact air pollutant concentrations. We first identified the effective range of influence for land cover and meteorological characteristics on air pollutants. Based on this, we predicted changes in air pollutant concentrations when these factors were altered around selected AQMSs. The stations were chosen based on their proximity to Dongjak-daero and their location within the minimum buffer distance to reflect the impact of land cover changes (Fig. 1).

In the first scenario, the road surface of Dongjak-daero was transformed into green space by converting it into land cover types, such as deciduous, coniferous, and mixed forests, ensuring no multicollinearity issues. The goal was to assess the potential reduction in air pollutant concentrations by replacing an urban area with green space. Predicted pollutant concentrations, such as SO_2, CO, NO_2, O_3, PM_{10}, and $PM_{2.5}$, were compared before and after the change (Table 2).

The analysis showed relatively modest improvements in air quality, suggesting that greening smaller urban roads, such as Dongjak-daero alone may have limited effectiveness in reducing pollutants. The largest observed reduction was in CO, with a notable decrease of 0.0123 ppm after conversion to coniferous forest, whereas changes in other

Hyperparameter	value
N_estimators	50
Learning_rate	0.01
Reg_alpha	0
Reg_lambda	1
Max_depth	3
subsample	0.8

Fig. 4. Air quality monitoring data used in air quality concentration analysis.

pollutants were minimal.

The second scenario assessed the impact of converting adjacent greenbelt areas into urban developments, such as single housing, apartment buildings, and industrial facilities. This scenario had more significant negative effects on air quality compared to road greening (Table 3). For instance, when the greenbelt was converted into industrial areas, there was a substantial increase in SO_2 (0.001936 ppm) and PM_{10} (5.2331 μg/m^3), demonstrating that urbanization has a considerable adverse impact on pollutant concentrations. Furthermore, converting greenbelts into single housing increased $PM_{2.5}$ concentrations by 4.314206 μg/m^3, underscoring the crucial role that green spaces play in maintaining better air quality.

4. Discussion

Linear regression analysis provided better explanatory power than XGBoost, a representative machine learning model, in explaining urban environmental factors, including land use within a defined spatial range, based on data from 351 to 365 AQMSs. This result is probably due to the small sample size used in the analysis, which is consistent with previous studies that explored the correlation between urban air quality and urban environments (Johnson et al., 2010; Lai et al., 2021). This suggests that rather than unconditionally applying machine learning or deep learning models for prediction and relationship analysis, researchers should assess experimental conditions, such as the sample size, and determine the appropriate model through empirical testing. This also implies that simple linear regression analysis is a viable candidate model for such analyses.

Furthermore, the analysis results indicate that the explanatory power of air quality concentration varies with the spatial range and urban environment, including land use and land cover, depending on the characteristics of each air quality parameter. For example, the explanatory power for O_3 was significant at 6000 m, whereas that for NO_2 and PM_{10} was significant at 18,000 m. This suggests that urban and environmental planning should consider spatial influences on air quality. Previous studies (Weng and Yang, 2006; Li et al., 2015) have often considered narrow buffer zones of only a few hundred meters. Conversely, this study analyzed the relationship between the air quality and urban environment over a broader buffer range of 6–18 km. This broader spatial analysis highlights the fact that extensive urban environments can significantly affect air quality and should be considered in future urban and environmental planning. Additionally, this study found that certain air quality parameters exhibited a strong explanatory relationship with the surrounding urban environment, including land use, whereas others exhibited a weaker association. Therefore, future land-use changes or land-use planning should consider specific air quality parameters that can be effectively analyzed for environmental impact assessments.

In this study, the analysis was performed based on the direct distance that demonstrated the highest explanatory power. Similar analyses often consume several computing resources to calculate the urban environment within large areas that span a radius of several kilometers. Therefore, when there is a notable difference in explanatory power, it is necessary to select a suitable consensus range for analysis that considers various factors, such as computing resources and analysis time. For example, for the case of SO2 conducted in this study, the optimal buffer distance for SO_2 was determined to be 12,000 m, yielding the highest explanatory power ($R^2 = 0.3402$). However, the differences in R^2 values among the 11,000-m ($R^2 = 0.34$), 12,000-m ($R^2 = 0.3402$), and 13,000-m ($R^2 = 0.34$) distances were minimal. If there is no significant difference in explanatory power, 1100 m can be considered the optimal analysis range, as it balances efficiency in analysis time and the utilization of computing resources.

The analysis of the urban environment within a certain range around

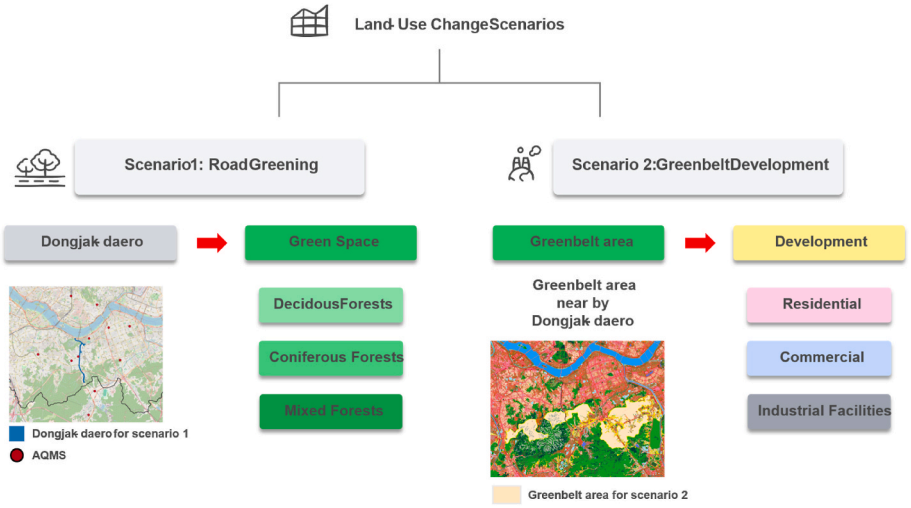


Fig. 5. Land-use change Scenarios(Scenario 1: Road greening, scenario 2: Greenbelt development).

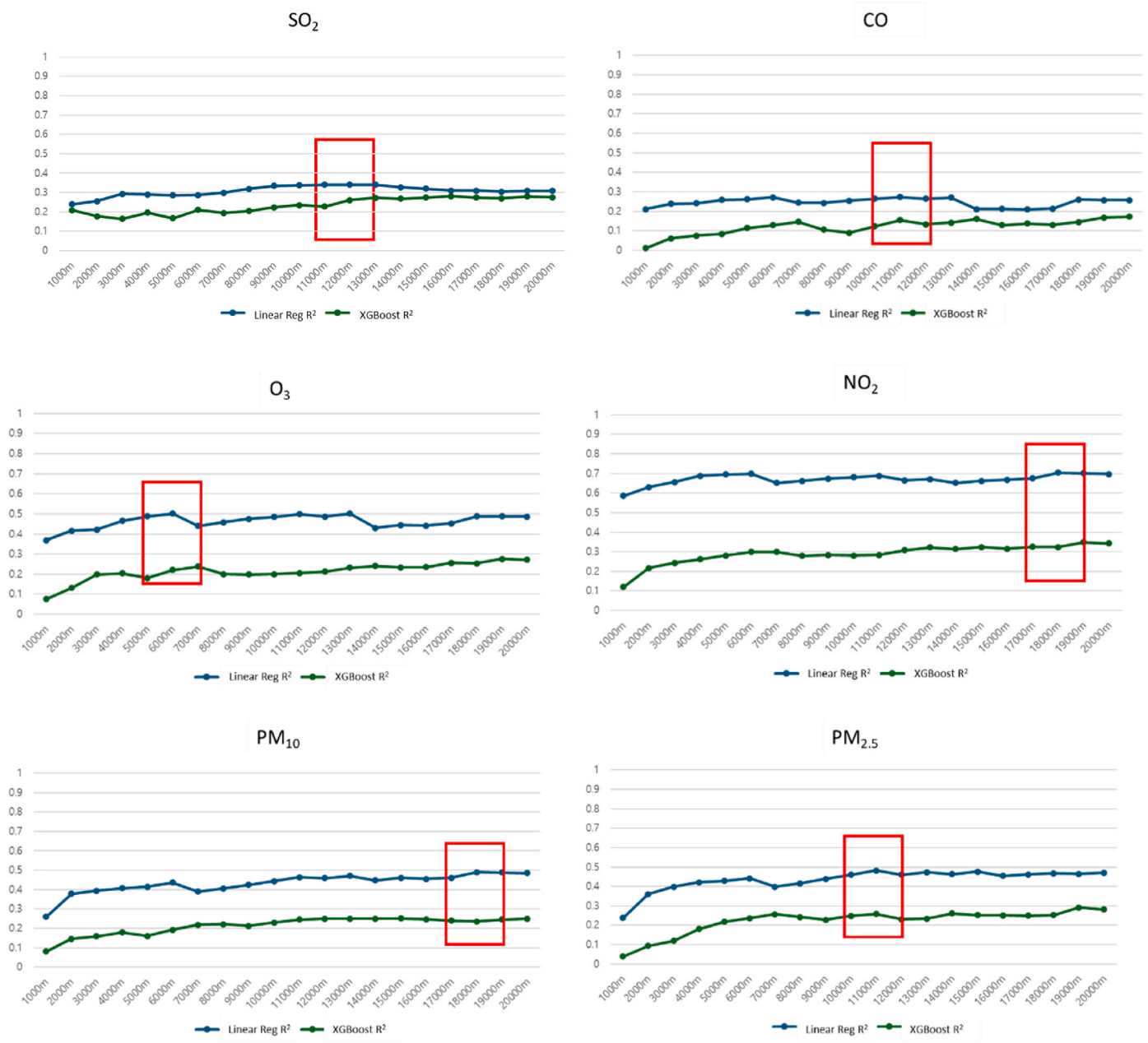


Fig. 6. Explanatory Power (R^2) of Multiple Linear Regression and XGBoost by Buffer Distance for each air pollutant.

each AQMS and the characteristics of the air pollutant concentrations showed that the impact on air quality varied according to the characteristics of the urban environment. This implies that urban planning should consider the urban environment and spatial range, including surrounding land use and land cover.

Consistent with previous research, NO2 showed a strong positive correlation with surrounding apartment housing, suggesting that NOx, although naturally occurring, significantly increased owing to human industrial activities, especially from combustion facilities, such as boilers in commercial and residential buildings (Yue et al., 2018). For PM_{10}, agricultural areas, such as paddy fields (Li and Huang, 2020), and apartment housing showed a strong positive correlation. Conversely, the forest area exhibited a negative relation with PM_{10}. Further, $PM_{2.5}$ concentrations were higher in areas with a high density of single-family housing and, similar to PM_{10}, lower in areas with extensive forest cover. Additionally, paddy fields, unconsolidated upland fields, and consolidated paddy fields were positively correlated with $PM_{2.5}$, similar to other barren lands. This result is consistent with those of previous studies, indicating that barren lands contribute to fine dust generation (Pinho et al., 2008).

However, for $PM_{2.5}$, the same type of forest showed different relations: coniferous forests exhibited a negative correlation, whereas deciduous forests showed a positive correlation. When constructing land-cover maps in South Korea, security facilities, such as military installations, airports, and power plants, are often reclassified as agricultural or barren land. This classification may have influenced the analysis, necessitating additional reviews to validate the impact of these factors. Therefore, future research should focus on examining the specific effects of these facilities on air quality.

Many of our findings align with existing knowledge, such as the positive association between industrial facilities with SO2, NO2, and particulate matter, and the mitigating effect of forested areas on PM10 and PM2.5. However, certain results appear less intuitive. In particular, air pollutants, such as CO and SO_2 exhibited low explanatory power in

the linear regression analysis, lacking sufficient explanatory power in terms of their relationship with land cover. For example, deciduous forests showed a positive correlation with CO, however, a negative correlation with SO_2. Examining the results of models with low explanatory power reveals that several predictor variables had low p-values, indicating weak explanatory relationships between specific land cover types and air pollutant concentrations. Similarly, factors expected to reduce pollution levels, such as high elevation or proximity to rivers, may correspond to high-traffic transport corridors or small residential combustion areas, contributing to increased levels of CO, NO_2, and PM_{10}, as observed in Jain et al. (2021). These cases suggest that land cover variables may act as proxies for unmeasured factors (e.g., heating fuel, traffic intensity, or tourism activities), complicating purely linear interpretations. This underscores the need to consider nonlinearity and regional drivers of air quality, as well as the necessity of incorporating additional emissions and meteorological data to completely explain the localized surges of specific pollutants, such as CO. Therefore, when predicting the concentrations of air pollutants that have weak explanatory power in relation to land cover, additional predictor variables must be incorporated and more sophisticated modeling approaches must be considered to better capture the complex interactions between land cover and air pollution.

Using the model developed in this study, we predicted changes in air pollutant concentrations caused by changes in land cover. However, restoration performed in small areas provided minimal improvement in air quality. Conversely, the significant development in the area was predicted to increase air pollutant concentrations. Notably, when large-scale development projects were undertaken in previously green spaces, air pollutant concentrations were predicted to increase significantly. These results are consistent with those of most previous studies that have examined changes in air quality due to land use and land cover changes. However, this study demonstrated the ability to apply Level-3 land use changes, distinguishing it from previous studies that primarily examined land use and land cover changes over broader areas.

Finally, vegetation indices, such as NDVI, which indirectly assess the quality of green spaces, were excluded from the regression analysis due to multicollinearity issues identified during the pre-analysis. This is because most forests in South Korea exhibit high vegetation vitality (Lee and Park, 2020). Additionally, as the forests of South Korea are primarily located in high-altitude mountainous areas, we included high-altitude regions in the analysis. Consequently, when predicting future land use changes, this could lead to the exclusion of green space quality. Therefore, future research should develop methodologies that incorporate the quality of green spaces into their analysis.

5. Conclusion

This study aimed to provide crucial baseline information for urban planning and policymaking by analyzing the impact of land cover- and land use-related independent variables on air pollutant concentrations. The findings offer valuable insights into urban air quality, including more effective urban planning and environmental management strategies that can be developed. However, several limitations exist in deriving and interpreting the results.

First, the limited number of AQMSs led to a higher explanatory power for linear regression analysis compared to machine learning models. Therefore, increasing the number of AQMSs and data samples will improve the performance of machine-learning models. Future research should focus on collecting additional data from various environments to construct more reliable models. Second, vegetation indices, such as NDVI, were excluded from the linear regression model due to high multicollinearity, which prevented the consideration of green space quality. Therefore, future research should develop methods to incorporate greenspace quality into the analysis.

Similar to other studies, this research also has some methodological constraints. First, because the dataset is confined to South Korea and the number of AQMS (360) is limited, it is difficult to generalize the results. Moreover, while this study mainly compared regression analysis and machine learning approaches, there is a potential for more in-depth findings by applying complex methodologies such as spatiotemporal data analysis, network analysis, or multilevel modeling. These methodologies could offer a more nuanced explanation of the complex interactions between urban structure and air pollution, warranting consideration in future research.

The findings of this study also provide important insights for urban planning and policymaking. First, the fact that air pollutant concentrations vary depending on spatial scale indicates that specific spatial ranges must be considered in urban design. For example, NO_2 and PM_{10} exhibited effects over a wider area of about 18 km, whereas O_3 showed measurable impacts even within a narrower range of around 6 km. This highlights the need to develop tailored policies that consider the diverse land-use patterns within a city.

Moreover, small-scale urban development efforts, such as green space restoration, had a limited impact on improving air quality, whereas the urbanization of green belts had a significantly negative effect. These findings emphasize the need to preserve green belts and enforce strict regulations to minimize environmental harm during large-scale development projects. Such insights can be used to inform various environmental impact assessments and formulate climate change adaptation policies.

As certain air pollutants showed low explanatory power in the model presented in this study, future research should utilize a broader range of predictor variables to better explain these air pollutant concentrations. Additionally, more sophisticated modeling approaches should be considered to better capture the complex interactions between land cover and air pollution. Given the limitations of currently available official datasets, including time-series data, future studies should focus on collecting more diverse and comprehensive datasets to enhance model accuracy and interpretability.

CRediT authorship contribution statement

Lee Gunwon: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Han Yuhan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Formal analysis, Data curation. **Geunhan Kim:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding sources

Declaration of generative AI in scientific writing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Geunhan Kim reports financial support was provided by Korea Agency for Infrastructure Technology Advancement (KAIA). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

TThis paper is based on the findings of the research project (2025-014(R)) which was conducted by the Korea Environment Institute (KEI) and supported by a Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2003-00242291).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apr.2025.102498.

References

Adams, M.D., Massey, F., Chastko, K., Cupini, C., 2020. Spatial modelling of particulate matter air pollution sensor measurements collected by community scientists while cycling, land use regression with spatial cross-validation, and applications of machine learning for data correction. Atmos. Environ. 230, 117479. https://doi.org/10.1016/j.atmosenv.2020.117479.

Akimoto, H., 2003. Global air quality and pollution. Science 302, 1716–1719. https://doi.org/10.1126/science.1092666.

AlThuwaynee, O.F., Kim, S.W., Najemaden, M.A., Aydda, A., Balogun, A.L., Fayyadh, M. M., Park, H.J., 2021. Demystifying uncertainty in PM10 susceptibility mapping using variable drop-off in extreme-gradient boosting (XGB) and random forest (RF) algorithms. Environ. Sci. Pollut. Res. Int. 28, 43544–43566. https://doi.org/10.1007/s11356-021-13255-4.

Baldauf, R.W., Heist, D., Isakov, V., Perry, S., Hagler, G.S.W., Kimbrough, S., Shores, R., Black, K., Brixey, L., 2013. Air quality variability near a highway in a complex urban environment. Atmos. Environ. 64, 169–178. https://doi.org/10.1016/j.atmosenv.2012.09.054.

Balew, A., Korme, T., 2020. Monitoring land surface temperature in Bahir Dar city and its surrounding using Landsat images. Egypt. J. Remote Sens. Space Sci. 23, 371–386. https://doi.org/10.1016/j.ejrs.2020.02.001.

Blocken, B., Vervoort, R., van Hooff, T., 2016. Reduction of outdoor particulate matter concentrations by local removal in semi-enclosed parking garages: a preliminary case study for Eindhoven city center. J. Wind Eng. Ind. Aerodyn. 159, 80–98. https://doi.org/10.1016/j.jweia.2016.10.008.

Bonilla-Bedoya, S., Zalakeviciute, R., Coronel, D.M., Durango-Cordero, J., Molina, J.R., Macedo-Pezzopane, J.E., Herrera, M.Á., 2021. Spatiotemporal variation of forest cover and its relation to air quality in urban Andean socio-ecological systems. Urban For. Urban Green. 59, 127008. https://doi.org/10.1016/j.ufug.2021.127008.

Bozdağ, A., Dokuz, Y., Gökçek, Ö.B., 2020. Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey. Environ. Pollut. 263, 114635. https://doi.org/10.1016/j.envpol.2020.114635.

Capello, R., Camagni, R., 2000. Beyond optimal city size: an evaluation of alternative urban growth patterns. Urban Stud. 37, 1479–1496. https://doi.org/10.1080/00420980020080221.

Chen, B., You, S., Ye, Y., Fu, Y., Ye, Z., Deng, J., Wang, K., Hong, Y., 2021. An interpretable self-adaptive deep neural network for estimating daily spatially continuous PM2.5 concentrations across China. Sci. Total Environ. 768, 144724. https://doi.org/10.1016/j.scitotenv.2020.144724.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system in *Proc.* In: 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 785–794

Chen, Y., Xu, Y., Wang, F., Shi, F., 2022. Mapping the emission of air pollution sources based on land-use classification: a case study of Shengzhou, China. Land Use Policy 117, 106083. https://doi.org/10.1016/j.landusepol.2022.106083.

Chossière, G.P., Xu, H., Dixit, Y., Isaacs, S., Eastham, S.D., Allroggen, F., et al., 2021. Air pollution impacts of COVID-19–related containment measures. Sci. Adv. 7 (21), eabe1178.

Chosun, Biz, 2023. A tunnel combining road and rainwater drainage to be built in South Korea to ease traffic congestion and prevent flooding. Chosun Biz. Retrieved from. https://biz.chosun.com/topics/topics_social/2023/12/26/EMLDMOOQ5REGJOEAA4I2GFUXNQ/.

Dong, J., Chen, Y., Yao, B., Zhang, X., Zeng, N., 2022. A neural network boosting regression model based on XGBoost. Appl. Soft Comput. 125, 109067. https://doi.org/10.1016/j.asoc.2022.109067.

Du, N., Ottens, H., Sliuzas, R., 2010. Spatial impact of urban expansion on surface water bodies—a case study of Wuhan, China. Landsc. Urban Plann. 94, 175–185. https://doi.org/10.1016/j.landurbplan.2009.10.002.

Dutheil, F., Baker, J.S., Navel, V., 2020. COVID-19 and air pollution: the worst is yet to come. Environ. Sci. Pollut. Control Ser. 27 (35), 44647–44649.

Faiz, A., 1993. Automotive emissions in developing countries: relative implications for global warming, acidification, and urban air quality. Transp. Res. A. 27, 167–186.

Gulati, S., Bansal, A., Pal, A., Mittal, N., Sharma, A., Gared, F., 2023. Estimating PM2.5 utilizing multiple linear regression and ANN techniques. Sci. Rep. 13, 22578. https://doi.org/10.1038/s41598-023-49717-7.

Hanna, S.R., Briggs, G.A., Hosker Jr, R.P., 1982. Handbook on atmospheric diffusion. National Oceanic and Atmospheric Administration, Atmospheric Turbulence and Diffusion Laboratory. Oak Ridge, Tennessee (DOE/TIC-11223).

Harrison, R.M., 2020. Airborne particulate matter. Philos. Trans. A Math. Phys. Eng. Sci. 378, 20190319. https://doi.org/10.1098/rsta.2019.0319.

Hien, P.D., Men, N.T., Tan, P.M., Hangartner, M., 2020. Impact of urban expansion on the air pollution landscape: a case study of Hanoi. Vietnam. Sci. Total Environ. 702, 134635. https://doi.org/10.1016/j.scitotenv.2019.134635.

Hsu, C.Y., Wu, C.D., Hsiao, Y.P., Chen, Y.C., Chen, M.J., Lung, S.C.C., 2018. Developing land-use regression models to estimate PM2.5-bound compound concentrations. Remote Sens. 10. https://doi.org/10.3390/rs10121971.

Hu, K., Rahman, A., Bhrugubanda, H., Sivaraman, V., 2017. HazeEst: machine learning based metropolitan air pollution estimation from fixed and mobile sensors. IEEE Sens. J. 17, 3517–3525. https://doi.org/10.1109/JSEN.2017.2690975.

Huang, C., Sun, K., Hu, J., Xue, T., Xu, H., Wang, M., 2022. Estimating 2013–2019 NO_2 exposure with high spatiotemporal resolution in China using an ensemble model. Environ. Pollut. 292, 118285. https://doi.org/10.1016/j.envpol.2021.118285.

Huang, Y.K., Luvsan, M.E., Gombojav, E., Ochir, C., Bulgan, J., Chan, C.C., 2013. Land use patterns and SO2 and NO2 pollution in Ulaanbaatar, Mongolia. Environ. Res. 124, 1–6. https://doi.org/10.1016/j.envres.2013.02.006.

Huang, Z., Du, X., 2018. Urban land expansion and air pollution: evidence from China. J. Urban Plan. Dev 144. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000476.

Irga, P.J., Burchett, M.D., Torpy, F.R., 2015. Does urban forestry have a quantitative effect on ambient air quality in an urban environment? Atmos. Environ. 120, 173–181. https://doi.org/10.1016/j.atmosenv.2015.08.050.

Jain, S., Presto, A.A., Zimmerman, N., 2021. Spatial modeling of daily PM2. 5, NO2, and CO concentrations measured by a low-cost sensor network: comparison of linear, machine learning, and hybrid land use models. Environ. Sci. Technol. 55 (13), 8631–8641.

Johnson, M., Isakov, V., Touma, J.S., Mukerjee, S., Özkaynak, H., 2010. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmos. Environ. 44, 3660–3668. https://doi.org/10.1016/j.atmosenv.2010.06.041.

Johnston, F., Hanigan, I., Henderson, S., Morgan, G., Bowman, D., 2011. Extreme air pollution events from bushfires and dust storms and their association with mortality in Sydney, Australia 1994–2007. Environ. Res. 111, 811–816. https://doi.org/10.1016/j.envres.2011.05.007.

Kim, M., Kim, G., 2022. Modeling and predicting urban expansion in South Korea using explainable artificial intelligence (XAI) model. Appl. Sci. 12, 9169. https://doi.org/10.3390/app12189169.

Lai, S.B.S., Binti Md Shahri, N.H.N.B.M., Mohamad, M.B., Rahman, H.A.B.A., Rambli, A. B., 2021. Comparing the performance of AdaBoost, XGBoost, and logistic regression for imbalanced data. Math. Stat. 9, 379–385. https://doi.org/10.13189/ms.2021.090320.

Lee, H.J., Chatfield, R.B., Strawa, A.W., 2016. Enhancing the applicability of satellite remote sensing for PM2.5 estimation using MODIS Deep Blue AOD and land use regression in California, United States. Environ. Sci. Technol. 50, 6546–6555. https://doi.org/10.1021/acs.est.6b01438.

Lee, P.S.H., Park, J., 2020. An effect of urban forest on urban thermal environment in Seoul, South Korea, based on landsat imagery analysis. Forests 11, 630. https://doi.org/10.3390/f11060630.

Li, J., Huang, X., 2020. Impact of land-cover layout on particulate matter 2.5 in urban areas of China. Int. J. Digit. Earth. 13, 474–486. https://doi.org/10.1080/17538947.2018.1530310.

Li, X., Liu, W., Chen, Z., Zeng, G., Hu, C., León, T., Liang, J., Huang, G., Gao, Z., Li, Z., Yan, W., 2015. The application of semicircular-buffer-based land use regression models incorporating wind direction in predicting quarterly NO2 and PM10 concentrations. Atmos. Environ. 103, 18–24.

Lim, C.C., Kim, H., Vilcassim, M.J.R., Thurston, G.D., Gordon, T., Chen, L.C., Lee, K., Heimbinder, M., Kim, S.Y., 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environ. Int. 131, 105022. https://doi.org/10.1016/j.envint.2019.105022.

Liu, Z., Guan, Q., Lin, J., Yang, L., Luo, H., Wang, N., 2021. A new buffer selection strategy for land use regression model of PM 2.5 in Xi'an, China. Environ. Sci. Pollut. Control Ser. 28, 21245–21255.

Luan, J., Zhang, C., Xu, B., Xue, Y., Ren, Y., 2020. The predictive performances of random forest models with limited sample size and different species traits. Fish. Res. 227, 105534. https://doi.org/10.1016/j.fishres.2020.105534.

Ma, M., Yao, G., Guo, J., Bai, K., 2021a. Distinct spatiotemporal variation patterns of surface ozone in China due to diverse influential factors. J. Environ. Manag. 288, 112368. https://doi.org/10.1016/j.jenvman.2021.112368.

Ma, R., Ban, J., Wang, Q., Zhang, Y., Yang, Y., He, M.Z., Li, S., Shi, W., Li, T., 2021b. Random forest model based fine scale spatiotemporal O3 trends in the Beijing-Tianjin-Hebei region in China, 2010 to 2017. Environ. Pollut. 276, 116635. https://doi.org/10.1016/j.envpol.2021.116635.

Ma, X., Zou, B., Deng, J., Gao, J., Longley, I., Xiao, S., Guo, B., Wu, Y., Xu, T., Xu, X., Yang, X., Wang, X., Tan, Z., Wang, Y., Morawska, L., Salmond, J., 2024. A comprehensive review of the development of land use regression approaches for modeling spatiotemporal variations of ambient air pollution: a perspective from 2011 to 2023. Environ. Int. 183, 108430. https://doi.org/10.1016/j.envint.2024.108430.

Mogollón-Sotelo, C., Casallas, A., Vidal, S., Celis, N., Ferro, C., Belalcazar, L., 2021. A support vector machine model to forecast ground-level PM2.5 in a highly populated city with a complex terrain. Air Qual. Atmos. Health 14, 399–409. https://doi.org/10.1007/s11869-020-00945-0.

Mun, D.C., Kil, S.H., 2024. Research on valuation of ecosystem services for water quality improvement using unmanned aerial vehicles-Focusing on Purchased land in Gwangdong-ri area, Gwangju city (Gyeonggi). J. Korean Soc. Environ. Restor. Technol. 27, 1–16.

Nagar, P.K., Singh, D., Sharma, M., Kumar, A., Aneja, V.P., George, M.P., Agarwal, N., Shukla, S.P., 2017. Characterization of PM2.5 in Delhi: role and impact of secondary aerosol, burning of biomass, and municipal solid waste and crustal matter. Environ. Sci. Pollut. Res. Int. 24, 25179–25189. https://doi.org/10.1007/s11356-017-0171-3.

Nowak, D.J., Crane, D.E., Stevens, J.C., 2006. Air pollution removal by urban trees and shrubs in the United States. Urban For. Urban Green. 4 (3–4), 115–123.

Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., Moon, T., 2020. Estimating PM2.5 concentration of the conterminous United States via interpretable convolutional neural networks. Environ. Pollut. 256, 113395. https://doi.org/10.1016/j.envpol.2019.113395.

Phillips, B.B., Bullock, J.M., Osborne, J.L., Gaston, K.J., 2021. Spatial extent of road pollution: a national analysis. Sci. Total Environ. 773, 145589.
Pinho, P., Augusto, S., Martins-Loução, M.A., Pereira, M.J., Soares, A., Máguas, C., Branquinho, C., 2008. Causes of change in nitrophytic and oligotrophic lichen species in a Mediterranean climate: impact of land cover and atmospheric pollutants. Environ. Pollut. 154, 380–389. https://doi.org/10.1016/j.envpol.2007.11.028.
Rajput, D., Wang, W.J., Chen, C.C., 2023. Evaluation of a decided sample size in machine learning applications. BMC Bioinf. 24, 48. https://doi.org/10.1186/s12859-023-05156-9.
Roberts, P.T., Fryer-Taylor, R.E.J., Hall, D.J., 1994. Wind-tunnel studies of roughness effects in gas dispersion. Atmos. Environ. 28, 1861–1870. https://doi.org/10.1016/1352-2310(94)90325-5.
Shao, Y., Ma, Z., Wang, J., Bi, J., 2020. Estimating daily ground-level PM2.5 in China with random-forest-based spatiotemporal kriging. Sci. Total Environ. 740, 139761. https://doi.org/10.1016/j.scitotenv.2020.139761.
Smith, P.F., Ganesh, S., Liu, P., 2013. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. J. Neurosci. Methods 220, 85–91. https://doi.org/10.1016/j.jneumeth.2013.08.024.
Su, X., An, J., Zhang, Y., Zhu, P., Zhu, B., 2020. Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. Atmos. Pollut. Res. 11, 51–60. https://doi.org/10.1016/j.apr.2020.02.024.
Tao, W., Liu, J., Ban-Weiss, G.A., Hauglustaine, D.A., Zhang, L., Zhang, Q., Cheng, Y., Yu, Y., Tao, S., 2015. Effects of urban land expansion on the regional meteorology and air quality of eastern China. Atmos. Chem. Phys. 15, 8597–8614. https://doi.org/10.5194/acp-15-8597-2015.
Tella, A., Balogun, A.L., 2021. GIS-based air quality modelling: spatial prediction of PM10 for Selangor State, Malaysia using machine learning algorithms. Environ. Sci. Pollut. Res. 1-17.
Tuckett-Jones, B., Reade, T., 2017. City Air Quality at Height – Lessons for Developers & Planners WSP | Parsons Brinckerhoff.
Van Der Waals, J., 2000. The compact city and the environment: a review. Tijdschr. Econ. Soc. Geogr. 91, 111–121. https://doi.org/10.1111/1467-9663.00099.
Van Roode, S., Ruiz-Aguilar, J.J., González-Enrique, J., Turias, I.J., 2019. An artificial neural network ensemble approach to generate air pollution maps. Environ. Monit. Assess. 191, 727. https://doi.org/10.1007/s10661-019-7901-6.
Wang, W., Zhao, S., Jiao, L., Taylor, M., Zhang, B., Xu, G., Hou, H., 2019. Estimation of PM2.5 concentrations in China using a spatial back propagation neural network. Sci. Rep. 9, 13788. https://doi.org/10.1038/s41598-019-50177-1.
Wang, Z., Bai, Z., Yu, H., Zhang, J., Zhu, T., 2004. Regulatory standards related to building energy conservation and indoor-air-quality during rapid urbanization in China. Energy Build. 36, 1299–1308. https://doi.org/10.1016/j.enbuild.2003.09.013.
Wei, Y.D., Ye, X., 2014. Urbanization, urban land expansion and environmental change in China *Stoch*. Stoch. Environ. Res. Risk Assess. 28, 757–765. https://doi.org/10.1007/s00477-013-0840-9.
Weng, Q., Yang, S., 2006. Urban air pollution patterns, land use, and thermal landscape: an examination of the linkage using GIS. Environ. Monit. Assess. 117, 463–489. https://doi.org/10.1007/s10661-006-0888-9.
World Health Organization, 2021. New WHO global air quality guidelines aim to save millions of lives from air pollution. https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution.
Yang, S., Kim, H., Kim, S.N., Ahn, K., 2018. What is achieved and lost in living in a mixed-income neighborhood? Findings from South Korea. J. Hous. Built Environ. 33, 807–828. https://doi.org/10.1007/s10901-017-9586-x.
Yang, W., Jiang, X., 2021. Evaluating the influence of land use and land cover change on fine particulate matter. Sci. Rep. 11 (1), 17612.
Yang, X.F., Zheng, Y.X., Geng, G.N., Liu, H., Man, H.Y., Lv, Z.F., He, K.B., de Hoogh, K., 2017. Development of PM2.5 and NO2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. Environ. Pollut. 226, 143–153. https://doi.org/10.1016/j.envpol.2017.03.079.
Yu, G.H., Park, S., 2021. Chemical characterization and source apportionment of PM2.5 at an urban site in Gwangju, Korea. Atmos. Pollut. Res. 12, 101092. https://doi.org/10.1016/j.apr.2021.101092.
Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., Byambaa, T., Barn, P., Henderson, S.B., Janes, C. R., Lanphear, B.P., McCandless, L.C., Takaro, T.K., Venners, S.A., Webster, G.M., Allen, R.W., 2019. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environ. Pollut. 245, 746–753. https://doi.org/10.1016/j.envpol.2018.11.034.
Yue, T., Gao, X., Gao, J., Tong, Y., Wang, K., Zuo, P., Zhang, X., Tong, L., Wang, C., Xue, Y., 2018. Emission characteristics of NOx, CO, NH3 and VOCs from gas-fired industrial boilers based on field measurements in Beijing city, China. Atmos. Environ. 184, 1–8. https://doi.org/10.1016/j.atmosenv.2018.04.022.
Zahari, M.A.Z., Majid, M.R., Ho, C.S., Kurata, G., Nadhirah, N., Irina, S.Z., 2016. Relationship between land use composition and PM10 concentrations in Iskandar Malaysia. Clean Technol. Environ. Policy 18, 2429–2439. https://doi.org/10.1007/s10098-016-1263-3.
Zhang, P., Ma, W., Wen, F., Liu, L., Yang, L., Song, J., Wang, N., Liu, Q., 2021. Estimating PM2.5 concentration using the machine learning GA-SVM method to improve the land use regression model in Shaanxi, China. Ecotoxicol. Environ. Saf. 225, 112772. https://doi.org/10.1016/j.ecoenv.2021.112772.
Zhao, B., Yu, L., Wang, C., Shuai, C., Zhu, J., Qu, S., Taiebat, M., Xu, M., 2021. Urban air pollution mapping using fleet vehicles as mobile monitors and machine learning. Environ. Sci. Technol. 55, 5579–5588. https://doi.org/10.1021/acs.est.0c08034.