보건학석사 학위논문

Estimation of High-Spatial Resolution of Ground-Level Ozone, Nitrogen Dioxide, and Carbon Monoxide in South Korea During 2002-2020 Using Machine-Learning Based Ensemble Model

머신러닝 모델을 사용한 2002~2020년 한국의 O_3 , NO_2 , CO 농도의 고해상도 추정

2023년 2월

서울대학교 보건대학원

보건학과 보건학전공

권 도 훈

Estimation of High-Spatial Resolution of Ground-Level Ozone, Nitrogen Dioxide, and Carbon Monoxide in South Korea During 2002-2020 Using Machine-Learning Based Ensemble Model

지도 교수 김 호

이 논문을 보건학석사 학위논문으로 제출함
2022년 11월

서울대학교 보건대학원
보건학과 보건학전공
권 도 훈

권도훈의 보건학석사 학위논문을 인준함
2022년 11월

위 원 장 이 승 묵 (인)

부위원장 이 우 주 (인)

위 원 김 호 (인)

Abstract

Backrgound : Long-term exposure to ozone (O_3), nitrogen dioxide (NO_2), and carbon monoxide (CO) is known to cause various diseases and increase mortality. For that reason, estimating ground-level O_3, NO_2, and CO concentrations with a high spatial resolution is crucial for assessing the health effects associated with these air pollutants. However, related studies are limited in South Korea. This study aimed to develop machine learning-based models to predict the monthly O_3 (average of daily 8-hour maximums), NO_2, and CO at a spatial resolution of 1 km × 1 km across South Korea from 2002 to 2020.

Methods : Approximately 80% of the monitoring stations were used to train the three machine learning models (random forest, light gradient boosting, and neural network) with a 10-fold cross-validation, and 20% of the monitoring stations were used to test the model performance. The author also applied ensemble models to integrate the variation in predictions among the models. Multiple predictors with satellite-based remote sensing data, inverse distance weighted ground-level air pollutants, land use variables, reanalysis datasets for meteorological variables, and regional socioeconmoic variables collected from various databases were

included in the prediction model.

Results : For O_3, the overall R^2 of the ensemble model was 0.841 during the entire study period. Urban areas showed a better model performance (R^2 = 0.845) than rural areas (R^2 = 0.762). For NO_2, the highest overall R^2 was 0.756, which best fit in autumn (R^2 = 0.768). For CO, the overall R^2 value was 0.506. This study provides high spatial resolution monthly average O_3 and NO_2 estimates with excellent performance ($R^2 > 0.75$).

Conclusion : The author's predictions can be used to analyze the spatial patterns in pollutants in relation to population characteristics and studies on the health effects of long-term exposure to air pollution using geocode-based health information and local health data.

Keywords : Gaseous air pollution, Exposure assessment, High spatial resolution, Machine learning model, Ensemble model

Student Number : 2021-24226

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Numerous studies have consistently identified that exposure to ground-level gaseous air pollutants, such as ozone (O_3), nitrogen dioxide (NO_2), and carbon monoxide (CO), affects various health outcomes. Exposure to O_3 can cause diabetes mellitus (Li et al., 2021) and respiratory diseases (Lin et al., 2008; Rhee et al., 2019). Exposure to NO_2 is associated with cardiopulmonary system disorders (Dijkema et al., 2016), and mortality rate could be elevated by exposed to O_3 and NO_2 (Heinrich et al., 2013; Huang et al., 2021b; Lim et al., 2019; Niu et al., 2022). Also, long-term exposure to O_3, NO_2, and CO is related to cardiovascular diseases (Kim et al., 2017). Other studies found that the risk of lung and liver cancers might be associated with the gaseous air pollutants (Bălă et al., 2021; So et al., 2021; Yazdi et al., 2019), and the negative effects on health have been observed across all age groups: from newborns to elderly (Dimakopoulou et al., 2020; Heinrich et al., 2013; Huang et al., 2021b; Lin et al., 2008; Rhee et al., 2019). Recent studies have suggested that an increase in air pollution can affect Coronavirus 2019 infection (COVID-19) (Travaglio et al., 2021; Zheng et al., 2021) and its fatality (Garcia et al., 2022; Konstantinoudis et al., 2021).

Monitors of air pollution have been used in numerous studies to investigate how exposure to air pollution impacts human health outcomes, monitoring networks generally are disproportionately located in urban areas and even within cities do not fully capture spatial heterogeneity of air pollution exposure. Further, several studies have shown that air pollution monitors in some countries such as the United States and Brazil are disproportionately located in some communities, providing less information for other communities (Bravo et al., 2016; Ebisu et al., 2014). Further, some monitoring networks do not provide daily data. Regional air quality modeling, such as the Weather Research and Forecasting Model – Community Muldiscale Air Quality Modeling System (WRF–CMAQ) (Wong et al., 2012) can provide full spatial and temporal coverage, but are often time consuming and computationally costly to conduct for large areas at high spatial resolution.

Therefore, to estimate and prevent health impacts attributable to these gas pollutants, many studies have developed models to predict the group–level concentration of gaseous air pollutants in order to provide estimates of concentrations at times and locations for which monitoring data are not available. Most of these studies used spatial interpolation approaches with dispersion models with O_3, NO_2, and CO (Liu et al., 2019b), land use regression models

(LUR) for O_3 and NO_2 (Kerckhoffs et al., 2015; Rosenlund et al., 2008), LUR with chemical transport modeling approaches for O_3 (Wang et al., 2016), and LUR with satellite-based models for NO_2 (Chen et al., 2020; Vienneau et al., 2013; Young et al., 2016). In the case of NO_2, LUR with traffic-related factors has been widely used to estimate ground-level concentrations (Bechle et al., 2015; Chen et al., 2020; Larkin et al., 2017; Vienneau et al., 2013; Young et al., 2016).

However, these modeling approaches have limitations. First, spatial interpolation methods, such as inverse distance weighting (IDW) and kriging, are based on the hypothesis that air pollutants have a distance-decay relationship over the study area. However, considering only spatial correlation with a variogram is inadequate for considering complex geographical information (Lu and Wong, 2008) and the interpolation methods do not address geographical and meteorological factors that could affect air pollutants (Chen et al., 2012; Rosenlund et al., 2008). Further, such approaches are limited by the existing air pollution monitoring network, which may not well represent all areas. As an alternative approach, LUR with geographical and meteorological predictors has been widely performed (Chen et al., 2012; Chen et al., 2020), and while this approach is useful, it has disadvantages because it is based on

linear regression methods. Particularly, limitations can exist if some of the predictors have complex nonlinear relationships with air pollutants, if some variables are not fully available for the whole study area and time period, or if there are high-order interactions among predictors and pollutants (Zhan et al., 2018). Also, the same problem can arise with mixed-effect models and geographically weighted regression because these models assumed a linear relationship between predictor variables and outcome variable (Di et al., 2019b).

To address the limitations of conventional prediction models, recent studies have conducted prediction modeling based on machine-learning methods (Araki et al., 2021; Chen et al., 2021; Zhan et al., 2018). Nonetheless, few studies can address various types of predictors that may crucially contribute to the performance of prediction models because of limited data sources and problems in computational time and memory storage capacity, especially in relation to satellite-based remote sensing data that include multiple environmental, land-use, demographic, and meteorological variables (Gorelick et al., 2017).

This study aimed to develop machine learning-based prediction models for the monthly average concentrations of gaseous air

pollutants covering O_3, NO_2, and CO at a resolution of 1 km × 1 km across South Korea from 2002 to 2020. Satellite-based remote sensing data were mainly obtained from Google Earth Engine (GEE) (Tamiminia et al., 2020) and other predictors to increase prediction performance were collected from the Socioeconomic Data and Applications Center (SEDAC) and a database of community health outcomes and health determinants (hereafter, regional socioeconomic database) provided by the Korean Disease Control and Prevention Agency. To the best of the author's knowledge, this is the first study to develop machine learning-based air pollution prediction models that cover all areas in South Korea with high spatial resolution and long timeframe.

Chapter 2. Materials and Methods

2.1. Study area

This study covered the entire region of South Korea for January 2002 to December 2020. Because there were limitations in collecting reliable remote sensing data, the author excluded island areas from this study. The total number of grids was 97,653 in the entire study area, with gridcell resolution at 1 km × 1 km.

2.2. Air pollution monitoring data

As response variables for prediction models, the author collected ground-level hourly measured O_3, NO_2, and CO concentrations from the Air Korea database provided by the Korea Environment Corporation (URL is presented in Table S1). To reduce potential observation biases, the author used concentration data from monitoring sites with observations for ≥ 9 months per year for a given pollutant. The total number of selected monitoring sites was 480 for O_3 and NO_2, and 447 for CO. From the selected monitoring sites, the author calculated the monthly average of daily maximum 8-h O_3, monthly average of daily NO_2 values, and monthly average of daily CO.

Because monitoring stations are not equally distributed across the study area and the monitoring data from nearby monitoring sites are more correlated than data from faraway sites, the author used the IDW, a commonly applied spatial interpolation method. Specifically, the author used monitoring data to compute the IDW for O_3, NO_2, and CO at each 1 km × 1 km grid and added these estimations as predictor variables in the author's model.

2.3. Satellite-based remote sensing data

The author extracted multiple remote sensing variables from the GEE and SEDAC, including meteorological data (with AOD), land-use data, and surface reflectance. The author aggregated all collected predictor variables at each 1 km × 1 km grid cell, and calculated the monthly averages or categorical value that appears most in the month for each grid cell. If provided resolution of a variable is coarser than 1 km × 1 km (e.g. 11 km × 11 km), the author assigned the value of the coarser resolution grid cell to all 1 km × 1 km grid cells within that larger grid cell. A full list of 78 remote sensing variables can be found in the supplementary material (Section 1.2 and Table S1).

2.3.1. Meteorological data

O_3, NO_2, and CO concentrations can be affected spatially and temporally by meteorological factors such as temperature, wind speed and direction, precipitation, humidity, and cloud droplets (Requia et al., 2020; Yinusa et al., 2019; Zhan et al., 2018). Meteorological variables were collected from various reanalysis datasets, and monthly aggregates of air temperature, soil temperature, surface pressure, 10-m u-component, and v-component of wind (eastward and northward components of the 10-m wind) were collected from the 5th generation European Center for Medium-Range Weather Forecasts atmospheric reanalysis (ERA5) and surface-based reanalysis (ERA5-Land). The author aggregated the temporal resolutions of these datasets from daily values to month. The author also obtained the total water column density, which is the percentage of total cloud cover, from the National Centers for Environmental Prediction (NCEP). To obtain more information about sky coverage, the author also collected day and night clear-sky coverage from MOD11A1 v061, cloud cirrus area fraction, and liquid water cloud optical thickness from MOD08_M3 v061. The author retrieved merged satellite-gauge precipitation estimates and accumulation-weighted probabilities of the liquid precipitation phase from global precipitation measurement (GPM).

Due to the influence of aerosols on UV flux and photochemical reaction (Bian et al., 2007) the fact that sharing the same source of emissions between CO and AOD (e.g. biomass burning emission) (Andreae, 2019; Buchholz et al., 2021), O_3 and CO are generally considered to be related to AOD (Buchholz et al., 2021; Liu et al., 2019a). The Moderate Resolution Imaging Spectroradiometer (MODIS) is a widely used satellite-based sensor that provides various remote sensing data types, including AOD. Since AOD is dependent on wavelength, the author obtained AOD data retrieved at 0.47 μm and 0.55 μm from Terra & Aqua combined Multi-angle Implementation of Atmospheric Correction (MAIAC) Land Aerosol Optical Depth (MCD19A2 v006). AOD at 0.55 μm for both ocean and land, and corrected AOD (land) at 0.47 μm were collected from MOD08_M3 v061. Besides AOD, by referring to the variables used in the previous ozone estimation study (Requia et al., 2020), the author collected the total column O_3 from the Total Ozone Mapping Spectrometer (TOMS) and Ozone Monitoring Instrument (OMI) data (Parsons et al., 2010) as satellite-based air quality data to potentially account for ground-level ozone concentrations (Colombi et al., 2021).

The SEDAC dataset provides global annual $PM_{2.5}$ estimates by combining AOD from various data sources by combining MODIS.

Since this dataset is accessible online with fine spatial grid resolution (0.02° × 0.02°), the author extracted the annual global surface $PM_{2.5}$ concentrations for each gridcell from the SEDAC. Detailed information is presented in the supplementary material (Section 1.2 and Table S1).

2.3.2. Land-use data

Land-use information is important to enhance prediction performance, especially for estimates of air pollutant concentrations, because it can partly explain the fine-scale spatial pattern or distribution of air pollutants (Huang et al., 2021a). Previous studies have reported that land types and land covers, such as vegetation index and various types of land cover fractions, were also considered relevant variables for estimating air pollutants (Chen et al., 2020; Kerckhoffs et al., 2015; Zhu et al., 2022). Thus, the author extracted land-use variables related to greenness from MODIS, Copernicus Global Land Cover Layers, and the Global Land Cover Map. The normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) were derived (MOD13A2 v006), and the leaf area index (LAI) and the fraction of absorbed photosynthetically active radiation (FPAR) were derived (MCD15A3H v061). The leaf area index with high/low vegetation

was obtained from the ERA5-Land reanalysis dataset.

The author also accessed several land cover layer datasets with information about land types (non-vegetated barren, forests, water bodies, shrubland, and others). The MCD12Q1 V6 product provides the global land cover types for each year. Land-cover types 1-5 were collected from MCD12Q1 V6. The FAO-Land Cover Classification System 1 (LCCS1) land cover layer, FAO-LCCS2 land use layer, FAO-LCCS3 surface hydrology layer, and their confidence levels (0-100%) were collected from MCD12Q1 V6. Other types of land cover layers from the Copernicus Global Land Cover Layers and land cover map variables from the Global Land Cover Map were included as predictor variables in the modeling procedure. More detailed information about the land-use data is provided in the supplementary material (Section 1.2.1 and Table S1).

2.3.3. Surface reflectance

Surface albedo and reflectivity may be associated with ground-level O_3 and NO_2 concentrations through interactions with other materials (Jandaghian and Akbari, 2020; Taha, 1997). To consider this in the process of estimating gaseous pollutants, the author retrieved the black/white sky and bidirectional reflectance

distribution function (BRDF) albedo from MCD43A3 V6. The author also collected Band 1–5, Band 7 of surface reflectance, and Band 6 of surface temperature from Landsat 7 created using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) algorithm. Similarly, the emissivities from bands 31 and 32 were obtained from MOD11A1 v061.

2.4. Regional socioeconomic predictors

Various regional demographic, socioeconomic, and environmental factors from various data sources have been collected across all district-level regions annually by the Korea Disease Control and Prevention Agency, and then incorporated into the community health outcomes and regional socioeconomic database, with thousands of variables since 2008. Since this database consists of the district-level regions, the grid values were set as allocated district-level variables by finding the grids included in each district. The author selected from this database 24 regional variables that could potentially be related to air pollutant concentrations.

O_3 is produced by a chemical reaction between NO_x and VOCs under various meteorological conditions. The main factors generating NO_2 and CO are emissions from traffic and industrial

sources (Kim et al., 2013; Rosenlund et al., 2008). These factors can be considered urbanized characteristics, such as population density, greenness area, number of cars, wastewater, and organic material load discharge (Araki et al., 2021; Carslaw and Rhys-Tyler, 2013; Khalid, 2021). These factors are also associated with differences in the demographic structure and infrastructure among districts (Glover and Simon, 1975; Khalid, 2021). Thus, the author obtained variable that represents urbanization from the regional socioeconomic database. A more detailed description and the full list of variables are presented in the supplementary material (Section 1.2.3 and Table S1).

2.5. Modeling procedures

The author adopted three machine learning-based models, namely random forest, light gradient boosting, and neural network, to predict monthly O_3, NO_2, and CO averages using a 1 km × 1 km grid during 2002-2020. Previous studies have also used these models to predict air pollutant concentrations in other locations (Di et al., 2019a; Di et al., 2019b; Requia et al., 2020). A total of 112 predictor variables collected from the GEE, SEDAC, regional socioeconomic database, and others were used as input variables,

and each air pollutant concentration was predicted as the outcome value. Missing values in the predictor variables were replaced with values from the imputation procedure (supplementary material Section 2.1). Randomly selected 80% of the monitoring stations were used to train the model, and the remaining 20% of the monitoring stations were used to test the model performance. To avoid overfitting and the possibility that the dataset was extracted by chance, the author trained each of the three machine learning models with 10-fold cross-validation (CV) in the training set. With these trained models, the author checked the model performance of three machine learning models and simple averages among models (ensemble model) in the test set using R^2 and root mean squared error (RMSE). Finally, the author predicted monthly averages of each air pollutant with a 1 km × 1 km grid during 2002-2020 using the three machine learning-based and ensemble models. The overall process is illustrated in Fig. 1.

2.5.1. Data Preprocessing

Missing values were addressed prior to the modeling procedures (Table S2). The author imputed missing values using the random forest model and linear interpolation method following previous studies (Di et al., 2019a; Di et al., 2019b; Requia et al., 2020).

Detailed information about the missing data imputation procedure is provided in supplementary material (Section 2.1). After the missing value imputation process, the author standardized all the predictor variables using average and standard deviation, separately for each variable, to control the variation within the individual variable. For detail, the author let "X" be a predictor variable, then the author transformed this variable to $X_{standardized} = \frac{X - mean(X)}{sd(X)}$. The author also added yearly and monthly terms, seasonal terms, binary indicators of the COVID-19 pandemic period, monthly terms of the fourth highest value month for each air pollutant, urban binary indicators, and binary indicators of metropolitan city areas. (Supplementary material section 1.2.4 and Fig. S1).

2.5.2. Machine learning-based model

In previous studies, random forest, gradient boosting, and neural networks were used for estimating $PM_{2.5}$ (Di et al., 2019b), O_3 (Requia et al., 2020), and NO_2 (Di et al., 2019a). A random forest is operated by aggregating decision trees from bootstrapped data to reduce the correlation between the trees; therefore, the random forest can reduce the variance of estimations (Hastie et al., 2009). Otherwise, gradient boosting focused on reducing the bias of estimations by adding week learners sequentially to fit residuals

from the previous model prediction. In the neural network case, which comprises several hidden layers with various activation functions, repeatedly updates the weights on the hidden layers across every epoch to reduce bias. Given these characteristics, prediction performance can differ according to the model used. Additionally, within each model, the prediction performance can be affected by hyperparameter settings. For example, the number of trees and the maximum depth of each tree can affect the random forest model performance, and the learning rate in gradient boosting and the number of layers and units in a neural network can also influence the model performance. Thus, the author optimized the best parameters with a 10-fold CV for each model in a grid search process. Detailed information about machine learning models and results from the grid search process are shown in supplementary material (Section 2.2 and Table S3).

2.5.3. Ensemble Model

Given the differences in the characteristics of each model, the performance and estimation results appeared slightly different by space and time. To aggregate the results, the author calculated the simple averages of each machine learning estimation.

$$\hat{Y}_{SA_{ij}} = \frac{\hat{Y}_{rf_{ij}} + \hat{Y}_{gb_{ij}} + \hat{Y}_{nn_{ij}}}{3}$$

$\hat{Y}_{rf_{ij}}$, $\hat{Y}_{gb_{ij}}$, and $\hat{Y}_{nn_{ij}}$ are air pollution estimations from the random forest, light gradient boosting, and neural network, respectively, at location i at time j; $\hat{Y}_{SA_{ij}}$ is the simple average estimation derived by averaging the three estimations at location i at time j. The author also trained a generalized additive model (GAM) to consider the geographical variation of each of the three machine-learning estimations. Detailed information about the GAM is presented in the supplementary material (Section 2.3 and Table S4).

2.5.4. Model Prediction

The monthly concentrations of each air pollutant were predicted using three trained machine learning models, the average prediction from the machine learning models, and GAM. Consequently, the author generated five datasets of predicted values of monthly O_3, NO_2, and CO averages at a 1 km × 1 km resolution across the study area from 2002 to 2020.

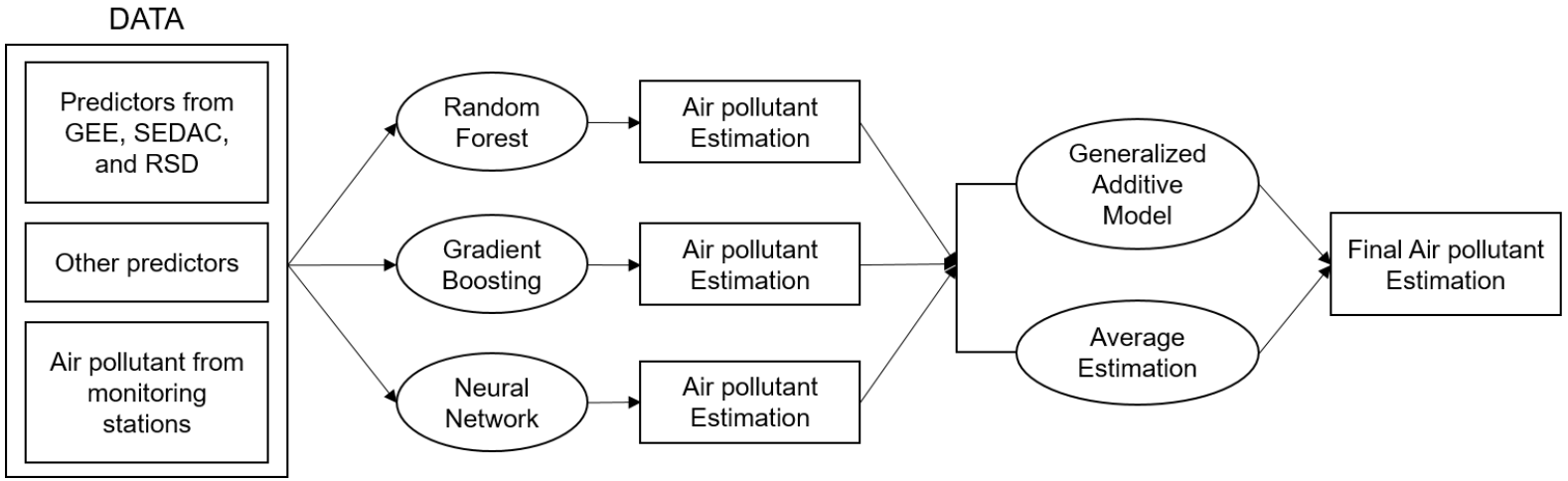

Fig. 1. Flowchart of the modeling process. GEE: Google Earth Engine, SEDAC: Socioeconomic Data and Applications Center, RSD: Regional Socioeconomic Database from Korean Disease Control and Prevention Agency

Chapter 3. Results

Table 1 presents the overall R^2. The overall R^2 for O_3, NO_2, and CO was 0.841, 0.756, and 0.506, respectively. The overall RMSE for O_3, NO_2, and CO were 5.435 (ppb), 4.867 (ppb), and 0.152 (ppm), respectively. The author's models showed excellent performances for O_3 and NO_2. Among the results of the three machine learning and ensemble models (SA: Simple Average), the ensemble model outperformed the other models for O_3 and NO_2, whereas the random forest model showed slightly better performance than the ensemble model for CO. The author also fit the GAM; however, because the predictive performance of the GAM was lower than that of the SA, the author does not include it in Table 1 and show its performance in Table S4.

Table 1 also presents the R^2 for overall and for three- or four-year time periods. For O_3, the annual R^2 of ensemble model (SA) varied from 0.732 to 0.874 across the three- and four-year time periods. For NO_2, the annual R^2 of ensemble model ranged from 0.538 to 0.861. For CO, the R^2 of ensemble model varied from 0.302 to 0.553. The author's model performance was higher in more recent years for O_3 and NO_2. Except for the ensemble model, the light gradient boosting and random forest models usually

showed high R^2 values for NO_2 and CO. Among the seasons, for O_3, NO_2, and CO the predictive R^2 was highest in autumn (Table S5). The author's ensemble model showed the best performance for the whole season for O_3 and NO_2, while the random forest was the best model for CO.

Table 1. Model performance for O_3, NO_2, and CO overall and in three- and four-year periods

		R^2				RMSE			
O_3 (ppb)	years	RF	GB	NN	SA	RF	GB	NN	SA
	2002~2005	0.715	0.724	0.736†	0.732	6.582	6.532	6.319†	6.402
	2006~2008	0.805	0.799	0.801	0.808†	5.762	5.802	5.766	5.684†
	2009~2011	0.838	0.836	0.839	0.843†	5.030	5.048	5.017	4.948†
	2012~2014	0.853	0.853	0.845	0.854†	5.186	5.215	5.343	5.172†
	2015~2017	0.868	0.874	0.869	0.874†	5.259	5.133	5.23	5.119†
	2018~2020	0.842	0.841	0.827	0.843†	5.249	5.234	5.488	5.201†
	overall	0.837	0.837	0.834	0.841†	5.524	5.500	5.557	5.435†
NO_2 (ppb)		R^2				RMSE			
	years	RF	GB	NN	SA	RF	GB	NN	SA
	2002~2005	0.527	0.541†	0.489	0.538	6.864	6.838	7.189	6.768†

	2006~2008	0.733	0.735	0.736	0.746†	5.181	5.114	5.134	5.037†
	2009~2011	0.75	0.756	0.763	0.768†	4.900	4.778	4.785	4.717†
	2012~2014	0.713	0.733	0.733	0.735†	5.343	5.113†	5.147	5.129
	2015~2017	0.715	0.736†	0.731	0.736	4.836	4.634†	4.722	4.668
	2018~2020	0.844	0.864†	0.849	0.861	3.595	3.266†	3.482	3.352
	overall	0.741	0.754	0.741	0.756†	5.035	4.877	5.010	4.867†
		R^2				RMSE			
	years	RF	GB	NN	SA	RF	GB	NN	SA
CO (ppm)	2002~2005	0.320†	0.283	0.212	0.302	0.228†	0.234	0.246	0.231
	2006~2008	0.505	0.502	0.431	0.508†	0.190	0.189†	0.209	0.193
	2009~2011	0.545	0.544	0.499	0.553†	0.148	0.148†	0.162	0.150
	2012~2014	0.517	0.515	0.492	0.527†	0.134†	0.134	0.145	0.135
	2015~2017	0.430	0.424	0.442	0.45†	0.120	0.120	0.121	0.118†
	2018~2020	0.489	0.470	0.451	0.491†	0.095	0.096	0.100	0.095†

overall	0.506†	0.492	0.438	0.505	0.152†	0.153	0.164	0.153

* RF: Random forest, GB: light gradient boosting, NN: neural network, SA: simple average estimation of RF, GB, and NN.

* The performance for O_3 and NO_2 was calculated based on ppb and for CO on ppm.

† A model that performs better than other models during the period.

Table S5 also presents the R^2 by urbanicity (urban or rural), with higher R^2 values in urban areas than in rural areas for all air pollutants. The simple average estimations showed the highest R^2 values among the models. The prediction performances of the random forest and light gradient boosting models were similar. Fig. S2–S4 shows the spatiotemporal patterns of the O_3, NO_2, and CO prediction distributions across the study period. Overall, annual and seasonal O_3 concentrations increased consistently over time, whereas decreasing patterns were observed for NO_2 and CO.

Fig. 2 displays the density scatter plot for the monthly averages of the monitored and predicted concentrations for each air pollutant. Although most points approximate a 1:1 straight line of monitored and predicted relationships for O_3 and NO_2, representing equal agreement, this was less so for CO, especially at very high and very low observed concentrations. Fig. 3 shows a map of the monitored and predicted concentrations.

The author reported the percentage decrease in R^2 when omitting each grouped predictor variable from each model (Fig. 4). The overall impact of IDW was more significant than the other grouped variables for O_3 and NO_2; however, it was not critical in the random forest model. The variable with the greatest impact on CO varied by model. Meteorological and regional variables were slightly more

important than other grouped variables in the random forest and gradient boosting models, but in the neural network model, IDW had a very strong effect.

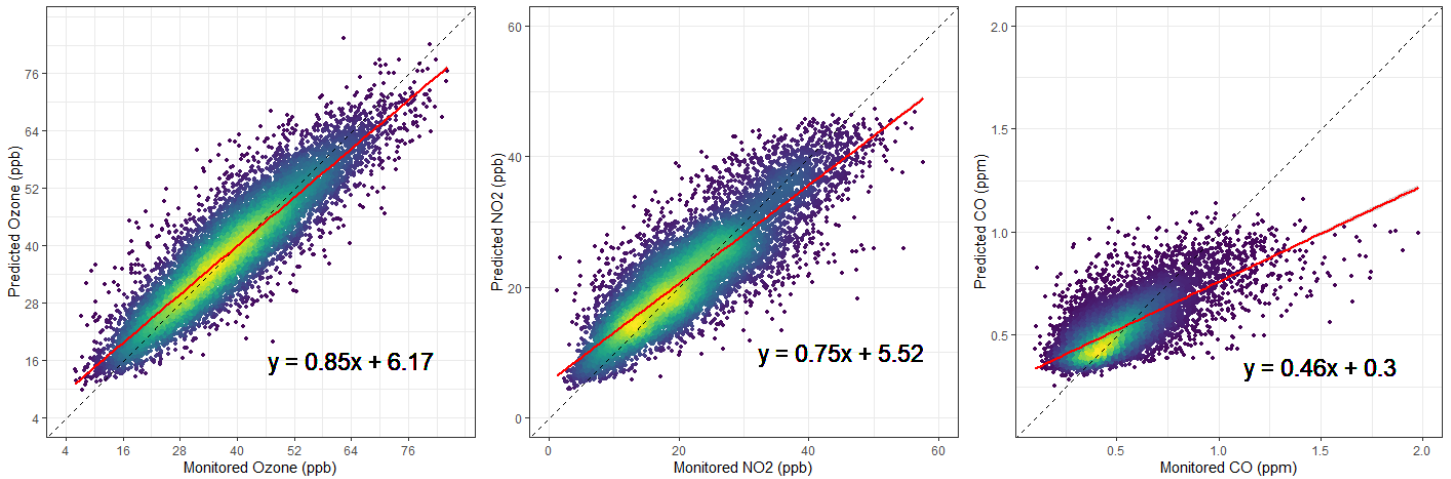

Fig. 2. Density scatter plot for monthly averages of the monitored and predicted concentrations of O_3, NO_2, and CO

* Dashed lines represent that the monitored and predicted estimations are the same for each air pollutant.

* Red lines represent the fitted line. x : ground-based measurements. y : estimated surface concentrations.

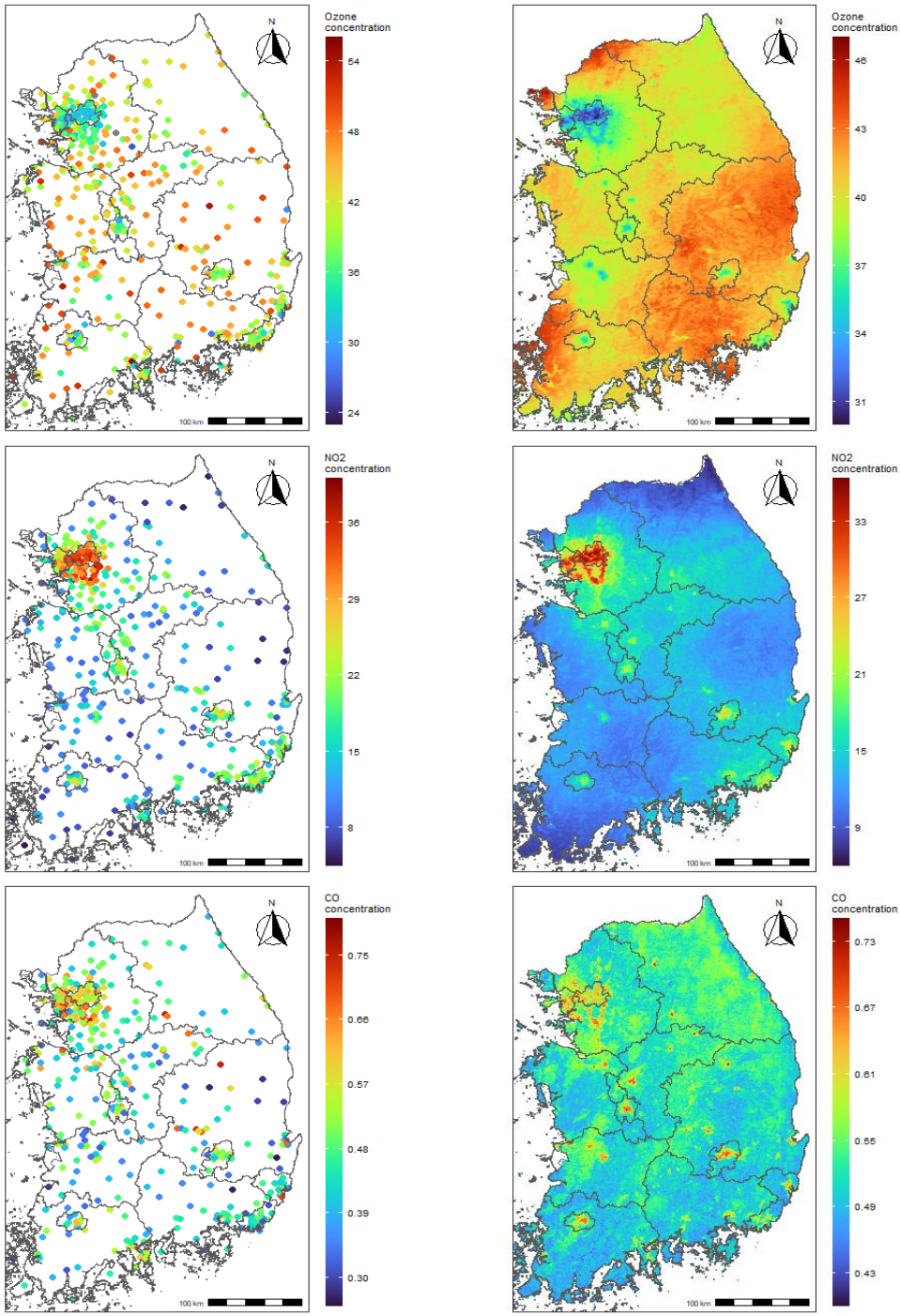

Fig. 3. Maps of monitored and predicted O_3, NO_2 and CO during 2002~2020

Left figures: Overall monitoring observations at point locations.

Right figures: Overall predicted estimations for each 1 × 1 km grid cell across entire South Korea.

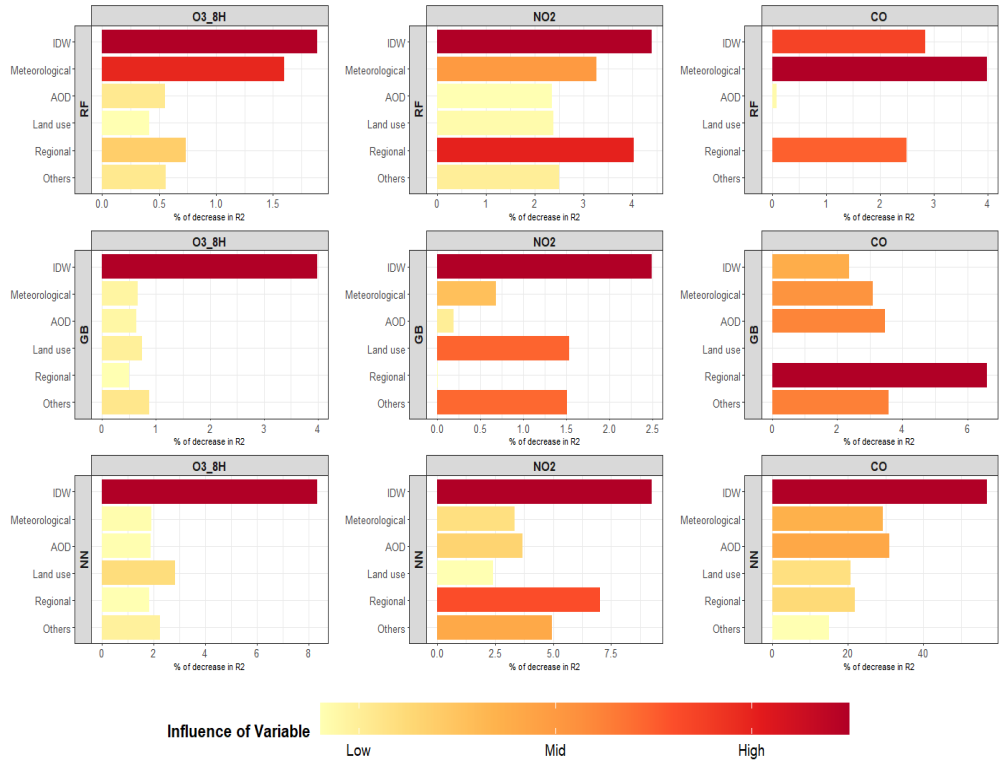

Fig. 4. Percentage decrease in R^2 when excluding grouped variables from each machine learning model of O_3, NO_2, and CO. The closer the color is to red, the greater the effect of the variables on the model performance

* The main vertical axis represents models, and the main horizontal axis represents air pollutants.

* X-axis (for each figure): % decrease in R^2.

* Y-axis (for each figure): Classified group variables

  - IDW: Inverse Distance Weighted estimations for O_3, NO_2, and CO

  - Meteorological: Meteorological variables (e.g. temperature, humidity, precipitation)

  - AOD: Aerosol Optical Depth variables

  - Land-use: Land-use variables (e.g. forest type, landcover, lc_prop1_categorical)

  - Regional: District-level variables (e.g. population density, park area per capita),

  - Others: Other satellite-based variables and added terms (e.g. landsat, albedo, surface reflectance, urban binary indicator)

* RF: Random Forest, GB: light Gradient Boosting, NN: Neural Network

Chapter 4. Discussion

In this study, the author developed machine learning models (random forest, light gradient boosting, and neural network) to predict monthly O_3, NO_2, and CO concentrations. Consequently, for the first time in South Korea, the author estimated the monthly O_3, NO_2, and CO averages across the contiguous region of South Korea at each 1 km × 1 km grid cell for 2002 to 2020. The model performed well to predict O_3 and NO_2, with R^2 values of 0.841 and 0.756, respectively.

Many studies have estimated gaseous pollutants using machine learning models. In the U.S., the estimated daily maximum of 8-h O_3 and daily NO_2 at 1 km × 1 km across the continental United States using multiple machine learning models and a geographically weighted generalized additive model during 2000–2016, with an overall R^2 of 0.9 and 0.788, respectively (Di et al., 2019a; Requia et al., 2020). In China, the daily maximum of 8-h O_3 and daily NO_2 concentrations were predicted at 0.0625° × 0.0625° and 0.1° × 0.1° grids across mainland China for 2008–2019 and 2013–2016, respectively (Chen et al., 2021; Zhan et al., 2018), using hybrid random forest models with site-based monthly R^2 at 0.82 and 0.65, respectively. Another study conducted in China estimated

ground-level monthly O_3 using extreme gradient boosting for regression with a site-based monthly R^2 of 0.68 (Liu et al., 2020). A previous study for Japan predicted national-scale 1 km × 1 km monthly O_3, NO_2, and other air pollutants by LUR structure adopting a random forest model during 2010-2015, with a R^2 of 0.86 and 0.84 (Araki et al., 2021). The overall predicted performances for O_3 and NO_2 were almost equivalent to or outperformed those of other related studies.

Most studies estimating gaseous air pollutants for South Korea used LUR for modeling and focused on specific regions with relatively short time periods (Choi et al., 2017; Kim and Guldmann, 2011; Kim and Guldmann, 2015). A previous study estimated the concentrations of O_3 and NO_2 in South Korea using machine learning models for 2018-2020 (Kang et al., 2021); however, the spatial resolution was relatively coarse (6 km × 6 km) and the study period was not sufficient to consider the long-term health impact of air pollutants. The author addressed the weaknesses of LUR using multiple machine learning models and their ensemble results by averaging each prediction estimate. To the best of the author's knowledge, this study is the first to cover South Korea with fine resolution and over a relatively long period.

The performance of author's model rapidly improved after 2006 for all the air pollutants. The author postulate that this might be due to an increase in the number of monitoring stations. Before 2006, the total number of monitoring stations was less than 200. In 2007, the number of monitoring stations in urban areas surpassed 200 and the number in rural areas was 10. Since 2007, a larger number of observation stations are in operation to better estimate the distribution of air pollution concentrations across contiguous South Korea, with over 200 urban areas and about 100 monitors in rural areas in 2020. Additionally, differences in monitoring networkslikely explain differences in performance between urban and rural areas, with higher model performances in urban areas than in rural areas (Table S5). There have been fewer monitoring stations in rural areas than in urban areas in the past (Table S6), which means the author's models are better able to estimate pollution in urban areas in the earlier years of the author's study time period. However, as the highest R^2 of the rural area was over 0.75 for O_3 and NO_2, this study also performed well, even in rural areas.

This study had several limitations. First, for NO_2 and CO, location-based emission information derived from local industrial and traffic sources are usually considered primary predictor

variables (Kim et al., 2013; Rosenlund et al., 2008), such as power plants, road length, and distance to the road (Araki et al., 2021; Wong et al., 2021). These factors have been important in estimating NO_2 and CO concentrations in previous studies, although datasets with sufficient information on point sources were not available. However, the author calculated the R^2 of each model by removing district-level variables, including the number of vehicle registrations, wastewater, and organic material load generation and discharge, which did not appear to have a substantial effect on model performance. Second, the model performances in each season were lower than the overall performances for O_3 and CO. This finding was consistent with a previous study for O_3 (Araki et al., 2021); thus, this issue should be addressed in future studies by adding variables considering the seasonal variation of each air pollutant. Third, the author did not include some island regions in the author's study area to focus on improving air pollutant estimation performance in South Korea, due to the lack of data for some of the study variables. Further research should consider these islands. Fourth, as noted above the monitoring network better reflected urban areas than rural areas, especially in the earlier years of this study period. Finally, the absence of location-based emission-related information may affect the prediction performance

of this models, such as lower predicted CO performance compared to the other pollutants. Also, measurement error could occur in CO due to the measurement analyzer. Non-Dispersive Infrared (NDIR) analyzer are used for monitoring CO concentration, and Gas Filter Correlation (GFC) are adopted on NDIR for detecting lower CO concentration (< 1 ppm) to cover the shortcomings of NDIR, which has a problem of detecting low CO concentrations. However, due to the interference effects by other gases, the accuracy of GFC analyzer could be reduced (Dinh et al., 2017). It can be associated with unstable variation of observed and predicted carbon monoxide.

Chapter 5. Conclusion

To author's knowledge, this is the first nationwide study of South Korea to estimate monthly averages of O_3, NO_2, and CO for a long timeframe (from 2002 to 2020) across contiguous South Korea by aggregating remote sensing data and regional socioeconomic databases. Random forest, light gradient boosting, and neural network algorithms were used to train the model with CV. The author integrated the prediction estimate of each machine learning method by using simple averaging and GAM, and finally, machine learning and ensemble models produced monthly averages of O_3, NO_2, and CO at each 1 km $\times$ 1 km grid cell. The author's ensemble model showed excellent performance compared to previous studies, with R^2 values for O_3 and NO_2 of 0.841 and 0.756, respectively. The author's predictions can be utilized to estimate the health impact of each air pollutant with both individual-level geocodes and regional datasets in South Korea, by providing highly spatially resolved monthly estimates for times and locations without monitors.

Bibliography

Andreae MO. Emission of trace gases and aerosols from biomass burning – an updated assessment. Atmos. Chem. Phys. 2019; 19: 8523–8546.

Araki S, Hasunuma H, Yamamoto K, Shima M, Michikawa T, Nitta H, et al. Estimating monthly concentrations of ambient key air pollutants in Japan during 2010–2015 for a national–scale birth cohort. Environmental Pollution 2021; 284: 117483.

Bălă G–P, Râjnoveanu R–M, Tudorache E, Motișan R, Oancea C. Air pollution exposure—the (in) visible risk factor for respiratory diseases. Environmental Science and Pollution Research 2021; 28: 19615–19628.

Bechle MJ, Millet DB, Marshall JD. National spatiotemporal exposure surface for NO2: monthly scaling of a satellite–derived land–use regression, 2000–2010. Environmental science & technology 2015; 49: 12297–12305.

Bian H, Han S, Tie X, Sun M, Liu A. Evidence of impact of aerosols on surface ozone concentration in Tianjin, China. Atmospheric Environment 2007; 41: 4672–4681.

Bravo MA, Son J, de Freitas CU, Gouveia N, Bell ML. Air pollution and mortality in São Paulo, Brazil: Effects of multiple

pollutants and analysis of susceptible populations. Journal of Exposure Science & Environmental Epidemiology 2016; 26: 150-161.

Buchholz RR, Worden HM, Park M, Francis G, Deeter MN, Edwards DP, et al. Air pollution trends measured from Terra: CO and AOD over industrial, fire-prone, and background regions. Remote Sensing of Environment 2021; 256: 112275.

Carslaw DC, Rhys-Tyler G. New insights from comprehensive on-road measurements of NOx, NO2 and NH3 from vehicle emission remote sensing in London, UK. Atmospheric Environment 2013; 81: 339-347.

Chen G, Chen J, Dong G-h, Yang B-y, Liu Y, Lu T, et al. Improving satellite-based estimation of surface ozone across China during 2008-2019 using iterative random forest model and high-resolution grid meteorological data. Sustainable Cities and Society 2021; 69: 102807.

Chen L, Wang Y, Li P, Ji Y, Kong S, Li Z, et al. A land use regression model incorporating data on industrial point source pollution. Journal of Environmental Sciences 2012; 24: 1251-1258.

Chen T-H, Hsu Y-C, Zeng Y-T, Lung S-CC, Su H-J, Chao HJ, et al. A hybrid kriging/land-use regression model with Asian

culture-specific sources to assess NO2 spatial-temporal variations. Environmental Pollution 2020; 259: 113875.

Choi G, Bell ML, Lee J-T. A study on modeling nitrogen dioxide concentrations using land-use regression and conventionally used exposure assessment methods. Environmental Research Letters 2017; 12: 044003.

Colombi N, Miyazaki K, Bowman KW, Neu JL, Jacob DJ. A new methodology for inferring surface ozone from multispectral satellite measurements. Environmental Research Letters 2021; 16: 105005.

Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, et al. Assessing NO2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. Environmental science & technology 2019a; 54: 1372-1384.

Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, et al. An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. Environment international 2019b; 130: 104909.

Dijkema MBA, van Strien RT, van der Zee SC, Mallant SF, Fischer P, Hoek G, et al. Spatial variation in nitrogen dioxide concentrations and cardiopulmonary hospital admissions.

Environmental Research 2016; 151: 721-727.

Dimakopoulou K, Douros J, Samoli E, Karakatsani A, Rodopoulou S, Papakosta D, et al. Long-term exposure to ozone and children's respiratory health: Results from the RESPOZE study. Environmental research 2020; 182: 109002.

Dinh T-V, Ahn J-W, Choi I-Y, Song K-Y, Chung C-H, Kim J-C. Limitations of gas filter correlation: A case study on carbon monoxide non-dispersive infrared analyzer. Sensors and Actuators B: Chemical 2017; 243: 684-689.

Ebisu K, Belanger K, Bell ML. Association between airborne PM_{2.5}chemical constituents and birth weight—implication of buffer exposure assignment. Environmental Research Letters 2014; 9: 084007.

Garcia E, Marian B, Chen Z, Li K, Lurmann F, Gilliland F, et al. Long-term air pollution and COVID-19 mortality rates in California: Findings from the Spring/Summer and Winter surges of COVID-19. Environmental Pollution 2022; 292: 118396.

Glover DR, Simon JL. The effect of population density on infrastructure: the case of road building. Economic Development and Cultural Change 1975; 23: 453-468.

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R.

Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 2017; 202: 18-27.

Hastie T, Tibshirani R, Friedman J. Random forests. The elements of statistical learning. Springer, 2009, pp. 587-604.

Heinrich J, Thiering E, Rzehak P, Krämer U, Hochadel M, Rauchfuss KM, et al. Long-term exposure to NO2 and PM10 and all-cause and cause-specific mortality in a prospective cohort of women. Occupational and environmental medicine 2013; 70: 179-186.

Huang D, He B, Wei L, Sun L, Li Y, Yan Z, et al. Impact of land cover on air pollution at different spatial scales in the vicinity of metropolitan areas. Ecological Indicators 2021a; 132: 108313.

Huang S, Li H, Wang M, Qian Y, Steenland K, Caudle WM, et al. Long-term exposure to nitrogen dioxide and mortality: a systematic review and meta-analysis. Science of The Total Environment 2021b; 776: 145968.

Jandaghian Z, Akbari H. Effects of increasing surface reflectivity on aerosol, radiation, and cloud interactions in the urban atmosphere. Theoretical and Applied Climatology 2020; 139: 873-892.

Kang Y, Choi H, Im J, Park S, Shin M, Song C-K, et al. Estimation of surface-level NO2 and O3 concentrations using TROPOMI data and machine learning over East Asia. Environmental Pollution 2021; 288: 117711.

Kerckhoffs J, Wang M, Meliefste K, Malmqvist E, Fischer P, Janssen NA, et al. A national fine spatial scale land-use regression model for ozone. Environmental research 2015; 140: 440-448.

Khalid KM. Correlation between Air Quality and Wastewater Pollution. Environmental Sustainability-Preparing for Tomorrow. IntechOpen, 2021.

Kim H, Kim J, Kim S, Kang SH, Kim HJ, Kim H, et al. Cardiovascular effects of long-term exposure to air pollution: a population-based study with 900 845 person-years of follow-up. Journal of the American Heart Association 2017; 6: e007170.

Kim NK, Kim YP, Morino Y, Kurokawa J-i, Ohara T. Verification of NOx emission inventory over South Korea using sectoral activity data and satellite observation of NO2 vertical column densities. Atmospheric Environment 2013; 77: 496-508.

Kim Y, Guldmann J-M. Impact of traffic flows and wind directions on air pollution concentrations in Seoul, Korea. Atmospheric

Environment 2011; 45: 2803-2810.

Kim Y, Guldmann J-M. Land-use regression panel models of NO2 concentrations in Seoul, Korea. Atmospheric Environment 2015; 107: 364-373.

Konstantinoudis G, Padellini T, Bennett J, Davies B, Ezzati M, Blangiardo M. Long-term exposure to air-pollution and COVID-19 mortality in England: a hierarchical spatial analysis. Environment international 2021; 146: 106316.

Larkin A, Geddes JA, Martin RV, Xiao Q, Liu Y, Marshall JD, et al. Global land use regression model for nitrogen dioxide air pollution. Environmental science & technology 2017; 51: 6957-6964.

Li Y-L, Chuang T-W, Chang P-y, Lin L-Y, Su C-T, Chien L-N, et al. Long-term exposure to ozone and sulfur dioxide increases the incidence of type 2 diabetes mellitus among aged 30 to 50 adult population. Environmental Research 2021; 194: 110624.

Lim CC, Hayes RB, Ahn J, Shao Y, Silverman DT, Jones RR, et al. Long-term exposure to ozone and cause-specific mortality risk in the United States. American journal of respiratory and critical care medicine 2019; 200: 1022-1031.

Lin S, Liu X, Le LH, Hwang S-A. Chronic exposure to ambient

ozone and asthma hospital admissions among children. Environmental Health Perspectives 2008; 116: 1725-1730.

Liu Q, Liu T, Chen Y, Xu J, Gao W, Zhang H, et al. Effects of aerosols on the surface ozone generation via a study of the interaction of ozone and its precursors during the summer in Shanghai, China. Science of The Total Environment 2019a; 675: 235-246.

Liu R, Ma Z, Liu Y, Shao Y, Zhao W, Bi J. Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. Environment international 2020; 142: 105823.

Liu Y, Pan J, Zhang H, Shi C, Li G, Peng Z, et al. Short-Term Exposure to Ambient Air Pollution and Asthma Mortality. Am J Respir Crit Care Med 2019b; 200: 24-32.

Lu GY, Wong DW. An adaptive inverse-distance weighting spatial interpolation technique. Computers & geosciences 2008; 34: 1044-1055.

Niu Y, Zhou Y, Chen R, Yin P, Meng X, Wang W, et al. Long-term exposure to ozone and cardiovascular mortality in China: a nationwide cohort study. The Lancet Planetary Health 2022; 6: e496-e503.

Parsons MA, Duerr R, Minster JB. Data citation and peer review.

Eos, Transactions American Geophysical Union 2010; 91: 297-298.

Requia WJ, Di Q, Silvern R, Kelly JT, Koutrakis P, Mickley LJ, et al. An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. Environmental science & technology 2020; 54: 11037-11047.

Rhee J, Dominici F, Zanobetti A, Schwartz J, Wang Y, Di Q, et al. Impact of long-term exposures to ambient PM2. 5 and ozone on ARDS risk for older adults in the United States. Chest 2019; 156: 71-79.

Rosenlund M, Forastiere F, Stafoggia M, Porta D, Perucci M, Ranzi A, et al. Comparison of regression models with land-use and emissions data to predict the spatial distribution of traffic-related air pollution in Rome. J Expo Sci Environ Epidemiol 2008; 18: 192-9.

So R, Chen J, Mehta AJ, Liu S, Strak M, Wolf K, et al. Long-term exposure to air pollution and liver cancer incidence in six European cohorts. International Journal of Cancer 2021; 149: 1887-1897.

Taha H. Modeling the impacts of large-scale albedo changes on ozone air quality in the South Coast Air Basin. Atmospheric

Environment 1997; 31: 1667-1676.

Tamiminia H, Salehi B, Mahdianpari M, Quackenbush L, Adeli S, Brisco B. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS Journal of Photogrammetry and Remote Sensing 2020; 164: 152-170.

Travaglio M, Yu Y, Popovic R, Selley L, Leal NS, Martins LM. Links between air pollution and COVID-19 in England. Environmental pollution 2021; 268: 115859.

Vienneau D, De Hoogh K, Bechle MJ, Beelen R, Van Donkelaar A, Martin RV, et al. Western European land use regression incorporating satellite-and ground-based measurements of NO2 and PM10. Environmental science & technology 2013; 47: 13555-13564.

Wang M, Sampson PD, Hu J, Kleeman M, Keller JP, Olives C, et al. Combining land-use regression and chemical transport modeling in a spatiotemporal geostatistical model for ozone and PM2. 5. Environmental science & technology 2016; 50: 5111-5118.

Wong DC, Pleim J, Mathur R, Binkowski F, Otte T, Gilliam R, et al. WRF-CMAQ two-way coupled system with aerosol feedback: software development and preliminary results. Geosci. Model Dev. 2012; 5: 299-312.

Wong P-Y, Hsu C-Y, Wu J-Y, Teo T-A, Huang J-W, Guo H-R, et al. Incorporating land-use regression into machine learning algorithms in estimating the spatial-temporal variation of carbon monoxide in Taiwan. Environmental Modelling & Software 2021; 139: 104996.

Yazdi MD, Wang Y, Di Q, Zanobetti A, Schwartz J. Long-term exposure to PM2. 5 and ozone and hospital admissions of Medicare participants in the Southeast USA. Environment international 2019; 130: 104879.

Yinusa AA, Ogunwale SA, Sobamowo MG, Usman MA. Application of multi-step differential transform method to the nonlinear behaviour of cloud droplets on gaseous atmospheric pollutant removal. Thermal Science and Engineering Progress 2019; 14: 100422.

Young MT, Bechle MJ, Sampson PD, Szpiro AA, Marshall JD, Sheppard L, et al. Satellite-based NO2 and model validation in a national prediction model based on universal kriging and land-use regression. Environmental science & technology 2016; 50: 3686-3694.

Zhan Y, Luo Y, Deng X, Zhang K, Zhang M, Grieneisen ML, et al. Satellite-based estimates of daily NO2 exposure in China using hybrid random forest and spatiotemporal kriging model.

Environmental science & technology 2018; 52: 4180-4189.

Zheng P, Chen Z, Liu Y, Song H, Wu C-H, Li B, et al. Association between coronavirus disease 2019 (COVID-19) and long-term exposure to air pollution: Evidence from the first epidemic wave in China. Environmental Pollution 2021; 276: 116682.

Zhu L, Xing H, Hou D. Analysis of carbon emissions from land cover change during 2000 to 2020 in Shandong Province, China. Scientific Reports 2022; 12: 1-12.

Supplementary materials

1. Data Source

1.1. Air pollutants monitoring station

The author used 480 monitoring stations for O_3 and NO_2, and 447 monitoring stations for CO from 2002 to 2020 among entire contiguous region of Korea. Not all monitors were in operation for the entire study period.

1.2. Predictors

The author extracted predictor variables from Google Earth Engine (GEE), Socioeconomic Data and Applications Center (SEDAC), and regional socioeconomic database. More detailed information about each data source is shown in Table S1.

1.2.1. Google Earth Engine

Google Earth Engine (GEE) has been proposed as a feasible solution for obtaining satellite-based data. GEE is a cloud-based platform for geospatial datasets that allows researchers to access the petabyte scale of free-use remote sensing data, including

various raw and ready-to-use datasets (Tamiminia et al., 2020). Studies using GEE will be able to suggest a new and impactful study protocol to develop prediction models for air pollution concentration

1.2.1.1. AOD measurements and satellite-based air quality

Moderate Resolution Imaging Spectroradiometer (MODIS) were used to retrieve remote sensing data. The author collected 0.55 μm and 0.47 μm aerosol optical depth (AOD) over land from Terra & Aqua MAIAC Land Aerosol Optical Depth, called MCD19A2 v006 data product. From this data, the author accessed daily AOD with 1 km × 1 km spatial resolution. Also, the author selected monthly averages of aerosol optical depth at 0.55 μm for both ocean and land, and monthly averages of corrected AOD (land) at 0.47 μm from MOD08_M3 v061 with 1.0° × 1.0° spatial resolution.

The author extracted total column ozone from Total Ozone Mapping Spectrometer (TOMS) and Ozone Monitoring Instrument (OMI) data with 1.0° × 1.0° spatial resolution. TOMS provide satellite-based continuous observations available during long period for catching global and regional trends in total ozone, with 1.0° × 1.25° spatial resolution. OMI continue the TOMS record for total ozone and other climate parameters since OMI provide total ozone

with an improved spatial resolution with 1.0° × 1.0° spatial resolution, compared to TOMS.

1.2.1.2. Meteorological data

The author obtained air temperature at 2m height, soil temperature, surface pressure, 10m u-component and v-component of wind, and leaf area index with high/low vegetation from 5th generation European Centre for Medium-Range Weather Forecasts atmospheric reanalysis (ERA5) and surface-based one (ERA5-Land), produced from European Centre for Medium-Range Weather Forecasts (ECMWF) climate reanalysis. ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables with 27.83 km spatial resolution. ERA5-Land reanalysis dataset is the evolution version of land variables with an improved spatial resolution with 11.132 km spatial resolution.

The author collected total water column density from the National Centers for Environmental Prediction (NCEP)/National Center for Atmospheric Research (NCAR). The NCEP provides multiple type of reanalysis dataset including weather and climate data. NCEP conducted a joint project with the National Center for Atmospheric Research (NCAR), called NCEP/NCAR Reanalysis project, for producing future and current atmospheric analyses. NCEP/NCAR

datasets is produced by every 6 hours with 2.5° spatial resolution. Data produced from NCEP/DOE Reanalysis II, an improved version of NCEP/NCAR, was used to get % of total cloud cover. It provided reanalysis atmospheric data with same temporal and spatial resolution as NCEP/NCAR.

The author extracted cirrus area fraction and liquid water cloud optical thickness from MOD08_M3 v061 with 1.0° × 1.0° spatial resolution. Also, clear day and night sky coverage were retrieved from MOD11A1 v061 with 1 km × 1 km spatial resolution.

Merged satellite-gauge precipitation estimate and accumulation-weighted probability of liquid precipitation phase were retrieved from Global Precipitation Measurement (GPM), which produced dataset every 3 hours, with 0.1° spatial resolution. GPM used the Integrated Multi-satellitE Retrievals for GPM (IMERG) which is the combined algorithm that provides rainfall estimates by using all instruments in the GPM.

1.2.1.3. Land-use terms

The author extracted land-use variables from MODIS, Copernicus Global Land Cover Layers and Global Land Cover Map.

The author collected Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) from MOD13A2

v061, and Leaf Area Index (LAI) and fraction of absorbed photosynthetically active radiation (FPAR) from MCD15A3H v061. Those two datasets produce remote sensing data with fine spatial resolution, 1 km and 500 m respectively. The author also collected Band 1–5, and Band 7 surface reflectance, and Band 6 surface temperature from Landsat 7 Surface Reflectance with 30 m × 30 m spatial resolution.

Land Cover Type 1 (Annual International Geosphere–Biosphere Programme classification), Land Cover Type 2 (Annual University of Maryland classification), Land Cover Type 3 (Annual LAI classification), Land Cover Type 4 (Annual BIOME–Biogeochemical Cycles classification), Land Cover Type 5 (Annual Plant Functional Types classification) were collected from MCD12Q1 V6 with 500 m × 500 m spatial resolution. Also, FAO–Land Cover Classification System 1 (LCCS1) land cover layer, FAO–LCCS2 land use layer, FAO–LCCS3 surface hydrology layer and their confidence (0~100%) were retrieved from MCD12Q1 V6.

The author also selected coverfraction of bare, crops, grass, moss, shrub, tree, snow, urban, permanent water and seasonal water, and forest type from Copernicus Global Land Cover Layers with 100 m × 100 m spatial resolution, and land cover map variable from Global Land Cover Map (GlobCover) with 300 m ×

300 m spatial resolution.

1.2.1.4. Other terms

The author collected daily black/white sky and Bidirectional Reflectance Distribution Function (BRDF) albedo from MCD43A3 v006 with 500 m × 500 m spatial resolution. Surface reflectance for band 1, band 2, band 3, band 4, band 5, band 6, and band 7 were extracted from MOD09A1 v061. Also, the author retrieved bands 31 and 32 emissivity values from MOD11A1 v061 with 1 km × 1 km spatial resolution.

1.2.2. SEDAC

1.2.2.1. PM2.5

The author obtained the variable for concentrations of ground-level fine particulate matter ($PM_{2.5}$) using the dataset from NASA Socioeconomic Data and Applications Center (SEDAC) which provides global annual $PM_{2.5}$ grids for the period 1988 to 2016. It combines Aerosol Optical Depth (AOD) retrievals from various satellite datasets including the MODIS, Multiangle Imaging SpectroRadiometer (MISR), and the Sea-Viewing Wide Field-of-View Sensor (SeaWiFS). The GEOS-Chem chemical transport model and Geographically Weighted Regression (GWR) were used

with global ground-based measurements to approach actual $PM_{2.5}$ levels per grid cell. Annual $PM_{2.5}$ concentrations were extracted with 0.02° spatial resolution.

1.2.3. Regional socioeconomic database

Regional socioeconomic database (database of community health outcomes and health determinants in the Republic of Korea) was constructed to understand the impact of community characteristics on health outcomes and gaps. District-level demographic, socioeconomic, and environmental variables have been collected annually by the Korea Disease Control and Prevention Agency, with about 2,000 variables since 2008.

1.2.3.1. Socioeconomic variables

The author obtained number of vehicle registrations per capita, % of road area compared to urban area, national basic livelihood security recipients per 1,000 people, population density, population growth rate, % of the population that is urban, ratio of total population to area, and % of persons with financial independence for using predictor variables in the study model.

1.2.3.2. Environmental pollution variables

Household waste discharge per resident, wastewater generation

per 1,000 people, wastewater discharge per 1,000 people, number of wastewater dischargers, organic material load generated per 1,000 people, and organic material load discharge per 1,000 people were selected to predictor variables in this study.

1.2.3.3. Green area variables

The author consider variables of % of the area that is forest, forest area ratio to living sphere, total urban forest area ratio, park area per capita, total urban forest area per capita, number of parks per 1,000 people, urban park area per 1,000 people, forest area of living sphere per capita, ratio of roadside green space to urban area, and ratio of riverside green area to urban area were retrieved from regional socioeconomic database.

1.2.4. Additional predictor variables

The author calculated spatially weighted average of each air pollution measurements using inverse distance weighting method, and added them for predictor variables in the study model. Also, year, month and seasonal terms are included in the model as categorical variables to adjust for yearly, monthly and seasonal variation of each air pollutant. The author considered spring as March ~ May, summer as June ~ August, autumn as September ~

November, and winter as December ~ February.

During COVID-19 pandemic period, emission of air pollutants and precursors changed with decreases in personal transportation, but increases in emissions from delivery vehicles.. While the effect was not uniform across different locations, many areas experienced improved air pollution compared to pre-COVID-19 period (~2019) (Ju et al., 2021). Thus, the author added a binary variable for the year 2020 to reflect conditions of COVID-19. Also, the author considered a monthly term to reflect the months with the 1^{st} to 4^{th} highest concentrations for each air pollutant as a binary indicator. Lastly, the author added binary indicators of metropolitan city areas including Seoul, Incheon, Daejeon, Gwangju, Daegu, Busan, and Ulsan because they showed different spatiotemporal distribution for each air pollutant.

2. Method

2.1. Missing value processing

When extracting predictor variables from data sources, missing values occurred commonly for various reasons. (1) Remote sensing data may not be collected for a particular period or in a particular

area due to weather issues or equipment malfunction. In author's case, 0.47 μm aerosol optical depth obtained from MCD19A2 v006 had 1.61 % missing values. (2) Some variables were investigated for a specific time period. For example, % of forest area and land cover variables were obtained every 2 years or for specific periods. If the author ignores those missing values and remove the rows having missing values, the author's model may not fit properly into the entire study area and period. Thus, the author decided to estimate in missing values.

There are multiple ways to estimate missing values, using averages of nearby location or nearby time points, and global averages of each variable. The author referred to previous studies on this topic (Di et al., 2019a; Di et al., 2019b; Requia et al., 2020), and used random forest for estimating missing values since this model could fit linear and nonlinear relationship smoothly between predictor variables and outcome. The author selected predictor variables that have no missing values for the whole grid and time period, and regarded them as "X" for filling in missing values. For example, 0.47 μm aerosol optical depth contained 1.61 % missing values as mentioned above. The author regarded them as "Y", and consider non-missing "Y" as "Y_train". Afterwards, the author trained the random forest model by subtracting "X"

present in the same grid and time as "Y_train". The author calculated out-of-bag error for each column, and filled in missing values by using "X" present in the same grid and time as the "Y" missing values in the trained random forest model. This process was conducted for each predictor variable which were extracted from Google Earth Engine and SEDAC except for land cover variables.

Land cover datasets were available for specific time periods. The author adopted a linear interpolation method for land cover variables. For a regional socioeconomic database, the same method was adopted for each district since variables in this source were investigated yearly or every "n" years. In the case of Global Land Cover Map dataset, it was available only in 2009. That means only one value was available for each grid. Thus, the author regarded this variable as unchanged over time during estimating period.

2.2. Machine Learning Modeling

The author used random forest, light gradient boosting and neural network algorithm for estimating O_3, NO_2 and CO concentrations. Random forest is a well-known machine learning algorithm with low

variance of predictions, but its training speed usually can take a long time to fit. To overcome this problem, the author used "ranger" package, which improved training speed of random forest. For gradient boosting, the author adopted a light gradient boosting algorithm. Light gradient boosting is improved gradient boosting in training speed and accuracy with tree construction of parallelization. Many studies reported that light gradient boosting algorithms showed better performance with less training time than typical gradient boosting and extreme gradient boosting algorithms (Ke et al., 2017; Wei et al., 2021; Zhang et al., 2019). Moreover, the author fitted deep neural network by tuning parameters elaborately to capture spatiotemporal distribution for each air pollutant properly. Although there are various ways to fit neural network algorithms well and quickly, most studies estimating air pollutants have only adjusted well-known parameters, for example, learning rate, number of layers, and epochs (Di et al., 2019a; Requia et al., 2020). Especially, neural network model usually demands higher learning rate at the beginning steps, and through the learning process it should be tuned in more detail in the last steps. Therefore, the author added a learning speed scheduler to consider these characteristics, which reduced model training time. The "ranger" package with R, "lightgbm" package and "Pytorch"

framework with Python were used to model the random forest, light gradient boosting and neural network algorithm, respectively.

2.3. Ensemble model (GAM)

To aggregate the estimations of three machine learning models, the author fitted two types of generalized additive model by considering geographical location and estimations of each model.

$$(1)\quad \hat{Y} = f_1(Location_i) + f_2\left(\hat{Y}_{rf_{ij}}\right) + f_3\left(\hat{Y}_{gb_{ij}}\right) + f_4\left(\hat{Y}_{nn_{ij}}\right)$$

$$(2)\quad \hat{Y} = f_5\left(Location_i, \hat{Y}_{rf_{ij}}\right) + f_6\left(Location_i, \hat{Y}_{gb_{ij}}\right) + f_7\left(Location_i, \hat{Y}_{nn_{ij}}\right)$$

$Location_i$ is geographical information (i.e., longitude and latitude) of location i ; $\hat{Y}_{rf_{ij}}$, $\hat{Y}_{gb_{ij}}$, and $\hat{Y}_{nn_{ij}}$ are predicted air pollution concentrations from the random forest, gradient boosting, and neural network respectively at location i on time j; f_1 denote a thin plate spline function for location i; f_2~ f_4 denote linear functions; f_5 ~ f_7 denote the thin plate spline for interactions between location i and predicted air pollution concentrations at location i on time j from each model.

Formula (1) focused on linear relationship between three model estimation and monitored air pollutants by considering location. The author also tried to fit formula (2) model which covered interaction between geographical characteristics and prediction estimates, and

compared their performances for each air pollutant (Table S4). The author found that formula (1) showed better performance in O_3 and NO_2 compared to formula (2). In CO, the performance of formula (1) and (2) was almost the same.

Table S1. Detailed information about data sources

Cloud platform	Data source	Predictor variables	Collection Period	Spatiotemporal Resolution
Google Earth Engine	ERA5 Monthly Aggregates	mean_2m_air_temperature	2002~2020	27.83 km, Monthly
		minimum_2m_air_temperature		
		maximum_2m_air_temperature		
		dewpoint_2m_temperature		
		total_precipitation		
		surface_pressure		
		mean_sea_level_pressure		
		u_component_of_wind_10m		
	ERA5-Land Monthly Averaged	dewpoint_temperature_2m	2002~2020	11.132 km, Monthly
		leaf_area_index_high_vegetation		
		leaf_area_index_low_vegetation		
		soil_temperature_level_1		
		soil_temperature_level_2		
		soil_temperature_level_3		
		soil_temperature_level_4		
		temperature_2m		
		total_precipitation		
		u_component_of_wind_10m		
		v_component_of_wind_10m		
	MCD43A3.006 MODIS Albedo	Black sky Albedo	2002~2020	500 m, Daily
		White sky Albedo		
		BRDF Albedo		
	MOD09A1.006 MODIS Terra Surface Reflectance	sur_refl_b01	2002~2020	500 m, 8-day
		sur_refl_b02		
		sur_refl_b03		
		sur_refl_b04		
		sur_refl_b05		
		sur_refl_b06		
		sur_refl_b07		
	MOD11A1.006 Terra Land Surface Temperature and Emissivity	Emis_31	2002~2020	1 km, Daily
		Emis_32		
		Clear_day_cov		
		Clear_night_cov		
	MCD15A3H.006 MODIS Leaf Area Index/FPAR	Fpar	2002.07~2020	500 m, 4-day
		Lai		
	MOD13A2.006 Terra Vegetation Indices	NDVI	2002~2020	1 km, 16-day
		EVI		
	MOD08_M3.061 Terra	Aerosol_Optical_Depth_Land_Ocean_Mean_Mean	2002~2020	111.32 km, Monthly

	Atmosphere Monthly Global Product	Aerosol_Optical_Depth_Land_QA_Mean_Mean_470		
		Cirrus_Fraction_SWIR_FMean		
		Cloud_Optical_Thickness_Liquid_Log_Mean_Mean		
		Cloud_Optical_Thickness_Liquid_Mean_Uncertainty		
	MCD19A2.006: Terra & Aqua MAIAC Land Aerosol Optical Depth	Optical_Depth_047	2002~2020	1 km, Daily
		Optical_Depth_055		
	NCEP/NCAR Reanalysis Data, Water Vapor	pr_wtr	2002~2020	278.3 km, 6-hour
	NCEP-DOE Reanalysis 2 (Gaussian Grid), Total Cloud Coverage	tcdc	2002~2020	278.3 km, Monthly
	TOMS and OMI Merged Ozone Data	ozone	2002~2020	111 km, Daily
	GPM: Monthly Global Precipitation Measurement (GPM) v6	Precipitation	2002~2020	11.132 km, Monthly
		probabilityLiquidPrecipitation		
	USGS Landsat 7 Level 2, Collection 2, Tier 1	SR_B1	2002~2020	30 m, 16-day
		SR_B2		
		SR_B3		
		SR_B4		
		SR_B5		
		ST_B6		
		SR_B7		
	Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3	forest_type	2015~2019	100 m, Yearly
		bare_coverfraction		
		crops_coverfraction		
		grass_coverfraction		
		shrub_coverfraction		
		tree_coverfraction		
		urban_coverfraction		
		water_permanent_coverfraction		
		water_seasonal_coverfraction		
	GlobCover: Global Land Cover Map	landcover	2009	300 m
	MCD12Q1.006 MODIS Land Cover Type Yearly Global 500m	lc_prop1_categorical	2002~2019	500 m, Yearly
		lc_prop1_assessment		
		lc_prop2_categorical		
		lc_prop2_assessment		
		lc_prop3_categorical		

		lc_prop3_assessment		
		lc_type1_categorical		
		lc_type2_categorical		
		lc_type3_categorical		
		lc_type4_categorical		
		lc_type5_categorical		
	Link : https://developers.google.com/earth-engine/datasets/			
SEDAC	Global (GL) Annual PM2.5 Grids from MODIS, MISR and SeaWiFS Aerosol Optical Depth (AOD), v4.03	PM2.5 (AOD)	2002~2019	0.02 degree, Yearly
	Link : https://sedac.ciesin.columbia.edu/data/set/sdei-global-annual-gwr-pm2-5-modis-misr-seawifs-aod-v4-gl-03			
RSD	Regional socioeconomic database	number of registered vehicles per person	2008~2019	District-level, Yearly
		% of road area compared to urban area	2008~2019	District-level, Yearly
		national basic livelihood security recipients per 1,000 people	2008~2019	District-level, Yearly
		population density	2008~2019	District-level, Yearly
		population growth rate	2008~2019	District-level, Yearly
		% of urban population	2008~2019	District-level, Yearly
		ratio of total population to area	2008~2019	District-level, Yearly
		% of persons with financial independence	2008~2019	District-level, Yearly
		household waste discharge per resident	2008~2019	District-level, Yearly
		wastewater generation per 1,000 people	2018~2015, 2017, 2018	District-level, Yearly
		wastewater discharge per 1,000 people	2018~2015, 2017, 2018	District-level, Yearly
		number of wastewater dischargers	2018~2015, 2017, 2018	District-level, Yearly
		organic material load generated per 1,000 people	2018~2015, 2017, 2018	District-level, Yearly
		organic material load discharge per 1,000 people	2018~2015, 2017, 2018	District-level, Yearly
		% of the area that is forest	2009, 2011, 2013, 2015, 2017, 2019	District-level, every 2 Years
		forest area ratio to living sphere	2009, 2011, 2013, 2015, 2017, 2019	District-level, every 2 Years
		total urban forest area ratio	2009, 2011, 2013, 2015, 2017, 2019	District-level, every 2 Years
		park area per capita	2008~2019	District-level, every

				2 Years
		total urban forest area per capita	2009, 2011, 2013, 2015, 2017, 2019	District-level, every 2 Years
		number of parks per 1,000 people	2008~2019	District-level, Yearly
		urban park area per 1,000 people	2008~2017	District-level, Yearly
		urban forest area of living sphere per capita	2009, 2011, 2013, 2015, 2017, 2019	District-level, every 2 Years
		ratio of roadside green space to urban area	2009, 2011, 2013, 2015, 2017, 2019	District-level, every 2 Years
		ratio of riverside green area to urban area	2009, 2011, 2013, 2015, 2017, 2019	District-level, every 2 Years
	Link : https://chs.kdca.go.kr/chs/recsRoom/dataBaseMain.do			
Monitoring Station	Inverse Distance Weighting (IDW)	IDW of O_3	2002~2020	By station, Hourly
		IDW of NO_2		
		IDW of CO		
	Link : https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123			
Others		Dummy variables for each year		
		Dummy variables for each month		
		Dummy variables for each season		
		Binary indicator whether metropolitan city or not		
		Binary indicator whether COVID-19 year or not		
		Binary indicator whether the fourth highest month for each air pollutant or not		

Table S2. Variables sorted by % missing values

Predictor variables	% of missing values
forest_type	73.81
bare_coverfraction	73.79
crops_coverfraction	73.79
grass_coverfraction	73.79
shrub_coverfraction	73.79
tree_coverfraction	73.79
urban_coverfraction	73.79
water_permanent_coverfraction	73.79
water_seasonal_coverfraction	73.79
total urban forest area ratio	73.65
forest area ratio of living sphere	73.65
ratio of riverside green area to urban area	73.02
ratio of roadside green space to urban area	68.55
% of the area that is forest	68.42
total urban forest area per capita	68.39
urban forest area of living sphere per capita	68.39
organic material load discharge per 1,000 people	52.83
urban park area per 1,000 people	52.59
wastewater generation per 1,000 people	47.6
wastewater discharge per 1,000 people	47.6
organic material load generated per 1,000 people	47.57
number of wastewater dischargers	42.25
household waste discharge per resident	42.11
number of parks per 1,000 people	42.1
national basic livelihood security recipients per 1,000	38.4
% of persons with financial independence	37.06
population growth rate	36.94
number of vehicle registrations per capita	36.87
% of the population that is urban	36.84
population density	36.84
park area per capita	36.84
% of road area compared to urban area	36.84
ratio of total population to area	36.84
$PM_{2.5}$	7.64
st_b6	7.23
sr_b1	6.15
sr_b2	6.15
sr_b3	6.15
sr_b4	6.15
sr_b5	6.15
sr_b7	6.15
dewpoint_temperature_2m_land_0020	4.29
leaf_area_index_high_vegetation_land_0020	4.29
leaf_area_index_low_vegetation_land_0020	4.29
soil_temperature_level_1_land	4.29
soil_temperature_level_2_land	4.29
soil_temperature_level_3_land	4.29

soil_temperature_level_4_land	4.29
temperature_2m_land_0020	4.29
total_precipitation_land_0020	4.29
u_component_of_wind_10m_land_0020	4.29
v_component_of_wind_10m_land_0020	4.29
fpar	4.11
lai	4.11
dewpoint_2m_temperature	3.07
maximum_2m_air_temperature	3.07
mean_2m_air_temperature	3.07
mean_sea_level_pressure	3.07
minimum_2m_air_temperature	3.07
surface_pressure	3.07
total_precipitation	3.07
u_component_of_wind_10m	3.07
v_component_of_wind_10m	3.07
aerosol_optical_depth_land_qa_mean_mean_470	1.88
brdf_albedo_band_mandatory_quality_band6_0020	1.82
brdf_albedo_band_mandatory_quality_shortwave_0020	1.8
brdf_albedo_band_mandatory_quality_nir_0020	1.77
optical_depth_055_0020	1.66
optical_depth_047_0020	1.61
aerosol_optical_depth_land_ocean_mean_mean	1.54
albedo_bsa_band6	1.14
albedo_wsa_band6	1.14
albedo_bsa_shortwave	1.12
albedo_wsa_shortwave	1.12
albedo_bsa_nir	1.11
albedo_wsa_nir	1.11
brdf_albedo_band_mandatory_quality_band5_0020	1.04
brdf_albedo_band_mandatory_quality_band3_0020	1.03
brdf_albedo_band_mandatory_quality_vis_0020	1.03
brdf_albedo_band_mandatory_quality_band1_0020	1.01
brdf_albedo_band_mandatory_quality_band2_0020	1.01
brdf_albedo_band_mandatory_quality_band4_0020	1.01
brdf_albedo_band_mandatory_quality_band7_0020	1.01
clear_day_cov	0.77
clear_night_cov	0.76
albedo_bsa_band3	0.68
albedo_bsa_band5	0.68
albedo_bsa_vis	0.68
albedo_wsa_band3	0.68
albedo_wsa_band5	0.68
albedo_wsa_vis	0.68
albedo_bsa_band1	0.67
albedo_bsa_band2	0.67
albedo_bsa_band4	0.67
albedo_bsa_band7	0.67
albedo_wsa_band1	0.67
albedo_wsa_band2	0.67

albedo_wsa_band4	0.67
albedo_wsa_band7	0.67
emis_31	0.66
emis_32	0.66
cirrus_fraction_swir_f_mean	0.44
cloud_optical_thickness_liquid_log_mean_mean	0.44
cloud_optical_thickness_liquid_mean_uncertainty	0.44
ndvi	0.33
evi	0.32
lc_prop2_assessment	0.07
lc_prop1_assessment	0.06
lc_prop3_assessment	0.06
landcover_noyear_categorical	0
lc_prop1_categorical	0
lc_prop2_categorical	0
lc_prop3_categorical	0
lc_type1_categorical	0
lc_type2_categorical	0
lc_type3_categorical	0
lc_type4_categorical	0
lc_type5_categorical	0
ozone	0
pr_wtr	0
precipitation	0
probability_liquid_precipitation	0
sur_refl_b01	0
sur_refl_b02	0
sur_refl_b03	0
sur_refl_b04	0
sur_refl_b05	0
sur_refl_b06	0
sur_refl_b07	0
tcdc	0

Table S3. Results of parameter grid search using 10-fold cross-validation for O_3, NO_2 and CO

O_3					
Random Forest		**Light Gradient Boosting**		**Neural Network**	
Parameter Name	Parameter Value	Parameter Name	Parameter Value	Parameter Name	Parameter Value
Number of trees	3,000	Number of trees	500	Epochs	500
Maximum tree depth	30	Maximum tree depth	5	Hidden layer & number of hidden units for each layer	2, 16
Minimum node size	6	Minimal node size	20	Activation function	Rectifier (ReLU)
Sample rate	0.9	Learning Rate	0.01	Optimizer & learning rate	Adam, 0.01
		Column sample rate	0.7	Scheduler & decay rate	Exponential , 0.992
		L1, L2 regularization	0.1, 1	Dropout rate	0
NO_2					
Random Forest		**Light Gradient Boosting**		**Neural Network**	
Parameter Name	Parameter Value	Parameter Name	Parameter Value	Parameter Name	Parameter Value
Number of trees	1,500	Number of trees	500	Epochs	500
Maximum tree depth	12	Maximum tree depth	7	Hidden layer & number of hidden units for each layer	2, 64

Minimum node size	8	Minimum node size	20	Activation function	Rectifier (ReLU)
Sample rate	0.9	Learning Rate	0.01	Optimizer & learning rate	Adam, 0.008
		Column sample rate	0.55	Scheduler & decay rate	Exponential , 0.96
		L1, L2 regularization	0.2, 4	Dropout rate	0.2
CO					
Random Forest		Light Gradient Boosting		Neural Network	
Parameter Name	Parameter Value	Parameter Name	Parameter Value	Parameter Name	Parameter Value
Number of trees	3,000	Number of trees	400	Epochs	500
Maximum tree depth	24	Maximum tree depth	7	Hidden layer & number of hidden units for each layer	3, 16
Minimum node size	6	Minimal node size	20	Activation function	Rectifier (ReLU)
Sample rate	0.8	Learning Rate	0.01	Optimizer & learning rate	Adam, 0.1
		Column sample rate	0.55	Scheduler & decay rate	Exponential , 0.985
		L1, L2 regularization	0.3, 1	Dropout rate	0.2

Table S4. Yearly ensemble (GAM) performance for O_3, NO_2, and CO

		R^2		RMSE	
		GAM 1	GAM 2	GAM1	GAM2
O_3 (ppb)	year				
	2002~2005	0.725	0.694	6.469	6.772
	2006~2008	0.808	0.795	5.692	5.903
	2009~2011	0.842	0.828	4.97	5.162
	2012~2014	0.854	0.846	5.175	5.309
	2015~2017	0.871	0.855	5.195	5.513
	2018~2020	0.843	0.831	5.205	5.455
	overall	0.84	0.825	5.463	5.706
		R^2		RMSE	
		GAM 1	GAM 2	GAM1	GAM2
NO_2 (ppb)	year				
	2002~2005	0.511	0.268	7.336	11.549
	2006~2008	0.716	0.67	5.335	5.752
	2009~2011	0.738	0.682	5.024	5.532
	2012~2014	0.702	0.65	5.486	5.942
	2015~2017	0.706	0.653	5.056	5.436
	2018~2020	0.84	0.787	3.693	4.087
	overall	0.721	0.603	5.236	6.456
		R^2		RMSE	
		GAM 1	GAM 2	GAM1	GAM2
CO (ppm)	year				
	2002~2005	0.326	0.328	0.227	0.227
	2006~2008	0.476	0.474	0.193	0.193
	2009~2011	0.51	0.514	0.153	0.153
	2012~2014	0.49	0.491	0.138	0.138
	2015~2017	0.402	0.405	0.123	0.123
	2018~2020	0.474	0.476	0.096	0.097
	overall	0.488	0.488	0.154	0.154

* GAM : Generalized Additive Model which aggregating estimations of Random Forest, Light Gradient Boosting, and Neural Network.

* GAM 1 : Formula (1) in section 2.3

* GAM 2 : Formula (2) in section 2.3

* Performance for O_3 and NO_2 were calculated based on ppb, and ppm for CO.

Table S5. Model performances for O_3, NO_2, and CO by season and urbanity

		R^2					RMSE				
	Season	RF	GB	NN	GAM	SA	RF	GB	NN	GAM	SA
O_3 (ppb)	Spring	0.651	0.655	0.658	0.661	0.665†	6.697	6.632	6.606	6.585	6.546†
	Summer	0.742	0.74	0.74	0.743	0.747†	6.068	6.077	6.08	6.045	5.994†
	Autumn	0.744	0.739	0.724	0.746†	0.745	4.753	4.791	4.938	4.73†	4.739
	Winter	0.73	0.736	0.713	0.738	0.74†	4.222	4.138	4.306	4.136	4.113†
	Area										
	Urban	0.84	0.84	0.84	0.842	0.845†	5.468	5.449	5.449	5.416	5.372†
	Rural	0.762†	0.755	0.715	0.761	0.753	6.197	6.114	6.792	6.03†	6.176
		R^2					RMSE				
	Season	RF	GB	NN	GAM	SA	RF	GB	NN	GAM	SA
NO_2 (ppb)	Spring	0.688	0.7	0.661	0.674	0.702†	5.291	5.144	5.461	5.386	5.138†
	Summer	0.662	0.67	0.677	0.643	0.681†	4.332	4.251	4.211	4.594	4.194†
	Autumn	0.753	0.766	0.76	0.736	0.768†	4.827	4.652	4.751	5.039	4.656†
	Winter	0.663	0.684	0.675	0.656	0.686†	5.608	5.39†	5.512	5.853	5.401
	Area										
	Urban	0.738	0.742	0.727	0.718	0.747†	4.902	4.839	4.987	5.135	4.801†
	Rural	0.606	0.723	0.752†	0.678	0.716	6.138	5.217†	5.221	6.098	5.447
		R^2					RMSE				
CO (ppm)	Season	RF	GB	NN	GAM	SA	RF	GB	NN	GAM	SA
	Spring	0.311†	0.286	0.241	0.292	0.307	0.134	0.135	0.139	0.135	0.133†
	Summer	0.184†	0.158	0.117	0.165	0.177	0.124†	0.126	0.133	0.125	0.126
	Autumn	0.471†	0.446	0.415	0.451	0.471	0.138†	0.14	0.146	0.14	0.139

Winter	0.392†	0.384	0.312	0.362	0.392	0.201†	0.201	0.222	0.204	0.203
Area										
Urban	0.526	0.51	0.478	0.508	0.53†	0.53	0.152†	0.154	0.154	0.154
Rural	0.126	0.135†	0.074	0.112	0.118	0.153	0.15†	0.164	0.155	0.153

* RF : Random Forest, GB : light Gradient Boosting, NN : Neural Network, GAM : Generalized Additive Model (Formula (1) in Section 2.3), SA : Simple average estimation of RF, GB, and NN

* Performance for O_3 and NO_2 were calculated based on ppb, and ppm for CO.

† A model that performs better than other models during the period

Table S6. Number of monitoring stations by year for O_3, NO_2 and CO in urban and rural areas

Year	O_3			NO_2			CO		
	Urban	Rural	Overall	Urban	Rural	Overall	Urban	Rural	Overall
2002	137	8	145	135	8	143	133	7	140
2003	157	8	165	156	8	164	153	8	161
2004	172	7	179	169	8	177	167	7	174
2005	185	8	193	181	8	189	184	8	192
2006	194	9	203	192	8	200	190	9	199
2007	207	9	216	207	9	216	206	9	215
2008	217	10	227	217	10	227	217	10	227
2009	220	10	230	220	10	230	220	10	230
2010	222	10	232	222	10	232	218	10	228
2011	223	14	237	223	14	237	221	14	235
2012	231	15	246	229	15	244	227	15	242
2013	236	16	252	237	16	253	232	16	248
2014	237	16	253	238	16	254	233	16	249
2015	238	17	255	238	17	255	233	16	249
2016	242	18	260	242	18	260	236	17	253
2017	249	27	276	249	27	276	239	19	258
2018	280	45	325	279	45	324	276	45	321
2019	309	86	395	309	86	395	288	73	361
2020	357	106	463	357	106	463	330	100	430

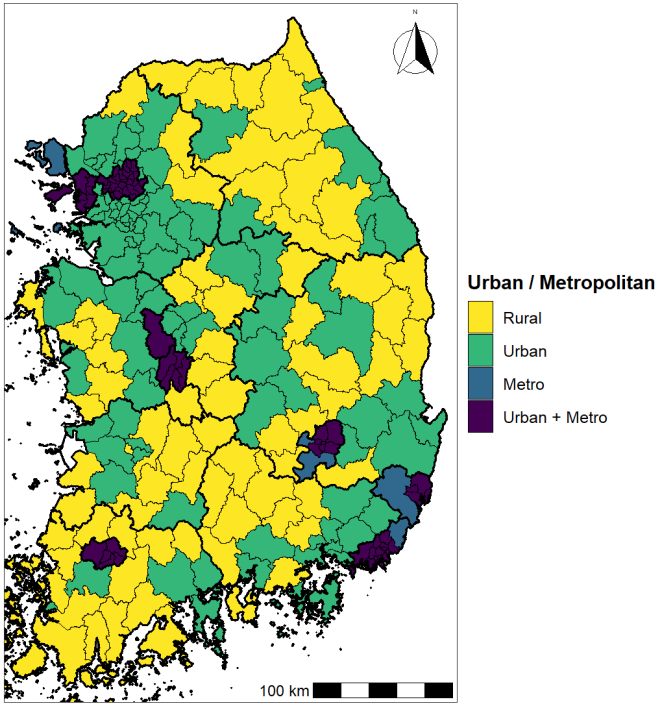

Fig. S1. Urban/Rural and Metropolitan (Metro) area for entire contiguous regions of South Korea

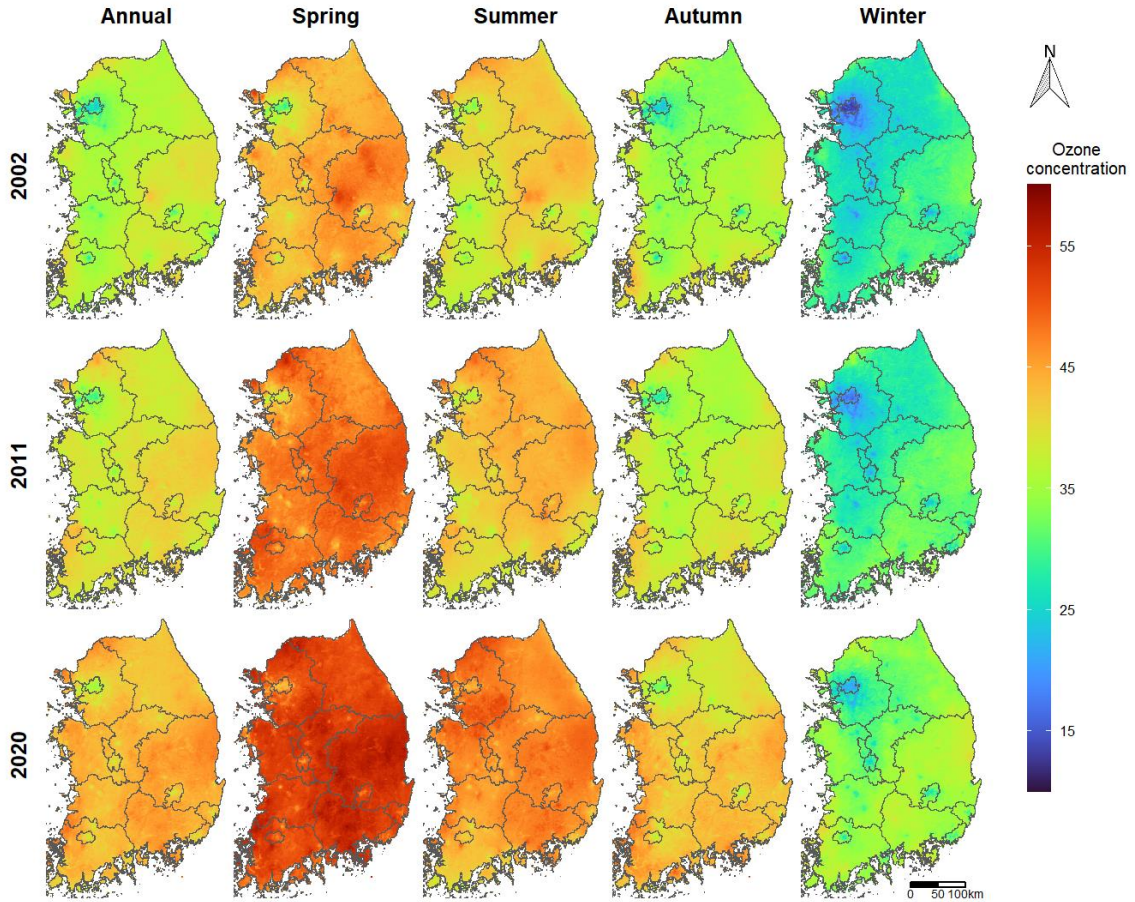

Fig. S2. Distribution maps of predicted O_3 (ppb) by year and season for contiguous South Korea

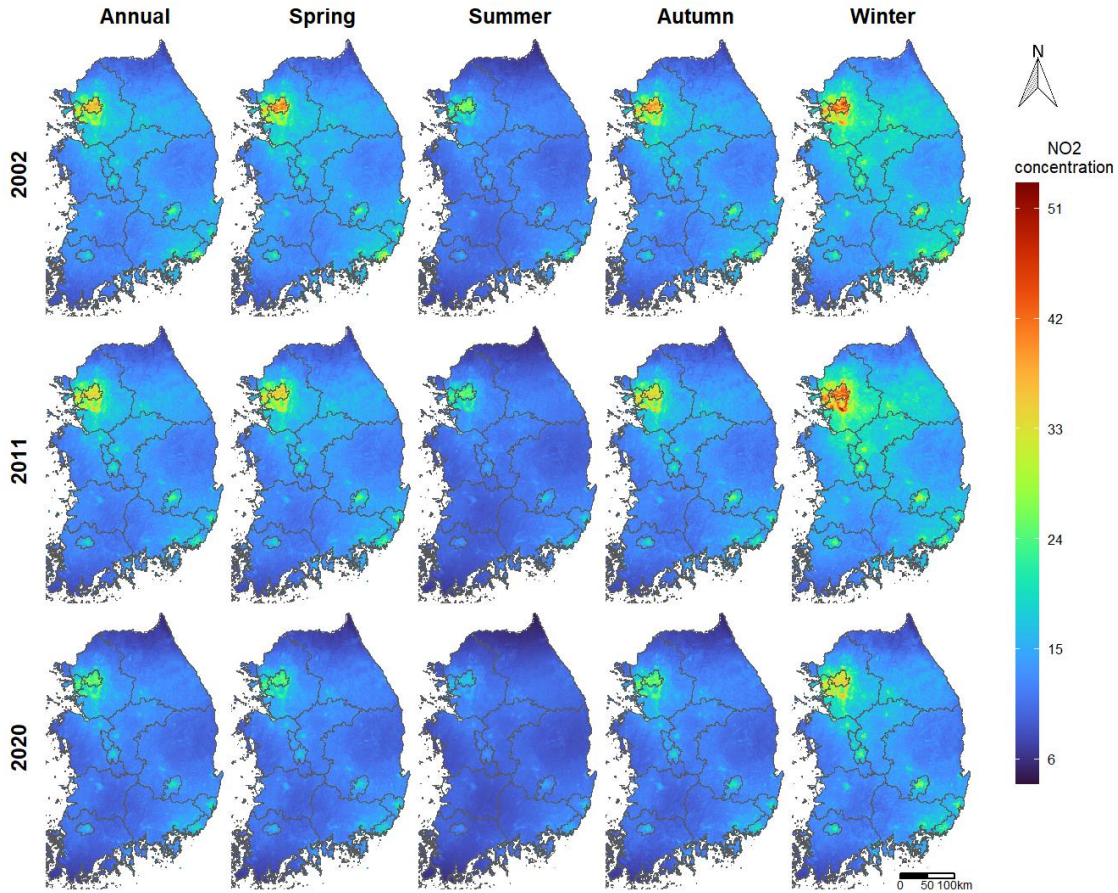

Fig. S3. Distribution maps of predicted NO_2 (ppb) by year and season for contiguous South Korea

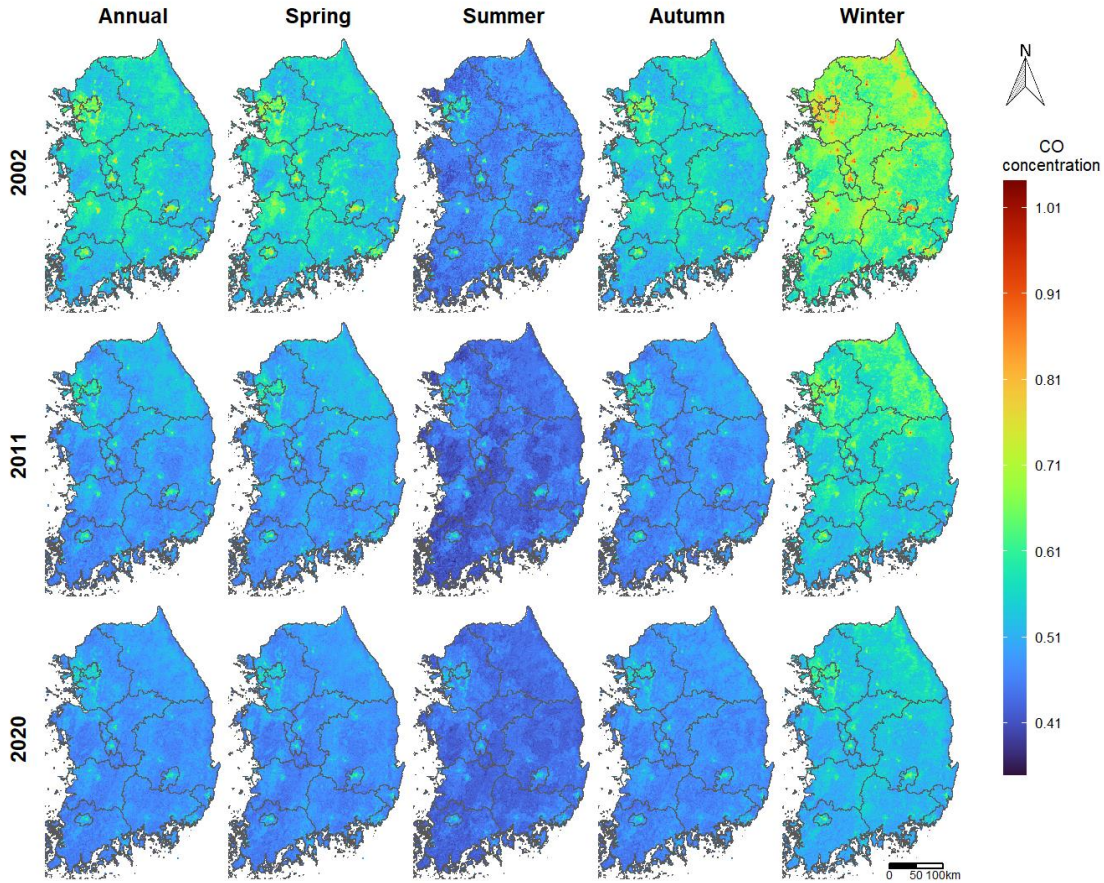

Fig. S4. Distribution maps of predicted CO (ppm) by year and season for contiguous South Korea

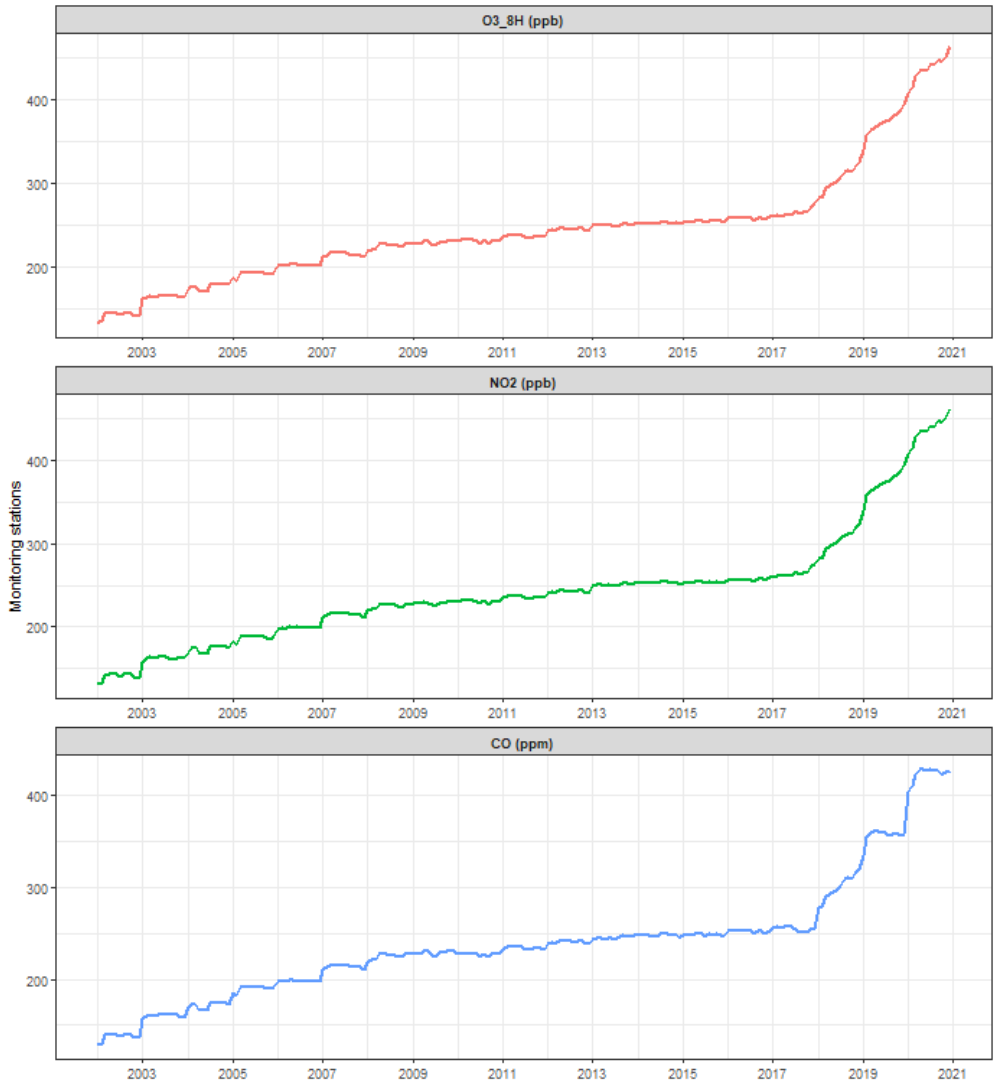

Fig. S5. Monthly fluctuations in the number of monitoring stations for O_3, NO_2, and CO between 2002 and 2020

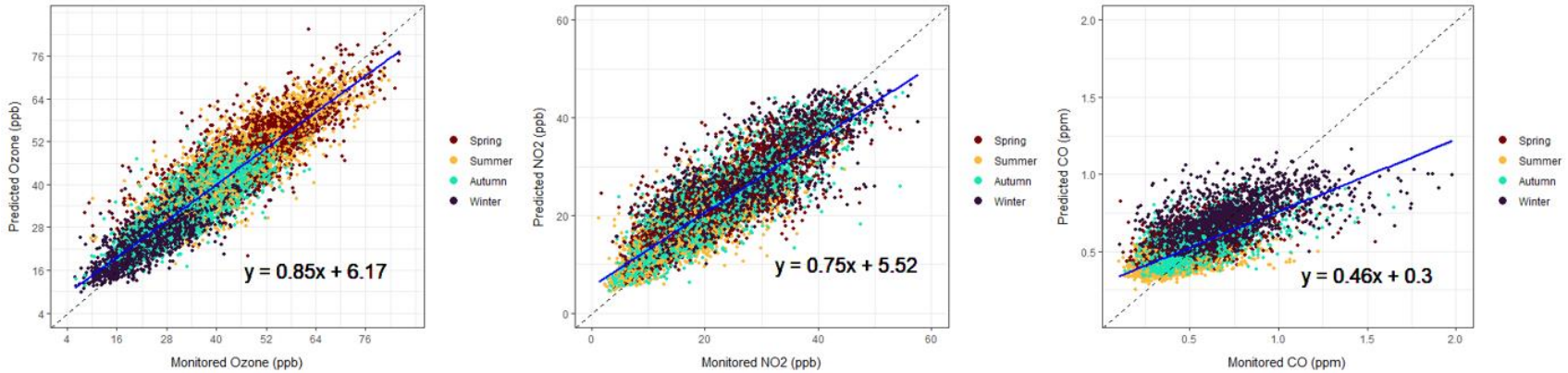

Fig. S6. Density scatter plot for monthly averages of the monitored and predicted concentrations of O_3, NO_2, and CO with seasonal discrimination

Bibliography (Supplement)

Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, et al. Assessing NO2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. Environmental science & technology 2019a; 54: 1372-1384.

Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, et al. An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. Environment international 2019b; 130: 104909.

Ju MJ, Oh J, Choi Y-H. Changes in air pollution levels after COVID-19 outbreak in Korea. Science of the Total Environment 2021; 750: 141521.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 2017; 30.

Requia WJ, Di Q, Silvern R, Kelly JT, Koutrakis P, Mickley LJ, et al. An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. Environmental science & technology 2020; 54: 11037-11047.

Tamiminia H, Salehi B, Mahdianpari M, Quackenbush L, Adeli S, Brisco B. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS Journal of Photogrammetry and Remote Sensing 2020; 164: 152-170.

Wei J, Li Z, Pinker RT, Wang J, Sun L, Xue W, et al. Himawari-8-derived diurnal variations in ground-level PM 2.5 pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). Atmospheric Chemistry and Physics 2021; 21: 7863-7880.

Zhang Y, Wang Y, Gao M, Ma Q, Zhao J, Zhang R, et al. A predictive data feature exploration-based air quality prediction approach. IEEE Access 2019; 7: 30732-30743.

국 문 초 록

머신러닝 모델을 사용한 2002~2020년 한국의 O_3 , NO_2 , CO 농도의 고해상도 추정

서울대학교 보건대학원
보건학과 보건통계학전공
권 도 훈

연구배경 : 오존(O_3), 이산화질소(NO_2), 일산화탄소(CO)에 장기간 노출되면 각종 질병을 유발하고 사망률을 높이는 것으로 알려져 있다. 그렇기에, 고해상도로 지표면 O_3, NO_2, CO 농도를 추정하는 것은 이러한 대기오염물질과 관련된 건강 영향을 평가하는 데 매우 중요하다. 하지만, 장기간에 걸쳐 고해상도로 가스상 대기오염물질(O_3, NO_2, CO)를 추정한 연구는 국내에서 아직 진행된 바가 없다. 따라서, 본 연구는 2002년부터 2020년까지 대한민국 전역에서 1km × 1km의 공간해상도로 월별 O_3(일평균 8시간 최대치), NO_2, CO를 머신러닝 기반 모델 및 그들의 앙상블 모형을 통해 예측하고자 한다.

연구방법 : 3가지 머신러닝 모델(랜덤 포레스트, 라이트 그래디언트 부스팅, 신경망)의 최적의 파라미터를 찾기 위해 모니터링 스테이션의 약 80%를 훈련 데이터로 사용하였고, 10-fold 교차검증을 통해 훈련 데이터 내에서 훈련/평가 단계를 거쳤으며, 나머지 모니터링 스테이션의 20%를 모델 평가에 사용하였다. 여기에 추가로 머신러닝 모델 간의 예측 변동을 통합하기 위해 앙상블 모델을 적용했다. 데이터에는 위성 기반 원격 감지 데이터, 역거리 가중치 기반 대기오염농도, 토지 이용 변수, 기상 재분석 자료, 다양한 데이터베이스에서 수집된 지역 사회경제적 변수 등이 포함되었다.

연구결과 : O_3의 경우, 전체 연구 기간 동안 앙상블 모델의 R^2가 0.841을 기록했으며, 도시 지역이 농촌 지역(R^2 = 0.762)보다 우수한 예측 성능(R^2 = 0.845)을 보였다. NO_2의 경우, 앙상블(평균) 모델의 R^2가 0.756으로 가장 높았으며, 계절로 보면 가을에 예측 성능이 가장 높았다(R^2 = 0.768). CO의 경우, R^2가 0.506 을 기록했다. 본 연구는 O_3 및 NO_2 에서 R^2 > 0.75 으로 높은 예측력의 고해상도 월평균 추정치를 제공한다.

결론 : 본 연구에서 얻어진 대기오염 추정 결과는 인구 특성과 관련된 가스상 대기오염물질의 공간 패턴을 분석하거나, 위치 기반 건강 정보와 행정구역 단위 건강 데이터와 엮여서 장기간 대기오염 노출의 건강 영향을 평가하는 연구에 사용될 수 있을 것으로 기대된다.

주 요 어 : 가스상 대기오염물질, 장기간 노출 평가, 고해상도(공간), 머신러닝 모델, 앙상블 모델

학 번 : 2021-24226