Estimation of Ground SO_2 Concentration through The Synergistic Use of Satellite Data and Numerical Models

Yoo-jin Kang (1), Hyun-young Choi (1), Jung-ho Im (1), Eun-na Jang (1), Jung-hee Lee (1), Yeon-su Lee (1)

[1] 50, UNIST-gil, Ulsan 44919, Republic of Korea
Email: kangyj@unist.ac.kr; hyong56@unist.ac.kr; ersgis@unist.ac.kr; enjang@unist.ac.kr; olive7861@unist.ac.kr; leeysu0423@unist.ac.kr

Abstract: Sulfur dioxide, SO_2, is a precursor that generates secondary air pollutants through chemical reactions in the atmosphere and they have negative effects on human health. Thus, people want to know about ground level SO_2 concentrations. Although the station-based monitoring has been conducted and accurate, it is difficult to provide continuous concentration for the large areas due to the sparse distribution of stations. In addition, satellite data that can provide continuous information has only atmospheric vertical column density. To compensate for these limitations, this study tried to estimate the ground concentration of SO2 through the synergistic use of satellite data and numerical models using machine learning method. In this study, Ozone Monitoring Instrument (OMI) satellite data which has O_3, NO_2, SO_2 and HCHO vertical column density with the 25km spatial resolution is used to quantify the ground SO_2 concentration based on machine learning approaches. In addition to these vertical column densities, total 57 variables containing other satellite, meteorological variables extracted from numerical model, emission model data are used together. For the machine learning model, Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boost (XGB) and Artificial Neural Network (ANN) models were constructed as base models and stacking ensemble method was applied. The stacking ensemble model consists of two-step with different input variables. To do this, training samples are divided into two parts for each step of the stacking ensemble model, and input variables for each step are selected by the Boruta algorithm. Then, 33 variables are used in step 1 models and the rest 24 variables are used in step 2 models. Our final model consists of best combination of base models and has R^2 = 0.85, RMSE = 0.0015 ppm for training sample, R^2 = 0.48, RMSE = 0.0023 ppm for the validation sample and R^2 = 0.47 and RMSE = 0.0030 ppm for test sample, respectively about 1km spatial resolution in South Korea (Figure 1). Although it shows an underestimated pattern, it has better performance than individual models and they well simulated time series patterns except for several high concentration cases. When this model is applied for mapping the ground concentration, high concentration cases are observed near the large cities and plant areas and the result maps showed the seasonal variations of SO_2 concentration well (Figure 2).

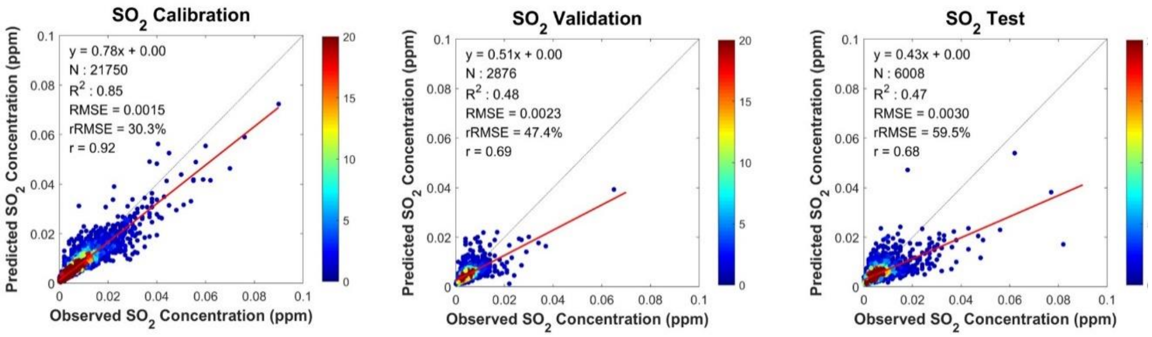

Figure 1 Stacking ensemble model calibration, model validation and model test result in scatter plot.

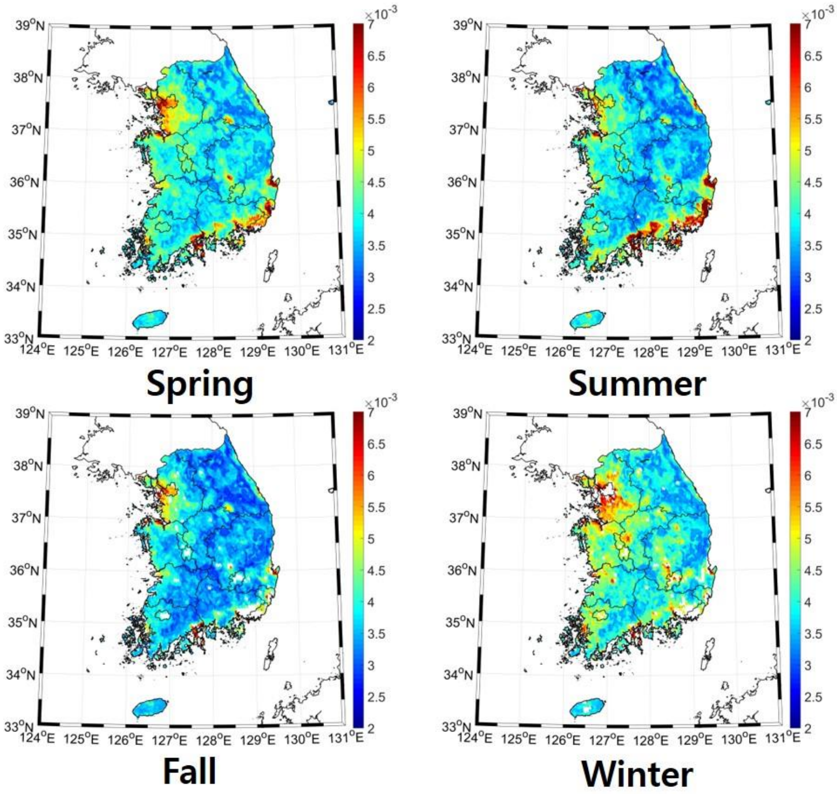

Figure 2 Seasonal map based on stacking ensemble model.

Keywords: Satellite data, air quality, sulfur dioxide, machine learning