Using land-use machine learning models to estimate daily NO_2 concentration variations in Taiwan

Chih-Da Wu, Ph.D.

Adjunct Associate Research Fellow / Associate Professor

National Institute of Environmental Health Sciences, National Health Research Institutes, Miaoli, Taiwan

Department of Geomatics, National Cheng Kung University, Tainan, Taiwan

E-mail: chidawu@mail.ncku.edu.tw

ABSTRACT

Exposure surrogates from limited monitoring stations may not be enough for epidemiological study in a large area. However, air pollution modelling approaches still require to improve the spatiotemporal resolution for better estimation. Otherwise, the exposure assessment would not be feasible for health risk assessment. To deal with this challenge, this study featured a Land-Use Regression (LUR) models that uses machine learning to assess the spatial-temporal variability of Nitrogen Dioxide (NO_2). Daily average NO_2 data was collected from 70 fixed air quality monitoring stations on the main island of Taiwan that belonged to the Taiwan EPA. Arou nd 0.41 million observations from 2000 to 2016 were used for the analysis. Several datasets were collected to determine spatial predictor variables, including the EPA environmental resources dataset, the meteorological dataset, the land-use inventory, the landmark dataset, the digital road network map, the digital terrain model, MODIS Normalized Difference Vegetation Index database, and the power plant distribution dataset. Conventional LUR and Hybrid Kriging-LUR were firstly used to identify the important predictor variables. Deep neural network, random forest, and XGBoost algorithms were then used to fit the prediction model based on the variables selected by the LUR models. Data splitting, 10-fold cross validation, external data verification, seasonal-based and county-based validation methods were used to verify the robustness of the developed models. The results showed that the proposed conventional LUR and Hybrid Kriging-LUR models respectively captured 65% and 78% of NO_2 variation. When XGBoost algorithm was further incorporated, the explanatory power of the model was respectively increased to 84% and 91%. The Hybrid Kriging-LUR with XGBoost algorithm outperformed the other integrated

methods. This study demonstrated the value of combining Hybrid Kriging-LUR model and an XGBoost algorithm to estimate the spatial-temporal variability of NO_2 exposure. Emission reduction strategies and land-use management could be made according to the association of specific land-use/land cover types selected in the final mode.

Keywords: NO_2; Land-use regression; Ordinary Kriging; Machine learning; Taiwan