GEO-AI PREDICTION MODEL ON ESTIMATING SPATIOTEMPORAL VARIATION OF $PM_{2.5}$ CONCENTRATIONS IN MORNING AND EVENING RUSH HOURS- A CASE STUDY IN TAIPEI METROPOLITAN, TAIWAN

Pei-Yi Wong (1), Huey-Jen Su (1), Chih-Da Wu (2,3,4*)

[1]Department of Environmental and Occupational Health, National Cheng Kung University, Tainan, Taiwan.
[2]Department of Geomatics, National Cheng Kung University, Tainan, Taiwan.
[3]National Institute of Environmental Health Sciences, National Health Research Institutes, Miaoli, Taiwan.
[4]Innovation and Development Center of Sustainable Agriculture, National Chung Hsing University, Tainan, Taiwan
Corresponding author: Chih-Da Wu, chidawu@mail.ncku.edu.tw

ABSTRACT

Urban air pollution has been a critical issue worldwide. It is indicated that $PM_{2.5}$ poses many adverse health effects. Since the $PM_{2.5}$ level during rush hours significantly contributes to overall exposure, it is crucial to investigate the effects during commutes. This study aimed to propose a geospatial artificial intelligence (Geo-AI) prediction model to estimate the spatiotemporal variation of $PM_{2.5}$ concentrations during morning and evening rush hours in Taiwan. Hourly $PM_{2.5}$ measurements from 2006 to 2020 were collected from 74 air quality monitoring stations established by the Taiwan Environmental Protection Administration. Total of 0.35 million observations were involved in analysis. We further aggregated hourly $PM_{2.5}$ into morning (7:00-9:00) and evening (16:00-18:00) averages. Potential predictor variables included co-pollutants (NO_2 and SO_2), land-use/land cover information, landmarks, satellite images. Totally, around 400 potential variables were included in the modelling process. For feature selection, we adopted SHapley Additive exPlanations (SHAP) index as a reference to remove the redundant features. To develop the Geo-AI prediction model, Kriging interpolation, land-use regression, machine learning, and ensemble stacking approach were utilized. For model validation, 10-fold cross validation (CV) and temporal external data were included to test overfitting issue and the model extrapolation capability. The Geo-AI model captured 90% and 95% of $PM_{2.5}$ variability, and the root mean square error (RMSE) was 4.85 and 3.75 $\mu g/m^3$ for morning and evening periods, respectively. Similar results were obtained from 10-fold CV and external data validation with R^2 of 0.90 and 0.83 for morning, 0.94 and 0.77 for evening periods. The selected variables showed that Kriged $PM_{2.5}$, distance to industrial parks, and the density of roads explained most of $PM_{2.5}$ variation for morning rush hour. It's also discovered that Kriged $PM_{2.5}$, distance to industrial parks, and the density of temples affected $PM_{2.5}$ variation in evening rush hour. The developed Geo-AI could estimate the spatiotemporal variation of $PM_{2.5}$ concentrations with a high prediction accuracy for morning and evening rush hours. Important features were identified by the explainable SHAP index through machine learning process. The spatial distribution of estimated $PM_{2.5}$ could provide information for the government to manage urban air pollution controlling strategies.

Keywords: Air Pollution, $PM_{2.5}$, Rush hour, Ensemble model, Geo-AI

1. INTRODUCTION

$PM_{2.5}$ has evidenced many adverse health effects on human health. Exposure to short-term $PM_{2.5}$ might lead to the increasing risk of cardiovascular disease, respiratory diseases, asthma exacerbation, and mortality (Chen et al., 2018, 2017; Liu et al., 2021). It is inevitable to tackle the challenge in controlling air pollution exposures. In addition, $PM_{2.5}$ levels during commuting periods were indicated a two to three times higher compared than the background situations (Correia et al., 2020; de Nazelle et al., 2017). Although people spend less time in commuting, the exposure might attribute higher health risk. A study assessing the health risk from dynamic $PM_{2.5}$ exposure in a metropolitan indicated that high-level $PM_{2.5}$ was often underestimated in morning rush hours, which make a severe situation for commuters (Li et al., 2021). An accurate air pollution exposure assessment for rush hours is in an urgent need for the population.

To address this issue, air pollution modelling strategies have been used to depict air pollution variations in a large surface based on geospatial technologies, including Kriging interpolation, Land-use Regression (LUR), satellite-derived data, and machine learning models (de Hoogh et al., 2016; Eeftens et al., 2012; Wong et al., 2021). With the ability of capturing nonlinear relationship between potential emission sources and target air pollution, machine learning provides a great prediction accuracy in air pollution spatial modelling. Furthermore, studies integrate the advantages in considering spatial variance of geospatial information and assembling the predictions from different modelling procedures to bring in geospatial-artificial intelligencebased (Geo-AI-based) machine learning models (Babaan et al., 2023; Wong et al., 2023). For example, the Geo-AI models explained 90% of variation in daily $PM_{2.5}$ concentration and improve 38% than traditional LUR model (Wong et al., 2023).

The aim of this study was to develop Geo-AI prediction models to estimate the spatial variation of $PM_{2.5}$ concentration during morning and evening commuting rush hours.

2. MATERIALS AND MEHODS

2.1 Air pollution

Hourly $PM_{2.5}$ measurements from 2006 to 2020 were collected from the monitoring stations established by the Taiwan Environmental Protection Administration (Taiwan EPA). Total of 0.35 million observations were involved in analysis. We further aggregated hourly $PM_{2.5}$ into morning (7:00-9:00) and evening (16:00-18:00) averages. The categorized rush hours were adopted to better reflect the commuting periods of elementary school children, who were a representative of vulnerable group. In addition, Kriged $PM_{2.5}$ level from a leave-one-out Ordinary Kriging approach was used as a predictor to consider the impact of surrounding sites on the target monitoring station (Wu et al., 2018).

2.2 Potential predictor variables

Potential predictor variables included co-pollutants (NO_2 and SO_2), land-use/land cover information, landmarks, roads, and satellite images. All the potential spatial predictors were calculated the density surrounding each monitoring station within the circular buffer range from 50-m, 150-m, 250-m, 500-m, 750-m, 1000-m, 1250m, 1500-m, 1750-m, 2000-m, 2500-m, 3000-m, 4000-m, and 5000-m. Totally, around 400 potential variables were included in the modelling process. For feature selection, we adopted SHapley Additive exPlanations (SHAP) index as a reference to remove the redundant features. All predictors were fitted in XGBoost to generate SHAP values. The absolute SHAP index of each predictor was sorted, and only the important features with an R^2 change greater than 0.001 were retained as key influential factors (Hsu et al., 2023).

2.3 Development of Geo-AI prediction models

To develop the Geo-AI prediction models, Kriging interpolation, land-use regression, machine learning, and ensemble stacking approach were utilized. The dataset was split 80% into training set for model building and the remaining 20% for model testing. This study utilized five machine learning algorithms including XGBoost, LightGBM, CatBoost, gradient boosting machine, and random forest. These algorithms were trained at first, and the predictions were used to fit the stacking ensemble model. For model validation, 10-fold cross validation (CV) and temporal external data were included to test overfitting issue and the model extrapolation capability. R^2 and root mean square error (RMSE) were applied as indicators for performance evaluation. The final Geo-AI prediction model was selected based on the indicators that outperformed the other models.

2.4 Spatial modelling of $PM_{2.5}$ during morning and evening rush hours

The spatial distribution of $PM_{2.5}$ concentrations was further investigated using the finalized Geo-AI prediction model. Seasonal variations during morning and evening rush hours were depicted for the period from 2006 to 2020. The spatial differences between morning and evening were further investigated.

3. RESULTS AND DISCUSSION

3.1 Key influential factors

The results showed that Kriged $PM_{2.5}$, the distance to industrial parks, and road density within a 150-m circular buffer explained most of $PM_{2.5}$ variations during morning rush hours in Taipei metropolises. It's also discovered that Kriged $PM_{2.5}$, the distance to industrial parks, and temple density within a 150-m circular buffer affected $PM_{2.5}$ variations in evening rush hours. It was illustrated that surrounding $PM_{2.5}$ levels would have a certain association with concentration from the target monitoring site (Wu et al., 2018). Taking this into account, we incorporated Kriging interpolated $PM_{2.5}$ as a predictor and it indeed contributed a large effect on the Geo-AI models. On the other hand, the worshipping events and incense burning in temples were reported to be associated with the increment of $PM_{2.5}$ levels (Lung and Kao, 2003; Wang et al., 2021; Xu et al., 2020). $PM_{2.5}$ levels would also be contributed form traffic related emissions (Hassanpour Matikolaei et al., 2019). Hence, the density of roads and temples were included as key influential factors.

3.2 Performance of Geo-AI prediction model

The Geo-AI prediction model captured 90% and 95% of $PM_{2.5}$ variability, and the root mean square error (RMSE) was 4.85 and 3.75 $\mu g/m^3$ during morning and evening periods, respectively. The testing model also showed a high prediction accuracy with a R^2 value of 0.87 during both morning and evening rush hours. The results from 10-fold CV and external data validation were consistent with that of training models with R^2 values of 0.90 and 0.83 for the morning models, 0.94 and 0.77 for the evening models.

3.3 Spatial distribution of estimated $PM_{2.5}$

Estimation maps were made to show the spatial and seasonal variations of $PM_{2.5}$ concentrations during morning and evening rush hours (Figure 1 and 2). For the seasonal variations, it was observed that $PM_{2.5}$ level was highest in winter, followed by spring, fall, and summer in both morning and evening rush hours. This phenomenon was consistent with that of measurements from the monitoring stations. The spatial distribution showed that $PM_{2.5}$ was concentrated in the central and western parts of Taipei metropolises in the morning. Conversely, $PM_{2.5}$ diffused to surrounding areas in the evening. The difference of $PM_{2.5}$ spatial distribution between morning and evening might be due to the difference of population mobility and the lag effects of air pollution (Lu, 2023; Yang et al., 2021).

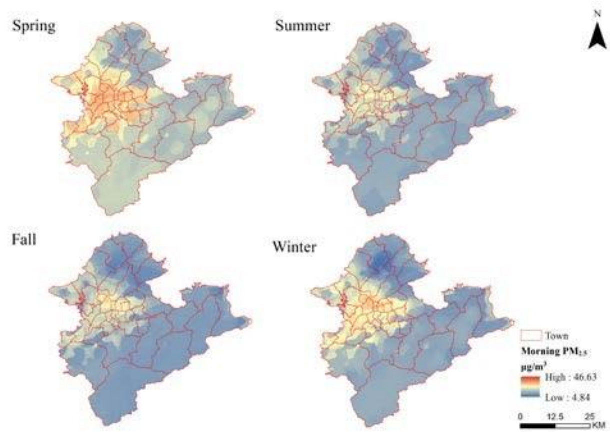

Figure 1. Spatial distribution of seasonal $PM_{2.5}$ estimation during morning rush hours

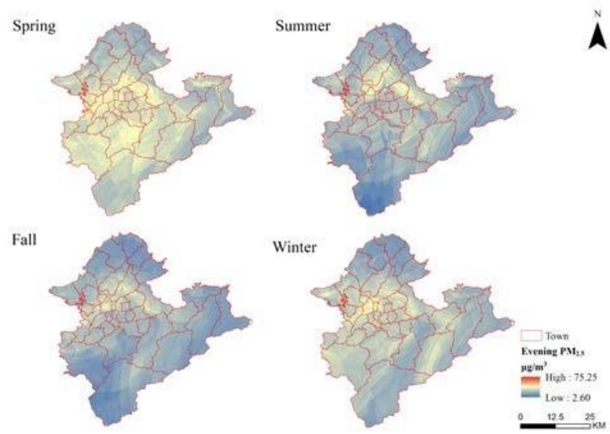

Figure 2. Spatial distribution of seasonal $PM_{2.5}$ estimation during evening rush hours

4. CONCLUSIONS

The developed Geo-AI prediction models capture the spatiotemporal variation of $PM_{2.5}$ concentrations with a high prediction accuracy of 0.90 and 0.95 for morning and evening rush hours, respectively. Kriged $PM_{2.5}$, density of roads and temples are the main factors affecting $PM_{2.5}$ levels during both morning and evening rush hours in Taipei metropolises. The prediction maps could be applied to investigate the transport of $PM_{2.5}$ in large areas.

References

Babaan, J., Hsu, F.T., Wong, P.Y., Chen, P.C., Guo, Y.L., Lung, S.C.C., Chen, Y.C., Wu, C.D., 2023. A GeoAI-based Ensemble Mixed Spatial Prediction Model with Fine Spatial-Temporal Resolution for Estimating Daytime/Nighttime/Daily Average Ozone Concentrations Variations in Taiwan. J. Hazard. Mater. 130749.

Chen, C., Zhu, P., Lan, L., Zhou, L., Liu, R., Sun, Q., Ban, J., Wang, W., Xu, D., Li, T., 2018. Short-term exposures to $PM_{2.5}$ and cause-specific mortality of cardiovascular health in China. Environ. Res. 161, 188–194.

Chen, R., Yin, P., Meng, X., Liu, C., Wang, L., Xu, X., Ross, J.A., Tse, L.A., Zhao, Z., Kan, H., Zhou, M., 2017. Fine Particulate Air Pollution and Daily Mortality. A Nationwide Analysis in 272 Chinese Cities. Am J Respir Crit Care Med 196, 73–81. https://doi.org/10.1164/rccm.201609-1862OC

Correia, C., Martins, V., Cunha-Lopes, I., Faria, T., Diapouli, E., Eleftheriadis, K., Almeida, S.M., 2020. Particle exposure and inhaled dose while commuting in Lisbon. Environ. Pollut. 257, 113547.

de Hoogh, K., Gulliver, J., van Donkelaar, A., Martin, R.V, Marshall, J.D., Bechle, M.J., Cesaroni, G., Pradas, M.C., Dedele, A., Eeftens, M., 2016. Development of West-European $PM_{2.5}$ and NO_2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environ. Res. 151, 1–10.

de Nazelle, A., Bode, O., Orjuela, J.P., 2017. Comparison of air pollution exposures in active vs. passive travel modes in European cities: A quantitative review. Environ. Int. 99, 151–160.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., de Nazelle, A., 2012. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46, 11195–11205.

Hassanpour Matikolaei, S.A.H., Jamshidi, H., Samimi, A., 2019. Characterizing the effect of traffic density on ambient CO, NO_2, and $PM_{2.5}$ in Tehran, Iran: an hourly land-use regression model. Transp. Lett. Int. J. Transp. Res. 11.

Hsu, C.Y., Lin, T.W., Babaan, J.B., Asri, A.K., Wong, P.Y., Chi, K.H., Ngo, T.H., Yang, Y.H., Pan, W.C., Wu, C.D., 2023. Estimating the Daily Average Concentration Variations of PCDD/Fs in Taiwan Using a Novel Geo-AI Based Ensemble Mixed Spatial Model. J. Hazard. Mater. 131859.

Li, Xin, Yang, T., Zeng, Z., Li, Xiaodong, Zeng, G., Liang, J., Xiao, R., Chen, X., 2021. Underestimated or overestimated? Dynamic assessment of hourly $PM_{2.5}$ exposure in the metropolitan area based on heatmap and micro-air monitoring stations. Sci. Total Environ. 779, 146283.

Liu, L., Liu, C., Chen, R., Zhou, Y., Meng, X., Hong, J., Cao, L., Lu, Y., Dong, X., Xia, M., 2021. Associations of short-term exposure to air pollution and emergency department visits for pediatric asthma in Shanghai, China. Chemosphere 263, 127856.

Lu, Y., 2023. Assessing air pollution exposure misclassification using high-resolution $PM_{2.5}$ concentration model and human mobility data. Air Qual. Atmos. Heal. 1–14.

Lung, S.C., Kao, M., 2003. Worshippers' exposure to particulate matter in two temples in Taiwan. J. Air Waste Manag. Assoc. 53, 130–135.

Wang, C.Y., Lim, B.S., Wang, Y.H., Huang, Y.C.T., 2021. Identification of High Personal $PM_{2.5}$ Exposure during Real Time Commuting in the Taipei Metropolitan Area. Atmosphere (Basel). 12, 396.

Wong, P.Y., Lee, H.Y., Zeng, Y.T., Chern, Y.R., Chen, N.T., Lung, S.C.C., Su, H.J., Wu, C.D., 2021. Using a Land Use Regression Model with Machine Learning to Estimate Ground Level $PM_{2.5}$. Environ. Pollut. 116846.

Wong, P.Y., Su, H.J., Lung, S.C.C., Wu, C.D., 2023. An ensemble mixed spatial model in estimating longterm and diurnal variations of $PM_{2.5}$ in Taiwan. Sci. Total Environ. 161336.

Wu, C.D., Zeng, Y.T., Lung, S.C.C., 2018. A hybrid kriging/land-use regression model to assess $PM_{2.5}$ spatial-temporal variability. Sci. Total Environ. 645, 1456–1464.

Wu, C.D., Zeng, Y.T., Lung, S.C.C., 2018. A hybrid kriging/land-use regression model to assess $PM_{2.5}$ spatial-temporal variability. Sci. Total Environ. 645, 1456–1464.

Xu, H., Ta, W., Yang, L., Feng, R., He, K., Shen, Z., Meng, Z., Zhang, N., Li, Y., Zhang, Y., 2020. Characterizations of $PM_{2.5}$-bound organic compounds and associated potential cancer risks on cooking emissions from dominated types of commercial restaurants in northwestern China. Chemosphere 261, 127758.

Yang, Z., Yang, J., Li, M., Chen, J., Ou, C.Q., 2021. Nonlinear and lagged meteorological effects on daily levels of ambient $PM_{2.5}$ and O_3: Evidence from 284 Chinese cities. J. Clean. Prod. 278, 123931.

Acknowledgements

This study was grant supported by the National Science and Technology Council, Taiwan (MOST 111-2121-M006-004-; MOST 110-2628-M-006 -001 -MY3), the National Health Research Institutes (NHRI-109A1EMCO-02202212), Academia Sinica, Taiwan (AS-SS110-02) under "Trans-disciplinary $PM_{2.5}$ Exposure Research in Urban Areas for Health-oriented Preventive Strategies (II)", Environmental Protection Administration, Taiwan (EPA112F090), and the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. The authors are grateful to the National Aeronautics and Space Administration (NASA) and to the U.S. Geological Survey (USGS) for remote sensing data supports.