EnvironmentAsia
The international journal by the Thai Society of Higher Education Institutes on Environment

EnvironmentAsia 15(2) 2022 15-22
DOI 10.14456/ea.2022.29
ISSN 1906-1714; ONLINE ISSN: 2586-8861

Land Use Random Forests Model Incorporating WRF/CMAQ for Estimating Daily $PM_{2.5}$ Concentration in Bangkok, Thailand

Tin Thongthammachart*, Shin Araki, Hikari Shimadera, Tomohito Matsuo, and Akira Kondo

Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan

*Corresponding author: tin.thongthammachart@ea.see.eng.osaka-u.ac.jp

This article received the best oral presentation award in Transboundary Pollution, Climate Change and Atmospheric Research Session in the 6th EnvironmentAsia Virtual International Conference; December 20-21, 2021 at Bangkok, Thailand
Revised: January 27, 2022; Accepted: February 7, 2022

Abstract

Fine particulate matter ($PM_{2.5}$) level in Bangkok and vicinity areas, Thailand, has increased for a half decade and exceeds the national air quality standard in dry season. However, there are a limited number of air quality monitoring stations measuring $PM_{2.5}$ concentration. A land use regression (LUR) model can be applied to an estimate of $PM_{2.5}$ concentrations at non-monitored locations. This is the first study that develops the LUR model with the random forests (RF) algorithm incorporating WRF/CMAQ for predicting daily $PM_{2.5}$ concentration for the year 2017 in Bangkok and vicinity areas. CMAQ-simulated $PM_{2.5}$ concentrations, WRF-simulated meteorological parameters, population density, and land use variables were used as predictors of the LUR model. The predictor variables were assessed by variable importance measurements. A cross-validation (CV) technique was performed to evaluate the prediction performance in spatial and temporal aspects and obtained the coefficient of determination (R^2) and root mean square error (RMSE) values. The daily $PM_{2.5}$ concentrations were estimated by the developed model at 1×1 km resolution in the study area. The predictive performance of the developed land use random forests (LURF) model was compared with that of the conventional LUR model built with multiple linear regression to reveal the advantages of the RF algorithm. Accordingly, the LURF model obtains a spatial CV-R^2 of 0.63 and a temporal CV-R^2 of 0.70, which are higher than that of the conventional LUR model (a spatial and a temporal CV-R^2 of 0.46). The findings demonstrate that the LURF model incorporating WRF/CMAQ is advantageous to accurately estimate ambient $PM_{2.5}$ level in Thailand.

Keywords: Fine particulate matter; Land use regression; Random forests; WRF-CMAQ

1. Introduction

Air quality in central region of Thailand has become worse for a half decade due to the increase of fine particulate matter ($PM_{2.5}$) level. Moreover, the number of air quality monitoring stations for $PM_{2.5}$ measurement are limited (PCD, 2018). To estimate $PM_{2.5}$ at non-monitored locations, a land use regression (LUR) model can be applied for this demand (Hoek *et al.,* 2008). The conventional LUR model is developed from multiple linear regression (MLR) technique. Chemical transport model (CTM) can be applied to the LUR model as a predictor variable to enhance the predictability (Di *et al.,* 2019, Gariazzo *et al.,* 2020). The CTM provides spatially and temporally resolved estimates of air pollutant concentrations. Nevertheless, the conventional LUR model bears the drawback of difficulty in capturing non-linearity between the objective

pollutant concentration and the predictors.

The machine learning technique was introduced to the LUR framework since it can efficiently handle non-linearity and deal with complex interactions among the predictive variables (Gupta and Christopher, 2009). A random forests (RF) technique developed from a decision tree-based machine learning is effectively used for solving regression and classification problems (Breiman, 2001). The RF performs several advantages, such as: 1) better forecasting performance compared to artificial neural networks (ANN), support vector machine (Liaw and Wiener, 2002), Naïve Bayes classifiers and logistic regression (Yu *et al.,* 2016); 2) robust against overfitting (Breiman, 2001); and 3) it has only two user-defined parameters: the number of variables in the subset at each node and the number of trees in the forest (Breiman, 2001).

In this study, a land use random forest (LURF) model incorporating WRF/CMAQ was developed to estimate daily $PM_{2.5}$ concentration in Thailand. The cross-validation (CV) method was applied to evaluate the model predictive performance. The predictive performance of a LURF model was compared to that of the conventional LUR model built with MLR technique to investigate whether the RF algorithm improves $PM_{2.5}$ predictability in the study domain. Finally, the advantages of this study were discussed and outlined its implications. This study would be beneficial for developing models for health and ecological risk assessment associated with $PM_{2.5}$ exposure, and for making science-based environmental policies to mitigate $PM_{2.5}$ levels in the future.

2. Materials and Methods

2.1 Study area and air quality monitoring stations

The study area is a 300 × 300-km domain in and around the central region of Thailand, Which includes the largest metropolitan area of Bangkok and vicinity provinces (Figure 1). The daily average $PM_{2.5}$ concentrations were obtained for the year 2017 from the database of the regulatory monitoring network provided by Pollution Control Department of Thailand and Bangkok Metropolitan Administration. The locations of the air monitors are shown in Figure 1. The observations only from the general ambient air-quality monitoring stations are utilized in this study since they are positioned to prevent the effects of traffic or specific emission sources. Moreover, the data from air quality monitoring stations with a temporal coverage of more than 80% in a year were only used to build the LUR model to guarantee that it was temporally representative. Therefore, a total of 11 monitoring stations were considered in this study.

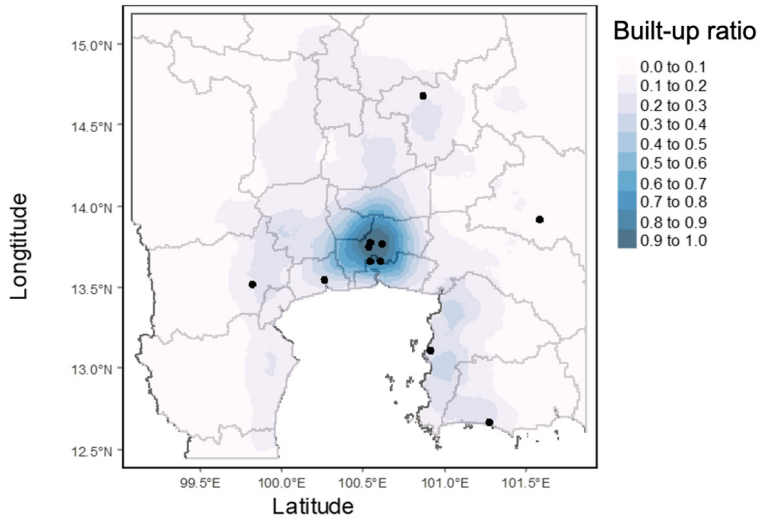

Figure 1. Study area - The dots represent the ambient $PM_{2.5}$ monitoring stations and their locations. The contour color represents the built-up area ratio.

2.2 Dataset and predictor variables

The datasets were selected by considering potential factors that can influence the spatiotemporal concentration of the air pollutant. Some of the gridded data including WRF/CMAQ features were resampled to conform to an origin and resolution of 1-km grid scale. The predictor variables are shown in Table 1.

2.2.1 WRF/CMAQ simulations

The Weather Research and Forecasting model (WRF) v3.8 is used for simulating meteorological features (Skamarock and Klemp, 2008). The Community Multiscale Air Quality Modeling System (CMAQ) v5.2.1 is used for simulating $PM_{2.5}$ concentration with meteorological fields produced by the WRF. The domain for the WRF/CMAQ simulation covers the central region of Thailand with 5-km grids which is identical to the study area (Figure 1). The WRF-simulated hourly meteorological fields were processed to prepare for the predictor variables, including planetary boundary layer (PBL) height, wind speed, air temperature, and relative humidity. The CMAQ-simulated hourly air-quality fields were processed to prepare the predictor variables.

2.2.2 Road, land use and population data

The road dataset obtained from Open Street Map (https://www.openstreetmap.org) was classified as trunk, motorway, primary, secondary, and tertiary. The trunk, motorway, and primary are extracted and included as a major road. The secondary and tertiary roads are extracted and included as a minor road. The road length in a grid cell was calculated for each classifier using road network data. The agricultural area ratio was extracted from land use data which is provided by Geo-Informatics and Space Technology Development Agency: GISDA, Thailand. The built up area ratio is download from the database of European Commission, Joint Research Centre (JRC) (Corbane *et al.,* 2020). The forest area was downloaded from the forest cover database (https://earthenginepartners.appspot.com/) by Hansen *et al.* (2013). The elevation data is obtained from Jarvis (2021). The population data is obtained from the WorldPop database, University of Southampton, UK (https://data.humdata.org/dataset/worldpop-population-counts-for-thailand). The focal-sum method was applied to the road, land use and population variables with the purpose of considering the distance decay effect as following the proposed technique by Vienneau *et al.* (2009).

Table 1. Predictor variables

Predictor Variables	Unit
CMAQ-simulated $PM_{2.5}$ concentration (CMAQ_PM2.5)	$\mu g/m^3$
WRF-simulated planetary boundary layer (WRF_PBL)	m
WRF-simulated wind speed (WRF_Wind)	m/s
WRF-simulated relative humidity (WRF_RH)	%
WRF-simulated ambient temperature (WRF_Temp)	K
Built-up area ratio	Unitless
Forest area ratio	Unitless
Agriculture area ratio	Unitless
Population density ratio	Unitless
Road length, Major Road	km/km^2
Road length, Minor Road	km/km^2
Land elevation	m

2.3 Model building

A spatiotemporal land use random forest (LURF) model was constructed with the predictor variables, as shown in Table 1. In this study, R statistical software version 4.0.5 was used for building model (R Core Team, 2020). The ranger package was used for the implementation of RFs (Wright and Ziegler, 2017). The number of variables in the subset at each node (m_{try}) was set to 2p/3, where p is the number of predictor variables in the entire data set (Wright and Ziegler, 2017). The other parameters were set to the default values. The R^2 of the model was calculated as 1 − MSE/var(Y), where MSE is the mean of the out-of-bag errors for all the prediction points, and Y is the observed values (Brokamp *et al.,* 2017). After obtaining the LURF model, we ranked the predictor variables by sorting the variable importance measure in the descending order. A land use linear regression model (LULR) was conducted for the comparison benchmark with the LURF model. The LULR models were constructed using a supervised stepwise MLR method (Beelen *et al.,* 2013). The selected potential predictor variables of the LULR models were the same as those of the LURF model presented in Table 1. We use the adjusted model R^2 for the model R^2 of the LULR model.

2.4 Model evaluation

The performance of the models were evaluated using CV in two different aspects: spatial and temporal CVs (Thongthammachart *et al.,* 2021a). These two validation features aim to test the model's effectiveness in estimating in either the other locations or the other periods. The "leave-one-monitoring station-out" was applied for spatial CV and the "leave-one-month-out" was applied for the temporal CV. The predicting accuracy from these CVs were investigated using the R^2 values between the predicted and observed values. The precision of the prediction was determined via the root mean square error (RMSE) between the predicted and observed values.

3. Results and Discussion

The measurement of the prediction potential of the variables for the LURF model are sorted by their order of importance in Figure 2. The CMAQ-simulated $PM_{2.5}$ concentration emerges as the most important variable. Relative humidity (WRF_RH) and ambient temperature (WRF_Temp) are the second and the third most influential variables. The model-R^2 value of the LURF model is 0.78. On the other hand, the final variables for the LULR were the CMAQ-simulated $PM_{2.5}$ concentration. The stepwise variable selection of the LULR can assume that the CMAQ variable is the most important predictor contributing to model accuracy. The adjusted R^2 value of the LULR model is 0.49.

In the model evaluation results (Figure 3), The cross-validated R^2 values, categorized by spatial and temporal CVs, of the LURF and LULR models, were 0.63 and 0.70, and 0.46 and 0.46, respectively. Furthermore, the cross validated RMSE values of these aspects were 9.0 and 8.1, as well as 10.8 and 10.8, respectively. Consequently, the LURF model reveals a higher model-R^2 and CVs–R^2 as well as lower CV–RMSE values at both spatial and temporal aspects than that of the LULR model. Accordingly, the results imply that the LURF model is superior performance to the LULR model for daily $PM_{2.5}$ prediction in the study domain.

Furthermore, the temporal CVs of the models reveal a higher R^2 and a lower RMSE values than that of the spatial CVs. To investigate the CV behavior, a coefficient of variation from the residuals of the daily $PM_{2.5}$ concentration was calculated in two aspects: the spatially averaged temporal variation (concentration on each day) and temporally averaged spatial variation (concentration at each station) (Thongthammachart *et al.,* 2021a). The coefficient of variation value of spatially averaged temporal variation is 126.9, while the value of temporally averaged spatial variation is 160.5. Hence, the residuals of the observed $PM_{2.5}$ are larger in terms of spatial variation than temporal variation. Additionally, the limited number of $PM_{2.5}$ monitoring stations probably causes weak goodness of the fit of the LUR model

for the site-to-site validation (Chalermpong *et al.,* 2021). Accordingly, these reasons could support the CV results why the temporal CVs exhibit higher predicting performance than the spatial CVs.

In addition, the LUR model in this study achieved high CV-R^2 with well agreement between the observed and the predicted $PM_{2.5}$ levels. The integration of RF and CTM to the LUR framework for accurately estimating daily mean $PM_{2.5}$ levels has been effectively deployed in numerous developed and/or rich countries for enhanced prediction at national and regional scales. The CV–R^2 obtained by similar studies are 0.85 for Continental United States (Di *et al.,* 2019, Di *et al.,* 2016), 0.61 for China (Xue *et al.,* 2019), and 0.81 for Japan (Thongthammachart *et al.,* 2021b). Besides, the previous study was successfully used the LURF model to forecast $PM_{2.5}$ at the city scale and their model results obtained CV–R^2 value of 0.72 (Gariazzo *et al.,* 2020). Accordingly, the LURF incorporating CTM is effectively applicable for daily $PM_{2.5}$ prediction at varying geographical scales.

This study makes valuable contributions as evidenced from its outcomes. The developed LUR model accurately predicts daily $PM_{2.5}$ levels, which have high spatiotemporal variability. The application of the RF technique to LUR model can enhance the predictability of $PM_{2.5}$ concentration rather than using the conventional LUR model. The model can be confidently applied to future studies to achieve accurate predictions which are advantageous for human and ecological risk assessment.

Although our study exhibits the advantages, it has some limitations. First, the number of $PM_{2.5}$ monitoring stations is limited. It possibly affects the performance of the fit LUR model. The LUR model's performance could possibly be improved by increasing the number of air measurement stations. Second, this study only utilized the daily average $PM_{2.5}$ concentration for one year (2017), which is insufficient for a comprehensive investigation of the long-term trend of $PM_{2.5}$. Apart from the limitations, this study successfully developed the LURF model incorporating WRF/CMAQ to estimate daily $PM_{2.5}$ levels in Bangkok, Thailand.

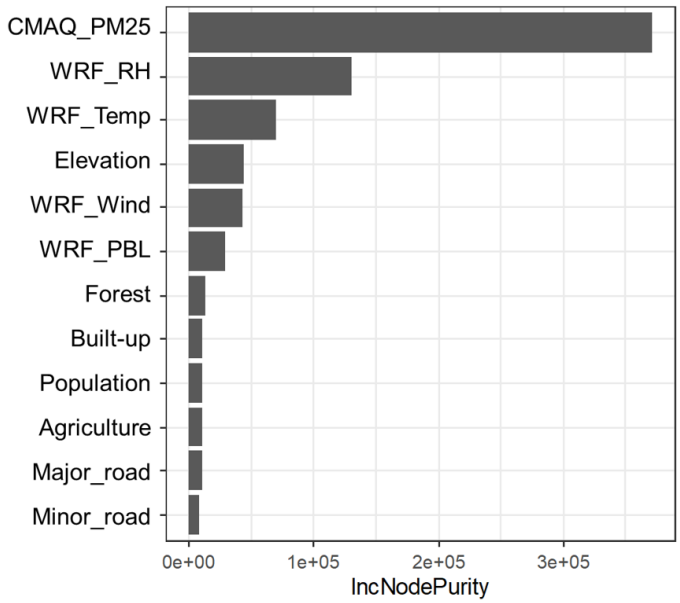

Figure 2. Variable importance plot for the LURF model. The variables are listed based on the importance from top to bottom

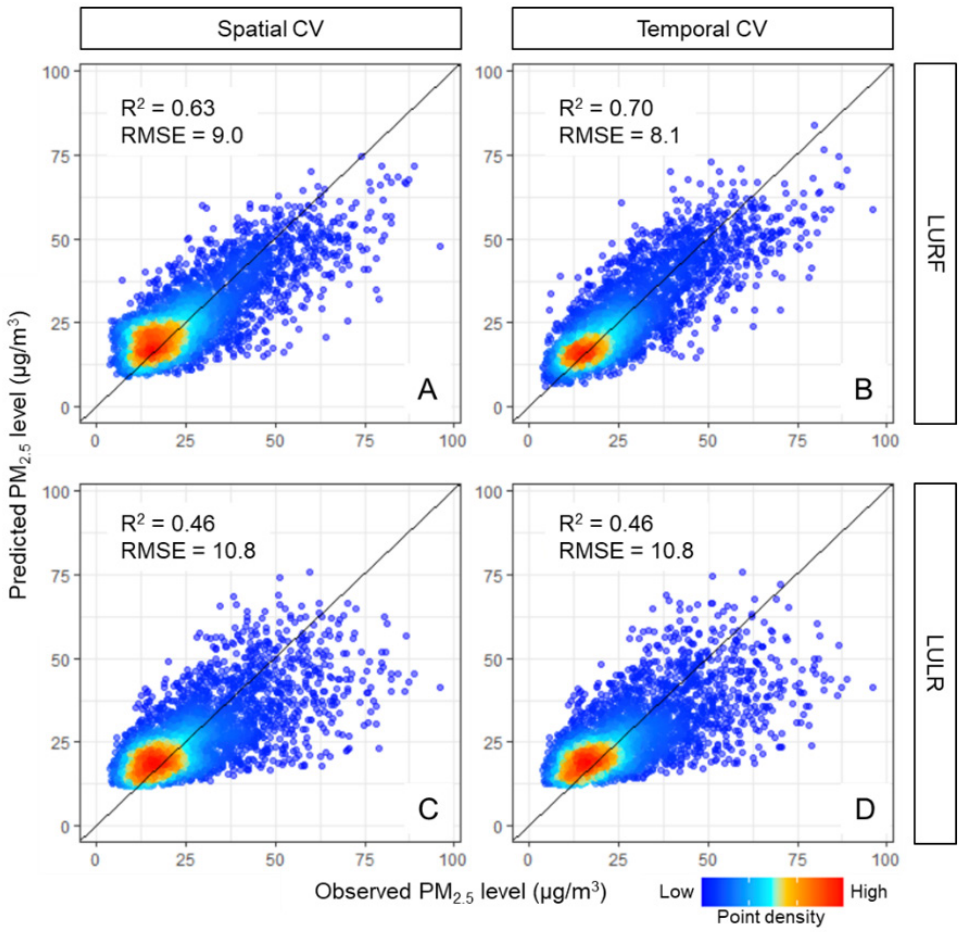

Figure 3. Scatter plots of the predicted and observed daily average concentrations of $PM_{2.5}$ obtained by cross validation. (A) and (B), and (C) and (D) show the LURF and LULR models, respectively. Similarly, (A) and (C), and (B) and (D) show spatial and temporal cross-validations. The red and blue colors indicate higher and lower point densities, respectively

4. Conclusion

This study was successfully implemented the RF technique with WRF/CMAQ to LUR model to accurately estimate ambient $PM_{2.5}$ concentrations for the first time in Thailand. The important advantages of applying the RF technique of capturing the non-linearity between the predicted $PM_{2.5}$ concentration and predictor variables was illustrated in this study. The application of the WRF/CMAQ into the LURF model obviously improves air pollutant predicting performance in terms of accuracy and precision as indicated by the higher CV-R^2 and lower CV-RMSE values compared to the conventional LUR model. In future studies, the LURF model will be intended to apply to a more extensive area and a longer time, as well as investigate spatiotemporal distribution of $PM_{2.5}$ levels in Thailand and other developing countries.

Acknowledgement

We extend my sincere thanks to the Pollution Control Department of Thailand and Bangkok Metropolitan Administration for providing air pollution data. This study was supported by the Environment Research and Technology Development Fund of the Environmental Restoration and Conservation Agency of Japan [grant number JPMEERF20195055]; and the JSPS KAKENHI [grant Number 19K12370].

References

Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli X, Tsai M-Y, Künzli N, Schikowski T, Marcon A, Eriksen KT, Raaschou-Nielsen O, Stephanou E, Patelarou E, Lanki T, Yli-Tuomi T, Declercq C, Falq G, Stempfelet M, Birk M, Cyrys J, von Klot S, Nádor G, Varró MJ, Dėdelė A, Gražulevičienė R, Mölter A, Lindley S, Madsen C, Cesaroni G, Ranzi A, Badaloni C, Hoffmann B, Nonnemacher M, Krämer U, Kuhlbusch T, Cirach M, de Nazelle A, Nieuwenhuijsen M, Bellander T, Korek M, Olsson D, Strömgren M, Dons E, Jerrett M, Fischer P, Wang M, Brunekreef B, de Hoogh K. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. Atmospheric Environment 2013; 72:10-23.

Breiman L. Random Forests. Machine Learning 2001; 45(1): 5-32.

Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmospheric Environment 2017; 151: 1-11.

Chalermpong S, Thaithatkul P, Anuchitchanchai O, Sanghatawatana P. Land use regression modeling for fine particulate matters in Bangkok, Thailand, using time-variant predictors: Effects of seasonal factors, open biomass burning, and traffic-related factors. Atmospheric Environment 2021; 246:118128.

Corbane C, Sabo F, Politis P, Syrris V. GHS-BUILT-S2 R2020A - GHS built-up grid, derived from Sentinel-2 global image composite for reference year 2018 using Convolutional Neural Networks (GHS-S2Net). In: European Commission JRCJ, editor. 2020.

Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, Sabath MB, Choirat C, Koutrakis P, Lyapustin A, Wang Y, Mickley LJ, Schwartz J. An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. Environment International 2019;130:104909.

Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang YJ, Schwartz J. Assessing $PM_{2.5}$ Exposures with High Spatiotemporal Resolution across the Continental United States. Environmental Science & Technology 2016; 50(9): 4712-21.

Gariazzo C, Carlino G, Silibello C, Renzi M, Finardi S, Pepe N, Radice P, Forastiere F, Michelozzi P, Viegi G, Stafoggia M. A multi-city air pollution population exposure study: Combined use of chemical-transport and random-Forest models with dynamic population data. Science of The Total Environment 2020; 724: 138102.

Gupta P, Christopher SA. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. Journal of Geophysical Research. 2009;114(D20).

Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, Kommareddy A, Egorov A, Chini L, Justice CO, Townshend JRG. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science 2013; 342(6160): 850-3.

Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, Briggs D. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment 2008; 42(33): 7561-78.

Jarvis A. SRTM 90m DEM Digital Elevation Database; 2021.

Liaw A, Wiener M. Classification and regression by random forest. R News 2002; 2: 18-22.

PCD. Thailand state of pollution report 2017. Bangkok, Thailand: Pollution Control Department; 2018.

R Core Team. R: A Language and Environment for Statistical Computing Vienna, Austria2020 [Available from: https://www.R-_project.org/.

Skamarock WC, Klemp JB. A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. Journal of Computational Physics 2008; 227(7): 3465-85.

Thongthammachart T, Araki S, Shimadera H, Eto S, Matsuo T, Kondo A, editors. High Spatiotemporal NO_2 Estimates by Land Use Random Forests Integrated with WRF/CMAQ. A&WMA's 114th Annual Conference & Exhibition; 2021a June 14 - June 17, 2021; Florida, USA.

Thongthammachart T, Araki S, Shimadera H, Eto S, Matsuo T, Kondo A. An integrated model combining random forests and WRF/CMAQ model for high accuracy spatiotemporal $PM_{2.5}$ predictions in the Kansai region of Japan. Atmospheric Environment 2021b; 262: 118620.

Vienneau D, de Hoogh K, Briggs D. A GIS-based method for modelling air pollution exposures across Europe. Science of the Total Environment 2009; 408(2): 255-66.

Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software; Vol 1, Issue 1 (2017); 2017.

Xue T, Zheng Y, Tong D, Zheng B, Li X, Zhu T, Zhang Q. Spatiotemporal continuous estimates of $PM_{2.5}$ concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. Environment International 2019;123: 345-57.

Yu R, Yang Y, Yang L, Han G, Move OA. RAQ-A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors (Basel) 2016; 16(1): 86.