A NEURAL NETWORK-BASED LAND USE REGRESSION MODEL TO ESTIMATE SPATIAL-TEMPORAL VARIABILITY OF NITROGEN DIOXIDE

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

[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.
Email: aa6624tw@gmail.com; chidawu@mail.ncku.edu.tw; hjsu@mail.ncku.edu.tw

KEY WORDS: Air Pollution, Nitrogen Dioxide, Land-Use Regression, Deep Neural Network

ABSTRACT: Nitrogen dioxide (NO_2) is a kind of highly reactive gas and secondary pollutant mainly from burning fossil fuels, which were predominant species in vehicle exhaust. Since traffic volume density is heavy and large number of temples and restaurants were densely distributed in Taiwan. The high concentration of NO_2 may cause adverse effects on respiratory system. To estimate NO_2 concentration more accurately, this study aimed to utilize a neural network-based land use regression model to assess the spatial-temporal variability. Daily average NO_2 data were collected from 70 fixed air quality monitoring stations in Taiwan main island which were established by Taiwan Environment Protective Administration. Totally, around 0.41 million observations were collected for our analysis. Several datasets were collected for obtaining spatial predictor variables, including EPA environmental resources dataset, meteorological dataset, land-use inventory, landmark dataset, digital road network map, DTM, MODIS NDVI dataset, and thermal power plant distribution dataset. To establish the integrated approach, conventional land-use regression (LUR) was first used to identify the important predictors variables. Then a deep neural network (DNN) algorithm was applied to fit the prediction model. 10-fold cross validation and external data verification methods were used to further confirm the robustness of model performance. The results showed that, the developed conventional LUR model captured 60% of NO_2 variation. Of the 11 variables selected by the stepwise variable selection procedure, PM_{10}, SO_2, O_3 explained 18%, 7% and 5% NO_2 variation, respectively. After integrating DNN algorithm with conventional LUR method, the model explanatory power was increased to 85%, with a 25% improved in model performance. Consistent findings were obtained from the 10-fold cross validation, while the cross-validated R^2 was increased from 61% to 83%, and root-mean-square error (RMSE) was decreased from 6.56 ppb to 4.34 ppb. This study demonstrates the value of incorporating the conventional LUR model and DNN algorithm in estimating spatial-temporal variability of NO_2 exposure.

1. Introduction

Nitrogen dioxide (NO_2) is a kind of highly reactive gas and secondary pollutant mainly from burning fossil fuels, which were predominant species in vehicle exhaust (Brunekreef and Holgate 2002; Kampa and Castanas 2008). The high concentration of NO_2 may cause adverse health effects on respiratory system and lung disease on a global scale (Ierodiakonou et al. 2016; Wu et al. 2016).

For the purpose of controlling exposure to air pollution, previous studies had investigated about intra-urban variability in exposure to NO_2, among these studies, a few fixed monitoring stations were used to obtain air pollution concentration as an indicator to personal exposure (Ding et al. 2017; Liu et al. 2018; Zhang et al. 2019), but this method can only obtain risk value represented for an entire region which is not generally feasible for large scale health studies (Adams and Kanaroglou 2016). Hence, take spatial-temporal effects on air pollution levels into account, spatial interpolation and land-use regression based on geographic information system (GIS) were widely used as intra-urban exposure assessment methods in Asia or European (Beelen et al. 2013; Chan et al. 2009; Eeftens et al. 2012; Young et al. 2016). Spatial interpolation is a commonly used method for capturing air pollution level based on a limited number of monitoring sites, considering the distance between different emission points and using spatial statistical method to estimate the variation of air pollution on the ground. Without an information about local emission sources and land use patterns, the estimated air pollution concentrations may lead to exposure misclassification, resulting in over- or underestimate on health risk. Residents can only make less informed decisions in daily activities without getting correct health risk information. In the field of urban health and epidemiology studies, LUR combines a set of geographic sources and monitoring network as independent variables to build up multiple linear regression to estimate air pollution levels in entire study area (Achakulwisut et al. 2019; Sbihi et al. 2016). LUR has been proved having a prominent role for characterizing spatial relationships between local emissions and intra-urban air pollution variations (Michanowicz et al. 2016; Wu et al. 2018). As artificial neural network algorithm has been applied to capture non-linear relationships which present in the data by training machine learning models. This technique can predict air pollution concentration by training a group of input data, hidden layers and nodes to obtain the similar predictions as observations (Adams and Kanaroglou 2016; Liu et al. 2015).

To estimate NO_2 concentration more accurately, this study aimed to utilize a neural network-based land use regression model to assess the spatial-temporal variability. Conventional LUR approach will be used to select independent predictors variables for deep neural network developing. This integrated method was expected to have a better performance in NO_2 estimating compared with conventional LUR approach.

2. Materials and Method

2.1. Study area

Taiwan is located at the southeast of China, the geographical area of Taiwan amounts to 36,193 km^2. It comprises 14 counties and 368 townships with a total population of 23,476,640. There are about 21 million registered motor vehicles (including both motorcycles and cars) for a vehicle density of 92.9 per hundred people (MOTC, 2019). Traffic density and degrees of urbanization were in relation with intra-urban air pollution exposures (Rijnders et al. 2001). Moreover, culture-specific emission sources such as incense burning in temples and Chinese restaurants with gas cooking may emit diverse air pollutants (Lee and Wang 2004; Yu et al. 2015). Since traffic volume density is heavy and large number of temples and restaurants were densely distributed in Taiwan. The high level of NO_2 from local emission sources may lead to adverse effects to environment and human health.

2.2. Air pollutant database

Daily average NO_2 data were obtain from 2000-2016 and collected from 70 fixed air quality monitoring stations in Taiwan main island which were established by Taiwan Environment Protective Administration. The monitoring site types including 54 general stations, five traffic stations, four industrial stations, two national park stations, four background stations, and one other type stations. General stations are established to represent the ambient air condition for general residents (Fig. 1). Totally, around 0.41 million observations were collected for model development. Ozone, sulfur dioxide and PM_{10} also collected from EPA environmental resources were used as predictors variables.

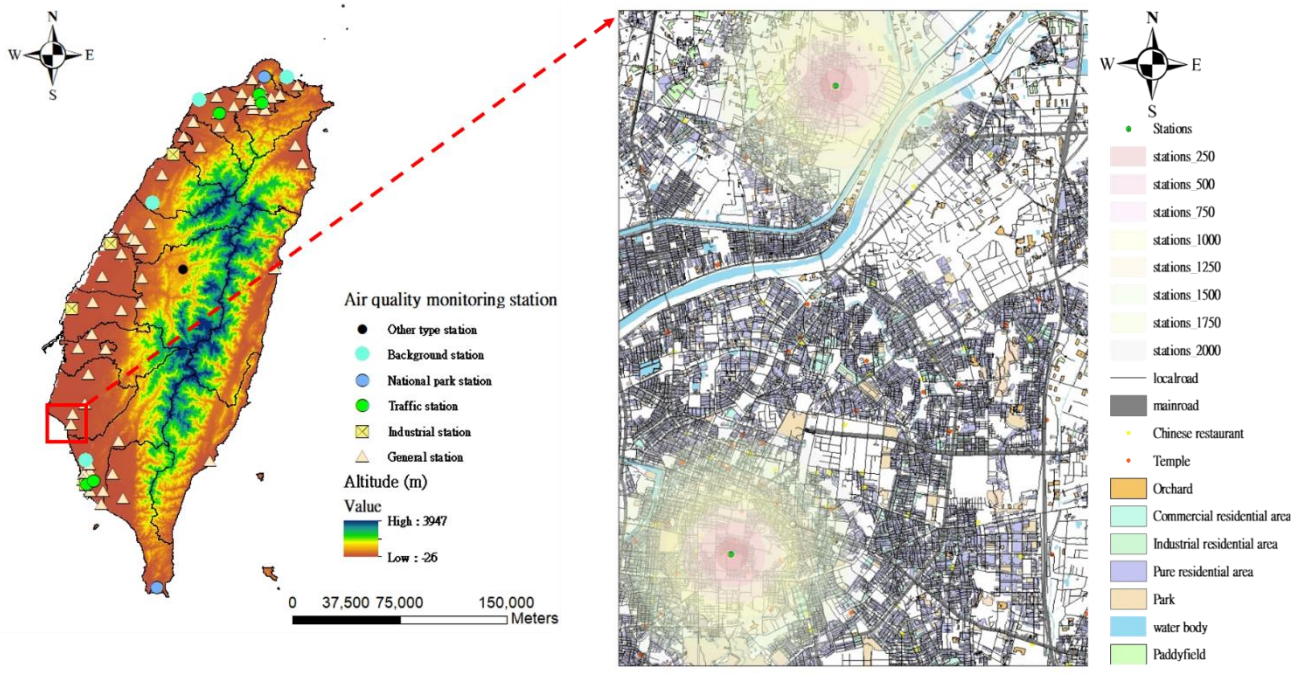

Figure 1. Monitoring stations and land-use database within the study area

2.3. Geo-spatial database

Several databases were collected for obtaining spatial predictor variables, including meteorological dataset, land-use inventory, landmark database, digital road network map, DTM, MODIS NDVI databases, and thermal power plant distribution dataset. Meteorological predictors variables including temperature, relative humidity, wind speed, wind direction, precipitation and UV were used as explanatory variables. Land-use inventory were used to derive land-use/landcover variables, including residential area, green space, water body and so on; temples, Chinese restaurants and manufactories from landmark database; road patterns from the digital road network map; and topographic altitudes of EPA monitoring sites from the Digital Terrain Model with 20 m × 20 m resolution; Normalized Difference Vegetation Index (NDVI) from NASA MODIS NDVI database with 250 m × 250 m spatial resolution. All of these geo-spatial variables were abstracted from 50 m to 5000 m circular buffer ranges surrounding each air quality monitoring station to represent the neighborhood land-use/landcover allocations.

2.4. NO_2 variation prediction using various approaches

To establish the integrated approach, conventional land-use regression (LUR) was first used to identify the important predictors variables. Then a deep neural network (DNN) algorithm was applied to fit the prediction model. After both approaches were developed, we will make a comparison to verify the difference of model performance between conventional and NN-based predictive model.

In the conventional LUR approach, air pollutant database and geo-spatial databases will be selected first, Spearman correlation coefficients were applied to assess the bivariate association between NO_2 and all the potential predictors. Only correlation coefficient of variables with a slope of the pre-specified direction was regard as the start model. The supervised stepwise multiple regression procedure was used to select variables with entered and removed p value criterion 0.1 and 0.3, respectively. With criterion following the previous paper (Beelen et al. 2013), fulfilling the increase in adjusted R^2 was more than 1%, the coefficient conformed to the predefined direction of effect and direction of effect for predictors already in the model did not change. Additionally, to assess the collinearly of variables in the developed model, Variance Inflation Factor (VIF) was used. Predictor variables with VIF criterion < 3 were included in the final model to develop conventional LUR approach.

Neural networks are excellent mathematical methods in forecasting air pollution concentration (Adams and Kanaroglou 2016; Solaiman et al. 2008). In this study, we will utilize deep neural network (DNN) algorithm integrated with LUR to generalize the linear regression model for estimating daily NO_2 concentration. Totally, around 0.41 million observations were collected for developing DNN model. For model establishing, around 80% of the data were for model training and 20% for model testing which were applied

following all fitting processes. During processing the DNN algorithm, 3 hidden layers were used. The weights in a neural network model, which map the data to the hidden layer neurons, are partially analogous to the coefficients in a regression model. Learning rate formula of the model was LearnRate_value * 1 /(1 + decay_value * epoch_value). Epoch value of model development was set for 1000, during each epoch descending gradient method was used for minimizing the error. To optimize the model performance, package 'Adam' was used for optimizing the model and the loss function package 'mean_absolute_error' was used for identifying the mean absolute error (MAE) between model predictions and observations within the model.

After both conventional and NN-based LUR approaches has developed, R^2 and root-mean-square error (RMSE) were used to determine the model predictive abilities and the residuals between predictions and observations. Based on these two parameters, this study compared conventional LUR approach with NN-based LUR model. In addition, 10-fold cross validation method was further used to confirm the reliability and robustness of model performance. Land use/landcover were extracted by ArcGIS 10.5. Abovementioned approaches were analyzed using SPSS 22.0 and R 3.5.2.

3. Results and Discussions

3.1. Descriptive statistics of measured NO_2 concentrations

NO_2 concentration yearly average shows a decreasing trend during the entire study period. Fig.2 shows that general stations, traffic stations, industrial stations, national park stations, background stations, and other type stations had different NO_2 levels. Traffic station had the highest value of NO_2 concentration (29.20 ± 12.01 ppb), national park station had the lowest (3.42 ± 2.39 ppb), and other station types had similar level of NO_2 among the study period. Comparing with Taiwan EPA's air quality standards of NO_2 (50 ppb) and WHO ambient air quality standard of NO_2 (40 μg/m^3), the measured NO_2 level of the traffic station was under the two official criterions.

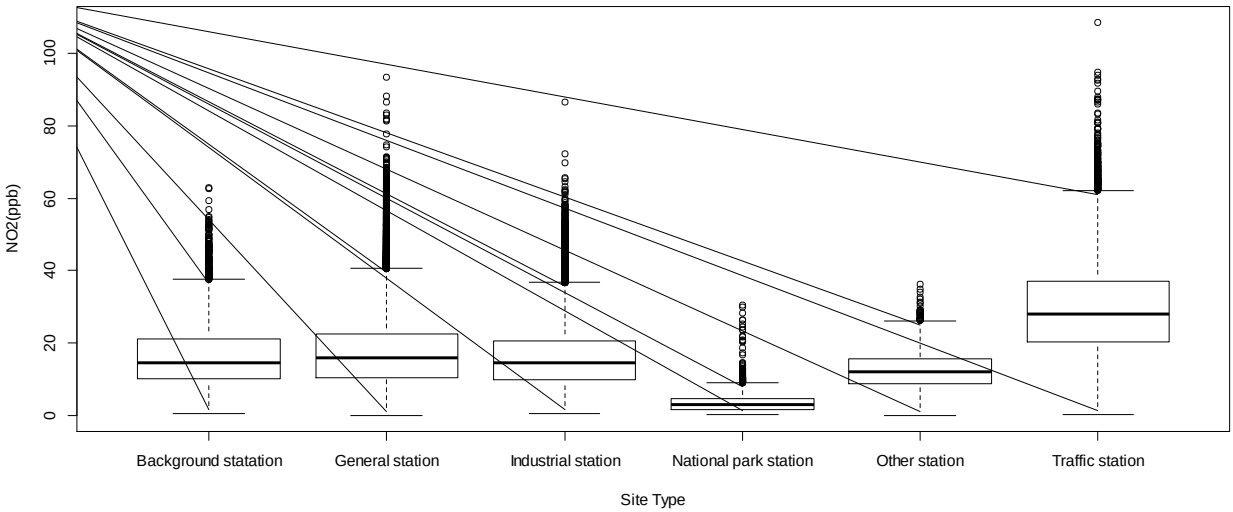

Figure 2. NO_2 concentration within different air monitoring site types

3.2. Comparison of conventional LUR model and NN-based LUR

Table1 lists the coefficient of predictors variables selected in the conventional LUR and NN-based LUR approach. R^2, adjusted R^2, RMSE were used to present model performance. The higher the R^2 and adjusted R^2 indicate the better model prediction ability, and the lower the value of RMSE indicates the residuals comes smaller. Results showed that 11 variables were selected by the stepwise variable selection procedure, PM_{10}, SO_2, O_3 explained 18%, 7% and 5% NO_2 variation within the model. R^2, adjusted R^2 and RMSE of conventional LUR approach were 0.60, 0.60 and 6.56, respectively. And R^2 and RMSE of NN-based LUR approach was 0.85 and 4.13. As for checking model robustness, 10-fold cross validation was utilized within both approaches and captured about 61% and 83% NO_2 variation. This verifies the reliability and robustness of the constructed model. Further comparing the results between the two model development approaches, the NN-based LUR model consistently performed better than the conventional LUR model in all cases, again showing the improvement by integrating neural network algorithm method into LUR modelling procedures.

Table 1. Coefficient estimates of the developed LUR model, and the comparison of model performance using different approaches

	Coefficient (95% CI)	Conventional LUR approach	NN-based LUR approach
Intercept	17.735 (17.6 - 17.87)	R^2: 0.60 Adjusted R^2: 0.60 RMSE: 6.56 10-fold CV R^2: 0.61	R^2: 0.85 RMSE: 4.13 10-fold CV R^2: 0.83
SO_2	0.855 (0.845 - 0.865)		
O_3	-0.138 (-0.14 - -0.136)		
PM_{10}	0.13 (0.129 - 0.131)		
Temple150m	0.217 (0.212 - 0.222)		
Wind speed	-1.562 (-1.581 - -1.544)		
Temperature	-0.396 (-0.4 - -0.392)		
$Manufactory_{5000m}$	0.015 (0.015 - 0.015)		
$Funeral\ industry_{4000m}$	0.037 (0.036 - 0.038)		
Industrial and commercial residential $area_{4000m}$	0.108 (0.107 - 0.109)		
$Majorroad_{50m}$	301.188 (296.6 - 305.775)		
Distance to the nearest bus station	-0.00012 (-0.00012 - -0.00011)		

Several improvement ways were proposed and used in previous studies. For example, regression kriging (RK) was applied to assess the spatial distributions of NO_2 and O_3 in Japan (Araki et al. 2015). In their approach, the residuals of the developed land-use regression model were interpolated using ordinary kriging and used to adjust the pollutant estimations obtained from LUR models. In the case the distribution of residuals cannot be explained by the spatial autocorrelation models of kriging, the applicability of RK might be limited (Hengl et al. 2007). The proposed DNN/LUR approach did not face the same challenges since DNN provided an efficient methodology to improve the prediction performance for a non-linear distribution data characteristic. The other studies have considered remote sensing estimates such as aerosol optical depth (AOD) and AOD based $PM_{2.5}$ estimates as the variable of LUR model (Beckerman et al. 2013; Hystad et al. 2011; Yang et al. 2017). In Taiwan, it is not easy to acquire clear satellite images because of the cloudy and rainy weather conditions especially in summer (Wu et al. 2013). Most of the western part of the island are missing observations on the AOD image. Weather conditions limits the applicability of satellite estimates on exposure assessment not only in Taiwan, but also in the other tropical/subtropical regions with similar weather characteristics.

The burning of joss paper and incense in temples and funeral industry are activities particular to Asian communities. (Lung et al. 2014) evaluated the influence of various spot pollution sources on exposure levels within communities, such as temples. They found that PM levels at locations near spot pollution sources could be increased by 3.5 to 4.9 times compared with those at background locations. (Lung and Kao. 2003) assessed the $PM_{2.5}$ exposures of temple worshippers in Taiwan. Their results found that the geometric mean of personal exposure was 444 $\mu g/m^3$. The later was approximately 4 to 6 times that of roadside concentrations. In our study, both temples and funeral industry were statistically associated with NO_2 concentrations and included in the developed model, our study provides insight into NO_2 exposure predictions in studies of Asian community.

4. Conclusion

This study demonstrates the value of incorporating the conventional LUR model and DNN algorithm in estimating spatial-temporal variability of NO_2 exposure. Model validation shows the robustness of DNN approach in capturing daily concentration. This method can be used to predict air pollution levels within-city in different areas.

References

Achakulwisut P, Brauer M, Hystad P, Anenberg SC. 2019. Global, national, and urban burdens of paediatric asthma incidence attributable to ambient no2 pollution: Estimates from global datasets. The Lancet Planetary Health 3:e166-e178.

Adams MD, Kanaroglou PS. 2016. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. Journal of environmental management 168:133-141.

Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli X, et al. 2013. Development of no2 and nox land use regression models for estimating air pollution exposure in 36 study areas in europe–the escape project. Atmospheric Environment 72:10-23.

Brunekreef B, Holgate ST. 2002. Air pollution and health. The Lancet 360:1233-1242.

Chan TC, Chen ML, Lin IF, Lee CH, Chiang PH, Wang DW, et al. 2009. Spatiotemporal analysis of air pollution and asthma patient visits in taipei, taiwan. International Journal of Health Geographics 8:26.

Ding L, Zhu D, Peng D, Zhao Y. 2017. Air pollution and asthma attacks in children: A case–crossover analysis in the city of chongqing, china. Environmental pollution 220:348-353.

Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, et al. 2012. Development of land use regression models for pm2. 5, pm2. 5 absorbance, pm10 and pmcoarse in 20 european study areas; results of the escape project. Environmental science technology 46:11195-11205.

Ierodiakonou D, Zanobetti A, Coull BA, Melly S, Postma DS, Boezen HM, et al. 2016. Ambient air pollution, lung function, and airway responsiveness in asthmatic children. Journal of Allergy Clinical Immunology 137:390-399.

Kampa M, Castanas E. 2008. Human health effects of air pollution. Environmental pollution 151:362-367.

Lee SC, Wang B. 2004. Characteristics of emissions of air pollutants from burning of incense in a large environmental chamber. Atmospheric Environment 38:941-951.

Liu W, Li X, Chen Z, Zeng G, León T, Liang J, et al. 2015. Land use regression models coupled with meteorology to model spatial and temporal variability of no2 and pm10 in changsha, china. Atmospheric Environment 116:272-280.

Liu Y, Chen S, Xu J, Liu X, Wu Y, Zhou L, et al. 2018. The association between air pollution and outpatient and inpatient visits in shenzhen, china. International journal of environmental research public health 15:178.

Michanowicz DR, Shmool JL, Cambal L, Tunno BJ, Gillooly S, Hunt MJO, et al. 2016. A hybrid land use regression/line-source dispersion model for predicting intra-urban no2. Transportation Research Part D: Transport Environment 43:181-191.

Rijnders E, Janssen N, Van Vliet P, Brunekreef B. 2001. Personal and outdoor nitrogen dioxide concentrations in relation to degree of urbanization and traffic density. Environmental health perspectives 109:411-417.

Sbihi H, Tamburic L, Koehoorn M, Brauer M. 2016. Perinatal air pollution exposure and development of asthma from birth to age 10 years. European Respiratory Journal 47:1062-1071.

Solaiman T, Coulibaly P, Kanaroglou P. 2008. Ground-level ozone forecasting using data-driven methods. Air Quality, Atmosphere Health 1:179-193.

Wu CD, Zeng YT, Lung SCC. 2018. A hybrid kriging/land-use regression model to assess pm2.

5 spatial-temporal variability. Science of The Total Environment 645:1456-1464.
Wu S, Ni Y, Li H, Pan L, Yang D, Baccarelli AA, et al. 2016. Short-term exposure to high ambient air pollution increases airway inflammation and respiratory symptoms in chronic obstructive pulmonary disease patients in beijing, china. Environment international 94:76-82.
Young MT, Bechle MJ, Sampson PD, Szpiro AA, Marshall JD, Sheppard L, et al. 2016. Satellite-based no2 and model validation in a national prediction model based on universal kriging and land-use regression. Environmental science technology 50:3686-3694.
Yu KP, Yang KR, Chen YC, Gong JY, Chen YP, Shih HC, et al. 2015. Indoor air pollution from gas cooking in five taiwanese families. Building Environment 93:258-266.
Zhang Y, Ni H, Bai L, Cheng Q, Zhang H, Wang S, et al. 2019. The short-term association between air pollution and childhood asthma hospital admissions in urban areas of hefei city in china: A time-series study. Environmental research 169:510-516.