COMPARISON OF LUR-BASED AND ANN-BASED $PM_{2.5}$ CONCENTRATION ESTIMATION OVER TAIPEI METROPOLIS

Dewinta Heriza (1), Chao-Hung Lin (1), Chih-Da Wu (2)

[1] Geospatial Artificial Intelligence Laboratory, Geomatics Department, National Cheng Kung University, 1 University Road, 70101, Tainan City, Taiwan, R.O.C.
[2] Geomatics and Environmental Health Laboratory, Geomatics Department, National Cheng Kung University, 1 University Road, 70101, Tainan City, Taiwan, R.O.C.
Email: dewintaheriza@gmail.com; linhung@mail.ncku.edu.tw; chidawu@mail.ncku.edu.tw

KEY WORDS: Fine Particulate matter, Land Use-Regression, Artificial Neural Networks.

ABSTRACT: Fine particulate matter ($PM_{2.5}$) is an air pollutant that has been becoming one of the major environmental issues in national governments. Air quality monitoring and prediction are thus necessary for management and control. In previous studies, a land-use regression (LUR) model with several factors such as chemical particles, meteorological information, greenness environments, and landmarks combined with interpolation techniques is used to predict $PM_{2.5}$ concentrations using data from Taipei metropolis, which exhibits typical Asian city characteristics. Recently, a lot of attention was paid to the improvement of methods which are used to predict air quality especially $PM_{2.5}$. This study proposes utilizing artificial neural networks to predict $PM_{2.5}$ concentrations and the built $PM_{2.5}$ prediction model is compared with that using LUR. To obtain the resulted, cross-validation is adopted in the proposed method. 17 air quality monitoring stations established by environmental protection administration of Taiwan with annual average $PM_{2.5}$ concentrations from 2006-2012 were used for model development. In experiments, quantitative accuracy assessments were conducted to evaluate the performance of proposed methods, in term of determination coefficients (R^2) and root means square error (RMSE), compared with LUR-based method. The result show that LUR-based still perform better than ANN-based. The R^2 of LUR-based was 0.9 while the R^2 ANN-based was 0.8.

1. INTRODUCTION

Air pollution is one of the most pressing environmental problems and the critical challenges facing modern society. The effects of air pollution are disrupted health (Janssen et al., 2013), environmental degradations (Munir et al., 2017), and visibility (Environmental Protection Agency (EPA) of US). The most dangerous and most frequently studied air pollution is particulate matter (PM). There are two types of PM, there are fine particles and coarse particles where represented as PM2.5 and PM10 particles, respectively (Xu et al., 2017). Fine particles ($PM_{2.5}$) consist of primary and secondary anthropogenic combustion products derived from traffic and energy production (Li et al., 2004). Coarse particles ($PM_{2.5-10}$) mainly result from natural processing, such as resuscitation of local soils, dust storms, and anthropogenic sources such as road dust and various industrial processes (Querol et al., 2004). Compared to PM10, $PM_{2.5}$ gets more attention because of its smaller size, longer atmospheric life and worse health risks (Dominici et al., 2014).

In Taiwan (Republic of China), air pollution is created both domestically and from China (the People's Republic of China), although the majority comes from domestic sources rather than across borders. Taiwan's topography has been noted as a factor in the problem of air pollution, which causes poor pollution and traps pollutants.

Air quality in Taipei is generally in global standards, with an Air Quality Index (AQI) value of 100 or less on a 500-point scale. In system color coding, it is routinely measured as green (good) or yellow (in a healthy range). But elsewhere along the western plains of Taiwan - as far north as Taoyuan and

even parts of New Taipei City - pollution levels are often a serious problem, falling into the red zone or worse. Air quality is a very local phenomenon and can fluctuate significantly throughout the year and even throughout the day, depending on the weather. A quick inspection of the Taiwan EPA website throughout the day on January 31, 2018 revealed significant fluctuations. Taipei City reaches orange levels in some areas during the day, while Taoyuan stays red all day (international.thenewslens.com). This indicates the air pollution are necessarily monitored and estimated, especially $PM_{2.5}$ particles.

Previous studies have successfully predicted $PM_{2.5}$ observations by means of linear regression with several factors, such as chemical particles (Tunno et al., 2016), green environment (Shi et al., 2017), and meteorological information (Masoudi et al., 2018). Further, Wu et al. (2018) combined the use of land use linear regression and Kriging interpolation in predicting the $PM_{2.5}$ particles over Taiwan area by considering surrounding landmarks, such as temple, Chinese restaurant, parks, industrial area, etc. There are papers devoted to prediction of air pollution level based on artificial neural networks. Artificial neural networks (ANNs) belong to the group of statistical models. It is means that worthwhile to remarks the possibilities using artificial neural networks to predict and modelling air quality. In this paper, we predict $PM_{2.5}$ using LUR-based and ANN-based and compare which one that better to predict air quality in Taipei Metropolis.

2. MATERIALS AND METHODS

The study area was Taipei metropolis, which consist of Taipei City and New Taipei City. This region compromises 41 townships with population density is 11848 people/km2, which means that more than 25% of Taiwan's population lives in this are (DGB, 2011; Wu et al., 2017). Government reported in 2009, 13.2% of Taipei Metropolis is covered by building and 68.3% is covered by forest (NLMSC, 2009). In addition to that, there are 4945 thousand and 7027 vehicle/km2 of the registered motor vehicle number (both motorbike and car) and vehicle density in the region (MOTC, 2016). There are as many as 1173 Buddhist and Taoist temples throughout Taipei metropolis, ranging in size from single room shrines to huge multi-story buildings (CRGIS, 2016). And more than thirty-two thousand restaurants are distributed within the region, with almost fourteen restaurants/km2 (CGIS, 2016). Religious and cooking emissions under such a high temple and restaurant density contributes significantly to urban air pollution (Yu et al., 2015; Kuo et al., 2015).

Table 1. Details of spatial database over Taipei Metropolis

GIS Database	Type
National land-use inventory	Pure residential, commercial residential, and industrial-commercial residential
Map of industrial park	Industrial parks in year of 2010
Landmark database	0.25 million landmark points including Chinese restaurant (night market included), temple, and any other landmarks
Digital road network	Local roads, major roads, and express ways

2.1 Data gathering and refinement

Four categorizes of databases are utilized to LUR modelling and Artificial Neural Network for $PM_{2.5}$ predictions over Taipei Metropolis. Those are chemical particle database, meteorological database, spatial database, and greenness environment database. This subsection covers the process of collecting data as well as their refinements.

The observations of chemical and meteorological databases were obtained from 17 automatics monitoring stations established by Environmental Protection Administration (EPA) of Taiwan. The stations distributed within the study area, including 14 general station, two traffic stations, and one

national park stations. Daily observations from 2006 to 2012 were aggregated into annual averages with non-data and outlier removals are applied.

GIS databases was collected as potential predictors for model development. Four GIS databases were used, including the national land-use inventory, map of industrial park, landmark database, and digital road network map. Table 1 defines the details of this database. In this study, a "Chinese restaurant" is defined as an establishment that serves Chinese/Taiwanese cuisine, local foods, or Chinese caterer with stir-frying as the primarily cooking style. Night markets, with Chinese restaurants gathering, were also included for GIS Database. The spatial distributions of temples, Chinese restaurants, and transportation related landmarks, including auto detailing, parking lots, gas stations, but stops, and bus stations were extracted from the landmark database of 2006, 2008, and 2010 for database.

As for the remote sensing database, the amount of "greenness" (e.g. trees and vegetation) was obtained from NASA's Earth Observing System data - the global Moderate Resolution Imaging Spector-radiometer (MODIS) NDVI database. With a 16 days of temporal resolution, NDVI images were obtained monthly from 2006 to 2012 where there were two NDVI measurements for each cell in every month. Images with the acquisition date closer to mid-month (the fifteenth) were collected. NDVI images from each month were then aggregated to annual average.

2.2 Model development (LUR) and accuracy analysis

A LUR model was then developed using a stepwise procedure modified from Kerckhoffs et al. (2015) and Lee et al. (2015). A stepwise factor selection was utilized to extract the factors which have a high-significance effects to the $PM_{2.5}$ concentration prediction. In the pre-processing step, we removed factors with negative correlation to $PM_{2.5}$ concentrations and kept the factors with opposite condition. The factors with positive correlation and the interpolation result of $PM_{2.5}$ concentrations by modified IDW in all stations were further included into the stepwise factor selection. First, we defined an initial model which consisted only the $PM_{2.5}$ concentrations by modified IDW. Then, add a factor with the highest p-value and find the direction effect matching; if the direction is incorrect, then remove the factor and go to the first step. If the direction is right, perform the significance test; if the significance is larger than 0.1, removed the factor and go the first step. If the significance is lower than 0.1, then calculate the VIF to detect the redundancy; if VIF is larger than 3, then remove factor and go to first step. If smaller than 3, then update the initial model and go to the first step.

Variables with VIF < 3 were included to establish the final model. The determination of coefficient (R2), adjusted R2, and Root Mean Square Error (RMSE) were used to assess the model performance and accuracy. The final equation resulting from the stepwise regression procedures is given as follows:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \qquad (1)$$

where Y is $PM_{2.5}$ concentrations; b_0 is constant intercept; b_1 to b_n are regression coefficients; and X_1 … Xn are potential predictors. Two methodologies were employed for model reliability assessment. First, the developed model was evaluated with a 10- fold cross-validation methodology (Wang et al., 2014; Kerckhoffset al., 2015). Of the air quality monitoring sites, 90% were randomly selected for model development and the remaining 10% were used for model validation. This procedure was repeated ten times, such that each monitoring site was in a test set once. The R^2, adjusted R^2, and RMSE of the tests were recorded and compared with those obtained from the final model for evaluating the goodness of fit and robustness of the model. In the second methodology, data collected from 2013 were treated as out-of-sample data. The $PM_{2.5}$ concentrations of the out-of-sample observations from 2013 were estimated and then compared with the known observations to derive the accuracy of the external verification.

2.3 Artificial Neural Networks

Recently, artificial neural networks are intensive development of algorithm that can be used to solve problem in the environmental problem, especially air quality. This feature cause ANNs have complex unknown relationship between the variables. In environmental protections, they can be used providing missing data from environmental protections, they can be used to providing missing data from environmental monitoring, predicting air and water pollutions levels and sound levels, automatic image analysis and interpretation of biological monitoring result, environmental assessment, determination of ore lithological composition and many other issues (Haupt et al. 2009, Kwiechen and Pawul 2012, Krawczykowski et al. 2009, singh et al. 2009, Szczepanska and Kmiecik 2001, tadeusiewiccz and Doborowolski 2004)).

There are many types of the artificial neural networks which differ in structrure and principle of operation. In this research, using the fully connected feedforward networks known as multi layerpercepton (MLP). Artificial neural netwoks model are mathematical models inspired by the functioning of nervous system (Gardner and Dorling, 1998; Cobourn et al., 2000; Agirre-Basurko et al., 2006), which are composed by a number of interconnected entities.

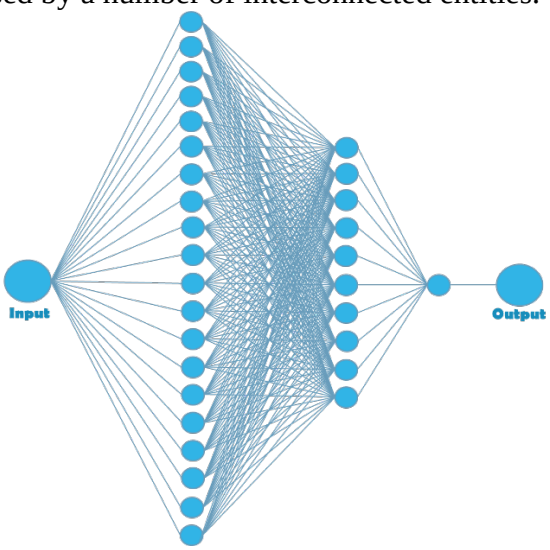

Figure 1. Neural Networks Architecture

2.3.1 Input data for ANN training and target data

The ANN input data sets consists of all the variables from data gathering such as chemical meteorological, spatial database and greenness. Total of the input data for ANN training are 143 variables from 2006 to 2012. In this study, 2 hidden layers have been used with 10 neurons in each layer. Algorithm for this training is using random data division with Levenberg-Marquardt training with epoch progress 105 iterations and validation check was 100. The target data we consider daily measurement of $PM_{2.5}$ concentrations in the Taipei Metropolis with 17 automatics stations. The data were recorded from 2006 to 2012 by (EPA) Environmental Protections Administration.

3. RESULT

3.1 LUR development

In LUR-based stepwise regression will be used at the pre-processing step, 143 variables are considered in total. Correlation coefficient between each variables and $PM_{2.5}$ observations from EPA are calculated. Those which have positive correlation and p-value less than 0.2 will be considers in

the main step of the stepwise factor selection. With the right association and high significance, six variables have survived from the stepwise factor selection where their details are shown in Table 2. However, several of those have a high VIF value which indicates a redundancy between those factors. Thus, those factors are removed one by one so that all final factors have VIF value less than 3. Table 2 lists the details of those factors which further be called as predictors of $PM_{2.5}$ concentrations prediction.

Table 2. coefficient estimate of the develop LUR model

Variable	β	VIF	P-Value	R^2	Adj R^2	RMSE
(Intercept)	19.3	-	<0.01	0.9	0.89	1.66
NOx	0.08	1.80	<0.01			
SO_2	1.64	2.17	<0.01			
LR_750[a]	$7.81x10^{-5}$	1.65	0.01			
CR_1750[b]	0.01-1.86	1.87	<0.01			
Temple_750[c]	0.24	2.93	0.06			
NDVI_1750[d]	-10.5	1.12	<0.01			

a Length of local roads within a 750 m circular buffer.
b number of Chinese restaurants within a 1750 m circular buffer.
c Number of temples within a 750 m circular buffer.
d Average NDVI within a 1750 m circular buffer.

Where $PM_{2.5}$ is $PM_{2.5}$ mass concentration (mg/m3); $NDVI_{1750m}$ is the average NDVI within a 1750 m circular buffer; NO_X is the concentration of NOx (ppm); CR_{1750m} is the number of Chinese restaurants within a 1750 m circular buffer; SO_2 is the concentration of SO_2 (ppb); LR_{750m} is the length of local roads within a 750 m circular buffer (m); $Temple_{750m}$ is the number of temples within a 750 m circular buffer. The regression coefficients for CR_{1750m} and $Temples_{750m}$ are 0.01 and 0.24, respectively, indicating that for every additional Chinese restaurant and temple within the specific buffer ranges, $PM_{2.5}$ concentrations is expected to increase by an average of 0.01 mg/m3 and 0.24 mg/m3, respectively. The adjusted R^2 of the 0.89 indicates that the development model had a high explanatory power for $PM_{2.5}$ variations. NDVI had a negative association with $PM_{2.5}$ revealing that can mitigate neighbourhood the $PM_{2.5}$ concentrations magnitude. By the variables can be survive the resultant LUR model is shown as the following equations:

$$PM_{2.5} = 19.3 + 0.08NO_x + 1.64SO_2 + 138LR_{750} + 0.01CR_{1750} + 0.24Temple_{750} - 10.5NDVI_{1750} \quad (2)$$

10-Fold cross validation was applied to verify the developed LUR model. For each test, eight to nine (10%) sample data were employed to determine the model accuracy, as shown in Fig. 2. The results reveal that the average accuracy from 2006 to 2012 was over 0.9 (averaged R2 ¼ 0.91; averaged adjusted R2 ¼ 0.90), and the RMSE is very similar to that obtained from the developed model (averaged CV RMSE ¼ 1.58; model RMSE ¼ 1.66). Overall this indicates that the developed model is able to provide highly accurate PM2.5 predictions. Fig. 3 shows the comparison between model predictions and on-site observations from 2013. The obtained external validated R2 of 0.83 again confirms the reliability and robustness of the constructed model.

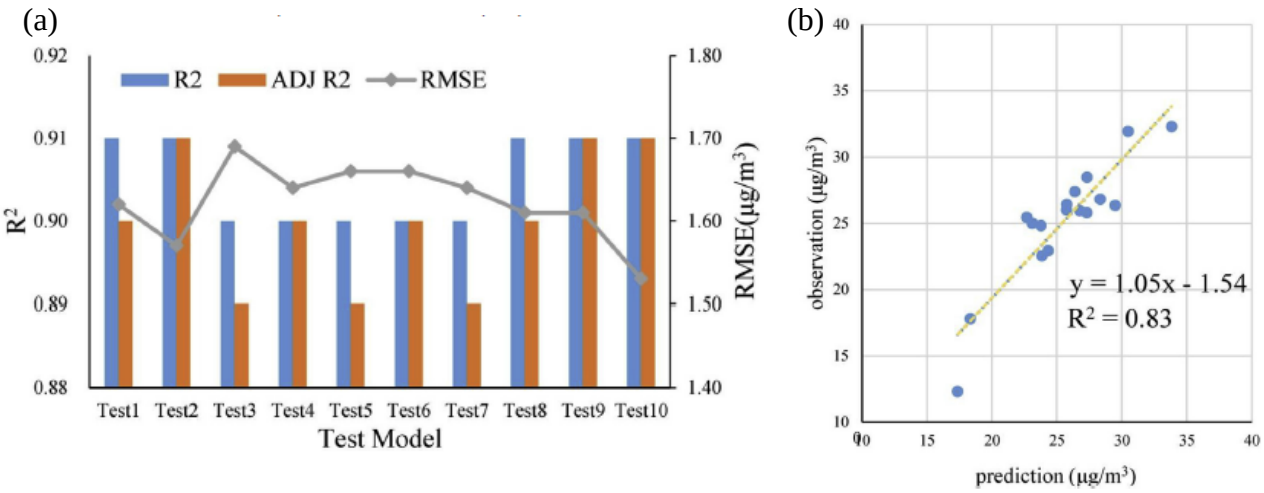

Figure 2. Result of 10-fold cross validation and observations regressed against predictions

3.2 Artificial neural networks development

From the above processing, results were obtained on the mean squared error (MSE) was 2.1209 at epoch 5 and the root mean squared (RMSE) was 1.4563. For the plot regression (Training: R=0.90428, Validation: R=0.94219, Test: R=0.94663) fig 3 shown the result of the processing using artificial neural networks model with two hidden layer and 10 neurons in each layer.

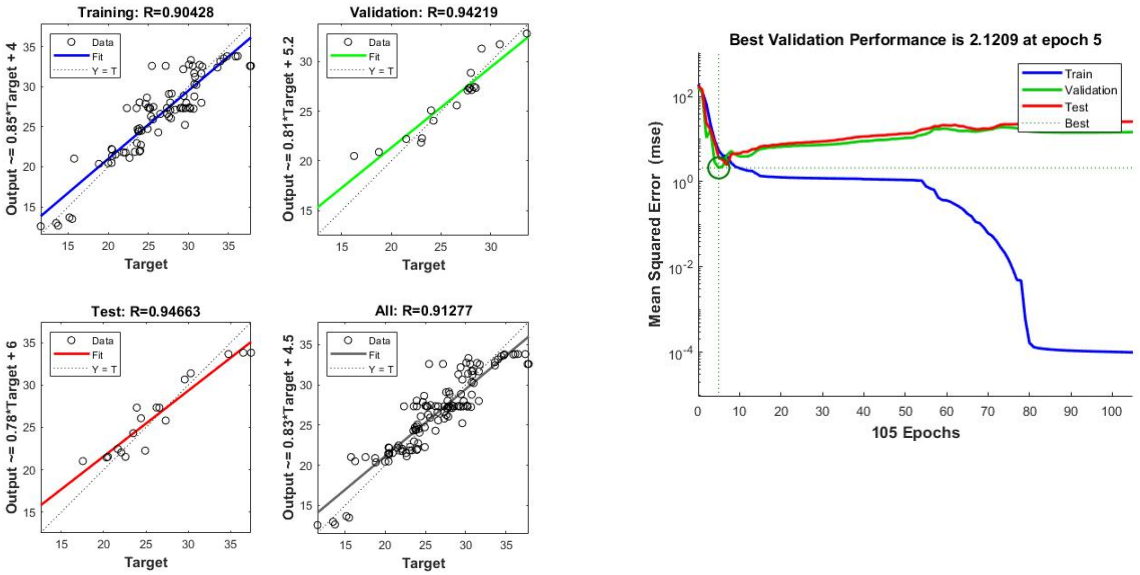

Figure 3. Result of artificial neural networks

3.3 Discussion

In this study, we present the result to demonstrate the effectiveness of LUR-based and ANN-based for solving the predictions of fine particulate matter concentrations levels. The input data for all the proposed method consist of meteorological data with daily average from 17 stations from 2006 to 2012. Result from LUR shown the RMSE was 1.66 and the RMSE from ANNs was 1.45 the result each methodology is significant. RMSE-LUR-based is higher than RMSE-ANN-based. The R squared from each method shown the different result too, R squared from ANN was 0.82 and LUR

was 0.9. It is important to notice that for all experiment the data are randomly partitioned into different validation. Land use regression used cross validation with 10-fold validation for training and testing data, and artificial neural networks used randomly training and testing data.

Several studies have applied LUR to assess fine particulate pollutants in Asia (Liu et al., 2015; Lee et al., 2015; Wu et al., 2015). Most model have been driven by significant predictors and achieve good models with the adjusted R2 was higher than 0.5. several studies have applied ANN for prediction air pollutant level because the construction a simple and effective tool for air quality. Comparing between two method had a good experience because the result shows the advantage and the disadvantage for each other. Based on the result and the effectiveness, land use regression still has a better result even LUR method has a long processing and consume time. It can be advantage for LUR-based. ANN-based has a good result too even in this study the result from LUR based in higher.

4. CONCLUSION

The main purpose of this study is to compare ANN-Based and LUR-based to predict fine particulate matter. Instead of using an LUR which has a better result than ANN- based but ANN-based need to do further assessment. In the future, when we already have a bigger data for air quality, is not possible the best method to predict air quality is using artificial neural networks. Proper selection of input and output data with clear dependence between them is necessary to get good result.

REFERENCE

DGB (Directorate General of Budget), 2017. National Statistics-current Index. Available. http://www.stat.gov.tw/point.asp?index=4.

Dominici, F., Greenstone, M. and Sunstein, C.R. (2014). Particulate matter matters. Science 344: 257–259.

EPA (Environmental Protection Agency). Health and environmental effects of particulate matter (PM).

Ferry, T. (2018). What will it take to improve Taiwan's air? https://international.thenewslens.com/article/90010.

Guerra, S.A., Lane, D.D., Marotz, G.A., Carter, R.E. and Hohl, C.M. (2006). Effects of wind direction on coarse and fine particulate matter concentrations in Southeast Kansas. Air & Waste Manage. 56:1525.

Janssen, N.A.H., Fischer, P., Marra, M., Ameling, C. and Cassee, F.R. (2013). Short-term effects of PM2.5, PM10 and PM2.5-10 on daily mortality in the Netherlands. Science of The Total Environment. 463-464:20-26.

Kuo, S.C., Tsai, Y.I., Sopajaree, K., 2015. Emission identification and health risk potential of allergy-causing fragrant substances in PM2.5from incense burning. Build. Environ. 87, 23–33.https://doi.org/10.1016/j.buildenv.2015.01.012.

Li, Z., Hopke, P.K., Husain, L., Qureshi, S., Dutkiewicz, V.A., Schwab, J.J., Drewnick, F. and Demerjian, K.L. (2004). Sources of fine particle composition in New York city. Atmos. Environ. 38: 6521–6529.

Munir, S., Habeebullah, T.M., Mohammed, A.M.F., Morsy, E.A., Rehan, M. and Ali, K. (2017). Analysing PM2.5 and its association with PM10 and meteorology in the arid climate of Makkah, Saudi Arabia. Aerosol and Air Quality Research. 17: 453-464.

MOTC (Ministry of Transportation and Communications), 2017. Vehicle Statistics. Available.http://www.motc.gov.tw/ch/home.jsp?id=6&parentpath=0.

Querol, X., Alastuey, A., Ruiz, C.R., Artiñano, B., Hansson, H.C., Harrison, R.M., Buringh, E., Ten Brink, H.M., Lutz, M., Bruckmann, P., Straehl, P. and Schneider, J. (2004). Speciation and origin of PM10 and PM2.5 in selected European cities. Atmos. Environ. 38: 6547–6555.

Tunno, B.J., Dalton, R., Cambal, L., Holguin, F., Lioy, P. and Clougherty, J.E. (2016). Indoor source apportionment in urban communities near industrial sites. Atmospheric Environment. 139:30-36.

Wu, C.D., Zeng, Y.T. and Lung, S.C.C. (2018). A hybrid kriging/land-use regression model to assess PM2.5 spatial-temporal variability. Science of the Total Environment. 645:1456-1464.

Xu, G., Jiao, L., Zhang, B., Zhao, S., Yuan, M., Gu, Y., Liu, J. and Tang, X. (2017). Spatial and temporal variability of the PM2.5/PM10 ration in wuhan, Central China. Aerosol and Air Quality Research. 17: 741-751.

Yu, K.P., Yang, K.R., Chen, Y.C., Gong, J.Y., Chen, Y.P., Shih, H.C., Lung, S.-C.C., 2015. Indoor air pollution from gas cooking in five Taiwanese families. Build. Environ. 93, 258–266. https://doi.org/10.1016/j.buildenv.2015.06.024.

Zhang, B., Jiao, L., Xu, G., Zhao, S., Tang, X., Zhou, Y. and Gong, C. (2018). Influences of wind and precipitation on different-sized particulate matter concentrations (PM2.5, PM10, PM2.5-10). Meteorology and Atmospheric Physics. 130:383-392.