TEHRAN AIR POLLUTANTS PREDICTION BASED ON RANDOM FOREST FEATURE SELECTION METHOD

A.Shamsoddini [*a], M.R.Aboodi [b], J.Karami [c]

[a] Department of RS and GIS, Tarbiat Modares University, Tehran, Iran- ali.shamsoddini@modares.ac.ir

[b] Department of RS and GIS, Tarbiat Modares University, Tehran, Iran-mohammadreza.aboodi@gmail.com

[c] Department of RS and GIS, Tarbiat Modares University, Tehran, Iran-jl.karami@modares.ac.ir

KEY WORDS: AIR POLLUTION, RANDOM FOREST FEATURE SELECTION, ARTIFICIAL NEURAL NETWORKS, MULTIPLE-LINEAR REGRESSION, HUMAN HEALTH, TEHRAN

ABSTRACT

Air pollution as one of the most serious forms of environmental pollutions poses huge threat to human life. Air pollution leads to environmental instability, and has harmful and undesirable effects on the environment. Modern prediction methods of the pollutant concentration are able to improve decision making and provide appropriate solutions. This study examines the performance of the Random Forest feature selection in combination with multiple-linear regression and Multilayer Perceptron Artificial Neural Networks methods, in order to achieve an efficient model to estimate carbon monoxide and nitrogen dioxide, sulfur dioxide and $PM_{2.5}$ contents in the air. The results indicated that Artificial Neural Networks fed by the attributes selected by Random Forest feature selection method performed more accurate than other models for the modeling of all pollutants. The estimation accuracy of sulfur dioxide emissions was lower than the other air contaminants whereas the nitrogen dioxide was predicted more accurate than the other pollutants.

1- INTRODUCTION

Air quality is considered as an important factor determining quality of life and public health in urban areas, especially in densely populated areas. Air pollution possesses the most serious threat to human health as far as the environmental problems are concerned. Air pollution leads to environmental instability, and has harmful and undesirable effects on the environment (Akbari et al, 2015&Valverde et al, 2015). Urbanization, population growth, industrial expansion, increased consumption of fossil fuels along with the low quality of fuels, lack of efficient transport systems, and traffic congestion have led to a daily discharge of large amounts of pollutants, which are incompatible with the natural mechanisms, into the air (Antanasijević et al, 2013). Monitoring and evaluation of various emissions and their sources, the implementation of standards and practical strategies of pollution reduction are necessary to solve this problem. In this regard, the use of modern methods to predict the concentration of pollutants can improve decision making and provide appropriate solutions. During last decades, several studies have been conducted to predict spatial-temporal concentration of pollutants in the air. Perez and Trier (2001) showed that Artificial Neural Network (ANN) model performed better than multiple-linear regression model for prediction of the nitrogen monoxide (NO) and nitrogen dioxide (NO_2) concentration (Perez et al, 2011). These results were confirmed by Grivas and Chaloulakou (2006) where PM_{10} was predicted using multi-layer perceptron (MLP) ANN fed by all data as input at all stations with the index of agreement between 0.8-0.89 (Grivas et al, 2006). Antanasijević et al (2013) used ANN model fed by the attributes selected through genetic algorithm for PM_{10} prediction. The predicted mean absolute error was 10 percent which was three times better than that derived from multiple-linear regression and principal component regression (Antanasijević et al, 2013). According to the studies mentioned, in recent years, statistical models especially ANN have been commonly applied for predicting air quality. As far as air pollution prediction is concerned, selecting effective input variables called feature selection is necessary for modelling, as measurements are achieved from various pollution sources which are often complex and show a nonlinear relationship(Jiang et al, 2004). The feature selection methods are able to reduce the complexity of the input attributes. Therefore, this study examines the performance of the Random Forest feature selection (RFFS) method in combination with multiple-linear regression (MLR) and artificial neural network MLP (called MLP hereafter), to achieve efficient models for prediction of the emissions. This study aims to:

- Develop optimized ANN models for prediction of carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and particulate matter 2.5 ($PM_{2.5}$) concentrations at 5 pollutant detection stations in Tehran.
- Identify the effective attributes for air pollution modeling using RFFS method.

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-W4-483-2017

- Compare the performance of the developed neural network model as a non-linear approach with multiple-linear regression model as a linear method.

2- DATA AND METHODOLOGY

2-1 THE USED DATA

The concentration of four pollutants, CO, NO_2, SO_2 and $PM_{2.5}$ measured in 2012 by five stations, Shaheed Beheshti, Chesmeh, Elm o sanat, park razi and pasdaran, distributed around Tehran were derived from Department of Environment, Islamic republic of Iran. Air pollution in an urban area has a complicated spatial pattern and pollution levels dramatically differ in small distances and at different times (Chaloulakou et al, 2003&Lee et al, 2011). Therefore, to achieve an efficient prediction model, it is necessary to take into account the potential factors affecting the pollutant concentration in the air as much as possible. In this study, the main factors responsible for the concentration of pollutants were organized in four categories including traffic index, concentration of pollutants during the previous days, meteorological and spatial factors. Meteorological attributes including temperature, relative humidity, wind speed, wind direction, cloud cover and surface pressure, as daily average, were collected from https://www.ecmwf.int/. Wind speed and direction were merged through the following equation (Siwek et al, 2012).

$$Wx = |W| * \cos\varphi$$

$$Wy = |W| * \sin\varphi \qquad (1)$$

In equation 1, W represents the wind speed and φ is wind direction indicator. In order to improve the performance of models, month related parameter indicating the changes in emissions due to changes in atmospheric conditions during the year was calculated using Equation 2 (Arhami et al, 2013).

$$MOY = \cos(\frac{2\pi m}{12}) \qquad (2)$$

In Equation 2, m represents the respective month. To calculate the traffic index defined in this study, buffers centered by each station and their radiuses incrementally increased by 100 m intervals up to 1 km were drawn. It is obvious that the concentration of pollutants varies with distance from the road. To calculate traffic index for each buffer, the distance between each station and each street within the buffer was calculated and multiplied by the width of the street and then the weighted average was calculated for each buffer. Spatial factors used in this study contains the coordinates of the stations and the vegetation density index defined in this study. To calculate the vegetation density index, within a radius of 1 km buffers centered by each stations were drawn at intervals of 100 m within a radius of 1 km distance from the station. For each buffer, normalized difference vegetation index (NDVI) was calculated using Landsat-7 image and average NDVI for each buffer was calculated as vegetation density index.

2-2-MODELING

Air pollution modeling methods are divided into two groups, linear and nonlinear methods (Azid et al, 2014&Siwek et al, 2012&Cai et al, 2009&Lu et al, 2004). In the recent years, different approaches including Multi-layer perceptron (Gardner et al (1999)), neural network model radial basis functions (Wang et al (2003)), the multiple-linear regression (Cassmassi (1998); Kardlynv (2001)), and support vector machine (Osowski et al. (2007)) were used for air pollution prediction. Accordingly, in this study, multiple-linear regression as a linear method and MLP, as a non-linear method are compared to indicate which method performs more efficient to predict the maximum concentration of pollutants in the air. MLP consists of a network of simple elements and relations. The number of input and output neurons of the artificial neural network is determined by the nature of the problem. The performance of the MLP depends on its architecture and parameters which are derived from training, and its activation functions (Grivas et al, 2006). In this study a three layer MLP comprising of input, hidden and output layers was used.

The attributes introduced in section 2.1 as along with maximum and mean concentrations of each air pollutant measured for 4 days before were used as the input attributes of the models for prediction of maximum concentration of pollutants for each day. In this study, the data set includes 1780 samples derived from the average of the hourly pollutant concentration measured for each day at each station in Tehran. The data set was randomly grouped into three sets including 70 percent for train, 15 percent for the test and 15 percent for the validation which are used for optimizing the MLP parameters and RFFS.

In addition to use all attributes, RFFS was used for the feature selection in order to select the appropriate attributes before performing modeling techniques. The RFFS is executed as follows:

1. Random Forest is fitted for an air pollutant using all independent attributes; the validation dataset is used to calculate the mean square error for the fitted model.

2. The absolute values of residual values of the predicted and measured pollutant are calculated.

3. The variable importance (VIF) of each independent attribute is calculated by permutation of the attributes.

4. Removal of 20% of the less important attributes.

5. Repeat steps 1 to 4, without repeating step 3.

The suitable number of trees should be preset prior to use of random forest, for this purpose, the number of iterations was set to 10,50,100 and 200 and the validation dataset was used to determine the appropriate number of iteration. Then RFFS was run on validation dataset to determine the suitable input attributes (Shamsoddini et al, 2013). To optimize MLP parameters including learning rate, momentum, number of iteration, minimum error and number of hidden layer neurons, the MLP was run several times and appropriate values were found for each pollutant.

Multicollinearity and overfitting can reduce the efficiency of the models derived by MLR. For this reason, tolerance, p-value and the condition index were calculated for each model derived from MLR to check these problems. Then, the models with conditional index greater than 30, the tolerance less than 0.1 or p-value greater than 0.05 were removed. Root mean square error (RMSE), the standard error of estimation (SSE), the coefficient of determination (R^2) and the percentage of error were calculated for evaluation of the model performance. The estimation error was obtained from the standard error of estimation divided by the mean value of each pollutant multiplied by 100. Paired samples t-tests were used for statistically comparing the performance of different models. These tests were applied on the absolute values of the residuals derived from subtracting measured pollutant values from predicted pollutant values.

3. RESULTS AND DISCUSSION

As mentioned, MLR was applied as a linear method compared to MLP as a nonlinear one. Random forest method which is a nonlinear method was not used in this paper since RFFS was used for feature selection and it could be biased if the random forest was applied on the attributes selected by RFFS for modelling; although, it is believed that the results of RFFS should not have bias towards random forest. After, applying MLR and MLP on the training data and developing the appropriate models based on the optimized MLP and conditions mentioned for MLR, the developed models were applied on the test data. Table 1 shows the results of air pollutant predictions derived from applying MLR and MLP with and without RFFS on the test data. The scatterplots of the forecasts vs. the observed pollutants values are illustrated in Figure 1. As Figure 1 indicates the predictions values almost coincide with the observed concentrations of the air pollutants, especially in the cases of NO_2 and SO_2. The best models derived from MLR, are shown as equations 6 to 9 for predicting CO, NO_2, SO_2 and $PM_{2.5}$, respectively.

$$Z_{CO} = 0.44(a_1) + 0.1(a_2) - 14.73(a_3) - 0.47(a_4) - 0.01(a_5) - 0.64(a_6) + 3.62 \tag{6}$$

$$Z_{NO_2} = 0.37(\beta_1) + 0.22(\beta_2) - 9(\beta_3) - 0.38(\beta_4) - 85.4(\beta_5) - 0.1(\beta_6) \tag{7}$$

$$Z_{SO_2} = 0.40(\gamma_1) + 0.12(\gamma_2) + 0.57(\gamma_3) + 0.36(\gamma_4) - 6.61(\gamma_5) + 1.88(\gamma_6) - 10.13 \tag{8}$$

$$Z_{PM_{2.5}} = 0.34(\theta_1) + 0.26(\theta_2) + 0.41(\theta_3) - 0.29(\theta_4) - 0.07(\theta_5) - 0.03(\theta_6) - 0.08(\theta_6) + 30.77 \tag{9}$$

Effective parameters for estimating CO according to equation (6) are shown with coefficients a_i, in which a_1 and a_2 respectively represent the maximum concentrations of CO one and four days before the day for which the prediction is conducted, a_3 shows the vegetation density index at a distance of 1,000 meters from the station, a_4 represents the month index of the year, a_5 represents the traffic indicators about 700 meters from the station and the average cloud cover is shown by a_6. According to Equation 7, the prediction of NO_2, is based on the variables denoted by β. β_1 and β_2, coefficients represent the maximum and mean NO_2 concentrations at one and three days prior to the day for which the prediction is conducted. β_3 represents the average cloud cover and β_4 is the mean NO_2 concentration in a day before the day for which the prediction is conducted . Variables β_5 and β_6 also represent the vegetation maximum NO_2 concentrations in the four days before the day for which the prediction is conducted, respectively. In Equation 8 developed for the prediction of SO_2, γ_1 and γ_2 and γ_3 represent the maximum SO_2 concentration at one and four days before the day for which the prediction is conducted and the mean SO_2 concentrations of a day before the day for which the prediction is conducted. γ_4 indicates traffic index for a buffer with 1000 m radius and γ_5 represents the mean total cloud cover in the region. Finally, γ_6 stands for average wind blowing in the direction Y. According to equation 9 derived for $PM_{2.5}$ prediction, θ_1 and θ_2 represent the maximum $PM_{2.5}$ concentrations of pollutant one and three days prior to the day for which the prediction is conducted, θ_3 and θ_4 respectively represent the mean $PM_{2.5}$ concentration in one and three days prior to the day for which the prediction is conducted, θ_5 and θ_6 show the traffic index within the buffers with radiuses of 100 m and 1000 m and the average indoor temperature is shown by θ_7. As mentioned, RFFS was applied on the attributes to prediction of each air pollutant and the results are given in Table 2. According to Table 2 the most important variable for prediction of the CO concentration is the maximum CO concentration in a day prior to the day for which the prediction is conducted. After this variable, mean CO concentration a day before the day for which the prediction is conducted and maximum CO concentration in two days prior to the day for which the prediction is conducted are the second and third important variables. Index of months of the year also plays an important role in prediction of CO. According to the results presented in Table 2, maximum and mean NO_2 concentrations in a day prior to the day for which the prediction is conducted are the most important variables for the prediction of NO_2 concentration. Also, maximum and mean concentration of SO_2 in a day prior to the day for which the prediction is conducted are the most important variables for SO_2 prediction. As seen in Table 2, similar to the other air pollutants, $PM_{2.5}$ prediction highly depends on the SO_2 concentration in a day prior to the day for which the prediction is conducted. Moreover, mean temperature is another variable which is important for $PM_{2.5}$ prediction. The attributes selected by RFFS for prediction of $PM_{2.5}$ concentration are 30 from which 9 important characteristics have been shown in Table 2. The selection of the pollutant concentration in the days prior to the day for which the prediction is conducted as the most effective variable for prediction of the air pollutants, may indicates high stability of pollutants in the atmosphere. Selection of index of months of the year reflects of the impact of emission sources on predictive models (Shamsoddini et al, 2013). As mentioned, paired samples t-test was used to statistically compare the performance of different methods and the results are shown in Table 3. P-values less than 0.05 indicate a significant difference between the performances of the methods used to predict the air pollutants.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-W4-483-2017 | © Authors 2017. CC BY 4.0 License.

Emissions	Method	R^2	RMSE	SEE	Estimation error
carbon monoxide [1](PPM)	MLR	0.56	2.3	2.33	50.43
	MLR/RFFS	0.48	1.38	1.39	30.3
	MLP	0.53	1.41	1.52	32.9
	MLP/ RFFS	0.47	1.39	1.39	30.39
nitrogen dioxide [2](PPB)	MLR	0.74	15.54	15.66	27.16
	MLR/ RFFS	0.74	12.2	12.27	22.82
	MLP	0.74	14.92	15.12	26.9
	MLP/ RFFS	0.76	12.04	12.22	22.72
sulfur dioxide (PPB)	MLR	0.61	12.93	13.96	39.57
	MLR/ RFFS	0.59	13.04	13.12	35.65
	MLP	0.61	12.75	13.77	39.03
	MLP/ RFFS	0.61	12.81	12.88	34.95
PM2.5 (PPB)	MLR	0.42	18.77	20.77	35.25
	MLR/ RFFS	0.42	19.85	20.15	32.51
	MLP	0.46	18.35	19.72	33.46
	MLP/ RFFS	0.49	18.13	19.24	32.32

Table 1. Results of MLR and MLP applied on test data

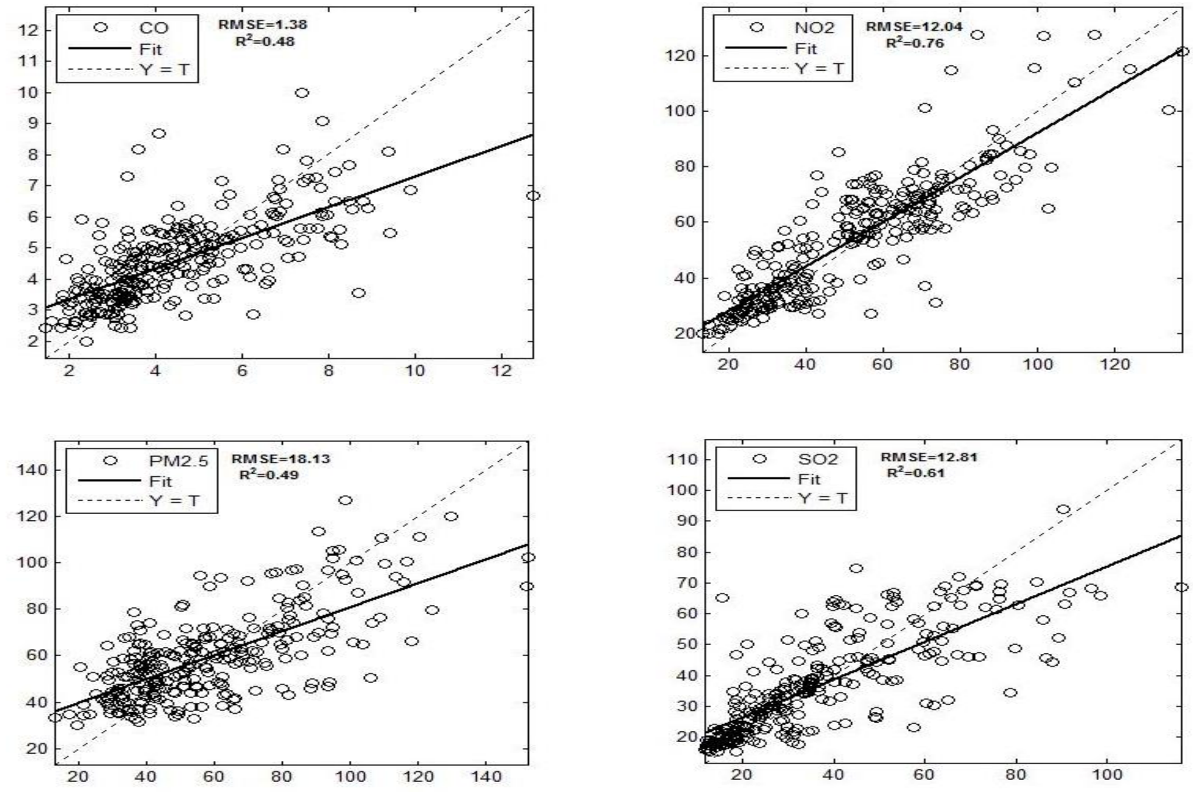

Figure1. Scatterplots of measured (horizontal axes) and predicted (vertical axes) air pollutants derived from the best prediction models

[1] - Parts per million

[2] - Parts per billion

Parameter	VIF(CO)	VIF(SO_2)	VIF(NO_2)	VIF($PM_{2.5}$)
maximum concentration the day before	20.3	26.2	26	26.7
mean concentrations of the days before	13.7	17.6	20/9	23.1
maximum concentration two days before	4.8	6.4	7.6	4.6
Month in year	4.7	-	-	2.4
mean concentrations of two days before	-	-	6.9	7.4
mean concentrations of four days before	-	-	6.2	2.1
mean concentrations of three days before	-	-	5.3	3
maximum concentration three days before	-	-	4	2.7
maximum concentration four days before	-	-	4	-
average temperature	-	-	-	4.8

Table 2. Results of the RFFS method to estimate the importance of each attribute represented in percent

Method	CO	NO_2	SO_2	$PM_{2.5}$
MLR/RFFS & MLR	0.00	0.00	0.63	0.56
MLP & MLR	0.00	0.72	0.37	0.51
MLP/ RFFS & MLP	0.86	0.00	0.59	0.67
MLP/ RFFS & MLR	0.00	0.00	0.87	0.9
MLR/ RFFS & MLP	0.77	0.00	0.89	0.39
MLP/ RFFS & MLR RFFS	0.09	0.88	0.03	0.33

Table 3. The p-values derived from the paired samples t-test

According to the results shown in Table 1, for the prediction of NO2, the best results derived from MLP in combination with RFFS with R^2 0.76 and estimation error 22.7 %. Based on the results shown in Table 1, SO2 was predicted by MLP in combination with RFFS with R^2 0.61 and estimation error 35% better than that derived from the other methods. Based on the results given in Table 1, PM2.5 was predicted using MLP fed by the attributes selected by RFFS with R^2 0.49 and estimation error 32.3% better than the other models. According to Tables 1 and 3, Co and NO_2, predictions are significantly improved using linear method fed by the attributes selected by RFFS whereas combination of MLP and RFFS significantly affects the prediction of NO2. According to Tables 1 and 3, the best predictions for CO, NO_2 and $PM_{2.5}$ can be derived using the combination of MLR or MLP with RFFS while the SO_2 prediction derived from the combination of MLP and RFFS resulted in the best accuracy compared to that derived from the other methods. According to the statistical tests applied to compare the accuracies of the air pollutant predictions, while NO_2 is predicted with the highest accuracy, SO_2 prediction is associated with the most errors than the other air pollutants. It seems that simpler and more linear relationships between the input attributes and NO_2 changes, can result in more accurately prediction of this air pollutant compared to that derived for other pollutants. It can be discussed that SO_2 concentration changes and the factors affecting these changes are more complex than that is the case for other air pollutants.

This complexity can be attributed to the following reasons:

- SO_2 is produced by a variety of the sources such as transportation, and fossil fuel used by power stations and factories which are highly changeable.

- The inability of input attributes to form an efficient model for prediction of this pollutant (Arhami et al, 2013). In fact air pollutant prediction accuracy depends upon the accuracy of the measured data, sources of release and spatial variation of pollutant concentration (Perez et al, 2011).

4. CONCLUSION

This study examined the feasibility of RFFS method for improving MLR and MLP performance when the prediction of the air pollutants including CO, NO_2, SO_2, and $PM_{2.5}$ is the aim. Different attributes including meteorological parameters, and the air pollutant concentration in the days before the day for which the prediction is conducted, were utilized. Also, two indices including traffic and vegetation density were proposed in this study to use along with the common attributes mentioned. The findings of this research can be concluded as follow:

- RFFS indicated that the proposed indices are useful for air pollutant predictions.

- According to RFFS, the concentrations of the air pollutants during the previous days are very important attributes for the prediction of air pollutant concentration for a day.

- Use of RFFS can result in reduction of the difference between the performance of the linear and non-linear methods for prediction of the air pollutants.

- MLP method in combination with RFFS is able to predict all pollutants better than that derived from MLR.

- While the prediction accuracy of SO_2 concentration is significantly lower than the other air pollutants, NO_2 concentration can be predicted with the highest accuracy.

REFERENCE

Akbari, M. Samadzadegan, F., 2015. Identification of air pollution patterns using a modified fuzzy co-occurrence pattern mining method, Int. J. Environ. Sci. Technol, Vol . 12, pp. 3551–3562.

Arhami,M., Kamali, N., Rajabi, M., 2013. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations, Environ Sci Pollut Res, Vol 20, pp. 4777–4789.

Antanasijević, D., Pocajt, V., Povrenović, D., Ristić,M., Perić-Grujić, A., 2013. PM10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization, Science of the Total Environment, Vol. 443, pp. 511–519, 2013.

Azid, A., Juahir, H., Toriman,M., 2014. Prediction of the Level of Air Pollution Using Principal Component Analysis and Artificial Neural Network Techniques: a Case Study in Malaysia, Water Air Soil Pollut, 225:2063.

Cai, M., Yin, Y., Xie,M., 2009. Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach, Transportation Research Part D, Vol. 14, pp. 32–41.

Chaloulakou, A., Saisana, M., Spyrellisa,N., 2003. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens, The Science of the Total Environment 313, 1–13.

Grivas, G. and Chaloulakou,A., 2006. Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece, Atmospheric Environment, Vol. 40, pp. 1216 – 1229.

Jiang, D., Zhanga, Y., Hua, X., Zenga, Y., Tanb,J., Shao,D., 2004. Progress in developing an ANN model for air Pollution index forecast. Atmospheric Environment, Vol. 38, pp. 7055–7064.

Lee, S., Ho, CH., Choi, Y.S., 2011. High-PM_{10} concentration episodes in Seoul, Korea: background sources and related meteorological conditions, Atmos Environ, Vol. 45(39), pp. 7240–7247.

Lu, W., Wang, W., Wang,X., Yan,S., Lam,J., 2004. Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong, Environmental Research 96:79 – 87.

Perez, P. and Trier, A., 2011. Prediction of NO and NO_2 concentrations near a street with heavy traffic in Santiago, Chile, Atmos. Environ., Vol. 35, pp. 1783-1789.

Shamsoddini, A., Trinder, J.C., Turner,R., 2013. Non-linear methods for inferring lidar metrics using SPOT-5 textural data, Photogramm. Remote Sens. Spatial Inf. Sci., II-5/W2, 259-264.

Siwek, K. and Osowaski.S., 2012. Improving the accuracy of predict ion of PM_{10} pollution by the wavelet transformation and an ensemble of neural predictors, Engineering Applications of Artificial Intelligence, Vol. 25, pp. 1246–1258.

Valverde, V., Pay,M., Baldasano, J., 2015. A model-based analysis of SO_2 and NO_2 dynamics from coal- fired power plants under representative synoptic circulation types over the Iberian Peninsula, Science of the Total Environment,Vol. 541, pp. 701–713.

https://doi.org/10.5194/isprs-archives-XLII-4-W4-483-2017