Prediction of PM_{10} concentration 24h in advance using neural networks in Bangkok, Thailand.

Presented in 12th International Conference on Integrated Diffuse Pollution Management (IWA DIPCON 2008).
Research Center for Environmental and Hazardous Substance Management (EHSM)

Saran Pansripong[1], Sudjit Karuchit[1] and Thongplew Kongjan[2]*

Abstract

Artificial neural network (ANN) technique, whose performances in coping with pattern recognition problems is well-known, were proposed to predict of PM_{10} concentration 24 hours in advance and to compare with traditional multiple regression (MR) technique. The average daily air pollution and meteorological data during 2000-2004 in four different areas in Bangkok were utilized. For ANN model generation, the multi-layer feed forward approach (MLFF) and the error-back propagation algorithm (BP) with sigmoid function were used with the selected predictor variables. For MR model generation, the stepwise and the backward selection techniques were employed for selecting appropriate variables into the models. The best models from the 2 approaches were evaluated for their predicting ability using a new set of data from the year 2005 - 2006.

The ANN models developed had the R^2 values in the range of 0.538 - 0.758 and could predict with high accuracy of R value between 0.908 and 0.919. They were effective in predicting the trends in the PM_{10} data set, but not the peak values. Overall, the performance of the neural network models was better than that of the regression models with the same inputs. The predicting ability of models from both approaches rely on variables such as PM_{10}, NO_2, temperature, and relative humidity.

Keywords: PM_{10}, artificial neural network, prediction model

[1]*School of Environmental Engineering. Suranaree University of Technology, Nakhonratchasima 30000*
[2]*Irrigation Development Institute, Royal Irrigation Department, 78 Moo 1 Tiwanon Rd. Bangtalad sub-district, Pakkred district, Nonthaburi 11120*
**corresponding author, e-mail: pansripong.s@hotmail.com*

Introduction

High level of particulate matters which are smaller than 10 micron (PM_{10}) in the atmosphere is a crucial problem in many cities in Thailand. PM_{10} has serious adverse effects on people's health, especially the lower respiratory system. The ability to manage the air quality in a city can be enhanced by the use of prediction models for PM_{10}. The main aim of the research is to develop artificial neural network (ANN) models for forecasting the PM_{10} concentration one day in advance at four different municipality areas in Bangkok. The resulting models were evaluated and compared with multiple regression (MR) models developed using the same set of data.

Particulate matters smaller than 10 micron will cause an effect on people in that they can dissipate to respiration system by breathing. In general, PM10 in urban area is originated from an internal combustion process of on-road vehicle and suspends in atmosphere for hours. Additionally, PM_{10} happen as an effect of processes from industries.

Perez et al. (Perez and Reyes, 2002) proposed that it is necessary to have models that can predict a maximum concentration level of 24-h moving average (24MA) at least 24 hours in advance so as to establish measures for reducing emissions. Selecting independent variables is one of important procedures for developing efficient models. They found that nonlinear models provided more accuracy of prediction than linear regression. In a review article, Gardner and Dorling (1998) described the usefulness of a multilayer neural network (a nonlinear model) for applications in atmospheric sciences. Perez et al. (Perez and Reyes, 2002) showed that a three-layer neural network was a useful tool to predict hourly averages of PM2.5 concentrations in the atmosphere of downtown Santiago, Chile, several hours in advance, when hourly concentrations of the previous day and forecasted temperature, relative humidity, and wind speed were used as input. Predictions generated with this network were more accurate than those produced with a linear regression that has the same inputs. Silva et al. (2001) used a MARS algorithm (which is nonlinear) to forecast PM10 concentrations in Santiago, Chile, and they reported an accuracy that appears greater than what can be obtained with multilayer neural networks and linear regressions. Van der Wal and Janssen (2000) showed that 45% of the variance of PM_{10} concentrations may be explained by changes in wind direction, temperature and duration of precipitations.

Methodology

The secondary data used for model development were obtained from 4 Pollution Control Department (PCD) monitoring stations during 2000 to 2004, namely, 10T station (Klong Chan Housing Community), 11T station (Huaykwang - National Housing Authority Stadium), 12T station (Nonsi Witthaya School) and 15T station (Singharaj Pittayakom School). The air pollution data used were the average daily measurements of sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), Ozone (O_3), and PM_{10}. The meteorological data used were the average daily measurements of temperature (T), relative humidity (RH), atmospheric pressure (P), rain (Rain), net radiation (NR), globe radiation (GR), wind speed (WS), and wind direction (WD). Thus, there are 5 air pollution variables and 8 meteorological variables. To

predict the average daily PM_{10} concentration 24 hours in advance, the above data of the present day and the last two days were employed as predictors. Therefore, there are 39 independent variables in each case.

Significant relationship between the predictor variables and the PM_{10} concentration of the next day were identified using Pearson correlation coefficients. Appropriate variables were then chosen for formulating models by using artificial neural network and multiple regression techniques.

For ANN model generation, the multi-layer feed forward approach (MLFF) and the error-back propagation algorithm (BP) with sigmoid function were used with the selected predictor variables. The sigmoid function according to the equation:

$$f(x)=\frac{1}{1+e^{-x}} \qquad (1)$$

The ANN procedure starts from collecting the data and separating them into 3 sets: training data (60%), test data (20%) and validation data (20%). Then, the structure is initialized with a simple network (1 hidden layer and a few hidden nodes). Consequently, the networks are trained. If a local optimum is found on the testing error, it is recorded and the training is temporally stopped. By adding a new node to the hidden layer, the training is restarted. The testing error is calculated again and the new local optimum is compared with that of the previous structure. This procedure is repeated. When the better performance, respective to testing error, is found on the best structure, the training is considered to be complete and the process is stopped. The best structure is assumed as the final optimal. Finally, the best structure is validated with the validation set. This algorithm is shown in Figure 1.

The procedures of training step include: 1) random selection of one, two, and three hidden layers, 2) random selection of the number of hidden nodes, and 3) random selection of initial weight, momentum, and learning rate. The best models are those with minimum value of statistical measurement of mean absolute percentage error (MAPE), according to the equation:

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{Predict_i-Actual_i}{Actual_i}\right|\times 100 \qquad (2)$$

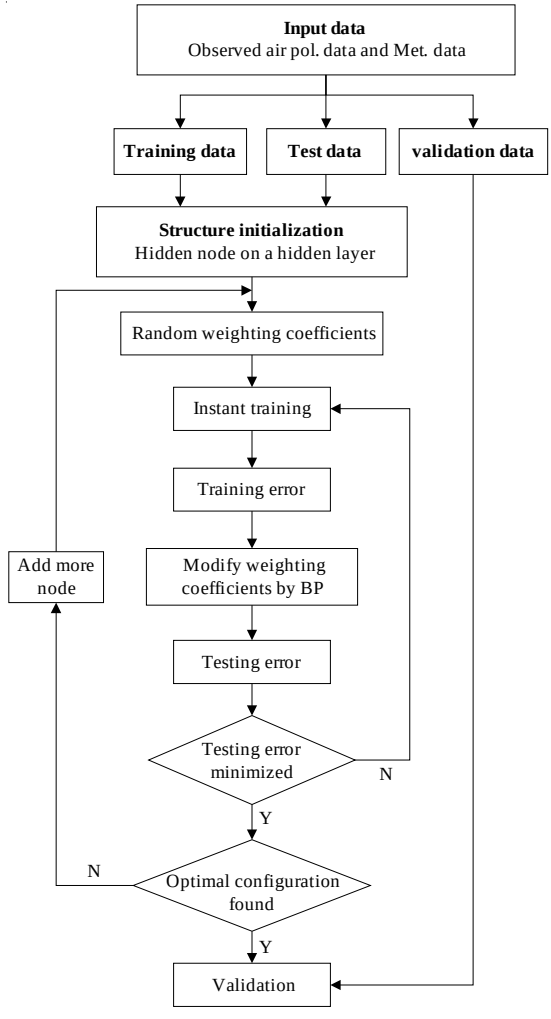

Figure 1. Training procedure of the MLFF

For MR model generation, the stepwise and the backward selection techniques were employed for selecting appropriate variables into the models. If X_i is the value of the input variable i and Y is the actual PM_{10} concentration, then the constant b_0 and the regression coefficients b_i are computed by the ordinary least-squares equation:

$$Y = b_0 + \sum_{i=1}^{n} b_i X_i + \varepsilon_i \qquad (3)$$

Consequently, the best models from the 2 approaches were evaluated for their predicting ability using a new set of data from the year 2005 - 2006.

Results and discussions

Appropriate predictors

PM_{10} concentrations of the next day of all stations had significant positive correlation with other air pollution concentrations both transformed and untransformed previous day data at 95% confidence level. It showed that the next day PM_{10} concentrations correlate closely with the air pollution data of the present day and the last two days. In addition, transformation into logarithm form did not reflect on the relationship pattern of PM_{10} and other variables. Statistic results revealed that the concentration of PM_{10} in the next day had strong relationship with PM_{10} of the days before, as perceived from correlation values ranging from 0.543 to 0.856. The second variable that affects to PM_{10} forecasting was NO_2, as shown by the correlation values in the ranges of 0.307 to 0.740. For other air pollution parameters, the correlation values were in the range of 0.186-0.532, 0.129-0.500 and 0.079-0.215 for CO, O_3 and SO_2, respectively.

In case of meteorological data, they showed negative correlation with the next day PM_{10} concentrations except pressure data (P). This could be explained by the fact that the pressure causes air mass to flow down to the earth's surface and confines the air dispersion. This brought about the direct relationship between the particulate matters and the atmospheric pressure.

The statistic results were congruent with the air pollution distribution theory relating to the effect of temperature (T), relative humidity (RH) and (P) on wind formation. The faster the wind blows, the more dispersion of the PM_{10}. According to the study of Thorpe (Thorpe, 2007), fast-flowing and strong-flowing wind could disperse PM_{10} to wherever it has gone and the wind velocity has inverse variation with the quantity of particulate matter. For the relationship between PM_{10} in the next day and GR and NR, it was found to be inversely correlated. The particulate matters suspended in the atmosphere have an impact on the amount of sun light penetrated into the earth. The particulate matters can absorb and scatter the sun light. Therefore, when the atmosphere is replete with particulate matters, the amount of light traveled to the earth is decreased. However, it has no statistical effect in some monitoring station because there are other factors that have a repercussion on GR and NR such as the cloud amount or the ray reflection from materials. The major factors that impacts on PM_{10} concentration are RH and P. The secondary ones are T, Rain, WD, WS, NR and GR.

Artificial neural network models

The best ANN model for predicting the PM_{10} concentration 24 hours in advance at each monitoring station is shown in Table 1. They are mostly found with natural logarithm transformation of variables. In every model, there are five independent variables

that is important in PM_{10} prediction including previous- PM_{10}, NO_2, P, RH and T. Because of the fact that air pollution data are intricate, it is necessary to generate models having at least 2 hidden layers and a number of hidden nodes. The R^2 stemming from 10T station, 11T station, 12T station and 15T station showed reasonably high value of 0.707, 0.758, 0.538 and 0.716, respectively. These results are comparable to the study of Grivas and Chaloulakou (2006). They obtained the R^2 values ranging between 0.50 and 0.67 for hourly-predicted PM_{10} concentrations.

Most distributions of environmental factors are in the log-normal form and the modeling of environmental data is generally complicated. Perez et al. (2000). proposed more than one hidden layer in every station on account of data complication. It could be seen from Table 1 that variables in the efficient models are the ones that were transformed into natural logarithm. By adjusting network parameters, learning rate (η) should be higher than 0.05. This is because the network would be trained until 10,000 loops (epochs) before predetermined error value was obtained resulting in obtaining a high error value of prediction. For optimal-selected weight (w) and momentum (α) values, it is dependent upon characteristics and relations of each set of data (Gardner and Dorling, 1998; Dimopoulos et al.,1999; Nagendra and Khare, 2006) as noticed from great weight values in many stations. From figure 2, it was found that artificial neural network models were able to capture the trends in the PM_{10} data set; they were not able to forecast the peak values.

Multiple regression models

Appropriate models were selected by taking Adjusted R^2 and residual analysis into account. It was found that efficient models were derived from backward approach and the dependent and independent variables must be in form of natural logarithm. The four-selected models from four monitoring stations, thereafter, were validated by another set of data to compare the actual PM_{10} in the next day with predicted PM_{10}. Correlation coefficient value (R) was used as determining indicator. If R value is close to unity, that model has high accuracy of prediction. MSPR was also taken into account-MSPR values less than or close to MSE values indicate that the models have high performance of prediction.

Table 1. Neural network model for prediction of PM_{10} concentration 24 hours in advance.

Station	Data transformation[1]	Architecture[2]	Parameters[3]			R^2
			η	α	w	
10T	Log-Log	32-63-63-63-1	0.1	0.3	0.4	0.707
11T	Log-Log	31-22-22-1	0.3	0.7	0.6	0.758
12T	Normal-Log	36-43-43-1	0.1	0.3	0.8	0.538
15T	Log-Log	38-26-26-26-1	0.3	0.3	0.8	0.716

[1]Independent variable - Dependent variable, [2]Input - Hidden layer(s) - Output, [3]η: learning rate, α: momentum, w : initial weight

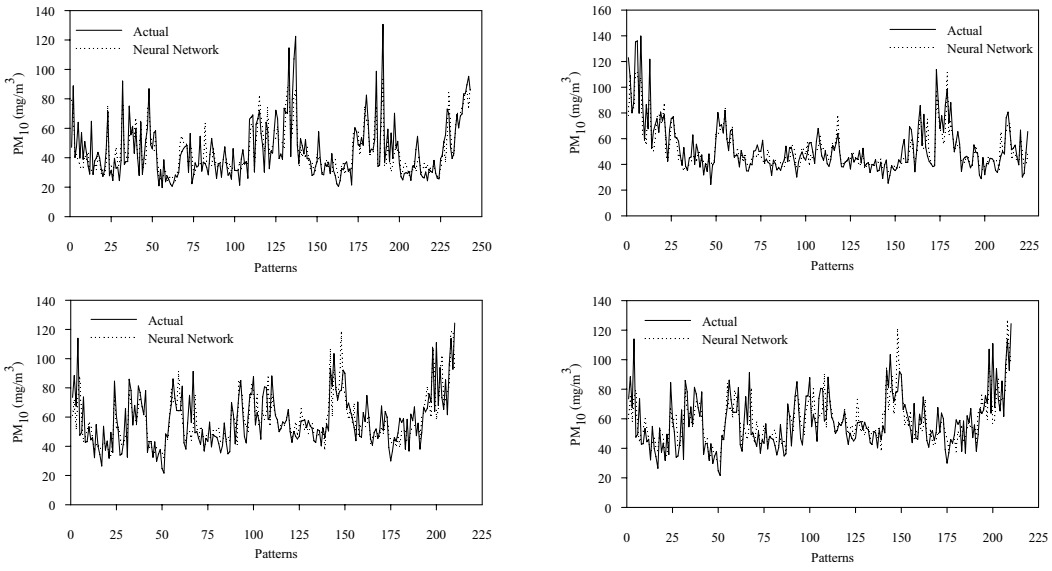

Figure 2. Plot between of PM_{10} actual and predicting values.

Table 2. Multiple regression model for prediction of PM_{10} concentration 24 hours in advance.

Station	Number of predictors	Regression		Validate	
		Adj.R^2	MSE	R	MSPR
10T	16	0.699	0.044	0.844	0.046
11T	22	0.760	0.024	0.877	0.031
12T	17	0.722	0.031	0.774	0.040
15T	21	0.727	0.051	0.849	0.061

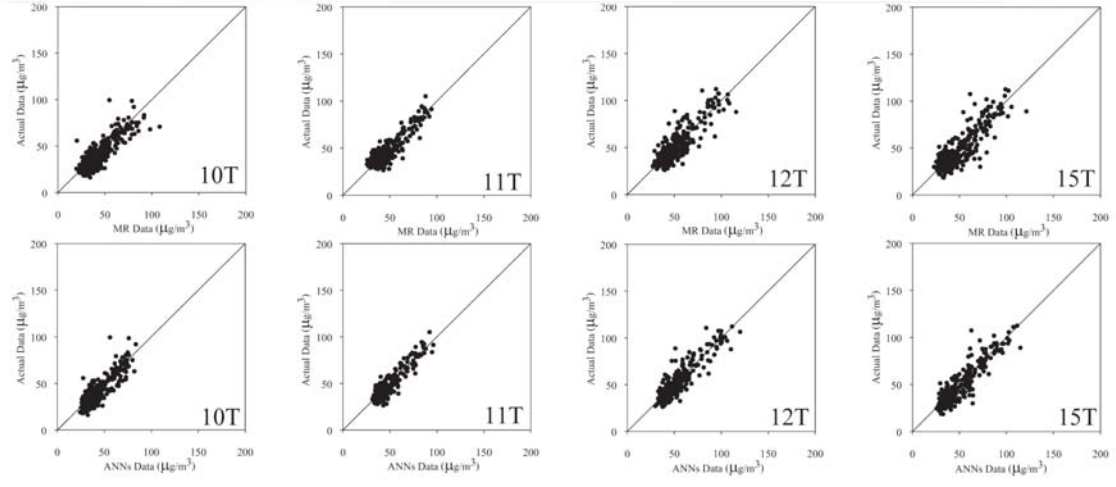

Figure 3. Scatter plot between of PM_{10} actual and predicting values.

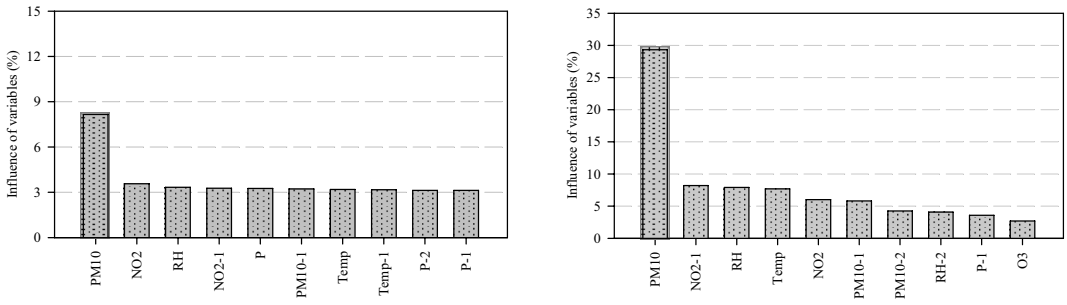

Figure 4. Influence of predicting values.

Table 2 shows that the Adjusted R^2 of the MR models ranged from 0.699 to 0.760. In other words, the models can predict the PM10 variation with accuracies between 69.9% and 76.0%.

After finishing the model development, the other set of data which accounts for 20% of the total was used to assess the finished models. It was found that the predicted and the actual PM_{10} data had good relation as perceived from R values and MSPR. The R values of 10T, 11T, 12T and 15T monitoring station were 0.844, 0.877, 0.774, 0.849, respectively. The MSPR was close to MSE as shown in Table 2.

Models performance comparison

To assess and compare the model performance, the actual values were plotted against the predicted values (Figure 3). If the distance of the scattering points are close to the bisectrix (equality line), it indicates that the model is well fitted with the data (Tabach et al.,2007). It was found that the plots of the multiple regression models had slightly more distribution of data from the equality line. The plots of the neural network models, however, shows better accuracy and correlation coefficient (R), with values of 0.908, 0.919, 0.910 and 0.913 for 10T, 11T, 12T and 15T, respectively.

Table 3. Comparison of correlation coefficients and RMSE between actual and predicting value.

Station	RMSE (g/m^3)		R	
	MR	ANNs	MR	ANNs
10T	8.700	6.649	0.854	0.908
11T	6.461	5.581	0.896	0.919
12T	9.139	7.595	0.863	0.910
15T	10.275	8.202	0.860	0.913

Table 3 shows the performance of the next 24 hours PM_{10} forecasting with the new data set. Using the neural network technique, the models for the monitoring stations 10T, 11T, 12T and 15T were found to have moderate uncertainty with RMSE values of 6.649, 5.581, 7.595 and 8.202 $\mu g/m^3$, respectively. The RMSE values of the MR models were 13.6 – 23.62% higher than those of the ANN model.

Influence of selected independent variables

For models from both ANN and MR approaches, the influence of each variable in each model to the dependent variable was computed and adjusted to have the sum of 100%. The influence of each variable was then averaged over the 4 stations and ranked. Figure 4 shows the 10 highest influence

independent variables from the ANN models and the MR models obtained in this study. It is apparent that PM_{10} and NO_2 of the present day and the previous day are essential to the prediction, regardless of the approach used to develop the model. Temperature and relative humidity are also effectual predictors. Notice that the ANN models include all the independent variables used in the model development, and the influence of each variable is relatively close.

Conclusion

From the results, it can be concluded that the ANN models derived from the average daily air pollution and meteorological measurements of the present day and the last two days can perform at high accuracy, with the R values between 0.908 and 0.919, which are better than those of the MR models. The ANN models were capable of predicting the trends in the PM_{10} data set, but not the peak values. The predicting ability of models from both approaches relies on variables such as PM_{10}, NO_2, temperature, and relative humidity. The models from this study can be used for predicting the PM_{10} concentration in Bangkok area, and can be useful tools for management of control and public warning strategies for PM_{10} level.

For recommendation, approaches for efficient selection of the variable should be developed. This is because the neural network approach takes all initial variables in generating a model, some of which have less influence on model prediction. Reducing model variables would reduce the monitoring workload and increase the emphasis on the truly important factors.

References

Dimopoulos, I., Chronopoulos, J., Chronopoulos-Sereli A and Lek, S. 1999. Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece). **Ecol. Model.** 120 , pp. 157-165.

Gardner, M.W. and Dorling, S.R. 1998. Artificial neural networks (the multilayer perceptron) - a review of applications in atmospheric sciences. **Atmospheric Environment 32:** 2627-2636.

Grivas, G. and Chaloulakou, A. 2006. Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens, Greece. **Atmospheric Environment 40** 1216 - 1229.

Nagendra, S.M.S. and Khare, M. 2006 Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. **Ecol. Model.** 190 (1-2), pp. 99-115.

Perez, P., Trier, A. and Reyes, J. 2000. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. **Atmospheric Environment 34:** 1189-1196.

Perez, P. and Reyes, J. 2002. Prediction of maximum of 24-h average of PM10 concentrations 30 hr in advance in Santiago, Chile. **Atmospheric Environment 36:** 4555-4561.

Silva, C., Perez, P. and Trier, A. 2001. Statistical modelling and prediction of atmospheric pollution by particulate material: two nonparametric approaches. **Environmetrics 12,** 147-159.

Tabach, E.E., Lancelot, L., Shahrour I. and Najjar, Y. 2007. Use of artificial neural network simulation metamodelling to assess ground-water contamination in a road project. **Mathematical and Computer Modelling 45.** 766-776.

Thorpe, A.J., Harrisona, R.M., Boulter, P.G. and McCrae, I.S. 2007. Estimation of particle resuspension source strength on a major London Road. **Atmospheric Environment 41:** 8007 - 8020.

Van der Wal, J.T., Janssen, L.H.J.M. 2000. Analysis of spatial and temporal variations of PM_{10} concentrations in the Netherlands using Kalman filtering. **Atmospheric Environment 34,** 3675-3687.