Prediction of Ambient PM$_{10}$ Concentration with Artificial Neural Network

L. H. Lam *, K. M. Mok

Department of Civil and Environmental Engineering, Faculty of Science and Technology, University of Macau, Macau SAR, China

Email: LHLam@umac.mo

Abstract This paper presents the design of building economic and flexible artificial neural network (ANN) model for the prediction of PM$_{10}$ concentration 24-hour-ahead. Instead of relying purely on historical information of the pollutant, the developed model would incorporate the effects from local meteorological conditions and other related pollutants explicitly. The ANN used was a three-layer feed-forward network (TLFN) of the back-propagation type. Computation efficiency was achieved by limiting the size of the input data to six parameters per input set, assuming variation of the predicted hourly PM$_{10}$ concentration to depend only on meteorological and air quality conditions within the last 72 hours, and separating the model development according to prevailing seasons. Two sets of model for the summer and winter seasons were developed and tested based on one full year of measurements in Macau. Selections of input parameters for models were determined by analyzing correlation coefficients among the hourly concentrations of measured pollutants and seven meteorological parameters. The number of neurons used in the hidden layer for each model was then determined by systematic trials and selecting that with minimum root mean square error. Five and six neurons were determined for the summer and winter models, respectively. Predictions for seven days by the trained models were compared with measurements. Results show that on average half of the predictions achieved accuracy on absolute relative percentage error of less than 50% with the summer model performed slightly better. Further studies on model selection technique are recommended for improvement of prediction accuracy.

Keywords: PM$_{10}$, prediction, criteria pollutants, meteorological effect, artificial neural network

INTRODUCTION

It is known that the concentrations of air pollutants are highly related to the variations of the local and regional meteorology conditions which dictate the dispersion and transport routes of them. Many correlation studies and models simulating the fate and impacts of the released pollutants are these bases [6-15]. Meanwhile, ambient air pollutants may also affect the concentrations of each other therefore making prediction or modeling of their behaviors a very complex problem. Previous studies applying artificial intelligent techniques on air quality prediction show success even the models developed use only temporal data of air concentrations; hence mainly time series analyses are performed [1-5]. This study attempts at designing economic and flexible artificial neural network (ANN) models for predictions on hourly concentrations of respiratory suspended particulates (PM$_{10}$) 24-hour-ahead, taking into account explicitly the effects from local meteorological conditions and concentrations of other criteria pollutants. They are then tested in Macau where has been experiencing gradual air-quality degradation in the recent decade due to rapid growth of economy, traffic, energy consumption and population of itself as well as the Pearl River Delta Region at large. The Macau Meteorological and Geophysical Bureau (SMG) started in 1987 to monitor the ambient air quality. In 1999, SMG started to report the last 24-hour air quality situation to the publics based on records of three continuous ambient air concentration monitoring stations inaugurated that year. In 2001 SMG also added one more continuous monitoring station in their network and started to announce air quality forecast of the next day. Details of the background information on air quality and its forecast in Macau could be found in [1]. Therefore, development of air quality prediction techniques and tools are important for Macau.
ARTIFICIAL NEURAL NETWORK

Prediction on the concentration of a pollutant is assumed to base on the history of its own concentration, concentration of other pollutant species and values of some key meteorological parameters. Considering that \( y_t \) is a stationary time series denoting the hourly variation of the concentration of a pollutant, precisely there is a function \( F \) such that:

\[
y(t) = F\{y(t-\tau_1), ..., y(t-\tau_n), a(t-\tau_1), ..., a(t-\tau_n), b(t-\tau_1), ..., b(t-\tau_n), ..., i(t-\tau_1), ..., i(t-\tau_n)\}
\]

where \( t \) is the present hour, \( [a], [b], ..., [i] \) are values of the related meteorological parameters and concentrations of other pollutants, and \( [\tau] \) are the hour delays. The task is to find a good approximation of this unknown function \( F \) through an artificial neural network (ANN). The ANN applied in this study is a three-layer feed-forward network (TLFN) of the back-propagation type. It was proved that a TLFN could approximate all continuous functions on any high dimension compact domains uniformly [16]. This means that \( F \) can be estimated by placing almost no condition. A definition sketch of a typical TLFN with \( n \) inputs, \( r \) neurons, and one output is shown in Fig. 1.

![Figure 1: Definition sketch of a typical three-layered feed-forward ANN with one input layer, one hidden layer, and one output layer.](image)

In the input layer, pattern \( P(1), P(2), ..., P(n) \) is taken. Then they are linearly transformed by the neurons of the input layer. The output signal generated is weighted with \( W1(n, r) \) and fed to each of the \( r \) neurons in the hidden layer. Then each neuron in the hidden layer sums all the received weighted input signals and bias \( b(r) \) and produces an overall unit activation which is then converted to an output signal by utilizing a sigmoid transfer function. They are weighted with \( W2(r, 1) \) next and fed to the neurons of the output layer. At last the weighted input signals and the bias \( b \) of the output layer are summed to produce a unit activation which is linearly transformed to yield the calculated target value \( T \). The learning method used is the well known supervised back-propagation method [17]. Briefly, this method lets the network learn from the examples sequentially and the values of the adaptive parameters of the network, the weights and the bias, are updated so that the cost function and the sum-squared error can be stochastically minimized. The TLFN of the back-propagation type used here is designed with the MATLAB Neural Network toolbox.

APPLICATION IN MACAU

The main objective is to predict the hourly concentrations of PM\(_{10}\) in the air of the next twenty four hours based on values of the historical data of some selected key meteorological parameters and criteria pollutants. There are currently four continuous ambient air quality monitoring stations collecting concentrations of criteria pollutants in Macau. The measured pollutants and categories of these stations are listed in Table 1. Note that the operation of the Taipa Grande and the Northern stations were started in March 1999, the Rua do Campo station was started in July 1999 and the Taipa station was started in March 2001.

As the station Taipa Grande is located at 114m above sea level and is also where SMG situated at, information recorded there experiences less local interference and can represent the general ambient air quality and background meteorological status of Macau. Data recorded at this station was selected for assessment on the present developed models. However, carbon monoxide CO is not monitored at this station so that the CO records from the Northern station were used in order to include all five criteria pollutants for completeness. It is commented that information from Taipa station was not yet available during the study, otherwise CO values recorded there would be more appropriate to use since it is much closer to the Taipa Grande station than the Northern one. The data used in the present study were the hourly air concentrations and the meteorological data from May 1st 1999 to April 30th 2000 obtained from SMG and they are listed below.
Table 1: Measured pollutants and site category of the four monitoring stations

<table>
<thead>
<tr>
<th>Station Name</th>
<th>Category</th>
<th>Measured Pollutants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taipa Grande</td>
<td>General Ambient</td>
<td>PM$_{10}$, SO$_2$, NO/NO$_2$/NO$_x$, O$_3$</td>
</tr>
<tr>
<td>Taipa</td>
<td>School, High Density Residential Area</td>
<td>PM$_{10}$, NO/NO$_2$/NO$_x$, O$_3$, CO</td>
</tr>
<tr>
<td>Northern</td>
<td>Commercial, High Density Residential Area</td>
<td>PM$_{10}$, SO$_2$, NO/NO$_2$/NO$_x$, O$_3$, CO</td>
</tr>
<tr>
<td>Rua do Campo</td>
<td>Commercial, Residential, Roadside</td>
<td>PM$_{10}$, NO/NO$_2$/NO$_x$, CO</td>
</tr>
</tbody>
</table>

**Hourly meteorological data**

- Dew: Dew point (°C);
- Dire: Wind direction (degree), [0, 360), (0 =North, 90 =East, 180 =South, 270 =West);
- Humi: Relative humidity (%), [0, 100];
- Rain: Precipitation during one hour (mm), (>= 0);
- Speed: Wind speed (km/h), (>= 0);
- Sun: Fraction of hour with sunshine, [0, 1]
- Temp: Outdoor ambient temperature (°C);

**Hourly air concentrations**

- CO: Carbon monoxide concentration (mg/m$^3$)
- NO$_2$: Nitrogen dioxide concentration (µg/m$^3$)
- O$_3$: Ozone concentration (µg/m$^3$)
- PM$_{10}$: Respiratory suspended particulate concentration (µg/m$^3$)
- SO$_2$: Sulfide dioxide concentration (µg/m$^3$)

Previous studies indicated that significantly different behaviors of air concentration in Macau are controlled by the prevailing monsoon climate which exhibits mainly two seasons, namely summer and winter [6, 7]. Therefore, it was decided that models of PM$_{10}$ for the two seasons would be treated independently. The hourly data available were then separated into a summer series from 01/May/1999 to 30/Sep/1999 (3672 hours) and a winter series from 01/Oct/1999 to 30/Apr/2000 (5112 hours) based on suggestion of [7]. For model training, testing and validation, each data series was divided into three segments for different purposes as indicated in Table 2.

**Table 2  Training, testing and validation (prediction) periods**

<table>
<thead>
<tr>
<th></th>
<th>Summer</th>
<th>Winter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1/5/99,00:00 - 16/9/99,23:00 (3336 hours)</td>
<td>1/10/99,00:00 – 16/4/00,23:00 (4776 hours)</td>
</tr>
<tr>
<td>Testing</td>
<td>17/9/99,00:00 – 23/9/99,23:00 (168 hours)</td>
<td>17/4/00,00:00 – 23/4/00,23:00 (168 hours)</td>
</tr>
<tr>
<td>Validation</td>
<td>24/9/99,00:00 – 30/9/99,23:00 (168 hours)</td>
<td>24/4/00,00:00 – 30/4/00,23:00 (168 hours)</td>
</tr>
</tbody>
</table>

The main structural elements needed to be determined for the TLFN models are the input pattern (layer) and the number of neurons in the hidden layer. Nevertheless, exact analysis of the TLFN structure is still inconclusive and it is usually determined experimentally. To limit the size of input data for better computation efficiency in this study, it was assumed that the variation of the predicted hourly PM$_{10}$ concentration depended only on meteorological and air quality conditions of the same hour one day (24 hours) to three days (72 hours) back at maximum. At the same time, a total of six parameters at most would be used as the input set.

To select the proper input parameters for each model, correlation coefficients among the listed hourly pollutant concentrations and meteorological data in the two seasons were calculated. The six parameters of different time stamps that had the highest absolute values of correlation coefficient with the target PM$_{10}(t)$ concentration are listed in Table 3. The linear correlations are not strong but the best ones are found between values within the last 48 hours which is within the original assumption. Therefore, input patterns selected according to these results for both summer and winter seasons are:
Summer Season

\[ PM_{10}(t) = F\{PM_{10}(t-24), \text{NO}_2(t-24), \text{O}_3(t-24), \text{DEW}(t-24), \text{NO}_2(t-48), \text{HUMI}(t-24)\} \] (2)

Winter Season

\[ PM_{10}(t) = F\{\text{HUMI}(t-24), \text{PM}_{10}(t-24), \text{HUMI}(t-48), \text{NO}_2(t-24), \text{PSEA}(t-48), \text{DEW}(t-24)\} \] (3)

The corresponding numbers of neurons in the hidden layer for each model with the above input pattern were tried systematically. Using data in the stated periods in Table 2 as the training and testing sets, the numbers of neuron used in the hidden layer were tested from 1 to 10. Each case of a certain neuron number was run four times and a test root mean square error (RMSE) was determined for each run. The root mean square error is defined as

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \frac{\text{model}_i - \text{measurement}_i}{\text{measurement}_i} \right)^2}
\] (4)

where \(N\) is the number of data used for testing. The final RMSE for the case of that given neuron number is the average of the four RMSEs. The case of certain neuron number that gave the least final RMSE in the testing period was adopted for final prediction or model validation. It is noted that the key parameters set in the program for calculation were 150,000 for maximum number of iteration, 0.1 for initial learning rate, 0.9 for learning momentum and 0.001 for learning error goal. Variations of RMSE with the number of neurons for all cases are plotted in Fig. 2.

Table 3: Correlation coefficients among the listed hourly pollutant concentrations and meteorological data in the two seasons

<table>
<thead>
<tr>
<th>Summer</th>
<th>Winter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation Between</td>
<td>Correlation Coef.</td>
</tr>
<tr>
<td>(PM_{10}(t)) &amp; (PM_{10}(t-24))</td>
<td>0.47</td>
</tr>
<tr>
<td>(PM_{10}(t)) &amp; (\text{NO}_2(t-24))</td>
<td>0.37</td>
</tr>
<tr>
<td>(PM_{10}(t)) &amp; (\text{O}_3(t-24))</td>
<td>0.30</td>
</tr>
<tr>
<td>(PM_{10}(t)) &amp; (\text{Dew}(t-24))</td>
<td>–0.24</td>
</tr>
<tr>
<td>(PM_{10}(t)) &amp; (\text{NO}_2(t-48))</td>
<td>0.22</td>
</tr>
<tr>
<td>(PM_{10}(t)) &amp; (\text{Humi}(t-24))</td>
<td>–0.22</td>
</tr>
</tbody>
</table>

The final choices of neuron number for the models of the summer season and winter season were 5 and 6 with their corresponding test RMSEs being 45.6% and 11.9%, respectively. It is noted that the RMSE value in the summer season was the least when the number of neuron was just one. However, the average RMSE value was also comparatively small when the number of neuron was five. In such case, five neurons were still adopted for the process of further prediction so that it was more consistent with that of the winter season case.

![Figure 2: Variation of root mean square error with number of neurons used in the hidden layer during testing](image)

When the structures (i.e. the input pattern and the number of neuron used in the hidden layer) of all the TLFN were determined, the models were re-trained with the complete data sets including both the training and testing periods before. They were then applied to do prediction for model validation. The learning performances of the trained models
are shown in Figs. 3 and 4. It is shown in Fig. 3 that the summer season model in general could simulate the variations of the PM$_{10}$ concentration by capturing both the low and high fluctuations quite well. However, the winter season model does not perform as well as the summer one as shown in Fig. 4. It in general shows lower fluctuated values than the measured ones. Nonetheless, it could still follow the main variation trend of a longer time scale.

![Figure 3: Learning performance of the summer season model](image1)

![Figure 4: Learning performance of the winter season model](image2)

The trained models were then used to forecast the hourly PM$_{10}$ concentrations 168 hours (seven days) ahead for model verification. Since the model was designed to do 24 hours ahead forecast, the input data would be updated with the actual measurements for new forecasts as the prediction process advanced through the verification period.

**RESULT AND DISCUSSION**

The model-prediction results comparing with the actual measurements are shown in Figs. 5 and 6. The performance of each model was evaluated qualitatively by visual inspection and quantitatively by calculating the absolute relative percentage error (ARPE) and the mean absolute percentage error (MAPE) between prediction values and actual measurements:

\[
ARPE = \frac{|\text{prediction}_i - \text{measurement}_i|}{\text{measurement}_i},
\]

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|\text{prediction}_i - \text{measurement}_i|}{\text{measurement}_i}.
\]

In appearance, there are some differences between the predictions and measurements in both cases. The summer model seems to perform better than the winter one. This may be due to the variation of the PM$_{10}$ concentration within the predicted summer period was relatively “regular”. In other words, large prediction errors occurred, especially in winter, at the hours where high and sharp peaks (a rapid and large increase and decrease) appeared in actual PM$_{10}$ measurements. It seems that the present models could not effectively predict extreme fluctuation of the measurement.
However, the sudden relatively large increase and decrease in pollutant concentration measured hourly in an ambient air quality monitoring station may be considered as ‘noise’ and could be caused by some instantaneous, individual incidents occurred in the vicinity of the monitoring station. It is reasonable that these abnormal individual incidents could not be predicted according to just limited past knowledge. Quantitative performance of the models is tabulated in Table 4.

Table 4: Percentage of prediction hours with different ARPEs

<table>
<thead>
<tr>
<th>Absolute Relative Percentage Error (ARPE)</th>
<th>&lt;10%</th>
<th>&lt;20%</th>
<th>&lt;30%</th>
<th>&lt;40%</th>
<th>&lt;50%</th>
<th>&gt;50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of prediction hours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer</td>
<td>19%</td>
<td>29%</td>
<td>49%</td>
<td>60%</td>
<td>64%</td>
<td>36%</td>
</tr>
<tr>
<td>Winter</td>
<td>10%</td>
<td>17%</td>
<td>30%</td>
<td>36%</td>
<td>41%</td>
<td>59%</td>
</tr>
</tbody>
</table>

Results from the summer season model show that it could achieve accuracy in ARPE of less than 30% for about half of the prediction hours. However, there are still more than one-third of the validation hours having ARPE larger than 50%. The overall performance of the summer season model viewed by the mean absolute percentage error (MAPE) over the total 168 prediction hours is 59%. As for the performance of the winter model, the results of prediction are not as good as that in the summer model. It could only achieve accuracy in ARPE of less than 30% for about one-third of the prediction hours. There are 60% of the validation hours still having ARPE larger than 50%. The overall performance of this model viewed by the mean absolute percentage error (MAPE) over the total 168 prediction hours is 99% which is almost 1.8 times of that given by the summer season model. Nonetheless, these results could be considered satisfactory given that the criteria used for the input pattern selection were quite restrictive. Performance of the models may be improved if more input parameter combinations or inclusion of more related parameters could be tested. The
concentration of PM_{10} is considered to depend highly on meteorological factors in this study; selection of the input meteorological data sets is of vital importance. For example, it is known that PM_{10} could be formed in the presence of VOC, NO_{x}, O_{3} and sun light, while the formation of NO_{2} and O_{3} is highly correlated to the intensity of ultraviolet (UV) ray from the sun. However, the meteorological data used in the present study related to sun light was record of the fraction of hour with sunshine, but not UV intensity. Hence effect of the sun on the variation of PM_{10} may not be properly accounted for as it was excluded in the model due to low correlation found in the correlation analysis. Therefore, different models may result if measurements of UV were used in replacement of the sunshine hour. It is suggested for future consideration. Furthermore, the selection of input meteorological and pollutant data sets by means of correlation coefficient alone may not be sufficient as it could only provide information on linear interdependence between two sets of variables. The relationship among air pollution concentrations and meteorological factors is obviously non-linear. Other techniques should be explored for model selection.

CONCLUSION

Studies on the 24-hour-ahead predictions of ambient PM_{10} concentration explicitly incorporated the effects of meteorological factors and related air pollutant conditions with artificial neural network have been worked out and tested in Macau. Results were satisfied given that the set criteria for model input pattern selection were restrictive and the information used was limited, but they were not as good as expected. It was found that the designed neural networks could capture the general trends of the real measurement, but large errors occurred at those hours where high and sharp peaks appeared in the actual measurements. These were suspected to be caused by instantaneous local disturbances occurring in short time scales which the input patterns of the trained models selected based on restrictive conditions did not account for. Therefore, the choice of model clearly has predominant effect on the prediction. It is necessary to explicitly address the issue of selecting the optimal model, that is, the one that fits the data well and is insensitive to modeling error in future study.

Acknowledgements

The financial support from the Fundo para o Desenvolvimento das Ciências e da Tecnologia (FDCT) under grant 052/2005/A is gratefully acknowledged.

REFERENCES


