Optimizing neural networks by genetic algorithms for predicting particulate matter concentration in summer in Beijing

WANG Fang, CHENG Shui-yuan, LI Ming-jun, FAN Qing, CHEN Dong-sheng, WANG Hai-yan, ZHANG Wei

College of Environmental & Energy Engineering, Beijing University of Technology, Beijing 100124, P. R. China
Beijing Academy of Environmental Sciences, Beijing 100037, P. R. China
Ministry of Transportation’s Waterborne Transportation Institute, Beijing 100088, P. R. China

Received 12 May 2010; received in revised form 22 July 2010

Abstract: We developed and tested an improved neural network to predict the average concentration of PM$_{10}$ (particulate matter with diameter smaller than 10 µm) several hours in advance in summer in Beijing. A genetic algorithm optimization procedure for optimizing initial weights and thresholds of the neural network was also evaluated. This research was based upon the PM$_{10}$ data from seven monitoring sites in Beijing urban region and meteorological observation data, which were recorded every 3 h during summer of 2002. Two neural network models were developed. Model I was built for predicting PM$_{10}$ concentrations 3 h in advance while Model II for one day in advance. The predictions of both models were found to be consistent with observations. Percent errors in forecasting the numerical value were about 20%. This brings us to the conclusion that short-term fluctuations of PM$_{10}$ concentrations in Beijing urban region in summer are to a large extent driven by meteorological conditions. Moreover, the predicted results of Model II were compared with the ones provided by the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. The mean relative errors of both models were 0.21 and 0.26, respectively. The performance of the neural network model was similar to numerical models, when applied to short-time prediction of PM$_{10}$ concentration.

Keywords: PM$_{10}$ concentration; neural network; genetic algorithm; prediction

1 Introduction

The adverse effects of ambient particulate matter (PM) have become a well-recognized problem in environmental sciences. Besides the reduction of visibility and the deposition of trace elements, the direct impact on human health via inhalation is an important issue [1]. Accurate air pollution prediction is thus important to protect ourselves from air pollution in advance.

Nowadays, the most frequently used tools for atmospheric pollution prediction include statistics forecast and numerical forecast. Statistical prediction is the most traditional prediction method. It is based on the regression analysis and directly connects meteorological conditions to the level of pollutants. This method lacks consideration of physical and chemical mechanisms of air pollutants. The prediction accuracy will decrease when the weather condition becomes unstable. Numerical forecast can provide 3D forecasting results for different pollutants. However, it is time-consuming and requires meteorological and emission databases with high resolution to initialize and run the model, which is hard to realize in small cities with poor monitoring devices.
Since their first application for ambient pollutant concentrations modeling [2], neural networks have been used for the forecasting of a wide range of pollutants and their concentrations at various time scales, with very good results [3-5]. In recent years, the use of neural networks has been extended also to the prediction of PM mass concentrations [6]. It is supposed that modeling PM concentrations should be more difficult as compared to the forecasting of common gaseous pollutants due to the complexity of the processes, which control the formation, transportation and removal of aerosol in the atmosphere. This is a good reason for which neural networks are expected to produce good predictive results, given their ability to capture the highly non-linear character of those processes [7]. The findings of numerous research studies exhibit that the performance of neural networks is generally superior in comparison to traditional statistical methods, such as multiple regression, classification and regression trees and autoregressive models [8].

The diurnal cycles of PM$_{10}$ (PM with diameter smaller than 10 µm) concentrations frequently reveal large variations within a day. It has been reported that health effects are more linked with short-term exposure, rather than with the 24 h averaged exposure [9]. There have been only a few published research studies involving the prediction of short-term PM concentrations. As an application, we developed an improved neural network to predict average PM$_{10}$ concentrations several hours in advance in Beijing urban areas.

2 Study area and data collection

Beijing is the political and cultural center of China. During the past decades, PM$_{10}$ pollution has become an important environmental concern for Beijing. The annual average of PM$_{10}$ concentration in Metro Beijing from 2000 to 2005 is substantially higher than the National Air Quality Standard of 100 µg m$^{-3}$. Many studies focus on the understanding of the PM phenomenon in Beijing and the ability to forecast ambient PM concentrations [10-11].

This study is based upon the PM$_{10}$ monitoring data from seven monitoring sites in Beijing urban areas (Fig. 1). Since our concern is to predict PM$_{10}$ concentrations in urban Beijing, the average value of the 7 monitoring stations covering urban Beijing was selected. Meteorological observation data from MICAPS (meteorological information comprehensive analysis and process system) were recorded in a 3-hourly basis for surface during July and August 2002.

3 Neural network modeling system

3.1 Neural network models

Artificial neural network models are computer programs that are designed to emulate human information processing capabilities such as knowledge processing, speech, prediction, classifications, and control. They have generated increasing acceptance in various engineering fields [3,12].

The major building block for any neural network architecture is the processing element or neuron. These neurons are located in one of three types of layers: the input layer, the hidden layer, or the output layer. The input neurons receive data from the outside environment, the hidden neurons receive signals from all of the neurons in the preceding layer, and the output neurons send information back to the external environment. These neurons are connected together by a line of communication called connection [3]. The most frequently used training algorithm is the so-called back propagation. This kind of training is relatively easy and offers good support for prediction applications. In this study, the neural network is based upon back propagation algorithm.

3.2 Genetic algorithm (GA) initial weights and thresholds selection

The randomly selected initial weights and thresholds of neural networks are lack of reasonable basis, thus
there is a small possibility to get the global optimal result. In the past few years, GAs were frequently used in approximation problems for the selection of initial parameters in various fields [7,13], and therefore obtained the optimal values. The multi-point search procedure in the GA focuses its attention on the most promising parts of the solution space and, consequently, a global, near optimal value can be rapidly and efficiently sought from a very large search space by utilizing natural evolution based on genetic operators. GAs were used for finding optimal initial weights and thresholds of the neural networks in this study.

To use GAs, an “individual” for evolution should be defined at the first step. Each individual represents a candidate for the optimal value which is determined by randomly selected initial weights and thresholds. Those initial weights and thresholds are then orderly arrayed. The procedure of selection drives the retained for the next generation based on “the elitist-strategy selection”. The procedure optimized by GAs. The procedure is as follows:

**Step 1.** At the first step, an initial population consisting of 50 types of individuals is generated at random. The individuals are all coded as real numbers, each of which is treated as a chromosome in GAs. We defined the number of nodes in the hidden layer as $S_1$, the number of nodes in the output layer as $S_2$, and the number of nodes in the input layer as $R$. The length of the chromosome is $S\{S=R*S_1+S_1*S_2+S_1+S_2\}$.

**Step 2.** The fitness values $F\{F=1/\sigma, \sigma \text{ is the network error}\}$ of all individuals are calculated using the neural network model, and their performances are evaluated. The individuals with higher fitness values are selected, and they pass down to the next generation.

**Step 3.** Crossover and mutation operations are applied to the individuals selected at random and new individuals are generated. Superior individuals are selected and retained for the next generation based on “the elitist-strategy selection”. The procedure of selection drives the solution search process towards more fit members, according to crossover rate and mutation rate.

Steps 2 and 3 are repeated until fitness continues to generate the same maximum value with increasing the generation number. An optimal value is given by the individual with the maximum fitness.

**Step 4.** The neural network starts to training with normalized input data and GAs optimized initial weights and thresholds. Network errors of every time point are calculated. This step repeats until the network errors meet the termination criteria. Finally, the results are renormalized as output.

### 3.3 Selection of input parameters

It is known that the temporal and spatial variations of PM concentrations are governed by a complex interplay of many parameters. Atmospheric particulates can be both of primary or secondary origin. The primary particulates are mainly emitted by anthropogenic sources like mechanical friction, smelting or combustion of fossil fuels, but also natural phenomena like wildfires can emit PM. Secondary particulates originate from chemical reactions, condensation and coagulation in the atmosphere. This formation is influenced by concentrations of other atmospheric pollutants and by meteorological conditions like humidity and temperature. The amount of atmospheric PM is further determined by deposition and transport by winds [1]. While most of the precipitation in Beijing occurring in July and August, rainfall is also a significant factor in this study. Moreover, PM$_{10}$ concentration is also influenced by mixing layer height and solar radiation [7], which are unavailable from observation stations.

Multiple regression was applied to the selection of input parameters. Results show that average concentration of PM$_{10}$ in the first several hours and temperature of the forecast time are two main leading factors to the predicted PM$_{10}$ concentration. Finally, seven parameters were selected from more than ten items of meteorological observation data by ways of multiple regression and experience. The input parameters are listed in Table 1.

### 3.4 Softwares

Software applications used in the present study include MATLAB 7.0. Functions in neural network toolbox and GA toolbox of MATLAB were used for the development of the neural network models and the GA optimization procedure.

### 4 Results and discussion

#### 4.1 Model I

492 groups of data set was selected and divided in separate subsets for the models’ development and evaluation. The training of the neural networks was conducted using the bulk of the data set (396 groups of data set from Jul. 1 to Aug. 19), while the remaining was used as testing data (96 groups of data set from Aug. 20 to Aug. 31). The output is the PM$_{10}$ average concentration every 3 h.
Fig. 2 Flow chart of the neural network procedure optimized by genetic algorithms

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial conditions</td>
<td>PM₁₀&lt;sub&gt;t-1&lt;/sub&gt;</td>
<td>Average PM₁₀ concentration at hours t-1, μg/m³.</td>
</tr>
<tr>
<td>Future conditions</td>
<td>Wind speed &lt;t&gt;</td>
<td>Wind speed at hours t, m/s. (MICAPS observation of wind speed at 10 m height)</td>
</tr>
<tr>
<td></td>
<td>Wind direction &lt;t&gt;</td>
<td>Wind direction at hours t, deg. (MICAPS observation of wind direction at 10 m height)</td>
</tr>
<tr>
<td></td>
<td>Relative humidity &lt;t&gt;</td>
<td>Relative humidity at hours t, %. (MICAPS observation of relative humidity at 1.5 m height)</td>
</tr>
<tr>
<td></td>
<td>Cloud cover &lt;t&gt;</td>
<td>Cloud cover at hours t. (MICAPS observation of total cloud amount)</td>
</tr>
<tr>
<td></td>
<td>Temperature &lt;t&gt;</td>
<td>Temperature at hours t, °C. (MICAPS observation of temperature at 1.5 m height)</td>
</tr>
<tr>
<td></td>
<td>Rainfall &lt;t&gt;</td>
<td>Rainfall at hours t, mm. (MICAPS observation of rainfall at 70 m height)</td>
</tr>
</tbody>
</table>

Notes: <t-1> represents the first several hours; <t> represents the forecasted hours; PM₁₀: particulate matter with diameter smaller than 10 µm

Optimal parameters in the neural networks were finalized after repeated computation. The number of the hidden layer node was tested from 1 to 30 and finally set to 10. The expected precise was 0.010 000. Regarding the GA parameters, the initial weights and thresholds were selected in the range of [−1, 1] and the initial population was 50. The selection rate, crossover rate and mutation rate were set equal to 0.09, 0.90 and 0.08, respectively.

In the process of network training, the network error corresponding to the initial weights and thresholds of the original network was 0.317 941. It achieved 0.093 316 after 80 times of training and met the expected precision after 163 times of training. The prediction errors (mean relative error) of these two models were 0.31 and 0.26, respectively. It is obvious that strinignty of network and the accuracy of prediction are effectively improved after the GA optimizing procedure. The result of the optimized neural network simulation was presented in Fig. 3. It revealed that the model successfully matched the trend of pollutants concentration although slightly delayed.

The same network system was then applied to simulate PM₁₀ concentrations in summer 2004 and its effectiveness was further proved (Fig. 4). Here, the testing data is from Aug. 20 to 31 in 2004.
4.2 Model II

To predict the next-day average concentration of PM$_{10}$ and make a comparison with numerical models, we made a few alterations to the input and output of the network. The former 50 days of datasets (data from Jul. 1 to Aug. 19) were selected as training data, and the left 12 days of datasets (data from Aug. 20 to 31) as testing data. The output was the next-day average concentration of PM$_{10}$. The same procedures used in Model I were then used in the development of Model II. The best results for Model II were obtained when the network completed 80 epochs.

The Models-3 Community Multiscale Air Quality (CMAQ) modeling system, as a part of the National Basic Research Program of China (973 Program, 2005CB724201), was considered as a reference for comparison with the improved neural network models [14]. The three dimensional meteorological fields required by CMAQ are provided by the MM5 for the outer domains (D-1 and D-2), and the ARPS (advanced regional prediction system) for the innermost domain (D-3). Fig. 5 shows the modeling domain. It had a nested system of 36-km-grid (D-1), 12-km-grid (D-2) and 4-km-grid (D-3) centered over Beijing region. Emission inventories were obtained from provincial environmental protection bureaus of the neighboring regions of Beijing, while the detailed ones for the innermost domain (Beijing region) were from CSSBAP (the Control Strategy Study of Beijing Air Pollution, 1999-2000) [10].
The predicting results of the two models are shown in Fig. 6. Prediction errors (mean relative error) of the optimized neural network model and the CMAQ model system are 0.21 and 0.26, respectively. It shows that the neural network is of similar prediction accuracy compared with CMAQ, when applied to the prediction of one day ahead PM$_{10}$ concentrations in urban areas of Beijing in summer.

Fig. 5 Three-level nested modeling domain and terrain height

Fig. 6 Comparison between results of the neural network model and the Models-3 Community Multiscale Air Quality (CMAQ) modeling system

5 Conclusions

In this paper, a modeling system was developed to predict PM$_{10}$ short-term concentrations, in order to investigate the potential of artificial neural network methods. GAs were used to optimize the initial weights and thresholds of the neural network in simulation. Astringency of the neural network and the accuracy of prediction were effectively improved.

We identified the local variables that are relevant to the prediction of PM$_{10}$ short-term concentrations in summer in Beijing urban areas. Those variables can be used as reference to define air quality. Accordingly, meteorological forecast results of the forecasted time play a key role in PM$_{10}$ concentration prediction by ways of neural networks.

We compared an early established CMAQ modeling system with the neural network models in the one-day-ahead prediction of PM$_{10}$ concentrations. The results show well prediction capability of both models. When applied in short-term PM$_{10}$ concentration prediction, the neural network, which is more convenient and time-saving, is of similar prediction accuracy compared with CMAQ. Neural network is a useful tool for air pollution management in urban areas.

Since meteorological conditions in each season in China are quite different, PM$_{10}$ prediction models should be developed for each season separately. Neural network prediction models were also established in other seasons. From the multiple regression results, we found that there are differences in decisive meteorological conditions of different seasons. For example, rainfall amount is important for PM$_{10}$ prediction in summer, but not important in spring.

Meteorological models (i.e. MM5 (the fifth-generation NCAR / Penn state mesoscale model), WRF (weather research and forecasting model)) can be applied to provide the meteorological inputs for neural networks. The accuracy of the meteorological forecast directly influences the accuracy of air quality forecast results. In real-time forecasting conditions, some compromise in performance should be expected, due to the possibility of less accurate meteorological forecasts. Therefore, the research effort in the immediate future should be supplemented by a sensitivity analysis for the meteorological predictors and by the development of models based on forecasted meteorological parameters.

References

39(18): 3279-3289.


Edited by XUE Jing-yuan