Forecasting Urban Air Quality via a Back-Propagation Neural Network and a Selection Sample Rule

Yonghong Liu 1,* Qianru Zhu 2, Dawen Yao 1 and Weijia Xu 3

1 School of Engineering, Sun Yat-Sen University, Guangzhou 510275, China; E-Mail: yaodawen@mail2.sysu.edu.cn
2 Guangdong Provincial Academy of Environmental Science, Guangzhou 510045, China; E-Mails: zhuqr2006@163.com (Q.Z.); xuwjia@mail.sysu.edu.cn (W.X.)
3 Institute of Advanced Technology, Sun Yat-Sen University, Guangzhou 510275, China

* Author to whom correspondence should be addressed; E-Mail: liu_its@163.com.

Academic Editor: Rebecca Sheesley

Received: 28 April 2015 / Accepted: 24 June 2015 / Published: 9 July 2015

Abstract: In this paper, based on a sample selection rule and a Back Propagation (BP) neural network, a new model of forecasting daily SO2, NO2, and PM10 concentration in seven sites of Guangzhou was developed using data from January 2006 to April 2012. A meteorological similarity principle was applied in the development of the sample selection rule. The key meteorological factors influencing SO2, NO2, and PM10 daily concentrations as well as weight matrices and threshold matrices were determined. A basic model was then developed based on the improved BP neural network. Improving the basic model, identification of the factor variation consistency was added in the rule, and seven sets of sensitivity experiments in one of the seven sites were conducted to obtain the selected model. A comparison of the basic model from May 2011 to April 2012 in one site showed that the selected model for PM10 displayed better forecasting performance, with Mean Absolute Percentage Error (MAPE) values decreasing by 4% and R² values increasing from 0.53 to 0.68. Evaluations conducted at the six other sites revealed a similar performance. On the whole, the analysis showed that the models presented here could provide local authorities with reliable and precise predictions and alarms about air quality if used at an operational scale.
1. Introduction

Air quality has recently become a serious issue in several of the large cities in China. This problem has significant potential for adverse impacts on human health and the environment [1–3]. Therefore, it is extremely important to accurately forecast the concentrations of pollutants to provide guidance for travel advice and governmental policies.

Forecasting the concentrations of air pollutants represents a difficult task due to the complexity of the physical and chemical processes involved. However, many researchers have been focusing on these types of forecasts [4–8]. The most common forecasting approaches are numerical models and statistical models. Numerical models do not require a large quantity of measured data, but they demand sound knowledge of pollution sources, the chemical composition of the exhaust gases, and the physical processes in the atmospheric boundary layer. This crucial knowledge is often limited. Thus, approximations and simplifications are often employed in the modeling process.

In contrast, statistical models usually necessitate a large quantity of measurement data under a large variety of atmospheric conditions. By applying regression and machine learning techniques, a number of functions can be used to fit the pollution data in terms of selected predictors. Neural networks, a subset of statistical models, are usually presented as systems of interconnected neurons that can compute values from inputs by feeding information through the network. Unlike other statistical models, neural networks make no prior assumptions concerning the data distribution. They can model highly nonlinear functions and can be trained for accurate generalization. These features of the neural network make it an attractive alternative to numerical and other statistical models [9–12].

There have been many applications of neural networks in air quality forecasting since the 1990s, and researchers have obtained fairly good results [13–16]. Despite the successful applications of neural networks in the area of atmospheric science, the method has its own weakness and limitations. Studies have shown that there are three main factors that affect neural network effectiveness: network topology, learning algorithm, and learning samples [17,18]. Previous research mainly concentrated on the network structure and learning algorithm, which improved the forecasting accuracy of the network [19–24]. However, when improvements in the network structure and learning algorithm reach a certain degree, improvements in the accuracy of the air quality forecasting models plateau. Therefore, the selection of learning samples has become a vital factor that determines the mapping ability and generalization of the network. This is because the selection can ensure the representativeness of the learning samples and remove unnecessary interference, and thereby improve the forecasting accuracy of the model. Harri Niska et al. [21] used a genetic algorithm for selecting the inputs and designing the high-level architecture of a multi-layer perceptron model for forecasting NO₂ concentrations. Sousa et al. [22] predicted hourly ozone concentrations based on feed-forward artificial neural networks using principal components as inputs, and they improved the predictions of models by reducing their complexity and eliminating data collinearity.
The main objectives of this paper are to develop a sample filter method for the prediction of the daily NO$_2$, SO$_2$, and PM$_{10}$ concentration in the Guangzhou Pearl River Delta region based on a similarity principle of weather and pollutant background concentration. During the development of the prediction models, the selection of parameters is conducted by means of sensitivity experiments and the Back Propagation (BP) neural network is used for data-driven computation. The above actions are all part of an integrated environmental strategy designed and run by the local authorities of Guangzhou, according to the demands of the Action Plan on Prevention and Control of Air Pollution. Currently, this action plan is the most rigorous and systematic framework for improving air quality in China.

2. Data

A significant quantity of observational data under a wide variety of atmospheric conditions was required for this study. The dataset in this paper includes meteorological parameters and pollutant concentrations in Guangzhou, which is located in the south central part of Guangdong Province, China (23°06′ N Latitude, 113°15′ E Longitude).

Real-time monitoring meteorological parameters, including temperature, wind speed, wind direction, rainfall, atmospheric pressure, relative humidity, and solar radiation intensity, were obtained from an automatic air quality monitoring station at Sun Yat-Sen University, located in the Haizhu District of Guangzhou City. Forecasting meteorological data, including temperature, wind speed, wind direction, and rainfall, were obtained from Guangzhou Weather Forecasts [25]. All the data were processed into the daily mean value as needed, according to the National Ambient Air Quality Standards (GB 3095-2012) issued by Environment Protection Administration (EPA) of China [26]. The monitoring meteorological data were used as historical meteorological data in the model, and the forecasting meteorological data were used as the meteorological data of the forecasting day. To reduce the interference of different geographic locations on the monitoring meteorological data, pollutant concentration forecasting of seven state-controlled air quality monitoring sites in urban Guangzhou was performed. Thus, the applied monitoring data of the atmospheric environment were derived from the daily pollutant concentration data from seven state-controlled air quality monitoring sites as reported by the Guangzhou Environmental Protection [27]. These state-controlled air quality monitoring sites are the Guangya Middle School (Num. 1), the Guangzhou No. 5 Middle School (Num. 2), the Guangzhou Environmental Monitor Station (Num. 3), the Experimental Kindergarten of Tianhe Vocational School (Num. 4), Luhu Park (Num. 5), Guangdong University of Business Studies (Num. 6), and the Guangzhou No. 86 Middle School (Num. 7). The data span the period from January 2006 to April 2012, and a total of 23,195 valid samples were used for the paper.

3. Methods

In view of the small variation in weather during our study period, a similarity principle of weather and concentration parameters was applied. The multilayer selection rule for historical samples from Guangzhou was then constructed. This step is very important for the development of predictive models. The selection of historical samples can improve the similarity between the occurrence of historical pollution and future pollution, and a proper selection can improve the efficiency of
data-driven models (e.g., BP neural networks). This is also in line with the pollution formation, where the main factor affecting the diffusion and transport of pollutants is the different meteorological parameters, and every meteorological parameter has a different influence on NO$_2$, SO$_2$, and PM$_{10}$ [28,29]. Thus, the sample selection was based on meteorological similarity and the consistency of the variation trend. The rule was divided into two parts, namely the identification of meteorological parameter similarity and the consistency of the variation trend, i.e., the identification of similarity in background concentrations.

First, a comprehensive correlation analysis of pollutant concentration and meteorological parameters was performed to determine the key factors of the selection rule, and these parameters were also used as inputs into the BP neural network. Next, the three-layer selection sample rule was applied. Finally, we utilized the improved BP neural network for data-driven computation to establish the air quality forecasting model of urban Guangzhou.

3.1. Identification of the Key Factors

A comprehensive correlation analysis of pollutant concentration and meteorological factors was conducted. The number of related days was set to two: the meteorology for the forecasting day and for the day before the forecasting day. Meanwhile, the daily mean value of pollutant concentration two days before the forecasting day was used as an input factor in an attempt to counteract the lack of pollutant emission source data.

A comprehensive analysis of pollutant concentration and meteorological factors was conducted for different pollutants, mainly through correlation analysis and weight analysis of the influencing factors in each pollution scenario. The analysis was intended to identify the degree of influence of each meteorological factor on pollutants, thus resulting in the selection of the factors with the greatest impact on pollutants and the allocation of the corresponding influencing weights. The correlation analysis started with the comparison of two typical pollution scenarios, namely, the ascending or descending periods of each pollutant, and the serious pollution or slight pollution periods. In this way, the degree of influence that the meteorological factors had on pollutants under these two situations was obtained. The average value of the two scenarios was calculated and multiplied with a correlation coefficient to obtain the comprehensive weight of the influence of each meteorological factor on different pollutants.

The ascending and descending periods of each pollutant are defined as the periods when the change in the pollutant concentration between consecutive days exceeds 0.05 mg/m$^3$. Serious pollution or slight pollution are defined as periods when the Air Pollution Index of the pollutant exceeds 100 or is lower than 20, respectively.

The identification of the influencing weight of each meteorological factor under the above-mentioned periods was achieved using the following steps:

(a) Obtaining the representative data for the meteorological factor

The specific data include the average value of the ascending period $M_{ua}$, the average value of the descending period $M_{ud}$, the maximum value of the analysis period $M_{\text{max}}$, the minimum value $M_{\text{min}}$ of the analysis period, and the overall average value $M_{\text{adv}}$. The $i$ represents the specific meteorological factor.
(b) Numerical normalization

(c) Variation analysis of the meteorological factor ($D_i$)

\[ D_i = \frac{M_{w} - M_{nd}}{M_{adv}} \]  

(1)

(d) Computation of the influencing weight

\[ w_i = \frac{D_i}{\sum_{i=1}^{n} D_i} \]  

(2)

Finally, the comprehensive influencing weights between meteorology factors and pollutant concentrations were determined by the following equation:

\[ r = R \times (w_1 + w_2) / 2 \]  

(3)

where $r$ is the comprehensive influencing weight between the meteorology factor and the pollutant concentration; $R$ is the correlation coefficient between the meteorology factor and the pollutant concentration; $w_1$ is the influencing weight in the ascending or descending period; and $w_2$ is the influencing weight in the serious or slight pollution periods.

3.2. A Selection Sample Rule Based on the Similarity Principle

Multiple meteorological factors create a variety of meteorological parameter spaces that impose different impacts on the transport and diffusion of pollutants. During air quality forecasting, if the appropriate meteorological space is found, the intrinsic relationship between multiple physical quantities and the pollutant will have a reference. An appropriate set of samples was selected for the main influencing factors such that forecasting could be targeted, and the mapping ability and generalization of the network could be improved. Thus, three-layer sample screening principles based on meteorological similarity criteria were proposed.

3.2.1. The Basic Description

The first level of screening identifies samples where the similarity of each meteorological factor reaches a certain threshold value range. The screened samples should conform to the following formula:

\[ \Delta y_j \leq y_{j, set} \text{, where, } \Delta y_j = |y_{j, pre} - y_{j, sam}| \]  

(4)

where $y_{j, pre}$ is the meteorological factor on the day of forecasting; $y_{j, sam}$ is the meteorological factor of the sample; $\Delta y_j$ is the meteorological similarity of the meteorology factors between the sample and the day of forecasting; $j$ is the specific meteorological factor; and $y_{j, set}$ is the threshold value screened by the meteorological factor, forming a primary threshold matrix $Y$. In this matrix, the threshold value can change dynamically according to the sample size demanded.

The second level of screening applies a threshold value range for total weighted meteorological similarity. The screened samples should conform to the following formula:
\[
S \leq S_{set}, \text{ where, } S = \sum_{j \leq M_{num}} (w_j \cdot \Delta y_j)
\]

where \(S\) is the entire meteorological similarity; \(S_{set}\) is the threshold value screened by the entire meteorological similarity; \(w_j\) is the weight of each meteorological factor, forming the weight matrix \(W\); and \(M_{num}\) is the number of meteorological factors.

The third level of screening identifies the \(n\) samples with the highest meteorological similarity. The screened samples should conform to the following formula:

\[
Q_{num} \leq n
\]

where \(Q_{num}\) is the number of samples in the sequenced sample column, and \(n\) is the number of samples needed.

Among these criteria, the selection of the weight matrices and the threshold matrices is key to obtaining high quality samples. Hence, the following identification approaches for weight matrices and threshold matrices were adopted.

3.2.2. Identification of \(w_j\)

The establishment of the weight matrix \(w_j\) was integrated with the selection of model input factors, and a comprehensive correlation analysis of pollutant concentration and meteorological factors was performed. While choosing the input parameters of the neural network, the weight matrix of the selection sample rule was also established.

3.2.3. Identification of \(y_{set}\)

The establishment of the threshold matrix \(y_{set}\) was accomplished via the orthogonal test method, which is a highly efficient experimental design method used for the arrangement of multi-factor experiments and the search for optimal horizontal combinations [30]. For the different pollutants, we set different levels of factors and selected some representative experimental points (horizontally mixed) for the experiments. The optimal horizontal combination was selected to generate the threshold matrix of the selection sample rule [31].

Based on the results of the above weight matrix \(w_j\), the tested experimental factors were identified. In accordance with prior knowledge, the level of each experimental factor was confirmed. The minimum absolute error of the forecasting model was adopted as the experimental objective to seek the optimal combination and finally identify the sample optimization threshold matrix.

3.3. Identification of the Variation Trend Consistency

There will be some scenarios in which wind speed decreases in history but increases on the prediction day compared with the previous day, based on the selection rule stated above in Section 3.2. Such a scenario will lead to an error in the prediction model for use in the BP neural network. Therefore, it is necessary to identify the variation trend consistency.

The factors considered were deduced according to the weight matrix of the selection rule (see Section 3.1) and the principles of the pollution formation. The chosen factors were rainfall, wind speed, and background concentration. However, sensibility experiments were still needed to determine the key
factor for NO$_2$, PM$_{10}$, and SO$_2$. The details of the experimental results will be introduced in the following section.

3.3.1. Variation Trend Consistency for Wind Speed

Because wind speed is a vector, wind speed is described as $w_x, w_y$.

$$w_x = w_x \cdot \cos(w_d) \quad \text{and} \quad w_y = w_x \cdot \sin(w_d)$$

where $w_x$ is the recorded wind speed and $w_y$ is the recorded wind direction.

Thus, the steps for the identification of the variation trend consistency for wind speed are as follows:

1. Calculate the variation between the forecasting day and the day before.

$$\Delta(w_x) = \left[\left(w_{x-p}\right)^2 + \left(w_{y-p}\right)^2\right] - \left[\left(w_{x-p-1}\right)^2 + \left(w_{y-p-1}\right)^2\right]$$

2. Calculate the variation between the two adjacent days in the samples selected in Section 3.1,

$$\Delta(w_x) = \left[\left(w_{x-t}\right)^2 + \left(w_{y-t}\right)^2\right] - \left[\left(w_{x-t-1}\right)^2 + \left(w_{y-t-1}\right)^2\right]$$

3. Identify whether the wind speed in the forecasting data shows the same tendency of ascending or descending as that in the selected samples. If the tendency is the same, the samples are reserved; otherwise, the samples are removed.

3.3.2. The Variation Trend Consistency Identification of Rainfall

The variation in the rainfall levels in the forecasting data was calculated using the following formula:

$$\Delta RF_f = RF_p - RF_{p-1}$$

The variation in the historical rainfall levels was calculated using the following formula:

$$\Delta RF_2 = RF_t - RF_{t-1}$$

We then identified whether the rainfall level in the forecasting data showed the same tendency of ascending or descending as that in the sample data. If similar, the samples are reserved; otherwise, the samples are removed.
3.3.3. Similarity Identification of Background Concentration

The following steps were used to conduct the similarity identification of the background concentration:

1. The background concentration on the day of forecasting is calculated as follows:
   \[ BC_1 = 0.6BC_{p,1} + 0.4BC_{p,2} \]  

2. The background concentration in the sample data is calculated as follows:
   \[ BC_2 = 0.6BC_{t,1} + 0.4BC_{t,2} \]  

3. Identify whether the background concentration in the forecasting data and the absolute difference of the background concentration on the day of forecasting is in the range of the threshold value. If they are in the range, the samples are reserved; otherwise, they are removed.
   \[ ABS(BC_1 - BC_2) <= Set \]

3.4. Improvements in BP Neural Network

Due to its strong learning and generalization ability, a BP neural network was used as the data-driven computation method [32]. In this paper, a BP neural network with three layers was applied to predict the daily concentrations of NO\textsubscript{2}, PM\textsubscript{10}, and SO\textsubscript{2}. The layers included an input layer, a hidden layer, and an output layer. The data described in Section 2 were divided into training, validation and test sets. The training and validation sets were from January 2006 to April 2011 in seven air quality monitoring sites, of which 80% of these data were randomly selected for the training set; the remaining 20% of the data comprised the validation set. In addition, the data from May 2011 to April 2012 were used for the test set, aiming to test and compare the model performance in seven air quality monitoring sites. There are two main components affecting pollutant concentration: emission sources and pollutant transmission and diffusion conditions. The key factor that affects pollutant transmission and diffusion in a city is the meteorological conditions. Therefore, the meteorological factors identified in Section 3.1 were considered as the major input factors for the BP neural network. According to the conclusions in the literature [33,34], the daily concentrations of NO\textsubscript{2}, PM\textsubscript{10}, and SO\textsubscript{2} for the two days before the forecasting day were also used as input factors for the BP neural network to reduce the influence for lacking emissions data. The final number of variables used in the input layer (NInput) in each forecast model is shown in Table 1.

The neuron number of the hidden layer is half that of the input layer [35]. Different neural network structures were established for NO\textsubscript{2}, PM\textsubscript{10}, and SO\textsubscript{2}. The neuron in the output layer was regarded as the forecasted daily concentration of NO\textsubscript{2}, PM\textsubscript{10}, and SO\textsubscript{2}.

The training termination conditions in the BP neural network were also changed to improve the overall accuracy of the forecasting model. When the average relative error of all training samples reached a specified error value, the training would cease. The specified error value was determined by experiments for different error. For NO\textsubscript{2}, PM\textsubscript{10}, and SO\textsubscript{2}, the optimal specified error values were 0.5, 0.4, and 0.35, respectively. Every group training sample was processed five times, which means that five groups of models were developed. The model with the least average relative error was selected as the prediction model, reducing the randomness of the BP neural network.
### Table 1. Forecasting results of the seven groups of sensitivity experiments.

<table>
<thead>
<tr>
<th>Pollutants</th>
<th>Experiments</th>
<th>N Input</th>
<th>Mean (mg/m³)</th>
<th>MAE (mg/m³)</th>
<th>MAPE</th>
<th>R</th>
<th>TFA</th>
<th>Ef</th>
<th>Af</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO₂</td>
<td>Basic (Group 1)</td>
<td>10</td>
<td>0.027</td>
<td>0.009</td>
<td>37.4</td>
<td>0.422</td>
<td>0.500</td>
<td>-0.322</td>
<td>1.513</td>
</tr>
<tr>
<td></td>
<td>RF * (Group 2)</td>
<td>10</td>
<td>0.027</td>
<td>0.009</td>
<td>36.6</td>
<td>0.510</td>
<td>0.536</td>
<td>0.010</td>
<td>1.543</td>
</tr>
<tr>
<td></td>
<td>WS (Group 3)</td>
<td>10</td>
<td>0.027</td>
<td>0.010</td>
<td>43.2</td>
<td>0.304</td>
<td>0.464</td>
<td>-0.583</td>
<td>1.693</td>
</tr>
<tr>
<td></td>
<td>BC (Group 4)</td>
<td>10</td>
<td>0.027</td>
<td>0.009</td>
<td>40.3</td>
<td>0.345</td>
<td>0.483</td>
<td>-0.937</td>
<td>1.577</td>
</tr>
<tr>
<td></td>
<td>RF + WS (Group 5)</td>
<td>10</td>
<td>0.027</td>
<td>0.009</td>
<td>38.5</td>
<td>0.430</td>
<td>0.482</td>
<td>-0.192</td>
<td>1.501</td>
</tr>
<tr>
<td></td>
<td>RF + BC (Group 6)</td>
<td>10</td>
<td>0.027</td>
<td>0.011</td>
<td>49.7</td>
<td>0.118</td>
<td>0.464</td>
<td>-1.726</td>
<td>1.575</td>
</tr>
<tr>
<td></td>
<td>WS + BC (Group 7)</td>
<td>10</td>
<td>0.027</td>
<td>0.012</td>
<td>52.8</td>
<td>0.178</td>
<td>0.393</td>
<td>-1.174</td>
<td>1.716</td>
</tr>
<tr>
<td>PM₁₀</td>
<td>Basic (Group 1)</td>
<td>7</td>
<td>0.105</td>
<td>0.025</td>
<td>26.6</td>
<td>0.536</td>
<td>0.492</td>
<td>0.210</td>
<td>1.297</td>
</tr>
<tr>
<td></td>
<td>RF (Group 2)</td>
<td>7</td>
<td>0.105</td>
<td>0.026</td>
<td>28.8</td>
<td>0.476</td>
<td>0.433</td>
<td>0.108</td>
<td>1.319</td>
</tr>
<tr>
<td></td>
<td>WS (Group 3)</td>
<td>7</td>
<td>0.105</td>
<td>0.025</td>
<td>26.2</td>
<td>0.527</td>
<td>0.483</td>
<td>0.190</td>
<td>1.289</td>
</tr>
<tr>
<td></td>
<td>BC (Group 4)</td>
<td>7</td>
<td>0.105</td>
<td>0.024</td>
<td>24.6</td>
<td>0.563</td>
<td>0.500</td>
<td>0.225</td>
<td>1.280</td>
</tr>
<tr>
<td></td>
<td>RF + WS (Group 5)</td>
<td>7</td>
<td>0.105</td>
<td>0.025</td>
<td>27.8</td>
<td>0.479</td>
<td>0.417</td>
<td>0.159</td>
<td>1.315</td>
</tr>
<tr>
<td></td>
<td>RF + BC * (Group 6)</td>
<td>7</td>
<td>0.105</td>
<td>0.023</td>
<td>22.7</td>
<td>0.672</td>
<td>0.550</td>
<td>0.348</td>
<td>1.269</td>
</tr>
<tr>
<td></td>
<td>WS + BC (Group 7)</td>
<td>7</td>
<td>0.105</td>
<td>0.024</td>
<td>26.9</td>
<td>0.581</td>
<td>0.417</td>
<td>0.317</td>
<td>1.290</td>
</tr>
<tr>
<td>NO₂</td>
<td>Basic (Group 1)</td>
<td>10</td>
<td>0.073</td>
<td>0.020</td>
<td>25.0</td>
<td>0.680</td>
<td>0.550</td>
<td>0.261</td>
<td>1.340</td>
</tr>
<tr>
<td></td>
<td>RF (Group 2)</td>
<td>10</td>
<td>0.073</td>
<td>0.020</td>
<td>24.1</td>
<td>0.660</td>
<td>0.533</td>
<td>0.199</td>
<td>1.345</td>
</tr>
<tr>
<td></td>
<td>WS (Group 3)</td>
<td>10</td>
<td>0.073</td>
<td>0.018</td>
<td>22.7</td>
<td>0.702</td>
<td>0.533</td>
<td>0.352</td>
<td>1.291</td>
</tr>
<tr>
<td></td>
<td>BC (Group 4)</td>
<td>10</td>
<td>0.073</td>
<td>0.019</td>
<td>23.7</td>
<td>0.715</td>
<td>0.517</td>
<td>0.337</td>
<td>1.315</td>
</tr>
<tr>
<td></td>
<td>RF + WS (Group 5)</td>
<td>10</td>
<td>0.073</td>
<td>0.018</td>
<td>23.7</td>
<td>0.723</td>
<td>0.617</td>
<td>0.386</td>
<td>1.298</td>
</tr>
<tr>
<td></td>
<td>RF + BC (Group 6)</td>
<td>10</td>
<td>0.073</td>
<td>0.019</td>
<td>24.3</td>
<td>0.716</td>
<td>0.483</td>
<td>0.380</td>
<td>1.306</td>
</tr>
<tr>
<td></td>
<td>WS + BC * (Group 7)</td>
<td>10</td>
<td>0.073</td>
<td>0.018</td>
<td>22.5</td>
<td>0.688</td>
<td>0.567</td>
<td>0.397</td>
<td>1.271</td>
</tr>
</tbody>
</table>

Note: * the Selected Model determined by making experiments.

### 3.5. Indices of Model Evaluation

We used the following indicators to evaluate the models: Mean absolute error (MAE), Mean Absolute Percentage Error (MAPE), Correlation coefficient (R), tendency forecasting accuracy (TFA), Nash–Sutcliffe coefficient of efficiency (Ef), and Accuracy factor (Af) [36]. The TFA is the forecasting accuracy rate determination for the upward or downward trend of pollutant concentrations over two consecutive days on the basis of monitoring results. Ef, an indicator of the model fit, is a normalized measure (−∞ to 1) that compares the mean square error generated by a particular model simulation to the variance of the target output sequence. An Ef value closer to 1 indicates better model performance; an Ef value of zero indicates that the model is, on average, performing only as good as the use of the
mean target value for prediction, and an Ef value < 0 indicates an altogether questionable choice of the model. Af is a simple multiplicative factor indicating the spread of the results around the prediction. The larger the Af value, the less accurate the average estimate.

The MAE, MAPE, TFA, Af and Ef are defined as follows:

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |y_{pre,i} - y_{mon,i}|
\]

(14)

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{|y_{pre,i} - y_{mon,i}|}{y_{mon,i}} \right) \times 100
\]

(15)

\[
TFA = \frac{A}{N}
\]

(16)

\[
Ef_f = 1 - \frac{\sum_{i=1}^{N} (y_{pre,i} - y_{mon,i})^2}{\sum_{i=1}^{N} (y_{pre,i} - \bar{y}_{mon})^2}
\]

(17)

\[
A_f = 10 \sum_{i=1}^{N} \left( \frac{y_{pre,i}}{y_{mon,i}} \right)
\]

(18)

where \(y_{pre}\) and \(y_{mon}\) are the predicted and measured values, respectively, and \(\bar{y}_{mon}\) is the mean of the measured values of the response variable. \(N\) is the total number of the observations. \(A\) is the number of correct forecasts for the upward or downward trend of pollutant concentrations over two consecutive days.

4. Results and Discussion

4.1. The Results of the Sensitivity Experiments in Guangzhou No. 5 Middle School (Num. 2)

As described in Section 3.3, sensitivity experiments were performed to determine the key factors. The data were obtained from the Guangzhou No. 5 Middle School site. Seven group experiments were performed for SO\(_2\), PM\(_{10}\), and NO\(_2\). The first experiment (called “Group 1”) was made by the model based on the selection rules described in Section 3.2. That is to say, Group 1 was run using the Basic Model. Besides these selection rules, the second to fourth experiments were conducted based on the variation trend consistency identification of rainfall (RF), wind speed (WS), and background concentration (BC), while the fifth to seventh experiments were considerations of RF + WS, RF + BC, and WS + BC. These experiments were referred to as Group 2, Group 3, Group 4, Group 5, Group 6, and Group 7, respectively. Table 1 summarizes the results of the seven groups of sensitivity experiments. The models with the best performance were selected (termed the Selected Models).

For PM\(_{10}\), the value of Ef and Af of Group 6 were much closer to 1.0 compared with the other models. Compared with Group 1, the Mean Absolute Percentage Error (MAPE) of Group 6 was 4% lower (0.227), R increased by almost 14%, and TFA increased by nearly 6% (0.550). For NO\(_2\), Group 7 had the best results with an MAPE of only 0.225, an R value of 0.688, and a TFA value of 0.567. The Ef and Af of Group 7 were 0.397 and 1.271, respectively, which were much closer to 1.0 than the other experiments. Group 2 had the most ideal experimental results for SO\(_2\); the MAPE was
0.366, R and TFA were both higher than 0.5, and Ef was the only positive value. In contrast to the PM$_{10}$ and NO$_2$ results, the SO$_2$ experiments based on BC did not produce the best results. This scenario is perhaps due to a non-obvious variation in daily SO$_2$ concentrations.

4.2. Errors of the Selected Models of Num. 2 for May 2011 to April 2012

The forecasting results are shown in a scatter diagram of the predicted versus the observed concentrations (Figure 1). The distribution of SO$_2$ is relatively dispersed, which is due to the diversity of the influencing factors and the complexity of dynamic processes. Singh et al. [36] forecast respirable suspended particulate matter (RSPM), SO$_2$, and NO$_2$. The results showed that compared with the two other pollutants, the degree of dispersion in the scatter diagram of the predicted and predicted SO$_2$ values was higher. Kurt et al. [37] used a neural network to build models of SO$_2$, PM$_{10}$, and CO. The error distribution of the SO$_2$ forecasting model based on the data from two days prior ranged from 37% to 40%, and the model was the least accurate of the three. The distributions of PM$_{10}$ and NO$_2$ were relatively better, i.e., the line fitted the correlation at 0.5 and above. The forecasting results were stable and the model performed well.

![Figure 1](image)

**Figure 1.** (a-c) Scatter plots of predicted versus observed NO$_2$, PM$_{10}$, SO$_2$ concentrations for Num. 2.

Errors in the selected model for Guangzhou No. 5 Middle School (Num. 2) from May 2011 to April 2012 are shown in Figure 2. Overall, the monthly prediction accuracy of SO$_2$ was higher than PM$_{10}$ and NO$_2$, and the NO$_2$ model performed better than the PM$_{10}$ model. The highest errors for SO$_2$,
PM$_{10}$, and NO$_2$ were observed in February, where the daily concentrations were almost the highest due to the bad weather; the BP neural network is not sensitive to extremely high or low values [33,34]. However, the MAPE of the SO$_2$, PM$_{10}$, and NO$_2$ models were 0.383, 0.353, and 0.290, respectively. These MAPE values are acceptable for operational forecasts.

![Figure 2. MAPE of models for Num. 2 from May 2011 to April 2012.](image)

4.3. Errors in the Selected Models for Others Sites

The selected model for SO$_2$, NO$_2$, and PM$_{10}$ was tested in the remaining six sites (detailed description in Section 2) in the urban district of Guangzhou, and a comparison was made between the Selected Model and the Basic Model. The results are shown in Table 2. On the whole, the Selected Model was equal to or better than the Basic Model for SO$_2$, NO$_2$, and PM$_{10}$. As for SO$_2$, the MAPE of the Selected Model decreased from 0.417 to 0.377, the correlation increased from 0.409 to 0.477, the TFA increased from 0.490 to 0.517. In addition, the Ef and Af were closer to 1 compared with the Basic Model. Adding the sample optimization rules to the variation tendency identification of the rainfall level changes improved the forecast accuracy of the different pollutants to different degrees at every site. For PM$_{10}$, the MAPE of the Selected Model was 0.250 for the six sites, which was almost 0.10 lower than that of the Basic model. The correlation was greater than 0.7, and the TFA increased by 24%, from 0.421 to 0.523. Adding the variation tendency identification of the rainfall level changes and the similarity identification of the background concentrations to the model resulted in an effective improvement of the forecast accuracy of PM$_{10}$. Regarding NO$_2$, adding the variation tendency identification of the wind speed changes and the similarity identification of the background concentrations did not greatly improve the forecast results. The Selected Model is useful for the six sites, and the errors of the model are acceptable for application purposes.
Table 2. Comparisons between the Selected and Basic Model in the remaining six sites.

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Site</th>
<th>Model</th>
<th>Mean (mg/m³)</th>
<th>MAE (mg/m³)</th>
<th>MAPE</th>
<th>R</th>
<th>TFA</th>
<th>Ef</th>
<th>Af</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO₂</td>
<td>Num. 1</td>
<td>Basic</td>
<td>0.024</td>
<td>0.008</td>
<td>36.8</td>
<td>0.525</td>
<td>0.506</td>
<td>0.159</td>
<td>1.459</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.024</td>
<td>0.008</td>
<td>34.9</td>
<td>0.614</td>
<td>0.525</td>
<td>0.237</td>
<td>1.451</td>
</tr>
<tr>
<td></td>
<td>Num. 3</td>
<td>Basic</td>
<td>0.027</td>
<td>0.010</td>
<td>43.6</td>
<td>0.418</td>
<td>0.511</td>
<td>-0.164</td>
<td>1.539</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.027</td>
<td>0.010</td>
<td>40.4</td>
<td>0.409</td>
<td>0.475</td>
<td>-0.181</td>
<td>1.548</td>
</tr>
<tr>
<td></td>
<td>Num. 4</td>
<td>Basic</td>
<td>0.023</td>
<td>0.009</td>
<td>44.2</td>
<td>0.394</td>
<td>0.509</td>
<td>-0.301</td>
<td>1.567</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.023</td>
<td>0.009</td>
<td>41.3</td>
<td>0.456</td>
<td>0.527</td>
<td>-0.332</td>
<td>1.541</td>
</tr>
<tr>
<td></td>
<td>Num. 5</td>
<td>basic</td>
<td>0.022</td>
<td>0.007</td>
<td>35.6</td>
<td>0.441</td>
<td>0.455</td>
<td>-0.019</td>
<td>1.468</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.022</td>
<td>0.007</td>
<td>31.6</td>
<td>0.472</td>
<td>0.515</td>
<td>0.059</td>
<td>1.408</td>
</tr>
<tr>
<td></td>
<td>Num. 6</td>
<td>Basic</td>
<td>0.027</td>
<td>0.011</td>
<td>42.8</td>
<td>0.355</td>
<td>0.466</td>
<td>0.055</td>
<td>1.587</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.027</td>
<td>0.010</td>
<td>39.6</td>
<td>0.451</td>
<td>0.508</td>
<td>0.019</td>
<td>1.551</td>
</tr>
<tr>
<td>PM₁₀</td>
<td>Num. 1</td>
<td>Basic</td>
<td>0.067</td>
<td>0.018</td>
<td>32.1</td>
<td>0.604</td>
<td>0.509</td>
<td>0.397</td>
<td>1.335</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.067</td>
<td>0.018</td>
<td>26.8</td>
<td>0.694</td>
<td>0.542</td>
<td>0.459</td>
<td>1.322</td>
</tr>
<tr>
<td></td>
<td>Num. 3</td>
<td>basic</td>
<td>0.067</td>
<td>0.016</td>
<td>24.9</td>
<td>0.713</td>
<td>0.509</td>
<td>0.348</td>
<td>1.354</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.067</td>
<td>0.016</td>
<td>22.7</td>
<td>0.729</td>
<td>0.531</td>
<td>0.487</td>
<td>1.267</td>
</tr>
<tr>
<td></td>
<td>Num. 4</td>
<td>Basic</td>
<td>0.063</td>
<td>0.018</td>
<td>30.4</td>
<td>0.583</td>
<td>0.493</td>
<td>0.301</td>
<td>1.358</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.063</td>
<td>0.019</td>
<td>29.2</td>
<td>0.589</td>
<td>0.492</td>
<td>0.247</td>
<td>1.390</td>
</tr>
<tr>
<td></td>
<td>Num. 5</td>
<td>basic</td>
<td>0.087</td>
<td>0.022</td>
<td>25.4</td>
<td>0.682</td>
<td>0.467</td>
<td>0.408</td>
<td>1.308</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.087</td>
<td>0.022</td>
<td>23.3</td>
<td>0.717</td>
<td>0.525</td>
<td>0.431</td>
<td>1.288</td>
</tr>
<tr>
<td>NO₂</td>
<td>Num. 1</td>
<td>Basic</td>
<td>0.061</td>
<td>0.013</td>
<td>20.9</td>
<td>0.688</td>
<td>0.483</td>
<td>0.392</td>
<td>1.248</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.061</td>
<td>0.013</td>
<td>20.5</td>
<td>0.715</td>
<td>0.500</td>
<td>0.448</td>
<td>1.243</td>
</tr>
<tr>
<td></td>
<td>Num. 3</td>
<td>Basic</td>
<td>0.068</td>
<td>0.016</td>
<td>22.2</td>
<td>0.596</td>
<td>0.463</td>
<td>0.226</td>
<td>1.272</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.068</td>
<td>0.015</td>
<td>21.5</td>
<td>0.676</td>
<td>0.557</td>
<td>0.320</td>
<td>1.266</td>
</tr>
<tr>
<td></td>
<td>Num. 4</td>
<td>Basic</td>
<td>0.052</td>
<td>0.010</td>
<td>21.9</td>
<td>0.685</td>
<td>0.511</td>
<td>0.456</td>
<td>1.232</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.052</td>
<td>0.010</td>
<td>19.3</td>
<td>0.722</td>
<td>0.541</td>
<td>0.502</td>
<td>1.215</td>
</tr>
<tr>
<td></td>
<td>Num. 5</td>
<td>Basic</td>
<td>0.038</td>
<td>0.009</td>
<td>25.8</td>
<td>0.613</td>
<td>0.454</td>
<td>0.363</td>
<td>1.285</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.038</td>
<td>0.009</td>
<td>23.0</td>
<td>0.599</td>
<td>0.462</td>
<td>0.308</td>
<td>1.267</td>
</tr>
<tr>
<td></td>
<td>Num. 6</td>
<td>Basic</td>
<td>0.053</td>
<td>0.014</td>
<td>26.8</td>
<td>0.757</td>
<td>0.497</td>
<td>0.405</td>
<td>1.337</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.053</td>
<td>0.015</td>
<td>24.6</td>
<td>0.728</td>
<td>0.528</td>
<td>0.310</td>
<td>1.334</td>
</tr>
<tr>
<td></td>
<td>Num. 7</td>
<td>Basic</td>
<td>0.041</td>
<td>0.010</td>
<td>27.4</td>
<td>0.668</td>
<td>0.476</td>
<td>0.435</td>
<td>1.305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selected</td>
<td>0.041</td>
<td>0.009</td>
<td>23.1</td>
<td>0.700</td>
<td>0.505</td>
<td>0.465</td>
<td>1.269</td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper, based on a selection sample rule and BP neural network, a new model of forecasting daily SO₂, NO₂, and PM₁₀ concentrations in seven Guangzhou sites was developed.

(1) A meteorological similarity principle was applied in the development of the selection sample rule. Key meteorological factors influencing the daily SO₂, NO₂, and PM₁₀ concentrations were determined and weight matrices and threshold matrices were generated. A basic model was
then developed based on the improved BP neural network. The selection sample rule consisted of three layers.

(2) In improving the basic model, identification of the variation consistency of some factors was added in the rule, and seven sets of sensitivity experiments (one in each of the seven sites) were conducted to obtain the selected model. These experiments determined that the variation consistency of the rainfall level added to the SO2 forecast model, the rainfall level variation tendency and the background concentration similarity identification added to the PM10 forecast model, while wind speed variation identification and background concentration similarity identification added to the NO2 forecast model. The improved BP neural network was also used for data-driven computation.

(3) Evaluations in the site by comparison of the basic model from May 2011 to April 2012 showed the selected model for PM10 displayed better forecasting performance, with MAPE values decreasing by 4% and R2 values increasing from 0.53 to 0.68. The selected model for NO2 had little improvements compared with the basic model, while the MAPE values of the selected model for SO2 were as high as 36.6% with R2 values of 0.51.

(4) Evaluations conducted at the six other sites revealed similar performances. The MAPE values of the selected models for SO2, PM10, and NO2 were 37.7%, 25.0%, and 22.0%, respectively. Of course, the above results showed that the SO2 model may be further improved in future research, by developing a combined model or by considering the interaction of atmospheric pollutants.

Acknowledgments

This work was completely supported by the National Natural Science Foundation of China (No. 51108471).

Author Contributions

Yonghong Liu, Qianru Zhu conceived and designed the model and experiments; Dawen Yao and Weijia Xu collected and analyzed the data; Yonghong Liu wrote the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References


© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).