The Prediction Model of Air Quality in Changsha City Based on Artificial Neural Network

Sicong Li[1, †], Wei Xu[2, *, †]

[1] School of Civil Engineering; Sichuan University; Chengdu; Sichuan; China

[2] School of Mathematics and Statistics; Changsha university of science and technology; Changsha; Hunan; China

* Corresponding Author Email: brucehsu@stu.csust.edu.cn

† These authors contributed equally.

Abstract. The air quality is a hot topic noticed by all the present society. Although modern technology has helped to collect data of the air position, the huge amount of irregular data also makes it hard to detect the pattern of air pollution. Therefore, this research is to apply data mining to the analysis of the air quality. Data mining is a comprehensive mining of information including trends, features and correlations hidden in data. Data mining can extract the value of data which was not known from numerous random data. This research is based on the daily average values of six major pollutants in Changsha from January 1, 2015, to December 31, 2017, and analyzes the relationship between the air quality and socioeconomic factors of Changsha to establish a prediction model using B-P back propagation neural network. The final model of the research fits the actual data well and predicts the future value of API precisely. However, considering the slow convergence rate and the addition of new samples which may affect the previously learnt samples, further exploration on the B-P neural network model can be launched with other mathematical theories.

Keywords: Data mining, Air quality, API, B-P neural network.

1. Introduction

Ambient air is the outdoor air that animals and plants, humans live with. Ambient air quality is to evaluate the pollution degree and clarify the ambient air condition. Ambient air quality standard is the standard stipulated by the national laws of various environmental air quality, which is the upper and lower limit of the average concentration of air pollutants in a certain period of time. Air pollutants are substances produced by human activities and discharged into the atmosphere that are harmful to themselves and the environment. And air pollution can affect human health and comfort, and have a negative impact on the environment. The equivalent diameter is similar to the target to be tested in geometry, optics, electricity and aerodynamics. Particulate matter (particle diameter less than or equal to 10μm), refers to the ambient air equivalent diameter less than or equal to 10 particles, and inhaled particulate matter. Particulate matter (with a particle diameter of less than or equal to 2.5μm) is a particulate matter with an environmental equivalent diameter of 2.5μm, also known as fine particulate matter. The Air Quality Index (API) is an index that quantitatively describes air quality conditions. The Air Quality sub-index (IAPI) is the air quality index for each pollutant[1].

At present, there are still many deficiencies in the research on air quality. For example, the analysis of the factors affecting air quality mostly stays in the descriptive statistical analysis, without further correlation analysis and prediction. Most of the mathematical models for air quality prediction are mostly simple least squares models, without further training and improvement of the model with the training sample. And the air quality has been highly concerned by people, and is closely related to people's lives. Such research will further improve people's lives and the economic development of the country.

Based on the existing literature. This paper will conduct the analysis of the correlation of air quality factors, and build the model for many training learning according to the characteristics of the artificial neural network, and will not stop adjusting the weights of each layer, until the error reaches the

acceptable range. In the end, improve the accuracy of the prediction model. The method mentioned above is called back propagation neural network.

2. Data analysis

2.1. Data Collection

Data of this paper are collected from National Population Health Data Center[2]. The research is based on the data of air quality of Changsha from 2015 to 2017. The data record the index values of PM2.5, PM10, SO2, CO, O3, NO2 and API. Each daily average values is obtained from the hourly average values. The monthly average values and yearly average values can be calculated on the basis of the daily average values. Among these index values, the values of API reflect the comprehensive air quality of Changsha, so API is selected as the output variable. The rest six variables are selected as the input variables. According to the daily average API, the air quality can be divided into six levels: excellent, good, light pollution, moderate pollution, heavy pollution and serious pollution. The API intervals corresponding to these six levels are 0~50, 51~100, 101~150, 151~200, 201~300, 300 and above[3]. A high number of the value of API means the air pollution is serious.

2.2. Data Reprocessing

The data are imported into R studio. There are no missing values upon inspection. Therefore, Excel is used to standardize the data. In excel the formula is used as follows[4]:

$$x_{ij}^{*} = \frac{x_{ij} - min\{x_{ij}\}}{max\{x_{ij}\} - min\{x_{ij}\}} \quad (1)$$

The data is arranged in chronological order from January 1, 2015 to December 31, 2017. The data is used to draw a box plot shown as follows.

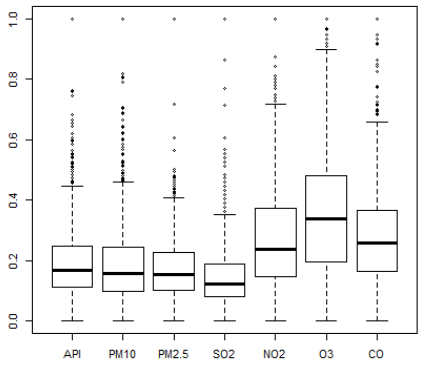


Fig. 1 Box plot of indexes of Changsha air quality data.

Figure 1 represents the box plot of these six index. It can be seen from figure1 that many data exceed the upper limit. However, considering the complicated and uncertain weather changes, abnormal values are not removed. Under the circumstance, the abnormal data are used to calculate.

2.3. Descriptive Statistics

2.3.1. Descriptive Statistics of Air Composition

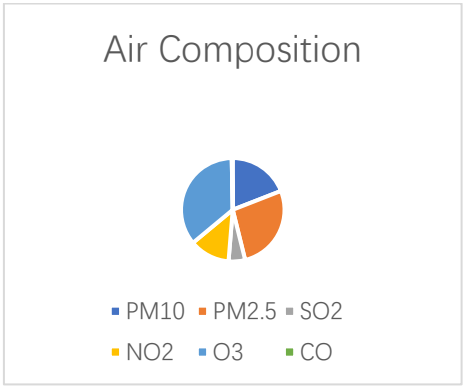


Fig. 2 Pie chart of the composition of the polluting gases in Changsha.

Figure 2 represents the composition of the polluting gases in Changsha. It can be seen from the figure that among the six values, the values of PM10 and PM2.5 take up the most proportion. It indicates that the value of PM10 and PM2.5 may be the most important factor for us to consider. Moreover, the values of NO2 and O3 are also not to be ignored. However, the values of SO2 and CO seem to be small, so they may be less important in our model. Admittedly, they cannot be excluded in the analysis, otherwise the result will be less accurate and convincing[5].

The kurtosis of SO2 is 10.27301 while the kurtosis of API values is 4.5691, which means the distribution of SO2 is quite sharp. That is to say, there are many extreme values in the distribution of SO2 values. Besides, the kurtoses of NO2 and O3 are 0.6148 and -0.2891 respectively, which means the distributions of NO2 and O3 are relatively flat. So, there are not many extreme values.

Through calculation, we've reached the conclusion that the distributions of the values are right skewed since the skewness of all the values is bigger than 0. The skewness of API values is 1.663, while the skewness of SO2 values is 2.324, which is bigger than the other values' skewness, representing the value of SO2 is skewed to a larger extent. That is to say, the distribution of SO2 has a longer tail on the right than the other values.

2.3.2. Distribution of API Values

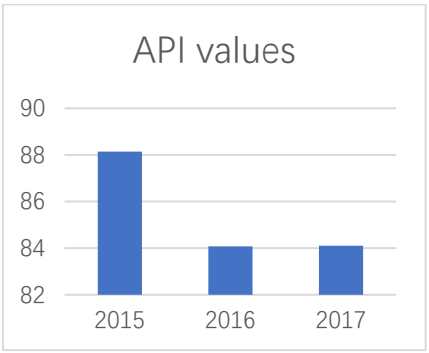


Fig. 3 Average value of API from 2015 to 2017 in Changsha city.

Figure 3 represents the annual average API values in Changsha from 2015 to 2017. Based on the daily average air quality data, the annual average API values can be calculated. Because API index values reflect the comprehensive air quality, comparing the annual average API values can show the changes of air pollution in the three years, as is shown in Figure 3.

On the whole, the average values of API of Changsha in the three years are distributed between 80 and 90. This fact corresponds to an air quality level of good. It can be inferred that air pollution in Changsha is not too serious. Since 2015, it is obvious that the values of API show a downward trend, indicating that the air quality of Changsha has been improved at a significant level. The improvement

is attributed to the unremitting efforts of the government and people in Changsha during the three years.

We use Z-value to calculate the confidence interval of the value of API. With the sample number to be 1053, the average value to be 85.39 and standard deviation to be 45.66, we can calculate that the 95% confidence interval is (82.65, 88.13) and the mean standard error is 1.40. This confidence interval further proves that the air quality in Changsha is good on the whole.

Figure 4 represents the daily API values. Although the annual average values of API decreases, there is a difference in the variances of the API values in the three years. It shows that data differ from the mean at a significant level, which can be seen from Figure 4. Figure 4 is a bar chart of the complete distribution of API values from 2015 to 2017.

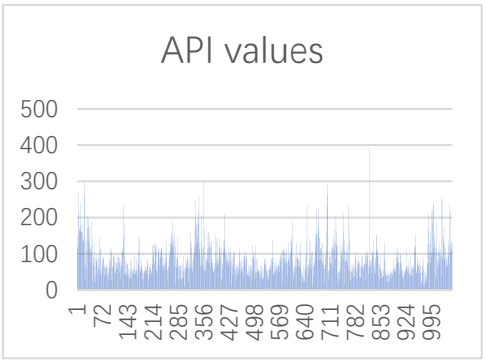


Fig. 4 Bar chart of daily API values of Changsha city

It can be calculated that in 2015, the variance of API values is 2157.054. In 2016, the variance of API values is 1613.211. In 2017, the variance of API values is 2440.234. The variance of API values from 2015 to 2017 is 2054.624. The difference indicates that API values have a greater dispersion in 2017 and a smaller dispersion in 2016. It means the air pollution in Changsha had bigger changes in 2017 and smaller changes in 2016.

2.3.3. Influence of Economic Factors on Air Quality

Economic factors will have a certain influence on the air quality. Considering that the economic data of Changsha are all annual data, we select the annual average API values and five annual economic data of

Changsha from 2015 to 2017. The economic data include GDP, output of secondary industry, green area coverage, total population and automobile ownership ratio. The data is shown in Table 1.

Table 1. Values of GDP, output of secondary industry, green area coverage, population, automobile ownership ratio and API of Changsha City

	GDP	Output	Coverage	Population	Ownership	API
2015	85025984	36083905	12278	6803579	1688299	88.14201
2016	91664027	37650910	12928	6959998	1942362	84.07735
2017	100501981	39912911	14877	7087939	2177513	84.10765

Table 1 represents the values of GDP, output, coverage, population, ownership ratio and API of Changsha[6].

By using the Pearson correlation test in SPSS, it can be found that the annual average API shows a significant negative correlation with GDP and automobile ownership ratio. The correlation coefficient of GDP is -0.819 and the correlation coefficients of automobile ownership ratio is -0.838. There is no significant correlation with green area coverage, total population and output of secondary industry. The relationship strongly proved that although Changsha is developing its GDP and automobile ownership ratio steadily in the three years, it pays attention to develop green economy which will not cause serious air pollution. Meanwhile, it is also a proof that Changsha is encouraging the use of green and low-carbon energy, and is promoting the consumption of new energy vehicles

which produce less air pollution than the traditional ones. In this way, the air quality of Changsha has been improved since 2015[7].

3. Prediction of Api Value based on Back Propagation Neural Network

3.1. Brief description of artificial neural network

Artificial neural network (ANN) is an abstract computing model of the human brain, which is a computer modeling mode of simulating the human brain thinking. Since the 1940s, it was hoped to achieve the simulation of human brain systems through computer programs. Through processing units similar to biological neurons, and through the organic connections between processing units, real-world complex problems such as pattern recognition, associative memory, and optimized computation are solved[8].

At present, the application research of artificial neural network is gradually stepping from artificial intelligence to the data mining field focusing on data analysis, and is widely applied to data classification and regression prediction, and can also be applied to cluster analysis.

B-P neural network is a typical artificial neural network, which is a feedforward multi-layer perceptron model. It is the learning process of error back propagation algorithm, which consists of two processes: forward propagation of information and back propagation of error. The neurons in the input layer receive the input information and pass it to the middle layer; the internal information processing layer is responsible for the information transformation, which can be designed as a single hidden layer or multiple hidden layer structure according to the requirements of information change ability; the last hidden layer to the output layer neurons, after further processing, the output output information processing results to the outside world. When the actual output does not match the desired output, it will begin the back propagation phase. The error is corrected by the output layer to the hidden layer and the hidden layer and the input layer layer by layer. The process of repeated information forward propagation and error back-propagation is the process of continuous weight adjustment of each layer, and also the learning and training process of neural network. This process proceeds until the error of the network output is reduced to an acceptable degree, or when it reaches the predetermined number of learning. Finally, the prediction process is completed by using the completely trained neural network to predict certain value. The B-P back-propagation model is a multi-layer perceptron structure, which contains not only the input and output nodes, but also one layer or multiple layer hidden layer, also known as the multi-layer perceptron model (multilayer perception, MLP)[9].

Based on the analysis of air quality in Changsha City, the Back Propagation Neural Network is used to predict the API value of Changsha City.

3.2. Prediction process of the air quality index

3.2.1. Back Propagation Neural Network modeling thinking

First, the air quality data of Changsha from 2015 to 2017 was standardized, and samples with trainable function are taken as the input and output of the feedforward network. B-P algorithm is used for training these samples. Finally, the nonlinear mapping model between the original API data and the predicted value obtained by the model was obtained to judge the rationality of the model. The types of pollutants are chosen as the hidden layer of the model, and the corresponding input variables are selected according to the hidden layer. And to avoid redundant calculation, 600 data were selected as the training samples in the original 1,051 data. In this model, six input variables are selected as PM2.5, PM10, SO2, CO, O3, and NO2; three hidden layer nodes; and one output variable API. Input variables and hidden layer index are shown in Table 2.

Table 2. Selection of input variables and hidden layer indicators.

Influencing factor (hidden layer index)	Input variables	unit
Particulate matter	PM2.5	$\mu g/m^3$
	PM10	$\mu g/m^3$
Oxide	SO2	mg/m^3
	CO	mg/m^3
	NO2	mg/m^3
Pure substance	O3	mg/m^3

3.2.2. Prediction of the API values based on Back Propagation Neural Networks

The 'neuralnet' function in R Studio's 'neuralnet' package is used to establish the B-P neural network. The specific process is to read the input and output variables in the data set first, standardize the data by Excel before reading, and then set the value of hidden layer as 3, 'threshold' value as 0.01, representing the iteration stops when the maximum adjustment of the weight is less than 0.01. 'Stepmax' value is 2000000, representing when the number of iterations reaches 2000000, it will stop. 'Linear.output' value is 'FALSE', representing the output node is a nonlinear function. 'Learningrate' value is 0.01, and 'algorithm' value is 'backprop', representing the traditional B-P neural network[10].

3.3. Analysis of model prediction results

The output results are shown in Figure 5

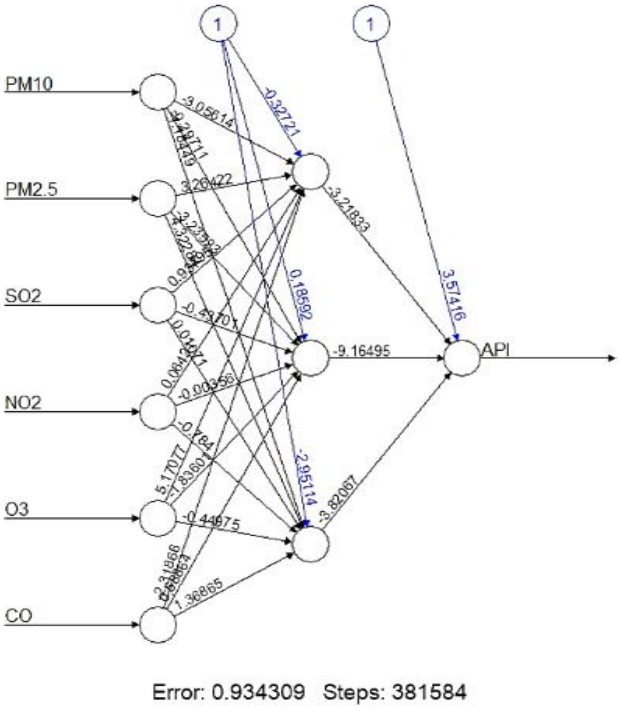


Fig. 5 Three layer B-P neural network structure.

Figure 5 shows that 381584 iterations were performed, and when the iteration finished, the loss function was 0.9343091 and the maximum adjustment of weights was 0.0009. Figure 4 shows the connection weight values of all input nodes (including deviation nodes), hidden nodes, and output nodes.

It is also necessary to consider that what input variables are more important for the prediction of the output variables in application, and they define the generalized weight for measuring the importance of the explanatory variables. In order to show the importance of the input variables more intuitively, the gwplot function is used to plot the scatter plot of specified input variable and its generalized weights at each observation. The drawing results are shown in Figure 6.

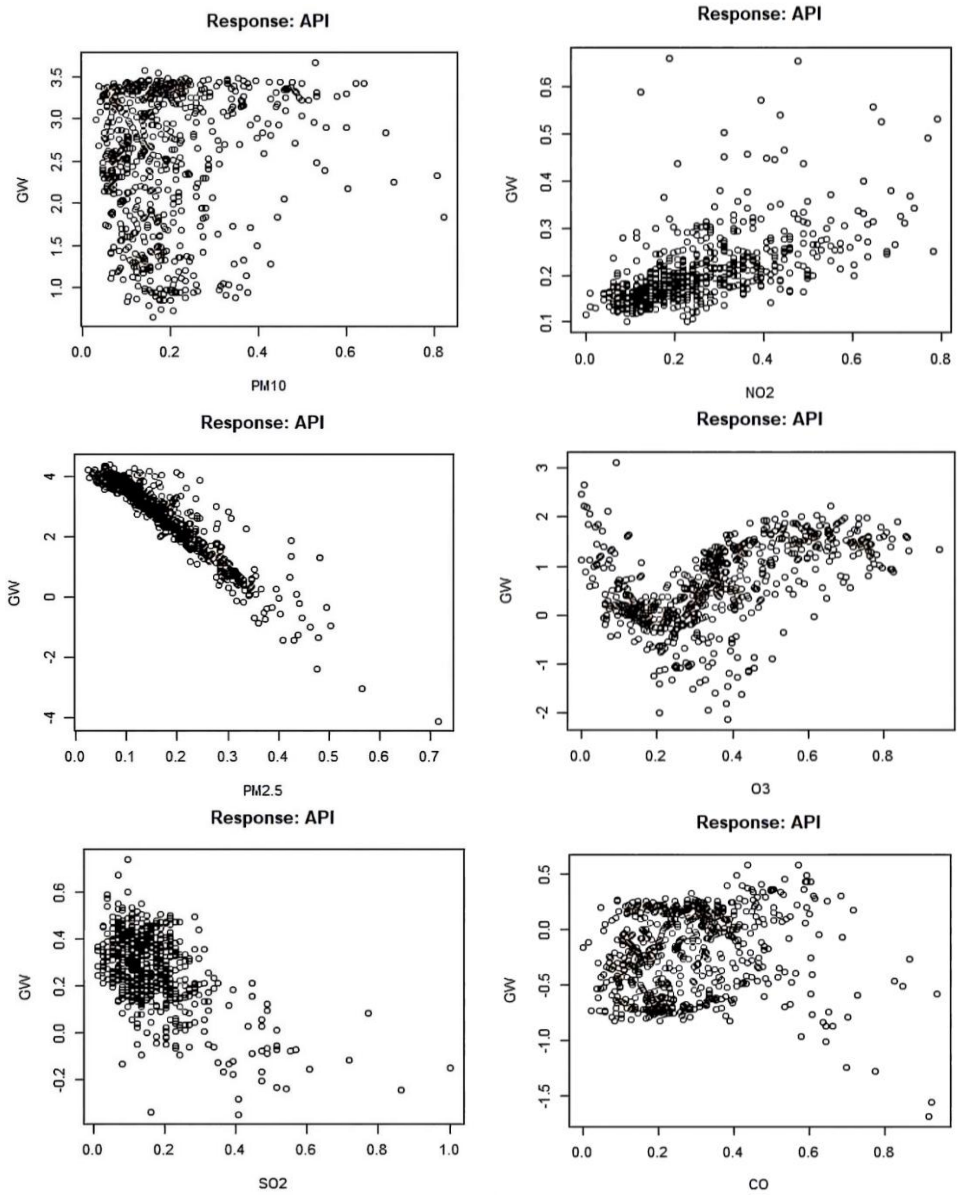


Fig. 6 Plot of generalized wright of six indexes

Figure 6 shows the generalized weight of all six indexes, which presents a certain nonlinear relationship. As a result, all six variables need to be considered when calculating the predicted values of the output nodes.

At this point, the predicted values compared with the actual values are shown in Figure 7

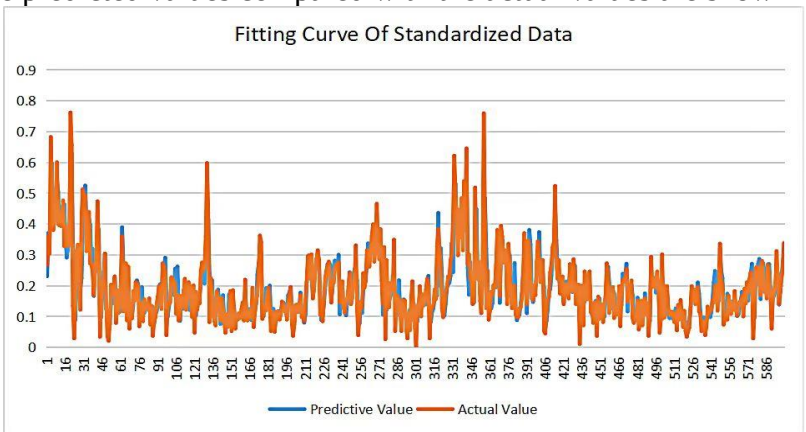


Fig. 7 Fitting curve between the predictive output value of the model and the actual value

We compared the predictive value with the actual values with the fitting curve from the program. It is easy to figure out that the distribution of the average daily API values does not follow a linear distribution, because there are many abrupt changes in the curve. But the neural network system is well fitted. For most of the mutation values, the curve of the predictive value and the actual value basically coincides, showing a high level fitting. The relative error between the predicted output value and the actual value is 28.51% on average through calculation, which indicates that the prediction is relatively reliable. So the prediction of API in the next few years based on the training model is quite convincing.

4. Conclusion

Air quality issues have always been drastically debated among people. It is vital for the society to find a way to predict the air quality. With the predictive value of API through B-P neural network, we can obtain the air quality in the short term, thus providing the government with ways to better set up the economic development strategy of Changsha, environmental protection strategy, people's livelihood development strategy. However, because the convergence rate of B-P neural network is slow, and the addition of new samples will affect the previously learned samples, it is necessary to further explore the practical application model combining B-P neural network model with other mathematical theories. We believe if the B-P neural network model is updated with more accurate calculation, it can have a contribution to improving the air condition in Changsha.

References

[1] Li Yuan. Research on development and monitoring data of daily environmental air quality report [J]. Heilongjiang environmental bulletin, 2022,35 (02): 22-24

[2] Chengdu University of information engineering. Sharing cup version - air pollution data. Data warehouse of national population health science data center phda, 2020. https://doi.org/10.12213/11.A000Q.202008.5.V1.0.

[3] [3] Ambient Air Quality Evaluation Method Based on Class 6 Pollutants in Beijing [J]. Liu Jie, Yang Peng, Lu Wensheng, Wang Jinhai. The Journal of Safety and the Environment.2015(01)

[4] Study on air quality prediction based on the Neural network model [J]. Jin Renhao, Zeng Guojing, Wang Sha. Heilongjiang Science.2021(12)

[5] Xie Yongqin, Liang Xiaolin. Probability theory and mathematical statistics [M]. China Renmin University Press, June 2014

[6] Jin Yongjin. Sampling: Theory and application (2nd Edition) [M]. Higher education press, January 2016

[7] National Bureau of statistics of the people's Republic of China China Statistical Yearbook [M] Beijing: China Statistics Press, 2021

[8] He Xiaoqun, Liu Wenqing. Applied regression analysis [M]. China Renmin University Press, April 2015

[9] Data Mining Fundamental Tutorials, K.P. Soman et al., translated by Fan Ming et al., Mechanical Industry Press, 2010

[10] Zhou weiteng, Han Baoming, Li Dewei. Prediction model of Beijing subway traffic volume based on B-P neural network [J]. Urban rapid rail transit, 2011,24 (02): 50-54

[11] Zhu Shunquan. Application of R software for data statistical analysis [M]. Tsinghua University Press, August 2016