Research on Air Quality Index Prediction Based on Neural Network

Taking Beijing as an Example

Cong Zhao
School of Economics and Management
Beijing Jiaotong University
Beijing, China

Xuemei Li
School of Economics and Management
Beijing Jiaotong University
Beijing, China

***Abstract*—With the continuous improvement of the level of industrialization, the air pollution situation in China has become more and more serious. In many places, extreme weather such as haze has appeared, which seriously threatens people's health. Therefore, it is necessary to establish a scientific and reasonable air quality index prediction model. However, there are significant differences in air quality indices in different quarters, that is, the AQI values are significantly seasonal. Therefore, in order to improve the prediction accuracy, the data of different quarters are distinguished, and models of different quarters are established. In this paper, the principal component analysis method is used to analyze the correlation between API value and various meteorological factors, and the correlation factor is used as the input variable of neural network. The number of neurons in different quarters is determined according to the mean square error, and Bayesian normalization is established. The neural network is a model of the algorithm. Finally, the corresponding model was used to predict the air quality index of the corresponding quarter in 2018 based on the winter, spring, summer and autumn air conditions of Beijing from 2014 to 2017. The results show that the forecasting accuracy of each quarter is 88.27%, 92.28%, 94.04%, and 91.01%, respectively. The prediction accuracy of most studies is 70%~90%, and the prediction accuracy is high, which has certain reference value.**

***Keywords*—neural network; air quality index; linear regression; principal component analysis; Bayesian algorithm**

I. INTRODUCTION

The Air Quality Index (AQI) is an index that quantitatively describes the air quality status. The pollutants involved in the AQI index include six items such as PM10, PM2.5, SO_2, NO_2, CO, and O_3 [1]. However, the air quality is not only affected by the above-mentioned pollutants, but also related to the local meteorological conditions. Under different meteorological conditions, even if the above pollutants are the same, the air quality may be different. At present, the research on environmental quality prediction includes: He Xiaoyun adds time parameter or momentum factor to neural network model [2], Zhang Huiyi uses universal gravitation search algorithm to improve neural network with prediction accuracy of 78.79% [3], Zhang Mengyao improved weighted Markov chain The accuracy of AQI prediction is 85.96% [4], and Niu Yuxia's relative error of genetic algorithm combined with neural network is 20%-40% [5]. Lin, Yi proposed a new air quality prediction algorithm based on cloud model granulation (CMG) and optimized the experimental parameters by grid search, reaching 71.43% prediction accuracy [6]. Kostas Karatzas established linear regression models and artificial neural network models, combined with random forest models for prediction [7]. J. Struzewska selected the parameters predicted by the GEM-AQ model as the predictors of the equation and used a representative time series to construct a regression function to predict the air quality [8]. Haiyan Chen proposed a method of uncertainty allocation (UA) for arbitrary prediction models to minimize the uncertainty of prediction results [9]. At present, the prediction accuracy of most studies is between 70% and 90%, which needs further improvement. At the same time, considering the neural network technology has good mapping ability for uncertain, multi-input and complex nonlinear problems, and superiority in forecasting, I establish a neural network prediction model based on Bayesian normalization algorithm [10].

In this paper, the principal component analysis method is used to analyze the correlation between API value and various meteorological factors and the mean square error to establish a neural network model based on Bayesian normalization algorithm. The model is used in winter and spring of Beijing from 2014 to 2017. Summer and autumn air conditions are the training samples for the corresponding quarterly air quality index for 2018.

II. BEIJING AIR QUALITY ASSESSMENT

According to the Beijing 2018 Air Quality Report, the statistics of Beijing's air quality status in 2018 are shown in "Fig. 1".

Copyright © 2019 The Authors. Published by Atlantis Press SARL.
This is an open access article distributed under the CC BY-NC 4.0 license (http://creativecommons.org/licenses/by-nc/4.0/).

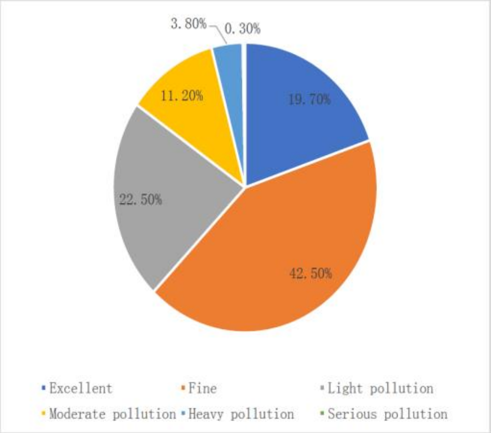

Fig. 1. Beijing air quality distribution map.

It can be found that the air quality in Beijing is mainly excellent, but there is also extreme weather such as heavy pollution. The total number of days above moderate pollution accounts for 15.3% of the whole year. In addition, Beijing SO_2 meets national standards (60µg/m3); NO_2 exceeds national standards (40µg/m3) 5%; PM10 exceeds national standards (70µg/m3) 11%; PM2.5 exceeds national standards (35µg/m3) 46%. Therefore, it is necessary to forecast the Beijing Air Quality Index to provide a basis for the treatment of air pollution.

III. Determine Factors of Influencing Air Quality

The air quality data used in this paper is derived from the data released by the post-weather report, including PM10, PM2.5, SO_2, NO_2, CO, O_3 and other data for each calendar year. The meteorological data comes from the global weather network, which contains a large amount of information such as temperature, wind level, wind direction and weather conditions (cloudy, sunny, rain, snow, etc.).

According to the AQI definition, in addition to the six factors of PM10, PM2.5, SO_2, NO_2, CO, and O_3 involved in air quality assessment, meteorological conditions will also affect air quality. According to the "New Compilation Dictionary", meteorological factors include temperature, pressure, humidity and other phenomena and condensation, precipitation, wind, clouds, atmospheric light and other phenomena. However, in the actual experimental process, the study could not obtain all the factors. Therefore, I choose temperature, humidity, weather conditions, rainfall (snow) and wind as indicators of weather, that is, choose the highest temperature, the lowest temperature, the average temperature, the wind level, day and night weather conditions (such as rain, snow, yin, etc.), wind direction, wind, relative humidity, and rainfall (snow) as an alternative to the air quality model. The meteorological factors are shown in "Table. I".

TABLE I. Symbols and Meanings of Various Meteorological Factors

Factor	Mean
X1	Maximum temperature
X2	Minimum temperature
X3	average temperature
X4	Daytime weather
X5	Daytime weather
X6	wind direction
X7	Wind power
X8	Relative humidity
X9	Rainfall(snow) amount

In the acquired data, the wind direction and weather condition data are qualitative descriptions rather than magnitudes, making it difficult to make direct comparisons. Therefore, I require quantify these qualitative description factors to perform linear regression test [11]. The quantified values of weather conditions and wind direction are shown in "Table II" and "Table III".

TABLE II. Weather Status Quantification Table

Weather	Heavy rain	Rainy	Thunder storm	Light rain	Sleet	Light snow	Sunny	Cloudy	Fog	Overcast	Smog
Quantity	1	2	3	4	5	6	7	8	9	10	11

TABLE III. Wind Vectorization Table

Wind Direction	North	Northeast	East	Southeast	South	Southwest	West	Northwest
Quantity	1	2	3	4	5	6	7	8

The experimental data for linear regression test is the data from 2014 to 2017. Therefore, in order to better predict the API value, this section uses SPSS to analyze the correlation between API values and various meteorological factors for the meteorological characteristics, and to find the connection between API and various meteorological factors. The linear correlation between the API and each meteorological factor is described by calculating the Pearson correlation coefficient and the Spearma rank correlation coefficient. The calculation results are shown in "Table IV":

TABLE IV. Meteorological Factors Related to AQI Values in Each Quarter

Season	Factor
Winter	X1、X2、X3、X4、X5、X6、X7、X8
Spring	X4、X5、X6、X7、X8
Summer	X1、X2、X4、X6、X7、X8、X9
Autumn	X4、X5、X6、X7、X8

According to the basis of judgment, the relationship between X1, X2, X3, X4, X5, X6, X7, X8 and AQI is more significant in winter; and X4, X5, X6, X7 and X8 are better correlated with air quality index AQI in spring. In summer, X1, X2, X4, X6, X7, X8, and X9 have good correlation with air quality index AQI, which is significant for AQI index; X4, X5, X6, X7, and X8 are related to air quality index AQI in spring.

First, the analysis is based on the corresponding candidate factors in four quarters to determine whether it is suitable for factor analysis. After analysis, the KMO values for the four quarters were found to be: 0.700, 0.550, 0.414, 0.624. According to the KMO metric given by Kaiser, it is known that the original factor in winter is suitable for factor analysis, while other quarters are not suitable for factor analysis.

TABLE V. FACTOR TOTAL VARIANCE

Factor	Sum	Feature Value	
		Percentage of variance	Accumulated (%)
1	2.971	37.14	37.14
2	1.802	22.525	59.67
3	1.278	15.969	75.63
4	0.799	9.989	85.62
5	0.72	9.003	94.63
6	0.356	4.444	99.07
7	0.045	0.564	99.63
8	0.029	0.367	100.00

For the winter neural network model, eight factors including maximum temperature, minimum temperature, average temperature, wind, wind level, daytime and night weather conditions, and relative humidity were used as the candidate factors for principal component analysis (PCA). According to the principle that the cumulative contribution rate reaches 90% or more, the data is further processed [12]. The result is shown in "Table V".

TABLE VI. MAIN COMPONENT COEFFICIENT TABLE

	Factor				
	1	2	3	4	5
X1	0.351	0.021	-0.106	-0.041	0.021
X2	0.321	-0.006	0.096	0.016	-0.001
X3	0.352	0.037	-0.111	-0.007	0.014
X4	0.018	0.023	-0.098	-0.089	1.026
X5	-0.021	-0.013	-0.093	1.022	-0.088
X6	0.054	0.559	-0.022	-0.080	0.047
X7	0.007	-0.580	0.165	-0.058	0.007
X8	-0.067	-0.103	1.035	-0.085	-0.090

It can be seen that the cumulative variance contribution of the first five factors has exceeded 90%, reaching 94.625%. Therefore, it can be considered that it is feasible to replace the original eight factors with the first five factors (Z1, Z2, Z3, Z4, Z5). Therefore, the principal component coefficients corresponding to the five factors Z1, Z2, Z3, Z4, and Z5 are shown in "Table VI"

$$Z1 = 0.351 \times X1 + 0.321 \times X2 + 0.352 \times X3 + 0.018 \times X4 - 0.021 \times X5 + 0.054 \times X6 + 0.007 \times X7 - 0.067 \times X8$$

$$\mathrm{Z2} = 0.021 \times X1 - 0.006 \times X2 + 0.037 \times X3 + 0.023 \times X4 - 0.013 \times X5 + 0.559 \times X6 - 0.580 \times X7 - 0.103 \times X8$$

$$\mathrm{Z3} = -0.106 \times X1 + 0.096 \times X2 - 0.111 \times X3 - 0.098 \times X4 - 0.093 \times X5 - 0.022 \times X6 + 0.165 \times X7 + 1.035 \times X8$$

$$\mathrm{Z4} = -0.041 \times X1 + 0.016 \times X2 - 0.007 \times X3 - 0.089 \times X4 + 1.022 \times X5 - 0.080 \times X6 - 0.058 \times X7 - 0.085 \times X8$$

$$\mathrm{Z5} = 0.021 \times X1 - 0.001 \times X2 + 0.014 \times X3 + 1.026 \times X4 - 0.088 \times X5 + 0.047 \times X6 + 0.007 \times X7 - 0.090 \times X8$$

In summary, the factors affecting air quality include: 6 pollutants participating in the AQI index evaluation, namely PM10, PM2.5, SO_2, NO_2, CO, O_3 and related meteorological factors.

IV. NEURAL NETWORK BASED AIR QUALITY PREDICTION MODEL

The theory has proved that a BP neural network with only one hidden node can achieve arbitrary precision approximation for any nonlinear function. Therefore, from the perspective of simple and convenient training, only one hidden layer is generally selected. Moreover, as the number of hidden layers in the network increases, the reverse transmission efficiency of the error becomes very low, which is disadvantage for the learn of the neural network. Therefore, the neural network with only one hidden layer is used in the research.

According to the definition of AQI, the six factors of PM10, PM2.5, SO_2, NO_2, CO and O_3 are the influencing factors of air quality. In the third section, SPSS is used to analyze the correlation between API value and various meteorological factors. Therefore, the input neurons of the winter model include: PM10, PM2.5, SO_2, NO_2, CO, O_3, X1, X2, X3, X4, X5, X6, X7, X8; the spring model and the autumn model contain PM10, PM2. 5, SO_2, NO_2, CO, O_3, X4, X5, X6, X7, X8; summer model contains PM10, PM2.5, SO_2, NO_2, CO, O_3, X1, X2, X4, X6, X7, X8 , X9. The number of neurons in the model output layer is 1 (AQI value).

In general, the number of neurons in the hidden layer is determined by the following formula:

$$\mathrm{n} = \sqrt{\mathrm{n_i} + \mathrm{n_0}} + a \qquad (2)$$

Where:

n — the number of nodes in the hidden layer

n_i — the number of nodes in the input layer

n_0 — the number of nodes in the output layer

a — constant, value between 1 and 10

Therefore, the number of neurons in the hidden layer is between 5 and 14. In this paper, the mean square error (MSE) is the selection criterion of the number of neurons in the hidden layer: if the prediction performance of the network is unstable, the mean square error will be correspondingly larger; on the contrary, the prediction performance is stable. The formula for calculating the mean square error (MSE):

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \qquad (3)$$

Where $\hat{y}_i$ represents the training result value and y_i represents the actual value.

The relationship between the implicit neuron and the mean square error is shown in "Fig. 2". The abscissa indicates the number of neurons in the hidden layer, and the ordinate indicates the corresponding mean squared difference.

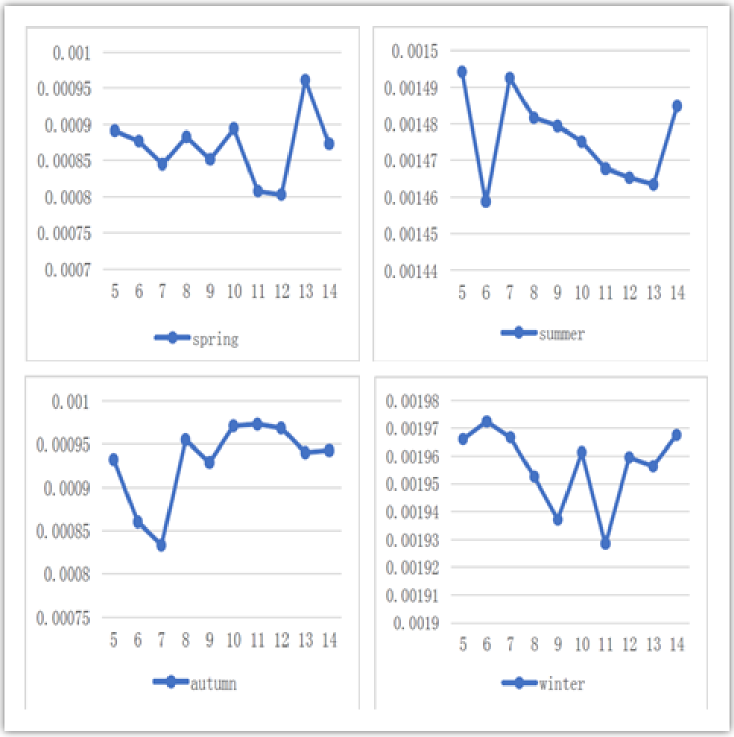

Fig. 2. Hidden layer neurons and mean square error plots for each quarter.

It can be found that when the number of neurons in the hidden layer is 11 in the winter neural network system, the mean square error is the smallest at this time, that is, the prediction performance of the model is relatively stable; it is implicit in the spring neural network system. When the number of layer neurons is 12, the mean square error is the smallest. In the summer neural network system, when the number of neurons in the hidden layer is 6, the mean square error is the smallest; in the autumn neural network system, it is implied. When the number of layer neurons is seven, the mean square error is the smallest. Therefore, for different quarters, I establish different neural network models: the neural network model in winter is 11-11-1, the neural network model in spring is 11-12-1, and the neural network model in summer is 13-6-1, autumn. The neural network model is 11-7-1.

In addition, the implicit layer transfer function of this study is the “tansig-function”; considering the output as the air quality AQI index, the output transfer function selects the “purelin-function” [13].

V. Experimental Results and Analysis

After filtering out the default value data, this paper forecasts the air quality index of the corresponding quarter in 2018 based on the winter, spring, summer and autumn air conditions of Beijing from 2014 to 2017. The correlation coefficient of training results and training is shown in "Fig. 3", "Fig. 4", "Fig. 5" and "Fig. 6".

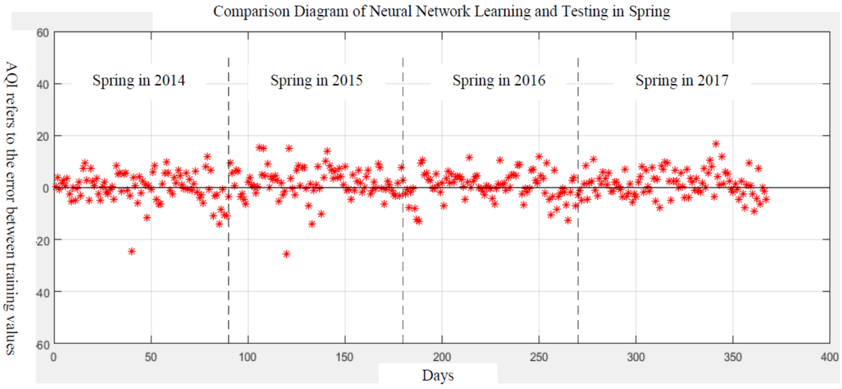

Fig. 3. Comparison of training results and actual data for each quarter: Spring.

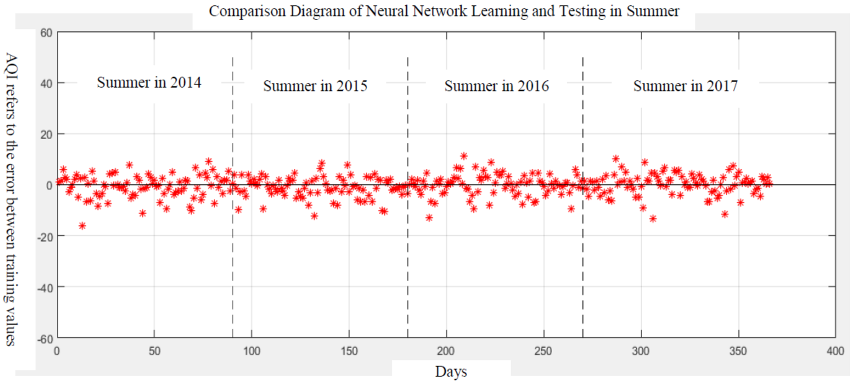

Fig. 4. Comparison of training results and actual data for each quarter: Summer.

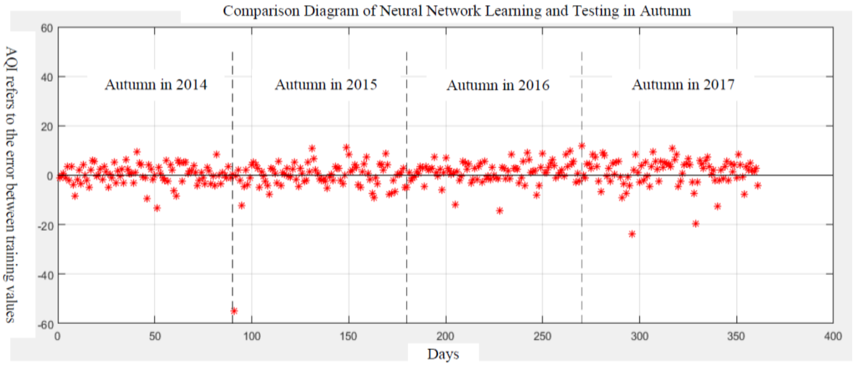

Fig. 5. Comparison of training results and actual data for each quarter: Autumn.

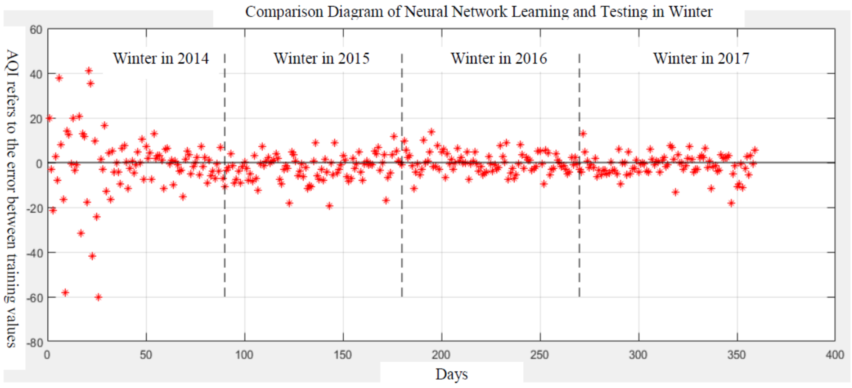

Fig. 6. Comparison of training results and actual data for each quarter: Winter.

It can be found that the trained AQI value fits well with the actual air quality index. The training correlation is more than 0.995, and the training effect is better.

The predicted air quality index is compared to the actual AQI index. Among them, the error can be expressed as:

$$\text{Error}=\frac{|\text{AQI actual}-\text{AQI forecasts}|}{\text{AQI actual}} \quad (4)$$

Among them, the air quality index AQI value and actual value predicted by the neural network model in each quarter are shown in "Fig. 7", "Fig. 8", "Fig. 9" and "Fig. 10".

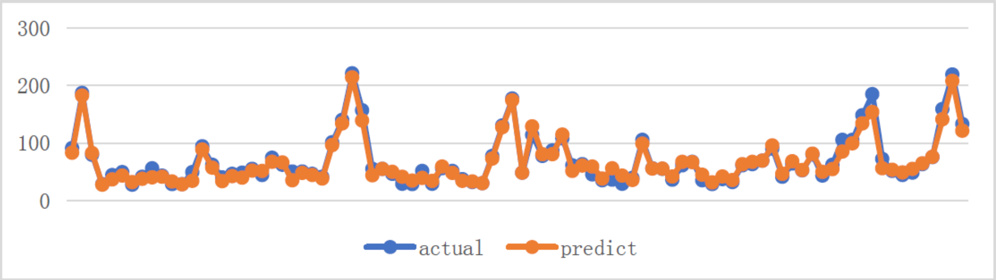

Fig. 7. Comparison of actual and predicted AQI: Spring.

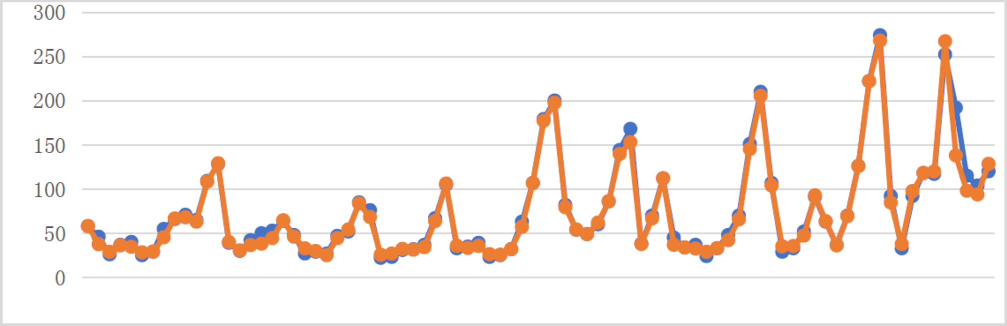

Fig. 8. Comparison of actual and predicted AQI: Summer.

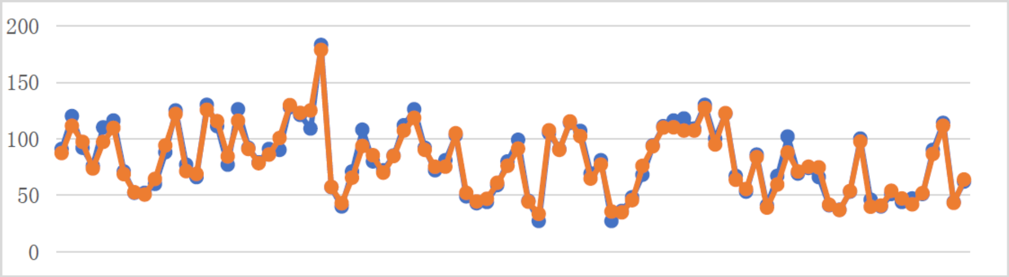

Fig. 9. Comparison of actual and predicted AQI: Autumn.

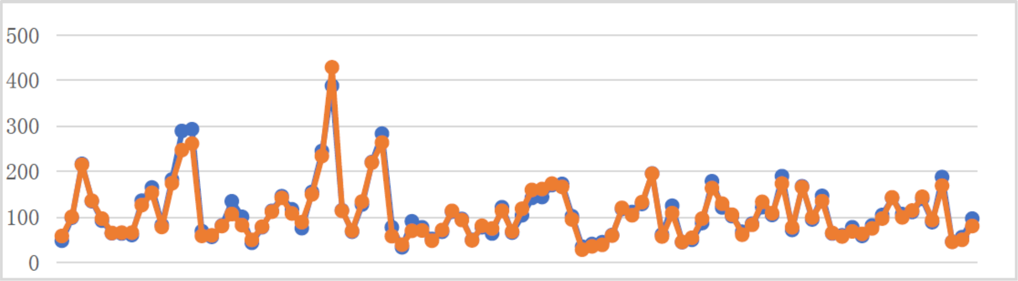

Fig. 10. Comparison of actual and predicted AQI: Winter.

The predicted air quality index is compared to the actual AQI index. Among them, the average error rate of winter prediction is 11.73%, the average error rate of spring prediction is 7.72%, the average error rate of summer prediction is 5.96%, and the average error rate of autumn prediction is 8.99%.

It can be seen from the above analysis that the air quality index obtained from the spring, summer and autumn predictions is in good agreement with the actual situation, while the winter prediction results are relatively poor.

VI. Conclusion

In this paper, a neural network model based on Bayesian normalization algorithm is established after determining the influence factor of air quality. The winter, spring, summer and autumn air conditions of Beijing from 2014 to 2017 are training samples for the corresponding quarter of 2018. The air quality index was predicted. It can be found that the forecasting accuracy of each quarter is 88.27%, 92.28%, 94.04%, and 91.01%, respectively. The prediction accuracy of most studies is 70%~90%, and the model prediction accuracy is high, which provides a model with higher prediction accuracy. It shows that BP network can be applied to Beijing's air quality forecast after effective training and has high prediction accuracy.

References

[1] Chen Zhe, Liu Jiankun, Li Bingjie. Air quality evaluation and prediction based on BP neural network [J]. Modern economic information, 2014 (24): 387-388.

[2] He Xiaoyun, Luo Zerong, Li Mingyue, et al. Modeling and analysis of air quality based on BP neural network [J]. Shandong Industrial Technology, 2018 (17): 239-240.

[3] Zhang Huiyi. Application of improved BP neural network in air pollution prediction in Guangzhou [J]. Energy and Environment, 2017 (01): 11-13.

[4] Zhang Mengyao, Huang Hengjun. Lanzhou air quality prediction based on improved weighted Markov chain [J]. Journal of Lanzhou University of Finance and Economics, 2018, 34 (03): 111-117.

[5] Niu Yuxia. Research on air quality prediction model based on genetic algorithm and BP neural network [J].Software, 2017,38(12): 49-53.

[6] LIN Y, ZHAO L, LI H, et al. Air quality forecasting based on cloud model granulation[J]. EURASIP Journal on Wireless Communications and Networking, 2018,2018(1).

[7] KARATZAS K, KATSIFARAKIS N, ORLOWSKI C, et al. Revisiting urban air quality forecasting: a regression approach[J]. Vietnam Journal of Computer Science, 2018,5(2).

[8] STRUZEWSKA J, KAMINSKI J W, JEFIMOW M. Application of model output statistics to the GEM-AQ high resolution air quality forecast[J]. Atmospheric Research, 2016,181:186-199.

[9] CHENG H, SANDU A. Uncertainty apportionment for air quality forecast models, 2009[C]. ACM, 2009.

[10] Zhang Tian, Mao Yanyan, Wang Jingyang, et al. An air quality prediction method based on BP neural network [J]. Information Communication, 2017 (11): 72-74.

[11] Xue Shiqiong. Realization of air quality prediction and visualization based on BP neural network [D]. Tianjin University, 2016.

[12] Tian Jingyi, Fan Zexuan, Sun Lihua. Air quality prediction and analysis based on BP neural network [J]. Journal of Liaoning University of Science and Technology, 2015, 38 (02): 131-136.

[13] Li Ying, Wang Aihua, Gu Jianwei, et al. Application of BP neural network in project management and Bayesian regularization optimization [J]. Value Engineering, 2009, 28 (08): 91-93.