Design of Risk Early Warning Model Based on Big Data

Qi Shen*
Heihe University, Heilongjiang Heihe, China
*Corresponding author: shenqi@163.com

Abstract. With the development of China's economy, pollution has become more and more serious, so total pollution control has become a priority for governments at all levels. The center of total pollutant emission reduction is to formulate scientific and reasonable pollutant emission reduction policies based on the forecast of pollutant emission in the next few years. According to the national requirements, the emissions of four pollutants, namely sulfur dioxide, nitrogen oxide, ammonia nitrogen and chemical oxygen demand, need to be counted. The research objective of this paper is to establish a combined model applicable to the emission prediction of sulfur dioxide, nitrogen oxide, ammonia nitrogen and chemical oxygen demand by analyzing the impact of population factors and industrial distribution on pollutant emissions and combining the grey system theory and neural network theory. According to the characteristics of pollutant discharge in a certain city, a pollutant discharge prediction model based on GM (1,1), GM (0,4) and BP neural network is proposed.

Keywords: The grey system theory, Neural network theory, GM (1,1), BP neural network.

1. Introduction
Predict the current domestic pollutants are mainly concentrated on the sulfur dioxide and chemical oxygen demand (cod), and less about prediction of nitrogen oxides and ammonia nitrogen, and most of the prediction algorithm for only one of the pollutants, according to the latest requirements of the state, provinces and cities to the four kinds of pollutant emission reduction plan, and accurately predict the emissions of pollutants, four is the precondition of emission reduction plan, existing models are not able to meet the needs for four kinds of prediction of discharge of pollutants. There are a large number of untrue data in the data obtained from each administrative district and county, which have a large deviation from the actual value and will seriously affect the prediction results. Therefore, data preprocessing should be carried out before the prediction to eliminate the messy data and data irrelevant to the prediction. Through to the domestic and foreign theories about total amount control and pollutant emission prediction research, summarizes the current in pollutant emission prediction model are low and need to be improved, in combination with the actual situation of a city, to provide administrative districts of the city environmental protection bureau of the pollutants discharged history data as the foundation, factors such as population and distribution of industrial structure in each district and county as a reference factor, further study of pollutant emission regularity and key factors, analyzing the main factors affecting the pollutants discharge, According to these analysis results, the grey system theory and BP neural network were used to build a grey neural network model to predict...
the emissions of sulfur dioxide, nitrogen oxide, ammonia nitrogen and chemical oxygen demand in the city, which provided decision support for the municipal Environmental protection bureau to make the total emission reduction plan[1,2].

In this paper, a series of optimizations are made to improve the accuracy of the prediction model, so that the average error rate of the prediction can be controlled within 5%.

2. Grey Prediction Method
In the gray system, although the data is not complete, the gray system still considers the data to be holistic. The model established by the grey theory is not the original data sequence model, but the data model generated through a series of data processing. Therefore, the prediction method based on the grey theory system idea is called the grey prediction method. Grey forecasting is also known as GM model forecasting.

The steps for establishing the GM(1, 1) prediction model are as follows. Let the given original data sequence be:

\[ X^{(0)}(0), X^{(0)}(1), \ldots, X^{(0)}(n) \]

Due to the weak regularity of these original data, the data fluctuates greatly, which will adversely affect the prediction results. Therefore, in gray prediction, cumulative generation is usually used to process the data, as shown below:

\[ X^{(1)}(0), X^{(1)}(1), \ldots, X^{(1)}(n) \]

\[ X^{(1)}(k) = \sum_{i=1}^{k} X^{(0)}(i), k = 1, 2, \ldots, n \]

Where, \( n \) is the number of raw data.

Through accumulation generation, the original data series with no or weak regularity can be transformed into a monotonically growing curve. Through this transformation, the regularity of data series can be strengthened and its fluctuation can be reduced[3].

3. BP Neural Network
BP neural network modeling generally includes collecting sample data, determining the topology of the neural network, training the neural network, initial connection weight of the neural network, ensuring the performance and generalization ability of the model, and finally determining a reasonable network model.

The establishment of BP neural network model is mainly divided into several steps, such as sample training, topological structure determination, neural network training, improvement of model generalization ability, determination of initial value and selection of the final model. Among them, a large enough number of samples with high precision and good typicality are the prerequisites for BP neural network modeling. But at the same time, the large sample space is bound to contain some abnormal sample values. Therefore, data preprocessing must be carried out on samples before training them. Results preprocessing samples can avoid abnormal training results caused by abnormal samples. Determine the topological structure of mainly includes the determination of number of hidden layer and the number of hidden layer nodes to determine two aspects, features in the hidden layer to extract data, to a large extent determine the reasonable number of hidden layer associated with BP neural network output is accurate or not, as is often the case, we can use a hidden layer, if the result does not meet the requirements, and then gradually increase the number of hidden layer, but the number of hidden layer can negatively impact the result of the network; The number of hidden layer nodes is very important in BP neural network. The principle of choosing the appropriate number of hidden layer nodes is less than not more. The different initial weights of BP neural network determine which local or global minimum the algorithm converges to, and the initial weights of BP neural network are generally required to be distributed between -0.5 and 0.5. The error and generalization ability of different network structures may also be different, so it is necessary to analyze the advantages and disadvantages of different network structure models[4].
4. Verification of Pollutant Emission Prediction Model

4.1. Data Preparation

Table 1 shows the demographic factors and industry data required for GM (0,4) model, which are obtained from the official websites of statistics bureaus of a city and all districts and counties.

Table 1 Distribution of population factors and industrial output value in a certain district and county

| Year | Population(ten thousand) | Per capita consumption (Yuan/person) | The first industry(one hundred million yuan) | The second industry(one hundred million yuan) | The third industry(one hundred million yuan) |
|------|--------------------------|-------------------------------------|--------------------------------------------|--------------------------------------------|---------------------------------------------|
| 2002 | 121.3                    | 8051.46                             | 16.92                                      | 12.56                                      | 13.86                                       |
| 2003 | 122.9                    | 9803.36                             | 17.18                                      | 15.93                                      | 20.83                                       |
| 2004 | 123.5                    | 11138.35                            | 22.2                                       | 18.05                                      | 24.10                                       |
| 2005 | 125.1                    | 12472.98                            | 27.65                                      | 25.20                                      | 28.16                                       |
| 2006 | 126.3                    | 13983.13                            | 31.3                                       | 35.9                                       | 32.8                                        |
| 2007 | 126.3                    | 15197.15                            | 30.1                                       | 40.0                                       | 41.1                                        |
| 2008 | 126.7                    | 16395.18                            | 31.1                                       | 50.6                                       | 49.3                                        |
| 2009 | 127.5                    | 17093.96                            | 35.0                                       | 69.5                                       | 69.7                                        |
| 2010 | 127.9                    | 18050.35                            | 36.2                                       | 85.26                                      | 84.04                                       |
| 2011 | 128.1                    | 19132.33                            | 37.1                                       | 89.51                                      | 93.17                                       |
| 2012 | 128.1                    | 20165.14                            | 38.9                                       | 95.36                                      | 99.28                                       |

4.2. Data Preprocessing

Due to non-standard statistics and other reasons, the historical data of pollutant discharge in some districts and counties are abnormal, which will have a negative impact on the predicted results. Therefore, when the model is used for simulation, the original data must be preprocessed.

Let the original emission data series of sulfur dioxide be

\[ X^{(0)} = (1257.08, 1638.97, 1538.1, 1689.40, 1580.23, 1667.93, 1768.05, 1896.55) \]

The data sequence was preprocessed by using relevant formulas, and the preprocessing results were shown in table 2.

Table 2 Pollutant history data after pretreatment

| Year | Sulfur dioxide(t) | Ammonia nitrogen (t) | Nitrogen oxides(t) | Chemical oxygen demand(t) |
|------|------------------|----------------------|-------------------|--------------------------|
| 2002 | 1032.03          | 774.46               | 735.58            | 2073.29                  |
| 2003 | 1286.54          | 1091.97              | 823.47            | 1846.21                  |
| 2004 | 1521.55          | 1092.51              | 927.83            | 2088.51                  |
| 2005 | 1612.78          | 999.56               | 985.69            | 2333.54                  |
| 2006 | 1597.03          | 1040.10              | 1066.32           | 2461.07                  |
| 2007 | 1589.65          | 1243.65              | 1167.19           | 2552.65                  |
| 2008 | 1615.38          | 1473.45              | 1223.45           | 2645.84                  |

4.3. The Establishment of Pollutant Emission Prediction Model

The combined prediction model of GM(1, 1) and GM(0,4) requires the data of pollutant emissions in a city from 2002 to 2012, as well as population factors and industrial distribution in corresponding years, namely, table 1 and table 2.
Where, the input parameter of GM (1,1) in the combined model is the historical data of the discharge of a single pollutant generated from Table 2. The historical discharge of this pollutant is selected for different pollutants, and the output parameter is the predicted value 1. The input of the GM (0,4) model is the historical data of the discharge of a single pollutant generated from table 2 and the historical data of the three related factors with the highest degree of correlation, that is, a total of four inputs, and the output parameter is the predicted value 2. The weight of predicted value 1 and predicted value 2 was calculated to obtain the preliminary predicted value and residual sequence. The residual sequence is used as the input of BP neural network to obtain the residual correction value and finally obtain the required predicted value.

Take the prediction of sulfur dioxide as an example, the input parameter of GM (1,1) is the historical emission of sulfur dioxide, the output parameter is the predicted value of sulfur dioxide 1, the input parameter of GM (0,4) is the historical emission of sulfur dioxide and the three related factors with the highest degree of correlation with sulfur dioxide, and the output parameter is the predicted value of sulfur dioxide 2. The input parameter of BP neural network is the residual sequence of the preliminary predicted value calculated by the predicted value 1 and the predicted value 2, and the output is the modified residual value.

According to the above description, the prediction of pollutant discharge prediction model can be roughly divided into three steps: first, the combined gray prediction model of the modified GM (1,1) model and GM (0,4) model is used to make a preliminary prediction of pollutant discharge, calculate the weight coefficient, and obtain the preliminary prediction value and residual sequence; Then, the residual sequence is input into BP neural network for residual correction, and the residual correction value is obtained. Finally, combined with the preliminary predicted value and residual correction value, the final predicted value is obtained.

4.4. Residual Error Correction
The correction residual sequence is shown in table 3.

| Year | Sulfur dioxide(t) | Ammonia nitrogen(t) | Nitrogen oxides(t) | Chemical oxygen demand(t) |
|------|------------------|---------------------|--------------------|---------------------------|
| 2009 | 122.8            | 8.11                | 65.01              | 64.11                     |
| 2010 | 55.21            | 2.08                | 77.36              | -33.17                    |
| 2011 | 77.36            | -6.36               | 137.88             | -34.07                    |
| 2012 | -43.55           | -9.14               | -18                | 13.9                      |

4.5. The Final Predicted Value Is Calculated
Suppose $\bar{x}^{(0)}(i)$ is the original predicted value, $\bar{e}^{(0)}(l)$ is the modified residual value, and $\bar{x}^{(0)}(i,l)$ is the final predicted value, then

$$\bar{x}^{(0)}(i,l) = \bar{x}^{(0)}(i) + \bar{e}^{(0)}(l)$$

The predicted value of pollutant discharge prediction model is obtained. As shown in tables 4 through 7.

| Year  | The actual value | Simulation value (t) | Residual | Error rate |
|-------|------------------|----------------------|----------|------------|
| 2009  | 1768.05          | 1814.14              | 46.09    | 2.61%      |
| 2010  | 1896.55          | 1844.79              | -51.76   | 2.73%      |
| 2011  | 1931.43          | 1890.92              | -40.51   | 2.10%      |
| 2012  | 1899.37          | 1941.37              | 42       | 2.21%      |
Table 5 Ammonia nitrogen simulation results

| Year | The actual value | Simulation value (t) | Residual | Error rate |
|------|------------------|----------------------|----------|------------|
| 2009 | 1673.21          | 1597.35              | -75.86   | 4.53%      |
| 2010 | 1637.98          | 1599.79              | -38.19   | 2.33%      |
| 2011 | 1598.24          | 1631.14              | 32.9     | 2.06%      |
| 2012 | 1585.47          | 1642.15              | 56.68    | 3.57%      |

Table 6 Chemical oxygen demand simulation results

| Year | The actual value | Simulation value (t) | Residual | Error rate |
|------|------------------|----------------------|----------|------------|
| 2009 | 2890.86          | 2839.24              | -51.62   | 1.79%      |
| 2010 | 2965.07          | 2863.56              | -101.51  | 3.42%      |
| 2011 | 2899.03          | 2901.97              | 2.94     | 0.10%      |
| 2012 | 2831.89          | 3013.43              | 181.54   | 6.41%      |

Table 7 Nox simulation results

| Year | The actual value | Simulation value (t) | Residual | Error rate |
|------|------------------|----------------------|----------|------------|
| 2009 | 1308.09          | 1328.94              | 20.85    | 1.59%      |
| 2010 | 1498.67          | 1479.11              | -19.56   | 1.31%      |
| 2011 | 1503.98          | 1539.37              | 35.39    | 2.35%      |
| 2012 | 1547.97          | 1581.35              | 33.38    | 2.16%      |

Calculate the average relative error rate, as shown in table 8.

Table 8 The average relative error rate

| Pollutants | The actual value | Simulation value (t) | Residual | Error rate |
|------------|------------------|----------------------|----------|------------|
| Average error rate | 2.4%     | 3.1%                 | 1.81%    | 2.9%       |

Using improved after the result of the combination model to forecast the pollutant emissions and the actual value error range is relatively small, the maximum error is 6%, the minimum error is 0.1%, and a little change, average relative error rate is lower than other two kinds of prediction model, the combination of a variety of prediction methods and the optimized model is compared with other models, the most close to the real value.

Acknowledgements
Department of Education under Grant No. 2020-KYYWF-0881, Research on Risk Evaluation and Countermeasures of Brain Drain in Universities in the Border Areas of Heilongjiang Province.

References
[1] Smith L E D, Porter K S, "Management of catchments for the protection of water resources: drawing on the New York City watershed experience", Regional Environmental Change, vol.10, no. 4, pp. 311-326, 2010.
[2] Sandoval-Solis S, McKinney D C, Loucks D P. “Sustainability Index for Water Resources Planning and Management”. Journal of Water Resources Planning and Management, vol.137, no. 5, pp. 381-390, 2010.
[3] Commoner B. “The closing circle: nature, man, and technology”. New York, vol.141, no.1, pp.265-268, 1971.
[4] Grossman G M, Krueger A B. “Economic growth and the environment”. The quarterly journal of economics, vol.110, no.2, pp.353-377, 1995.