

ANALYSIS TECHNOLOGY OF ENVIRONMENTAL MONITORING DATA BASED ON INTERNET OF THINGS ENVIRONMENT AND IMPROVED NEURAL NETWORK ALGORITHM

ZHAI, W.

*Xi'an Aeronautical University, Xi'an 710077, China
(e-mail: 2587842805@qq.com)*

(Received 7th Jun 2019; accepted 10th Oct 2019)

Abstract. With the development of industrialization, the problem of environmental pollution has become increasingly serious. Environmental monitoring data, as a measurement index of environmental quality, is increasingly valued by governments and citizens. However, the environmental data accumulated through real-time monitoring at the current stage is mostly used to write basic reports, while the hidden laws or values still need to be further explored. This paper proposes two original environmental prediction models and improves upon these two methods separately to predict the quality of the atmospheric environment. The following research results are obtained: the multivariate linear equation is optimized through stepwise linear regression, which can accurately predict the short-term atmospheric environmental quality; the improved BP neural network can predict the mid-term and long-term atmospheric environmental quality through short-term training.

Keywords: *environmental monitoring, neural network, environmental quality, data analysis, prediction*

Introduction

With the rapid advancement of industrialization and the rapid increase of urban population, the environmental pollution problem is becoming more and more serious (Beck et al., 1961; Jeffrey et al., 2000), including atmospheric pollution, such as acid rain and fog; water environment pollution, such as black odorous water, water bloom and red tide; solid waste pollution, such as construction waste and tailings pond pollution. In recent years, the atmospheric pollution has been particularly prominent. It directly leads to a decline in the quality of people's living environment, affecting human health and causing huge economic losses (Lauth et al., 2010; Abramovitch et al., 2015). In order to avoid the occurrence of air pollution, large funds have been invested in the prevention of air pollution at various levels in our country (Bey et al., 2017) and an atmospheric environment monitoring Internet of Things has been established. Through the monitoring of the quality of the atmospheric environment, it can be controlled in real time, which guides the industrial production and the construction of related pollution prevention facilities (Rodgers et al., 2015; Lochner et al., 2011; Zhang et al., 2018a). At present, almost all atmospheric environmental monitoring data is used to prepare environmental reports such as daily newspapers and annual reports (Arsie et al., 2010; Marino et al., 2017) while the value of data needs to be further explored (Li et al., 2014). For example, through historical monitoring data, the future trend of atmospheric environmental quality can be predicted so as to better guide people's production activities (Chien et al., 2005, 2010). Meanwhile, it provides relevant scientific evidence for government decision-making and management departments in the formulation of systems concerning this area (Chien et al., 2003).

At present, the methods for predicting the quality of atmospheric environment mainly include numerical prediction and statistical prediction. However, the model precision of numerical prediction is lower than that of statistical prediction and the scope of application is limited (Chien et al., 2007; Song, 2018; Cooper and Ekström, 2005; Reifman et al., 2000; Zhang et al., 2018b; Zhao et al., 2018). Therefore, this paper adopts the statistical prediction with easy data acquisition and high prediction precision, including the back propagation (BP) neural network and multiple linear regression equation for the analysis of atmospheric environment prediction.

Materials and methods

Multiple regression model

When there is a linear relationship between two or more independent variables and dependent variables, it is a multiple linear regression. Its mathematical model is shown in *Equation 1*:

$$y = a_1 + a_2x_1 + a_3x_2 + \dots + a_{m+1}x_m + \varepsilon \quad (\text{Eq.1})$$

In this formula, $a_1, a_2, a_3, \dots, a_{m+1}$, are regression coefficients; ε is a random error.

The estimated value of regression coefficient is shown in *Equation 2*:

$$Q = \sum_{t=1}^n [y_t - (a_1 + a_2x_{1t} + a_3x_{2t} + \dots + a_{m+1}x_{mt})]^2 \quad (\text{Eq.2})$$

After obtaining the multiple linear regression equation, it needs to be tested to determine its precision. The commonly used test methods include correlation coefficient test, F test and t test.

(1) Correlation coefficient test

The correlation coefficient is an index used to measure the fitting degree of the linear models. The mathematical expression is the ratio of the regression sum of squares to the total sum of squares, as is shown in *Equation 3*:

$$R^2 = \frac{SSR / p}{SST / n - p - 1} \quad (\text{Eq.3})$$

(2) F test

The F test is used to test whether the relationship between the independent variable and the dependent variable is significant in the linear model and it is expressed by *Equation 4*:

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p - 1)} = \frac{SSR}{SSE / (n - 2)} = \frac{R^2}{1 - R^2} (N - 2) \sim F(1, n - 2) \quad (\text{Eq.4})$$

(3) *T test*

The t test can be used to determine whether a variable is retained as an independent variable in the model and it is expressed by *Equation 5*:

$$t = \frac{\hat{\beta}_1}{\frac{\sigma}{\sqrt{\sum_{j=1}^n (x_{ij} - \bar{x}_{1j})^2}}} \sim t(n - p - 1) \quad (\text{Eq.5})$$

In this equation:

$$\hat{\beta} \sim N \left[\frac{\sigma^2}{\sum_{j=1}^n (x_{ji} - \bar{x})^2}, \hat{\sigma}^2 = \frac{SSE}{(n - p - 1)} \right]$$

In *Equations 3–5*, SST represents the deviation sum of squares of the difference between the observed value and the mean value of the dependent variable. SSR is the deviation caused by the independent variable, which is the regression sum of squares. SSE is the residual sum of squares caused by the experimental error. The relationship is shown in *Equation 6*:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SSR + SSE \quad (\text{Eq.6})$$

BP neural network

BP neural network is a neural network of error inverse propagation, as is shown in *Figure 1*: after the data is input, it is forwardly propagated from the input layer to the output layer via the hidden layer. According to the set error, the output layer corrects the weight through the hidden layer, which is the error inverse propagation, thereby achieving a stepwise improvement of the precision of the output value of the neural network.

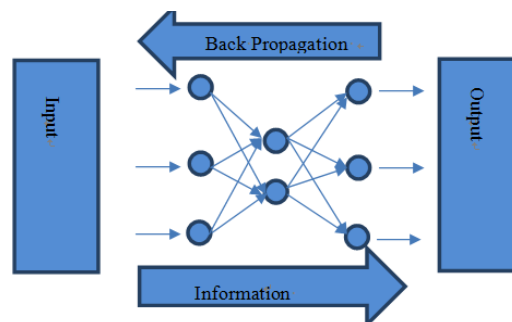


Figure 1. BP neural network topology

Genetic algorithm

The genetic algorithm is developed from the Darwin's theory of evolution and Mendel's genetics. It uses the imitation of biological gene coding to encode individuals, serving as the initial population. The selection, crossover and mutation are completed according to the principle of survival of the fittest. The new population is formed and the above operation is repeated, thereby realizing the retention of excellent genes and inheriting these genes to the offspring. Therefore, the population can better adapt to the environment and continue to breed and evolve.

Data collection and processing

We adopt the GB3095-2012 standard to evaluate the atmospheric environmental quality in the suburbs of Shijiazhuang. The air quality automatic monitor was used to measure the PM_{2.5}, PM₁₀, CO, SO₂, O₃ and NO₂ in the air; temperature-humidity sensor, anemometer and barometer were used to measure meteorological conditions such as temperature, humidity, wind speed, wind direction and air pressure, etc. For example, when the humidity is high, the degree of air pollution will increase. Therefore, this paper will collect 6 indexes of atmospheric pollutants and corresponding 5 meteorological indexes.

Data collection

This paper downloads the monitoring data of atmospheric environment of the environmental protection bureau in the Shi Jiazhuang by writing the crawler software. The data collection process is shown in *Figure 2*.

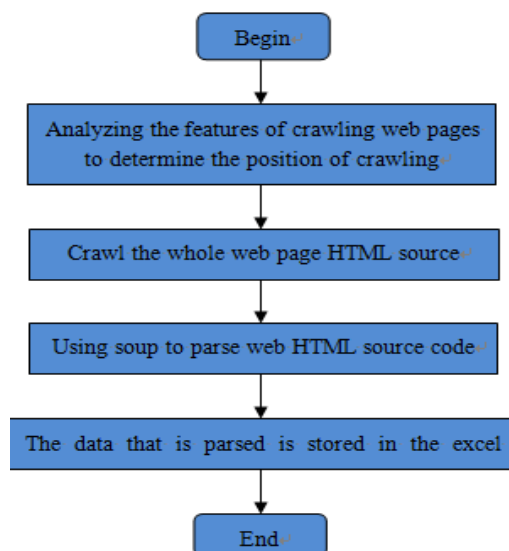


Figure 2. Flow chart of web crawler

Data processing

Data preprocessing plays an important role in the in-depth, accurate mining and analysis of data. In this paper, the collected data is preprocessed in the following two ways.

(1) Eliminating abnormal data

According to the reasonable distribution range and mutual relationship of each atmospheric data, it is checked whether there is abnormal or contradictory value in the data and the abnormal data is eliminated. In addition, the collected data may be invalid or missing and the mean value of variables is used for the estimation and supplement.

(2) Data normalization

In order to reduce the deduction in the precision of the prediction model caused by the difference in magnitude and dimension between the atmospheric monitoring data, the normalization is conducted on the data. The following two normalization methods are adopted for the multiple linear regression model and the neural network model respectively:

1) Normalized to the interval of [0, 1]

Set x_{max} and x_{min} represent the maximum and minimum value of the original data respectively; x_i is the actual data; and \hat{x}_i represents the normalized value. The normalization equation is:

$$\hat{x}_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (\text{Eq.7})$$

The output value is then converted using the formula $x_i = (x_{max} - x_{min})\hat{x}_i + x_{min}$.

2) Normalized to zero mean and unit variance

After all the raw data is calculated as the mean value of each dimension, the mean is subtracted from each dimension and finally each dimension of the data is divided by the standard deviation of the dimension. The equation is:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma} \quad (\text{Eq.8})$$

Then, the output value is then converted using the formula $x_i = \hat{x}_i\sigma + \mu$.

Results and discussion

The crawler software is used to obtain atmospheric pollutants such as PM2.5, PM10, CO, SO₂, O₃, NO₂, as well as the meteorological data such as temperature, humidity, wind speed, wind direction and pressure. There are 22,000 pieces of PM10 concentration data, of which 20,000 pieces of data are used as training data and the rest 2,000 pieces of data are used as test data at the monitoring point.

Traditional multivariate model

The preprocessed data is constructed into a multiple linear regression model, and correlation coefficient test, F test and t test are performed. The predicted result is then compared to the test data.

Traditional multiple linear regression prediction model

The dependent variable is PM10 and the independent variable is five types of meteorological data such as temperature and pressure. The modeling method is all input.

1) The correlation coefficient test results are shown in *Table 1* and R2 represents the fitting effect. The larger the value, the better the fitting effect.

Table 1. Test of correlation coefficient

Model	R	R2	R2 adjusted	Standard deviation rate error
1	0.438	0.509	0.507	26.73%

2) The result of the F test is shown in *Table 2*. As it can be seen from the table, the result of F test is < 0.01. Therefore, five meteorological indexes have a significant impact on the concentration of PM10.

Table 2. Test of significance

Model	Square	Df	Mean square	F	Significance
Regression	237971.932	5	475897.746	80.792	0.00
Residual	3424801.21	579	5883.049		
Statistics	5804671.02	596			

3) The result of the t test is shown in *Table 3*. The non-normalized coefficient is used to list the regression equation. The normalization coefficient is used to reflect the degree of influence of the independent variable on the dependent variable; the partial regression coefficient is used to determine whether the influence of an independent variable on the dependent variable is statistically significant. When it is < 0.05, it indicates significant statistical significance; when it is < 0.01, the statistical significance is extremely significant.

Table 3. T-test

Model	Unstandardized coefficients	Standardized coefficient	T	Significance
constant	398.212		5.461	0.00
Pressure/Pa	-3.796	-0.248	-5.377	0.00
Temperature/°C	-1.296	-0.0497	-0.981	0.032
Moisture/%	2.223	0.379	9.201	0.00
Wind speed/m/s	-55.059	-0.228	-7.068	0.00
Wind direction	-0.207	-0.179	-5.425	0.00

4) The linear regression equation of the prediction model is:

$$y = -3.796 \times 1 - 1.296 \times 2 + 2.223 \times 3 - 55.059 \times 4 - 0.207 \times 5 + 398.212.$$

The comparison between the true value and the predicted data is shown in *Figure 3*.

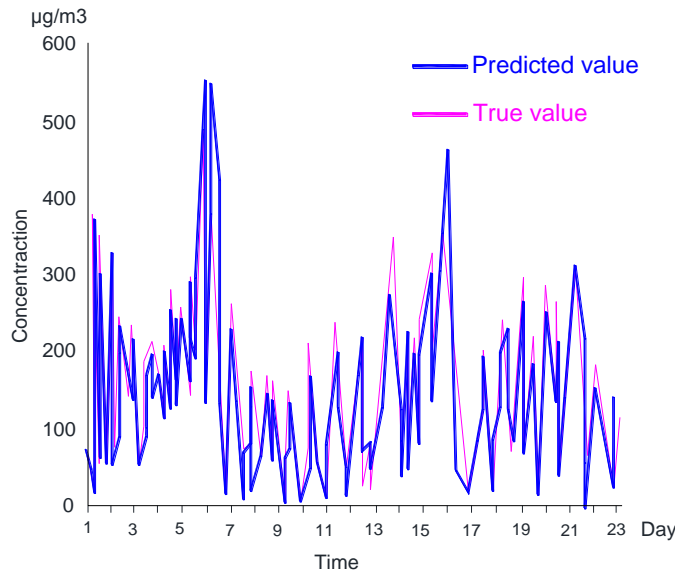


Figure 3. Comparison of predicted value by traditional multiple linear regression model and true value

Improved multiple linear regression prediction model

Considering the physical and chemical reactions between pollutants and the significant impact of seasonal factors on pollutants, other pollutants and seasonal factors are included in the multivariate equation for the optimization. The seasonal variables are: spring $108.1 \mu\text{g}/\text{m}^3$, weight 0.25; summer $97.8 \mu\text{g}/\text{m}^3$, weight 0.2; autumn $112.3 \mu\text{g}/\text{m}^3$, weight 0.25; winter $121.9 \mu\text{g}/\text{m}^3$, weight 0.3.

The significance test is performed by introducing independent variables one by one into the regression model until all significant independent variables are introduced into the regression model. The result of the stepwise regression model is shown in *Table 4*.

Table 4. Model abstract

Model	R	R2	R2 adjusted	Standard deviation rate error	Introduce variable
1	0.781	0.739	0.821	20.89%	Constant, PM2.5, season
2	0.813	0.769	0.825	20.75%	Constant, PM2.5, season, temperature, O ₃
3	0.836	0.808	0.825	20.17%	Constant, PM2.5, season, wind speed, temperature, O ₃
4	0.852	0.817	0.827	19.964%	Constant, PM2.5, season, wind speed, temperature, O ₃ , pressure
5	0.863	0.842	0.829	19.697%	Constant, PM2.5, season, wind speed, temperature, O ₃ , pressure, moisture
6	0.881	0.842	0.828	19.603%	

It can be seen from *Table 4* that PM2.5 has the most significant impact on PM10, while the impact of atmospheric pollutants SO₂, NO₂, CO and the meteorological factor, wind direction on PM10 can be negligible. Therefore, after eliminating these indexes, *Table 5* can be obtained.

Table 5. Test of correlation coefficient

Model	Unstandardized coefficients	Standardized coefficient	T	Significance
Constant	-90.087		-4.875	0.00
PM2.5/ppm	30.304	0.937	89.851	0.00
Temperature/°C	19.729	0.113	8.	0.032
O ₃ /ppm	-0.359	-0.10	-7.61	0.00
Wind speed/m/s	8.541	0.038	4.31	0.00
Pressure/Pa	0.897	0.059	5.081	0.00
Moisture	5.184	0.029	2.857	0.003
Season	10.280	0.269	20.351	0.005

The regression equation can be obtained:

$$y = 30.304 \times 1 + 19.729 \times 2 - 0.359 \times 3 + 8.541 \times 4 + 0.897 \times 5 + 5.184 \times 6 + 10.28 \times 7 - 90.087.$$

Using this model, the true value and the predicted value are shown in *Figure 4*.

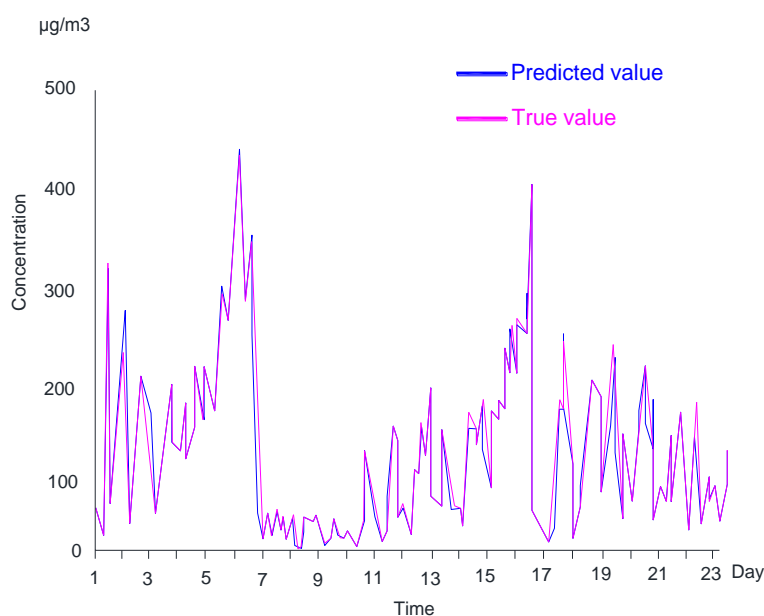


Figure 4. Comparison of predicted value by optimal multiple linear regression model and true value

It can be seen from *Figure 4* that the fitting degree of the optimized model reaches 0.828, which is significantly higher than that of the original model, indicating that the prediction for PM10 by the stepwise linear regression method is more precious after considering meteorological factors and other pollutants.

In addition, it can also be seen from *Figure 4* that the prediction error in the short term (4 days) is the smallest. At the same time, PM2.5, wind speed, air pressure, humidity and season have an enhancing effect on PM10 concentration. The impact of PM2.5 on PM10 is the greatest; while O₃ and temperature has a weakening effect on PM10.

Traditional BP neural network model

The neural network structure needs to be determined by the number of hidden layers, the number of nodes in the input layer, the number of nodes in the hidden layer, the number of nodes in the output layer, the activation function, training method and training parameters. In this paper, a three-layer neural network with a layer of hidden layer is used and it is set to be 11 and 1 according to the principle that the number of input and output nodes is as small as possible. The number of neurons in the hidden layer is determined by formula $M = \sqrt{n + m + a}$. In the formula, a is a constant between 0 and 10, and m and n are the number of neurons in the input layer and the output layer respectively. The number of hidden layers in this paper is 6. The input function in the hidden layer is $f(x) = \frac{1}{1 + e^{-x}}$ and the linear activation function is used on the output layer. The learning rate is 0.01.

The fitting relationship between the predicted value and the true value of the BP neural network is shown in *Figure 5*.

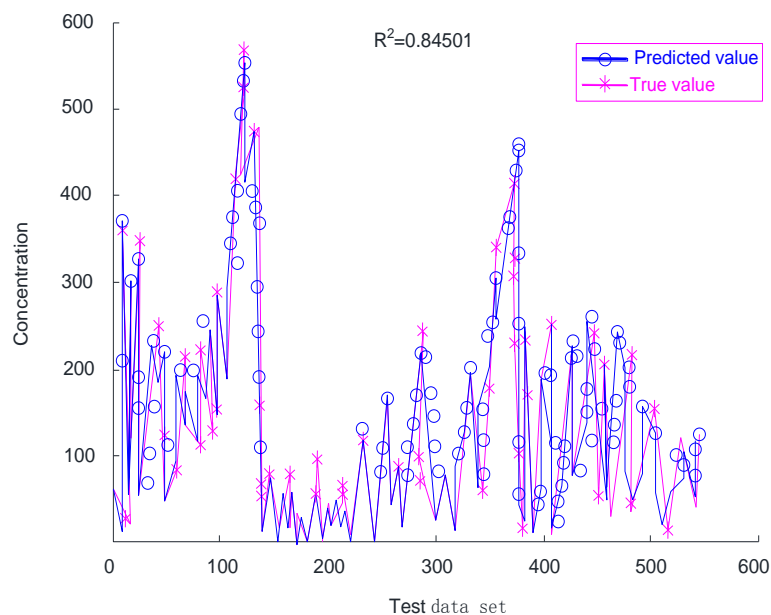


Figure 5. Comparison of predicted value of BP neural network and true value $\mu\text{g}/\text{m}^3$

It can be seen from the *Figure 5* that the traditional BP neural network can better reflect the variation trend of atmospheric pollution in the future and can relatively accurately predict the concentration of atmospheric pollutants. Its goodness of fit is 0.84501.

Improved BP neural network prediction model

The traditional BP neural network uses the gradient descent method to train the neural network, which may lead to problems such as insufficient neural network search ability and slow training speed. This problem can be solved by genetic algorithm. By calculating the fitness value of each individual, it can conduct three genetic operations of selection, crossing and mutation to improve its global search ability and find the individual with the best fitness.

The number of input and output layers in the neural network are 11 and 1, respectively. It is found through the experiment that when the number of nodes in the hidden layer is 11, the prediction effect is the best. There are a total of 120 weights and 11 thresholds in the optimized neural network. The individual coding length is 131; the population size is 22; the evolution number is 57; the crossover probability is 0.22; and mutation probability is 0.1, as is shown in *Table 6*.

Table 6. Learning and training of GA-BP

Network structure	Population size	Evolution times	Initial crossover probability	Initial mutation probability
11-11-1	22	57	0.22	0.1

The fitting of the predicted value in the optimized BP neural network is shown in *Figure 6*. The goodness of fitness in the optimized BP neural network is 0.87925.

The comparison of the effect of these four types of neutral network prediction model is shown in *Table 7*.

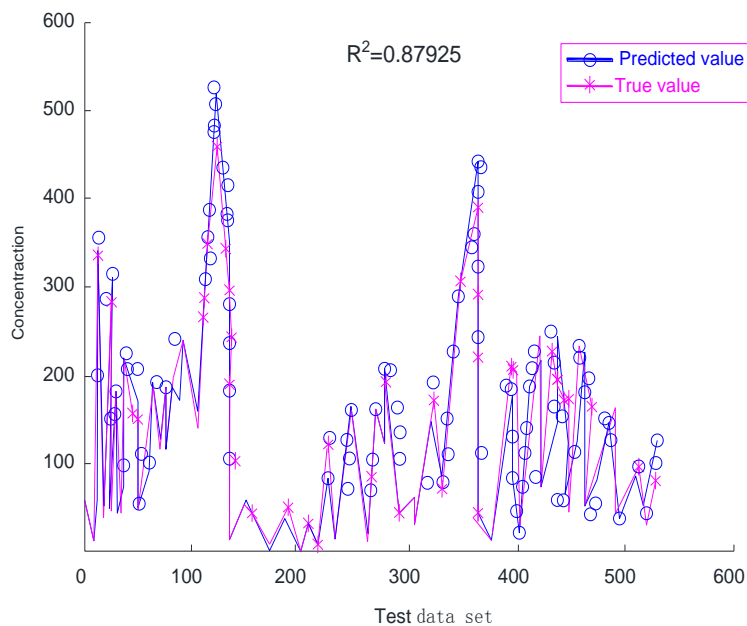


Figure 6. Comparison of predicted value of optimized BP neural network model and true value $\mu\text{g}/\text{m}^3$

Table 7. Comparison of prediction results of four models

	Traditional multiple regression model	Multiple regression model after optimization	Traditional BP neural network model	Optimized BP neural network model
Goodness of fit	0.509	0.819	0.862	0.882
PM10 mean square error	30.27	12.31	8.81	5.38
PM10 mean absolute error	0.163	0.085	0.069	0.047

Conclusions

With the increase of people's requirement for environmental quality, the prediction for the variation trend of environmental quality using the monitoring data becomes increasingly important. In this paper, the original multiple linear regression model, the original BP neural network and the optimized model are used to predict the atmospheric environmental quality. The following research conclusions are drawn:

(1) The traditional multiple linear regression model can only predict the variation trend of atmospheric environmental quality coarsely; while the other three models can predict the concentration of future atmospheric pollutants accurately.

(2) The stepwise linear regression can be used to predict PM₁₀ more accurately after considering meteorological factors and other pollutants. The prediction error for the short term (4 days) is the smallest.

(3) The prediction for the mid-term and long-term atmospheric environmental quality is the best using the optimized BP neural network model.

(4) This paper applied the IoT environment and the improved neural network algorithm to the analysis of environmental monitoring data, which had improved the analysis accuracy and provided a theoretical basis for subsequent data management, atmospheric environment prediction and map management.

REFERENCES

- [1] Abramovitch, A., Pizzagalli, D. A., Geller, D. A., Reuman, L., Wilhelm, S. (2015): Cigarette smoking in obsessive-compulsive disorder and unaffected parents of OCD patients. – *European Psychiatry* 30(1): 137-144.
- [2] Arsie, I., Marra, D., Pianese, C., Sorrentino, M. (2010): Real-time estimation of engine nox emissions via recurrent neural networks. – *IFAC Proceedings Volumes* 43(7): 228-233.
- [3] Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., Erbaugh, J. (1961): An inventory for measuring depression. – *Arch Gen Psychiatry* 4(6): 561-571.
- [4] Bey, K., Lennertz, L., Riesel, A., Klawohn, J., Kaufmann, C., Heinzl, S. (2017): Harm avoidance and childhood adversities in patients with obsessive-compulsive disorder and their unaffected first-degree relatives. – *Acta Psychiatrica Scandinavica* 135(4): 328-338.
- [5] Chien, C. F., Chen, W. C., Lo, F. Y., Lin, Y. C. (2007): A case study to evaluate the productivity changes of the thermal power plants of the Taiwan Power Company. – *IEEE Transactions on Energy Conversion* 22(3): 680-88.
- [6] Chien, T. W., Chu, H., Hsu, W. C., Tseng, T. K., Hsu, C. H., Chen, K. Y. (2003): A feasibility study on the predictive emission monitoring system applied to the Hsinta power plant of Taiwan Power Company. – *Air Repair* 53(8): 1022-28.
- [7] Chien, T. W., Chu, H., Hsu, W. C., Tu, Y. Y., Tsai, H. S., Chen, K. Y. (2005): A performance study of PEMS applied to the Hsinta power station of Taipower. – *Atmospheric Environment* 39(2): 223-30.
- [8] Chien, T. W., Hsueh, H. T., Chu, H., Hsu, W. C., Tu, Y. Y., Tsai, H. S. (2010): A feasibility study of a predictive emissions monitoring system applied to Taipower's Nanpu and Hsinta power plants. – *Air Repair* 60(8): 907-13.
- [9] Cooper, D. A., Ekström, M. (2005): Applicability of the PEMS technique for simplified NOx monitoring on board ships. – *Atmospheric Environment* 39(1): 127-37.
- [10] Jeffrey, B. H., Richard, J. D. (2000): Decreased responsiveness to reward in depression. – *Cognition & Emotion* 14(5): 711-24.
- [11] Lauth, B., Arnkelsson, G. B., Magnússon, P., Skarphéðinsson, G. Á., Ferrari, P., Pétursson, H. (2010): Validity of k-sads-pl (schedule for affective disorders and

- schizophrenia for school-age children - present and lifetime version) depression diagnoses in an adolescent clinical population. – *Nordic Journal of Psychiatry* 64(6): 409-409.
- [12] Li, P. K., Pan, R., Chen, C. (2014): A novel neural network based modeling for control of NO_x emission in power plant. – *Applied Mechanics & Materials* 643(643): 385-90.
- [13] Lochner, C., Serebro, P., Van, D. M. L., Hemmings, S., Kinnear, C., Seedat, S. (2011): Comorbid obsessive-compulsive personality disorder in obsessive-compulsive disorder (OCD): a marker of severity. – *Progress in Neuropsychopharmacology & Biological Psychiatry* 35(4): 1087-92.
- [14] Marino, C., Nucera, A., Nucera, G., Pietrafesa, M. (2017): Economic, energetic and environmental analysis of the waste management system of Reggio Calabria. – *International Journal of Heat and Technology* 35(S1): S108-S116.
- [15] Reifman, J., Feldman, E. E., Wei, T. Y., Glickert, R. W. (2000): An intelligent emissions controller for fuel lean gas reburn in coal-fired power plants. – *Air Repair* 50(2): 240-51.
- [16] Rodgers, S., Ajdacic-Gross, V., Kawohl, W., Müller, M., Rössler, W., Hengartner, M. P. (2015): Comparing two basic subtypes in OCD across three large community samples: a pure compulsive versus a mixed obsessive-compulsive subtype. – *European Archives of Psychiatry & Clinical Neuroscience* 265(8): 719-34.
- [17] Song, S. L. (2018): Application of gray prediction and linear programming model in economic management. – *Mathematical Modelling of Engineering Problems* 5(1): 46-50.
- [18] Zhang, J., Li, Y. B., Liu, B. X., Wu, Y. Q., Yi, H. C. (2018a): Forward modelling of circular loop source and calculation of whole area apparent resistivity based on TEM. – *Traitement du Signal* 35(2): 183-198.
- [19] Zhang, J. X., Sun, W. G., Niu, F. S., Wang, L., Zhao, Y. W., Han, M. M. (2018b): Atmospheric sulfuric acid leaching thermodynamics from metallurgical zinc-bearing dust sludge. – *International Journal of Heat and Technology* 36(1): 229-236.
- [20] Zhao, W., Li, Y. J., Ren, J. Y., Chen, S. G., Li, Y. Q. (2018): A novel operation state prediction method for servers in smart grids. – *European Journal of Electrical Engineering* 20(3): 379-392.