Method of outdoor PM2.5 concentration prediction based on MATLAB

Jiaqi Liu

School of Electrical Information, Jinan University, Zhuhai, Guangdong, 519070, China

*Corresponding author’s e-mail: angela@cas-harbour.org

Abstract. People will have a lot of time exposed to the outside air every day. If the air quality is very poor, such as PM2.5 content is excessive, it will cause great damage to the human body, for example, significantly increasing the incidence of cardiovascular and cerebrovascular diseases, reducing lung capacity and immunity, stimulating the respiratory tract, and increasing the incidence of respiratory diseases. Therefore, it is particularly important to predict the next day’s PM2.5 concentration in advance so that people can take precautions to reduce the inhalation of harmful substances such as PM2.5 and keep the body healthy. This paper includes using a specific machine to measure a certain amount of PM2.5 and concentration data in a fixed place, constructing a neural network model in MATLAB software to finally predict the next concentration, and comparing the actual values. Regardless of the concentration of PM2.5, the neural network described in this paper achieves good results that close to the true value.

1. Introduction
When the staff lived in the office, if the concentration of PM2.5 in the environment was too high, it would cause harm to the body such as increasing the risk of heart disease, reducing lung capacity, increasing the prevalence of lung cancer and so on. Therefore, in order to target the above-mentioned harmful substances, it is necessary to take measures early. The key factor in taking measures is the concentration of harmful substances of the day, hence predicting the concentration as soon as possible is of great significance. Firstly, the paper only needs to collect the relevant data from the previous two days to predict data for the next half day. The data collected is not difficult and with high feasibility. Thirdly, the paper adopts the MATLAB model with higher precision, so the result of the value predicted is more accurate. Fourthly, the paper has strong applicability, which means the system can be applied to any similar environment for prediction.

2. Literature Review
Nowadays, three conventional approaches are used to forecast smog concentrations: chemical transport, machine learning, and statistical models. Statistical models, mainly based on single variable linear regression, have shown a negative correlation between different meteorological parameters (wind and temperature) and PM concentrations. However, the accuracy of statistical models relies on an updated source list that is very difficult to produce. In the sense, the combination of statistical model and computer becomes necessary.

This article provides an idea for the use of computers for smog prediction. The smog-related parameters can be studied by the computer to predict the smog index. However, the article considers...
the influence of physical factors and does not use the processing method of big data that could result in lower prediction accuracy and higher cost.[2]

Professor Hou Putonghua and Professor Yang Hang use three exponential smoothing models to predict concentration. The article shows that it is possible to predict PM2.5 concentration by using computer to dispose of big data.[3]

This paper adopts the computer smog simulation model and it shows the probability of using Brevity (Brevity is a software which can build up a model and use the model to simulate actual environment) to simulate and collect the PM2.5 data.[4]

These articles all utilize big data processing methods to maintain high accuracy. Due to the use of the neural network, the operation is costively and speed is relatively slow. But these articles still give some inspiration to minimize the amount of computation while ensuring accuracy, so a simpler neural network model can be used to process the data.

3. Research Method
This paper adopts the dynamic neural network that comes with MATLAB to build a neural network model and collects the data of PM2.5 changes with time. It processes the time series by using a neural network and finally attains the predicted values. The data in this paper is divided into three parts. The first part is the training set, which is used for model training. The second part is the validation set, which is used to generate forecast data. The third one is the test set, which is used to compare with the predicted value and locate the error of the predicted data. Since the data in the dataset is collected for a month, it can be assumed that there are no major changes in the area, such as wind, temperature, humidity, and other factors. Therefore, the prediction results of the model have some significance. The first step is to collect data. Then, it shows how to pre-process the concentration data. The third step displays how to use MATLAB to establish a neuron model and to analyze the code.

3.1. The first step(collect data)
The concentration data of pm2.5 in the past thirty days were measured in Shijiazhuang City, Hebei Province, China. This table has the number of PM2.5 per hour for thirty days with microgram as the unit.

3.2. The second step(pre-process data)
Since the experiment uses a neural network model and the time axis is not set, the data needs to be processed. In the first place, it needs to ensure that the number of days of data in the training set and the validation set is an integer. So the data from January 21st to February 18th was intercepted as a training set, and the data from February 18th to 19th was used as a validation set. The following experiments enter the data from February 18th to 19th, and the predicted results will be set for the forecast from February 19th to 20th.

3.3. The third step(establish model)
Establish a MATLAB neural network model as follows.
Firstly, import the data and draw the initial data graph. The first 660 data is used as the training set, and the last 23 data is as the verification set. The 660 data includes exactly the whole day’s data, so does the 23 data. It is equivalent to artificially adding the abscissa to the data so that the future value can be predicted. Converting the data into the sequence data needs the NARNET function, and the function of the NUM2CELL is to treat each element in the matrix as an element of the cell. Set the delay to 10, that is, the calculated new value depends on the first 10 values. Set the number of hidden layer nodes to 10. NARNET function and PREPARES function are for data preparation. Train the

```matlab
1 - pressure = z(1:663);
2 - pressure = h(1:24);
3 - figure
4 - plot( pressure )
5 - % number of training and validation data
6 - num_all_data = length( pressure );
7 - % The first 660 data as a training set
8 - num_train = floor( num_all_data * 660/663 );
9 - % remaining as validation set
10 - num_valid = num_all_data - num_train;
11 - % Converted to the sequence data needed by narnet
12 - y_train_num = num2cell( pressure(1:num_train) ' );
13 - y_valid_num = num2cell( pressure(1+num_train:end) ' );
14 - % delay
15 - feedback_delays = 1:10;
16 - % Number of hidden layer nodes
17 - num_hd_neuron = 10;
18 - % Narnet build
19 - net = narnet( feedback_delays, num Hd_neuron );
20 - [Xs,Xi, Ai, Ts] = prepares( net, [], {}, y_train_num);
21 - net = train( net, Xs, Ts, Xi);
22 - view( net );
23 - Y = net( Xs, Xi );
```

Figure1. MATLAB code

```matlab
24 - perf = perform( net, Ts, Y );
25 - fprintf( ' neural network mse on training set : %8.4f \n', perf );
26 - % predict
27 - yini = y_train_num(end-max(feedback_delays)+1:end);
28 - [Xs, Xi, Ai] = prepares( net, [], {}, yini y_valid_num);
29 - y_prd_num = net( Xs, Xi, Ai );
30 - y_prd_num = cell2mat( y_prd_num );
31 - y_valid_num = cell2mat( y_valid_num );
32 - % draw picture, calculate mse
33 - figure
34 - title( ' NARNET PREDICTION ' )
35 - hold on
36 - plot( y_valid_num, 'r', 'linewidth', 2 );
37 - plot( y_prd_num, 'b--', 'linewidth', 2 );
38 - plot( pressure, 'g', 'linewidth', 2 );
39 - legend( ' the true value of the (n-1)th day',
40 - ' the predicted value of the nth day', ' the true value of the nth day' )
41 - mape_error = ( abs( y_prd_num - y_valid_num ) ./ y_valid_num );
42 - mape_error = mean( ( y_prd_num - y_valid_num ) .^ 2 );
43 - figure
44 - title( ' RELATIVE ERROR ' )
45 - hold on
46 - plot( mape_error, 'b' );
```

Figure2. MATLAB code
network with the TRAIN function that comes with MATLAB. The meaning of the CELL2MAT function is to convert the elements in the cell into a matrix. Finally, draw a comparison chart of the predicted results, the real results and the error ones.

3.4. The fourth step(input PM2.5 data and analyze charts)
This picture shows the internal model of the neural network. The input value enters the hidden layer, after a delay of 10, multiplied by the weight W, plus b, the result is put into the activation function (a nonlinear activation function), multiplied by the weight W, plus b, then put the result through a linear function and then receive the output.

![Model flow chart](image)

Figure 3. Model flow chart

This graph shows the change in PM2.5 over time. The abscissa is the data of the training set and the verification set, and the ordinate is the value of PM2.5 for this period. It can be seen that the PM2.5 situation in the region is severe, and more than half of the time this month is above 100 micrograms.

![Total PM2.5 data overview](image)

Figure 4. Total PM2.5 data overview

This graph shows the curves for validation set data, real data, and forecast data. The blue line is the predicted result of the system that processes the validation set data, that is, the predicted data. The red line is the verification machine data. The green line is the real data. It can be seen from this figure that the prediction results are similar to the data of the verification set. The prediction of the true value is not particularly ideal. The first 8 prediction data is relatively ideal and can be referred to. The first 8 data means that the model can roughly predict the PM2.5 status from 0 am to 8 am.
Figure 5. Comparison chart of PM2.5 predicted and true values

The abscissa of this graph is the validation set data, and the ordinate is the relative error. As can be seen from the figure, as the verification set continues to deepen, the trend of the relative error of the predicted values given by the network is gradually increasing. It is because that the network belongs to a network with delay, and the new predicted value is weighted by the first few one. Therefore, the later predicted value is based on the previously predicted value, which will lead to errors. Finally, the accumulation of the final error is getting bigger and bigger.

Figure 6. PM2.5 relative error chart

\[
\text{Pred\_error} = \frac{\text{abs}(y\_pred - y\_valid)}{y\_valid};
\]

\text{Pred\_error} is the absolute value of the relative error.
‘abs’ means absolute value.
y\_pred is the predicted value of the current point.
y\_valid is the valid value of the current point.

4. Research analysis
This article uses the RNN neural network to process the collected smog concentration data for the last month. It obtains the predicted image, draws the error analysis image, and carefully
analyzes the cause of the error and the feasibility of the whole model. Through the above error analysis and accuracy analysis, it can be deduced that the model can predict the last 8 hours’ data accurately. Therefore, the model can be applied to life practice and serve as guidance in actual life.

5. **Experiment conclusion**

1> High accuracy. It can be seen from the above result graphs that the whole system has high accuracy for predicting PM2.5 and TVOC concentration, and the error is basically below 0.2. When the predicted time is shorter, it is able to have higher accuracy. Therefore, accuracy is a highlight of this experiment.

2> Fast speed. The entire system processes quite fast, and it takes only 10 seconds to complete the analysis of 2050 data. Although it needs to ensure high precision, speed is as fast as possible. This brings hope to the practical application, in other words, through this system, it is possible to quickly predict the concentration of pollutants in real time and provide a good reference for the staff to take measures in advance.

3> Strong applicability. The data used in this experiment is collected in the field, so it is universal and can be predicted by using this method in a similar environment.

4> Low cost. Compared to many other complicated prediction methods, this method only needs past data without additional cost.

6. **Outlook**

1> The data in this experiment is derived from field measurements, but using computer software simulation can save a lot of manpower and material resources.

2> The parameters in the network model of this experiment have not yet reached the most optimal level, and theoretically, the error can be further reduced.

3> During the experiment, it was found that the method could be applied to other fields, and is currently actively experimented to observe the performance of the neural network in different fields. The next paper plans to use computer software for field simulations, adjust the parameters of the neural network, and finally complete the prediction.

7. **Conclusion**

From the above analysis, it can be seen that the method has great significance for the prediction of PM2.5 concentration, which is reflected in more accurate prediction of the data in the first 8 hours of a day after analyzing the previous month's data. The time when people are exposed to a large amount of PM2.5 is exactly 8 hours. Therefore, the progress of this method is reflected in the ability to predict the PM2.5 concentration on the next day, so that they have sufficient time to take protective measures. However, there are still some shortcomings in this method, for example, it can not accurately predict the PM2.5 data after 8 hours. Therefore, some improvements to the model are needed, such as adding more neurons or adding more layers to make the model more accurate.

**References**

[1] Wang Whiz. Study on low-visibility weather analysis and prediction methods of haze[A]. Chinese Meteorological Society, Meteorological Society of China Annual Meeting of Atmospheeric Composition Observation, Research and Forecasting Session [C]. Chinese Meteorological Society: China Meteorology Learning, 2007: 5.

[2] Hou Putonghua, Yang Hang. Analysis and prediction of haze weather based on cubic exponential smoothing model[J]. Environmental Science, 2014, 40(06):73-77

[3] Sun Zhengzhou, Bhang Biting, Thu Jungian, Wang Yang. Construction of Haze Model by Computer Simulation Technology[J]. Electronic Technology and Software Engineering, 2015(17): 175-176.
[4] Jan Kleine Deters, Rasa Zalakeviciute et al., “Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters”. Journal of Electrical and Computer Engineering Vol 2017, pp14-19, 2017. 6.

[5] Y. Li, Q. Chen, H. Zhao, L. Wang, and R. Tao, “Variations in pm10, pm2.5 and pm1.0 in an urban area of the Sichuan basin and their relation to meteorological factors,” Atmosphere, vol. 6, no. 1, pp. 150–163, 2015. 3.

[6] X. Xi, Z. Wei, R. Xiaoguang et al., “A comprehensive evaluation of air pollution prediction improvement by a machine learning method”, in Proceedings of the 10th IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI 2015 - In conjunction with ICT4ALL’15, pp. 176-181, Hammamet, Tunisia, November 2015.