Application research on PM2.5 concentration prediction of multivariate chaotic time series

Yun Zheng¹, Qiang Zhang¹*, Zhihe Wang¹ and Yunan Zhu¹
¹College of Computer Science & Engineering, Northwest Normal University, Lanzhou, Gansu, 730070, China
*Corresponding author’s e-mail: 157815084@qq.com

Abstract. PM2.5 is affected by complex factors such as meteorological elements in the air system and other pollutants in the air. So PM2.5 has chaotic property, which makes the prediction of PM2.5 concentration extremely difficult. In order to improve the prediction accuracy of PM2.5 concentration, this paper introduces the chaotic time series prediction method to establish multivariate time for PM2.5 concentration. The sequence prediction model achieves short-term predictions based on the hour concentration of PM2.5 in Beijing. Firstly, the chaotic time series phase space of the relevant unit is expanded into the multi-time sequence phase space, and the multi-time sequence phase space matrix of PM2.5 concentration is constructed. Then the RBF neural network is used to predict the state point in the multi-phase space system. The phase space points of the PM2.5 concentration sequence is separated for prediction. Finally, the comparison between the prediction model and the traditional prediction model is carried out. The results show that the root mean square error of the predicted PM2.5 concentration in the multivariate chaotic time series prediction model based on the phase space reconstruction is 4.92% in the next 5 hours. The average absolute error is 2.40%, which is more effective than the commonly used statistical prediction method.

1. Introduction
In recent years, haze is frequent in some areas, PM2.5 prediction based on time series is the most commonly used research method. In terms of modelling input dimensions, time series prediction methods can be divided into univariate and multivariate methods. Generally multivariate methods are more accurate than the univariate time series model predictions. The commonly used prediction models, such as AR and ARMAX. Yu Hui and others use ARMAX to achieve tracking and prediction of hour’s concentration of PM2.5[1]. Regarding the above prediction models, all of them are based on linearity. However, in the air pollution timing, a large number of time series are nonlinear time series. Using a linear model to describe a nonlinear time series often fails, therefore the prediction of nonlinear prediction model is for PM2.5 prediction, it is crucially important. Nonlinear time series modelling and prediction, such as threshold method[2], exponential model, local linear model[3], nearest neighbour regression[4], neural network, etc. are widely used, and the neural network has better nonlinear processing ability, which can better cope with the analysis and prediction of complex nonlinear systems such as air pollutants.

In these observations, there are nonlinear chaotic features, and the multivariate time series contains more information than the single variable sequence. By extracting the information of multiple variables in the system, the prediction accuracy can be improved. For example, Han Min et al. applied the multivariate chaotic time series prediction model to the single-step prediction of monthly average
temperature in Dalian[5]. A good prediction result has been obtained. Applying multivariate chaotic time series prediction method to PM2.5 concentration prediction can fully exploit the evolution law of air pollutants, which will further improve the prediction accuracy of PM2.5. It has broad prospects, and the current chaotic time series prediction method research on the prediction of air pollutants is in absence. Therefore, this paper introduces chaos theory and RBF neural network, and proposes a multivariate chaotic time series of PM2.5 concentration prediction model based on phase space reconstruction to overcome the problem of low prediction accuracy of traditional prediction models.

2. Construction of multivariate chaotic time series prediction model

In the actual forecasting application, the observed time series data of the system is multi-dimensional, especially in the subsequent empirical research in this paper. There are many time series that need to be included in the input of the prediction model, so in order to ensure accuracy and reliability in the actual prediction ,this section is based on phase space reconstruction and chaotic time series neural network prediction theory[6]. Firstly, the phase space reconstruction of multiple time series is carried out, and the phase space of a single time series is expanded into a multi-variable time series phase space. The RBF neural network algorithm is then used to construct a multi-chaotic time series prediction model based on phase space.

2.1 Phase space reconstruction of multivariate chaotic time series

In practical applications, the unit chaotic time series is limited and contains noise. Therefore, the system information obtained by phase space reconstruction based on single time series is inevitably missing, and cannot effectively restore the original system features. Accordingly, phase space reconstruction of multiple time series is required at the same time, and multiple time series is mapped into the high-dimensional phase space.

Survivability is a fundamental characteristic of large-scale network, and will not disappear due to system evolution or external environment changes. In the large-scale network, there is information exchange between the failed subsystems and other subsystems, which may cause the cascaded failures. The increase of failed subsystems may lead to the whole large-scale network failure. \( S \) time series is analysed \( X_1, X_2, \cdots, X_S \), among them:

\[
X_q = \{x_{q,1}, x_{q,2}, \cdots, x_{q,q}\}, q = 1, 2, \cdots, S
\]  

(1)

If the time delay of each time series is \( \tau_1, \tau_2, \cdots, \tau_S \), the embedded dimension is \( m_1, m_2, \cdots, m_S \). The state coordinates of the multivariate time series at the time of the phase space \( i \) can be expressed as:

\[
H_i = (x_{i,1}, x_{i,1-\tau_1}, x_{i,1-(m_1-\tau_1)}; x_{i,2}, x_{i,2-\tau_2}, x_{i,2-(m_2-\tau_2)}; \cdots; x_{i,S}, x_{i,S-\tau_S}, x_{i,S-(m_S-\tau_S)})
\]  

(2)

Each phase point represents the state of the air system at a certain moment, and the phase points are sequentially connected in time sequence to obtain the air system change trajectory of the PM2.5, which closely reflects the evolutional information of the entire air pollutant system.

2.2 Construction of multivariate chaotic time series model based on RBF neural network

The RBF neural network has the characteristics of approximating arbitrary nonlinear functions and can be processed within the system. It is difficult to analyse the regularity, and secondly it has a good generalization ability and relatively high learning convergence speed. Moreover, it has a good effect in dealing with nonlinear function approximation problems, and displays a good fitting ability to the original chaotic system evolution trajectory, especially in noisy sequences. The prediction error of RBF neural network model is relatively small. Therefore, RBF neural network is suitably used to predict chaotic time series. The radial basis neural network activation function selects the Gaussian function. When the multi-phase space state coordinates are input, the function goes as follows:

\[
R(x_p - c_i) = \exp\left(-\frac{1}{2\sigma^2}\|x_p - c_i\|^2\right)
\]

(3)
In equation (3), \( \|x_i - c_i\| \) is the European norm, \( \sigma \) is the variance of Gaussian function, \( c_i \) is the centre of the Gaussian function. The network output is:

$$y_j = \sum_{i=1}^{s} \omega_{ij} \exp\left(-\frac{1}{2\sigma^2}\|x_i - c_i\|^2\right)$$  \( j = 1, 2, \ldots, n \)  \( (4) \)

In equation (4), \( x_s = (x_{s1}, x_{s2}, \ldots, x_{sn}) \) is the first \( s \) input of multi-phase space state points, \( s = 1, 2, \ldots, S \) is the total status point, \( \omega_{ij} \) is the connection weight between the hidden layer and the output layer; \( c_i \) is the centre of the hidden layer node; \( i = 1, 2, 3, \ldots, h \) is the number of hidden layer nodes; \( y_j \) is the actual output of the jth output state point of the network corresponding to the input state point. Assuming \( d \) is the desired output of the multi-timed phase space state point, the basis function variance can be expressed as equation (5):

$$\sigma = \frac{1}{p} \sum_{j} |d_j - y_j, c_j|$$  \( (5) \)

In this step, all the state points of the multiphase space are taken as input samples, and the multiphase space is mined \( H \) phase points and their next evolution \( H + 1 \) first \( m \) dimensional relationship \( f \). The RBF neural network is used to predict the phase space points of the next evolution, and the predicted value of the time series to be predicted is separated from the predicted result. The specific description is as follows: in a multidimensional phase space, there is a smooth map \( f \), which is the map of the chaotic state space, and the phase space trajectory can be expressed as equation (6):

$$H(i + 1) = f(H(i)), i = 1, 2, \ldots, M$$  \( (6) \)

This kind of approximate mapping relationship in practical applications is usually constructed according to the given data, so that it approximates the theoretical (6) mapping relationship \( f \). In this paper uses RBF neural network to construct the mapping \( f' \) to approximate \( f \), and the next evolutionary phase point \( H_{i+1} \) of the phase point \( H_i \) is predicted by the state point in the fitted phase space.

3. Prediction and application based on PM2.5 concentration of multivariate chaotic time series

Taking Beijing PM2.5 concentration prediction as the research object for there are many cities suffering from haze in China, and Beijing is one of the most serious ones. The reason for the formation of haze in Beijing is the increase of air pollution. These pollutants are mainly PM2.5, PM10, sulfur dioxide and nitrogen dioxide, while the PM2.5 is the primary pollutant, which is harmful to the human body[7]. Second, the PM2.5 concentration is closely related to meteorological elements such as air humidity, pressure, dew point, and wind speed[8][9]. According to the current air pollutant index and meteorological element data, simple prediction modeling cannot effectively predict PM2.5 concentration, so the prediction model constructed in this paper is used to predict PM2.5 concentration in the perspective of system analysis.

3.1 Data source

Since PM2.5 concentration and other air pollutants and meteorological elements have strong dynamic variability, they are greatly affected by random factors. By quantifying the data and mining the hidden laws in the data, it requires a long time and a high frequency data, so the data used in this paper is historical air pollutants and historical meteorological elements. Because the long-term change trend of PM2.5 in Beijing urban area is relatively consistent, the pollution situation of PM2.5 in Beijing will be reflected by the data of the United States Embassy in China as a single observation point. In order to ensure the sufficient experimental data and data quality, the time series data of annual historical hourly data from 2010 to 2014, which was published by the U.S. Embassy in China, was used to acquire meteorological element data. The data was collected from UCI data sets, including air pressure, temperature, wind direction, wind speed, dew point and other indicators. In addition to PM2.5
concentration data from the above UCI data set, other pollutant data is derived from Data Open Network (http://beijingair.sinaapp.com/).

3.2 Air system time series missing data to fill
Due to many unavoidable reasons such as detection equipment failure, abnormal detection procedures, or improper handling of humans, the detection data is often incomplete. The lack of data usually causes great errors in the research results. Therefore, in order to ensure the continuity of the time series, subsequent analysis processing, model integrity, it is needed to deal with missing data. It is necessary to consider all time series comprehensively and adopt a multivariate time series filling method. According to the timing characteristics of the air system, the literature is adopted [10] the multivariate time series missing data filling method for PM2.5, dew point, temperature, air pressure, wind direction, wind speed, PM10, SO2, NO2, O3, from 0:00 on May 13, 2014 to 23:00 on November 12, 2014, were filled by using the missing data filling method of time series. “Figure. 1” and “figure 2” are randomized to set the missing point value filling effect. The combined threshold filling method is obviously better than the single univariate filling and multivariate filling methods. Therefore, based on the combined threshold filling method, the missing values of PM2.5, dew point, temperature, air pressure, wind direction, wind speed, PM10, SO2, NO2, O3, CO time series are filled. According to “figure. 1” and “figure. 2”, it is obvious that the multi-variable data filling method is better.

3.3 PM2.5 concentration and impact factor analysis
Before modelling and predicting the PM2.5, in order to ensure the prediction accuracy of the prediction model, it is necessary to consider the influence degree of PM2.5 multiple influencing factors on PM2.5 concentration. Here, the dew point, temperature, air pressure, wind direction, wind speed, PM10, SO2, NO2, O3, CO, PM2.5 index time series are set to \( X_1, X_2, \ldots, X_{10} \). PM2.5 sequence is \( Y \). From \( X_1, X_2, \ldots, X_{10} \) in filter out \( Y \) correlation is the strongest impact factor combination. So calculation is required \( X_1, X_2, \ldots, X_{10} \) all combinations of timing and \( Y \) the complex correlation coefficient between the sequences. Complex correlation coefficient can be used to measure the amount of closeness of a linear relationship between a variable and multiple variables, or to characterize the regression variance of a variable with a plurality of variables as a variable proportion of variance [11]. So calculate \( Y \) the correlation coefficient between the time series group of influencing factors and the independent variable group can be used to estimate the degree of correlation between the time series of the impact factor and PM2.5.

In order to ensure the accuracy of the calculation of the complex correlation coefficient, the literature has used [12] the single dependent variable PLS stepwise regression method \( Y \) variable group \( X \) absolute regression equation \( \hat{Y} \). Find the largest complex correlation coefficient according to the above method \( R \) for 0.8752.

3.4 Air system chaotic characteristics identification
In practical applications, it is very important to judge whether each time series has chaotic characteristics before the phase space reconstruction of the multivariate chaotic time series.
The largest Lyapunov exponent method is a quantitative method to judge the chaotic characteristics of time series by calculating the system attractor invariant parameters. For a complex nonlinear time series of air systems, it is usually sufficient to calculate its maximum Lyapunov exponent. In order to ensure the accuracy of the prediction model and avoid the subjective discriminant method, the maximum Lyapunov exponent method is used to judge the chaos of time series, several air system time series variables participating in phase space reconstruction are selected. The small data method is adopted in this paper [13] to determine the chaotic characteristics of the air system time series in this paper. Extract the \( X_i, X_2, \ldots, X_{10}, Y \) as sequence maximum Lyapunov exponent.

| Sequentially | \( \lambda \)  | Sequentially | \( \lambda \) |
|--------------|---------------|--------------|---------------|
| \( X_1 \)    | -0.7531       | \( X_n \)    | 0.0056        |
| \( X_2 \)    | 0.0351        | \( X_s \)    | -0.0658       |
| \( X_3 \)    | 0.0337        | \( X_{10} \) | 0.0053        |
| \( X_4 \)    | -0.0022       | \( Y \)      | 0.0301        |

According to the largest Lyapunov index \( \lambda > 0 \) the principle can be used to judge the chaotic characteristics of the corresponding time series. From the maximum given in Table 2, the fields satisfying this condition are five time series of temperature, air pressure, PM10, CO and PM2.5.

4. PM2.5 concentration prediction simulation

This section predicts and analyses the PM2.5 hourly concentration based on the constructed prediction model. To further verify the accuracy of the model for PM2.5 prediction, the conventional ARMA and multiple linear regression time series prediction models were used to predict the hourly concentration of PM2.5.

4.1 Prediction of PM2.5 concentration in multivariate chaotic time series based on phase space reconstruction

The 1 to 4343 sample points were used as multivariate phase space reconstruction datasets to normalize the air pollutants and meteorological data of different dimensions, that is by using the mapmimax function in matlab201b to normalize the temperature, pressure, PM10, CO and PM2.5 timing values of the four influencing factors to the \([0, 1]\) interval, we can obtain a phase space matrix of 4298*17 air system.

The trained training set is trained by using the newrb function in matlab. The PM2.5 concentration values of 72 sample points after 4343 are predicted by the sim function and the predicted values are denormalized. Set the best training parameters of RBF according to multiple experiments, learning target error \( err\_goal = 0.001 \), radial basis function expansion factor \( spread \) is 80, the maximum number of iterations is 100, and the number of neurons added between the two displays is 1. During the training, when the iteration reaches the 14th stop training, the learning target error reaches the preset value, and the prediction result is shown in table 2.

| Timing test sample point | Time         | Actual value | Predictive value | Relatively error (%) |
|--------------------------|--------------|--------------|------------------|----------------------|
| 4344                     | 2014-11-9-23 | 101          | 100.7            | 0.13                 |
| 4345                     | 2014-11-10-1 | 104          | 103.7            | 0.14                 |
| 4346                     | 2014-11-10-2 | 106          | 106.2            | 0.08                 |
| 4347                     | 2014-11-10-3 | 108          | 108.2            | 0.08                 |
| 4348                     | 2014-11-10-4 | 98           | 109.0            | 0.05                 |
| 4349                     | 2014-11-10-5 | 79           | 106.3            | 11.40                |
| 4350                     | 2014-11-10-6 | 44           | 103.4            | 24.84                |
| 4351                     | 2014-11-10-7 | 33           | 101.2            | 28.55                |
From the prediction results, the predicted value in the first 5 hours is more accurate than the predicted value after 5 hours, which also verifies that the chaotic time series prediction can only make short-term prediction conclusions.

4.2 Analysis of simulation results of PM2.5 concentration prediction

From the prediction results of each model prediction, the predicted values of PM2.5 concentration are all valid in the first 5 hours. In order to verify the advantages of the multivariate chaotic time series prediction model constructed in this paper in the PM2.5 concentration prediction.

According to the simulation experiments of PM2.5 concentration prediction of each model mentioned above, it can be predicted that the short-term PM2.5 concentration value in the next 5 hours, it was found that the predicted time is longer, the accuracy is lower. The table 4 is a comparison of prediction results of each prediction model within 5 hours, while table 3 is a comparison result of performance of each model.

Table 3. Comparison of PM2.5 prediction performance of each model.

| Predictive model type                          | RMSE  | MSE  |
|------------------------------------------------|-------|------|
| Multivariate chaotic time series prediction   | 4.92  | 2.40 |
| Cell chaotic time series prediction           | 10.05 | 9.27 |
| ARIMA time series prediction                  | 5.50  | 3.46 |
| Multiple linear regression time series prediction | 100.80 | 45.08 |

Through experimental comparison, the root mean square error and average absolute error of the multivariate chaotic time series prediction model based on phase space reconstruction are shorter than the unit chaotic time series prediction model ARIMA, multiple linear regression. The prediction accuracy is higher in the short-term prediction than the other three models, and the performance is better.

Table 4. Different models predict PM2.5 concentration \( (\mu g/m^3) \) value result

| Date        | PM2.5 true value | Predictive value | Relative error (%) | Predictive value | Relative error (%) | Predictive value | Relative error (%) | Predictive value | Relative error (%) |
|-------------|------------------|------------------|-------------------|------------------|-------------------|------------------|-------------------|------------------|-------------------|
| 2014-11-9-23| 101              | 100.7            | 0.13              | 96.7             | 1.79              | 100.4            | 0.37              | 145.3            | 41.94             |
| 2014-11-10  | 104              | 103.7            | 0.14              | 95.4             | 3.59              | 101.2            | 2.59              | 147.5            | 40.00             |
| 2014-11-10  | 106              | 106.2            | 0.08              | 94.1             | 4.97              | 101.2            | 4.44              | 147.4            | 38.33             |
| 2014-11-10  | 108              | 108.2            | 0.08              | 92.8             | 6.35              | 100.8            | 6.67              | 151.7            | 40.09             |
| 2014-11-10  | 98               | 109.0            | 0.05              | 91.5             | 2.70              | 100.1            | 1.94              | 150.2            | 48.33             |

5. Conclusion
In this paper, the PM2.5 concentration prediction model of multivariate chaotic time series prediction method based on phase space reconstruction is described in detail, and the model is applied to Beijing PM2.5 concentration prediction, and the simulation result is compared with the chaotic time series prediction model, ARIMA model and multiple linear regression model. The test sample was validated from 23:00 on November 9, 2014 to 3 o'clock on November 10, 2014. According to the prediction performance evaluation index, the prediction performance of the multivariate chaotic time series PM2.5 prediction model constructed in this paper is more advantageous than the traditional time series prediction model.

Acknowledgment
This work was supported by the National Natural Science Foundation of China under Grant No.71764025.

References
[1] Yu H, Yuan J, Yu X Y, Zhang L X, Chen W L. (2017) Tracking Prediction Model for PM2.5 Hourly Concentration Based on ARMAX, Natural Science and Engineering Technology Edition), 50 (01): 105-111.
[2] Tong H. Threshold models in time series analysis—Some reflections (2015) Journal of Econometrics, 189(2):485-491.
[3] Wang J. Research On Local Model-Based Time Series Prediction (2007) Harbin Institute of Technology, Harbin.
[4] Prokhorov M D, Ponomarenko V I, Khorev V S. (2013) Recovery of delay time from time series based on the nearest neighbour method, Physics Letters A, 377(43): 3106-3111.
[5] HAN M, XU M L, REN W J. (2014) Research on Multivariate Chaotic Time Series Prediction Using mRSM Model, Acta Automatica Sinica,40(05): 822-829.
[6] Lu J H, Lu J A, Chen S H. (2002) Chaotic Time Series Analysis and Its Application, Wuhan University Press, Wuhan.
[7] Yang X X, Feng L H, Wei P. (2012) Air Particulate Matter PM2.5 in Beijing and Its Harm, Frontier Science, (01):22-31.
[8] Li L Z, Cao L, Wang L, Liu H, Shi X F, Gong X L. (2013) On the Pollution Sources and Hazards of PM2.5 in China, Energy and Energy Conservation, (04): 77-78.
[9] Zhang H, Chen Q, Xie B, Zhao S Y. (2014) PM2.5 and Tropospheric Ozone in China and Pollutant Emission Controlling Integrated Analyses, Advances in Climate Change, (04): 289-296.
[10] Li Z X, Zhang F M, Wang Y, Tao Q, Li C.(2018) Method for missing data imputation for multivariate time series, Systems Engineering and Electronics, 40(01): 225-230.
[11] Meng X Z, Zhang H X, Zhang T Z, Wang H M, Liu Q, Shang H L.(2012) The Response Surface Approximation Method of HCPE Connecting Rod Design Optimization, Journal of Qingdao University(Engineering & Technology Edition), 27(02):48- 52.
[12] WANG Z H, CHEN X H, ZHANG Q. (2017) PLS Model of Single Dependent Variable Applied in Predicting PM2.5 Real-Time Concentration, Computer Simulation, (10): 387-391+413.
[13] Duan Z H. Application of chaos theory and combination theory in daily runoff and load forecasting (2004) Huazhong University of Science and Technology, Wuhan.