Urban environmental pulsation research based on Wavelet Decomposition enhanced ARMA model

Xiang Li[1,2,a] Ling Peng[1,b,*] Tianhe Chi[1] Congcong Wen[3] Yizhi Xu[1,2]

[1] Institute of Remote Sensing and Digital Earth, CAS, Beijing 100094, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] East China, China University of Petroleum, Qingdao 257061, China

[a]lixiang01@radi.ac.cn, [b]pengling@radi.ac.cn

Keywords: city pulsation; air pollutant concentration; water quality; Wavelet Decomposition; ARMA

Abstract. With the rapid development of Internet of Things technologies, smart city public information platform is collecting more and more city data. Make full use of these data, to analyze the pulsation rule of city development, is very important for solving the existing problems in cities. Relying on the Sino-Singapore Tianjin Eco-city (SSTEC) data collecting platform and city pulsation project requirements, this paper explored the data source for environmental pulsation analysis. Then by analysis the time series features of data, we chose the wavelet enhanced ARMA model to model and predict the SSTEC air pollutant concentration data and water quality data. The result shows that, wavelet enhanced ARMA model is a proper method for environmental prediction which has a relatively higher prediction accuracy. Our results provide a guidance for urban environmental management, also appropriate in some other smart city application fields.

Introduction

Rapid development of Internet of Things and cloud computing technologies vigorously promotes the transition from digital city to smart city. Smart city constructions are carrying out all over the country in full swing. Smart city public information platform is collecting more and more city data, including population, economy, transportation, energy, environment, etc. However, because of varying data quality, intricate data structure, inefficient data mining method, we get big data but few information. Make full use of these data to analyze the pulsation rule of city development is of great importance to improve citizen living quality as well as city management efficiency. City pulsation analysis based on spatial-temporal data mining has become an urgent needs for smart city construction [1].

SSTEC is the first batch of state-level wisdom city pilot, one of the government strategic cooperation project between the China and Singapore. It has a profound understanding of smart city construction and rich experience in construction of smart city public information platform, which has formed some preliminary results and practical application. SSTEC's next target is to deepen and optimize infrastructure construction of public information platform, strengthen the data analysis and information mining, and make the platform become the core infrastructure of eco-city informationization construction and the exhibition of city development.

In recent years, with the rapid development of national economy, the speeding up of urbanization and the expansion of industrial scale, the environmental problem has become increasingly serious [2, 3]. SSTEC is committed to build a "resource saving, environment friendly, economic boom, social harmony" new city. Based on the requirement of Ecological environment health, social harmony and progress, economy booming regional fusion of efficient, SSTEC determined 22 controlling indexes and four leading indexes, environment friendly, as one of the most important indicators ranked top. In this paper, we explored the urban environmental problem based on SSTEC city pulsation analysis project requirements. Our results provide a guidance for urban environmental management.

SSTEC Environmental Pulsation Requirements and Solutions

SSTEC Environmental pulsation requirements include:

1) Data collecting: Collecting the environmental data, using Data Base, SQL, and Hadoop technologies to properly organize the data.

2) Modeling: Through exploring the spatial-temporal features of environmental data, we can choose proper data mining models to reveal the pulsation rule of data. Common models includes multiple linear regression model, ARMA model, Grey forecast model, SVR model, ANN model, and so on.

3) Visualization: Using some visualization techniques, including traditional statistical graph, HTML5 technologies, open source tools, such as D3.js and ECharts, to vividly show the pulsation rule of our environmental data.

To satisfy SSTEC’s environmental pulsation requirements, we use air quality data and water quality monitoring data to conduct some experiments. First, we analyzed the time series features of data, then chose the proper time series prediction model based on these features. After that, we use our model to conduct the modeling and prediction experiments. The overall framework of environmental pulsation analysis shows in figure 1. This paper mainly talks about the data features exploration, model selection and modeling and analysis.

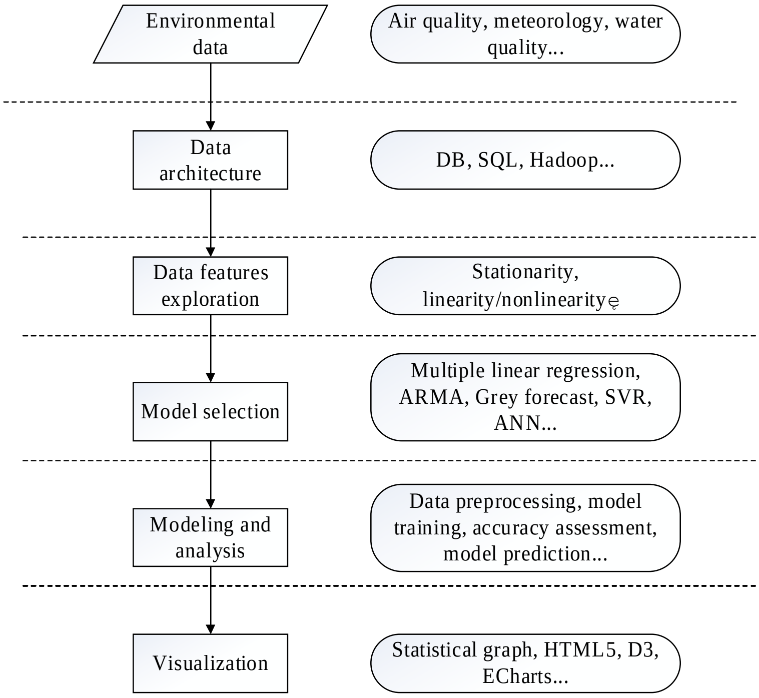

Environmental data
Air quality, meteorology, water quality...
Data architecture
DB, SQL, Hadoop...
Data features exploration
Stationarity, linearity/nonlinearity
Model selection
Multiple linear regression, ARMA, Grey forecast, SVR, ANN...
Modeling and analysis
Data preprocessing, model training, accuracy assessment, model prediction...
Visualization
Statistical graph, HTML5, D3, ECharts...

Fig.1. SSTEC environmental pulsation analysis framework

Time Series Features of Environmental Data

Data profile

SSTEC air quality monitoring data profile:

Sites summary	1 (Luzhuang Garden air quality automatic monitoring station)
Monitoring time	2009/6/24 to 2015/5/14
Monitoring factors	SO2[mg/m^3]、NO2[mg/m^3]、CO[mg/m^3]、O3[mg/m^3]、PM10[mg/m^3]、PM2.5[mg/m^3]

SSTEC water quality monitoring data profile:

Sites summary	13(6 water supply monitoring sites and 7 surface water monitoring sites)
Monitoring time	2012/11/1 to 2015/5/14
Monitoring factors	pH value, suspended solids[mg/l], COD[mg/l], TOC[mg/l], nitrogen ammonia[mg/l], temperature[degree], phosphorus[mg/l], flow[m^3/h], TDS[mg/l], nitrate[mg/l], conductivity[s/m], dissolved oxygen[mg/l]. Varying from different sites.

Some of the air quality and water quality monitoring data are missing, we adopted the simple linear interpolation method to fill them.

Stationarity

Stationarity means no trend in data series, that is, joint distribution and conditional distribution don’t change with time. In statistical perspective, which means the mean value, variance and covariance of data series don’t change with time [4]. Unit root test is the standard method for stationarity test, including Augmented Dickey-Fuller test (ADF), Dickey-Fuller Test with GLS (DFGLS), Phillips-Perron, KPSS, ERS and NP method, etc.[5].

In this paper, we used ADF test to validate the stationarity of air quality data and water quality data. The results show that, test statistics P-Value are all less than 0.001, which means significant stationarity of the time series data.

Linearity/nonlinearity

Common nonlinearity detecting method includes McLeod-Li test, Bispectral test, BDS test, RESET test, F test, and Neural Network nonlinearity test. McLeod-Li test is the most common way. We used the McLeod-Li test to conduct a linearity test for air quality data and water quality data. The results show that, test statistics P-Value are all less than 0.001, which means significant nonlinearity of the time series data.

Models

At present, the commonly used time series data forecasting methods mainly include multiple linear regression models, Auto-Regressive and Moving Average Model (ARMA), gray prediction, support vector regression model, artificial neural network model, etc. Among them, the ARMA model is an effective tool for the modeling of stationary time series, and it has been widely used in the fields of environmental forecasting [6, 7]. However, the ARMA model is a linear model which can only extract the global rule of time series data, and it can hardly show the influence of any short-term events and unexpected events [8]. In order to solve this problem, some scholars [5, 8, 9] proposed the improved ARMA model based on wavelet multiscale analysis. Wavelet multi-resolution analysis can extract the multiscale features of the sequence. In this way, the research of the complex multiscale time series is transferred into the study of multiple time series of different frequency. So the decomposed different sequences can be effectively analyzed by using different methods according to their characteristics.

In this paper, the process of the air pollutant concentration and water quality prediction shows in figure 2.

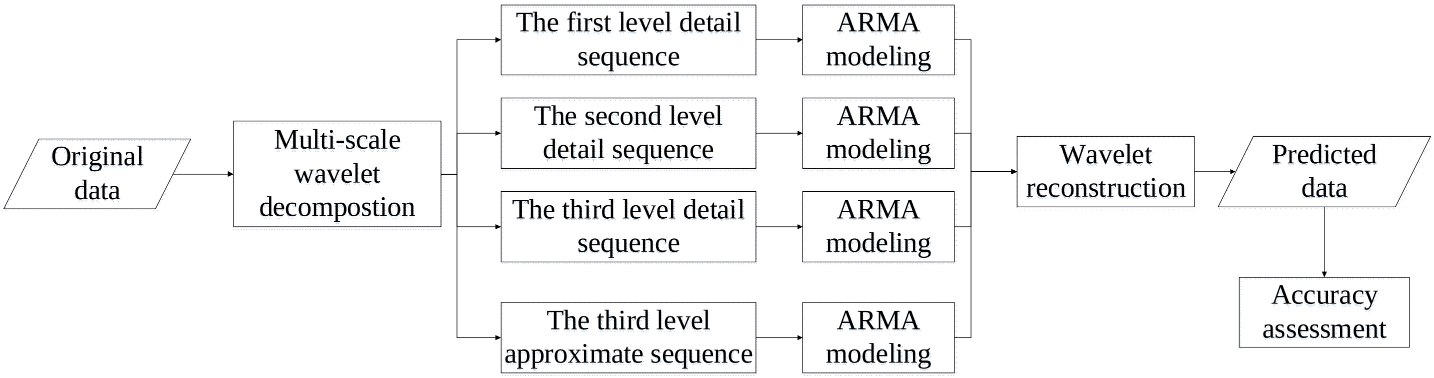

Original data
Multi-scale wavelet decompostion
The first level detail sequence
The second level detail sequence
The third level detail sequence
The third level approximate sequence
ARMA modeling
ARMA modeling
ARMA modeling
ARMA modeling
Wavelet reconstruction
Predicted data
Accuracy assessment

Fig. 2. Process of air pollutant concentration and water quality prediction

The wavelet multiscale decomposition

Multiscale analysis is an important method in wavelet analysis and application, and it is a method of spatial decomposition of signal. General using Mallat algorithm [10] to carry out wavelet multiscale decomposition. Using Mallat algorithm, the signal can be decomposed by layer by layer, the last time decomposed low frequency signal is decomposed into low frequency and high frequency two parts again which is the result of each layer decomposition. The calculation formula is as follows in Eq. 1 and Eq. 2.

$$\mathrm{A}_{\mathrm{j+1,k}} = \sum_m h_0(m-2k)A_{j,m} \tag{1}$$

$$\mathrm{D}_{\mathrm{j+1,k}} = \sum_m h_1(m-2k)A_{j,m} \tag{2}$$

j denotes decomposition scale, k 、 m denotes translation variable, $A_{j,m}$ denotes the approximation coefficient and the low frequency part, $\mathrm{D}_{\mathrm{j,k}}$ denotes the detail coefficient and the high frequency part, h_0 is the low pass filter and h_1 is the high pass filter. After decomposed wavelet coefficients can be used to reconstruct the original sequence, the reconstruction formula is Eq. 3.

$$\mathrm{A}_{\mathrm{j-1,m}} = \sum_k h_0(m-2k)A_{j,k} + \sum_k h_1(m-2k)D_{j,k} \tag{3}$$

ARMA model

ARMA model is a one of commonly used model that is used to modeling stationary random sequences, and its modeling and prediction steps are as follows:

1. data preprocessing

Data preprocessing includes zero mean processing and differential stationary processing.

2. model identification

Model identification is based on the autocorrelation function and partial correlation function. If the autocorrelation function or the partial autocorrelation function is truncated, that is p=0 or q =0, we can directly judge model order by using truncated characteristics. If autocorrelation function and partial autocorrelation function are not truncated, that is p≠0 and q≠0, generally using the Akaike Information Criterion of minimum selection model order.

3. parameter estimation

The estimation of model parameters is usually carried out by the least square method which has the advantages of simplicity, high accuracy, fast convergence and strong robustness.

4. model checking

After the establishment of the model, we shell test the model before prediction. If the test does not pass, then we adjust the model orders, and the parameters will be re-estimated and retest until the test passes. Autocorrelation function chart and DW coefficient are generally used to conduct the test.

5. forecast

According to the established model to conduct prediction analysis and accuracy evaluation.

Experiments

Air pollutant concentration forecast

1. data introduction

In this paper, two sets of experiments were carried out using the daily average data of air pollutant concentration from 2009/6/24 to 2015/5/14 in Tianjin city. Experiment one: select the daily average data of pollutant concentration in the first 358 days of 2014 as training data, and the remaining 7 days' data as the test data. Experiment two: select the first 65 months of monthly average data of pollutant concentration from June 2009 to April 2015 for training, and the remaining 7 months' data for accuracy assessment. In order to illustrate the prediction accuracy of our method, we use the daily PM2.5 data of SSTEC to conduct another experiment (experiment three), by adopting ARMA model, Support Vector Regression model, and Artificial Neural Network model to predict the data. In data processing, the wavelet multiscale decomposition and ARMA prediction model are implemented by Matlab software.

2. data preprocessing

The data preprocessing was carried out by using the maximum and minimum value normalization method. Using the db3 wavelet as the basis function to achieve wavelet multiscale decomposition. The time series is decomposed into an approximate sequence A and three wavelet transform sequences D1, D2 and D3. Among them, A is the approximate sequence of the original sequence, reflecting the trend of data; D1, D2 and D3 are the details of the original sequence, which reflects the small fluctuations of the sequence [11].

3. model training

Before the ARMA modeling, we use ADF test to validate the stationarity of decomposed sequence. The test statistic P-value of approximate sequence and detail sequence are all less than 0.001. It means that these sequence are stationary, and the ARMA model can be used to modeling these sequence. The model order is determined by the AIC criterion.

4. model prediction

The ARMA model is used to predict the approximate sequence and the detail sequence, then the prediction results are reconstructed through Mallat algorithm, and the final pollutant concentration is predicted.

5. experimental results

The results of the experiment one are shown in figure 3 (a) ~ (f), and the prediction accuracy shows in table 1. The results of experiment two are shown in figure 4 (a) ~ (f), and the prediction accuracy shows in table 2. The prediction accuracy of experiment three shows in table 3. In table 1 to table 4, the RMSE represents the root mean square error, the MAE is mean absolute error. The calculation formulas are as follows.

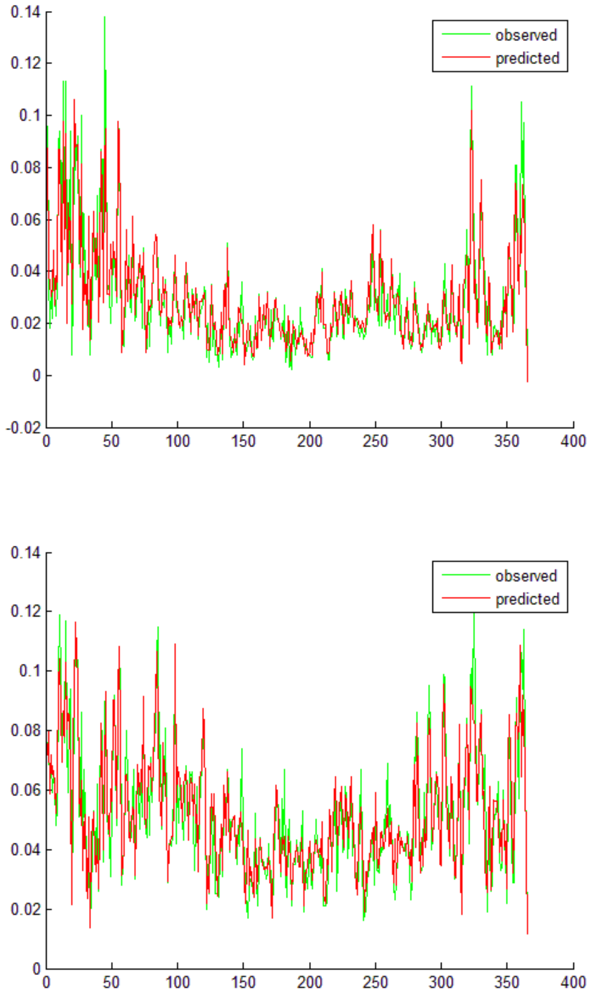
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| y_i - \hat{y}_i | \qquad (5)$$

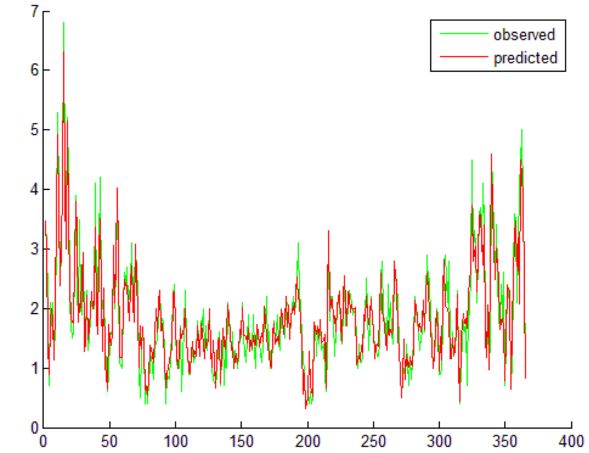
$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{| y_i - \hat{y}_i |}{y_i} \qquad (6)$$

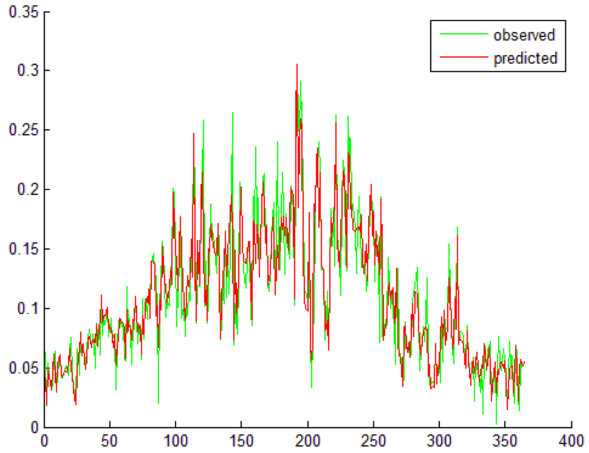
y_i denotes the observed values of the sample *i*, $\hat{y}_i$ denotes the fitting (prediction) value of the sample ***i***, and n for the sample number.

Through these graphs, we can figure out four main conclusions: 1) the wavelet decomposition enhanced model has an observably higher prediction accuracy, in contrast to ARMA model, SVR model and ANN model; 2) the wavelet decomposition enhanced model can fit the training data well, the mean relative error (MRE) is no more than 25 percent; 3) the prediction errors are relatively high than training errors, which means slightly over-fitting of our model; 4) the monthly prediction errors are relatively high than daily prediction errors, probably due to average effect.

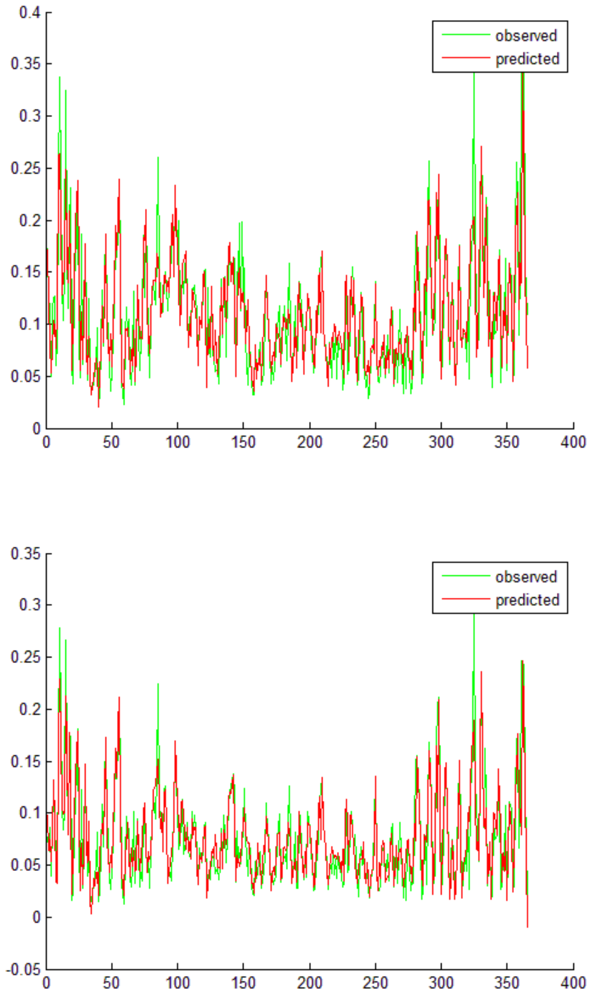


(a) (b)


observed
predicted
0
1
2
3
4
5
6
7
0
50
100
150
200
250
300
350
400

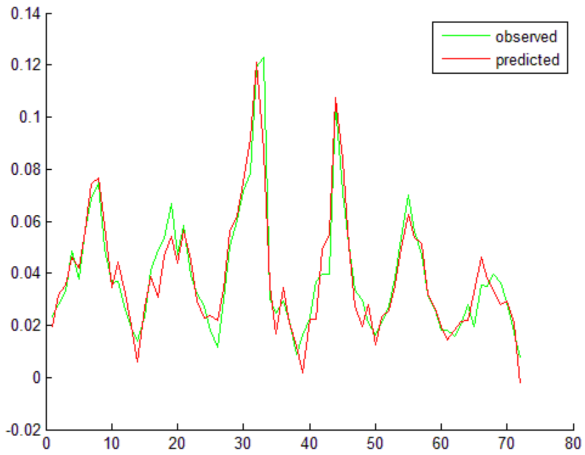

observed
predicted
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0
50
100
150
200
250
300
350
400

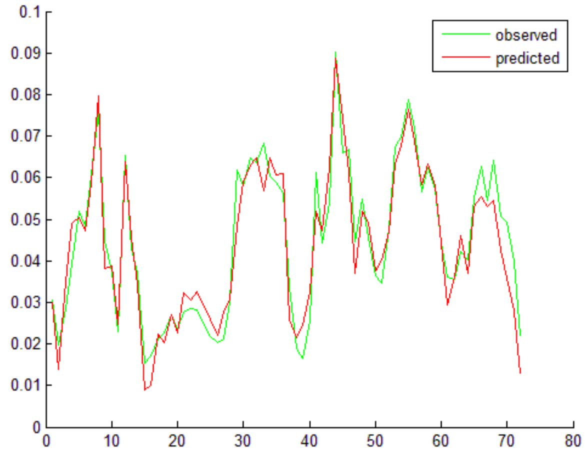
(c) (d)



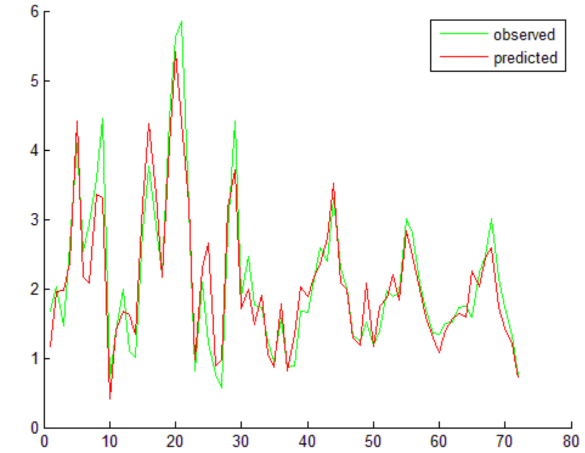
(e) (f)

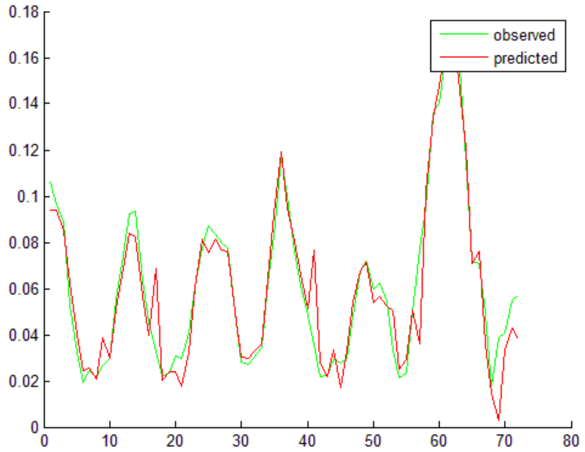
Fig. 3 (a) ~ (f). Experiment one. Observed and predicted values of the pollutants SO2 (a), NO2 (b), CO (c), O3 (d), PM10 (e), PM2.5 (f) in SSTEC, 2014.


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
0
10
20
30
40
50
60
70
80
observed
predicted


0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
40
50
60
70
80
observed
predicted

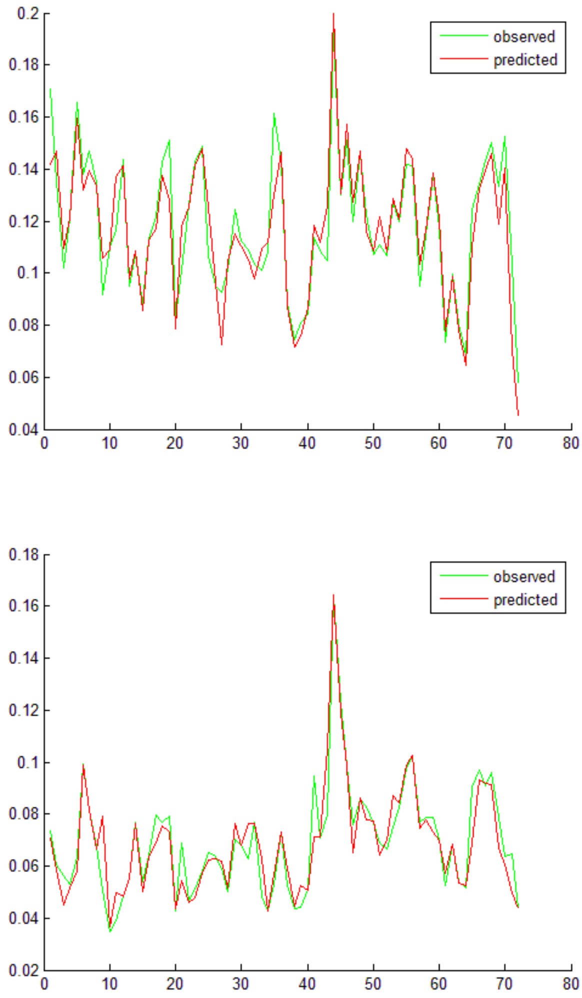
(a) (b)


observed
predicted
0
1
2
3
4
5
6
0
10
20
30
40
50
60
70
80


observed
predicted
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0
10
20
30
40
50
60
70
80

(c)

(d)



(e) (f)

Fig. 4 (a) ~ (f). Experiment two. Observed and predicted values of the pollutants SO2 (a), NO2 (b), CO (c), O3 (d), PM10 (e), PM2.5 (f) in SSTEC from June-2009 to April-2015.

Table 1. Experiment 1(Air pollutant concentration prediction accuracy)

	model fitting			model testing		
	RMSE	MAE	MRE	RMSE	MAE	MRE
SO2	0.008	0.005	25.92%	0.0264	0.0181	38.56%
NO2	0.3304	0.2054	13.07%	0.0216	0.0143	16.88%
CO	0.3304	0.2054	14.31%	0.8645	0.5849	18.81%
O3	0.0217	0.0126	16.99%	0.0145	0.0084	42.21%
PM10	0.0263	0.0163	18.93%	0.0902	0.0575	24.28%
PM2.5	0.0202	0.0124	23.06%	0.0678	0.0415	44.05%

Table 2. Experiment 2(Air pollutant concentration prediction accuracy)

	model fitting			model testing		
	RMSE	MAE	MRE	RMSE	MAE	MRE
SO2	0.0081	0.0058	18.00%	0.0069	0.0059	32.89%
NO2	0.0048	0.0038	10.73%	0.0094	0.0087	20.53%
CO	0.4276	0.3093	16.85%	0.2637	0.2154	10.62%
O3	0.0104	0.0065	14.19%	0.0171	0.0137	31.30%
PM10	0.0095	0.0064	5.47%	0.0151	0.0117	11.34%
PM2.5	0.008	0.0047	7.34%	0.0078	0.0058	7.78%

Table 3. Prediction accuracy of different models

	model fitting		model testing	
	RMSE	MAE	RMSE	MAE
SVR	0.051	0.030	0.097	0.085
ARMA	0.037	0.028	0.076	0.071
ANN	0.040	0.031	0.088	0.072
our method	0.020	0.012	0.068	0.042

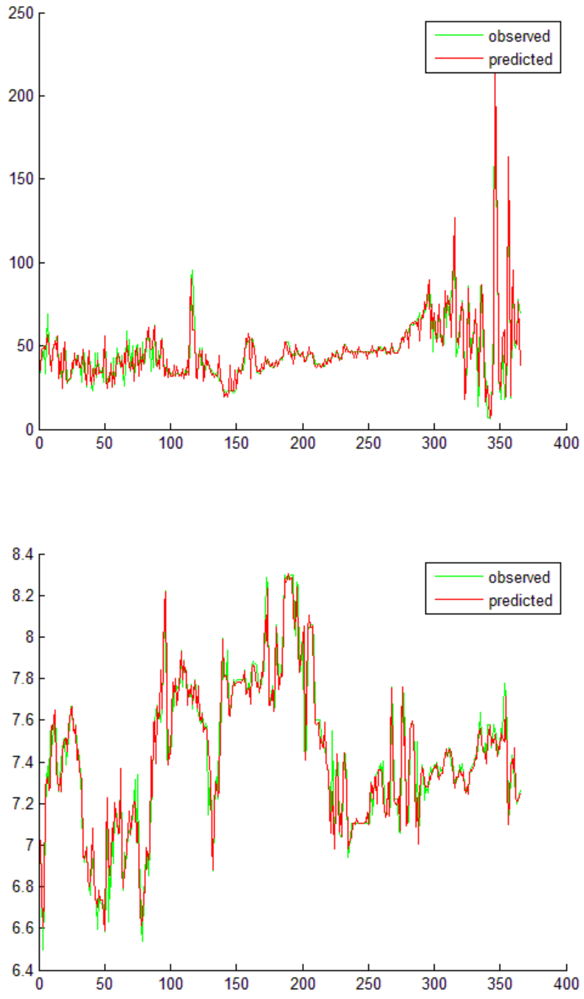
Water quality forecast

Using the water quality monitoring data of the SSTEC Jiyun River sewage factory outlet site in 2014 for experiments, the daily average data of pollutant concentration in the first 358 days as training data and the remaining 7 days data as test data. Experiment process is same as air pollutant concentration forecast. The results are shown in figure 5, the accuracy assessment is shown in table 4.

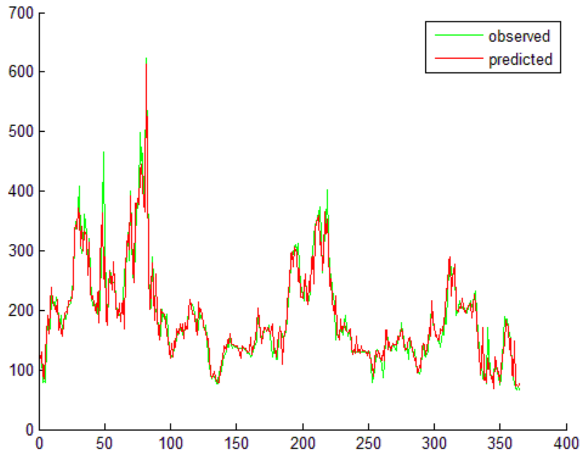
Through these graphs, we can easily find out that: 1) the training and testing accuracy are generally satisfying, which means that the wavelet decomposition enhanced model is a proper way for water quality prediction; 2) the prediction errors are relatively high than training errors, which means slightly over-fitting of our model; 3) nitrogen ammonia and phosphorus get relatively higher prediction accuracy, probably due to good stationarity of data sequence.

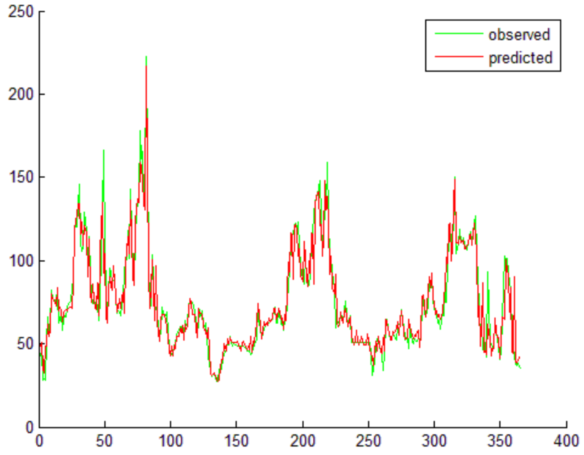
Table 4. Water quality prediction accuracy assessment

	model fitting			model testing		
	RMSE	MAE	MRE	RMSE	MAE	MRE
PH value	0.0879	0.0468	18.00%	0.0588	0.0313	32.89%
suspended solids	6.4893	3.3514	10.73%	12.3986	7.137	20.53%
COD	22.0881	11.7944	16.85%	26.6208	14.2558	10.62%
TOC	8.8051	4.8588	14.19%	18.2141	9.9829	31.30%
nitrogen ammonia	3.9304	2.1502	5.47%	1.8324	1.1844	11.34%
phosphorus	0.2283	0.1264	7.34%	0.4017	0.2641	7.78%

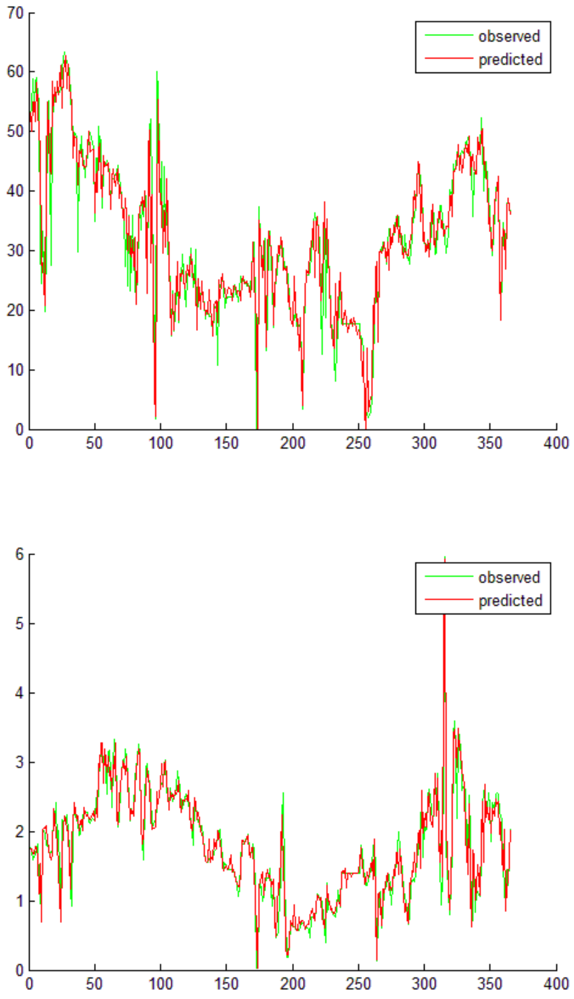

250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
observed
predicted
8.4
8.2
8
7.8
7.6
7.4
7.2
7
6.8
6.6
6.4
0
50
100
150
200
250
300
350
400
observed
predicted

(a) (b)


700
600
500
400
300
200
100
0
0
50
100
150
200
250
300
350
400
observed
predicted


250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
observed
predicted

(c) (d)



(e) (f)

Fig. 5 (a) ~ (f). Observed and predicted values of PH (a), suspended solids (b), COD (c), TOC(d), nitrogen ammonia (E), phosphorus (f) in SSTEC Jiyun River sewage factory outlet site in 2014

Conclusion

This paper gives a brief discussion about smart city development. Relying on the Sino-Singapore Tianjin Eco-city (SSTEC) data collecting platform and city pulsation project requirements, we proposed an environmental pulsation analysis framework in smart city, and talked about the three mainly steps: data features exploration, model selection and modeling and analysis. We introduced the wavelet enhanced ARMA model in detail, and used this model to conduct some prediction experiments on air quality data and water quality data in SSTEC. The result shows that, wavelet enhanced ARMA model is a proper method for environmental prediction. Our results provide some guidance for urban environmental management, also applicative in some other smart city application fields.

Reference

[1] Zheng Y., Capra L., Wolfson O., et al. Urban Computing, J. ACM Transactions on Intelligent Systems and Technology. 3(2014) 1-55.

[2] Zheng Y., Liu F., Hsieh H. U-Air: When urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. (2013) 1436-1444.

[3] Ayari S., Nouira K., Trabelsi A, A Hybrid ARIMA and Artificial Neural Networks model to forecast air quality in urban areas: case of Tunisia. Advanced Materials Research. 2012, pp.2969-2979.

[4] Jiaqiu Wang. Spatial-temporal sequence data analysis and modeling, Beijing, 2008.

[5] Zheng Gu, Baojin Chu, Huikun Jiang, Wavelet and ARMA combinatorial method and application of non-stationary time series analysis, J. System Engineering. 01(2010) 73-77.

[6] Fanqiang Meng, The application of ARIMA model in Air pollution index prediction, J. Statistic and decision. 07(2009) 33-35.

[7] Li Wang, Yuan Zhao, Xianming Yang, etc, The air quality research in Lanzhou city based on time series analysis model and residual control chart, J. Plateau Meteorology. 01 (2015) 230-236.

[8] Weimin Tong, Yijun Li, Yongzheng Dan, The time series data mining based on wavelet analysis, J. Computer Engineering. 01(2008) 26-29.

[9] Huicheng Zhou, Yong Peng, The monthly runoff prediction calibration model based on wavelet decomposition, J. JOURNAL OF SYSTEM SIMULATION. 05(2007) 1104-1108.

[10] Mallat S. G, A theory for multiresolution signal decomposition: the wavelet representation, J. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 7(1989) 674-693.

[11] Lingjuan Ding, J. The stock index prediction analysis based on wavelet analysis and ARMA-SVM model (Master Thesis). East China Normal University, Shanghai, China, 2012.