Influent property forecasting of sewage treatment based on big data analysis approach

Zhiwei Cheng¹*, Zhihong Xie¹, Yongsheng Deng¹ and Xiaodan Wang²

¹Information Engineering Institute, Chongqing Vocational and Technical University of Mechatronics, Chongqing 402760, China
²School of Big Data and Software Engineering, Chongqing College of Mobile Communication, Chongqing 401520, China
*Corresponding author’s e-mail: zhiweicheng_zc@163.com

Abstract. Precise influent property forecasting is very important to maintain the stable operation of sewage treatment procedure. A big data analysis method of combining the wavelet packet transform (WPT) and adaptive network-based fuzzy inference system (ANFIS) is reported to solve this problem. In this approach, the WPT is used to decompose the influent property data in different cycles. The time sub-series, which are results of wavelet coefficients reconstruction, are employed to establish the forecasting system. The forecasting sub-results of each cycle are eventually integrated into an overall forecasting result. Furthermore, chaos theory is introduced to obtain the input structure of the multi-cycle regression models. The reported approach is verified by the historical influent property. A back propagation neural network and the standard ANFIS are used for a comparison test. The results demonstrate that the reported method has best ability in the peer models.

1. Introduction

In the sewage treatment procedure, the state of the influent property directly affects the technological parameters, the usage of additives and the treatment efficiency. The influent property plays a significant role in stable running and sewage treatment energy management. Therefore, it is meaningful to accurately obtain the value of the influent property ahead for running optimization and operation efficiency. As a result, lots of methods based on big data analysis have been proposed in the literatures. Various time series models have been developed to forecast influent property. Models used parameters describing the physical and biological influent sewage properties as the forecasting target, and achieved good results. However, the forecasting performance of the model is decreased when data are not distributed normally or nonlinear [1]. Artificial neural network (ANN), which can forecast the influent property by building the nonlinear relationship of input and output without considering the basic theoretical mechanism, are an efficient alternative approach for addressing the nonlinear, dynamic, and complex problems [2, 3]. However, ANN have the obvious drawback in terms of covering over training, easy to trap in local minima, hard to find the appropriate model parameters, initializing the weights randomly and complex structure. The ANFIS, which has the advantages of both the fuzzy inference system and the ANN [4], is an effective means to address the regression issue. Compared with the ANN, the ANFIS has adaptability, robustness, and better self-learning capabilities in regression issue. Because of its superior performance, the ANFIS is then introduced to forecast influent property related to sewage treatment process [5]. Moreover, the feature engineering in modeling process have a critical effect on
the forecasting performance. The feature may provide more valuable information in multiple cycle so that models can learn more deeply. To improve the forecasting performance, the WPT method is employed to obtain more information for helping improving the ability of the prediction model. Having the mentioned constituents, one proposed the multi-cycle fuzzy system (MCFS) based on the WPT and the ANFIS. It was reported to forecast the influent property of the sewage treatment plant. In addition, the proposed approach used the chaos property analysis to determine the input structure of the multi-cycle regression model.

2. Methodologies

2.1 Influent property decomposition

In this work, wavelet packet transform (WPT) is employed to decompose the influent property. WPT, which is the further optimization of the wavelet transform, has the characteristics of multi-scale decomposition and multi-band analysis, and provide a more convenient method for signal processing [6]. Decomposing a discrete signal \( D(i) \) is expressed as

\[
C_{0,0} = D(i), \quad C_{p,n}(i) = \sum_{k} H_{k-2^{p}} C_{p,n}(k), \quad C_{p+1,2n}(i) = \sum_{k} G_{k-2^{p}} C_{p,n}(k),
\]

(1)

where \( k \) represents the number of wavelet coefficients, and \( C_{p,n}(i) \) is the wavelet coefficients in the \( p \)-th level and \( n \)-th sub-frequency band. Just as mentioned above, the reconstruction algorithm is denoted as

\[
C_{p,n}(i) = \sum_{k} l_{i-2^{p}} C_{p+1,2n}(k) + \sum_{k} e_{i=2^{p}} C_{p+1,2n}(k),
\]

(2)

where \( j_{i-2^{p}} \) and \( e_{i-2^{p}} \) are the reconstruction filters.

The corresponding wavelet coefficients is utilized to obtain signals \( y_{0}(i), y_{1}(i), \ldots, y_{2^{p}-1}(i) \) of each sub-frequency band. \( D(i) \) is established with the above-mentioned time sub-series by

\[
D(i) = y_{0}(i) + y_{1}(i) + \ldots + y_{2^{p}-1}(i).
\]

(3)

What is more, the ‘db5’, as the widely used mother wavelet, is employed to transform the original signals.

2.2 Input structure determination

After the above processing, the time series of different cycles have been obtained. The chaos property analysis is introduced to provide the theoretical basis for the time series forecasting. The phase space reconstruction (PSR) technology and largest Lyapunov exponent are utilized to analyze the chaotic property of the time sub-series. Once the chaos property is confirmed, the input structure of regression model is also determined.

The PSR, proposed by Packard et al. [8] and Takens [9], is a very important step in chaotic time series processing. The time sub-series \( y_{i} (i = 1, 2, \ldots, N) \) can be transformed to

\[
Y_{t} = \{y_{i}, y_{i+\tau}, y_{i+2\tau}, \ldots, y_{i+(m-1)\tau}\},
\]

(4)

where \( \tau \) represents the delay time and \( m \) represents the embedding dimension. \( \tau \) is first computed by the autocorrelation approach. Subsequently, the optimal \( m \) is obtained with the false nearest neighbors (FNN) based on the calculated \( \tau \).

Based on the reconstructed phase space, the largest Lyapunov exponent \( l_{max} \) is computed to identify the chaos property. According to the used data characteristics, the small data sets method [10] is employed to compute the \( l_{max} \), expressed as

\[
l_{max} = \frac{1}{s_{M-2}^{2} \sum_{K=1}^{M} \log_{2} \frac{L'(S_{K})}{L(S_{K-1})}}
\]

(5)

where \( L(S_{K}) \) represents the distance between the \( Y_{t} \) and its interest neighbor, and \( L'(S_{K}) \) is the evolved length of \( L(S_{K}) \) at the time \( S_{K} \). \( M \) denotes the replacement steps number. If \( l_{max} > 0 \), the sub-series of the influent property has a chaos property.

2.3 ANFIS for forecasting

Having analyzed the chaos property, adaptive network-based fuzzy inference system is adopted to forecast the influent property. Developed by Jang, the ANFIS is a combination of ANN and fuzzy logic [11]. To learn the training data and generate the membership function of the input and output, the
The learning mechanism of the ANN is introduced into the fuzzy system. In this work, the forecasting problem of the time sub-series $y_i$ is expressed as

$$y_{i+1} = F[y_i, y_{i-\tau}, \ldots, y_{i-(m-1)\tau}],$$

(6)

where $F$ represents ANFIS regression model, $y_{i+1}$ denotes the value at $i+1$.

### 2.4. Overview of the MCFS

As previously stated, the implementation procedure of the MCFS is showed in figure 1.

The main procedure of the MCFS is summarized as follows:

1): Use the WPT to decompose the collected influent property data.

2): Multi-cycle time sub-series are produced by branch reconstruction method.

3): Perform the chaos analysis method to calculate the input structure of regression in each cycle.

4): Built the multi-cycle ANFIS for the sub-series regression.

5): The forecasting sub-results are correspondingly integrated into the result.

### 3. Experiments

The BOD, as a representation of the influent property, is collected from one STP during 2017 and is displayed in figure 2. The STP treats domestic sewage with the capacity of 45,000 cubic meters per day. The STP used the traditional Anoxic/Oxic technology with low operating cost, easy operation process and high efficiency. In order to evaluate the performance of the comparison models, the mean absolute error (MAE), root mean square error (RMSE) mean absolute percentage error (MAPE) are introduced in this study [12, 13].

As mentioned in Section 2, The WPT uses the ‘db5’ and the level 2 decomposition. The obtained wavelet coefficients are rebuilt to be the sub-series of the influent property in different cycles. As already stated, the delay time $\tau$ and the optimal embedding dimension $m$ is obtained with the autocorrelation method and the FNN approach. The largest Lyapunov exponents are then used to recognize the chaos property after the phase space reestablishment of the sub-series for the influent property. Subsequently, the small data sets method is used to computed the largest Lyapunov exponents. Positive exponents demonstrate that the influent series sub-series have the chaos property. The chaotic properties of the other sub-results are calculated, as displayed in table 1.

#### Table 1. Results of chaos analysis for influent property sub-series.

| Sub-series | $\tau$ | $m$ | $l_{max}$ |
|------------|-------|-----|---------|
| $Y_1$      | 1     | 4   | 0.2126  |
| $Y_2$      | 1     | 3   | 0.1832  |
| $Y_3$      | 1     | 4   | 0.1968  |
| $Y_4$      | 1     | 5   | 0.2117  |
The conclusion is drawn that all the sub-series have the chaos properties. As a result, the PSR technology can be employed to obtain the input structure of the regression models. Furthermore, the ANFIS and the BPNN and are adopted to validate the performance of the proposed approach. The peer models used the same data. The forecasting results using different models are displayed in figure 2, figure 3 and figure 4.

Fig. 2 MCFS forecasting results for BOD: (a) is the fitting curves between the forecasting values and the observed values; (b) is correlogram analysis on the forecasting values and the observed values.

Fig. 3 BPNN forecasting results for BOD: (a) is the fitting curves between the forecasting values and the observed values; (b) is correlogram analysis on the forecasting values and the observed values.

Fig. 4 ANFIS forecasting results for BOD: (a) is the fitting curves between the forecasting values and the observed values; (b) is correlogram analysis on the forecasting values and the observed values.
Figure 2 (a) and figure 2 (b) reveal that the forecasting trendline can clearly follow the variations of the observed trendline. It is obvious that there is little difference between the forecasting values and the observed values, and the MCFS has a good fitting ability. Figure 3 (a) shows that the BPNN has a worse forecasting performance. It can only sketchily trace the variations of observed trendline. Specifically, the gap between each value is relatively large. Figure 4 (a) shows that the ANFIS can keep track of the primary periodic change of the forecasting trendline, and could follow the general trend. It has better forecasting performance than the BPNN. Whereas, it can’t be ignored that the forecasting values of the ANFIS cannot accurately consistent with the observed values. Obviously, the information obtained the ANFIS is not inadequate due to the raw data is not processed. Through the above, the MCFS has the superior forecasting ability both in the detailed and the overall trend.

Moreover, the scatter analysis displayed in figure 2 (b), figure 3 (b), and figure 4 (b). The scatter of the MCFS model is more concentrated on the ideal line compared the BPNN and the ANFIS. In conclusion, the forecasting performance of these models are ranked from low to high as BPNN, ANFIS, and MCFS. It suggests that the performance of the MCFS is significantly improved due to the model can obtain more information representations of the influent property. Furthermore, the quantitative comparison of the forecasting performance for the three models are illustrated in Table 2.

| Model  | MAE   | MAPE (%) | RMSE  |
|--------|-------|----------|-------|
| BPNN   | 4.123 | 5.788    | 5.541 |
| ANFIS  | 4.012 | 5.291    | 5.126 |
| MCFS   | 1.298 | 1.789    | 1.802 |

Table 2 reveals that the MCFS has the best performance compared with the comparison models in term of the used criteria. The differences in the MAE, MAPE and RMSE between the MCFS and the BPNN are 68.52%, 69.09%, and 67.48%, respectively, and between the MCFS and the ANFIS are 67.65%, 66.18%, and 64.85%, respectively. The difference of these criteria indicates that the model could obtain deeper influent property representations and reduce the interference of random noise.

4. Conclusions
This work developed the MCFS method combining the WPT for multi-cycle representation and the ANFIS for regression problem to precisely forecast the influent property of the STP. The reported MCFS approach is applied to the influent property data from one STP. The experimental results indicate that the MCFS has a superior forecasting ability over the peer models, and has the better forecasting performance.

Acknowledgements
This work was supported by the National High Skilled Talents Training Base (Big Data Technology and Application) of Chongqing Vocational and Technical University of Mechatronics, and the Chongqing Higher Education Teaching Reform Research Project (203717).

References
[1] Bai, Y., Wang, P., Li, C., Xie, J., and Wang, Y. (2015) Dynamic Forecast of Daily Urban Water Consumption Using a Variable-structure Support Vector Regression Model. J. Water Resour. Plann. Manage, 141(3).
[2] Bai, Y., Li, Y., Wang, X., Xie, J., and Li, C. (2016) Air Pollutants Concentrations Forecasting Using Back Propagation Neural Network Based on Wavelet Decomposition with Meteorological Conditions, Atmos. Pollut. Res., 7(3): 557-566.
[3] Khaki, M., Yusoff, I., and Islami, N. (2015) Application of the Artificial Neural Network and Neuro-fuzzy System for Assessment of Groundwater Quality. Clean - Soil Air Water, 43(4): 551-560.
[4] Jang, J.-S.R. (1993) Anfis: adaptive-network-based fuzzy inference system. IEEE Trans on Smc., 23(3): 665-685.

[5] Gaya, M.S., Wahab, N.A., Sam, Y.M., Anuar, A.N., and Samsuddin, S.I. (2013) Anfis modelling of carbon removal in domestic wastewater treatment plant. Applied Mechanics & Materials, 372: 597-601.

[6] Mallat, S.G. (1989) A Theory of Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Trans. pattern Anal. machine Intell., 11(7): 674-693.

[7] Coifman, R.R., and Wickerhauser, M.V. (1992) Entropy-based algorithms for best basis selection. IEEE Transactions on Information Theory, 38(2): 713-718.

[8] Packard, N.H., Crutchfield, J.P., and Shaw, R. S. (2008) Geometry from a time series. Physical Review Letters, 45(9):712-716.

[9] Takens, F. (1981) Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, 366-381

[10] Rosenstein, M.T., Collins, J.J., and Luca, C. J.D. (1993) A practical method for calculating largest Lyapunov exponents from small data sets. Physica D, 65(1-2): 117-134.

[11] Jang, J.S.R, Sun, C.T, and Mizutani, E. (1997) Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Automatic Control IEEE Transactions on, 42(10): 1482-1484.

[12] Li, X.J., Cheng, Z.W., Yu, Q.B., Bai, Y., and Li, C. (2017) Water-quality prediction using multimodal support vector regression: case study of Jialing River, China. Journal of Environmental Engineering, 143(10): 04017070.

[13] Cheng Z, Li X, Bai Y, et al. (2018) Multi-scale fuzzy inference system for influent characteristic prediction of wastewater treatment. Clean Soil Air Water, 46(7):1700343.