Desalination and Water Treatment
www.deswater.com
doi: 10.1080/19443994.2014.986202

57 (2016) 3452–3465
February

Taylor & Francis
Taylor & Francis Group

Seasonal artificial neural network model for water quality prediction via a clustering analysis method in a wastewater treatment plant of China

Ying Zhao[a,b,*], Liang Guo[a,b,*], Junbo Liang[c], Min Zhang[d]

[a]*School of Municipal and Environmental Engineering, Harbin Institute of Technology, 150090 Harbin, China, Tel. +86 045186283109; emails: zhaoying@hit.edu.cn (Y. Zhao), guoliang0617@hit.eud.cn (L. Guo)*
[b]*State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, 150090 Harbin, China*
[c]*School of Astronautics, Harbin Institute of Technology, 150090 Harbin, China, Tel. +86 18545001274; email: 1219493517@qq.com*
[d]*College of Civil and Architectural Engineering, Heilongjiang Institute of Technology, 150050 Harbin, China, Tel. +86 18045105095; email: 14194418@qq.com*

Received 19 December 2013; Accepted 4 November 2014

ABSTRACT

For recovering the water quality of a river, it is a key factor to improve purifying capacity of wastewater in wastewater treatment plants (WTPs). The relational model for some key parameters of WTP processes is important for it can reveal the current situation and handling ability of the WTP and offer managers more useful information to design the processes for the optimized operation. The seasonal artificial neural network (ANN) models were designed for improving purifying ability of wastewater in a WTP of Harbin, northeast of China. The ANN models revealed the relationship of raw water quality, energy consumption, and effluent water quality. The effluent water quality could be predicted by the models. The clustering analysis method, an important data mining method, was used to classify the WTP data for building seasonal models. Meanwhile, an annual model was built by the whole data. It indicates that the prediction accuracy of seasonal models is better than the annual model by contrasting the errors. Seasonal models should be a more effective tool to reveal the relationship of WTP data. So it can offer managers more precise information to control and design the processes of WTPs, which result in better purifying ability of wastewater.

Keywords: ANN model; Clustering analysis; Wastewater treatment plant; Water quality prediction; Optimization; Seasonal model

1. Introduction

Surface water pollution has been a serious problem due to rapid industrial development, population growth, and urbanization in the last decades in China [1]. It is necessary to set up more wastewater treatment plants (WTPs) to protect surface water. However, many WTPs are usually inefficient to handle wastewater. Though the number and scale of Chinese WTPs are expanding in recent years [2], rivers have been polluted by point and non-point pollution sources due to inefficient operation of WTPs, which has resulted in ecological destruction [3,4].

For recovering the water quality of river, it is a key factor to improve the purifying capacity of wastewater

*Corresponding authors.

in WTPs. For improving the purifying capacity, managers must grasp and accurately design all the processes of WTPs and make them optimally operate. The relational model for some key parameters of the processes is important for optimized operation. It can reveal the current situation and handling ability of WTPs and offer managers more useful information to design the processes for the optimized operation. Contrasting with the data of other fields, such as finance and business, the characteristics of WTPs data are [5]: (1) larger data; (2) more complicated and nonlinear relationship; and (3) polytrope and variety. It is difficult to use static mathematic models based on mechanism analysis to handle the WTPs data with such characteristics. For stronger learning ability and fault tolerance, artificial neural network (ANN) is well done in handling data with complicated and nonlinear relationships. Furthermore, ANN integrated with data mining methods can be used to address more complicated problems.

In recent years, ANN has been used to design and optimize the processes of WTPs. The simulation of external carbon addition was studied by a back propagation neural network (BPNN) model in the continuous flow anoxic/oxic (A/O) nitrogen removal process for domestic wastewater with low carbon nitrogen (C/N) [6]. An integrated neural-fuzzy process controller was developed to control aeration in an Aerated Submerged Biofilm Wastewater Treatment Process [7]. An adaptive fuzzy neural network controller was proposed to realize the control of Dissolved Oxygen (DO) in the activated sludge model, and to adjust the measured factor so as to reduce static error [8]. On-line pH and oxidation–reduction potential monitoring and ANN models were applied to dynamically control the wastewater chlorination and dechlorination dosages for reuse purposes [8]. An ANN modeling was applied to predict the performance of a laboratory-scale batch-fed reactor in terms of COD removal and TKN reduction from real-life slaughter house wastewater as output parameters in correspondence to various input functions of initial concentrations, microbial concentration, contact time, pH, DO, etc. [9]. An ANN was applied for prediction of performances in competitive adsorption of phenol and resorcinol from aqueous solution by conventional and low-cost carbonaceous adsorbent materials [10]. These studies indicated that ANN was a robust tool to design and optimize many key processes of WTPs.

ANN was also used to predict water quality and other parameters of WTPs. An ANN model was developed to estimate daily BOD in the inlet of wastewater biochemical treatment plants [11]. An ANN approach and a software sensor were proposed for the real-time estimation of nutrient concentrations and overcoming the problem of delayed measurements. In order to improve the neural network performance, a split network structure applied separately for anaerobic and aerobic conditions was employed with dynamic modeling methods such as auto-regressive with exogenous inputs [12]. The BP ANN model was used to predict the effluent stream quality at a WTP [13]. Three kinds of ANN models were contrasted to get the best model to predict the effluent COD concentration of the WTP [14]. A powerful aeration energy consumption monitor model was set up by BPNN in the WTP [15]. These researches indicated that ANN models could predict water quality and other parameters of the WTPs.

In these studies, though ANN model was a robust tool to predict water quality and other parameters of WTPs, prediction accuracies of the ANN models were still not satisfied. Furthermore, the ANN models were trained directly by the data of whole year (annual model) even if there were obvious changes in the data of different seasons. So they should be called the annual models. In fact, detailed seasonal strategies are necessary for wastewater treatment because many characteristics of wastewater always change seasonally. Seasonal ANN models can offer managers more detailed management strategies in each period resulting in better purifying capacity of wastewater in WTPs. Seasonal models are trained by corresponding seasonal data, which are relatively concentrated and less fluctuant resulting in small uncertainty. The model with small uncertainty usually has a good structure and prediction performance. So the prediction accuracy of seasonal model should be better than the annual model. However, seasonal ANN modeling has seldom been used in WTPs. Data mining, an interdisciplinary subfield of computer science [16,17], is the computational process of discovering patterns in large data-sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data-set and transform it into an understandable structure for further use [16]. It can help us to understand many scientific problems by preprocessing and analysis of raw data. Data mining is also seldom used in optimizing in ANN models.

In this paper, seasonal ANN models were built for improving purifying ability of wastewater in a WTP of Harbin, northeast of China. The ANN models revealed the relationship of raw water quality, energy consumption, and effluent water quality. The effluent water quality could be predicted by the models. Furthermore, data mining method was applied to improve prediction accuracy of the models. All the

data were analyzed and classified by clustering analysis method, an important data mining method. Firstly, an annual model was built by the whole data. Secondly, the whole data was classified into several classes by clustering analysis method based on seasonal characteristics of the data. Seasonal models were built, respectively, by the classified data-sets. The prediction accuracies of the seasonal model and annual model were contrasted to verify the superiority of the seasonal model. Seasonal ANN model should be a useful tool for improving purifying ability of wastewater because effective relational model was used to predict effluent water quality and control energy consumption.

2. Materials and methods

2.1. Data analysis

The studied data were collected from a WTP in Harbin City, Heilongjiang Province, China from 2009 to 2011. The consumption of electricity and reagent reflect the cost of purifying wastewater. Therefore, besides raw wastewater and effluent water, energy consumption (electricity and reagent) was studied. The relational model of these data can help managers to operate the WTP optimally based on energy-saving mechanism. The WTP is located in northeast China and local climate has distinct seasonal variations. The difference of average water temperature between summer and winter exceeds 30˚C, so raw wastewater and other parameters of the WTP have obvious seasonal characteristics. For exploring seasonal characteristics of the data, two ways were discussed: (1) the annual difference among three years; and (2) the seasonal difference in one year. Three statistical analysis methods, including range, mean, and standard deviation, were used to identify the differences of data among months or years. In statistics and probability theory, the standard deviation (SD) (represented by the Greek letter sigma, σ) measures the amount of variation or dispersion from the average [18].

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{1}$$

where

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value); a high standard deviation indicates that the data points are spread out over a large range of values.

2.2. Clustering analysis method

2.2.1. Clustering analysis

Clustering analysis method is an important method of data mining [19–22]. Variables are classified based on their similarity by hierarchical cluster method. The weights of correlation among the parameters are obtained by the clustering procedures. In clustering analysis method, grouping objects (cases) are divided into several classes (clustering) so that the objects exhibit high internal (within-clustering) homogeneity and high external (between clustering) heterogeneity. The objects are grouped by linking intersample similarities and the outcome illustrates the overall similarity of variables in the data-set [23]. Hierarchical agglomerative clustering analysis was carried out on the normalized data by Ward's method, an extremely powerful grouping mechanism, which yields a larger proportion of correct classified observations [24], using squared Euclidean distances as a measure of similarity [23–26]. The distance is computed as:

$$\mathrm{D}(x, y) = \sum_{i} (x_i - y_i)^2 \tag{3}$$

where D is the distances between x and y, x and y are two variables in the variable space, and i is the ith component of x or y.

2.2.2. Classifying WTP data based on clustering analysis

For building seasonal model, the WTP data must be classified based on seasonal characteristics. Clustering analysis method was used to classify the data, and the results are an important evidence to divide each period. The data, including raw water quality, effluent water quality, and energy consumption parameters, were taken as clustering objects. Since the differences of data among months were studied by statistical analysis methods, the monthly raw data of three years were taken as initial classes. The clustering analysis was carried out by the software of SPSS. The clustering results, including clustering schedule and clustering dendrogram, were given to show the clustering

Table 1
Statistical result of the WTP data from 2009 to 2011

Month	Value	Raw water quality data						
		Raw water TP (mg/l)	Raw water NH_3-N (mg/l)	Raw water BOD (mg/l)	Raw water COD (mg/l)	Raw water SS (mg/l)	Raw water PH	Raw water temperature (mg/l)
January	Range	1.6–4.4	30.5–58.9	112.0–302.0	251.0–559.0	112.0–659.0	7.0–7.4	7.0–11.0
	Mean	2.5	43.3	186.1	365.4	235.5	7.3	9.7
	SD	1.2	5.1	24.3	49.7	83.7	0.1	1
February	Range	1.7–4.4	24.9–58.6	137.0–238.0	212.0–466.0	108.0–399.0	7.2–7.4	8.0–12.0
	Mean	2.5	41.5	183.9	358.6	209.9	7.3	9.9
	SD	1.2	6.5	19.9	46.1	75.4	0	1.2
March	Range	1.7–4.3	27.9–55.7	106.0–235.0	217.0–492.0	103.0–360.0	6.4–7.4	6.4–13.0
	Mean	2.5	40.3	187.3	354.2	218.1	7.2	10.3
	SD	1.2	6	18.7	49.8	65.5	0.2	1.5
April	Range	1.6–4.3	20.8–56.7	94.0–241.0	178.0–469.0	88.0–356.0	7.3–7.5	7.0–22.0
	Mean	2.5	38.2	183.2	340.4	223.7	7.4	11.6
	SD	1.2	6.6	25.6	61.9	57.7	0	2.1
May	Range	1.6–4.3	11.6–51.4	130.0–210.0	249.0–401.0	101.0–323.0	7.2–7.4	10.0–22.0
	Mean	2.5	34.9	183.5	344	222.8	7.4	13.9
	SD	1.2	9.4	12.8	41.9	53.6	0	2.4
June	Range	1.6–4.3	18.7–47.3	78.0–228.0	156.0–476.0	103.0–388.0	7.2–7.5	9.0–24.0
	Mean	2.5	32.7	164.5	324.2	223.6	7.3	16.7
	SD	1.2	7.1	41.8	89.9	64.4	0.1	2.9
July	Range	1.6–4.3	12.3–49.1	61.0–218.0	113.0–436.0	82.0–378.0	7.2–7.4	15.0–27.0
	Mean	2.5	28.8	146.9	291.5	211.8	7.3	20.5
	SD	1.2	8.7	52.8	105.9	73.3	0	2.7
August	Range	1.6–4.3	12.6–41.6	74.0–201.0	137.0–413.0	111.0–379.0	7.2–7.4	16.0–28.0
	Mean	2.5	28.6	140.6	281.3	209.4	7.3	20.6
	SD	1.2	6.5	36.7	80.4	61	0	2.8
September	Range	1.6–4.4	17.4–49.4	73.0–194.0	131.0–370.0	114.0–296.0	7.3–7.4	12.0–23.0
	Mean	2.5	31.1	134	257	192.4	7.3	17.8
	SD	1.2	8.4	30.8	55.7	50.8	0	2.4
October	Range	1.6–4.4	18.8–47.1	74.0–203.0	128.0–357.0	102.0–393.0	7.2–7.5	10.0–19.0
	Mean	2.5	32.9	139	258.9	208.1	7.3	13.5
	SD	1.2	7.9	35.5	63	64.4	0	2.2
November	Range	1.6–4.5	19.4–48.2	83.0–251.0	168.0–471.0	96.0–339.0	7.5–8.0	9.0–14.0
	Mean	2.6	34.5	155.8	282.7	221.9	7.8	10.9
	SD	1.3	5.8	30.4	53.3	54.9	0.1	1.1
December	Range	1.6–4.4	22.7–53.1	103.0–216.0	207.0–389.0	104.0–314.0	7.6–7.8	8.0–13.0
	Mean	3.4	40.2	167.9	326.7	216.9	7.7	10.2
	SD	1.2	6.2	23	44.7	51.9	0.1	0.8

Table 1
(*Continued*)

Month	Value	Effluent water quality data				
		Water purification TP (mg/l)	Water purification NH_3-N (mg/l)	Water purification BOD (mg/l)	Water purification COD (mg/l)	Water purification SS (mg/l)
January	Range	1.4–1.5	0.1–7.8	13.0–19.0	41.0–80.0	14.0–20.0
	Mean	1.4	4.8	16	51.3	17.3
	SD	0	2.1	1.1	6.2	1.5
February	Range	1.4–1.5	0.2–7.7	14.0–18.0	43.0–60.0	14.0–20.0
	Mean	1.4	4.4	16.2	50.6	17.1
	SD	0	2.4	1	5.1	1.3
March	Range	1.4–1.5	0.2–7.7	12.0–20.0	40.0–60.0	12.0–20.0
	Mean	1.4	4	16.4	50.6	16.8
	SD	0	2.4	1.3	5	1.6
April	Range	1.4–1.5	0.3–7.8	11.0–18.0	42.0–58.0	12.0–18.0
	Mean	1.4	4.1	15.5	49.6	16
	SD	0	2.2	1.5	3.6	1.6
May	Range	1.4–1.5	0.6–7.6	11.0–17.0	42.0–58.0	12.0–18.0
	Mean	1.4	4.2	15.1	49	16.2
	SD	0	2.2	1.4	3.2	1.3
June	Range	1.4–1.4	0.8–7.6	12.0–18.0	4.9–57.0	11.0–18.0
	Mean	1.4	4.2	14.9	49.5	15.5
	SD	0	2	1.6	5.7	1.4
July	Range	1.4–1.4	0.7–6.6	11.0–19.0	36.0–53.0	11.0–18.0
	Mean	1.4	3.7	15.1	46.7	15
	SD	0	1.6	1.8	3.4	1.7
August	Range	1.3–1.5	0.8–4.8	12.0–18.0	37.0–55.0	11.0–18.0
	Mean	1.4	2.9	14.9	47.3	14.4
	SD	0	0.9	1.7	3.7	1.6
September	Range	1.3–1.5	1.3–6.8	13.0–18.0	41.0–57.0	11.0–18.0
	Mean	1.4	3.7	15.6	49.5	15.3
	SD	0	1.1	1.4	3.3	1.4
October	Range	1.3–1.5	1.8–7.8	12.0–18.0	40.0–59.0	9.0–18.0
	Mean	1.4	4.2	15.9	50	15.4
	SD	0	1.5	1.3	3.1	1.7
November	Range	1.4–1.5	0.9–7.5	12.0–18.0	42.0–57.0	11.0–18.0
	Mean	1.4	4.7	16	49.3	16
	SD	0	1.7	1.3	2.9	1.7
December	Range	1.4–1.4	2.7–7.8	13.0–18.0	40.0–55.0	12.0–19.0
	Mean	1.4	5.4	16.2	48	15.9
	SD	0	1.6	1.2	2.4	1.6

Table 1
(*Continued*)

Month	Value	Energy consumption data		
		Electricity consumption (kwh/m^3)	The amount of coagulant (kg/km^3)	The dosage of flocculants (kg/km^3)
January	Range	0.2–67.2	23.5–92.5	0.5–2.2
	Mean	1	61.3	1.7
	SD	7	25.8	0.3
February	Range	0.2–0.4	23.6–87.7	1.1–2.2
	Mean	0.3	59.8	1.7
	SD	0.1	26	0.3
March	Range	0.2–0.4	24.0–83.3	1.4–2.2
	Mean	0.3	60.4	1.7
	SD	0	25.8	0.3
April	Range	0.2–0.4	25.0–113.6	1.0–2.7
	Mean	0.3	62.8	1.6
	SD	0	25.4	0.2
May	Range	0.2–0.4	64.2–122.9	0.3–2.6
	Mean	0.3	86.8	1.6
	SD	0	12.7	0.4
June	Range	0.2–0.4	30.6–96.1	0.7–2.4
	Mean	0.3	79.8	1.6
	SD	0	11.6	0.3
July	Range	0.2–0.4	27.3–142.9	0.8–2.9
	Mean	0.3	76.5	1.6
	SD	0	18	0.3
August	Range	0.2–0.4	27.2–84.0	0.5–2.5
	Mean	0.3	57.7	1.5
	SD	0	20.5	0.3
September	Range	0.2–0.4	25.7–89.4	0.8–2.7
	Mean	0.3	58.8	1.7
	SD	0	20.5	0.3
October	Range	0.2–0.4	10.4–86.8	1.5–4.4
	Mean	0.3	53	2.6
	SD	0	30.3	1.2
November	Range	0.2–0.4	10.4–82.8	0.4–5.4
	Mean	0.3	54.5	2.8
	SD	0	30.6	1.5
December	Range	0.2–0.3	10.1–98.6	0.5–5.6
	Mean	0.3	55.3	2.5
	SD	0	31.4	1.3

sequence and homogeneous classes. The data in homogeneous classes were used to build seasonal ANN models.

2.3. ANN modeling

In computer science and related fields, ANNs are computational models inspired by an animal's central nervous systems, in particular, the brain which is capable of machine learning as well as pattern recognition. ANNs are generally presented as systems of interconnected neurons which can compute values from inputs. ANN models are built by the way of learning from data. ANNs have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming. Therefore, the networks are applicable to a dynamic and unstable wastewater treatment system. ANN models were built to predict effluent water quality of a WTP by MATLAB.

2.3.1. BPNN and learning algorithm

The back propagation (BP) is the most widely used algorithm in ANN. BP, an abbreviation for "backward propagation of errors," is a common method of training ANNs used in conjunction with an optimization method such as gradient descent. The method calculates the gradient of a loss function with respect to all the weights in the network. The gradient is fed to the optimization method which in turn uses it to update the weights, in an attempt to minimize the loss function [27]. The convergence of BP with the common algorithm is linear and slow during training. In order to improve the disadvantages and decrease error, some algorithms have been proposed and applied [28]. Among the improving algorithms, the adaptive learning rate algorithm brings a better accuracy and can accelerate convergence of the model. So BP model and adaptive learning rate algorithm were used to build the ANN models.

2.3.2. Input and output variables

In the ANN models, the output of the model is the effluent water quality of the WTP and the inputs are raw water quality and energy consumption parameters. According to the comprehensive analysis of available data, the outputs were identified as Total Phosphorus (TP), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Suspended Solids (SS), and NH_3-N of the effluent water. The input variables included seven raw wastewater and three energy consumption parameters. Raw wastewater parameters were TP, BOD, COD, SS, NH_3-N, pH, and SS, and energy consumption parameters were electricity consumption, coagulant, and flocculants. Early stopping method was used to improve generalization ability of the network, so the data were divided into a training set, a validation set, and a test set.

2.3.3. ANN parameter selection

The basic structure of an ANN model is usually comprised of three distinctive layers, the input layer, where the data are introduced to the model and computation of the weighted sum of the input is performed, the hidden layer or layers, where data are processed, and the output layer, where the results of

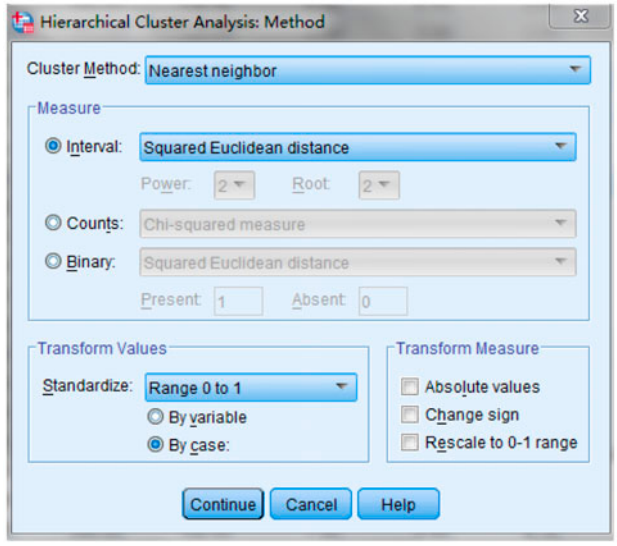

Fig. 1. The parameter design of clustering analysis by SPSS.

Table 2
Clustering schedule of the raw WTP data in different months

Initial classes (Month)	Clustering classes
Case 1	1
Case 2	1
Case 3	1
Case 4	1
Case 5	2
Case 6	1
Case 7	3
Case 8	3
Case 9	3
Case 10	4
Case 11	4
Case 12	1

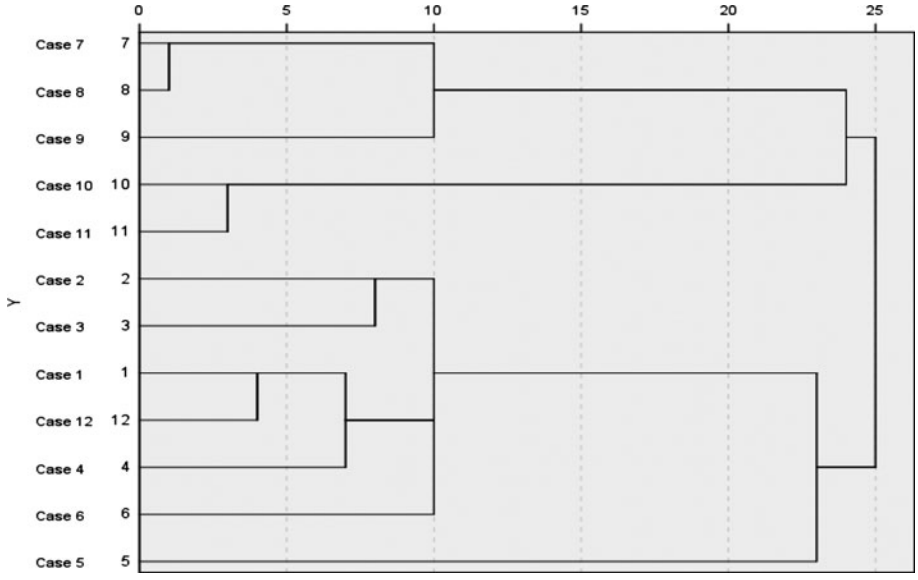

Fig. 2. Clustering dendrogram of the raw WTP data in different months.

ANN are produced [28]. Complex neural network forecast model is susceptible to bad training and might harm the performance of the network. It has been proved that BP neural network model with a single-hidden layer could approach any nonlinear functions with limited points of discontinuity with any accuracy, if there were enough neurons in the hidden layer [29]. Therefore, the BP model with a hidden layer was chosen in this study. Although the literatures report several methods for determining the number of neurons in a hidden layer, they are purely empirical and cannot be generalized. Thus, the number of neurons in the hidden layer was chosen from 11 to 20 by summarizing a lot of literatures [30–32], and hyperbolic tansig or logsig sigmoid functions were chosen as the neuron transfer function in the hidden layer. Trial and error method is used to find the best structure of the model.

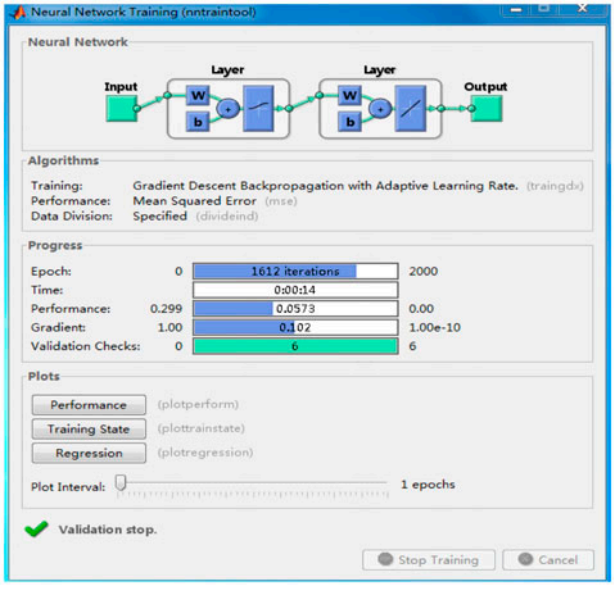

Fig. 3. The structure drawing and parameter design of a BP network model.

The transfer function in the last layer can greatly influence the output characteristic of the whole neural network. Sigmoid function in the output layer requires the estimated output converted back to the real world using the same sigmoid function. However, linear function estimates the output in the range from negative infinity to positive infinity, which avoids remapping of the outputs [33–35]. Therefore, a linear function was chosen as the transfer function for the neurons in the output layer.

2.3.4. The improvement of the network generalization ability

The generalization ability is an important factor to evaluate the performance of an ANN model [36]. Early

Table 3
Classified results of raw WTP data

Class	Month
First	January–April, December
Second	May–June
Third	July–September
Fourth	October–November

Table 4
Training results of the annual ANN model

The neuron transfer function in the hidden layer	The number of neurons in the hidden layer	Training error	Validation error	The actual times of the training
TANSIG	10	0.0528	0.050664	1,656
	11	0.0536	0.048032	1,633
	12	0.0527	0.051408	1,627
	13	0.0532	0.048065	1,627
	14	0.0534	0.052703	1,590
	15	0.0529	0.051611	1,589
	16	0.0535	0.048603	1,607
	17	0.052	0.052232	1,606
	18	0.053	0.052936	1,561
	19	**0.0513**	**0.051705**	**1,580**
	20	0.0539	0.049342	1,539
LOGSIG	10	0.114	0.10965	235
	11	0.0534	0.05798	1,797
	12	0.12	0.11498	174
	13	0.112	0.10659	167
	14	0.115	0.11595	169
	15	0.0549	0.057379	1,727
	16	0.115	0.11275	175
	17	0.055	0.051032	1,725
	18	0.0548	0.052379	1,816
	19	0.0563	0.051934	1,732
	20	0.113	0.10322	169

stopping method was used to improve the generalization ability of the network. The data were divided into a training set, a validation set, and a test set in the early stopping method. The training set was used to train the neural network. The validation set was used to supervise the error of the training set. In the beginning of training, the validation error of the validation set usually decreased with the drop in training error. But when the network was excessively trained, the validation error would increase gradually even if the training error was decreasing. Network training would be stopped at this time, and the network with the minimum validation error was the best model. Finally, the test set was used to validate the model.

3. Results and discussion

3.1. Data analysis result

The raw monthly data in the WTP were analyzed before the application for clustering analysis method and modeling, using range, mean, and standard deviation. The analysis results showed little difference among years for each parameter. And there were obvious differences among seasonal periods for most of the parameters. The monthly WTP data are given in Table 1, including raw water quality, effluent water quality, and energy consumption parameters. Table 1 showed that the monthly data in one seasonal period were similar, which indicated that seasonal ANN

Table 5
The optimized seasonal ANN models

Seasonal ANN model[a]	The number of neurons in the hidden layer	The neuron transfer function in the hidden layer	Training error	Validation error
1	19	TANSIG	0.071	0.0812
2	16	TANSIG	0.066	0.0712
3	14	TANSIG	0.1064	0.1219
4	15	TANSIG	0.0677	0.099

[a]The first seasonal model is used in January–April, December; the second seasonal model is used in May and June; the third seasonal model is used in July–September; the fourth seasonal model is used in October and November.

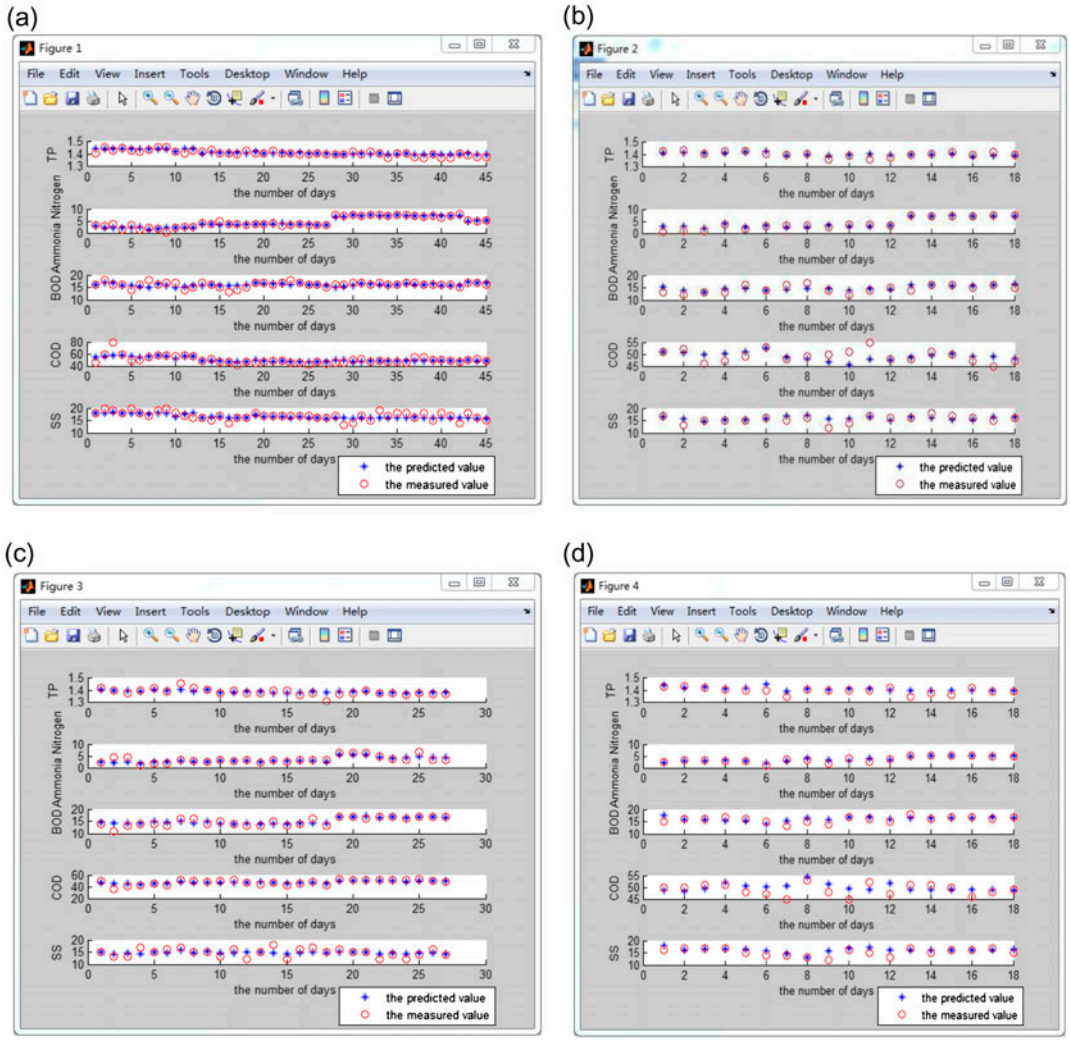

Fig. 4. The comparison curves of prediction values and measured values of seasonal models: (a) the first seasonal model, (b) the second seasonal model, (c) the third seasonal model, and (d) the fourth seasonal model.

models were necessary for effective management of the WTP.

3.2. Clustering analysis result

The monthly data of Table 1 were taken as initial classes in clustering analysis method, which formed 12 initial classes. The clustering analysis was implemented by the software of SPSS (Fig. 1). The clustering schedule and dendrogram are shown in Table 2 and Fig. 2.

Table 2 indicated that the data could be divided into three classes according to their similarity. January–June and December were the first class; the second class contained July–September; October and November formed the third class. Considering there was too much monthly data in the first class, further classification was done for the first class in Fig. 3. It indicated that December data had a close relation with the months from January to April for the early clustering of them. Relatively, the correlation between the data of May, June, and other months was weak in this class, so they were taken out from the first class. Thus, the raw WTP data are divided into four classes as shown in Table 3.

The WTP is located in the city of Harbin, northeastern China. Winter of Harbin is from December to February. Summer lasts from June to August. While spring is from March to May, September–November comes into autumn [37,38]. Contrasting the classified results and seasonal change of Harbin, the classified results obtained by the clustering analysis method

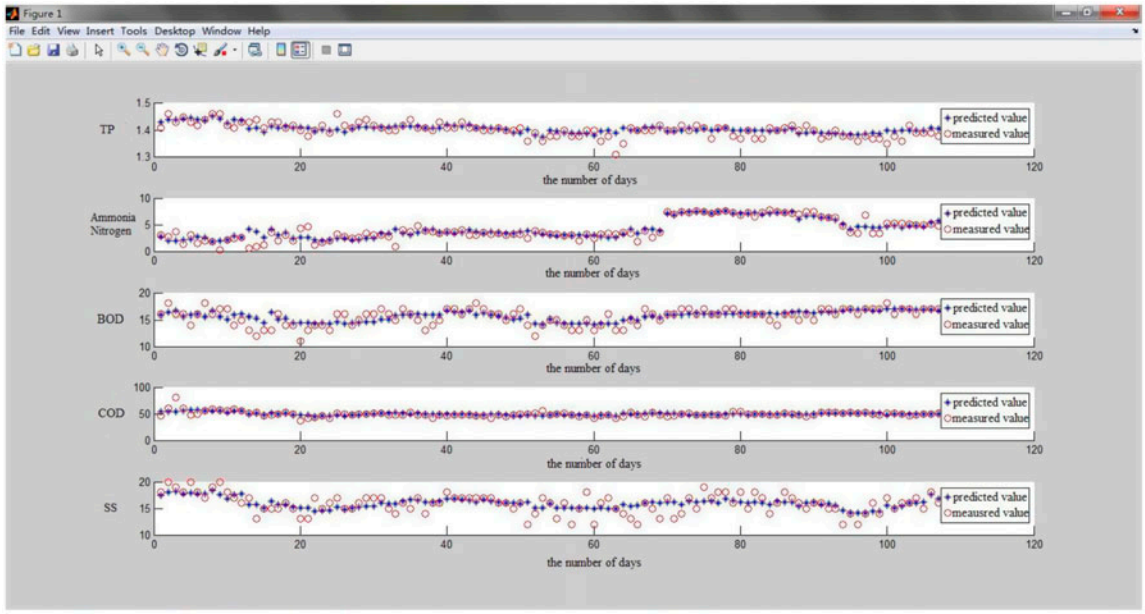

Fig. 5. The comparison curves of prediction values and measured values of the annual model.

were basically consistent with the local seasonal variation.

3.3. Annual ANN model and seasonal ANN model

For offering managers more robust tools to improve purifying ability of wastewater, four seasonal ANN models were built to predict the WTP effluent water quality based on the above-classified data of Table 3. At the same time, the annual model was built by the whole data for comparing the prediction accuracies of the seasonal models and annual model. The structure drawing and parameter design of a BP network model are shown as an example in Fig. 3. The annual model was trained with the whole data from 2009 to 2010 and the training results are shown in Table 4.

Contrasting the errors, the model with the minimum error is the optimized annual model. The neuron number of the hidden layer is 19, the transfer function of the hidden layer is TANSIG, the training error is 0.0513, and the validation error is 0.051705 in the optimized annual model. Most of the training times were less than 2,000 times in Table 4, which were the targeted training times. It showed that overtraining appeared before they met the targeted training times. The models were overtrained earlier resulting in bigger errors, while the models with more training times had lower errors. Thus, the study of preventing overtraining and extending training times is important for better modeling.

Four seasonal ANN models were built, respectively, and the optimized seasonal models were selected in a similar manner as shown above (Table 5).

3.4. Prediction accuracy of the two models

The prediction accuracies of the seasonal models and the annual model were contrasted to verify the superiority of the seasonal models. The test set of each seasonal model contained the first three day's data of the months in corresponding class, if there was no valid value on some day, the next day's value would be selected. The four test sets were, respectively, used to verify the prediction performance of the seasonal models. The comparison curves of prediction values and measured values of the seasonal models and the annual model are, respectively, shown in Figs. 4 and 5.

The prediction errors of the seasonal models are calculated, respectively, in Table 6, including maximum error, minimum error, and average error. At the same time, the prediction error of the annual model was also calculated in each seasonal period (class) and as shown in Table 6. The prediction errors of the annual model and seasonal models were contrasted to verify the superiority of the seasonal models.

In general, the errors between the prediction values and measured values of each parameter were small. It indicated that the ANN model was a robust tool to simulate and predict effluent water quality of the WTP.

Table 6
Prediction errors of the annual model and seasonal models

Effluent water quality parameter	Maximum error		Minimum error		Average error	
	Annual model	Seasonal model	Annual model	Seasonal model	Annual model	Seasonal model
(a) *The first seasonal period*						
TP	0.022551	0.02252	0.000282	0.0000661	0.009509	0.007539
Ammonia nitrogen	8.558185	9.053197	0.00549	0.003307	0.317877	0.207683
BOD	0.224012	0.211148	0.000545	0.000671	0.04992	0.059146
COD	0.327459	0.285498	0.000504	0.000169	0.067106	0.054403
SS	0.238789	0.243038	0.002086	0.001298	0.069704	0.07315
Population mean	1.874199	1.96308	0.001781	0.0011	0.1028232	0.0723842
(b) *The second seasonal period*						
TP	0.032197	0.034499	0.000238	0.000523	0.011991	0.011087
Ammonia nitrogen	5.996028	3.798688	0.008183	0.000833	0.74389	0.52463
BOD	0.265947	0.178537	0.006086	0.001624	0.102611	0.080304
COD	0.174997	0.129994	0.006342	0.000169	0.05392	0.042818
SS	2.985375	0.316323	1.669129	0.020298	2.154099	0.082313
Population mean	1.8909088	0.8916082	0.3379956	0.0046894	0.6133022	0.1482304
(c) *The third seasonal period*						
TP	0.0605692	0.056999	0.0000304	0.000824	0.0134451	0.008298
Ammonia nitrogen	0.7819429	0.725893	0.0169381	0.023996	0.3023342	0.116302
BOD	0.3275406	0.28498	0.008669	0.0000613	0.0676843	0.070251
COD	0.298975	0.253782	0.0005557	0.002417	0.0562858	0.034659
SS	0.2883168	0.241647	0.0045567	0.002651	0.0853155	0.046777
Population mean	0.3514689	0.3126602	0.00615	0.0059899	0.105013	0.0532574
(d) *The fourth seasonal period*						
TP	0.061606	0.034899	0.001174	0.004661	0.01348	0.007373
Ammonia nitrogen	0.89073	1.521605	0.008541	0.000417	0.223996	0.151507
BOD	0.311973	0.190663	0.009653	0.006834	0.063533	0.036066
COD	0.315435	0.122256	0.003643	0.005838	0.057136	0.049995
SS	0.251841	0.312979	0.002963	0.001835	0.090115	0.103536
Population mean	0.366317	0.4364804	0.0051948	0.003917	0.089652	0.0656954

Contrasting the population mean of errors, including maximum error, minimum error, and average error, most population means of the seasonal models were smaller than the annual model. It indicated that the prediction accuracy of seasonal models was better than the annual model. Seasonal models should be a more effective tool to reveal the relationship of raw water quality, energy consumption, and effluent water quality. So it can offer managers more precise information to control and design the processes of the WTP, which result in better purifying ability of wastewater.

4. Conclusions

The raw monthly data in the WTP were analyzed before the application for clustering analysis method and modeling, using range, mean, and standard deviation. The analysis results showed little difference among years for each parameter. And there were obvious differences among seasonal periods for most of the parameters. The differences are related to the local climate of the WTP. Furthermore, it showed that the monthly data in one seasonal period were similar, which indicated that seasonal ANN models were necessary for effective management of the WTP.

For building seasonal models, the WTP data must be classified based on seasonal characteristics of the data. Clustering analysis method was used to classify the data and the raw WTP data were divided into four classes. Contrasting the classified results and seasonal change of local climate, the classified results obtained by the clustering analysis method were basically consistent with the local seasonal variation.

The prediction error analysis indicated that the ANN model was a robust tool to simulate and predict effluent water quality of the WTP. Contrasting the population mean of errors, it indicated that the prediction accuracy of seasonal models was better than the annual model. Seasonal models should be a more effective tool to reveal the relationship of raw water quality, energy

consumption, and effluent water quality. So it can offer managers more precise information to control and design the processes of the WTP, which result in better purifying ability of wastewater.

Acknowledgment

This work was supported by China Postdoctoral Science Foundation (20110491056); supported by Postdoctoral Science Foundation of Heilongjiang Province of China (LBH-Z10172); Research and Innovation Fund project of Harbin Institute of Technology (HIT) in 2011; supported by Harbin Scientific and Technological Innovative Talents Research Special Fund Project of Harbin Municipal Science and Technology Bureau (2013RFQXJ121); Scientific and Technological Research General Program Project of Heilongjiang Provincial Department of Education (12531532).

References

[1] W. Yang, J. Nan, D. Sun, An online water quality monitoring and management system developed for the Liming River basin in Daqing, China, J. Environ. Manage. 88 (2008) 318–325.
[2] X. Gao, The study of scale and efficiency of Chinese wastewater treatment plant, Technological Pioneers 1 (2014) 195–196.
[3] Y. Zhao, S. Ashish, S. Bellie, M. Lucy, A Bayesian method for multi-pollution source water quality model and seasonal water quality management in river segments, Environ. Modell. Softw. 57 (2014) 216–226.
[4] L.B. Yang, S.Y. Zeng, Y.P. Ju, M. He, J. Chen, Statistical analysis and quantitative recognition of energy consumption of municipal wastewater treatment plants in China, Water Wastewater Eng. 34(10) (2008) 42–45.
[5] Y.S. Luo, R.H. Li, S.C. Mei, Data mining model research on complex industry process, Inf. Control 32 (1) (2003) 32–35.
[6] Y.Z. Peng, Z.H. Wang, S.Y. Wang, Simulation of external carbon addition to anoxic–oxic process based on back-propagation neural network, J. Chem. Ind. Eng. (China) 56(2) (2005) 296–300.
[7] M.Z. Huang, J.Q. Wan, Y.W. Ma, Y. Wang, W.J. Li, X.F. Sun, Control rules of aeration in a submerged biofilm wastewater treatment process using fuzzy neural networks, Expert Syst. Appl. 36(7) (2009) 10428–10437.
[8] C.B. Liu, J.F. Qiao, F.F. Zhang, Fuzzy neural network controls of dissolved oxygen in the wastewater treatment processes, J. Shandong Univ. (Engineering Science), 35(3) (2005) 83–87.
[9] R.F. Yu, H.W. Chen, W.P. Cheng, Y.C. Shen, Dynamic control of disinfection for wastewater reuse applying ORP/pH monitoring and artificial neural networks, Resour. Conserv. Recycl. 52(8–9) (2008) 1015–1021.
[10] P. Kundu, A. Debsarkar, S.N. Mukherjee, S. Kumar, Artificial neural network modelling in biological removal of organic carbon and nitrogen for the treatment of slaughterhouse wastewater in a batch reactor, Environ. Technol. 35(10) (2014) 1296–1306.
[11] R.M. Aghav, S. Kumar, S.N. Mukherjee, Artificial neural network modeling in competitive adsorption of phenol and resorcinol from water environment using some carbonaceous adsorbents, J. Hazard. Mater. 188 (1–3) (2011) 67–77.
[12] E. Dogan, A. Ates, E.C. Yilmaz, B. Eren, Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand, Environ. Prog. 27(4) (2008) 439–446.
[13] S.H. Hong, M.W. Lee, D.S. Lee, J.M. Park, Monitoring of sequencing batch reactor for nitrogen and phosphorus removal using neural networks, Biochem. Eng. J. 35(3) (2007) 365–370.
[14] S.I. Ri, D.G. Hou, Z.J. Zhang, W.L. Zhou, L. Chi, W.I. Choe, Prediction of water quality in biological wastewater treatment plant using BP artificial neural network, Mod. Chem. Ind. 29(12) (2009) 66–70.
[15] L.J. Wang, Analysis of sewage treatment aeration energy consumption based on PCA and BP networks, Comput. Technol. Develop. 21(3) (2011) 243–245.
[16] P.S. Kunwar, B. Ankita, M. Amrita, J. Gunja, Artificial neural network modeling of the river water quality—A case study, Ecol. Model. 220(6) (2009) 888–895.
[17] M.S.B. Phridvi Raj, C.V. Guru Rao, Data mining – past, present and future – a typical survey on data streams, Procedia Technol. 12 (2014) 255–263.
[18] J.M. Bland, D.G. Altman, Statistics notes: Measurement error, British Medical J. 312(7047) (1996) 1654.
[19] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: Data mining, inference and prediction, Springer, 2001.
[20] E. Pastuchova, S. Vaclavikova, Cluster analysis-data mining technique for discovering natural groupings in the data, J. Electr. Eng. 64(2) (2013) 128–131.
[21] M. Nikolaos, B. Basilis, G. Ioannis, A method for improving the accuracy of data mining classification algorithms, Comput. Oper. Res. 36(10) (2009) 2829–2839.
[22] Q.W. Chen, E.M. Arthur, Integration of data mining techniques and heuristic knowledge in fuzzy logic modeling of eutrophication in Taihu Lake, Ecol. Modell. 162 (2003) 55–67.
[23] Y.B. Yang, H. Lin, Z.Y. Guo, J.X. Jiang, A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis, Comput. Geosci. 33 (2007) 20–30.
[24] B.H. Chu, M.S. Tsai, C.S. Ho, Toward a hybrid data mining model for customer retention, Knowledge-Based Syst. 20(8) (2007) 703–718.
[25] W. John, Similarity and Clustering in Chemical Information Systems, Wiley, Research Studies Press, New York, NY, 1987.
[26] H. Razmkhah, A. Abrishamchi, A. Torkian, Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: A case study on Jajrood River (Tehran, Iran), J. Environ. Manage. 91(4) (2010) 852–860.
[27] L. Guo, Y. Zhao, P. Wang, Determination of the principal factors of river water quality through cluster analysis method and its prediction, Front. Environ. Sci. Eng. China 6(2) (2012) 238–245.
[28] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, NY, 1983.

[29] Y.T. Hong, M.R. Rosen, R. Bhamidimarri, Analysis of a municipal wastewater treatment plant using a neural network-based pattern analysis, Water Res. 37(7) (2003) 1608–1618.
[30] T.Y. Pai, Gray and neural network prediction of effluent from the wastewater treatment plant of industrial park using influent quality, Environ. Eng. Sci. 25(5) (2008) 757–766.
[31] T.Y. Pai, S.C. Wang, C.F. Chiang, H.C. Su, L.F. Yu, P.J. Sung, C.Y. Lin, H.C. Hu, Improving neural network prediction of effluent from biological wastewater treatment plant of industrial park using fuzzy learning approach, Bioprocess. Biosyst. Eng. 32(6) (2009) 781–790.
[32] N.B. Chang, W.C. Chen, W.K. Shieh, Optimal control of wastewater treatment plants via integrated neural network and genetic algorithms, Civ. Eng. Environ. Syst. 18(1) (2001) 1–17.
[33] T.Y. Pai, P.Y. Yang, S.C. Wang, M.H. Lo, C.F. Chiang, J.L. Kuo, H.H. Chu, H.C. Su, L.F. Yu, H.C. Hu, Y.H. Chang, Predicting effluent from the wastewater treatment plant of industrial park based on fuzzy network and influent quality, Appl. Math. Modell. 35(8) (2011) 3674–3684.
[34] D. Hanbay, I. Turkoglu, Y. Demir, Prediction of wastewater treatment plant performance based on wavelet packet decomposition and neural networks, Expert Syst. Appl. 34 (2008) 1038–1043.
[35] M.M. Hamed, M.G. Khalafallah, E.A. Hassanien, Prediction of wastewater treatment plant performance using artificial neural networks, Environ. Modell. Softw. 19 (2004) 919–928.
[36] F.S. Mjalli, S.A. Asheh, H.E. Alfadala, Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance, J. Environ. Manage. 83 (2007) 329–338.
[37] C.G. Yin, L. Rosendahl, Z.Y. Luo, Methods to improve prediction performance of ANN models, Simul. Model. Pract. Theor. 11(3–4) (2003) 211–222.
[38] C. Karul, S. Soyupak, A.F. Çilesiz, N. Akbay, E. Germen, Case studies on the use of neural networks in eutrophication modeling, Ecol. Model. 134(2–3) (2000) 145–152.
[39] N. Shi, J.Q. Chen, Q.P. Tu, 4-phase climate change features in the last 100 years over China, Acta Meteorol. Sin. 53(4) (1995) 431–439.
[40] Z.C. Zhao, Y. Luo, Projections of climate change over northeastern China for the 21st century, J. Meteorol. Environ. 23(3) (2007) 1–4.