TAO TIAN
HUI-FENG XUE
FENG ZHANG
and
JIAN-GANG DONG

Data validity analysis and hybrid soft computing prediction method of water resources in mining area

In order to solve the problem of detecting and correcting the outliers of mining area water resources monitoring data, a hybrid soft computing method based on PSO and BP network was proposed to detect the outliers of time series data. First, through the classification and screening of the initial outlier data, the mining area water resource big data "full data" associated model was constructed to fit the regression curve between each of the two different data parameters, then use the "full data" associated model of mining area water resources big data to define another parameter set whose linear correlation coefficient is the largest, which is the correlation monitoring data parameter set of the monitoring data parameter. Through the correlation analysis between data, the overall change trend of most data can be fitted, without the influence of abnormal values, and the abnormal values can be detected effectively according to the deviation ratio. Finally, taking an area's annual mining area water consumption data as an example, we make an empirical analysis. The results show that the hybrid soft calculation method proposed in this paper can effectively detect abnormal values and the corrected data can truly reflect the situation of mining area water consumption in the area, and can provide more real and reliable data for subsequent analysis.

Keywords: Mining area, outlier's analysis, soft computing, BP network, particle swarm optimization

1. Introduction

At present, the situation of mining area water resources in China is very serious. The serious water pollution, the deterioration of water ecological environment and the shortage of water resources have become the main bottleneck of the sustainable development of the economy and society[1]. The implementation of the "river chief system" is urgently needed to predict and deal with all kinds of basic data and monitoring data in the provinces, cities, counties. Since 2012, China's water resources monitoring and capacity building project has covered 14,034 national monitoring targets for water resources, and has achieved full coverage of water quality monitoring for more than 70% of watersheds, 80% of important water function areas, and major provincial water quality monitoring areas. And about 55% of the water is monitored and terabytes of massive data are generated in real time every day, but the accuracy of these data is not high. The availability of water resources monitoring data has a variety of manifestations[2]. In terms of data consistency, the correlation between monitoring indicators has been destroyed, and the correlation of neighbouring nodes has been destroyed; in terms of timeliness, the performance of indicators for monitoring data obsolete. In terms of data integrity, some monitoring indicators are missing data and data are abnormal; in terms of data accuracy, they are negative data, continuous long-term data change, data inversion, and data fraud at monitoring sites. There is a rich relationship between water resources big data, and an important basis for mining the value in big data is to be able to analyze the hidden network of relationships in the data set[3]. For some specific monitoring indicators of data outliers, that is, the data is invalid data, need to be deduced through water resources big data and predict the real and effective data. In this paper, a mining area water quality monitoring data validity analysis method based on hybrid soft calculation method was proposed to detect the abnormal value of mining area water resources data. By studying the relationship between data, or by specific algorithm, we can detect, correct or reject abnormal data.

Messrs. Tao Tian, Hui-feng Xue and Feng Zhang, China Aerospace Academy of Systems Science and Engineering, Beijing 100048, China, Feng Zhang and Jian-gang Dong, School of Information Engineering, Yulin University, 719000, Yulin, China
*Email of the corresponding author: tigertiantao@163.com

2. Theoretical method

2.1. BP neural network algorithm

BP neural network is a multi-front one-way transmission of the network, BP algorithm includes two basic aspects: When the counter-propagating signals prior to the propagation and error, that is calculated according to the actual output from the input to the direction of the output will be, and the correction weights and thresholds direction from the output to input will be shown in Fig. 1[4]. BP algorithm because of its simple, easy, the advantages of less computation, parallel

and strong, is currently training the neural network using the most mature one of the most training algorithm. However, BP's learning efficiency is low, slow convergence and easy to fall into local minimum state[5]. Due to the lack of historical data, collected data sample is limited. Training defective BP neural network based on the Yulin area for a limited environmental sample data, the use of stereotypes wavelet neural network and particle swarm algorithm (Particle swarm optimization, PSO) to optimize learning BP network to accelerate convergence and avoid local minima It has some effect, and predict its trend[6].

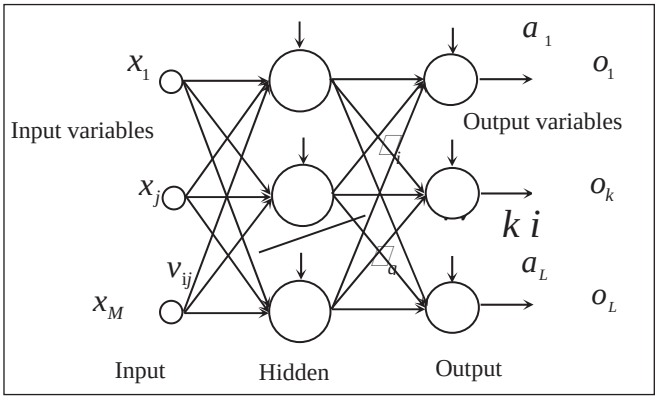

Fig. 1: BP network structure diagram

In Fig. 1, the calculation method for the input net_i of the first *i-th* node of the hidden layer is:

$$net_i = \sum_{j=1}^{M} v_{ij} x_j \tag{1}$$

The calculation method for the input y_i of the first *i-th* node of the hidden layer is:

$$y_i = f(net_i) = f\left(\sum_{j=1}^{M} v_{ij} x_j \right) \tag{2}$$

The calculation method for the input net_k of the first k node in the output layer is:

$$net_k = \sum_{i=1}^{q} w_{ki} y_i = \sum_{i=1}^{q} w_{ki} f\left(\sum_{j=1}^{M} v_{ij} x_j \right) \tag{3}$$

The calculation method for the input σ_k of the first k node in the output layer is:

$$o_k = f(net_k) = f\left(\sum_{i=1}^{q} w_{ki} f\left(\sum_{j=1}^{M} v_{ij} x_j \right) \right) \tag{4}$$

Before training in the network, we must first normalize the input and output samples. For input samples, the following normalization formula is used:

$$a_i^k = \frac{x_i^k - x_{i,\min}}{x_{i,\max} - x_{i,\min}} d_1 + d_2 \tag{5}$$

2.2. Particle swarm optimization algorithm

Particle swarm optimization algorithm was proposed in 1995 by Kennedy et al[7]. The basic PSO algorithm can be described as follows: Let the search space be N dimension, the number of particles is P, the position and velocity of the i-th particle in the N-dimensional search space are $X_i = (x_{i1}, x_{i2}, x_{iN})$ and $V_i = (v_{i1}, v_{i2} \ldots v_{iN})$. The individual extremes and group extremes of the particles are $P_i = (p_{i1}, p_{i2} \ldots p_{iN})$ and $P_g = (p_{g1}, p_{g2} \ldots p_{gN})$. The particle flight is shown in Fig. 2.

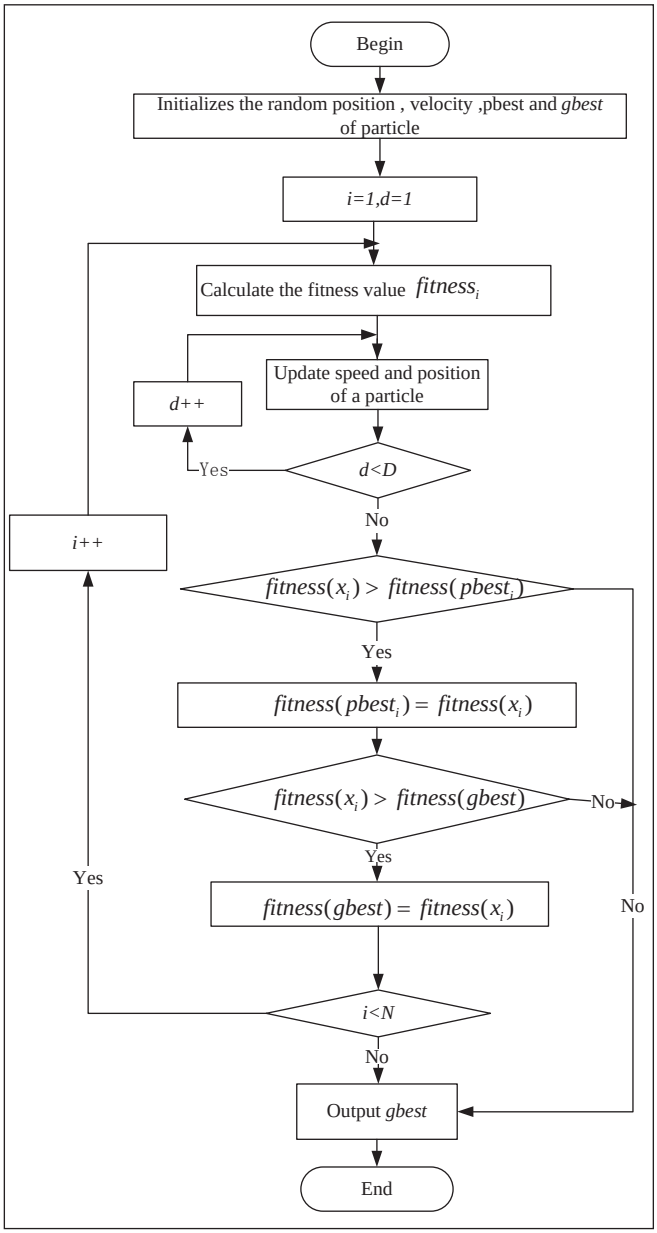

Fig. 2. Particle swarm optimization process

PSO algorithm mathematics as follows:

Its optimization problem model:min *f(x)*

Let *f(x)* search space is *D*-dimensional, the total number of particles is *N*. The position of the $(_{i = 1, 2, \ldots N})$ particles is $X_i = (X_{i1}, X_{i2}, \ldots X_{ij}, \ldots X_{iD})$, the flight speed of the *i* particles is $V_i = (V_{i1}, V_{i2}, \ldots V_{ij}, \ldots V_{iD})$, the optimal position of the *i* particle flight history is *pbest*, then $P_i = (P_{i1}, P_{i2}, \ldots P_{ij}, \ldots P_{iD})$, in this group, at least one particle is optimal, denoted *gbest*, then $P_{gbesti} = (P_{gbest1}, P_{gbest2}, \ldots P_{gbestD})$ is the global history

optimal position of the current group. $fitness_i = f(x_i)$ represent the position, velocity and fitness value of i particles in i.

The position update formula of each particle is:

$$v_{ij}(t+1) = \omega.v_{ij}(t) + c_1.r_1.(p_{ij} - x_{ij}(t)) + c_2.r_2.(p_{gbestj} - x_{ij}(t)) \quad (6)$$

$$x_{ij}(t+1) + x_{ij}(t) + v_{ij}(t+1)$$

Where, t represents the number of iterations, $i = 1, 2, ..., N$; $j = 1, 2, ..., D$; $c_1, c_2 > 0$ factor represents individual learning and social learning factor, r_1 and r_2 are both in the range between [0,1] independent random factor[8]; ω represents the inertia weight used to weigh the ability of local optimum and global optimum capacity. In order to balance global and local search capabilities, and its value should decrease linearly with the evolutionary algorithm, ω is defined as: $\omega = \omega_{\min} + (iter_{\max} - iter) \times (\omega_{\max} - \omega_{\min})/iter_{\max}$

Where, $\omega_{\min}$, $\omega_{\max}$ respectively maximum and minimum weight factor, *iter* is the current iteration number, $iter_{\max}$ is the total number of iterations. Particle swarm optimization process shown in Fig. 4, the process of the algorithm is as follows:

1) Random initialization position and velocity of the particle swarm.
2) Calculate the fitness value of each particle $fitness_i = f(x_i)$, corresponding initialization $pbest_i = fitness_i$, $gbest = \min(fitness_1, fitness_2, ..., fitness_N)$, i = 1, 2, ..., N
3) For each particle, its fitness compared to *pbest*, if it is the best, it is the best as the current position and update the *gbest* and *pbest*.
4) The adaptation values of each particle are compared to the adaptation values of *pbest*. If better, then as *gbest*.
5) Iterative update speed and position of a particle.
6) If the number of iterations unfinished or find a satisfactory adaptation value, will continue to calculate the fitness value of each particle.
7) Output *gbest*.

3. Hybrid soft calculation prediction model for water resources data validity prediction

The analysis of the effectiveness of water resources big data is an important means to prevent the generation of invalid data, and the analysis of the validity of data in a general sense can be based on the analysis of the validity of existing data[9]. This kind of analysis can only explain how much effective data there is and how many invalid data there are, but it cannot reduce the proportion of invalid data. Therefore, how to reduce invalid data is a key issue that needs to be solved in data validity analysis. It is very important to accurately predict the invalid probability of invalid data if it can predict and forecast the possibility of invalid data in advance, and alarm the maintenance personnel of the monitoring point in real time, cleaning, maintaining and replacing the monitoring equipment in time, so it can reduce the ratio of invalid data[10]. At the same time, the effective data of advance prediction can confirm each other with the actual monitoring data and invert each other, which play an important role in correcting the accuracy of data validity analysis. The dynamic observation data of the water resource system in the basin are ordered series according to the time series. The dynamic changes are mainly caused by the precipitation, the quantity of workers and the random factors. But traditional methods such as time series analysis and grey system prediction cannot reflect the dynamic mechanism of water resources system. The artificial neural network has the powerful function of dealing with the problems of nonlinearity, uncertainty and randomness. The artificial neural network is used to establish the dynamic prediction model of water resource data validity.

3.1. Setting and forecasting method of data validity threshold

In view of the actual demand for the accuracy of the water resource data effectiveness analysis, the system engineering method is used to combine the short-term and long-term, the static and dynamic combination, the history and the future. In this paper, we use the prediction model of large data validity based on hybrid soft computing to realize the analysis and prediction of water resources data.

(1) To select the last 10 years of historical monitoring data (including water, water pH, water temperature, turbidity, ammonia, total phosphorus, total nitrogen, dissolved oxygen, COD, BOD and other indicators) for statistical analysis. The regression data $Y = kx + b$ between every two data parameters is fitted by the "data association" model of water resources data. The linear correlation coefficients between them were obtained, and significant tests were performed to retain meaningful correlation.

(2) Participating in every monitoring data parameter of statistical analysis, and using the "full data" association model of large data of water resources to define the other parameter set with the maximum linear correlation coefficient as the interrelated monitoring data set of the monitoring data parameters.

(3) The latest two normal monitoring data of water resources data parameters is X_{Ai}, X_{Ai+1}, $i = 1, ..., n$, and the corresponding monitoring time is t_i, t_{i+1}. The average data rate of the monitoring data is $V_i = |X_{At+1} - X_{at}| / (t_{i+1} - t_i)$ from t_i to t_{i+1}, taking v_i as the monitoring data parameter and referring to the rate of change, that is $v_t = v_i$. When the latest two normal data of monitoring data A is updated, the reference change rate of the parameter needs to be recalculated. The water resources monitoring and control capacity construction project management center received the latest collected data X_{Ai+2} of the A parameters. After monitoring time t_{i+2}, calculate that the average rate is

$V_{i+2} = |X_{Ai+2} - X_{Ai+1}| / (t_{i+2} - t_{i+1})$ from t_{i+1} to t_{i+2}. If $v_{i+1} > 2*v_t$, then the *A* parameter fluctuates greatly. The validity of this monitoring data needs to enter step (4) for diagnosis; otherwise, the *A* parameter for this monitoring data is considered normal.

(4) When a monitoring data parameter *A* has a significant fluctuation, when the effectiveness of monitoring data (x_{A3}) needs to be diagnosed. Find out the parameter set *B* of the parameter determined in step (2). The *B* parameter set corresponds to the monitoring time. The corresponding set of monitoring data is Y_{B3}. Through the steps (1), fitting *A*, *B* linear regression curve $Y = kX + b$ and *B* parameters and the monitoring data Y_{B3}, we predict and estimate the *A* parameter with a credibility of 0.98, and the he confidence interval is calculated to be *(m,n)* if $X_{A3} \in (m,n)$ it is diagnosed as monitoring data parameter *A* and *B*. When the monitoring data X_{A3} and Y_{B3} are at 0.98 confidence level, it is satisfies the regression equation, that is, monitoring data parameters and appear similar fluctuations, and this fluctuation is due to the large change of original monitoring data, which is a normal change, the monitoring data of *A* parameters are normal and effective. Otherwise, it is considered that the monitoring data parameters *A* and *B* do not appear similar changes, A parameter fluctuation is due to the instrument failure, the A parameter of the monitoring data is abnormal data, and the effective monitoring data should be predicted through the correlation data set.

3.2. PSO NEURAL NETWORK OPTIMIZATION (PSO-NN) ALGORITHM

PSO neural network optimization (PSO-NN) algorithm is used to replace the traditional parameters of PSO training algorithms such as BP to optimize NN. Each particle is a vector that represents a group of parameters, process the process for the global optimal value is to obtain the optimal parameters.

The training error is used to calculate the fitness value *f(x)*, it is given

$$f(x) = \frac{1}{1 + \frac{1}{2n}\sum_{k=1}^{n}(y_k - t_k)} \tag{8}$$

Where, *k* is the number of samples, y_k is the actual output value, t_k is the output value. When it reach the maximum number of iterations or target error, the programme terminates, and get the global optimum value that a group of optimal parameters. In the absence of a priori information, if the *W* is too small, then all the incentive function are almost in a linear part, will also reduce the speed of convergence, The general consensus is that *W* obeys the exponential distribution, the formula is as follows:

$$P(W) = \frac{1}{Zw(\alpha)}\exp(-aEw) \tag{9}$$

4. Experimental analysis

There are many influencing factors before water resources big data, and there is a strong non-linearity in causality. Before predicting valid data, it is necessary to determine the relationship between the impact factors, and establish a predictive model of water resource big data effectiveness under natural and human activity conditions. Hybrid soft calculation method is a method to determine the dynamic influencing factors of water resources system, for a variety of factors included in an abstract system, it is necessary to identify whether the impact on the system is to be developed or suppressed. Its essence is to compare the time series reflecting the changes of various subsystems (or factors) in order to find the numerical relationship among subsystems (or factors) in the process of system development. To establish a PSO-BP network model for predicting the validity of water resources big data and comprehensively consider various factors. A_{out} as an output variable, *B* is a set of factor parameters associated with parameter *A*. The sample data comes from the observation data of the national water resources monitoring capacity building project from 2012 to 2016. All relevant samples participate in the training and inspection of the network. This paper deals with statistics data from January 2012 to March 2017 as input and output samples. Samples from January 2012 to December 2014 were selected as training samples for monthly water resources. Select 3 layers of BP neural network, each layer transfer function is S-type function, the MAT LAB neural network tool box is used to write the programme, and a data validity prediction model is established to train input and output samples. For the B-parameter identification problem, the most commonly used optimization methods are based on gradient search, and its defects lie in the sensitivity of the initial estimation of the model parameters and the local minimum problem. Compared to the traditional gradient-based search-based optimization method, the hybrid soft-computation method has good global convergence characteristics. The inverse problem of B-parameter identification was transformed into a combinatorial optimization problem, and a strategy of hybrid soft-computing method to identify two-dimensional B-parameter sets was proposed. Monitoring data comes from the database of water resources management system. The monitoring data of the time series of original monthly water consumption is $\{m_i\}$, as is shown in Fig. 3. As can be seen from Fig. 3, there are abnormally large values, anomalous small values, and 0 values, and there are multiple consecutive 0 values. In order to embody the advantages of the method proposed in this paper, we compared the traditional outlier detection method with the box plot, and we used a hybrid soft calculation method to predict the outliers. The actual monitoring data value predicted by hybrid software as is shown in Fig. 4. Using the model of 2012 to December 2017 was selected as training samples for monthly water resources level is simulated, and the test from 2012-2016,

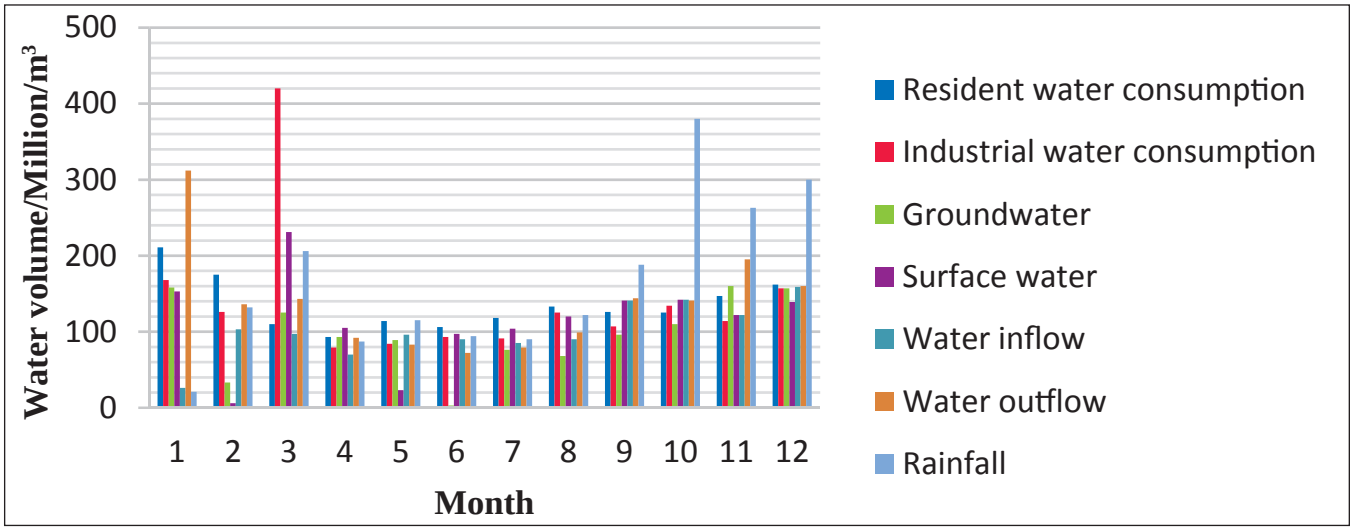

Fig. 3. Primitive monitor data (2012.1-2017.12)

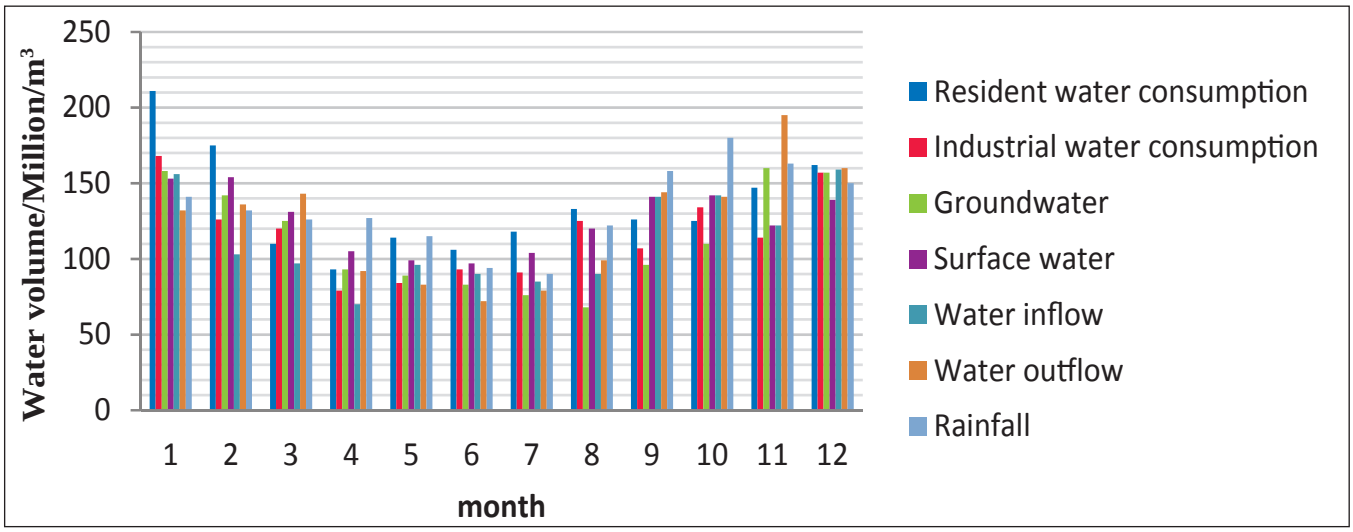

Fig. 4 The actual monitoring data value predicted by hybrid software (2012.1-2017.12)

the monitor water level, and the simulation values and the measured values and the relative error is shown in Table 1.

Table 1. Simulation results of BP-PSO model

Times	Actual value (Million/m^3)	Simulation value (Million/m^3)	Relative error %
2017	463.35	473.58	0.132
2016	484.35	471.21	-0.111
2015	595.54	589.99	0.165
2014	565.55	575.89	0.226
2013	613.33	623.87	0.203
2012	614.51	604.65	-0.231

5. Conclusions

In order to give full play to the role of monitoring data in water resources management business, effective data preprocessing is very important. In this paper, we use BP network and PSO algorithm to build a hybrid soft computing method for predicting the effectiveness of water resources data. Conventional outlier's detection and correction methods have certain requirements for the number of missing data in time series. If the data is missing, it is difficult to recover. The more missing data, the more difficult it is to recover. The mixed soft computing outliers detection and curve fitting outliers correction method proposed in this paper also need to consider data missing before using. Compared to the traditional methods of outlier detection and correction, the results show that the method presented in this paper is more superior. After the detection and correction of outliers, it can get closer to the actual daily water intake data.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 11641002, 51651901); Natural Science Basic Research Plan in Shaanxi Province of China (2015SF261, 2017NY-134, 2016KJXX-62), Funding Project for Department of Yulin University (16GK24,13YK50), and the authors thank for the help.

(Continued on page 628)