Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2018.22556

121 (2018) 351–356
July

Water quality anomaly detection approach based on a neural network prediction model

Wei Liu[a], Lisha Tan[b], Qicheng Xu[c], Zhijun Gao[d,*]

[a]*Network Center, Shenyang Jianzhu University, Shenyang, China*
[b]*Students' Affairs Division, Shenyang Jianzhu University, Shenyang, China*
[c]*School of science, Shenyang Jianzhu University, Shenyang, China*
[d]*Information and Control Engineering Faculty, Shenyang Jianzhu University, Shenyang, China, email: gzj_net@sjzu.edu.cn (Z Gao)*

Received 24 February 2018; Accepted 28 May 2018

ABSTRACT

There is too high false positive rate in water quality anomaly detection in water quality data processing with more impulsive noise, so an approach based on radial basis function neural network and wavelet denoising is presented. It introduces wavelet transform modulus maxima denoising method to process the residual sequence prediction of water quality. The quality anomaly of water is determined by the comparison between the distance from the origin at each moment and special threshold, to achieve anomaly detection with higher accuracy. Due to less abnormal data contained in daily water quality data, we perform simulations with a method of superimposing certain distribution based on actual data, to better simulate the variation of water quality parameters in sudden pollution accident of city. The simulation results indicate the improved detection scheme based on neural network, and wavelet analysis has strong on-line detection ability, especially for low-intensity abnormalities, and the accuracy of detection also achieves significant improvement.

Keywords: Water quality anomaly; RBF; Wavelet analysis; Denoising; Threshold

1. Introduction

With the rapid development of industrialization and urbanization, a large amount of pollutants and wastes are discharged into rivers, lakes, and seas, causing serious pollution to water environment for human survival [1–3]. To find and control water pollution in time, it is very necessary to improve the ability of water quality anomaly in water quality monitoring system. Because of complexity of water environment, most of current water monitoring systems, along with diversity of pollutants and randomness of unexpected events, acquire the water category to make warning according to setting thresholds, which is hard to meet the demand for automatic and intelligent detection on water quality events [4]. Therefore, water quality early warning system still has great research and development space in intelligence, accuracy, and timeliness of situation judgment [5–8]. Because of severe situation of water environment, it has become a research hotspot that needs further study and urgent solution to replace traditional warning method of water quality by advanced quality anomaly detection based on information processing technologies.

In recent years, many scholars have carried out research on water quality anomaly detection method. Noel et al. [9] processed experimental and adopted the method of reducing background data noise to improve the detection rate for anomaly. To overcome the defect of simple experimental platform, Zhou et al. and Schurer et al. [10,11] adopted on-line filed data for further analysis, and they proposed the concepts of time series increments, linear filtering, and multivariate nearest

[*] Corresponding author.

Presented at the 3rd International Conference on Recent Advancements in Chemical, Environmental and Energy Engineering, 15–16 February, Chennai, India, 2018.

neighbors. Wang et al. [12] presented a comprehensive evaluation method for surface water quality based on fuzzy set. It is believed that most of the current water quality monitoring system usually performs warning according to the difference between measured value and standard value. However, such approach neglected special changing trend of the parameters of water quality, and it cannot satisfy the demand for all abnormal water quality in time and accurately. Most of the above detection methods often contain two procedures: predicting current value and judging whether the difference between current value and measured value is beyond the threshold. Time series incremental method is based on moving time window with length of 1, which does not consider the trend of historical data sufficiently. Linear filtering method [13] is more suitable for linear stationary time series prediction by linear combination of historical data to predict current value. Multivariate nearest neighbor method [14] integrates different types of water quality indexes, and it clusters original water quality data directly to determine the anomaly with multidimensional Euclidean distance. Its performance is affected by normal fluctuation of background data greatly which will cause higher false negative rate and false positive rate.

This article introduces wavelet transform modulus maxima denoising method to process the residual sequence of water quality prediction, considering the complexity of water environment and non-linear water quality index time series. Then, the results are compared with given threshold to determine the anomaly of water quality, and to achieve anomaly detection with higher accuracy. During the experimental analysis process, contraposing to the problems on water quality anomaly data in real life that is hard to be acquired, so we adopt water anomaly simulation, to compare the water quality anomaly detection method based on prediction model with time series increment method. The simulations show the former has higher detection rate and lower false positive rate. The water quality anomaly detection methods with wavelet denoising are tested to verify its performance, resulting in reduction of false positives caused by outliers and better anomaly detection capability.

The rest of this paper is organized as follows: Section 2 presents a water quality anomaly detection model based on radial basis function (RBF) neural network and explains the procedures; Section 3 describes the improved water quality anomaly detection algorithm based on RBF and wavelet denoising; Section 4 provides the experiments of algorithm tests by simulation of water quality anomaly and the results are analyzed; and Section 5 draws some conclusions and presents the future work.

2. Water quality anomaly detection model based on RBF network

The process of water quality anomaly detection based on prediction model is: first, the historical monitoring data of several water quality parameters are selected as training samples to establish normal water quality mode by neural network training; second, the water quality parameters measured in real time are used as input parameters of the model, and output value of model is used as prediction value of water quality; the prediction of water quality parameters is compared with actual monitoring value to achieve the residual, which will also be compared with regulated threshold. If it is less than the threshold, it is believed to be normal background data; otherwise, it is believed to be abnormal data. Due to the characteristics of non-linear time series of water quality index, we adopt RBF neural network [15] to establish water quality index prediction model. Then the residual time series is acquired by the comparison between prediction value and actual value to make online wavelet denoising in special sliding window. Finally, the water quality anomaly is determined by case whether the judgment error exceeds the threshold value or not. The method of time series prediction with RBF predicts the quantitative relationship between origin and predictive horizon by nonlinear approximation ability of RBF. By the analysis of actual monitoring data, it is found that the change of water quality index is a gradual process.

The prediction model presented in this article is described as follows:

$$D(t) = F(D(t-1), D(t-1), \ldots, D(t-n)) \tag{1}$$

where $D(t)$ is the monitoring data of certain water quality index at hour t; n is the number of input layer nodes; and F is input–output mapping relation determined by neural network.

Matlab is used for simulation of RBF neural network: first, the data are normalized by $x_{\mathrm{i}} = (x - x_{\min})/(x_{\max} - x_{\min})$; second, determine the number of input nodes and generate input variables and output variables according to Eq. (1), to be divided into training samples and test samples; third, use function Newrbe to create accurate neural network and find the optimal basis function extension speed; use function to perform simulation on the network and output the prediction of test samples; finally, anti-normalization processing of the output.

3. Water quality anomaly detection approach based on neural network prediction and wavelet denoising

3.1. Improved water quality anomaly detection

The anomaly detection method of water quality based on neural network prediction compares residuals and sets threshold to judge whether anomaly exists there. The residual series contains more pulse noise, and they are usually caused by unstable sensor running or transmission failure, which should not be determined to be quality anomaly. Thus, these residual series should be processed to eliminate the pulse noise, and to reduce false positives. Because wavelet modulus maxima has better removal efficiency in pulse noise, we adopt such method to perform denoising. Then the results are compared with given threshold to judge the anomaly of water. The flowchart is depicted as Fig. 1:

3.2. Selection of abnormal threshold

A group of residuals are acquired by RBF neural network prediction, presenting the difference between predicted data and real data, as shown in Eq. (2):

$$e(t) = |y(t) - y'(t)| \tag{2}$$

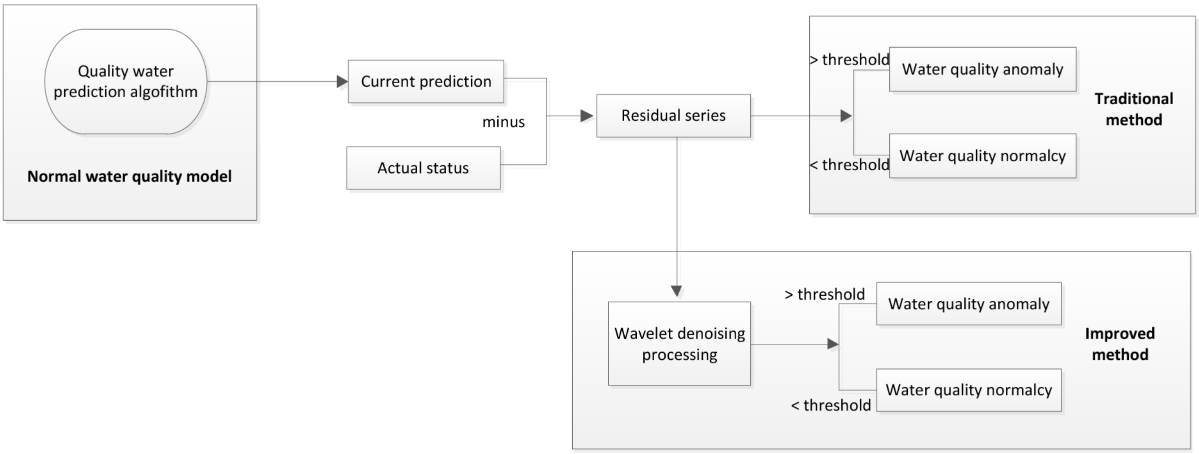

Fig. 1. Improved quality anomaly detection algorithm.

where $e(t)$ is residual at time t; $y(t)$ is measured value at time t; $y'(t)$ is predicted value at time t; $e(t)$ is used to determine the status of $y(t)$. If $e(t) \le U$, $y(t)$ is normal; otherwise, $y(t)$ is abnormal and it will be eliminated by historical data. U is threshold that is a positive constant set previously. The setting of U is very important for anomaly detection. When it is set too small, more false positives may be generated; otherwise, many important information may be missed, leading to false negatives.

In error theory, the elimination of abnormal data often adopts statistical method. Laita criterion [16] is a simple and common judging rule to be used. For repeated measurement of certain variable of nth power equal accuracy, we get arithmetic mean as $x_1,x_2, \ldots, x_n$. $\bar{x}$ is arithmetic mean and σ is standard deviation of measurement. If corresponding residual r_k of certain data x_k satisfies $|v_k| = |x_k - \bar{x}| > 3\sigma$, x_k is believed to have gross error, and it belongs to abnormal data. Because each measuring value only contains random error and it obeys normal distribution, the probability of residual that falls out of 3σ is 0.27%, that is, the possibility occurs in finite repeated measurement is very small. When strict examination is adopted 2σ can be taken as discriminant criterion.

Based on above ideas, we set the threshold as follows: the residual series is acquired by training samples as $e(t)$, 1,2, …, P, where P is the length of training samples. Then the mean and the standard deviation of residual series are computed as follows:

$$\bar{\varepsilon} = \frac{1}{P}\sum_{t=1}^{P} e(t) \tag{3}$$

$$\sigma = \sqrt{\frac{1}{P-1}\sum_{t=1}^{P}[e(t) - \bar{\varepsilon}]^2} \tag{4}$$

When $U = \bar{\varepsilon} + 2\sigma$ and $e(t)<U$, the detection is normal.

3.3. Parameters selection of denoising algorithm

During the denoising by modulus maxima of wavelet transform, the effect will be affected by decomposition level J, threshold T and signal reconstruction method, so it is necessary to choose appropriate parameter value:

3.3.1. Selection of J

Large decomposition level will cause loss of certain important local characteristics of signals, while small decomposition level may cause insufficient attenuation of module maximum corresponding to the noise, and hard distinction between signals and noises. Therefore, the value of J should be determined according to the size of signal-to-noise ratio (SNR). In general, when SNR is large, J will get smaller; otherwise, J will get bigger. There are some indications for common signals that when SNR is more than 20, $J = 3$; otherwise, $J = 4$.

3.3.2. Selection of T

When the noise is constant, if the wavelet modulus maximum amplitude of signals has litter difference, it indicates the signals are relatively stable and the value of T will get larger; otherwise, T will get smaller to decrease signal distortion. When the signal is stable, if SNR is large it indicates the noise power is smaller and its corresponding modulus maximum amplitude will be smaller. Thus, T will get a smaller value for denoising; otherwise, T will get a larger value. In general, the threshold is determined as follows:

$$T = \frac{\log_2(1+2\sqrt{N})}{J+Z}A \tag{5}$$

where $A = \max(w_{2'}^{d} f(n_i))$, that is, the amplitude of maximum modulus maxima; N is presented noise power; J is the maximum size to be selected; Z is a constant, whose value is set as 2 by experience.

3.3.3. Selection of reconstruction method

When the wavelet transform modulus maxima of signals of all scales are acquired, if we directly set zero for wavelet coefficients and adopt wavelet inverse transform for reconstruction, larger reconstructed signal error may be acquired.

We can adopt proper method to make reconstruction from all scales of modulus maxima and their locations. Then the wavelet coefficients are used to get reconstruction signal by inverse transformation.

4. Experimental analysis of water quality anomaly

In the water quality data acquired by on-line monitoring instruments, the data including abnormal events are less and the performance of water anomaly can be tested well. Because the current conditions of laboratories are limited, it is not appropriate to simulate abnormal events occurring by injecting different concentrations of pollutants. Then this article adopts *U*-distribution [17] to simulate the variation of water quality parameters caused by water pollution events, and the relative equation is follows:

$$y_E(t) = y_0(t) + E_{\text{ind}}(t)\cdot\delta\cdot E_{\max}\cdot\sigma_y \tag{6}$$

where $y_E(t)$ is the value of water quality parameter of superimposed exception event at time *t*; $y_0(t)$ is the value of water quality parameter under original background at time *t*; $E_{\text{ind}}(t)$ is event change indicator and its range is (0,1) which simulates the variation trend of event with time goes on; δ is the variation direction that influences the water quality parameters by contaminants. It equals to –1 or 1: when δ = 1, the water quality parameter is increasing; when δ = –1, the water quality parameter is decreasing; σ_y is standard deviation of water quality background data; $E_{\max}$ is the intensity of abnormal events and $E_{\max}\ \sigma_y$ is the maximum of deviation from initial data caused by contaminants.

We choose the ammonia nitrogen value monitored online of certain surface water source reservoir. The sensor output monitoring value every 15 min. The previous 3d data are taken as prediction training set, and the latter 3d data are taken as test set. To test the performance of algorithm, certain distribution is superimposed to simulate the change of water quality index caused by sudden pollution accident based on actual data. The simulated anomaly in this article is density function curve of standard normal distribution, denoted by $Ae - \frac{x^2}{2} / \sqrt{2\pi}$. Such type of curves can simulate the variation process of water quality index caused by pollution events, as depicted in Fig. 2. The anomaly intensity degree is set as $A = 1$ and the duration length is 10Δ*t*.

The prediction performance of RBF is tested by online monitoring of ammonia nitrogen value without superimposed anomaly. The former 3d data are used as training data set, and the input dimension is 3, whose results are depicted in Fig. 3. From this figure we can see, RBF neural network brings better eater quality prediction ability. Time series incremental method makes standardized processing on the data in sliding window. Whether there is anomaly or not, it will use the measured value of previous step as current prediction value. Such method creates the problem that it cannot predict background data when there is anomaly. However, RBF prediction mine the historical trend of water quality variation and it can predict the value of background data accurately. When the difference between predicted value and actual value is large, it believed to be abnormal.

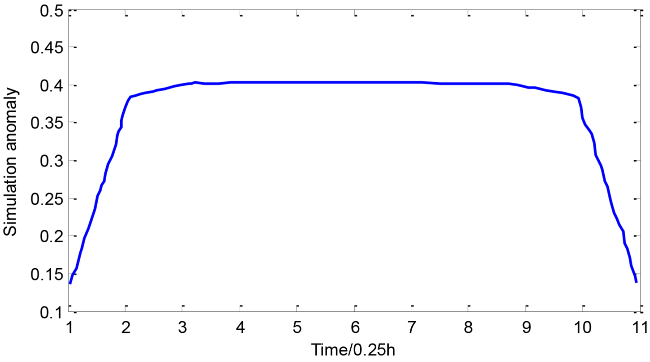

Fig. 2. Simulation of water quality anomaly with anomaly intensity degree *A* = 1.

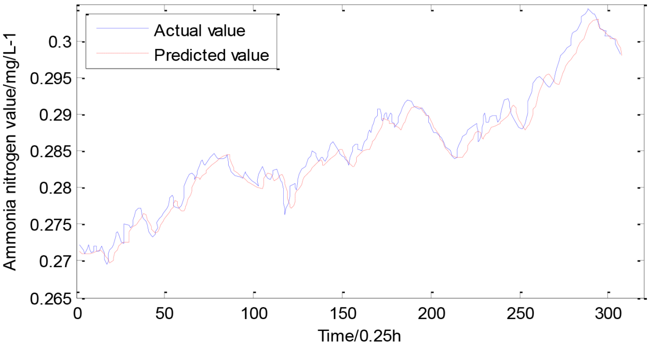

Fig. 3. Comparison curve of predicted value by RBF neural network and actual value.

We choose sym4 wavelet whose decomposition level *J* = 3 and the maximum threshold *T* = maximum modulus maxima*20/(*J*+2). Alternating projection algorithm is used to reconstruct wavelet parameters. In Fig. 4, there are comparisons of residual series before and after denoising with superimposed anomaly strength as 1, 1.5, and 2, respectively. Whatever large the anomaly strength is, our method can eliminate pulse noise contained in corresponding residual series, and save usefully information in reserved residuals simultaneously.

After the residual sequence is processed by cloud noise, it is compared with the threshold of residual chlorine: if it is bigger than the threshold, it is decided as anomaly; otherwise, it is normal. To verify the strength of testing ability we compute receiver operating characteristic (ROC) curve area, detection rate, and false positive rate of two algorithms. From the data listed in Table 1, we can infer that when anomaly intensity is low, the ROC area and detection rate will be bigger than that of improved algorithm. With increasing of anomaly intensity, the detection rate of improved algorithm also gets better. The reason is that low anomalies are considered to be normal fluctuations of water quality parameters. The improved algorithm filtered the residual series which reduces the residual value of relative moments, leading to decrease of detection rate. When anomaly intensity denoising is useful to extract useful information, the detection gets better than before. Because the wavelet modulus maxima denoising method adopted in this article aims to eliminate the effect of pulse noise, the pulse noise can always be removed without considering its mutation value or anomaly intensity.

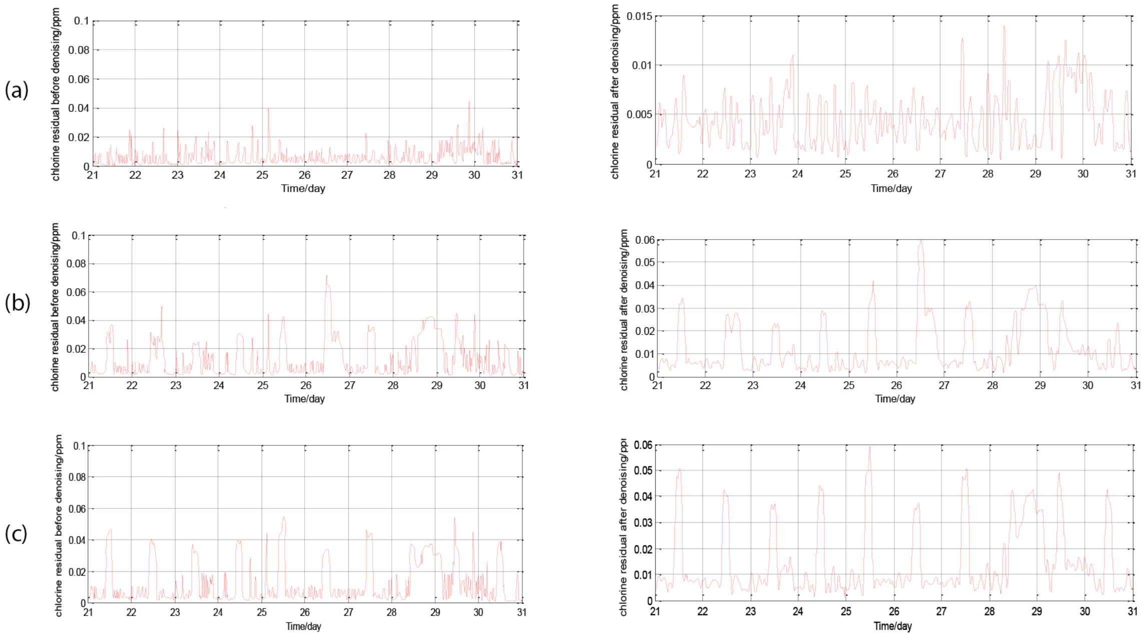

Fig. 4. Comparisons of residual series before and after denoising. (a) Chlorine residual series comparison before and after denoising when $A = 1$. (b) Chlorine residual series comparison before and after denoising when $A = 1.5$. (c) Chlorine residual series comparison before and after denoising when $A = 2$.

Table 1
Anomaly detection capability comparison of two algorithms

Indexes Intensity	Detection rate		False positive rate		ROC areas	
	Traditional algorithm	Improved algorithm	Traditional algorithm	Improved algorithm	Traditional algorithm	Improved algorithm
1	0.0168	0	0.0093	0	0.5281	0.5216
1.5	0.7543	0.7501	0.1419	0.1234	0.8466	0.8839
2	0.9168	0.9510	0.080	0.0693	0.9325	0.9812

5. Conclusion and future work

Rapid and accurate detection of water quality anomalies under natural or man-made events is of great significance to the protection of water environment and the protection of public health. For the problems of unsatisfactory anomaly detection performance under large water quality fluctuation of background data, a water quality detection approach based on RBF neural network and wavelet analysis is presented in this article. RBF is introduced to predict the water quality, and the residual series acquired by comparison between predicted value and actual value are denoised by wavelet. Then, the water quality status is determined by comparison between the distance deviated from the origin every moment and specific threshold. On-line monitoring ammonia nitrogen value of certain city water reservoir is taken as experimental object and the results indicate our algorithm has higher anomaly detection rate and lower false positive rate.

Abnormal water quality classification method in this paper is used to distinguish baseline variation and abnormal events. The final object of research is to further improve the performance of anomaly detection methods. Therefore, we can study the classification of anomalies caused by different pollutants in future work, to determine the types of contaminants, take effective measures in time and reduce the loss caused by pollution.

Acknowledgment

This study was supported by the scientific research project of Liaoning Provincial Education Department (WJZ2016016) and Liaoning Provincial Education Department (WJZ2016017).

References

[1] K. Miyabe, N. Taniguchi, A. Imura, Y. Tezuka, Kinetic study of the hydrodechlorination of trichloroethene in water using a platinum catalyst and hydrazine, Water Environ. Res., 75 (2003) 472–477.

[2] T. Wang, S. Xu, Dynamic successive assessment method of water environment carrying capacity and its application, Ecol. Indic., 52 (2015) 134–146.

[3] Y. Liu, H. Long, T. Li, S. Tu, Land use transitions and their effects on water environment in Huang-Huai-Hai Plain, China, Land Use Policy, 47 (2015) 293–301.

[4] I. Rojek, Models for better environmental intelligent management within water supply systems, Water Resour. Manage., 28 (2014) 3875–3890.

[5] D. Hou, X. Song, G. Zhang, H. Zhang, H. Loaiciga, An early warning and control system for urban, drinking water quality protection: China's experience, Environ. Sci. Pollut. Res. Int., 20 (2013) 4496–4508.

[6] Y. Wang, W. Zhang, B.A. Engel, H. Peng, L. Theller, Y. Shi, S. Hu, A fast mobile early warning system for water quality emergency risk in ungauged river basins, Environ. Modell. Software, 73 (2015) 76–89.

[7] H. Che, S. Liu, Contaminant detection using multiple conventional water quality sensors in an early warning system, Procedia Eng., 89 (2014) 479–487.

[8] C. Vélez, L. Alfonso, A.S. Torres, A. Galvis, G. Sepúlveda, Centinela: an early warning system for the water quality of the Cauca River, J. Hydroinf., 16 (2014) 1409–1424.

[9] V. Noel, H. Chepfer, C. Hoareau, M. Reverdy, G. Cesana, Effects of solar activity and geomagnetic field on noise in CALIOP profiles above the South Atlantic Anomaly, Atmos. Meas. Tech. Discuss., 6 (2013) 8589–8602.

[10] J.F. Zhou, C.F. Wang, D.F. Liu, J.W. Xiang, P. Zhao, F.L. Kou, Hydrology and water quality survey near the Gezhouba Dam in the winter, Adv. Mater. Res., 731 (2013) 3256–3261.

[11] R. Schurer, A. Janssen, L. Villacorte, M. Kennedy, Performance of ultrafiltration and coagulation in an UF-RO seawater desalination demonstration plant, Desal. Wat. Treat., 42 (2012) 57–64.

[12] W.C. Wang, D.M. Xu, K.W. Chau, G.J. Lei, Assessment of river water quality based on theory of variable fuzzy sets and fuzzy binary comparison method, Water Resour. Manage., 28 (2014) 4183–4200.

[13] P.L. Mills, M.P. Duduković, Deconvolution of noisy tracer response data by a linear filtering method, AIChE J., 34 (1988) 1752–1756.

[14] P. Facco, F. Bezzo, M. Barolo, Nearest-neighbor method for the automatic maintenance of multivariate statistical soft sensors in batch processing, Ind. Eng. Chem. Res., 49 (2010) 2336–2347.

[15] A.M. Aish, H.A. Zaqoot, S.M. Abdeljawad, Artificial neural network approach for predicting reverse osmosis desalination plants performance in the Gaza Strip, Desalination, 367 (2015) 240–247.

[16] V. Jormalainen, S. Merilaita, J. Tuomi, Male choice and male-male competition in *Idotea baitica* (Crustacea, Isopoda), Ethology, 96 (2015) 46–57.

[17] M. Shokoohi, M. Tabesh, S. Nazif, M. Dini, Water quality based multi-objective optimal design of water distribution systems, Water Resour. Manage., 31 (2017) 1–16.