Desalination and Water Treatment
www.deswater.com

doi:10.5004/dwt.2018.22353

111 (2018) 125–133
April

Modeling trihalomethanes concentrations in water treatment plants using machine learning techniques

Jongkwan Park[a], Chan ho Lee[a], Kyung Hwa Cho[a], Seongho Hong[b], Young Mo Kim[c,*], Yongeun Park[d,*]

[a]*School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, Korea*
[b]*Department of Chemical Engineering, Soongsil University, 369 Sangdo-Ro, Dongjak-Gu, Seoul 156-743, Korea*
[c]*School of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), 123 Cheomdan-gwagiro, Buk-gu, Gwangju 61005, Korea, email: youngmo@gist.ac.kr (Y.M. Kim)*
[d]*School of Civil and Environmental Engineering, Konkuk University, Seoul 05029, Korea, email: yepark@konkuk.ac.kr (Y. Park)*

Received 25 November 2017; Accepted 17 January 2018

ABSTRACT

Water disinfection process in a water treatment process results in the formation of disinfection by-products (DBPs), including total trihalomethanes (TTHMs). It takes a relatively long time to estimate TTHMs concentration level in the water treatment plants; thereby it is impossible to timely control operation parameters to reduce the TTHMs concentration. Here, we developed a predictive model to quantify TTHMs concentration using conventional water quality parameters from six water treatment plants in Han River. Before the developing the model, self-organizing map (SOM) and artificial neural network (ANN) restored missing values in input and output parameters. Then, an ANN model was trained to predict TTHMs by using relevant water quality parameters investigated from Pearson correlation. Pearson Correlation test selected six significant input parameters such as temperature, algae, pre-middle chlorine, post chlorine, total chlorine, and total organic carbon. Based on five-fold jackknife cross-validation, the ANN models built using different types of input data showed different performance in training (range of R^2 from 0.62 to 0.92) and validation (range of R^2 from 0.62 and 0.80) steps. This study can be a useful tool for predicting TTHMs concentrations using conventional water quality data in drinking water treatment plants. Machine learning models can be readily developed and utilized by managers working with drinking waters.

Keywords: Trihalomethanes (THMs); Drinking water treatment plant; Han River; Machine learning technique

1. Introduction

Chlorine dosing in water treatment plants is one of essential processes to achieve water disinfection for reducing potential risk of exposure to water-borne pathogens. Despite the benefit of chlorine dosing, the reaction of free chlorine with natural organic matters (NOM) is an issue that cause the formation of disinfection by-products (DBPs) [1,2]. DBPs pose a serious threat to human health because they are carcinogenic and mutagenic to humans [3,4]. In particular, the trihalomethanes (THMs) including chloroform, bromoform, bromodichloromethane, and dibromochloromethane are the most common DBPs detected in chlorinated drinking water [5]. In 1984, the World Health Organization reported a health-based guideline value for chloroform in the first edition of the Guidelines for Drinking Water Quality [7]. Various institutions from United Kingdom, United States, Canada, and Japan contributed to evaluate the risks for human health to chemicals including DBPs in the third edition of the guidelines [7]. For example, the United States Environmental Protection

*Corresponding author.

Agency (USEPA) regulates the maximum contamination limits (MCLs) for total trihalomethanes (TTHMs) at 0.08 mg/L [8].

Monitoring and modeling THM concentrations are very important for understanding variations in THM concentrations in drinking water distribution systems. However, rapid detection of THMs is difficult because measuring THMs requires relatively long time and expertise. That issue hinders operators to take quick actions for reducing the risk of increase in THMs concentrations. Modeling approaches have been used as an alternative rapid monitoring tool and/or predictive decision-making tools. Formation of TTHMs is closely related to water quality and operational parameters such as the amount of chlorine dosing, pH, organic matter composition, temperature, existence of algal matters, and reaction time between chlorine and organic matters [5,8–10]. Past studies have tried to model THM formation using numerical and statistical methods [11–18]. However, those methods were limited to accurately predict THM concentrations due to the complex interactions between influential factors and THMs.

Artificial neural networks (ANNs) are powerful tools to reflect the complex interactions using stochastic error minimizing algorithms [19,20]. Various researchers implemented ANNs to predict the formation of THMs [21–23]. Kulkarni and Chellam [21] predicted the formation of THMs, haloacetic acids (HAAs), and total organic halide (TOX) using seven input parameters including UV_{254} absorbance, chlorination conditions, and DOC and Br^- concentrations. Lewin et al. [22] applied 15 input parameters associated with raw water quality and post water quality after clarification. Milot et al. [23] used five input parameters such as DOC, reaction time, pH, chlorine dose, and temperature.

ANN models need a large number of observed and/or experimental data to train complex patterns of data, but incomplete or missing data is a common problem in water treatment plants. In order to deal with the problem, imputation method is applied for generating missing values. Francis et al. [24] used the multiple imputation approach to understand bromine substitution reaction in a trihaloacetic acid class. Bergman et al. [25] also imputed missing input data from water quality parameters collected at multiple locations in a watershed to predict THMs formation.

Therefore, the objectives of this study was to: 1) impute missing values for input and output parameters using ANN and Self-organizing map,2) develop an ANN-based prediction model to understand site-specific effects of water quality and chlorine dose conditions on the formation of THMs in finished water at six drinking water treatment plants and 3) evaluate the performance of the models that were built in terms of different types of input dataset. Influent water quality and operation conditions of chlorine dose in water treatment plants were used as input parameters.

2. Materials and methods

2.1. Site description and data acquisition

This study was focused on six water treatment plants in east part of the Han River Watershed; Ttukdo, Youngdeungpo, Guui, Amsa, Gangbuk, and Gwangam. The six plants supply tap water to a resident in Seoul city. More details on the treatment plants are given in Table 1. Fig. 1 shows a treatment process for drinking water supplies in the six water treatment plants. The process consisted of the three stages of chlorination: (1) pre-chlorination (pre-Cl) in the receiving well, (2) intermediate-chlorination (inter-Cl) before the filtration basin, and (3) post-chlorination (post-Cl) in the pure water reservoir. Five water quality parameters such as temperature, pH, algae, total organic carbon (TOC), and *chlorophyll-a* (*Chl-a*) were monthly monitored in intake water. THMs were monitored in clear well. In addition, the total amount of chlorine dose (Total-Cl) was estimated by summing up the amount of injected chlorine in each chlorination process.

2.2. Model development

Fig. 2 describes the process for applying Artificial Neural Network (ANN) and Self-Organizing map (SOM). ANNs were trained to predict *Chl-a* and TOC using water quality data. SOM was constructed to restore missing data for TTHMs concentration. After the process,we investigated the significant parameters to predict TTHMs using Pearson correlation. Finally, ANNs were trained to predict TTHMs with the significant parameters.

2.2.1. Pearson correlation

The Pearson correlation coefficient (PCC) is a measure of the linear correlation between two variables X and Y in statistics [26]:

$$Corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma X \sigma Y} \quad (1)$$

Table 1
Information of six drinking water treatment plants

Name	Location	Purification method	Advanced treatment method	Daily capacity† (m^3/day)
Gwangam	37°31′10.96″ N, 127°10′21.98″ E	Rapid sand filtration	Ozonation and granular activated carbon	400,000
Amsa	37°33′47.86″ N, 127°08′28.96″ E			1,600,000
Guui	37°32′49.73″ N, 127°05′33.03″ E			500,000
Ttukdo	37°32′22.50″ N, 127°02′31.38″ E			350,000
Youngdeungpo	37°33′00.60″ N, 126°52′51.61″ E			600,000
Gangbuk	37°35′44.49″ N, 127°11′10.36″ E			1,000,000

†Daily capacity was collected from http://data.seoul.go.kr/openinf/fileview.jsp?infId=OA-12880&t Menu=11.

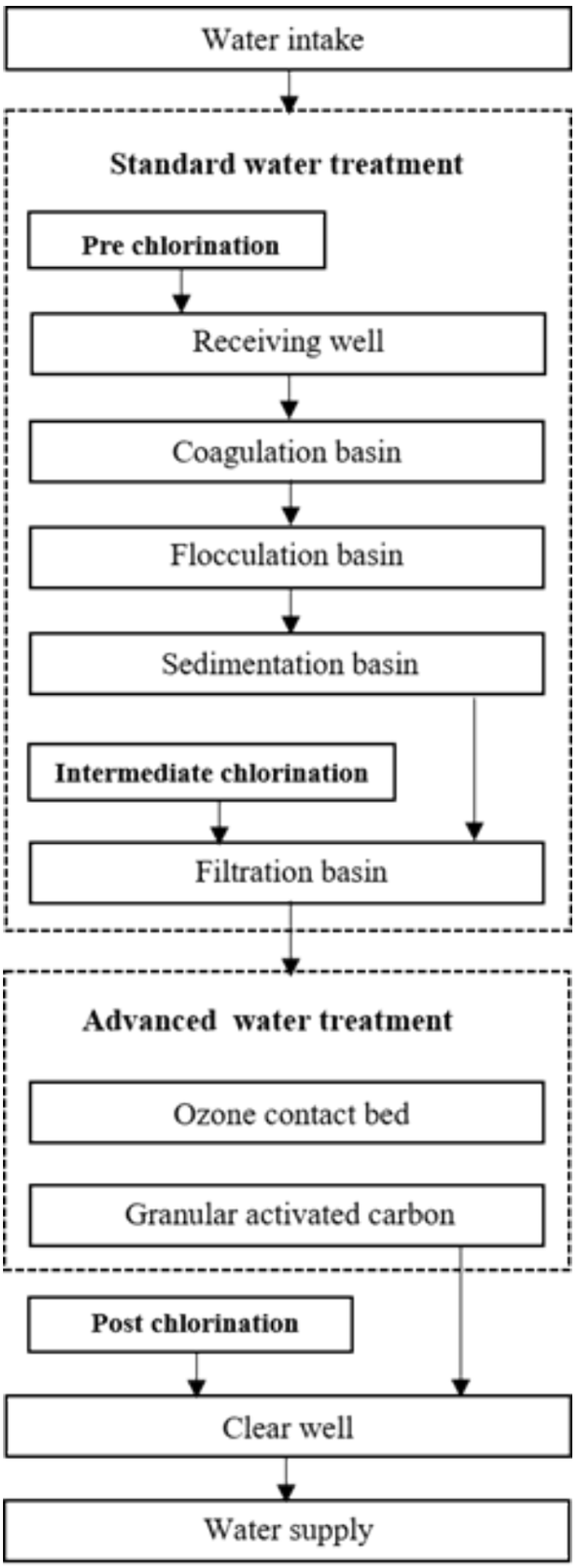

Fig. 1. Description of the water treatment process in the water treatment plants.

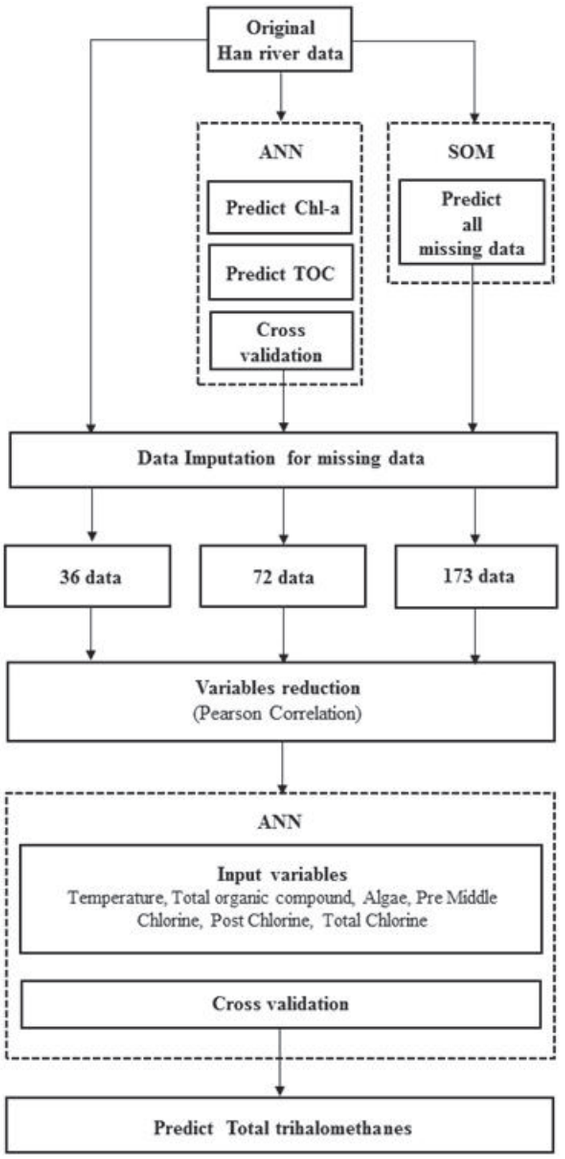

Fig. 2. Logical flow for predicting total trihalomethanes using two machine learning methods.

where *Cov* is the covariance between *X* and *Y*, σ*X* is the standard deviation of *X*, σ*Y* is the standard deviation of *Y*, and *Corr* is the Pearson correlation between *X* and *Y*. PCC has a value between +1 and −1; 1, 0, and −1 mean total positive linear correlation, no linear correlation, and total negative linear correlation, respectively [26]. PCC is a useful method to investigate the relationship between water quality and TTHMs. In this study, we used PCC to determine parameters, which have little relationship with TTHMs and then to reduce the number of input parameters for ANN model to predict TTHMs. *X* includes temperature, pH, algae, TOC, *Chl-a*, pre-Cl, inter-Cl, and post-Cl, and *Y* is TTHMs.

2.2.2. Artificial neural network

Artificial neural network, inspired by the biological neural networks, has performed to predict output by pattern recognition and complex processes [19,27–29]. An ANN consists of three layer (i.e., input, hidden, and output layers) and neuron (nodes) that are linked by weights in each layer [30]. Activation functions adjust signals between layers and then transfer signals to the next layer.

$$y = f\sum_{N}^{i=1} w_i \times x_i + b \tag{2}$$

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \tag{3}$$

where x_i is input data in the network, y is output data in the network, N is the number of neurons in the input vector, w_i is the connection weight between input and output, f is the transfer function, and b is the bias term, $f(x)$ is the tangent sigmoid function.

ANN applied a back-propagation algorithm to update the weight and bias in each neuron [31]. Weight and bias have random values in initial conditions and are updated by a back propagation step. A back-propagation algorithm has ability to adjust the learning rate by updating the learning rate's parameter. Eqs. (4) and (5) illustrate the back propagation step that used gradient descent with a momentum algorithm [32]:

$$\triangle w_i^{j+1} = -c \times \frac{\delta E}{\delta w_i^j}\left(w_i^j\right) + a \times \Delta w_i^j \tag{4}$$

$$w_i^{j+1} = w_i^j + \Delta w_i^{j+1} \tag{5}$$

where j is the iteration number, c is the learning rate, and a is the momentum constant. ANN stops repeating the back-propagation step when the error has smaller than a termination criterion for error goal or an iteration number has larger than maximum iterations.

To predict THMs concentration, many researchers have used various input water quality parameters including temperature, pH, TOC, and amount of chlorine dose to predict trihalomethane in a drinking water treatment plant. For instance, Abdullah et al. [33] considered pH, turbidity, ammonia, TOC, temperature, Cl_2 dosage, Cl_2 residue, and THM as input. In this study, we used two ANN models; the first ANN imputed TOC and *Chl-a* using seven input data such as TTHMs, pH, algae, pre-Cl, inter-Cl, post-Cl, and total chlorine. The second ANN model was developed to predict TTHMs using significant input parameters among temperature, algae, pre-Cl, inter-Cl, post-Cl, total chlorine, TOC.

2.2.3. Self-organizing map

Self-organizing map is one of unsupervised machine learning methods, which is based on a neural network to adjust weight values for matching input vector in training datasets [34]. SOM uses a competitive learning method, called winner-take-all. Winner-take-all declares a node, which is closest with input vector, as winner. And weights are updated to take node values nearby input vector. SOM repeats this process to relate output node with patterns of input data set. At the initial step, SOM initializes each node's weights with a random number between 0 and 1, and choose random input vector from training dataset. SOM identifies the best matching unit (BMU) that its weights are most similar to the input vector by Euclidean distance formula [29,35]:

$$c_j = \min_i \left\{\left\|w_i - x\right\|\right\} \tag{6}$$

where c_j is the winner unit, w_i is the weight vector, x is input vector, and $\|\cdot\|$ is the distance measure, typically Euclidian distance.

The SOM update rule for the weight vector of unit is [36]:

$$w_i(t+1) = w_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot \left(x(t) - w_i(t)\right) \tag{7}$$

where t denotes time, the x(t) is an input vector randomly drawn from the input data set at time t, h_{ci} (t) is the neighborhood kernel around the winner unit c, $\alpha(t)$ and is the learning rate at time t.

SOM followed above rule to update for weight vectors. SOM iteratively develops weight vectors of BMU and its neighboring units by using neighborhood function to minimize the distance between them. The Gaussian distribution is applied to update the weights, as follow [29,37]:

$$w_i(t+1) = \frac{\sum_{j=1}^{N} h_{c_j,i}(t) \cdot x_j}{\sum_{j=1}^{N} h_{c_j,i}(t)} \tag{8}$$

where $h_{cj,i}(t)$ is the neighborhood function around winner unit c_j.

We used two indicators, which are Quantization error and topological error, to check the quality of maps. The quantization error is given by [38]:

$$\varepsilon_q = \frac{1}{N}\sum_{i=1}^{N}\left\|X_i - M_b\right\| \tag{9}$$

where X_i is input data vector, M_b is best matching reference vector, and N is the total number of input samples.

The topological error is given by [38]:

$$\varepsilon_t = \frac{1}{N}\sum_{n}^{K=1}\eta\left(X_k\right) \tag{10}$$

where X_k is input data vector, η is matching unit, and N is the total number of input samples.

In this study, SOM was used to restore all missing TTHMs values using input parameters such as TTHMs, temperature, pH, TOC, algae, *Chl-a*, pre-Cl, inter-Cl, post-Cl, and total Cl.

2.3. Model performance

We implemented cross validation to verify the performance of ANNs [39]. Five-fold jackknife cross-validation was conducted for comparing the relative performance between ANN models to predict TTHMs. One of fifth of the data was implemented to test the models and four of fifth of the data was used to train the network of the models. In imputing the missing data, the models used the data that

were randomly shuffled datasets divided into two groups; one group was randomly selected 60% of data for training and the other group was the rest of data for validation.

Coefficient of determination (R^2) and root mean square error (RMSE) are used to estimate accuracy of models [40]:

$$R^2 = \frac{1}{N}\sum_{N}^{I=1}\left(Y_i^{obs} - Y_i^{sim}\right)^2 \quad (11)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{N}^{I=1}\left(Y_i^{obs} - Y_i^{sim}\right)^2} \quad (12)$$

where Y_i^{obs} is the i th observed data, Y_i^{sim} is i th simulated data, and N is the total number of observation.

3. Results and discussion

3.1. Missing data imputation

Input and output data consisted of nine water quality parameters collected from intake water and clear well. Among eight input parameters, a large number of TOC and *Chl-a* had missing data. ANN was constructed to restore missing data of TOC and *Chl-a*, which were used to develop prediction model for TTHMs. The ANN model used seven input parameters such as TTHMs, pH, algae, pre-Cl, inter-Cl, post Cl, and total chlorine and two output parameters, i.e., TOC and *Chl-a*. Five hidden neurons was used in the first and second layers, respectively. Activation functions between layers were set to be tangent sigmoid and back propagation function was set to be traingdm; the iteration number was set to be 200,000. 20 of sample data randomly chosen for training and 16 of sample data left were used for validating. The performance of the ANN model for restoring two parameters was satisfactory with NSE values ranging 0.90 to 0.99 and R^2 values ranging 0.94 to 0.99 for training and validation steps (Fig. 3). In Fig. 3d, the ANN model overestimated *Chl-a* when higher than 25 mg/L. Using the developed TOC and *Chl-a* models, we restored 35 missing data.

Like TOC and *Chl-a*, approximately 3.5% of TTHMs samples had missing data. SOM was applied to restore the missing TTHMs data. Final quantization error was 0.883 and final topographic error was 0.017 as SOM result. The performance of the SOM model was satisfactory to restore the missing data (Fig. 4).

3.2. Determination of the input parameters

Correlations between input parameters and TTHMs were investigated using Pearson Correlation Test to select the appropriate input parameters for constructing the ANN model. Table 2 presents the result of Pearson Correlation Test. Based on the significance values, TOC, total chlorine, and pre-Cl-middle chlorine had significant positive correlations with TTHMs ($p < 0.05$ in Table 2). Temperature and algae concentrations were also considered as input parameters because p-values were relatively low even though the p values were larger than 0.05. Oliver and Shindler [41] reported that algae may be potential THM precursors. Water temperature had a positive effect on increase in THM formation potential due to increasing the reaction rate between chlorine and TOC [8,42]). Saidan et al. [10] documented that the formation of THMs was affected by increase in the residence time, temperature, pH, and free chlorine and TOC concentrations. On the other hand, pH and *Chl-a* had relatively high p values ($p > 0.5$), compared to other parameters. Unlike high p-value for pH, past studies documented that pH had an influence on THM formation [9,43]. That is, the two parameters had little effects on the formation of TTHMs in this study. Therefore, six input parameters were determined for input parameters to the ANN model.

3.3. Model performance

Three types of data sets were prepared to evaluate the performance of ANN models for predicting TTHMs. One was the original data sets that included the missing data in both input parameters and TTHMs. Another was the revised data sets that included the restored data in TTHMs. The other were the revised data sets that included the restored data in input parameters. The optimized ANN model parameters that were obtained from five-fold jackknife cross validation are presented in Table 3. Different number of input data were used to develop the models in terms of types of input data (see the sixth row in Table 3), whereas same input parameters such as temperature,algae, pre-middle chlorine, post chlorine, total chlorine, and TOC were used to predict TTHM concentrations (see the eighth row in Table 3).

Based on comparison of the performance, the ANN models showed different R^2 and RMSE values in terms of types of input dataset in the training and validation steps (Table 4). The ANN models (0.918 and 0.876 of R^2 values) using imputing missing TTHMs and TOC data showed better training performance than that (0.617 of R^2 value) without imputation. However, the model (0.796 of R^2 value) using imputing TTHMs showed better validation performance than other two models (0.689 and 0.627 of R^2 values). Fig. 5 compares the one-to-one relationships between the predicted and observed TTHMs concentrations. The TTHMs concentrations were obtained from three models that were trained with optimized model parameters using whole datasets in each type of input. Chaib and Moschandreas [11] modeled the daily variations in volatile THM by-products such as chloroform, bromodichloromethane, and bromoform in drinking water using Box-Jenkins methods and the performance using an adjusted R^2 ranged from 0.80 to 0.94. Francis et al. [13] simulated DBPs such as THM, trihaloacetic acids, dihaloacetic acid and dihalacetonitrile using Bayesian statistical modeling. Their models had 0.81 of correlation coefficient between the predicted and observed bromine incorporation fraction, which was used as alternative for the THM class. Hong et al. [14] applied multiple linear regressions (MLRs) to predict TTHM, total concentrations of bromated THMs, chloroform, and bromodichloromethane concentrations in a river and R^2 values of the models ranged from 0.85 to 0.95; the R^2 value of the TTHM model was 0.90. Rodriguez and Sérodes [15] reported that the three tyeps of MLR model to predict the seasonal variation in THM showed R^2 values of 0.69, 0.92, and 0.52. Rodriguez et al. [16] modeled THM levels using three types of regression and R^2 values for linear, polynomial, and logarithmic models ranged from 0.43 to 0.84, 0.47

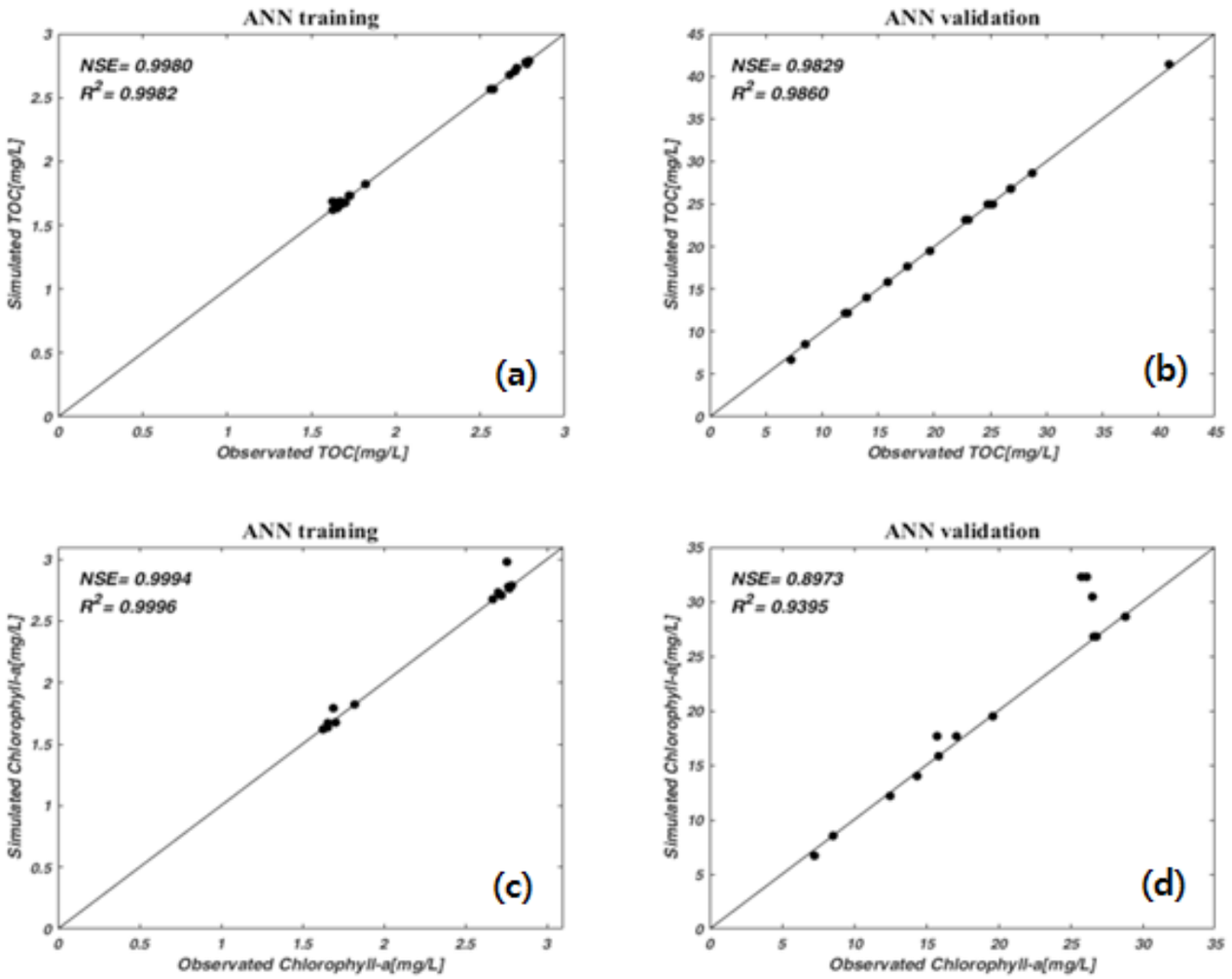

Fig. 3. Comparison between the predicted and observed TOC and Chl-a restored by ANN; (a) and (b) indicate TOC in training and validation steps; (c) and (d) indicate *Chl-a* in training and validation steps.

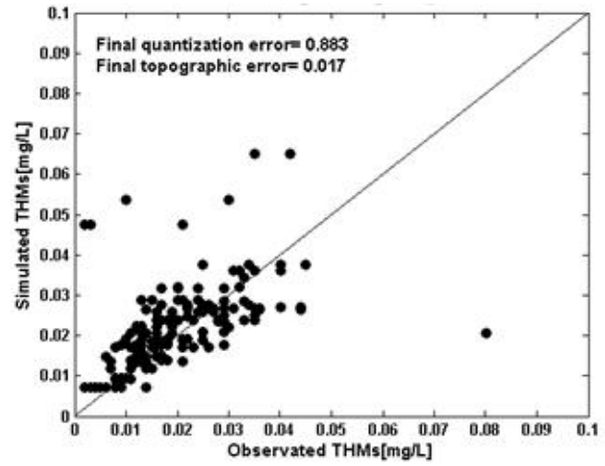

Fig. 4. Comparison between the predicted and observed TTHMs restored by self-organizing map.

Table 2
Results of Pearson correlation between input variables and TTHMs

Variable	Significance value (p)	Correlation coefficient
Temperature	0.202	0.049
pH	0.536	0.012
Total organic compound	0.034*	0.129
Algae	0.262	0.038
Chlorophyll-a	0.862	0.001
Pre-middle chlorine	0.003*	0.243
Post chlorine	0.203	0.049
Total chlorine	0.002*	0.260

The * sign for the p values indicates that the significance values are less than 0.05.

Table 3
Optimized ANN model parameters obtained from five-fold jackknife cross-validation and input data for predicting TTHMs

	Parameter	Original data with missing values	Original data with imputing missing TTHMs† data	Original data with imputing missing TOC† data
Optimized parameter	Learning rate	0.313 ± 0.189	0.312 ±0.128	0.498 ± 0.176
	Momentum constant	0.413 ± 0.165	0.597 ± 0.185	0.453 ± 0.200
	Number of hidden neurons	3 ± 0.3	8 ± 2.0	4 ± 1.5
	Activation function‡ 1 and 2	Tansig–Purelin	Logsig–Tansig	Logsig–Tansig
Input data	Number of input data	35	173	72
	Number of training and validation data	28, 7	139, 34	58, 14
	Input parameter	temperature, algae, pre-middle chlorine, post chlorine, total chlorine, and TOC		

†Missing TTHM and TOC data were restored by SOM and ANN, respectively.
‡Activation function presented the most frequently selected active function between layers; function 1 between the input layer and hidden layer; function 2 between the hidden layer and output layer.
The "±" sign separates the mean value and standard error obtained from the cross-validation.

Table 4
Comparison of performance of ANN models for predicting the TTHMs concentrations using the five-fold jackknife cross-validation

Types of input data	R^2		RMSE	
	training	validation	training	validation
Original data with missing values	0.617 ± 0.049	0.689 ± 0.101	0.0101 ± 0.0002	0.0095 ± 0.0032
Original data with imputing missing TTHMs data	0.918± 0.053	0.796 ± 0.106	0.0201 ± 0.0075	0.1640 ± 0.1337
Original data with imputing missing TOC data	0.876 ± 0.017	0.627 ± 0.061	0.0057 ± 0.0009	0.0095 ± 0.0041

The "±" sign separates the mean value and standard error obtained from the cross-validation.

to 0.85, and 0.52 to 0.87, respectively. Uyak et al. [17] develop logarithmic regression model to predict THMs using TOC, pH, temperature, and the amount of chlorine dose and R^2 value of the model ranged from 0.98 to 0.99. Overall, the performance of our models was found to be satisfactory as compared with that of literature (Table 4 and Fig. 5).

4. Conclusions

TTHMs, which are by-product of Chlorine disinfection process, are carcinogenic and mutagenic to human. Modeling TTHMs concentrations is important for understanding variations in TTHMs in drinking water treatment plants. This study proposed machine learning methods to predict formation of TTHMs in drinking water treatment plants using conventional water quality parameters. Self-organizing map (SOM) and artificial neural network (ANN) models were implemented to impute missing values for input and output parameters. Another ANN was used to develop a model for predicting TTHMs using important input parameters that were selected based on significance values from Pearson Correlation analysis. The major conclusions are as follows:

- SOM and ANN were practical to generate imputation data to improve model accuracy. The final quantization error of SOM was 0.883 and final topographic error was 0.017. The restored data for TOC and *Chl-a* from the trained ANN model showed high NSE and R^2 values.
- Based on Pearson Correlation analysis, significant parameters relating to THM formation were selected as temperature, algae, pre-middle chlorine, post chlorine, total chlorine, and total organic carbon.
- The ANN model was trained using different types of input data such as original data with missing values, with imputing TTHMs data by SOM, and with imputing TOC data by ANN. The performance was different among three ANN models in the training and validation steps. The ANN model using restored missing data for TTHMs showed better validation performance than that using original data with missing values and with imputing TOC data.

This study provided useful tools for reliably predicting TTHMs concentrations using available explanatory parameters in drinking water treatment plants. It is expected that machine learning models can be readily developed and utilized by managers working with drinking waters.

Acknowledgments

This work was supported by the 2017 Research Fund(1.170052.01) of UNIST (Ulsan National Institute of Science and Technology).

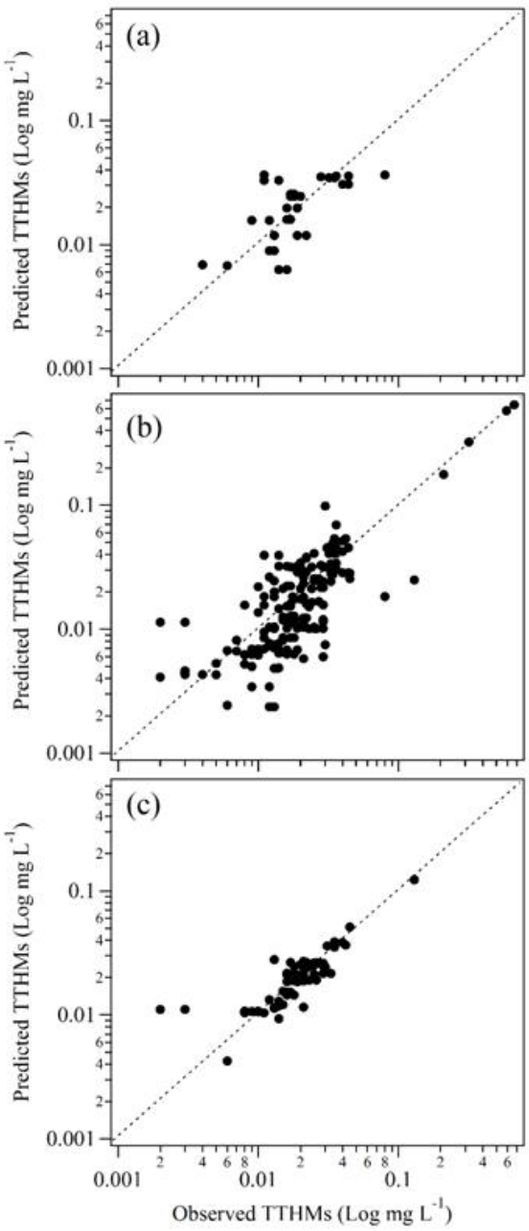

Fig. 5. Comparison between observed and predicted total trihalomethanes: (a), (b), and (c) indicate results obtained from the ANN models that were trained using original data with missing values, original data with imputing missing TTHMs by the SOM method, and original data with imputing missing TOC bythe ANN method, respectively.

References

[1] B. Kwon, S. Lee, J. Cho, H. Ahn, D. Lee, H.S. Shin, Biodegradability, DBP formation, and membrane fouling potential of natural organic matter: Characterization and controllability, Environ. Sci. Technol., 39 (2005) 732–739.
[2] D. Golea, A. Upton, P. Jarvis, G. Moore, S. Sutherland, S. Parsons, S. Judd, THM and HAA formation from NOM in raw and treated surface waters, Water Res., 112 (2017) 226–235.
[3] S. Kanitz, Y. Franco, V. Patrone, M. Caltabellotta, E. Raffo, C. Riggi, D. Timitilli, G. Ravera, Association between drinking water disinfection and somatic parameters at birth, Environ. Health Perspect., 104 (1996) 516.
[4] J. Zavaleta, F. Hauchman, M. Cox, Epidemiology and toxicology of disinfection by-products. Formation and control of disinfection by-products in drinking water, (1999) 95–117.
[5] R. Sadiq, M.J. Rodriguez, Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review, Sci. Total Environ., 321 (2004) 21–46.
[6] World Health Organization (WHO). (2004). Guidelines for drinking-water quality (Vol. 1). World Health Organization.
[7] USEPA (2006) National Primary Drinking Water standards.
[8] B. Ramavandi, S. Farjadfard, M. Ardjmand, S. Dobaradaran, Effect of water quality and operational parameters on trihalomethanes formation potential in Dez River water, Iran. Water Resour. Res., 11 (2015) 1–12.
[9] S. Navalon, M. Alvaro, H. Garcia, Carbohydrates as trihalomethanes precursors. Influence of pH and the presence of Cl- and Br- on trihalomethane formation potential, Water Res., 42 (2008) 3990–4000.
[10] M. Saidan, K. Rawajfeh, M. Fayyad, Investigation of factors affecting THMs formation in drinking water, Am. J. Environ. Eng., 3 (2013) 207–212.
[11] E. Chaib, D. Moschandreas, Modeling daily variation of trihalomethane compounds in drinking water system, Houston, Texas. J. Hazard. Mater., 151 (2008) 662–668.
[12] R.M. Clark, M. Sivaganesan, Predicting chlorine residuals and formation of TTHMs in drinking water, J. Environ. Eng., 124 (1998) 1203–1210.
[13] R.A. Francis, J.M. Van Briesen, M.J. Small, Bayesian statistical modeling of disinfection byproduct (DBP) bromine incorporation in the ICR database, Environ. Sci. Technol., 44 (2010) 1232–1239.
[14] H.C. Hong, Y. Liang, B.P. Han, A. Mazumder, M.H. Wong, Modeling of trihalomethane (THM) formation via chlorination of the water from Dongjiang River (source water for Hong Kong's drinking water), Sci. Total Environ., 385 (2007) 48–54.
[15] M.J. Rodriguez, J.-B. Sérodes, Spatial and temporal evolution of trihalomethanes in three water distribution systems, Water Res., 35 (2001) 1572–1586.
[16] M.J. Rodriguez, Y. Vinette, J.B. Sérodes, C. Bouchard, Trihalomethanes in drinking water of greater Québec region (Canada): occurrence, variations and modelling, Environ. Monit. Assess., 89 (2003) 69–93.
[17] V. Uyak, I. Toroz, S. Meric, Monitoring and modeling of trihalomethanes (THMs) for a water treatment plant in Istanbul, Desalination, 176 (2005) 91–101.
[18] G. Zhang, B. Lin, R.A. Falconer, Modelling disinfection by-products in contact tanks, J. Hydroinform., 2 (2000) 123–132.
[19] K.H. Cho, S. Sthiannopkao, Y.A. Pachepsky, K.W. Kim, J.H. Kim, Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network, Water Res., 45 (2011) 5535–5544.
[20] Y. Park, K.H. Cho, J. Park, S.M. Cha, J.H. Kim, Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea, Sci. Total Environ., 502 (2015) 31–41.
[21] P. Kulkarni, S. Chellam, Disinfection by-product formation following chlorination of drinking water: Artificial neural network models and changes in speciation with treatment, Sci. Total Environ., 408 (2010) 4202–4210.
[22] N. Lewin, Q. Zhang, L. Chu, R. Shariff, Predicting total trihalomethane formation in finished water using artificial neural networks, J. Environ. Eng. Sci., 3 (2004) S35–S43.
[23] J. Milot, M.J. Rodriguez, J.B. Sérodes, Contribution of neural networks for modeling trihalomethanes occurrence in drinking water, J. Water Resour. Plan. Manage., 128 (2002) 370–376.
[24] R.A. Francis, M.J. Small, J.M. Van Briesen, Multivariate distributions of disinfection by-products in chlorinated drinking water, Water Res., 43 (2009) 3453–3468.

[25] L.E. Bergman, J.M. Wilson, M.J. Small, J.M. Van Briesen, Application of classification trees for predicting disinfection by-product formation targets from source water characteristics, Environ. Eng. Sci., 33 (2016) 455–470.
[26] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient. In Noise reduction in speech processing. Springer Berlin Heidelberg (2009) 1–4.
[27] W.S. Mcculloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity (Reprinted from Bulletin of Mathematical Biophysics, Vol 5, Pg 115–133, 1943). Bull. Math. Biol., 52 (1990) 99–115.
[28] H. Guo, K. Jeong, J. Lim, J. Jo, Y.M. Kim, J.P. Park, J.H. Kim, K.H. Cho, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, J. Environ. Sci., 32 (2015) 90–101.
[29] M. Kim, S. Baek, M. Ligaray, J. Pyo, M. Park, K.H. Cho, Comparative studies of different imputation methods for recovering stream flow observation, Water, 7 (2015) 6847–6860.
[30] M.K. Gill, T. Asefa, Y. Kaheil, M. Mckee, Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique, Water Resour. Res., 43 (2007).
[31] M.B. Shukla, R. Kok, S.O. Prasher, G. Clark, R. Lacroix, Use of artificial neural networks in transient drainage design, Trans. ASAE, 39 (1996) 119–124.
[32] R. Rojas, Neural networks: a systematic introduction. Springer Science & Business Media (2013).
[33] M.P. Abdullah, C.H. Yew, M.S. bin Ramli, Formation, modeling and validation of trihalomethanes (THM) in Malaysian drinking water: a case study in the districts of Tampin, Negeri Sembilan and SabakBernam, Selangor, Malaysia, Water Res., 37 (2003) 4637–4644.
[34] S.J. Ki, J.H. Kang, S.W. Lee, Y.S. Lee, K.H. Cho, K.G. An, J.H. Kim, Advancing assessment and design of storm water monitoring programs using a self-organizing map: Characterization of trace metal concentration profiles in storm water runoff, Water Res., 45 (2011) 4183–4197.
[35] R.Q. Huang, L.F. Xi, X.L. Li, C.R. Liu, H. Qiu, J. Lee, Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods, Mech. Syst. Signal Process, 21 (2007) 193–207.
[36] T. Kohonen, T. Honkela, Kohonen network, Scholarpedia, 2 (2007) 1568.
[37] A.M. Kalteh, P. Hiorth, R. Bemdtsson, Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application, Environ. Modell. Softw., 23 (2008) 835–845.
[38] M.A. Malek, S.M. Shamsuddin, Restoration of hydrological data in the presence of missing data via Kohonen Self Organizing Maps. New Trends in Technologies. In Tech (2010).
[39] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, In Ijcai 14 (1995) 1137–1145.
[40] D.N. Moriasi, J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, T.L. Veith, Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Trans. ASAB, 50 (2007) 885–900.
[41] B.G. Oliver, D.B. Shindler, Trihalomethanes from the chlorination of aquatic algae, Environ. Sci. Technol., 14 (1980) 1502–1505.
[42] D. Baytak, A. Sofuoglu, F. Inal, S.C. Sofuoglu, Seasonal variation in drinking water concentrations of disinfection by-products in IZMIR and associated human health risks, Sci. Total Environ., 407 (2008) 286–296.
[43] M.A. El-Dib, R.K. Ali, THMs formation during chlorination of raw Nile river water, Water Res., 29 (1995) 375–378.