Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2021.27355

227 (2021) 86–92
July

Prediction of ultrafiltration membrane fouling using statistical models in pilot and full-scale operations

Nayoung Park[a], Hyungsoo Kim[a], Yongsoo Lee[b], Yongjun Choi[c,*], Sangyoup Lee[a,*]

[a]*Department of Water Resources, Graduate School of Water Resources, Sungkyunkwan University, 2066 Seobu-ro, Jangan-Gu, Suwon, Gyeonggi-do 440-746, Republic of Korea, Tel. +82-31-290-7542; Fax: +82-31-290-7549; email: sangyouplee@skku.edu*
[b]*Department of Civil and Environmental, Engineering Graduate School of Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea*
[c]*School of Civil and Environmental Engineering, Kookmin University, Jeongneung-Dong, Seongbuk-Gu, Seoul 136-702, Republic of Korea, Tel. +82-2-910-4529; Fax: +82-2-910-4939; email: choiyj1041@gmail.com*

Received 4 January 2021; Accepted 1 April 2021

ABSTRACT

Prediction of ultrafiltration membrane fouling using statistical models has been investigated. Statistical models employed in this study include artificial neural network (ANN), genetic programming (GP) and M5P tree model. Data obtained from pilot-scale (A plant) and full-scale (B plant) membrane plants were used for training and testing models. Fouling prediction by the classic Hermia model was also carried out for comparison with the statistical models. The Hermia model is used for simple estimating membrane fouling by data fitting but can provide information on the cause of membrane fouling according to the fitting trend. On the other hand, the statistical models can be used to predict the actual degree of membrane fouling rather than simple data fitting; however, these models do not provide information on the causes of membrane fouling. Therefore, complementary studies are possible by using these two types of models together. The ratio of the number of training and test data was varied to be 8:2, 6:4, 4:6, and 2:8 for prediction error control. As a result of applying the Hermia model, the ratio of training data to test data can be predicted up to around 8:2. Reliable predictions have been obtained up to the ratio of 4:6 in the ANN model, 6:4 in the GP model and 4:6 in the M5P tree model. Except for the summer period where the corrected trans-membrane pressure (TMP) at 25°C was unstable (in the full-scale plant B), the reliable prediction was obtained up to the ratio of 2:8 for the ANN model, 4:6 for the GP model and 6:4 for the M5P model. It has been demonstrated that the statistical models can make acceptable fouling predictions despite a small number of training data in both the pilot-scale and full-scale plant. In addition, the time for membrane cleaning can be scheduled in advance as the models also predict the proper cleaning time in combination with fouling prediction.

Keywords: Membrane fouling; Hermia model; Artificial neural network; Genetic programing; M5P tree model

1. Introduction

As regulations on the quality of drinking water become even more strict, demands for obtaining safe water quality, and efforts to secure stable water resources, are increasing. Various research on water purification processes, in order to secure more stable water quality, is in progress. In addition, the membrane filtration process, which is one of the current water treatment processes, is expanding in particulate matter and also in pathogenic microorganisms, such as Giardia and Cryptosporidium. The membrane filtration process can obtain stable membrane filtration water

* Corresponding authors.

quality, regardless of raw water and turbidity changes, compared to the existing water treatment process. In addition, since only a small space is required, it is not limited by geographical features and has the advantage of easy maintenance. However, membrane filtration involves complex interactions among the membrane surface, processing conditions, and effluents under treatment. These interactions can often affect each other, which results in a multifaceted effect on the surface of the membrane. This is known as the phenomenon of membrane fouling [1]. Therefore, in order to improve the stability and efficiency of operation in the water treatment process using a membrane, the most important consideration is to minimize all membrane fouling. Membrane fouling is a contaminant that occurs on the surface or the inside of the membrane. The fouling appears as contaminants, such as natural organic matter, colloidal and particulate matter, contained in raw water and separated by the membrane. Various studies are being conducted, such as identifying mechanisms for membrane fouling, and cleaning methods. In fact, there have been numerous studies on fouling at the lab scale, but relatively few studies have been conducted in the pilot-scale and full-scale system. The operation of the membrane filtration plant in the field compares the period of reaching the critical pressure to find data about the aging membrane or to perform clean-in-place (CIP) washing when known that the chemical cleaning efficiency is lower than before. However, this method creates problems such as decreased production efficiency, and also increased chemical cleaning costs.

Commonly, the ratio of training data to test data is the most common and accurate for 80% of training data and 20% of test data. Notably, reducing the ratio of training data does not significantly reduce accuracy. Therefore, in this study, after dividing the data by the CIP period, the training data and the test data were set at a ratio of 8:2, 6:4, 4:6, and 2:8, and presented as the ratio of the training data to the test data from these results of the minimal predictable level.

It can be expected to find the minimum number of data required to predict the CIP washing time.

2. Theory

2.1. Mathematical model

This blocking filtration model is mainly used to study porous membrane filtration contaminants of microorganisms, proteins, and natural organic matter as well as natural water and wastewater [2]. The blocking model equation for constant flow filtration can be obtained according to a procedure similar to that used for static pressure filtration [3]. The three membrane fouling models applied to the constant flow filtration system are as follows: the pore blocking, the pore constriction, and the cake formation models (see Table 1).

2.2. Artificial neural network model

The objective of a neural network is to compute output values from input values by internal calculations. Neurons are processing elements that carry out simple computations from a vector of input values. Neural networks are organized in several layers. Each layer is fully connected to the next one. Inputs are represented by χ_1, χ_2, and χ_i and the output is represented by y_j. A parameter ω (called weight) is associated with each connection between two cells or neurons [4]. Every input is multiplied by its corresponding weight, and the node uses the summation of these weighted inputs ($\omega_{ij} \times \chi_i$). Next, the weighted inputs are added to a threshold value (θ_j) [5]. The processes mentioned above can be described by the following equation:

$$y_j = \sum_{i=1}^{n}\left(\omega_{ij} \times \chi_i\right) + \theta_j \tag{1}$$

In most cases, an artificial neural network (ANNs) is an adaptive system that changes its structure based on the external or internal information which flows through the network

Table 1
Fouling mechanism in Hermia model

Fouling type	Equation	Constant parameter
Pore blocking	$1-\left(\frac{P}{P_0}\right)^{-1} = \alpha t$	α
Pore constriction	$1-\left(\frac{P}{P_0}\right)^{-0.5} = \beta t$	β
Cake formation	$P - P_0 = \gamma t$	γ

Note: P: transmembrane pressure; P_0: initial transmembrane pressure; t: operating time; α: pore blocking model parameter; β: pore constriction model parameter; γ: cake formation model parameter.

Table 2
Membrane specification in A pilot plant

Item	Content
Company	Hyosung
Membrane type	UF
Module type	Hollow fiber
Material	PVDF
Pore size	0.03 μm
Surface area	72 m^2

Table 3
Membrane specification in B full-scale plant

Item	Content
Company	Toray Industries
Membrane type	UF
Module type	Hollow fiber
Material	PVDF
Pore size	0.01 μm
Surface area	72 m^2

during the learning phase [6]. Also, this method does not need an explicit formulation of the physical relationship of the problem but does need to include available theoretical or empirical knowledge of the physics process [4].

2.3. Genetic programming model

The fundamental idea is that of emulating the Darwinian theory of evolution, where a population is progressively improved by selectively discarding the not-so-fit population, and then breeding new children from better populations [7]. The chromosomes in the genetic programming (GP) are represented in a hierarchical structure in the population. The representation of GP can be viewed as a tree-based structure composed of the function set and the terminal set. The function set is the operators, functions, or statements such as arithmetic operators ({+, −, ×, /}) or conditional statements ("If", "then"), which are available in the GP. The terminal set contains all inputs, constants, and other zero-arguments in the GP tree. Also, there are three main operators such as crossover, mutation, and reproduction, to show the procedures that help determine the (approximate) optimal generation. The operators are able to automatically discover any computer programming, mathematical functions, etc. [8].

2.4. M5P tree model

A model tree is used for numeric prediction, and at each leaf, it stores a linear regression model which predicts the target results [9]. It relates the observed inputs to the observed/estimated outputs by the process of deduction learning, which is applicable to categorical and numerical input–outputs. Model trees, though simple, are efficient and accurate tools for modeling the patterns and relationships for large datasets [10]. The M5P model specifically proceeds in the following three stages: construction, pruning, and smoothing.

In determining which attribute is the best for splitting the portion T of the training data that reaches a particular node, the splitting criterion is used. When splitting the branches, all criteria is based on standard deviation reduction. Furthermore, the attribute chosen for splitting maximizes the expected error reduction at that node [9]. This is represented in the following equation:

$$\mathrm{SDR} = \mathrm{sd}\left(T\right) - \frac{\left|T_i\right|}{\left|T\right|} \times \mathrm{sd}\left(T_i\right) \qquad (2)$$

3. Materials and methods

3.1. Membrane and experimental equipment

This study conducts the process configuration and operation method of each plant. The A pilot plant is a membrane pressurized operation system with a capacity of 2,000 m^3/d and is a simple process involving a sedimentation–filtration method that operates using an ultrafiltration (UF) membrane: 25 min filtration, and 2 min and 30 s backwashing (see Fig. 1). Also, the polyvinylidene fluoride (PVDF) hollow fiber microfiltration membrane (Hyosung Inc., Korea) was used (see Table 2). The B full-scale plant is a membrane pressurized operation system with a capacity of 8,000 m^3/d. It operates with a coagulation–flocculation–sedimentation UF membrane, and a total of 104 modules (26 modules/unit) are applied to the four arrays: at 35 min filtration and 1 min backwashing (see Fig. 2). The PVDF hollow fiber microfiltration membrane (Toray Industries, Inc., Japan) was used for the B full-scale plant (see Table 3).

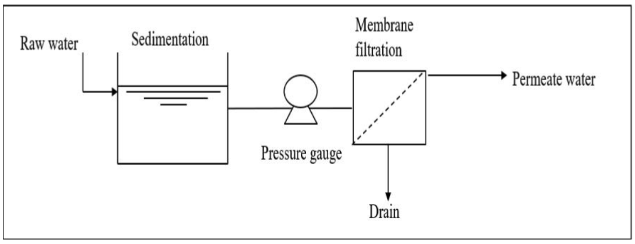

Fig. 1. Schematic diagram of A pilot plant system.

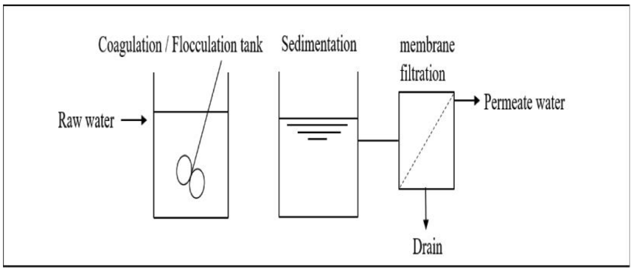

Fig. 2. Schematic diagram of B full-scale system.

3.2. Analysis methods

The database was built based on a series of experiments under different raw water compositions and operating conditions. A total number of 407 datasets of the A pilot plant, and 16,000 datasets of B full-scale plant, were converted for daily and hourly data use. In this study, based on the coefficient of determination (R^2), the root mean square error (RMSE) was used to find out the applicability of the Hermia model, the ANN model, the GP model, and the M5P model. Also, a 95% confidence level was used, based on the model with the highest R^2 value, in order to find the acceptable level. Finally, according to the number of data, the ratio of training data and test data was set to 8:2, 6:4, 4:6, 2:8.

4. Result and discussion

4.1. Raw water quality

Turbidity and temperature were investigated to determine any change in water quality characteristics according to the season of a river or lake. Figs. 3 and 4 are the raw water temperature and turbidity of plant A and plant B, respectively. Temperature and turbidity were used to show the quality of the raw water. The temperature and turbidity of the raw water of pilot plant A ranged from 5.68–114.23 NTU. The turbidity and temperature of the raw water of pilot plant B were 0.88–67.10 NTU. Also, the B plant had a raw water temperature of 2.49°C–31.78°C and 5.09°C–29.68°C, respectively.

4.2. Operating condition

Pilot plant A used data from May 2018 to October 2019 and operates a total of two CIP washing times. After the first CIP washing, the operation was conducted for 220 d, after stopping the operation for 3 months due to a machine failure. In addition, the full-scale plant B used the first array of data from January 2018 to December 2019. Cleaning enhanced backwashing (CEB) washing was performed on a monthly basis, and CEB washing was performed at the CIP washing level. In Fig. 5, TMP changes during the period of data collection for the pilot plant A (left) and the full-scale plant B (right) are shown.

4.3. Model prediction of the data using the Hermia model

The Hermia model (pore blocking, pore constriction, and cake formation) was applied in order to find the fouling mechanism. As a result of analyzing the main fouling mechanism, the cake formation model became the main mechanism. According to Table 4 and Fig. 6, it can be seen that when the ratio of training data and test data is 8:2, the R^2 value is relatively high (from 0.9001, 0.8385, 0.8168). However, if it is at 6:4, 4:6, 2:8, the predicted value is significantly lowered. This means that as the number of input data decreases, the predicted value decreases. The Hermia model shows that TMP increases with time. Therefore, it may be difficult to predict the application to a pilot plant or a full-scale plant of unstable data. For that reason, the result of plant B was omitted.

4.4. Model prediction of the data using the ANN model

The ANN model was created using MATLAB software. Tables 5 and 6, Figs. 7 and 8 are the R^2 value and RMSE, according to the input data ratio by applying the ANN

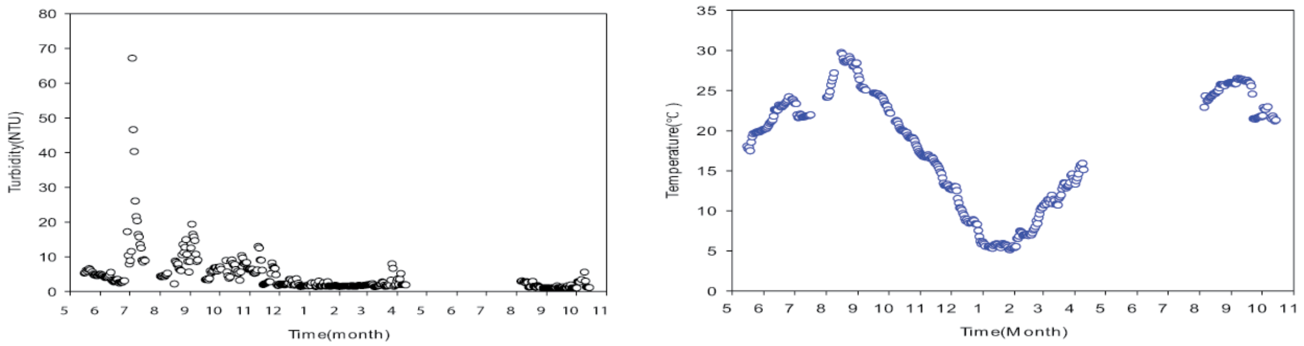

Fig. 3. Characteristic of raw water in A pilot plant (a) turbidity and (b) temperature.

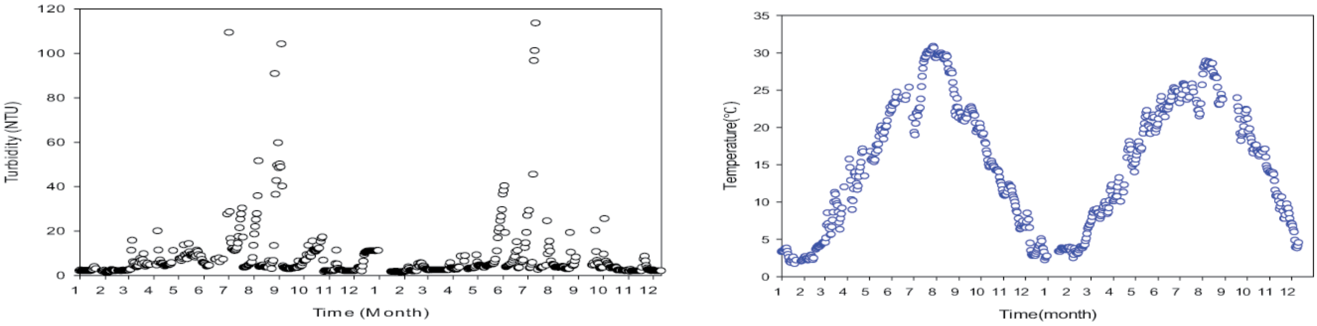

Fig. 4. Characteristic of raw water in B full-scale plant (a) turbidity and (b) temperature.

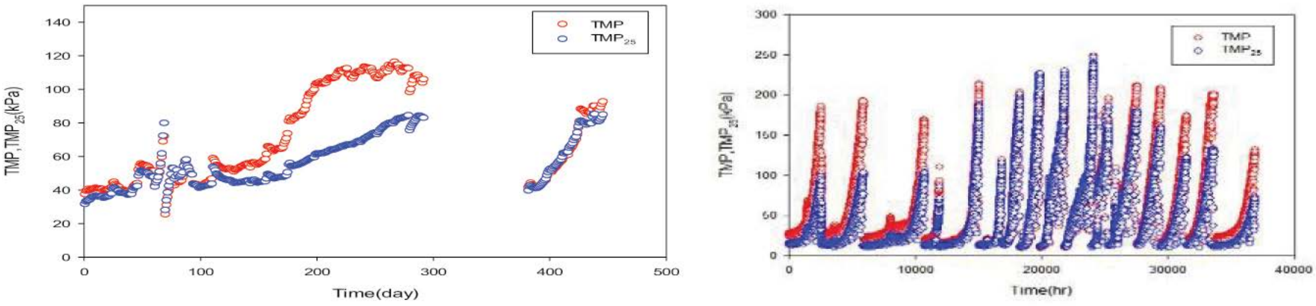

Fig. 5. TMP changes in A pilot plant, B full-scale plant.

Table 4
Cake formation model in A pilot plant

		2:8	4:6	6:4	8:2
Period 1	γ	0.2845	0.1982	0.1034	0.0672
	R^2	0.4468	0.6231	0.7187	0.8168
Period 2	γ	0.2329	0.1325	0.2003	0.0546
	R^2	0.5378	0.6631	0.7627	0.8385
Period 3	γ	0.2243	0.1209	0.1290	0.0705
	R^2	0.5513	0.6713	0.8432	0.9001

Table 5
ANN model in A pilot plant

		8:2	6:4	4:6	2:8
Period 1	RMSE	0.0377	0.0634	0.0037	0.2132
	R^2	0.9379	0.8267	0.8166	0.3274
Period 2	RMSE	0.0229	0.0230	0.0241	0.0420
	R^2	0.9864	0.9860	0.9850	0.9059
Period 3	RMSE	0.0139	0.0310	0.0474	0.1948
	R^2	0.9974	0.9876	0.9701	0.7380

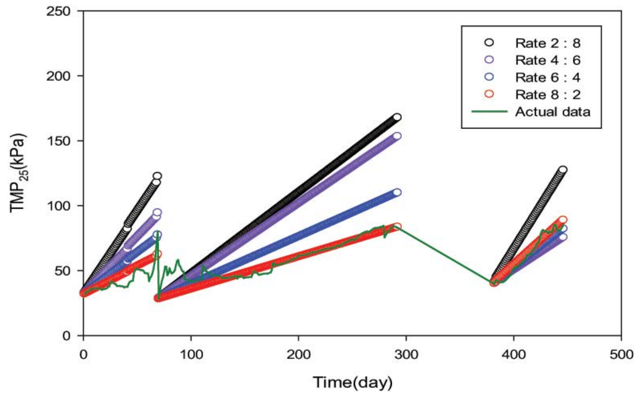

Fig. 6. Cake formation in A pilot plant.

model. Except for period 1 which is the initial operation, the pilot plant A, 4:6 (highest R^2 value = 0.9805), is the range that can be predicted up to the ratio. Also, in the full-scale plant B, it is shown that the TMP25 can be predicted up to 2:8 (highest R^2 value = 0.9953), excluding period 2 when the rapid increase occurs.

4.5. Model prediction data using the GP model

The GP model was created using the GP dot Net v5 software. Tables 7 and 8, Figs. 9 and 10 show the R^2 value and RMSE, according to the input data ratio by applying the GP model. Except for period 1, an acceptable level for predicting an acceptable is up to the 6:4 (highest R^2 value = 0.9686) ratio in the pilot plant A. In the full-scale plant B, it can be found that the TMP25 is predicted up to 4:6 (highest R^2 value = 0.9413), excluding period 2.

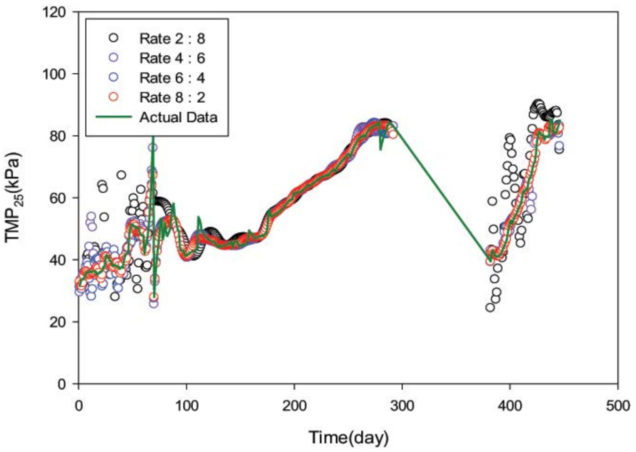

Fig. 7. ANN model in A pilot plant.

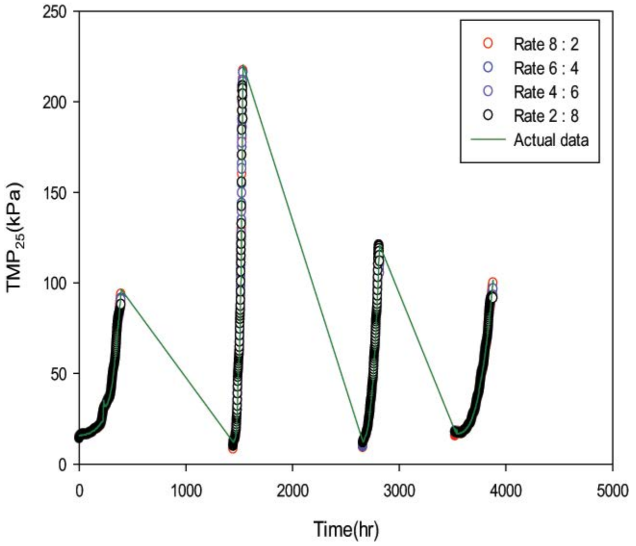

Fig. 8. ANN model in B full-scale plant

Table 6
ANN model in B full-scale plant

		8:2	6:4	4:6	2:8
Period 1	RMSE	0.0109	0.0125	0.0136	0.0160
	R^2	0.9973	0.9966	0.9959	0.9943
Period 2	RMSE	0.0143	0.0161	0.0232	0.0340
	R^2	0.9960	0.9949	0.9907	0.9809
Period 3	RMSE	0.0140	0.0156	0.0175	0.0209
	R^2	0.9965	0.9959	0.9947	0.9941
Period 4	RMSE	0.0084	0.0111	0.0145	0.0160
	R^2	0.9987	0.9978	0.9962	0.9953

4.6. Model prediction of the data using the M5P model

The M5P tree model was applied using WEKA 3.8.4. software. Tables 9 and 10, Figs. 11 and 12 show that the R^2 value and RMSE can be found, along with the input

Table 7
GP model in A pilot plant

		8:2	6:4	4:6	2:8
Period 1	RMSE	0.1402	0.1959	0.1986	0.2089
	R^2	0.7631	0.7253	0.6697	0.5732
Period 2	RMSE	0.0541	0.0771	0.1120	0.2026
	R^2	0.9294	0.9134	0.8580	0.4265
Period 3	RMSE	0.4818	0.0866	0.1169	0.2311
	R^2	0.9747	0.9686	0.8771	0.6426

Table 8
GP model in B full-scale plant

		8:2	6:4	4:6	2:8
Period 1	RMSE	0.0525	0.0972	0.1877	0.1379
	R^2	0.9394	0.9183	0.8955	0.8214
Period 2	RMSE	0.0354	0.1357	0.1771	0.2301
	R^2	0.9895	0.9369	0.8454	0.8449
Period 3	RMSE	0.0292	0.0874	0.1300	0.4655
	R^2	0.9972	0.9964	0.9030	0.8598
Period 4	RMSE	0.0181	0.0330	0.1847	0.1108
	R^2	0.9953	0.9867	0.9413	0.8155

Table 9
M5P model in A pilot plant

		8:2	6:4	4:6	2:8
Period 1	RMSE	0.0749	0.1185	0.1925	0.4549
	R^2	0.7353	0.6872	0.6836	0.6525
Period 2	RMSE	0.121	0.0544	0.1810	0.4223
	R^2	0.9673	0.9638	0.9532	0.8988
Period 3	RMSE	0.0889	0.0069	0.1400	0.4521
	R^2	0.9751	0.9689	0.9658	0.8995

Table 10
M5P model in B full-scale plant

		8:2	6:4	4:6	2:8
Period 1	RMSE	0.0845	0.0580	0.2065	0.2386
	R^2	0.9559	0.9442	0.8278	0.7745
Period 2	RMSE	0.0907	0.1382	0.2233	0.2872
	R^2	0.9418	0.8903	0.8755	0.8358
Period 3	RMSE	0.0481	0.1051	0.1627	0.2560
	R^2	0.9870	0.9657	0.9338	0.9066
Period 4	RMSE	0.0394	0.1147	0.2113	0.2904
	R^2	0.9915	0.9657	0.9211	0.8094

data ratio, by applying the M5P tree model. Except for period 1, the prediction level is up to the 4:6 (highest R^2 value = 0.9658) ratio in pilot plant A. In the full-scale plant B, it was observed that the TMP25 is predicted up to 6:4 (highest R^2 value = 0.9657), excluding period 2.

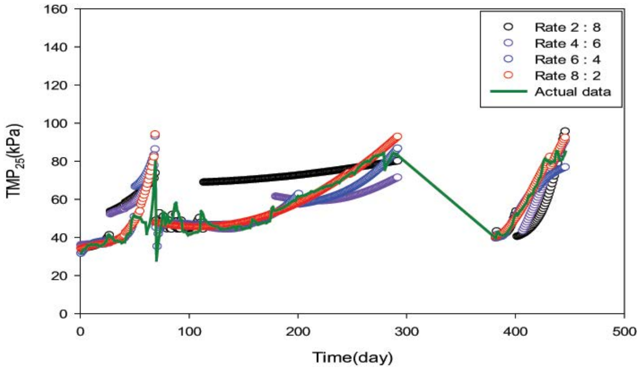

Fig. 9. GP model in A pilot plant.

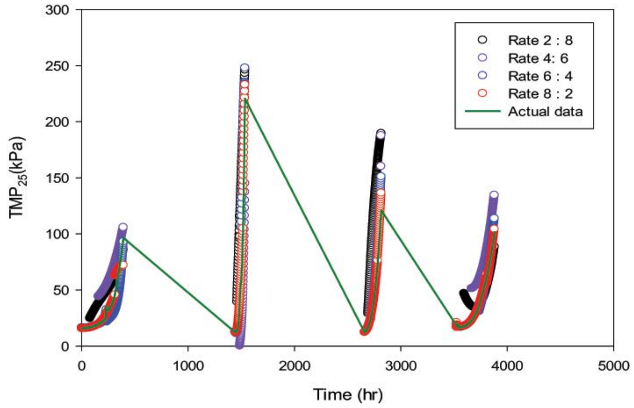

Fig. 10. GP model in B full-scale plant.

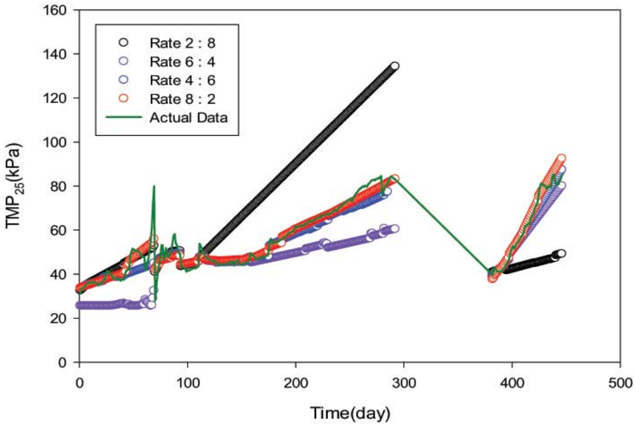

Fig. 11. M5P model in A pilot plant.

5. Conclusion

In current studies, the fitting and prediction concepts are mixed along with studies conducted at the lab scale. The fitting is a method of showing how reliable the model is itself, by extracting and fitting some data from other data. Therefore, fitting does not mean prediction. The Hermia model (pore blocking, pore constriction, and cake formation) was applied in order to find the main fouling mechanism. For pilot plant A, the highest level is the cake

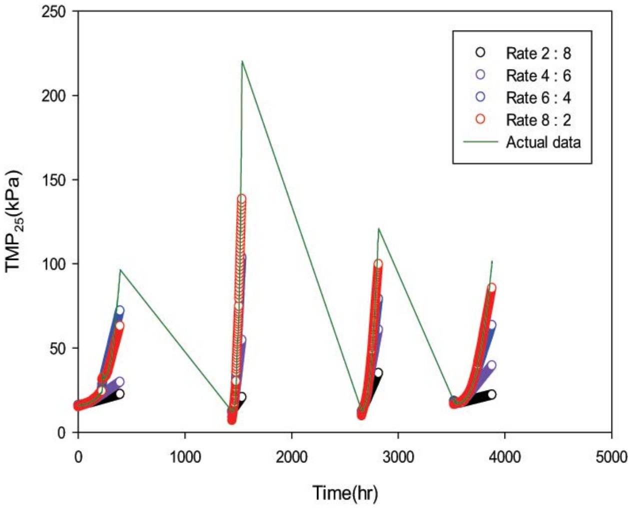

Fig. 12. M5P model in B full-scale plant.

formation model. In the case of the B full-scale plant, the highest level is the pore-blocking model. Mathematical and statistical models were used to predict membrane fouling and chemical periods using the pilot plant and full-scale data in this study. Predicting, defined as a method of dividing the data according to the chemical washing period, is also used in this study. To investigate the applicability of the Hermia model, the ANN model, the GP model, and the M5P model, the ratio of the training data to the test data was set to 8:2, 6:4, 4:6, and 2:8, according to the data number. The results show the application of the (cake formation) to the plant A data. It found that the R^2 value is relatively high when the ratio of the training data to the test data is 8:2. It gets considerably lower at 6:4, 4:6, and 2:8. This means that as the number of input data decreases, the predicted value decreases. The Hermia model is based on TMP that increases with time. It may be difficult to predict the application for a pilot plant or a full-scale plant with unstable data. Plant A is 4:6 for the ANN model, 6:4 for the GP model, and 4:6 except for period 1, which is the initial operation. In the case of plant B, except for period 2 (where the correct TMP at 25°C is unstable because it is summer), the ratio is 2:8 for the ANN model, 4:6 for the GP model. And for the M5P model, a range can be predicted up to a ratio of 6:4. Therefore, the main mechanism is the cake formation model, and with the Hermia model, TMP increases with time. So, it may be difficult to predict the application for a pilot plant or a full-scale plant with unstable data. The ANN model, the GP model, and the M5P model have the potential to predict the washing period of the membrane with a limited number of training data in the pilot plant or full-scale plant We concluded that the ANN model was the most ideal on account of its characteristics. The ANN model is based on human neural networks and is known to be effective in analyzing and predicting time series data. Also, it has a higher prediction accuracy than regression models such as GP, M5P models. It will also be possible to realize the operating conditions of the water treatment plant in advance, as it can predict the time of the next chemical washing.

Acknowledgment

This work was supported by the Korea Ministry of Environment as "Projects for Developing Eco-Innovation Technologies (2016002100001) as well as "Development of the big data platform and standardization techniques for the management of waterworks and sewer systems (2020002700001)".

References

[1] A. Bokhary, A. Tikka, M. Leitch, B.Q. Liao, Membrane fouling prevention and control strategies in pulp and paper industry applications: a review, J. Membr. Sci. Res., 4 (2018) 181–187.

[2] F.L. Wang, V.V. Tarabara, Pore blocking mechanisms during early stages of membrane fouling by colloids, J. Colloid Interface Sci., 328 (2008) 464–469.

[3] E. Iritani, N. Katagiri, Developments of blocking filtration model in membrane filtration, KONA Powder Part. J., 33 (2016) 179–202.

[4] N. Delgrange, C. Cabassud, M. Cabassud, L. Durand-Bourlier, J.M. Lainé, Neural networks for prediction of ultrafiltration transmembrane pressure – application to drinking water production, J. Membr. Sci., 150 (1998) 111–123.

[5] Y.J. Park, Y.J. Choi, S.H. Lee, Analysis of membrane fouling in a pilot-scale microfiltration plant using mathematical model and artificial neural network model, Desal. Water Treat., 77 (2017) 69–74.

[6] Y.-J. Choi, H.J. Oh, S.H. Lee, S.-H. Nam, T.-M. Hwang, Investigation of the filtration characteristics of pilot-scale hollow fiber submerged MF system using cake formation model and artificial neural networks model, Desalination, 297 (2012) 20–29.

[7] N. Muttil, J.H.W. Lee, Genetic programming for analysis and real-time prediction of coastal algal blooms, Ecol. Modell., 189 (2005) 363–376.

[8] T.-M. Lee, H.J. Oh, Y.-K. Choung, S.H. Oh, M.G. Jeon, J.H. Kim, S.H. Nam, S.H. Lee, Prediction of membrane fouling in the pilot-scale microfiltration system using genetic programming, Desalination, 249 (2009) 285–294.

[9] E.K. Onyari, F.M. Ilunga, Application of MLP neural network and M5P model tree in predicting streamflow: a case study of Luvuvhu Catchment, South Africa, Int. J. Innovation Manage. Technol., 4 (2013), doi: 10.7763/IJIMT.2013.V4.347.

[10] M.R. Nikoo, A. Karimi, R. Kerachian, H. Poorsepahy-Samian, F.H. Daneshmand, Rules for optimal operation of reservoir-river-groundwater systems considering water quality targets: application of M5P model, Eur. Water Resour. Assoc., 27 (2013) 2771–2784.