Front. Environ. Sci. Eng. 2016, 10(2): 299–310
DOI 10.1007/s11783-015-0825-7

RESEARCH ARTICLE

Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant

Minsoo KIM[1], Yejin KIM[2], Hyosoo KIM[3], Wenhua PIAO[1], Changwon KIM (✉)[1]

1 Department of Civil and Environmental Engineering, Pusan National University, Busan 609-735, Republic of Korea
2 Department of Civil and Environmental Engineering, Catholic University of Pusan, Busan 609-757, Republic of Korea
3 EnvironSoft Co., Ltd. #511 Industry-University Co., Bld., Pusan National University, Busan 609-735, Republic of Korea

Abstract The k-nearest neighbor (k-NN) method was evaluated to predict the influent flow rate and four water qualities, namely chemical oxygen demand (COD), suspended solid (SS), total nitrogen (T-N) and total phosphorus (T-P) at a wastewater treatment plant (WWTP). The search range and approach for determining the number of nearest neighbors (NNs) under dry and wet weather conditions were initially optimized based on the root mean square error (RMSE). The optimum search range for considering data size was one year. The square root-based (SR) approach was superior to the distance factor-based (DF) approach in determining the appropriate number of NNs. However, the results for both approaches varied slightly depending on the water quality and the weather conditions. The influent flow rate was accurately predicted within one standard deviation of measured values. Influent water qualities were well predicted with the mean absolute percentage error (MAPE) under both wet and dry weather conditions. For the seven-day prediction, the difference in predictive accuracy was less than 5% in dry weather conditions and slightly worse in wet weather conditions. Overall, the k-NN method was verified to be useful for predicting WWTP influent characteristics.

Keywords influent wastewater, prediction, data-driven model, k-nearest neighbor method (k-NN)

1 Introduction

For the stable operation of a wastewater treatment plant (WWTP), it is essential to consider major disturbances such as fluctuations in the influent flow rate and water qualities. The online monitoring of influent characteristics is limited by a lack of adequate equipment and high costs. Therefore, some mathematical models and data-driven models have been developed to predict the influent flow rate and the water quality. Traditionally, a few mathematical models have been applied to simulate physical phenomena such as the movement mechanism of contaminants in conduits [1]. The storm water management model (SWMM) has been widely used as a comprehensive package to predict the influent flow rate and quality changes within conduits [2]. This model can simulate hydraulic and hydrologic changes in runoff mechanisms in urban areas. Studies have reported the development of simple transfer functions that are simplified versions of integrative model equations for conduits in the SWMM [3]. Furthermore, the IWA (International Water Association) developed its benchmark simulation model No. 2 to predict WWTP influent characteristics [4]. For the model's increased applicability, a simple phenomenological model using only those parameters with significant effects on the target area has been developed [5]. However, these mathematical models require complicated parameter tuning and large-scale monitoring. To overcome these problems, many studies have investigated data-driven models such as the autoregressive integrated moving-average (ARIMA) model and the artificial neural network (ANN) [6].

A data-driven model can be classified as a linear or nonlinear time series model. The former uses a linear function with an error term for predictions based on a probability distribution. This group includes the autoregressive moving-average (ARMA), ARIMA, and seasonal autoregressive integrated moving-average (SARIMA) models, with the ARIMA model being the most widely used one. Kim et al. [7] applied the ARIMA model to forecast WWTP influent flow rate and compositions. Wang et al. [8] used the ARIMA model to predict precipitation by considering seasonal effects and stationarity. Valipour et al. [9] suggested the ARMA and ARIMA models for

Received July 8, 2015; accepted November 11, 2015

E-mail: cwkim@pusan.ac.kr

forecasting the monthly inflow of a dam reservoir by deriving an optimal model structure using an appropriate number of parameters. These linear time series models have also been applied to other research areas such as hydrological processes [10]. To obtain the order of linear time-series models and parameters, the autocorrelation function (ACF) or the partial autocorrelation function (PACF) should be used. If data do not appear to be stationary, then a nonlinear time series model should be considered instead of a linear time series model [11].

A nonlinear time series model can be developed using either a global or local approach [12]. The global approach is a method for developing a nonlinear time series model by deriving a function based on the whole attractor. ANN is a representative method in the global approach [13]. Kim et al. [6] and Solaimany-Aminbad et al. [14] used ANN to predict influent flow rate and compositions at WWTP. Bagheri et al. [15] forecasted the occurrence of activated sludge bulking using ANN with genetic algorithm to reduce the prediction error. Many other researchers have used ANN for monitoring, control, and predictions in activated sludge processes [16]. ANN has been used to establish model structures by selecting related input variables to forecast the patterns of a target variable. However, this method can lack objectivity because the number of layers of the model structure is typically determined based on subjective trial-and-error standards. ANN can assign high weight values to input variables with autocorrelation, resulting in a lag effect and the underestimation of information on major variables [17].

On the other hand, local approaches such as the k-nearest neighbor (k-NN) method can be applied. The k-NN method approaches a complex nonlinear time series by applying the concept of chaos theory, which posits that some part of a time series occurring in the past can occur in the future with highly similar characteristics [18]. The k-NN method at first selects data pairs among past data pairs such that selected pairs show characteristics that are very similar to those of past pairs located before the predicted time point. Then the weight is assigned to those selected data pairs, which are referred to as nearest neighbors (NNs), and the prediction is made by summing up weighted data pairs [19]. Therefore, the performance of the k-NN method is affected by the search range based on data size because the number of selected data pairs can vary according to the search range based on data size [20]. In addition, the performance of the k-NN method is affected by the approach used to determine the number of NNs or data pairs. Here the two possible approaches are the distance factor-based (DF) approach [21] and the square root-based (SR) approach [22], which are explained in detail later. The k-NN method can make predictions using a relatively straight-forward calculation and thus differs from the global approach. Therefore, unlike in mathematical modeling, the k-NN method requires no model development or verification and thus can be applied without recomposing data, unlike in the case of general data-based models. In addition, lag effects appear relatively seldom. However, because this method make predictions based on the current time period using past data, these predictions can be limited for events that did not occur in the past, such as certain peak influent flows [23].

WWTP influent flow rate and water qualities exhibit characteristics of a nonlinear time series because some part of the influent presents a periodical pattern of a linear time series according to the human life cycle. Another part is affected by irregular weather conditions showing a nonlinear time series. In addition, the influent exhibits characteristics such as autocorrelation, which shows a strong correlation between current, future, and past situations [24]. The flow rate of rivers and the WWTP influent show similar hydrological data, and therefore predicted data on this flow rate exhibit autocorrelation. Two studies reported that the k-NN method can predict the river flow much better than ANN because of this autocorrelation [25].

Therefore, this study evaluates the k-NN method by selecting conditions to forecast influent flow rate and four water qualities including COD, SS, T-N and T-P at a WWTP. These conditions were considered based on the following two factors: the search range based on data size and the number of NNs. Four search ranges were selected to consider weekly, monthly, seasonal, and annual variations. To determine the optimum number of NNs, the DF and SR approaches were taken. Then optimal conditions were evaluated for their potential application to a long-term of one week.

2 Material and methods

2.1 Data collection and performance evaluation

Target characteristics of the WWTP were the influent flow rate and water qualities: COD (chemical oxygen demand), SS (suspended solid), T-N (total nitrogen), and T-P (total phosphorus). Relevant daily data on these measures were collected from N WWTP in Busan, Korea, for three years from January 2008 to December 2010. Some part of the data were used for predictions, whereas the rest and predicted values were used to evaluate prediction results based on the root mean square error (RMSE). Because the influent wastewater of N WWTP was collected by a combined sewer system, influent characteristics were strongly affected by the weather condition, especially rain events.

2.2 Identification of the search range based on data size

The k-NN method was developed based on the assumption that some part of a past time series reappears in the future with similar patterns [26]. Therefore, NNs with a pattern

most similar to that of the current data pair were selected from past data.

The range of past data from which NNs were searched was defined as the search range. This was crucial because prediction results can be affected by data size based on the search range. To select an appropriate search range, four sets of data were tested: 30, 90, 365, and 731 day data. These were determined to consider weekly, monthly, seasonal, and annual variations, respectively, of influent wastewater. The rainfall effect was considered under the assumption of dry weather in January and February (59 days) and wet weather in July and August (62 days).

2.3 The calculation procedure for the k-NN method

First, within the pre-determined search range, the prediction of future data on day st + 1 (Y_{st+1}) is considered. The current time as the base is indicated as "st." Data for the calculation base are Y_{st} and Y_{st-1}, which are data pairs at times st and st-1, respectively. The past data pair, Y_{t-1} and Y_{t-2}, are selected to calculate the Euclidean distance (D_t) as follows Eq. (1)

$$D_t = \sqrt{(Y_{st-1} - Y_{t-2})^2 + (Y_{st} - Y_{t-1})^2}, \quad (1)$$

Then the value with the shortest distance (D_{MS}) is selected among all computed values of D_t, and some values of D_t that are close to D_{MS} are selected and named as NNs (D_{NN}). The number of selected D_{NN} is determined by an appropriate approach, as will be explained later.

Second, appropriate NNs that are selected to calculate weight values ($W_{NN,t}$) as follows:

$$W_{NN,t} = (D_{NN,t})^{-1} / (\sum_{1}^{t} D_{NN,t})^{-1}, \quad (2)$$

Third, future data Y_{st+1} are calculated using the following equation by multiplying the weight value by water quality data (Y_t) that correspond to NNs in that group and then summing up all results:

$$Y_{st+1} = \sum_{1}^{t} (W_{NN,t} \times Y_t). \quad (3)$$

2.4 A comparison of two approaches to determine an appropriate number of NNs

The number of selected NNs affects predictive performance. If too few NNs are selected, then the result is sensitive to noise, whereas if too many NNs are selected, then the predicted value is similar to that by linear regression. Therefore, to select an appropriate number of NNs, two approaches were applied in this study. In the DF approach, D_{NN} was selected such that it was located within values obtained by multiplying a certain factor by D_{MS}. In the SR approach, data as much as the square root of the number of past data provided were selected. These two approaches are now explained in detail.

DF approach: The distance corresponding to NNs (D_{NN}) can be determined by multiplying certain factors by the shortest Euclidean distance (D_{MS}), and predictive accuracy is affected by the selected number of NNs depending on the factor. Therefore, an optimum number of NNs should be considered before applying the developed k-NN method. According to some studies, a large number of NNs do not always increase predictive accuracy [27]. Similarly, other studies have reported that predictive accuracy is not controlled by a small number of NNs [28]. Therefore, seven tested factors were arbitrary selected in this study by considering a small range and a large range: the golden ratio (1.62 [29] and 2.0, 2.5, 5.0, 8.0, 15.0, and 30.0 [30]). The golden ratio of 1.62, the applied base in previous studies, was used to apply the k-NN method [31,32].

SR approach: This method selects the number of NNs according to the square root of the number of data sets used in the search range.

These two approaches differ as follows: In the DF approach, the number of selected NNs differs every time it is calculated. Therefore, predicted results are affected by assigned weight values and water quality data on selected NNs depending on various factors. In addition, additional analyses and calculation times are required for changes in final results. On the other hand, the SR approach uses a fixed number of NNs, and results are calculated using assigned weight values and water quality data on the fixed number of NNs. Therefore, the result can be easily understood if it is necessary to determine the estimation procedure.

2.5 An evaluation of the applicability of the k-NN method for long-term predictions

The aforementioned procedure predicts tomorrow's data based on past data, including data up to today. It is more valuable for planning WWTP operations if this method can be used for more long-term predictions. The k-NN method can be extended to predict data the day after tomorrow and beyond by incorporating predicted data for tomorrow into the mother data group. This procedure was tested in this study by expanding the prediction up to seven days. Because the number of data points was insufficient to consider any statistical analysis, accuracy was evaluated in terms of percentage differences between measured and predicted values.

2.6 An uncertainty analysis for the application of the k-NN method

For the application of the k-NN method, exploring uncertainty is required for understanding the selected

approach such as the DF and SR approaches in the calculation process. This is because a generated error in the calculation process for a predicted value may vary according to the applied approach. Therefore, an uncertainty analysis of the k-NN method was conducted in this study based on RMSE results for different data sizes and the two approaches.

3 Results and discussion

3.1 Statistical features of the N WWTP influent

Statistical features of measured data are listed in Table 1. The average influent flow rate was 306200 $m^3 \cdot d^{-1}$, ranging from 242953 $m^3 \cdot d^{-1}$ to 406461 $m^3 \cdot d^{-1}$. To eliminate statistical outliers in collected data, the lower and upper limits were set as plus/minus three times the standard deviation from the mean, respectively. Data located outside these two limits were replaced by two limits obtained from the straight-line interpolation.

3.2 Identification of an appropriate search range based on data size in the influent flow rate

Four potential search ranges in the K-NN method were tested to consider weekly, monthly, seasonal, and annual variations in influent wastewater under dry and wet weather conditions. One-day-ahead predictions were made continuously by considering effects according to the number of NNs. Predicted values were compared with measured data, and the RMSE was calculated as the basis for the evaluation. The results for dry and wet weather conditions are summarized in Tables 2.

In the dry weather condition, when the DF approach was applied with a factor of 1.62, the smallest search range of 30 days showed the best result. However, if the factor increased in the search range from 5 to 30, then larger search ranges such as one year (365 days) showed better results. The SR approach showed better results than the DF approach for all four search ranges, and it was also better in the one-year search range.

In the wet weather condition, the DF approach showed varying results. The SR approach showed better results than the DF approach in all four search ranges and was better in the one-year search range. The two-year (731 days) search range provided results no different from those for the one-year search range. This indicates that the influent flow rate had some annual pattern. Therefore, the one-year search range was considered good enough to predict the influent flow rate regardless of the weather condition.

3.3 Evaluation of effects of the number of NNs in the influent flow rate

Effects of the number of NNs were tested using both the DF approach and the SR approach for the influent flow

Table 1 Statistical properties of influent wastewater flow rate and compositions for the N WWTP

item	flow rate ($m^3 \cdot d^{-1}$)	COD ($mg \cdot L^{-1}$)	SS ($mg \cdot L^{-1}$)	T-N ($mg \cdot L^{-1}$)	T-P ($mg \cdot L^{-1}$)
maximum	406461.0	95.3	338.3	48.8	7.1
minimum	242953.0	36.9	56.7	13.9	1.8
average	306204.5	61.5	122.7	33.8	3.9
standard deviation	35681.7	10.3	25.6	5.9	0.8

Table 2 RMSE results of the effects by the search range considering the number of NNs in the influent flow rate under the dry and wet weather conditions

weather	search range (d)	DF approach							SR approach
		1.62	2.0	2.5	5.0	8.0	15.0	30.0	
dry	30	28528.9	28594.5	28623.7	28836.4	29428.1	29326.4	29280.2	27392.7
	90	31260.2	30795.7	30975.7	29914.3	30673.9	31465.0	31628.0	28222.6
	365	32280.4	31745.3	29986.7	27474.4	26769.9	27125.5	26992.9	26695.1
	731	34096.6	33300.4	32185.0	28667.8	27586.2	27763.7	28422.4	27166.0
wet	30	29731.8	28852.9	29273.3	31968.7	32676.8	33064.5	33838.4	27571.3
	90	29404.0	28898.4	29281.4	35437.7	37489.0	40878.6	44564.0	28683.9
	365	33752.9	31768.6	27258.4	27483.7	31354.8	41215.1	53193.6	26038.8
	731	31293.7	29472.2	28184.0	26949.2	27535.8	32491.2	45090.8	26280.7

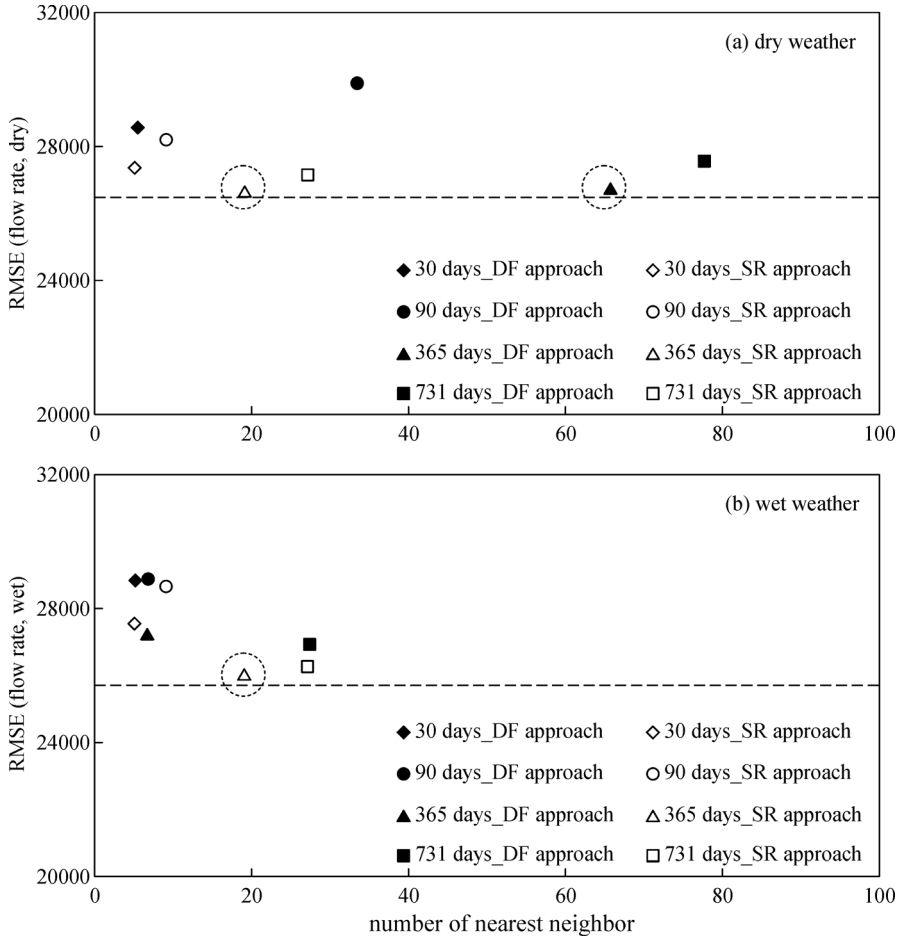

Fig. 1 Effects of the number of NNs on forecasting the flow rate analyzed using the RMSE based on the DF approach (◆, ●, ▲, ■) and the SR approach (◇, ○, △, □) in dry (a) and wet (b) weather conditions

rate. Fig. 1 shows the results analyzed in terms of the RMSE. In the dry weather condition, the two approaches showed similar results in the one-year search range. However, the number of NNs in the SR approach was less than a third of that in the DF approach. Therefore, the SR approach was better than the DF approach. These results suggest that if the number of NNs is too large, then the result may be closer to that obtained based on a linear time series model, which is very different from the real situation.

The optimal conditions for predicting future influent flow rate by using the k-NN method were the one-year search range and the number of NNs by the SR approach. With these optimal conditions, the influent flow rate was predicted for the period from January 1 to February 28 for dry weather and from July 1 to August 31 for wet weather. Then predicted outcomes were compared with measured values, as shown in Fig. 2. Differences between these two were within one standard deviation (35681.7 $m^3 \cdot d^{-1}$), verifying that the k-NN method could predict the WWTP influent flow rate.

3.4 Derivation of appropriate conditions in terms of the search range and the number of NNs to predict influent water qualities

To use the k-NN method to predict the four influent qualities, the number of NNs based on the two approaches was evaluated. Table 3 summarizes the results according to the weather condition and the water quality. Because each water quality had its own specific characteristic, the appropriate condition was different for each water quality. It was necessary to evaluate these conditions according to the water quality and the flow rate. The influent flow rate typically had an annual pattern, and the seasonal variation was caused by infiltrated water through the combined sewer system in the wet weather condition and by changes in water consumption during summer months as the temperature rose. Therefore, to predict the influent wastewater quality, it was crucial to determine similar patterns and investigate useful information from data collected over past few days or a long period of time depending on the condition.

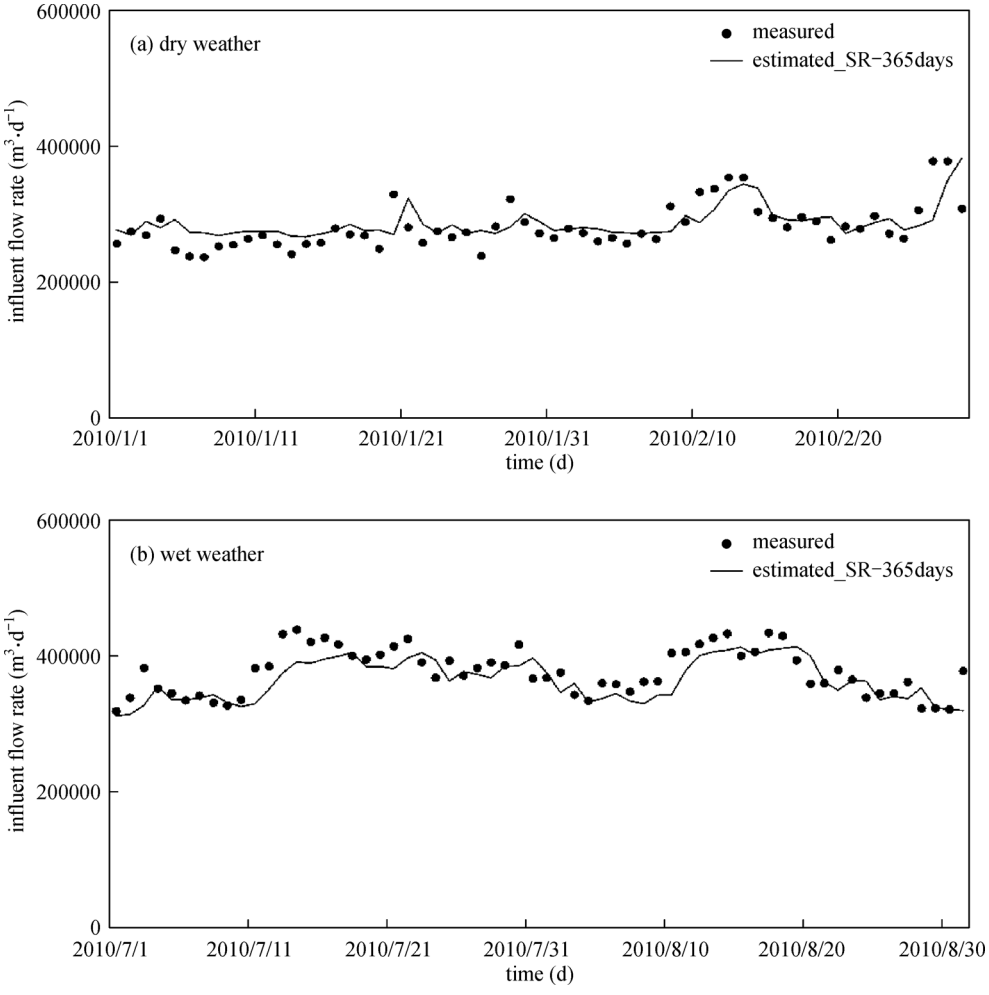

Fig. 2 Results for predicting the influent flow rate by using the k-NN method in dry and wet weather conditions

Table 3 Derivation of appropriate conditions for the search range and the number of NNs to predict influent water qualities (COD, SS, T-N, and T-P)

weather	subject	COD	SS	T-N	T-P
dry	search range	365 days	90 days	30 days	731 days
	approach	DF (8)	SR	DF (2.5)	SR
	number of NNs	74.4*	9.0	7.2*	27.0
	RMSE	3.91	7.69	2.02	0.15
wet	search range	731 days	365 days	731 days	365 days
	approach	DF (5)	SR	SR	DF (8)
	number of NNs	64.1*	19.0	27.0	137.6*
	RMSE	5.82	8.91	2.17	0.34

Note: * Average number of NNs.

COD: In general, characteristics of organic matter represented by influent COD are affected by variations in compositions and concentrations of households and industries and by seasonal variations in the influent flow rate. COD predictions showed the best performance in the one-year search range in dry weather and in the two-year search range in wet weather. In dry and wet weather forecasting, the average numbers of NNs were 74 and 64,

respectively, with DF values of 8 and 5, respectively. The size of the search range showed a difference of one year, but the referenced number of NNs according to the applied DF showed no substantial differences. This suggests that, because the COD pattern varies annually, it may be better to use long-term data (i.e., more than a year) for accurate predictions.

SS: SS variations are influenced by accumulated pollutants in the sewer system and streets according to variations in rainfall intensity in addition to the wastewater flow rate produced from various sources. In the case of SS, like the influent flow rate, the SR approach exhibited better performance. The four search ranges were quite different from one another depending on the weather condition, with 90 days in dry weather and 365 days in wet weather. These results indicate that the profile of dry-weather forecasting did not deviate greatly from recent trends. For wet-weather forecasting, past precipitation patterns were required, and therefore long-term data were referenced to obtain good prediction results.

T-N: Variations in T-N and T-P concentrations were related to variations in the SS concentration and COD variations. As in the case of SS, T-N recorded the lowest RMSE, with the search range of 30 days in dry weather and that of 731 days in wet weather. However, the DF approach with 2.5 in dry weather recorded the lowest RMSE. In the wet weather condition, the lowest RMSE was obtained by the SR approach. Because the weather condition affected variations in the T-N concentration, the search range was set to only 30 days in dry weather. In the wet weather condition, to minimize noise from the selection of data that did not reflect climate characteristics, the search range was extended to a longer period of about one year.

T-P: Absolute values and fluctuation ranges of concentration data were minimal. Therefore, the search range was selected to be relatively long, with two years for dry weather and one year for wet weather. In particular, the average number of NNs in wet weather was 137 for the DF of 8, which was larger than that in dry weather. This may be due to the fact that T-P in wet weather varied much more than that in dry weather. This case suggests that the forecasting was heavily dependent on NNs including fluctuation characteristics in the same year, not on those in the long search range of two years.

3.5 The capability to predict influent water qualities using the k-NN method

The k-NN method was used to predict influent water qualities in dry and wet weather conditions based on appropriate conditions that were derived as described earlier. Measured data were obtained from January 1 to February 28 and from July 1 to August 31, respectively. Predicted outcomes were compared with measured values, as shown in Fig. 3.

Predictions were statistically evaluated using the mean absolute percentage error (MAPE), which measures the difference between measured and estimated values [33]. All results were less than 8.9%, as shown in Fig. 3, verifying their statistical precision and the acceptability of using the k-NN method to predict WWTP influent qualities.

3.6 An evaluation of the applicability of the k-NN method to long-term predictions

Fig. 4 shows the results for the applicability of the k-NN method to long-term predictions in terms of differences in predictive accuracy. The results for the influent flow rate in Fig. 4(a) show good accuracy with less than 5% variations in dry weather. In wet weather, however, accuracy deteriorated, with the highest variation of 11.6%. This may be attributed to larger fluctuations in the influent flow rate in wet weather as a result of precipitation. The COD results in Fig. 4(b) show good accuracy, with less than 5% variations in dry weather and an acceptable level of accuracy of less than 6.5% in wet weather except for the two-day case, which may be explained by precipitation.

According to the SS results in Fig. 4(c), predictive accuracy in wet weather was better than that in dry weather. In dry weather with 90-day data and climate-change characteristics, the search range in dry weather was narrower than that in wet weather. Therefore, the rain occurrence within the search range influenced predictive performance and extended accuracy differences in continuous predictions. However, because wet weather employed 365-day data, it was considered to have lower prediction error rates because of buffer effects of referenced data because it reflected annual climate characteristics.

T-N and T-P showed high accuracy in dry weather, as shown in Fig. 4(d) and 4(e), respectively. In wet weather, however, accuracy deteriorated when the search range was more than a day.

3.7 An uncertainty analysis of the application of the k-NN method

In this study, optimal search ranges and application options for predicting influent flow rate and water qualities were derived by considering different data sizes and approaches. As demonstrated by the aforementioned results, predictive accuracy was affected by the number of data points selected as NNs. Any uncertainty in results may be due to two factors: data size as NNs used to make predictions and inherent uncertainty in data. Therefore, this section provides an uncertainty analysis by comparing RMSE results in dry and wet weather conditions to interpret data characteristics and explore the uncertainty originating from data. To explore the uncertainty from the number of NNs, RMSE variations were analyzed in each trial case. If the RMSE varies widely according to different factors in

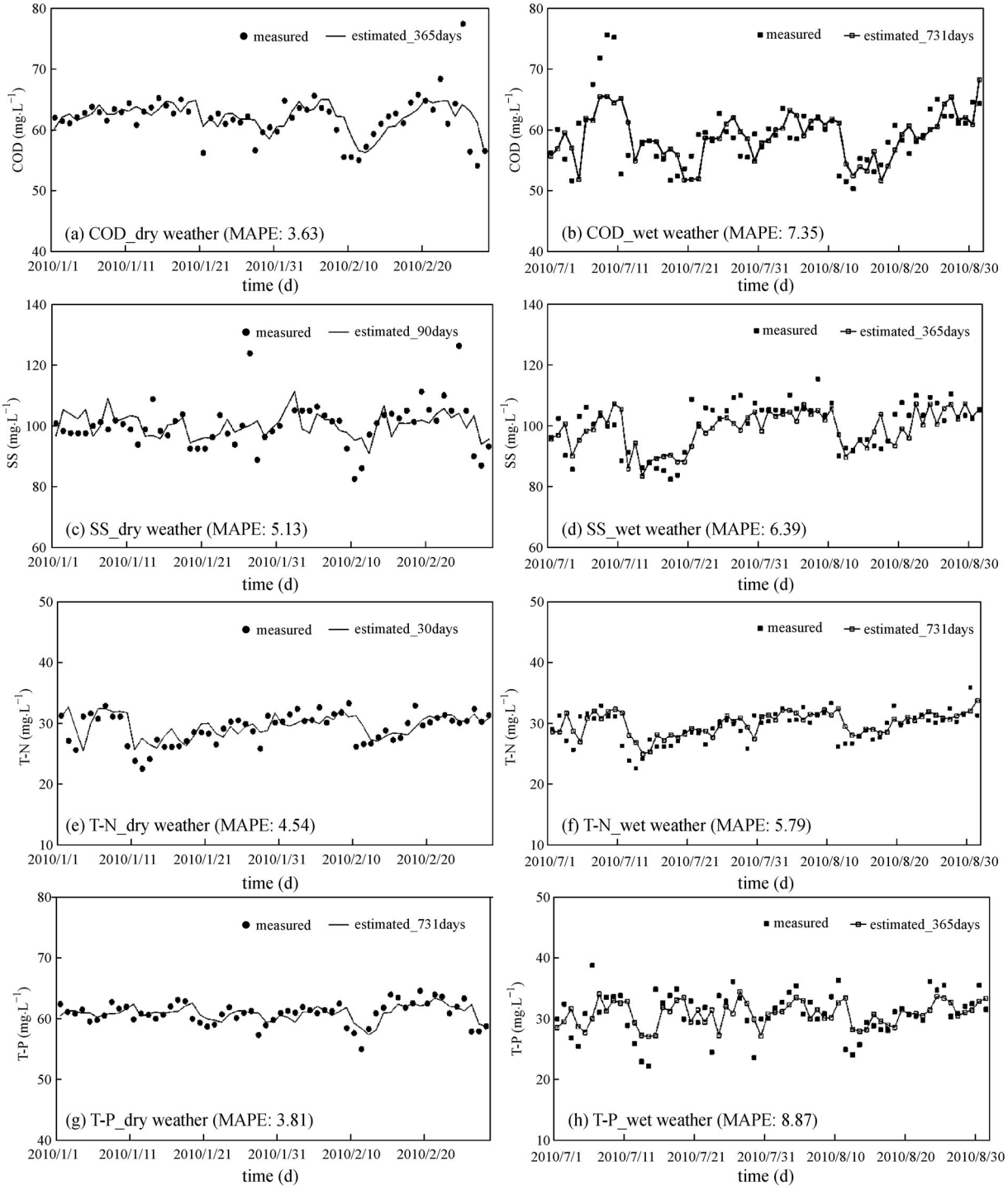

Fig. 3 A comparison of predicted influent qualities (COD, SS, T-N, and T-P) with measured data in dry and wet weather conditions

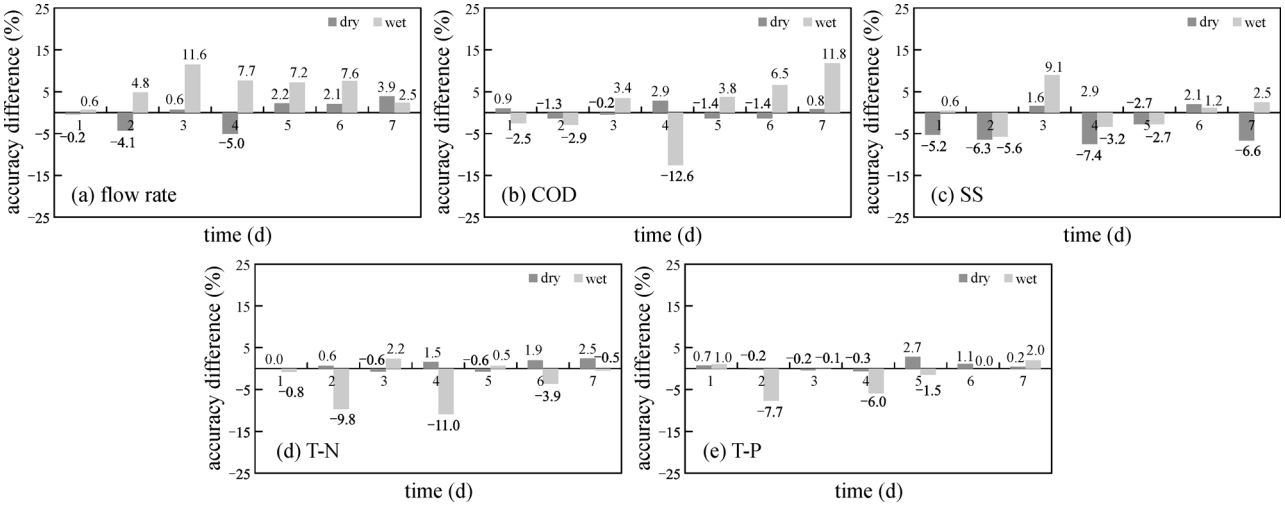

Fig. 4 An evaluation of the accuracy of long-term predictions of influent flow rates and water qualities based on the k-NN method

selecting NNs, then the DF approach can be assumed to reflect a high level of uncertainty.

Fig. 5 shows a box plot of RSME results for the two approaches for influent flow rates and water qualities.

For the uncertainty originating from data, Fig. 5 shows that the RMSE distribution varied between wet and dry weather conditions. Although it may be possible to determine an optimal approach, as discussed in Section 4, Fig. 5 shows that RMSE values can vary widely according to factor size in the DF approach.

Because of characteristics of data on the influent flow rate showing frequent peak values, the RMSE varied more widely in wet weather than in dry weather. However, this type of uncertainty was not observed in other variables. This suggests that the SR approach is a robust method for predicting the influent flow rate, whose RMSE remained relatively constant and low. On the other hand, other variables showed no distinct differences between dry and wet weather conditions, implying that the uncertainty from data was not a significant obstacle to using the k-NN method to make wastewater influent predictions.

For the uncertainty from options for k-NN application, namely the DF approach and the SR approach, both options generally showed robust performance except for influent flow rate predictions in wet weather. This implies that the uncertainty of original data is a more important factor than that of k-NN application conditions. However, it should be noted that the uncertainty from such factors can reduce RMSE variations but produce obvious differences, making it possible to select an optimal application condition.

4 Conclusions

The k-NN method was evaluated and applied to predict the influent flow rate and four water qualities at a WWTP, namely COD, SS, T-N, and T-P. To determine optimal conditions for the method, optimal search ranges based on data size and the number of NNs were examined.

Optimal search ranges varied depending on the influent flow rate, the water quality, and the weather condition. In most cases, at least one year of past data were required to reflect seasonal variations. In the dry weather condition, less-than-90-day data were useful. In terms of an appropriate number of NNs, the SR approach showed better results than the DF approach. However, different water qualities showed somewhat different preferences.

Once optimal conditions for the k-NN method were set, its prediction capability was evaluated. The influent flow rate was accurately predicted such that predicted values were located within one standard deviation of measured values. The influent water quality was statistically well predicted with a MAPE value of less than 8.9% for all four water qualities and in both wet and dry weather conditions. For seven-day predictions, the difference in predictive accuracy in dry weather was less than 5%. This result clearly demonstrates the applicability of this method to ordinary situations. In wet weather, however, the difference in predictive accuracy deteriorated such that the usefulness of predictions was limited.

Overall, the results suggest that the k-NN method may be readily used to predict the WWTP influent flow rates and water qualities in dry weather. In wet weather, however, predictions should be made only with caution. In addition, an uncertainty analysis was conducted using two approaches and different data sizes. Predictive performance was affected by inherent uncertainty in data size and characteristics. However, the optimal approach reduced the uncertainty of the k-NN method. In conclusion, the results suggest that the k-NN method can predict WWTP influent flow rates and water qualities.

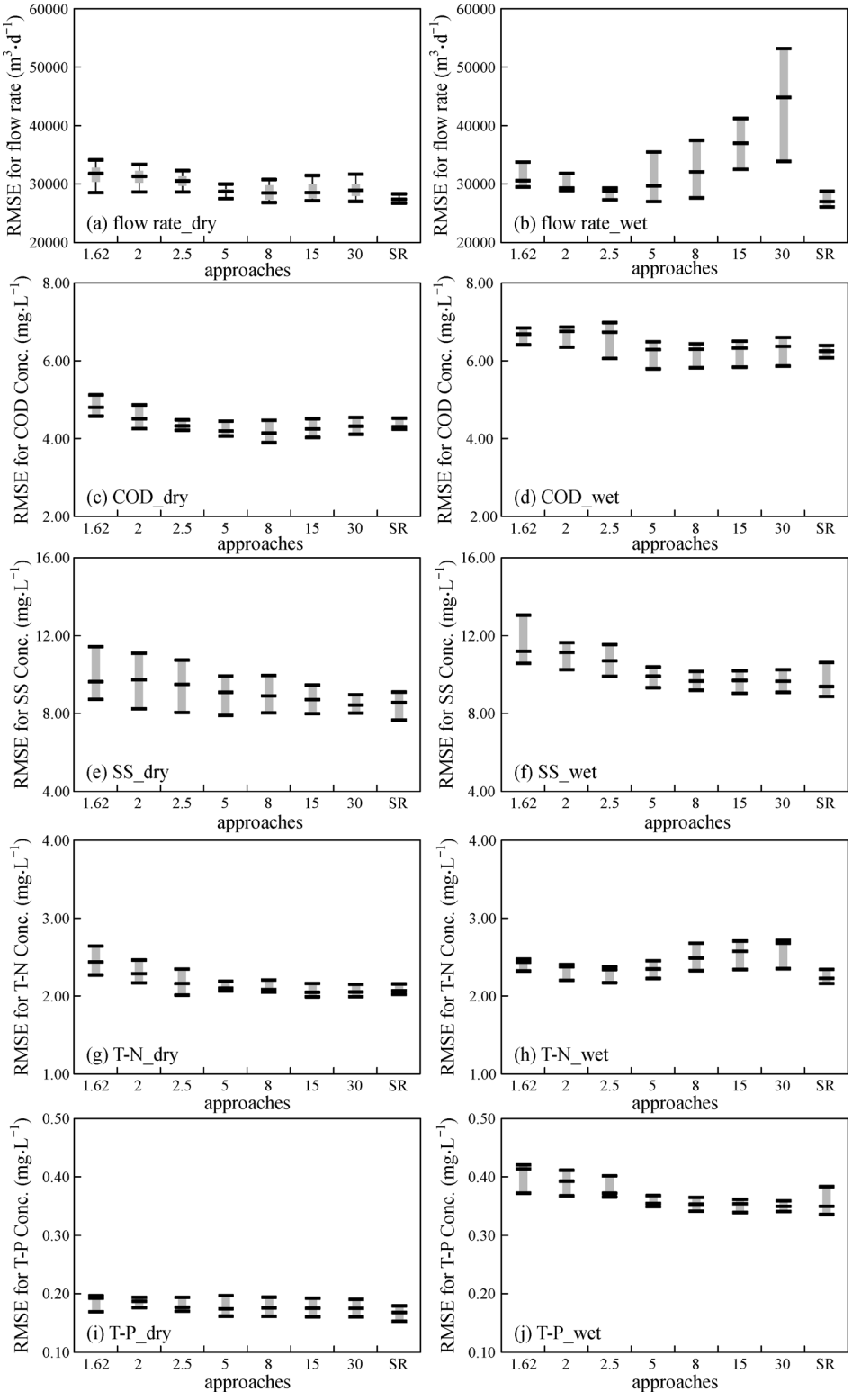

Fig. 5 A comparison of box plots for the application of two approaches in dry and wet weather conditions

Acknowledgements This research was supported by the Korea Ministry of Environment as part of the Eco-innovation Project. In addition, this work was financially supported by the second stage of the Brain Korea 21 Project in 2013.

References

1. Butler D, Graham N J D. Modeling dry weather wastewater flow in sewer networks. Journal of Environmental Engineering, 1995, 121 (2): 161–173
2. Lin S, Liao Y, Hsieh S, Kuo J, Chen Y. A pattern-oriented approach to development of a real-time storm sewer simulation system with an SWMM model. Journal of Hydroinformatics, 2010, 12(4): 408–423
3. Freni G, Mannian G, Viviani G. Urban storm-water quality management: centralized versus source control. Journal of Water Resources Planning and Management, 2010, 136(2): 268–278
4. Jeppsson U, Rosen C, Alex J, Copp J,Gernaey K V, Pons M N, Vanrolleghem P A. Towards a benchmark simulation model for plant-wide control strategy performance evaluation of WWTPs. Water Science and Technology, 2008, 53(1): 287–295
5. Gernaey K V,Flores-Alsina X, Rosen C, Benedetti L, Jeppsson U. Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. Journal of Environmental Modelling and Software, 2011, 26(11): 1255–1267
6. Kim H S, Kim Y J, Cheon S P, Baek G D, Kim S S, Kim C W. Evaluation of model-based control strategy based on generated setpoint schedules for NH_4-N removal in a pilot-scale A^2/O process. Chemical Engineering Journal, 2012, 203: 387–397
7. Kim J R, Ko J H, Im J H, Lee S H, Kim S H, Kim C W, Park T J. Forecasting influent flow rate and composition with occasional data for supervisory management system by time series model. Water Science and Technology, 2006, 53(4-5): 185–192
8. Wang H R, Wang C, Kin X, Kang J. An improved ARIMA model for precipitation simulations. Nonlinear Processes in Geophysics, 2014, 21(6): 1159–1168
9. Valipour M, Banihabib M E, Behbahani S M R. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez Dam Reservoir. Journal of Hydrology (Amsterdam), 2013, 476: 433–441
10. Mohammadi K, Eslami H R. Dayyani Dardashti Sh. Comparison of regression ARIMA and ANN models for reservoir inflow forecasting using snowmelt equivalent (a case study of Karaj). Journal of Agricultural Science and Technology, 2005, 7: 17–30
11. Khashei M, Bijari M. A new hybrid methodology for nonlinear time series forecasting. Modelling and Simulation in Engineering, 2011, 2011: 1–5
12. Laio F, Porporato A, Revelli R,Ridolfi L. A comparison of nonlinear flood forecasting methods. Water Resources Research, 2003, 39(5): 1129
13. Karunasinghe D S K, Liong Sh. Chaotic time series prediction with a global model: artificial neural network. Journal of Hydrology (Amsterdam), 2006, 3232(1-4): 92–105
14. Solaimany-Aminabad M, Maleki A, Hadi M. Application of artificial neural network (ANN) for the prediction of water treatment plant influent characteristics. Journal of Advances in Environmental Health Research, 2013, 1(2): 89–100
15. Bagheri M, Mirbagheri S A, Bagheri Z, Kamarkhani A M. Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach. Process Safety and Environmental Protection, 2015, 95: 12–25
16. Grieu S, Traoré A, Polit M, Colprim J. Prediction of parameters characterizing the state of a pollution removal biologic process. Engineering Applications of Artificial Intelligence, 2005, 18(5): 559–573
17. Wu C L, Chau K W. Data-driven models for monthly streamflow time series prediction. Engineering Applications of Artificial Intelligence, 2010, 23(8): 1350–1367
18. Arroyo J, Maté C. Forecasting histogram time series with k-nearest neighbours methods. International Journal of Forecasting, 2009, 25 (1): 192–207
19. Imandoust S B, Bolandraftar M. Application of k-nearest neighbor (KNN) approach for predicting economic events: theoretical background. Int. Journal of Engineering Research and Applications, 2013, 3(5): 605–610
20. Ponomarenko A,Avrelin N, Naidan B, Boytsov L. Comparative analysis of data structures for approximate nearest neighbor search. Journal of Mathematical Sciences, 2012, 181(6): 782–791
21. Batista G E A P A, Silva D F. How k-nearest neighbor parameters affect its performance. In Argentine Symposium on Artificial Intelligence, 2009, 1–12
22. Alkasassbeh M, Altarawneh G A, Hassanat A. On enhancing the performance of nearest neighbour classifiers using hassanat distance metric. Canadian Journal of Pure and Applied Sciences, 2015, 9(1): 3291–3298
23. Tongal H. Nonlinear forecasting of stream flows using a chaotic approach and artificial neural networks. Earth Sciences Research Journal, 2013, 17(2): 119–126
24. Nesmerak I, Blazkova S D. Analysis of the time series of waste water quality at the inflow of the wastewater treatment plant and transfer functions. Journal of Hydrology and Hydromechanics, 2014, 62(1): 55–59
25. Yu X Y, Liong S Y. Babovic V. EC-SVM approach for real-time hydrologic forecasting. Journal of Hydroinformatics, 2004, 6(3): 209–233
26. Toth E,Brath A, Montanari A. Comparison of short-term rainfall prediction models for real-time flood forecasting. Journal of Hydrology (Amsterdam), 2000, 239(1): 132–147
27. Gou J, Du L, Zhang Y, Xiong T. A new distance-weighted k-nearest neighbor classifier. Journal of Information and Computational Science, 2012, 9: 1429–1436
28. Hassanat A B, Abbadi M A, Altarawneh G A, Alhasanat A A. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. 2014
29. Livio M. The Golden Ratio: The Story of Phi, the World's Most Astonishing Number. New York: Broad books, 2002
30. Han J, Kamber M. Data mining: concepts and techniques. Morgan Kaufmann publishers, San Francisco, 2001
31. Karegowda A G, Jayaram M A, Manjunath A S. Combining

Akaike's information criterion and the golden-section search technique to find optimal numbers of k-nearest neighbors. Journal of Computer Applications, 2010, 2(1): 80–87

32. Yanxia S, van Wyk B J, Wag Z. A new golden ratio local search-based particle swarm optimization. In: Proceedings of 2012 International Conference on Systems and Informatics, China. 2012, 754–757
33. Teimouri M. Comparison of neural network and k-nearest neighbor methods in daily flow forecasting. Journal of Applied Sciences, 2010, 10: 1006–1010