4th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Classification of surface water quality using data-driven methods

Sahar Javidan [1], Shokouh Mohsenzadeh [1], Mohammad Taghi Sattari *[1]

[1]University of Tabriz, Faculty of Agriculture, Water Engineering Department, Tabriz, Iran

Keywords
Bagging Method
Total Dissolved Solid
Water Quality Index

Abstract

Access to clean and quality water resources has been one of the main concerns of human beings for a long time. Therefore, determining the quality of water for various uses, including irrigation is very important. River pollution is one of the most important problems in the world today, especially in developing countries. In the present study, using data related to water quality parameters of Bagh Kalayeh hydrometric station in the 23-year statistical period, first the WQI index was calculated, then using data mining technique, factors affecting water quality were determined. Finally, the results of data mining methods were compared with the results obtained from the qualitative index. Quantitative results of the models were evaluated by Correlation Coefficient (R), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and qualitative results of the models were evaluated by Kappa, RMSE and MAE statistics. The results showed that in quantitative modeling, scenario 5 including TH, K, SO_4, TDS and EC and in qualitative modeling, scenario 3 including TH, K and SO_4 were selected as the superior scenario.

1. Introduction

Surface and groundwater pollution is one of the most important problems in the world and environmental concerns. In recent decades, due to rapid population growth, water needs and consequently pollution load to water sources have increased. There are several methods for classifying groundwater and surface water quality according to the type of consumption, one of the most widely used methods is the use of quality indicators. Due to the lack of facilities in all water quality monitoring stations and the need to save time and money, the use of alternative methods such as modern data mining methods can be a good way to predict and classify water quality.

Sattari et al. (2017) used data mining methods to predict surface water quality. They concluded that the tree decision model using the four parameters of Electrical Conductivity (EC), pH, Sodium Adsorption Ratio (SAR) and Sodium (Na) is able to classify water quality very accurately.

Babbar and Babbar (2017) predicted the river Water Quality Index using data mining techniques. They found that decision tree classifiers and Support Vector Machines were the best predictive models in determining water quality.

Gakii and Jepkoech (2019) used the decision tree model to classify and analyze water quality in Kenya. They introduced the J48 and Decision Stump decision trees as the most accurate and least accurate models, respectively. They found that analysis of water alkalinity, pH level and Electrical Conductivity could play an important role in assessing water quality.

Othman et al. (2020) predicted the river Water Quality Index by considering the minimum number of input variables. The results showed the exceptional ability of the artificial neural network model to calculate WQI. They also introduced Dissolved Oxygen (DO) as the most effective parameter in determining water quality.

The aim of this study is to calculate the WQI index using data related to water quality parameters of Bagh Kalayeh station in Qazvin province and then to use data mining techniques to determine the factors affecting water quality.

2. Method

Qazvin province is located in the northwestern part of Iran and its area is about 15820 km^2. Bagh Kalayeh is

* **Corresponding Author**

(javidansahar77115@gmail.com) ORCID ID 0000 – 0001 – 6739 – 8242
(Shkmsn2000@gmail.com) ORCID ID 0000 – 0003 – 3982 – 3082
*(mtsattar@gmail.com) ORCID ID 0000 – 0002 – 5139 – 2118

Cite this study

Javidan S., Mohsenzadeh, S., & Sattari, M. T. (2022). Classification of surface water quality using data-driven methods. 4th Intercontinental Geoinformation Days (IGD), 78-81, Tabriz, Iran

a village in the Rudbar Alamut section of Qazvin city in Qazvin province. Bagh Kalayeh hydrometric station is located at latitude 36°23′ 38″, longitude 50° 29′ 51″ and altitude 1287m above sea level. The average rainfall for 20 years at this station is 423.06mm. The location of the station under study is shown in Figure 1.

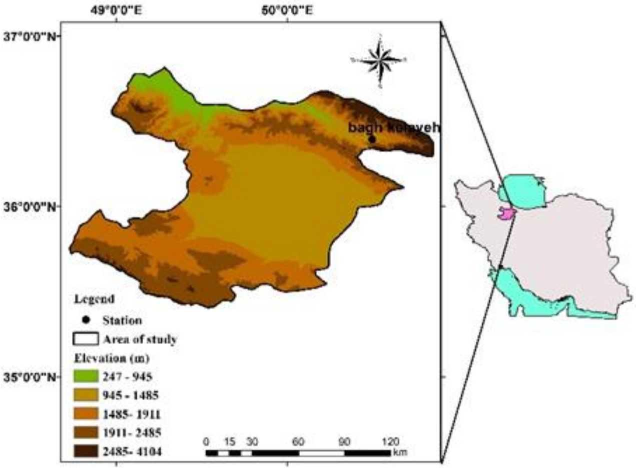

Figure 1. Location of the studied station

In the present study, to calculate the WQI index from the qualitative parameters of Bagh Kalayeh hydrometric station, including Total Hardness (TH), alkalinity (pH), Electrical Conductivity (EC), Total Dissolved Solid (TDS), Calcium (Ca), Sodium (Na), Magnesium (Mg), Potassium (K), Chlorine (Cl), Carbonate (CO_3), Bicarbonate (HCO_3) and Sulfate (SO_4) were used over a 23-year statistical period (1998-2020). The statistical characteristics of the variables used are presented in Table 1.

Table 1. Statistical characteristics of implemented parameters

Statistic	Minimum	Maximum	Mean
TH	95.00	481.50	278.15
PH	4.50	8.40	7.83
EC	279.00	1048.00	627.94
TDS	186.00	663.00	388.05
Ca	0.00	159.80	73.94
Na	0.46	60.49	17.26
Mg	2.76	58.20	22.18
K	0.39	19.50	1.99
Cl	0.00	89.60	27.36
CO_3	0.00	33.00	0.21
SO_4	22.08	374.88	140.21
HCO_3	50.02	391.62	163.60

Quantitative and qualitative values calculated with WQI index were considered as target outputs. Using the relief method, the types of input compounds (including the most effective parameters) were identified (Table 2). To estimate the quantitative values of WQI, the Bagging method was used with the Support Vector Regression algorithm (B-SVR) and for qualitative values, the Random Forest (RF) method was used. . Of the available data, 70% were considered for calibration and 30% for validation. Both methods were performed in Weka software.

Table 2. Parameters involved in each scenario

Scenario Number	Input Parameters
1	TH
2	TH, K
3	TH, K, SO_4
4	TH, K, SO_4, TDS
5	TH, K, SO_4, TDS, EC

2.1. Water Quality Index (WQI)

Drinking Water Quality Index was calculated using formulas 1 to 3. In these formulas, w is the weight of each parameter due to its importance in drinking and W is the relative weight of each parameter, C is the concentration of each parameter, S is the standard concentration of each parameter, q is the quality rank of each parameter and WQI is the drinking Water Quality Index (Singh 1992).

$$W_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

$$q_i = \left(\frac{C_i}{S_i}\right) \times 100 \tag{2}$$

$$WQI = \sum_{i=1}^{n} W_i q_i \tag{3}$$

Calculated WQI values are usually divided into five categories (Table 3).

Table 3. Water quality classification based on WQI value

Classification of Drinking Water Quality		
WQI Range	Class	Type of Water
below 50	I	Excellent water
50-100	II	Good water
100-200	III	Poor water
200-300	IV	Very poor water
above 300	V	Water unsuitable for drinking

2.2. Bagging Method

The Bagging method, first proposed by Breiman in 1996, connects several basic learners in parallel to reduce set variance. Each basic learner is trained on the same Bootstrap version using the same learning algorithm, then the output of these basic learners is aggregated by majority vote (for classification) or averaging (for regression) to obtain the final output. To achieve better and stronger performance, basic learners in a group must be precise and diverse (Breiman, 1996).

2.3 Support Vector Regression

Support Vector Machine (SVM) is a machine learning approach in data-driven research. This method is based on statistical learning theory and is used primarily for the best distinction between two data classes. Support Vector Machine models are divided into two main parts: (1) backup vector machine Bagging models, (2) backup vector regression model (SVR). The SVM model is used to solve the classification of data into different classes and the SVR model is used for forecasting (Demirci 2019).

2.4. Random Forest

The RF algorithm is a supervised classification algorithm. There is a direct relationship between the number of algorithm trees and the results that can be obtained. As the number of trees increases, a definite result is obtained. The difference between the RF algorithm and the decision tree algorithm is that root node detection and node splitting in RF are performed randomly. This is why the RF algorithm can be used in classification and regression tasks (Sachetana et al. 2017).

To compare the values obtained from data mining methods with the values calculated from the WQI index, the criteria of Correlation Coefficient (R), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Kappa statistics were used. The formulas of the above statistics are presented in Equations (4) to (7), respectively:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 .\ \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{6}$$

$$Kappa = p_i = (PA_0 - PA_E)/(1 - PA_E) \tag{7}$$

In the above relations, y_i is the estimated value of the model, x_i is the value calculated from the qualitative index, N is the number of data, M is the number of samples found in the wrong class, T is the total number of samples, PA_0 is the agreement of the two evaluators and PA_E is the expected agreement.

3. Results

First, the results obtained from the 5 input scenarios used in the Bagging method with the backup vector regression algorithm for estimating quantitative values and the Random Forest method for estimating qualitative values were presented in Tables 4 and 5, respectively:

Table 4. Evaluation criteria for estimating quantitative WQI values

Scenario	B-SVR		
	R	RMSE	MAE
1	0.96	3.01	2.28
2	0.96	2.82	2.08
3	0.97	2.70	2.08
4	0.98	2.06	1.39
5	0.98	1.92	1.23

Table 5. Evaluation criteria for estimating WQI quality values

Scenario	RF		
	Kappa	RMSE	MAE
1	1	0.0089	0.0008
2	1	0.0089	0.0018
3	1	0.0055	0.0007
4	1	0.0156	0.0026
5	1	0.0222	0.0046

According to Tables 4 and 5, however, the method has provided acceptable results in all scenarios. To select the best scenario, bar graphs of RMSE and MAE values for both methods are shown in Figures 2 and 3.

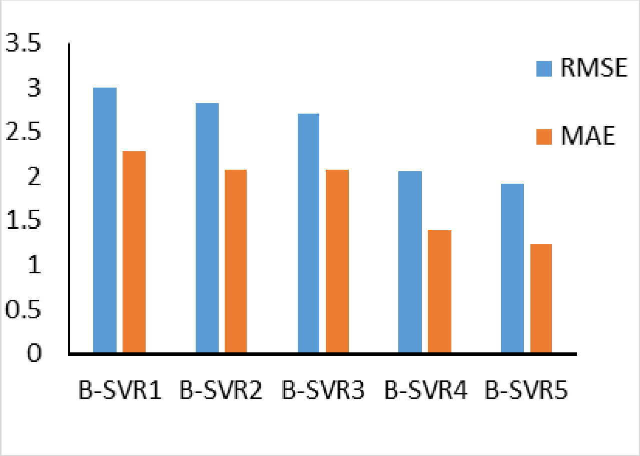

Figure 2. Bar chart of quantitative modeling errors

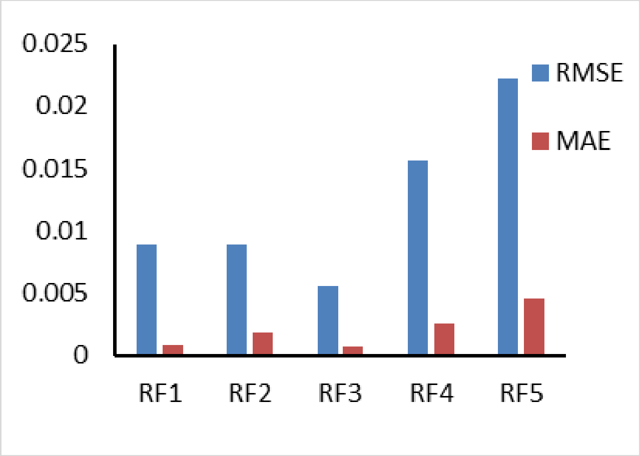

Figure 3. Bar chart of qualitative modeling errors

According to Figure 2 and 3, Scenario 5 (including TH, K, SO4, TDS, EC) with the lowest error rate as the superior scenario for estimating quantities of WQI and Scenario 3 (including TH, K, SO4) as Scenario Superior was selected to estimate WQI quality values.

4. Discussion

Calculating Water Quality Index with a large number of parameters is time consuming and difficult. However, data-based methods with a very small number of parameters provide more acceptable results, and this increases the popularity of data-based methods. The results of the present study showed that the data mining methods using the parameters TH, K, So_4, TDS, EC instead of the 12 parameters used in calculating the Water Quality Index, had considerable accuracy in estimating the quantitative and qualitative values of WQI.

5. Conclusion

In the present study, first Water Quality Index using parameters of Total Hardness (TH), alkalinity (pH), Electrical Conductivity (EC), Total Dissolved Solid (TDS), Calcium (Ca), Sodium (Na), Magnesium (Mg) , Potassium (K), Chlorine (Cl), Carbonate (CO_3), Bicarbonate (HCO_3) and Sulfate (SO_4) were calculated. Then, to estimate the quantitative values of WQI, the Bagging method was used with the basic vector regression algorithm, and to estimate the qualitative values, the Random Forest method was used, taking into account different scenarios. The results showed that B-SVR5 and RF3 methods had good accuracy for quantitative and qualitative estimation of WQI index in Bagh Kalayeh hydrometric station, respectively.

References

Babbar, R., & Babbar, S. (2017). Predicting river water quality index using data mining techniques. Environmental Earth Sciences, DOI 10.1007/s12665-017-6845-9.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24, 123–140.

Demirci, M. (2019). Estimation of rainfall-runoff relationship using support vector machines and M5 decision tree methods. DÜMF Mühendislik Dergisi, 10 (3), 1113–1124.

Gakii, C., & Jepkoech, J. (2019). A classification model for water quality analysis using desicion tree. European Journal of Computer Science and Information Technology, 7 (3).

Othman, F., Alaaeldin, M., Seyam, M., Ahmed, A., Teo, F., Ming Fai, Ch., Afan, H., Sherif, M., Sefelnasr, A., & Shafie, A. (2020). Efficient river water quality index prediction considering minimal number of inputs variables. Engineering Applications of Computational Fluid Mechanics, 14 (1), 751-763. https://doi.org/10.1080/19942060.2020.1760942.

Sattari, M. T., Mir Abbasi, R., & Nayebzadeh, M. (2017). Use of Mining in Predicting Surface Water Quality (Case Study: North Sahand Rivers). Echo Hydrology, 4 (2), 419-407.

Suchetana, B., Rajagopalan, B., & Silverstein, J. (2017). Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits using a regression tree model. Science of The Total Environment, 598, 249–257.