4th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Estimation of Mordagh Chay River water quality using gaussian process regression method

Kambiz Falsafian *1

[1]*University of Tabriz, Marand Technical College, Tabriz, Iran*

Keywords
Gaussian process regression
Mordagh Chay River
Water quality

Abstract

Accurate modeling of river quality parameters is essential for environmental planning, optimal operation of reservoirs, designing of hydraulic structures and irrigation planning. Considering that direct measurement of quality parameters is time consuming and costly, it is possible to predict these parameters with less time and cost and with high accuracy using artificial intelligence methods. In this regard and in the present study, the electrical conductivity parameter of Mordagh Chay River has been estimated using the gaussian process regression method. Based on this, the amounts of calcium, magnesium, sodium, chlorine and sulfate of this river on a monthly scale over a period of 47 years (1971-2018) were used as input parameters of the models. Statistical parameters of correlation coefficient, scattering index and Wilmott's index were also used to compare the obtained results with the observed values. Finally, the obtained results showed that the GPR5 model with SI of 0.093 and WI of 0.995 had the best performance. It was also generally concluded that using the models used in this study, the EC value in the Mordagh Chay River can be estimated with appropriate accuracy.

1. Introduction

Rivers are the main and most important source of fresh water for various industrial, drinking and agricultural usages, and all over the world, the high quality of river water is very important and vital. Various environmental factors such as the construction of industrial factories, population growth and the use of pesticides in agricultural lands have affected the quality of river water. Therefore, the study of variability of water quality criteria along a river has been researched and considered by researchers as modeling and predicting water quality to provide appropriate solutions to control and reduce river pollution. Also, one of the key factors in water resources management and fresh water processing in accordance with urban, industrial and agricultural needs is modeling and estimating water quality. Various criteria are used to indicate water quality. Water salinity is one of the most important criteria that is measured by the parameter of electrical conductivity (EC). In this regard, classical statistical methods have been used to develop water resources management and investigate changes in river water quality. In recent years, machine learning methods have been used significantly in predicting various parameters of water resources due to their accuracy and the need for low cost and time. Studies of Yesilnacar et al. (2008), Rankovic et al. (2010), Emamgholizadeh et al. (2013) and Shokoohi et al. (2017) are examples of research conducted in the field of modeling water quality parameters.

Haghiabi et al. (2018) examined the methods of artificial neural network (ANN), group method of data handling (GMDH) and support vector machine (SVM) in predicting the water quality of the Tireh River located in southwestern Iran. Comparison of the results by error indices showed the superiority of the SVM model over other models. Najah Ahmad et al. (2020) utilized techniques of Adaptive Neuro-Fuzzy Inference System (ANFIS), Radial Basis Function Neural Networks (RBF-ANN) and Multi-Layer Perceptron Neural Networks (MLP-ANN) for estimating water quality parameters in the Johor River Basin in Malaysia. They also used a hybrid method of Neuro-Fuzzy Inference System based augmented wavelet de-noising technique (WDT-ANFIS) to increase the accuracy and showed that in the validation section, the proposed model had a satisfactory

* **Corresponding Author**

*(kfalsafian@tabrizu.ac.ir) ORCID ID 0000 – 0002 – 1567 – 7976

Cite this study

Falsafian, K. (2022). Estimation of Mordagh Chay River water quality using gaussian process regression method. 4th Intercontinental Geoinformation Days (IGD), 58-61, Tabriz, Iran

performance in estimating all water quality parameters. Melesse et al. (2020) predicted salinity in the Babol River in northern Iran using random forest and M5P methods and eight new hybrid algorithms. In this research, the parameters PH, TDS, flow rate and some cations and anions were used as input of the models. Using the results, it was shown that hybrid models increase accuracy of the single models.

According to studies on water quality parameters, machine learning methods in most researches have had accurate and desirable results. Therefore, in this study, with the aim of modeling water quality in Mordagh Chay River located in Maragheh city, Gaussian process regression technique has been used and the amount of electrical conductivity in the study area has been estimated to manage and protect the water quality of this river using the obtained results.

2. Method

2.1. Study area

Mordagh Chay is a river located in northwestern Iran in the province of East Azerbaijan. This river originates from Sahand mountains and after passing through the lands of Maragheh and Malekan cities, it reaches a branch of Zarrineh river. In this study, water quality data Mordagh Chay at Gheshlagh Amir station, which is located at 46° 17′ longitude and 37° 18′ latitude, has been used. The data used in this study include the parameters of calcium (Ca), magnesium (Mg), sodium (Na), chlorine (Cl), sulfate (SO4) and electrical conductivity (EC) on a monthly scale and over a period of 47 years, from 1971 to 2018, so that the mentioned parameters in different combinations have been used as input of models to estimate EC. Table 1 shows the different combinations of input parameters of the studied models.

Table 1. Different combinations of input parameters of the studied models

Combination Number	Input Parameters	Output Parameter
1	Ca, Mg	EC
2	Ca, Na	EC
3	Mg, Na	EC
4	Ca, Mg, Cl	EC
5	Ca, Mg, Na	EC
6	Ca, Mg, Na, SO_4	EC
7	Ca, Mg, Na, Cl	EC
8	Ca, Mg, Na, Cl, SO_4	EC

2.2. Gaussian process regression (GPR)

Gaussian process regression models are based on the assumption that adjustment observations should carry information about each other. Gaussian processes are a way to specify a priority directly on the function space. This is a natural generalization of the Gaussian distribution which mean and covariance are vectors and matrices, respectively (Yang et al. 2018). The Gaussian distribution is on vectors while the Gaussian process is on functions. As a result, Gaussian process models do not need any validation process to generalize due to prior knowledge of functional dependencies and data, and Gaussian process regression models are able to understand the prediction distribution corresponding to the test input (Pal and Deswal, 2010).

Consider the set S with n observations $S = \{(x_i, y_i) | i = 1, \ldots, n\}$, in which $x_i$is the input vector with D dimension and $y_i$is the output with scalar or target. This set consists of two components, input and output, as sample or experimental points. For ease of operation, the inputs of the set are aggregated in $X = [x_1, x_2, x_3, \ldots, x_n]$matrix and the outputs are also collected in $Y = [y_1, y_2, y_3, \ldots, y_n]$matrix. The regression task is to create a new x^*input in order to achieve the predicted distribution for y^*corresponding values of the observational data and based on the S data set. The Gaussian process is a set of random variables, a limited number of which are integrated with Gaussian distributions. The Gaussian process is a generalization of the Gaussian distribution. The Gaussian distribution is actually the distribution between random variables, while the Gaussian process represents the distribution between functions. The Gaussian process is defined by the functions of mean and covariance in the form of Equations 1 and 2:

$$m(x) = E(f(x)) \tag{1}$$

$$k(x, x') = E(f(x) - m(x))(f(x') - m(x')) \tag{2}$$

which $k(x, x')$ is a function of covariance (or kernel) that is calculated in x and x'points. The Gaussian process can be expressed as Equation 3:

$$f(x) \square GP(m(x), k(x, x')) \tag{3}$$

Which is usually considered to be zero to simplify the value of the average function.

2.3. Performance evaluation criteria of models

Error values between the studied models and observational data were estimated and evaluated by correlation coefficient (R), scatter index (SI) and Wilmott's index (WI) using Equations 4 to 6.

$$R = \frac{\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right)} \tag{4}$$

$$SI = \frac{RMSE}{\overline{y}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)^2}}{\overline{y}} \tag{5}$$

$$WI = 1 - \left[\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{\sum_{i=1}^{n} \left(\left| x_i - \overline{y} \right| - \left| y_i - \overline{y} \right| \right)^2} \right] \qquad (6)$$

In Equations 4 to 6, $x_i$and $y_i$are the predicted and observed monthly electrical conductivity values, respectively and n is the number of observations.

3. Results

In this study, in order to model the electrical conductivity in the Mordagh Chay River, the parameters of calcium (Ca), magnesium (Mg), sodium (Na), chlorine (Cl) and sulfate (SO4) were used as input data and the EC parameter was estimated as the output of the models. Therefore, based on the correlation of input parameters with EC, eight different scenarios were defined by combining different input data for the studied models and the obtained results were compared using the statistical parameters of correlation coefficient, scatter index and Wilmott's index with observational data and superior models were introduced. Table 2 shows the statistical indicators of the studied models.

Table 2. Statistical indicators of studied models

Model	R	SI	WI
GPR1	0.992	0.105	0.995
GPR2	0.984	0.184	0.985
GPR3	0.913	0.396	0.881
GPR4	0.992	0.109	0.994
GPR5	0.991	0.093	0.995
GPR6	0.989	0.107	0.994
GPR7	0.989	0.103	0.994
GPR8	0.986	0.124	0.992

According to Table 2, the obtained results show that the GPR5 model with correlation coefficient of 0.991, scatter index of 0.093 and Wilmott's index of 0.995 had the best results. The second place was taken by GPR1 model with correlation coefficient of 0.992, scatter index of 0.55 and Wilmott's index of 0.995 with only two input parameters (Ca and Mg). Also, GPR7 model with correlation coefficient of 0.989, scatter index of 0.103 and Wilmott's index of 0.994 had good accuracy and was ranked third. In general, the results show that the developed models have appropriate and acceptable accuracy in modeling the EC values of the Mordagh Chay River.

4. Discussion

In order to better understand the performance of the superior models, the diagram of monthly changes of EC using the best models (Figure 1) and the graph of the distribution of computational EC values with the superior studied models in comparison with the observational EC (Figure 2) are given.

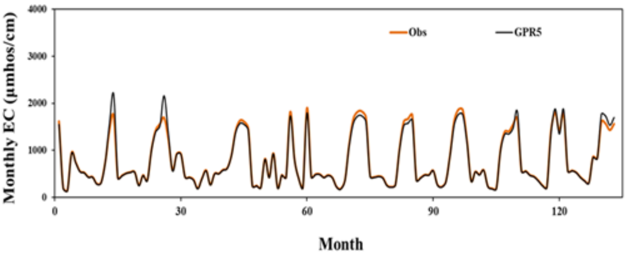

Figure 1. Diagram of monthly changes of EC using the best implemented model

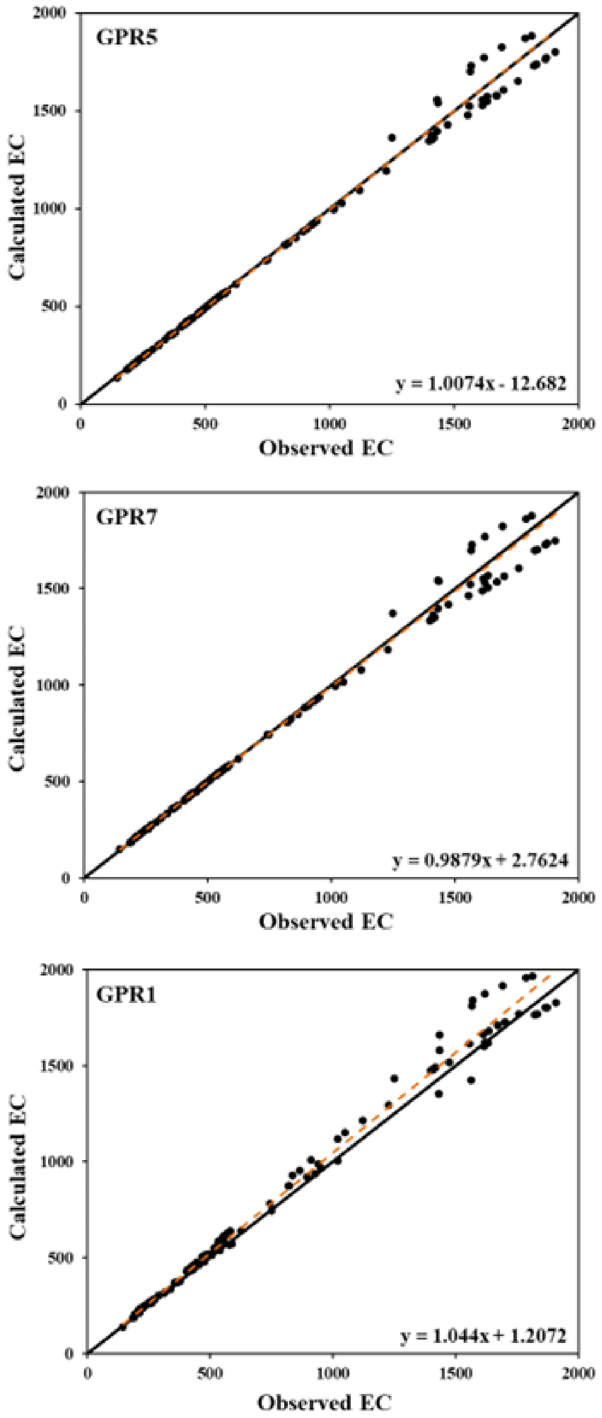

Figure 2. Distribution diagram of computational EC values with the superior studied models in comparison with observational EC

According to Figure 1, the high agreement of the GPR5 model with the observational data can be deduced. In Figure 2, the distribution of points around the bisector axis is lower in superior models which have less error and are closer to the observational values of EC.

5. Conclusion

Pollution and low water quality of rivers directly affect the environment and human life and estimation and determination of quality parameters of river water has a significant role in the management of water resources. Therefore, in the present study, machine learning method, Gaussian process regression was used to model and predict the electrical conductivity of the Mordagh Chay River. Hence, the anions and cations (calcium, magnesium, sodium, chlorine, sulfate) of this river were used as input data over a period of 47 years. The obtained results showed that GPR5 model with scatter index of 0.093 and Wilmott's index of 0.995 had the best results. It was also shown that all implemented models used were successful in estimating the EC value. In general, it is concluded that using GPR method, the EC parameter and water quality in the Mordagh Chay River can be modeled and estimated with low error and desirable accuracy.

References

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. Journal of Hydrology, 578, 124084, 1-18.

Emamgholizadeh, S., Kashi, H., Marofpoor, I., & Zalaghi, E. (2013). Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. International Journal of Environmental Science and Technology, 11 (3), 645-656.

Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. Water quality research journal of Canada, 53(1), 3-13.

Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Thai Pham, B. (2020). River water salinity prediction using hybrid machine learning models. Water, 12, 2951, 1-21.

Pal, M., & Deswal, S. (2010). Modelling pile capacity using Gaussian process regression. Computer. Geotechnical, 37, 942-947.

Ranković, V., Radulović, J., Radojević, I., Ostojić, A., & Čomić, L. (2010). Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. Ecological Modelling, 221 (8), 1239–1244.

Shokoohi, M., Tabesh, M., Nazif, S., & Dini, M. (2017). Water quality based multi-objective optimal design of water distribution systems. Water Resources Management, 31 (1), 93-108.

Yang, D., Zhang, X., Pan, R., Wang, Y., & Chen, Z. (2018). A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve. Journal of Power Sources, 384, 387–395.

Yesilnacar, M. I., Sahinkaya, E., Naz, M., & Ozkaya, B. (2008). Neural network prediction of nitrate in groundwater of Harran Plain, Turkey. Environmental Earth Sciences, 56(1), 19– 25.