DOI: 10.1515/eces-2019-0051 ECOL CHEM ENG S. 2019;26(4):727-742

Nurtac OZ[1*], Bayram TOPAL[2] and Halil Ibrahim UZUN[3]

PREDICTION OF WATER QUALITY IN RIVA RIVER WATERSHED

PROGNOZOWANIE JAKOŚCI WODY W ZLEWNI RZEKI RIVA

Abstract: The Riva River is a water basin located within the borders of Istanbul in the Marmara Region (Turkey) in the south-north direction. Water samples were taken for the 35 km drainage area of the Riva River Basin before the river flows into the Black Sea at 4 stations on the Riva River every month and analyses were carried out. Changes were observed in the quality of water from upstream to downstream. For this purpose, the spatial and temporal variations of water quality were investigated using 13 water quality variables with the ANOVA test. It was observed that COD, *DO*, S and BOD were important in determining the spatial variation. On the other hand, it was found out that all the variables were effective in determining the temporal variation. Moreover, the correlation analysis which was carried out in order to assess the relations between water quality variables showed that the variables of BOD-COD, BOD-*EC*, COD-*EC*, BOD-*T* and COD-*T* were correlated and the regression analysis showed that COD, *TKN* and NH_4-N explained BOD and BOD, NH_4-N, *T* and *TSS* explained COD by approximately 80 %. Consequently, the Artificial Neural Network (*ANN*), Decision Tree and Logistic Regression models were developed using the data of training set in order to predict the water quality classes of the variables of COD, BOD and NH_4-N. Quality classes were predicted for the variables by inputting the data of testing set into the developed models. According to these results, it was seen that the *ANN* was the best prediction model for COD, the Decision Tree for BOD and the *ANN* and Decision Tree for NH_4-N.

Keywords: Riva river, water quality, artificial neural network (*ANN*), decision tree, regression models

Introduction

Water quality in river ecosystems undergo rapid transformations based on natural factors affecting the basin (precipitation, weather, basin physiography, soil erosion, etc.) and anthropogenic factors (urbanization, industrial and agricultural activities, etc.) [1-4]. This situation negatively affects water quality criteria, biodiversity and the ecological health of rivers. Particularly nitrate and phosphate damage aquatic life by reducing water quality when they are present in excessive amounts although they naturally exist in freshwaters [5, 6].

[1] Environmental Engineering Department, Sakarya University, Esentepe Campus 54187 Sakarya, Turkey, phone +90 264 295 39 13, fax +90 264 295 56 01

[2] Business Administration Department, Sakarya University, Esentepe Campus 54187 Sakarya, Turkey

[3] Environmental Engineering Department, Yildiz Technical University, Davutpasa Campus, 34220 Istanbul, Turkey

[*]Corresponding author: nuroz@sakarya.edu.tr

The spatial and temporal variability of river ecosystems affects the physico-chemical characterization of water. Basin management plans are based on the assessment of water quality by using the field data in the study area. Therefore, it is inevitable to study temporal variations as well as spatial variations in water quality in order to better investigate and assess the water quality of basins [7-9]. Most water quality models reviewed in the literature analyze the spatial variability of different water constituents in relation to natural or anthropogenic factors in a specific catchment basin [10-12]. However, the analysis of temporal variability has also gained prominence in recent studies [13-16]. Seasonal variations in water quality can be used in order to increase our understanding as to how degradations in water quality occur and thus to design more effective restoration programs. Continuous and regular monitoring programs are required to gain reliable knowledge about natural characteristics of water quality and to understand the physico-chemical, spatial and temporal variations of water. However, the databases established are broad and complicated. Therefore, statistical techniques are widely used to assess spatial and temporal variations and to interpret large and complicated data sets of water quality [17-21]. The *ANN* has become a new tool and an effective model to predict various water quality variables in river systems [16, 22, 23].

In Turkey, water pollution and its effects are frequently observed particularly in the Marmara Region, where intensive industrial activities take place. Food and metal industry are among the most important industrial facilities within the Marmara Basin. The heavy industrial activities in the Basin lead to pollution and the water resources available to meet the demand are scarce, which requires the employment of a special water management approach within the basin.

The Riva River is a water basin located within the borders of Istanbul in the Marmara Region (Turkey) in the south-north direction. This study investigated how the anthropogenic factors of the Riva Basin influence water quality. Regression models were used to predict the contribution of potential pollution sources to the concentration of the selected water quality parameters and to determine the relationships between variables. Moreover, prediction models were established in order to determine the water quality classes of the parameters of COD (Chemical Oxygen Demand), BOD (Biochemical Oxygen Demand), NO_3-N (Nitrate Nitrogen), and NH_4-N (Ammonium Nitrogen) and their success was assessed.

Given the above considerations, the main purpose of this study is to better understand the spatial and temporal variability of the water quality of the Riva River. It is considered that the results will be beneficial for local authorities for pollution control and management and for better protection of the quality of river water.

Material and methods

The Marmara Region is one of seven geographical regions of Turkey and has a surface area of 67.000 km^2. It is the most developed region of Turkey in terms of industry because of its coastline, port facilities and the existence of sea, which meets its need of water.

Istanbul is the most populous and also the most important city of the Marmara region in economic, historical and socio-cultural terms. The Riva River Watershed is a drainage basin approximately 70 km long, located on the Anatolian side of Istanbul in the south-north direction. The lower 35 km of the river between Omerli Dam and the Black Sea is used for industrial wastewater discharge. There are no large rivers in Istanbul and the largest one is the Riva River. In this study, the water samples were collected monthly at four different stations chosen in the Riva River, which is around 35 km long.

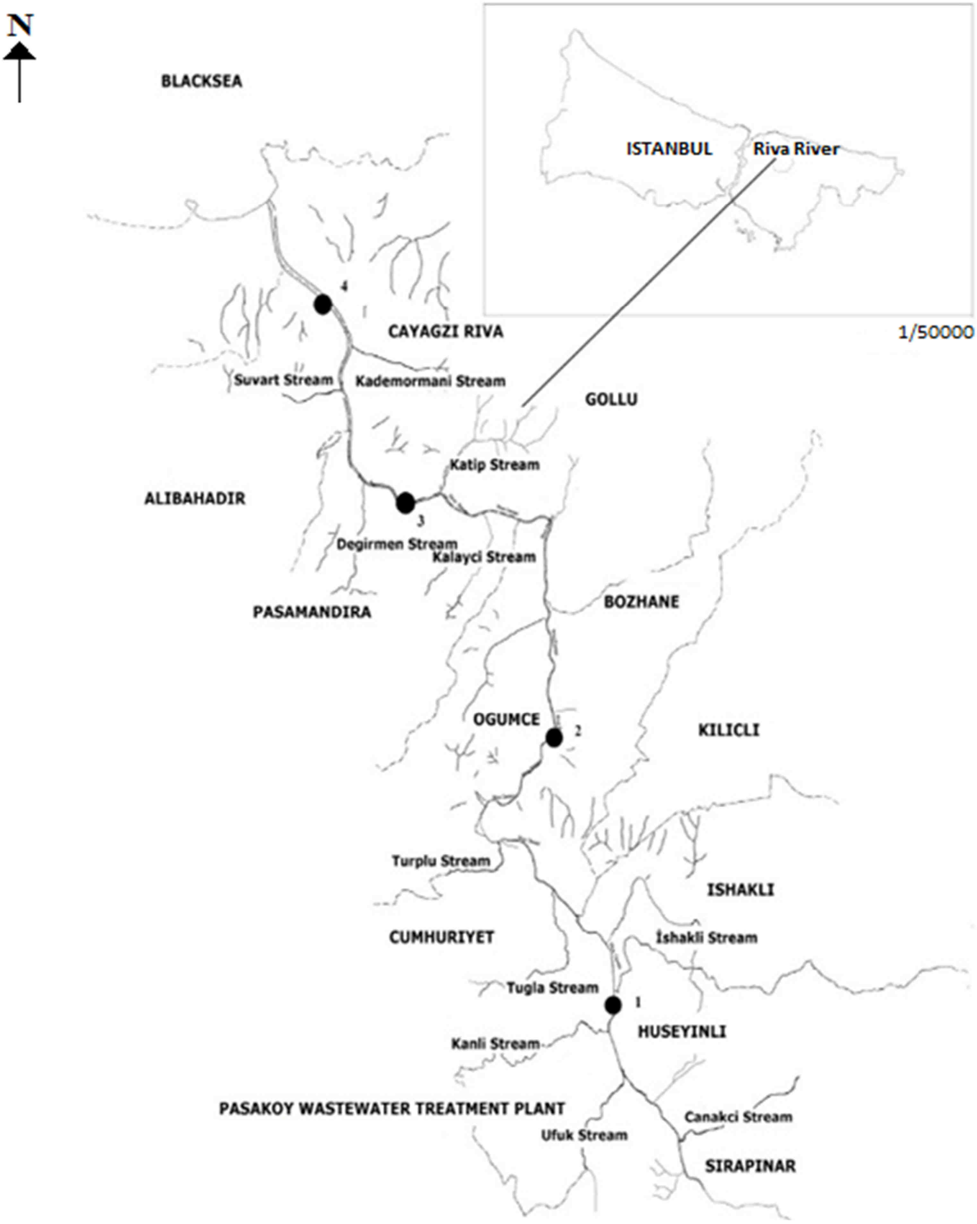


Fig. 1. Sampling stations of the Riva River

At the first sampling station of the study area, the existing water flow receives the discharge of the Pasakoy Advanced Biological Wastewater Treatment Plant, which treats the wastewater of Umraniye, Sancaktepe, Sultanbeyli and the neighborhood, tributary streams and surface water flows. There are croplands and private hobby gardens around the second sampling station. Furthermore, a dog farm, a paper mill and a plastics plant are located in the area. There are luxury residential areas around the third sampling station. Most of the wastewater from the buildings around the station directly enters into the river. Woodlands, dense reed beds, and recreational areas also exist around the station. There is a plant which produces detergents and chemical products before, as well as a restaurant and picnic areas around, the fourth sampling station. It is thought that the selected stations are representative of this section of the Riva Basin.

The Riva River System Drainage Basin and sampling stations are shown in Figure 1.

Water samples were taken at various stations on the Riva River and they were analyzed. The measurements for the water quality parameters of pH, temperature T [°C], dissolved oxygen DO [$\mathrm{mg \cdot dm^{-3}}$], electrical conductivity EC [$\mathrm{mS \cdot m^{-1}}$], salinity S [$\mathrm{g \cdot dm^{-3}}$], total suspended solids TSS [$\mathrm{mg \cdot dm^{-3}}$], total volatile suspended solids $TVSS$ [$\mathrm{mg \cdot dm^{-3}}$], chemical oxygen demand COD [$\mathrm{mg \cdot dm^{-3}}$], biochemical oxygen demand BOD [$\mathrm{mg \cdot dm^{-3}}$], total phosphorus TP [$\mathrm{mg \cdot dm^{-3}}$], nitrate nitrogen $\mathrm{NO_3}$-N [$\mathrm{mg \cdot dm^{-3}}$], ammonium nitrogen $\mathrm{NH_4}$-N [$\mathrm{mg \cdot dm^{-3}}$] and total Kjeldahl nitrogen TKN [$\mathrm{mg \cdot dm^{-3}}$] were performed on the samples.

A Lutron Oxygen Meter was used for the values of T and DO, a WTW Cond 315i Conductivity Meter for the EC values, a 315i WTW pH meter for the pH value and an YSI Model 30 Salinity Meter for the S value, and these values were measured in the field.

The gravimetric method was used for the analysis of TSS and $TVSS$ [24]. The open reflux method was used for the analysis of COD, the OxiTop method was used for the analysis of BOD, the colorimetric method for the analysis of TP, the cadmium reduction method for the analysis of $\mathrm{NO_3}$-N, the ion selective electrode method for the analysis of $\mathrm{NH_4}$-N, and the total Kjeldahl nitrogen method for the analysis of TKN [25].

Data evaluation

The spatial and temporal variations of water quality variables of the Riva River were analyzed through the analysis of variance. The relationships between the variables, on the other hand, were determined using correlation and regression analysis. Furthermore, prediction models for water quality class were established using the artificial neural networks, logistic regression and decision tree.

Samples were collected from four different sampling stations during four seasons for 18 months (once a week) in order to determine the variation of the Riva River water quality in different areas of the basin and in different seasons. The data of 13 quality variables analyzed were assessed using the analysis of variance. One of the important assumptions of the analysis of variance is that the distribution of the data is normal. Two important assumptions for applying variance analysis to a data are the normality of the data and the homogeneity of the variances. The normality of the data was investigated with the Kolmogorov-Smirnov and Shapiro-Wilk tests [21]. It was established that the original values of all the variables apart from COD did not distribute normally. The variables which did not display a normal distribution were normalized by applying logarithmic and square root transformation methods. Levene test was applied for homogeneity of variances. Most of the quality variables showed homogeneity assumption. Thus the analysis of variance was applied for all the variables that displayed a normal distribution and homogeneity of variances.

Predictions were made for the Artificial Neural Network model using the Clementine 10.5 software and the methods of Quick, Dynamic, Multiple, Prune, Radial Basis Function Network (*RBFN*), and Exhaustive Prune. The relevant methods were listed using the Clementine 10.5 software.

A randomly selected 60 % portion of the data set was set as the training set and the remaining 40 % as the test set. Prediction models for water quality classes using the water quality parameters of COD, BOD, NO_3-N and NH_4-N were formed with the training set. The remaining data were used as the test set. The quality class of each quality variable was predicted separately with the test set and the success of the models established was determined.

Results and discussion

Analysis of the variation in water quality with regard to sampling stations and seasons

It was investigated whether the water quality variables varied in terms of regions and seasons through the analysis of variance. It is first necessary that the data display a normal distribution for the implementation of the analysis of variance. It was observed that the variables except for COD did not distribute normally. The logarithmic and square root transformation methods were applied to normalize the other variables and thus it was seen that the variables of *DO*, *S*, *TSS*, *TVSS*, BOD, *TP*, NO_3-N and *TKN* were normalized. The results are given in Table 1.

Table 1

Tests of normality

Chemical quality variables	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	Degree of freedom	Significant	Statistic	Degree of freedom	Significant
COD	.067	132	.200*	.980	132	.045
ln pH	.201	132	.000	.715	132	.000
ln *T*	.133	132	.000	.876	132	.000
ln *DO*	.034	132	.200*	.992	132	.654*
ln *S*	.048	132	.200*	.988	132	.280*
ln *EC*	.168	132	.000	.918	132	.000
ln *TSS*	.068	132	.200*	.983	132	.101*
ln *TVSS*	.074	132	.103*	.978	132	.030
ln BOD	.087	132	.016	.986	132	.177*
ln *TP*	.076	132	.087*	.967	132	.002
ln NO_3-N	.060	132	.200*	.978	132	.029
Square root TKN	.064	129	.200*	.979	129	.039
NH_4-N	.121	132	.000	.964	132	.001

* this is a lower bound of the true significance

The data must be homogeneous for the analysis of variance. Table 2 gives the results of the Levene test for the equality of variance by measuring stations and seasons.

Table 2

Test of homogeneity of variances

	Levene test for measuring stations				**Levene test for seasons**			
	Levene statistic	**Degree of freedom^1**	**Degree of freedom^2**	**Significant**	**Levene statistic**	**Degree of freedom^1**	**Degree of freedom^2**	**Significant**
COD	1.163	3	128	.327*	2.082	3	128	.106*
pH^1	.699	3	128	.554*	.985	3	128	.402*
T^1	.802	3	128	.495*	20.680	3	128	.000
DO^1	2.609	3	128	.054*	.799	3	128	.497*
S^1	2.794	3	128	.042	2.480	3	128	.064*
EC^1	3.612	3	128	.015	7.351	3	128	.000
TSS^1	.640	3	128	.590*	15.415	3	128	.000
TVSS^1	.442	3	128	.723*	11.108	3	128	.000
BOD^1	.742	3	128	.529*	.677	3	128	.568*
TP^1	.275	3	128	.844*	1.126	3	128	.341*
NO_3-N^1	.274	3	128	.844*	13.198	3	128	.000
TKN^2	25.587	3	125	.000	2.190	3	125	.093*
NH_4-N^1	1.112	3	128	.347*	3.924	3	128	.010

$^{(1)}$ logarithmic values, $^{(2)}$ square root values, $^{(*)}$ homogeneous variance, the analysis of variance was applied on the normalized variables.

Analysis of the differences of water quality parameters with regard to spatial and temporal

It was investigated whether the quality variables displayed temporal and spatial variations using the analysis of variance (ANOVA). Therefore, it was revealed what type of variation the water quality of the river displayed in both regional and seasonal terms. As it is known, the *t* test is an analysis technique used to test the difference between the averages of two groups or two categories. If the number of groups or categories is higher than two, the difference between the averages is tested with the analysis of variance (ANOVA).

Two hypotheses were used in order to test whether the values of the quality variables differed according to the sampling stations and seasons.

Hypothesis 1 (H_1): The values of the related quality variable differs according to the regions.

Hypothesis 2 (H_2): The values of the related quality variable differs according to the seasons.

The results of the analysis of variance (ANOVA) showing the values of quality variables according to the sampling stations (spatial) and seasons (temporal) are given in Table 3.

As can be seen in Table 3, Hypothesis 1 was accepted for four water quality variables. COD and *S* displayed significant differences at a significance level of 1 % and *DO* and BOD at a significance level of 5 % according to the sampling stations. Then these four quality variables take on different values in different areas of the Riva River. The other water quality variables did not show significant differences according to the sampling stations. In that case, all the variables apart from COD, *S*, *DO* and BOD show similar quality characteristics throughout the river. BOD is the dominant chemical parameter that increases *DO* consumption in the river. As the BOD increases, the saturation level of *DO* in the river reaches the minimum. Therefore, BOD is important among water quality parameters [26]. According to other studies, COD and *DO* have excellent performance in reflecting the water quality of the basin [27].

Table 3
The ANOVA test for the difference of the values of water quality variables according to the sampling stations (spatial) and seasons (temporal)

Source	**Dependent variable**	**Sum of squares**	**Degree of freedom**	**Mean square**	**F value**	**Significance level**
Spatial	COD	7590.979	3;122	2530.326	7.275	.000**
	DO	.291	3;122	.097	2.851	.040*
	S	.640	3;122	.213	8.868	.000**
	TSS	.023	3;122	.008	.150	.930
	TVSS	.067	3;122	.022	.325	.807
	BOD	.296	3;122	.099	3.783	.012*
	TP	.038	3;122	.013	.269	.848
	NO_3 -N	.064	3;122	.021	1.422	.240
	TKN	.592	3;122	.197	.502	.681
Temporal	COD	17302.440	3;122	5767.480	16.581	.000**
	DO	1.430	3;122	.477	14.022	.000**
	S	2.596	3;122	.865	35.988	.000**
	TSS	1.370	3;122	.457	8.908	.000**
	TVSS	1.254	3;122	.418	6.130	.001**
	BOD	1.972	3;122	.657	25.244	.000**
	TP	.454	3;122	.151	3.202	.026*
	NO_3 -N	1.291	3;122	.430	28.521	.000**
	TKN	4.164	3;122	1.388	3.536	.017*

* Coefficient is significant at the 0.05 level, ** Coefficient is significant at the 0.01 level

The same Table also shows that Hypothesis 2 was also accepted for all the variables. According to this table, *TKN* showed significant differences according to the seasons at a significance level of 5 % and the other variables at a significance level of 1 %. Then it means all water quality variables were affected by seasonal variations. This is reported to be the highest in autumn, lowest in spring and summer and winter [27].

The unit difference was eliminated by converting the values of the quality variables measured with different units into a standard unit (*Z*). Thus, it became possible to display the spatial and temporal variation of quality variables on the same graph. The Z_i conversion (Eq. (1)); was performed as follows:

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{1}$$

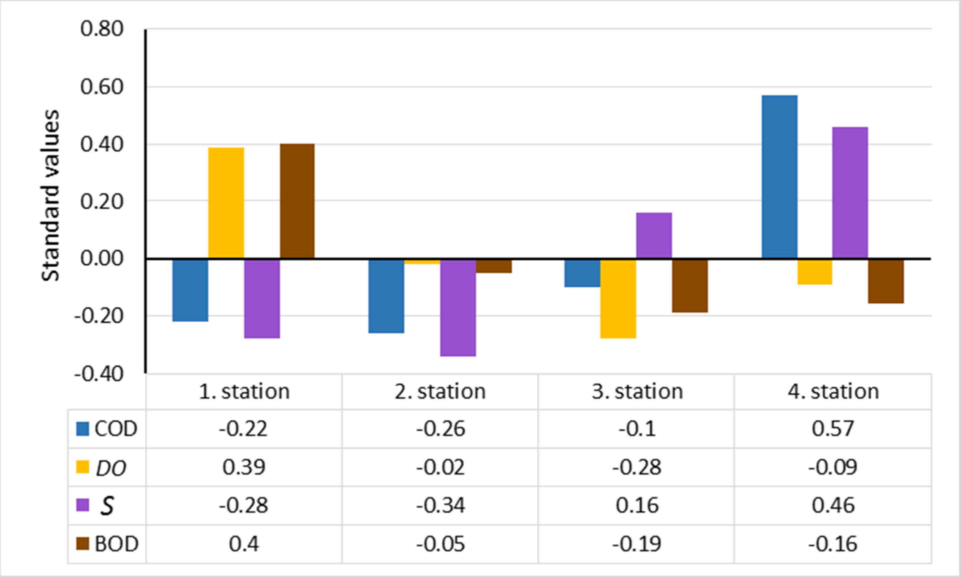
here X_i is the values of the quality variable, μ is the mean of the variable, and σ is the standard deviation of the variable.

The standardized values of the water quality variables which showed a significant difference according to the sampling stations are presented in Figure 2.

According to Figure 2, although the indicators of COD and *S* were low in sampling stations 1 and 2, they displayed a rapid increase in station 4. COD rises towards the downstream of the Riva River, which shows that the downstream of the river is polluted with regard to COD. There is a plant that produces detergent and chemical products before the sampling station 4. It is thought that the increase of the COD value in the downstream is related to this.

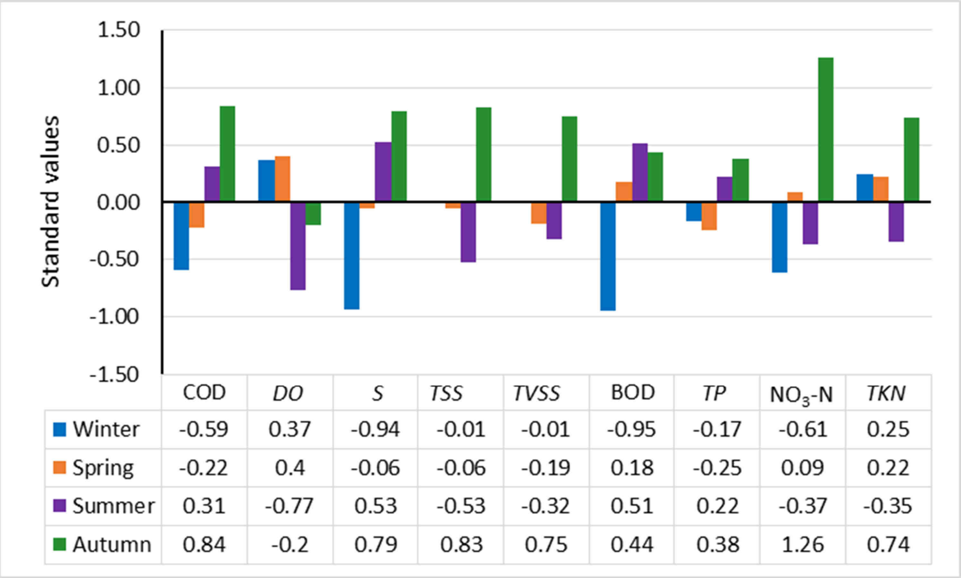
However, the indicators of BOD and *DO* were high in the station 1 while they showed a decreasing trend in the stations 3 and 4. BOD decreased at the downstream section of the river. The reason why BOD is high at the 1st sampling station is that the exit waters of

Pasakoy Advanced Biological Wastewater Treatment Plant are discharged into the river through this station. The same situation is also observed in similar studies [16]. This discharge caused a decrease in the value of *DO* and the value of *DO* dropped to the lowest level at the 3rd sampling station.



	1. station	2. station	3. station	4. station
🟦 COD	-0.22	-0.26	-0.1	0.57
🟨 *DO*	0.39	-0.02	-0.28	-0.09
🟪 *S*	-0.28	-0.34	0.16	0.46
🟫 BOD	0.4	-0.05	-0.19	-0.16

Fig. 2. Mean of standard values of water quality variables according to measurement stations



	COD	*DO*	*S*	*TSS*	*TVSS*	BOD	*TP*	NO_3-N	*TKN*
🟦 Winter	-0.59	0.37	-0.94	-0.01	-0.01	-0.95	-0.17	-0.61	0.25
🟧 Spring	-0.22	0.4	-0.06	-0.06	-0.19	0.18	-0.25	0.09	0.22
🟪 Summer	0.31	-0.77	0.53	-0.53	-0.32	0.51	0.22	-0.37	-0.35
🟩 Autumn	0.84	-0.2	0.79	0.83	0.75	0.44	0.38	1.26	0.74

Fig. 3. Mean of standardized values of water quality variables according to seasons

The change of standardized values of quality variables showing variations according to the seasons is given in Figure 3.

It can be seen in Figure 3 that all the water quality variables except for *DO* and *TKN* were low because of precipitation during winter and all the variables apart from *DO* and *TKN* took on high values in fall. During the summer months, the values of *DO* and *TSS* decreased and those of *S* and BOD increased [27].

The relationships among chemical variables

In this part, the relationships between chemical parameters were specified using correlation and regression analysis.

Correlations between chemical variables

The correlation coefficients obtained for the relationships among chemical variables are given in Table 4.

Table 4

Correlations (the correlations among the average monthly values of the parameters)

	pH	*T*	*DO*	*EC*	*S*	*TSS*	*TVSS*	BOD	COD	*TP*	NO_3-N	*TKN*	NH_4-N
pH	1												
T	.101	1											
DO	–.314	–.429	1										
EC	.044	**.751****	**–.532***	1									
S	–.227	.432	–.311	**.562***	1								
TSS	–.101	.036	–.244	.423	.119	1							
TVSS	–.104	.035	–.233	.394	.100	**.957****	1						
BOD	–.131	**.767****	**–.532***	**.759****	**.533***	.266	.206	1					
COD	.030	**.735****	**–.578***	**.781****	.270	**.477***	.433	**.800****	1				
TP	**.481***	.167	**–.472***	.248	.044	.111	.083	.223	.323	1			
NO_3-N	–.401	.180	–.244	**.545***	.186	**.471***	.398	.419	.547*	.092	1		
TKN	–.141	–.457	**.536***	**–.536***	–.296	–.050	.062	**–.528***	–.464	–.282	–.350	1	
NH_4-N	.215	–.046	–.002	.073	–.328	.150	.087	–.210	.177	.043	.362	–.240	1

*Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed)

The variables with significant relationships with each other according to the correlation matrix above can be summarized as follows. According to the correlation matrix, it is seen that the strongest relationship among the variables is naturally between *TSS* and *TVSS*. Apart from this, it can be seen in Table 4 that highly and positively related variables are COD-BOD, COD-*EC*, BOD-*T*, BOD-*EC*, *EC*-*T*, and COD-*T*, respectively [19, 28]. Though there is a positive relationship between the variables in general, *DO* is in a negative relationship with all the variables except *TKN*. A similar negative relationship is also observed between *TKN* and the other variables.

The prediction of relationship among the variables through Multiple Regression Model

Regression models are methods used to explain the relationship between independent variable and dependent variable(s). The direction and magnitude of the influence of each independent variable in the model can be determined with regression models. This part of the study tries to identify the relationships of significant quality variables which stand out in determination of water quality (dependent) with other variables (independent) through multiple regression models. For this, the variables of BOD, COD, NH_4-N and NO_3-N were selected as dependent variables. The other variables affecting each dependent variable were determined with the multiple regression model. Stepwise regression method was used to construct regression models. With this method, partial correlation coefficients are gradually added to the model starting with the highest and most significant variable. This process continues until there are no meaningful relationships. The results are given in Table 5.

Table 5

Multiple regression models for the variables of BOD, COD, NH_4-N and NO_3-N

Dependent variables		**Independent variables**									**Adjusted coefficient of determination** ***St. error***
		Con-stant	**BOD**	**COD**	***TKN***	**NH_4-N**	**pH**	***EC***	***T***	***TSS***	
BOD	b	7.929		0.157	–0.733	–0.641					0.791
	t	2.47*		5.9**	–2.2*	–3.6**					1.64
COD	b	11.3	2.43			1.906			0.88	0.292	0.8
	t	1.84	2.6*			2.18*			1.87	2.4*	7.61
NH_4-N	b	7.89	–0.743	0.123	–0.719						0.405
	t	2.2*	–3.6**	2.9*	–1.97						1.77
NO_3-N	b	16.27				0.168	–2.03	0.001			0.583
	t	3.6**				2.7*	3.2**	3.4**			0.573

* Coefficient is significant at the 0.05 level, ** Coefficient is significant at the 0.01 level

BOD is affected by the water quality variables of COD, *TKN* and NH_4-N and can be explained by these variables by 79.1 %. BOD is affected by COD in a positive direction [19] and by *TKN* and NH_4-N in a negative direction. COD is explained by BOD, NH_4-N, *T* and *TSS* by 80 % and the effect of all four variables is positive [29]. NH_4-N is explained by the variables of BOD, COD and *TKN* by 40.5 %. COD has a positive effect on NH_4-N while BOD and *TKN* affect it negatively. NO_3-N is explained by the variables of NH_4-N, pH and *EC* by 58.3 %. NO_3-N is affected by the variables of NH_4-N and *EC* positively and by pH negatively.

Prediction of water quality classes

The logistic regression and decision tree, known as classification methods, and neural network, an artificial intelligence method, were used in order to predict the water quality class of the Riva River.

Prediction of water quality classes with ANN

In this part of the study, predictions were made about the water quality variables of the Riva River with the *ANN* technique using the Clementine software. While monthly averages are used in regression analysis, all the measurement data were used in the *ANNs*. Emphasis was given to having a high amount of data so that the *ANN* is trained in a better way. The data were randomly divided into two, with approximately 65 % for training set and 35 % for test set. The water quality classes COD, BOD and NH_4-N were taken as the basis for the purpose of prediction. Since water quality classes did not exhibit much variation for the other parameters, they were not deemed worth making prediction.

As the input variable of the *ANNs*, the measurement values of chemical pollution variables were selected, and as the output variable, the water quality class of the relevant variables was selected.

Prediction of water quality class for COD with ANNs

The values that belong to chemical variables measured apart from COD (pH, *T*, *DO*, *EC*, *S*, *TSS*, *TVSS*, BOD, *TP*, NO_3-N, *TKN*, and NH_4-N) were used as the input variables for the prediction of the COD, and the water quality classes of the COD were used as the output variables. In the Artificial Neural Network model that was constructed, numerous hidden layers and neurons were used in order to make the best prediction for each different training method [16, 30, 31], as seen in Table 6.

Models are built in line with the network architecture above and predictions for the training and test sets for COD are given in Table 7.

Table 6

Neural Network Methods used in the study and the network architecture

Layers	**Training Methods Neuron Numbers**					
	Quick	**Dynamic**	**Multiple**	**Prune**	***RBFN***	**Exh. Prune**
Input neurons	12	12	12	12	12	12
Hidden layer 1	3	8	7	12	20	30
Hidden layer 2	-	6	5	-	-	20
Output neuron	1	1	1	1	1	1

Table 7

Predictions of water quality class with Neural Network Models

COD	Training Set					
	Quick	Dynamic	Multiple	Prune	*RBFN*	Exh. Prune
Correct %	93.75	93.75	100	**97.5**	70	95
	Test Set					
Correct %	46.15	55.77	59.62	**75**	51.92	59.62
BOD	Training Set					
	Quick	Dynamic	Multiple	Prune	*RBFN*	Exh. Prune
Correct %	**96.25**	93.75	98.75	97.5	75	98.75
	Test Set					
Correct %	**69.23**	67.31	67.31	67.31	67.31	65.38
NH_4-N	Training Set					
	Quick	Dynamic	Multiple	Prune	*RBFN*	Exh. Prune
Correct %	92.25	92.25	100	97.5	76.25	**96.25**
	Test Set					
Correct %	57.69	55.77	55.77	59.62	48.07	**65.4**

According to Table 7 it is seen that the most suitable model for identifying the water quality class for COD is the Multiple Model, when the training set is taken into consideration, and the Prune Model, when the test set is taken into consideration. Here, taking into consideration the accuracy rate obtained for the test set, it is possible to say that predictions to be made with the Prune model will achieve an accuracy rate of 75 %.

Prediction of water quality class for BOD with ANNs

Models similar to the ones built for the prediction of water quality class for COD were also constructed for BOD and predictions were made for water quality class of the training and test sets (Table 7).

According to the results of the test set, it is understood that the most suitable artificial neural network model for identifying the water quality class for COD is the Quick Model.

It can be said that a prediction that will be made with this model will have an accuracy rate of 69.23 %.

Prediction of Water Quality Class for NH_4-N with the ANNs

A similar model was built for the prediction of water quality class of the NH_4-N. The most suitable model for the prediction of the NH_4-N is the Exhausted Prune model with a 65.4 % accuracy rate in the predictions made for the test set (Table 7) [31].

Prediction of water quality classes with decision trees and logistic regression models

Decision tree models allow develop classification systems that predict or classify future observations based on a set of decision rules. In this study, predictions were made with four different decision tree models (Classification and Regression Tree (*CRT*) node, Chi-square Automatic Interaction Detector (*CHAID*) node, Quick Unbiased Efficient Statistical Tree (*QUEST*) node, C5.0 node) [32].

Logistic regression analysis is a regression method which assists classification and assignment operations. It is more advantageous than the other classification models because of its assumption of normal distribution and since there is no prerequisite for assumption of persistence.

Logistic regression models are categorized in three different ways according to the status of the dependent variable, which are binary, ordinal and nominal. In this study, the ordinal logistic regression model was used as there is more than two water quality class.

Prediction of water quality class for COD with decision tree and logistic regression model

In this part, the quality class was predicted with the decision tree models and logistic regression models that were established. The models were developed with the training set and predictions were made with the training set. The results related to the accuracy of the predictions are presented in Table 8 [33].

Table 8

Predictions of water quality class with decision tree and logistic regression models

COD	Decision Tree Models (Training set)				Logistic Regression
	C5	*CRT*	*QUEST*	*CHAID*	(Training set)
Correct %	88.75	**95**	67.5	76.25	75
COD	Decision Tree Models (Test set)				Logistic Regression
	C5	*CRT*	*QUEST*	*CHAID*	(Test set)
Correct %	50	**53.85**	51.92	51.92	48.8
BOD	Decision Tree Models (Training set)				Logistic Regression
	C5	*CRT*	*QUEST*	*CHAID*	(Training set)
Correct %	83.75	**97.5**	75	86.25	**82.5**
BOD	Decision Tree Models (Test set)				Logistic Regression
	C5	*CRT*	*QUEST*	*CHAID*	(Test set)
Correct %	**73.08**	59.61	73.08	67.3	**69.23**
NH_4-N	Decision Tree Models (Training set)				Logistic Regression
	C5	*CRT*	*QUEST*	*CHAID*	(Training set)
Correct	**93.75**	**93.75**	66.7	88.75	**67.5**
NH_4-N	Decision Tree Models (Test set)				Logistic Regression
	C5	*CRT*	*QUEST*	*CHAID*	(Test set)
Correct	**65.4**	50	50	57.7	**51.9**

CRT: Classification and Regression Trees, *QUEST*: Quick, Unbiased, Efficient, Statistical Tree, *CHAID*: Chi square Automatic Interaction Detection

It is understood that the most suitable decision tree model for the data of the training set is the *CRT* model. Table 8 shows that the accuracy rate of the predictions made with this model is 95 %. Again, the *CRT* model provides the best predictions for the test set with an accuracy rate of 53.85 %. The *CRT* Decision Tree model was given in Figure 4 as an example for the prediction of the quality class of COD.

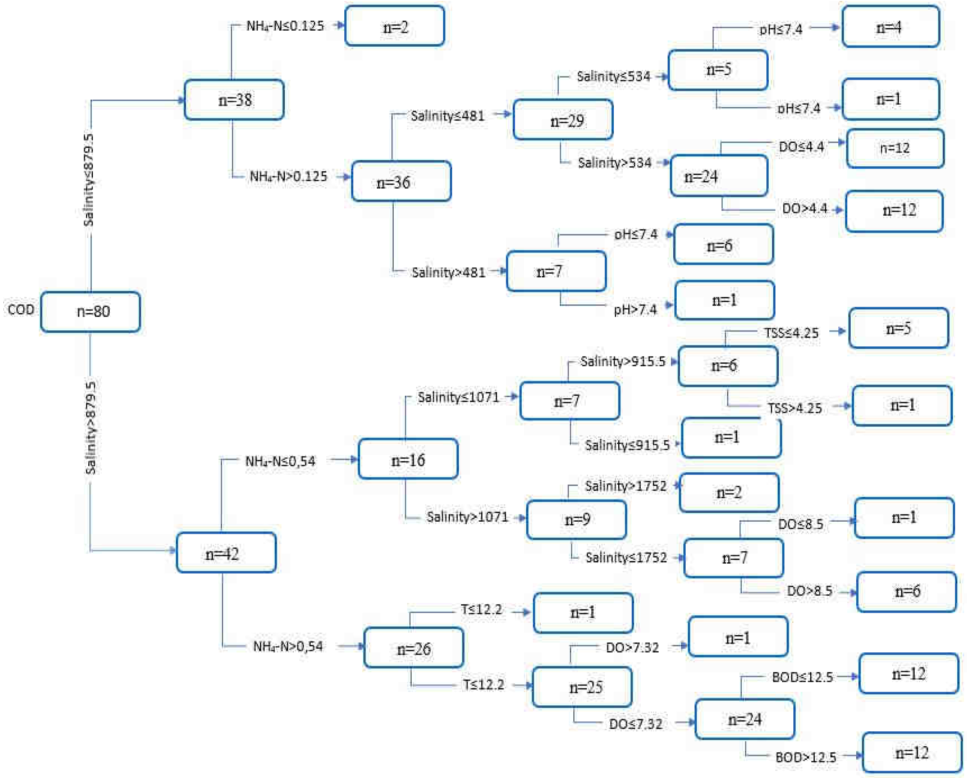


Fig. 4. *CRT* decision tree model for COD

A success rate of 75 % was achieved for the training set is in the predictions made with the logistic regression model, while the success rate was approximately 48 % for the test set.

Prediction of water quality class for BOD with Decision Tree and Logistic Regression Model

Among the decision tree models, the *CRT* model provided the best predictions for the training set with an accuracy rate of 97.5 %. However, it did not achieve this rate in the test sets. For the test sets, the C5 and *QUEST* models produced the best predictions with an accuracy rate of 73.08 %. Moreover, while the predictions of the logistic regression model display an accuracy rate of 82.5 % for the training set, this rate dropped to 69.23 % in the test set (Table 8).

Prediction of Water Quality Class for NH_4-N with the decision tree and logistic regression model

The C5 and *CRT* models provided the best predictions in the test set data for the prediction of water quality class of the NH_4-N with an accuracy rate of 93.75 %. The best predictions for the test set was produced by the C5 model with an accuracy rate of 65.4 %. While an accuracy rate of 67.5 % was achieved in the training set of the logistic regression model, this rate decreases to 51.9 % in the test set (Table 8) [16].

Conclusions

This study investigated the spatial and temporal variability of the existing water pollution conditions of the Riva River. The values of BOD and *S* were high in the upstream areas and COD was higher in the downstream areas of the river. Other studies in the literature point out that river pollution varies spatially and pollution level increases in the downstream basin [18]. The values of *DO* and *TKN* was high in winter and those of BOD and *S* was high in summer. *DO* values were reported to be lowest in August and highest in January [2, 28]. Therefore, it is necessary to monitor the pollution level of the river water in order to protect the water quality of the Riva Basin.

It was observed that COD, *S*, *DO* and BOD taken from the sampling stations on the Riva River showed variations according to the sampling stations. All of the water quality variables showed variations according to seasons. Thus, in this study it can also be recommended to study temporal variations as well as spatial variations in water quality in order to assess the water quality of the Riva River Basin. This situation is similar to other water quality studies [21-28]. In the studies evaluating the spatial-temporal changes of the water quality parameters in the river basin, it was found that the predictions in the spatial analyzes as well as the predictions in the temporal analyzes were correct [19].

According to the multiple regression model, there is a high relationship between BOD and COD. In general, studies in the literature confirm this [16]. An increase in the value of one rises the other one's value as well. In the same manner, rises in NH_4-N, *TSS* and *T* increase the value of COD. On the other hand, increases in *TKN* and NH_4-N reduce BOD. As the values of NH_4-N and *EC* increase, the value of NO_3-N rises; however, this value drops as pH increases.

As a result, it was understood that the *ANN* was suitable for the prediction of COD, the Decision Tree was suitable for the prediction of BOD and both models were suitable for the prediction of NH_4-N. Therefore, the developed models can be used to monitor and predict the water quality of the Riva River with reasonable accuracy. According to the literature, NH_4-N, COD, NO_3-N, *DO* and Turbidity are the most effective water quality parameters in assessing water quality [1, 10, 27, 34].

Acknowledgements

This research work was supported and granted by the Graduate School of Scientific Research, Project Commission (Project No: 2011-50-01-044) of Sakarya University.

References

[1] Singh KP, Malik A, Sinha S. Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques - a case study. Analytica Chim Acta. 2015;538(1-2):355-74. DOI: 10.1016/j.aca.2005.02.006.

[2] Mishra BK, Regmi RK, Masago Y, Fukushi K, Kumar P, Saraswat C. Assessment of Bagmati river pollution in Kathmandu valley: Scenario-based modeling and analysis for sustainable urban development. Sustain Water Qual Ecol. 2017;9-10:67-77. DOI: 10.1016/j.swaqe.2017.06.001.

[3] El-Tohamy WS, Abdel-Baki SN, Abdel-Aziz NE, Khidr AAA. Evaluation of spatial and temporal variations of surface water quality in the Nile River. Ecol Chem Eng S. 2018;25:569-80. DOI: 10.1515/eces-2018-0038.

[4] Taylor SD, He Y, Hiscock KM. Modelling the impacts of agricultural management practices on river water quality in Eastern England. J Environ Manage. 2016;180:147-63. DOI: 10.1016/j.jenvman.2016.05.002.

[5] Pirani M, Panton A, Purdie DA, Sahu SK. Modelling macronutrient dynamics in the Hampshire Avon River: A Bayasian approach to estimate seasonal variability and total flux. Sci Total Environ. 2016;572:1449-60. DOI: 10.1016/j.scitotenv.2016.04.129.

[6] Melching CS, Liang J, Fleere L, Wethington D. Modeling the water quality impacts of the separation of the Great Lakes and Mississippi River basins for invasive species control. J Great Lakes Res. 2015;41:87-98. DOI: 10.1016/j.jglr.2014.11.009.

[7] Mainali J, Chang H. Landscape and anthropogenic factors affecting spatial patterns of water quality trends in a large river basin, South Korea. J Hydrol. 2018;564:26-40. DOI: 10.1016/j.jhydrol.2018.06.074.

[8] Swain R, Sahoo B. Improving river water quality monitoring using satellite data products and a genetic algorithm processing approach. Sustain Water Qual Ecol. 2017;9-10:88-114. DOI: 10.1016/j.swaqe.2017.09.001.

[9] Shia Y, Xu G, Wang Y, Engel BA, Peng H, Zhang W, et al. Modelling hydrology and water quality processes in the Pengxi River basin of the Three Gorges Reservoir using the soil and water assessment tool. Agr Water Manage. 2017;82:24-38. DOI: 10.1016/j.agwat.2016.12.007.

[10] Decelis LDR, Igúzquiza EP, Andreo B. Spatial prediction of water quality variables along a main river channel, in presence of pollution hotspots. Sci Total Environ. 2017;605-606:276-90. DOI: 10.1016/j.scitotenv.2017.06.145.

[11] Vrebos D, Beauchard O, Meire P. The impact of land use and spatial mediated processes on the water quality in a river system. Sci Total Environ. 2017;601-602:365-73. DOI: 10.1016/j.scitotenv.2017.05.217.

[12] Jannin LB, Brito D, Sun X, Jauch E, Neves R, Sauvage S, et al. Spatially distributed modelling of surface water - groundwater exchanges during overbank flood events - a case study at the Garonne River. Adv Water Resour. 2016;94:146-59. DOI: 10.1016/j.advwatres.2016.05.008.

[13] Gorzel M, Kornijow R, Buczynska E. Quality of rivers: Comparison of hydro-morphological, physical-chemical and biological methods. Ecol Chem Eng S. 2018;25:101-22. DOI: 10.1515/eces-2018-0007.

[14] Wang Y, Zhang W, Zhao Y, Peng H, Shi Y. Modelling water quality and quantity with the influence of inter-basin water diversion projects and cascade reservoirs in the Middle-lower Hanjiang River. J Hydrol. 2016;541:1348-62. DOI:10.1016/j.jhydrol.2016.08.039.

[15] Whitehead PG, Bussi G, Bowes MJ, Read DS, Hutchins MG, Elliott JA, et al. Dynamic modelling of multiple phytoplankton groups in river with an application to the Thames river system in the UK. Environ Model Softw. 2015;74: 75-91. DOI: 10.1016/j.envsoft.2015.09.010.

[16] Ogleni N, Topal B. Water quality assessment of the Mudurnu River, Turkey, using biotic indices. Water Resour Manage. 2011;25:2487-508. DOI: 10.1007/s11269-011-9822-1.

[17] Ustaoğlu F, Tepe Y. Water quality and sediment contamination assessment of Pazarsuyu Stream, Turkey using multivariate statistical methods and pollution indicators. J Soil Water Conserv. 2019;7:47-56. DOI: 10.1016/j.iswcr.2018.09.001.

[18] Papazova P, Simeonova P. Long-term statistical assessment of the river quality of Tundja River. Ecol Chem Eng S. 2012;19:213-26. DOI: 10.2478/v10216-011-0016-9.

[19] Avila R, Horn B, Moriarty E, Hodson R, Moltchanova E. Evaluating statistical model performance in water quality prediction. J Environ Manage. 2018;206:910-9. DOI: 10.1016/j.jenvman.2017.11.049.

[20] Libera DA, Sankarasubramanian A. Multivariate bias corrections of mechanistic water quality model predictions. J Hydrol. 2018;564:529-41. DOI: 10.1016/j.jhydrol.2018.07.043.

[21] Tomas D, Čurlin M, Marić AS. Assessing the surface water status in Pannonian ecoregion by the water quality index model. Ecol Indic. 2017;79:182-90. DOI: 10.1016/j.ecolind.2017.04.033.

[22] Mahmoodabadi M, Arshad RR. Long-term evaluation of water quality parameters of the Karoun River using a regression approach and the adaptive neuro-fuzzy inference system. Mar Pollut Bull. 2018;126:372-80. DOI: 10.1016/j.marpolbul.2017.11.051.

[23] Kim SE, Seo IW. Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers. J Hydro-Environ Res. 2015;9:3325-39. DOI: 10.1016/j.jher.2014.09.006.

[24] Turkish Standards Institution (TSE), (1989). Water Quality - Total Solids Determination. TS 7093, 4. Ankara. https://en.tse.org.tr.

[25] APWA, AWWA, WEF, Standard Methods for the Examination of Water and Wastewater. 21th Ed. Washington, DC: American Public Health Association; 2005. ISBN: 0875530478.

[26] Arora S, Keshari AK. Estimation of re-aeration coefficient using MLR for modelling water quality of rivers in urban environment. Groundwater Sustainable Dev. 2018;7:430-5. DOI: 10.1016/j.gsd.2017.11.006.

[27] Wu Z, Wang X, Chen Y, Cai Y, Deng J. Assessing river water quality using water quality index in Lake Taihu Basin, China. Sci Total Environ. 2018;612:914-22. DOI: 10.1016/j.scitotenv.2017.08.293.

[28] Ewaid SH, Abed SA, Kadhum SA. Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis. Environ Technol Innov. 2018;11:390-8. DOI: 10.1016/j.eti.2018.06.013.

[29] Cui F, Park C, Kim M. Application of curve-fitting techniques to develop numerical calibration procedures for a river water quality model. J Environ Manage. 2019;1:109375. DOI:10.1016/j.jenvman.2019.109375.

[30] Kim HG, Hong S, Jeong KS, Kim DK, Joo GJ. Determination of sensitive variable regardless of hydrological alteration in artificial neural network model of chlorophyll a: Case study of Nakdong River. Ecol Modell. 2019;398:67-76. DOI: 10.1016/j.ecolmodel.2019.02.003.

[31] Singh KP, Basant A, Malik A, Jain G. Artificial neural network modeling of the river water quality - A case study. Ecol Modell. 2009;6:888-95. DOI:10.1016/j.ecolmodel.2009.01.004.

[32] Rokach L, Maimon O. Data Mining With Decision Trees Theory and Applications. Danvers USA: World Scientific Publishing Co. Pte. Ltd; 2008. ISBN: 9789812771711.

[33] Everaert G, Bennetsen E, Goethals PLM. An applicability index for reliable and applicable decision trees in water quality modelling. Ecol Inf. 2016;32:1-6. DOI: 10.1016/j.ecoinf.2015.12.004.

[34] Bostanmaneshrad F, Partani S, Noori R, Nachtnebel HP, Berndtsson R, Adamowski JF. Relationship between water quality and macro-scale parameters (landuse, erosion, geology, and population density) in the Siminehrood River Basin. Sci Total Environ. 2018;639:1588-600. DOI: 10.1016/j.scitotenv.2018.05.244.