Global NEST Journal, Vol 27, No 3, 06086

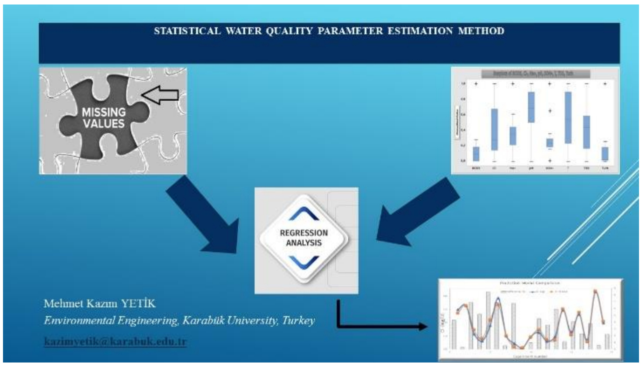
A statistical approach to estimating water quality parameter: a case study on Turkey's rivers

Mehmet Kazım Yetik[1]

[1]Environmental Engineering, Karabük University, Turkey
Received: 21/04/2024, Accepted: 28/12/2024, Available online: 15/01/2025, https://orcid.org/0000-0003-1150-3968

*to whom all correspondence should be addressed: e-mail: kazimyetik@karabuk.edu.tr, kazim.yetik@gmail.com
https://doi.org/10.30955/gnj.006086

Graphical abstract



Abstract

Monitoring water quality parameters in rivers is of critical importance for decision makers and researchers. However, difficulties such as missing data, test values exceeding certain limits and parameters that are not measured but need to be monitored are among the main problems in monitoring water quality. In this study, these three main problems in water quality monitoring processes were addressed and solution-oriented approaches were developed. Solutions were applied for missing and dirty data problems, and multivariate regression and prediction models were proposed for unmeasured parameters. A prediction model was developed using parameters such as BOD5, Cl^-, Na^+, pH, SO_4, temperature, TDS and turbidity, and the performance of the model was analyzed in detail for SO_4, Cl^-, Na^+ and BOD5 parameters. The accuracy of the model was evaluated with statistical indicators such as MSE, MAPE and correlation coefficient (r). The model showed high accuracy with MAPE values below 10% for most parameters. For example, for the Cl^- parameter, MSE was calculated as 118, MAPE as 3.6 and r value as 0.984. In addition, a user-friendly graphical user interface (GUI) has been developed and used to create an automatic regression model for the Cl^- parameter. This system, integrated with VBScripts, has been tested on the Kızılırmak River to prove its applicability within a single program framework for all rivers. The results obtained show the effectiveness and wide-ranging applicability of the proposed method.

Keywords: Water Quality Parameters, Missing data, Boxplot, Multivariate Analysis, user friendly GUI, Outlier Detection.

1. Introduction

Water is one of the most significant elements on our planet. No living species on earth can survive without it. It is a substance that is used by all living organisms. Industrialization and urbanization are rapidly developing near water resources, causing numerous contaminants from industry to impact water resources. These waste products from industrialization and urbanization change the natural concentration and characteristics of the water, increasing the level of water pollution in the region. Furthermore, river water affects every point of its route (Isaac & Siddiqui, 2022). Managing water resources is critical to basin management. In recent years, there has been an increasing interest in monitoring water quality indicators. In certain nations, collecting water samples manually and then analyzing them in a laboratory is mandatory (Islam Khan *et al.* 2022). The water quality index (WOI) is widely used by policymakers, decision-makers, and other stakeholders to obtain a clear and comprehensive picture of a water body's pollution status. It is one of the most commonly used concepts to express water quality in this area (Tripathi & Singal, 2019; Sadiq *et al.* 2022).

Monitoring water quality parameters in streams is very important to have information about the river, researchers and decision makers organize measures in the riverbed according to these parameters on the river. The most important problem that researchers and decision makers encounter at this stage is that incorrect data from analyses or online sensors cannot be determined immediately, or some information is missing due to sensor and analysis deficiencies, or a parameter that will be needed is not available in the system. In this study, solutions were applied based on a data set with these two problems and presented to researchers. In the third problem, a multiple regression model was created for some parameters selected from the data set to show that an existing parameter can be estimated correctly. In this stage, an interface was created with (Visual Basic Script) VBScript to make the processes fast and error-free, allowing researchers and decision makers to perform easy operations.

This article explores how dirty data can be removed from the dataset and how missing data can be retroactively added, as well as how a complete dataset can be created without outliers by applying relevant operations. Additionally, the effects of water quality parameters on each other were analyzed, and a model for estimating water quality parameters was presented. For this purpose, the water quality parameters of the data set of the Kızılırmak River for 5 years were used.

The literature review revealed that data sets were used in all analyses. However, measurement errors and data deficiencies are common in water quality test data due to seasonal and regional conditions. These shortcomings can directly influence the research results. In this study, methods to address these shortcomings were discussed and the proposed approach was applied to a data set. In addition, the process of estimating some parameters that are difficult to determine using the modelling method was explained in detail. The developed model was supported by statistical analyses and its accuracy and effectiveness were demonstrated. It is assessed that the methods presented in the study can provide important guide in the field of river research. In addition to proving the accuracy of the hypothesis supported by statistical analyses, the methods used contribute to the evaluation of the data obtained and to the planning of research processes to be carried out in different rivers. In this respect, the study has a quality that can serve as a guide for scientists and practitioners in the field of river research. The Kızılırmak river, which is particularly significant for Turkey and the countries bordering the Black Sea, has always been a focus of researchers, and scores of studies on its water have been undertaken. Because estimating or directly monitoring the water characteristics of the Kızılırmak river is of critical importance for the region. In this study, 8 different water quality parameters taken from the station on the Kızılırmak River were used, and the created prediction model allows the researcher to estimate a selected parameter. The model, which allows parameter change, was tested one by one with statistical methods for four different parameters, and the results were interpreted and made available to researchers. A modeling study was also carried out with normalized data, in order to compare the factor values of the parameters with each other (Lumb *et al.* 2011; Zhou *et al.* 2021). These effects of the parameters on each other will be especially useful for researchers doing Artificial Neural network (ANN) and artificial intelligence (AI) studies (Wenyan *et al.* 2023). For this purpose a user-friendly Graphical User Interface (GUI) was created by software coding in VBScripts, and this interface was used to create a model for a parameter chlorine ion (Cl).

1.1. Literature

Water plays a significant role in our lives. Water quality change or degradation endangers the aquatic environment, affects human health, and has an impact on the region's social and economic growth. It is necessary to gather data on water quality in order to avoid water contamination, especially in developing nations (Sun *et al.* 2019). Aquatic organisms, waterside soil erosion and seawater entering parameters are all monitored for water quality (Setshedi *et al.* 2021; Ma *et al.* 2020).

Human activities (agriculture, industry, and urbanization) and natural elements (soil, geology, and precipitation) in the basin have an impact on the quality of surface water in this region (Sundaray *et al.* 2006; Noori *et al.* 2009). Precipitation, hydrological conditions, and seasonal fluctuations on the streamline all contribute significantly to pollution (Z. M. Zhang *et al.* 2022). Assessing the quality of its water is therefore critical because it directly impacts public health (via drinking water) and aquatic life (via raw water) (Pande G. *et al.* 2015).

Monitoring is seen that water quality is decreasing in modern societies due to increasing water scarcity. It is not enough for communities to access and receive the required amount of water. In addition, the quality of the water reached must be appropriate (Horvat *et al.* 2021).

As a result of cultures' excessive usage of water, the problem of water body reduction and pollution has occurred. For this reason, due to environmental factors, completing this follow-up is a very costly and time-consuming operation in some circumstances, and it requires a large and expert team, and it is critical to have hardware that can provide full and accurate data. As a result, it is critical for researchers to evaluate water quality parameters or to improve the effectiveness of the current evaluation procedure. In order to develop a better management plan (Abyaneh, 2014), researchers are continuing their research.

According to Saikat Islam Khan and colleagues (2022), assessing water quality is one of the biggest challenges the world has faced in recent years, and their study aims to examine the research methods that researchers have used in the last few years for water quality management systems. Multiple linear regression, least squares method, decision tree, random forest method, wavelet neural network approach, and recursive wavelet network have been examined. The climate has been significantly impacted by these recently developed methods for monitoring water. Nevertheless, these models necessitate a considerable number of input parameters (Islam Khan *et al.* 2022; Uncumusaoğlu A. A., 2018).

Monitoring of water quality is essential for anticipating future hazards to the aquatic environment and for managing water resources. In the long-term research of accurate monitoring and evaluation of water quality, there are many obstacles. To reduce these challenges, several researchers used Multivariate Statistical Techniques (MST) and the Water Quality Identification Index (WQII) in their studies to examine the change in water quality in a river and identify key sources of pollution (Ma *et al.* 2020; Li *et al.* 2018).

Researchers have developed prediction models by using different numbers of parameters and incorporating different parameters into the model (Agegnehu *et al.* 2024; Li *et al.* 2018; Islam Khan *et al.* 2022; Ma *et al.* 2020). Some scientists, on the other hand, investigated strategies for developing a multivariate statistical model prediction model with fewer parameters for river estimate (Isaac & Siddiqui, 2022). In the index selection procedure, some scholars investigated the scientific basis of the multivariate statistical analysis approach (Liu *et al.* 2021; Gurjar & Tare, 2019; Anifowose & Odubela, 2018). Multivariate statistical methods have been used by many researchers to monitor and predict water quality pollution in their research. (Z. M. Zhang *et al.* 2022; Yidana & Yidana, 2010; Abyaneh, 2014; Noori *et al.* 2010; Kazi *et al.* 2009; Azhar *et al.* 2015; Fathi, Zamani-Ahmadmahmoodi, and Zare-Bidaki 2018; Sakizadeh, 2016).

On-site study of water quality parameters can be challenging due to geographical restrictions; also, inaccuracies in data may arise during on-site measurement due to meteorological factors such as air temperature, rain, etc. Because various factors might influence measurement values, different approaches are employed to discover and delete incorrect data from the data set. In outlier investigations, Tukey's (1977) boxplot method, which is described in detail in the book, is employed. The boxplot also produces excellent results in data extraction studies in many different fields. For example, the boxplot analysis method was used to evaluate data from surface water samples and discover outliers in water quality investigations (Ahmad *et al.* 2001).

Table 1 Table showing the statistical analysis values for the raw data set on which the study will be conducted.

Parameter	Unit	Mean	Max	Min	Variance	CV
BOD5	mg/L	2.928571	12	1	7.066327	90.76973
Cl	mg/L	247.5316	377.52	166.76	3711.257	24.61104
Na	mg/L	185.0211	307.64	117.82	1697.83	22.27029
pH	-	7.841176	8.5	6.6	0.298893	6.972306
SO4	mg/L	360.8211	1239	16.2	63964.29	70.09337
T	°C	15.625	21	9	16.73438	26.18091
TDS	mg/L	1141.533	1524	875	26937.58	14.37774
Turb	NTU	50.60667	418	0.4	10695.18	204.3555

Table 2 Outlier values and boxplot analysis values obtained as a result of Boxplot analysis for outlier value detection.

Parameter	Q1	Q3	Median	Bottom	Up
BOD5	1.75	3	2	1	4
Na	157.8	201.8	184.6	117.8	233.5
SO4	239.4	368.5	295.7	215.7	446.1
Turb	3.2	75.5	8.4	0.4	106

Missing data causes issues in the statistical analysis of water quality data obtained through machine learning. Some have investigated the success of various methods for dealing with defects in databases (Betrie *et al.* 2016; Y. Zhang & Thorburn, 2022). It is commonly used for the elimination of missing data in the creation of data sets, particularly in machine learning studies (Yang, 2022). Using an existing data set, researchers created estimation models using multivariate statistical approaches (Wang *et al.* 2012).

2. Material and procedure

2.1. Study area (case study)

This research was carried out with the data set of water quality parameters obtained from samples taken from the observation station on the Kızılırmak River.

The coordinates of the Station point, which is located on the (nearest point of the) coast in the Black Sea region and on the Kızılırmak, are $41^{0}35'02.84''$ North $35^{0}53'30.16''$ East, as seen in **Figure 1**. This station is one of the last stations on Kızılırmak due to its location. The location of the river and the station is important in revealing the parameters of pollutants discharged into the sea.

The longest river that starts in Turkey and discharges into the sea in Turkey is the Kızılırmak, which is 1151 kilometers long. Its basin, including its tributaries, covers an area of approximately 82 thousand square kilometers. The river runs through 45 different districts in 11 different provinces, with a total population of more than 11 million people living in the provinces it flows through. In addition to being a significant river for Turkey, it is also an important river for countries along the Black Sea coast due to its flow.

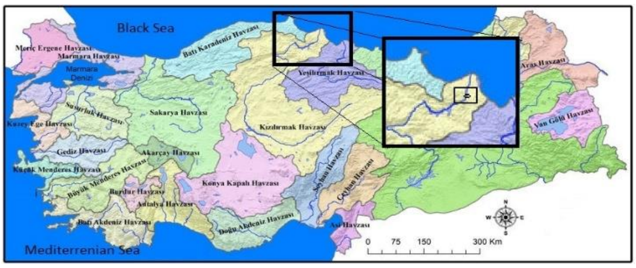


Figure 1. The station point's location (position within the basin) where the samples were taken to compile the data on the map of Turkey.

2.2. Data collection

Some water quality-determining parameters are analyzed on-site by experts from the General Directorate of State Hydraulic Works (is called as DSI in Turkey) using river samples, while others are measured in authorized laboratories. Analyzes of more than 50 parameters are conducted four times a year. This study was carried out using the findings of an analysis of data collected for 5 years. Biological oxygen demand (BOD5), chlorine ion (Cl), sodium ion (Na), pH, sulfate ion (SO4), Temperature (T), Total Dissolved Solids (TDS), and Turbidity (Turb) are the parameters employed in this research. The parameters, unit, Mean, Max, Min, Variance, and Coefficient of Variance (CV) are displayed in Table 1. The value of turbidity has the largest variance and CV in this table. The value variance and CV of pH are the lowest.

2.3. Method methodology

To develop a multivariate regression model based on the selected water quality parameters, the interactions and mutual influences among these parameters were analyzed. This preliminary analysis is critical for understanding the relationships between variables and identifying significant predictors. The observed effects of these interactions are presented in **Figure 2**, which provides a detailed visual representation of the correlations and dependencies within the dataset.

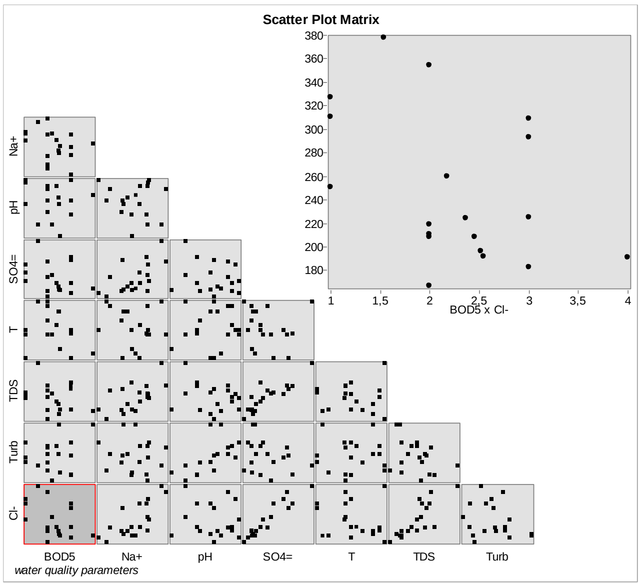


Figure 2. Scatter plot Matrix of selected water quality parameters

Some Researchers used variance and CV values while examining water quality parameters in their studies (Azhar *et al.* 2015; Abyaneh, 2014). As shown in Table 1, several maximum and minimum values are quite far from the mean (i.e. BOD5 and Turb values), and the variance value for this parameter is also very high. In such cases, Tukey (1977) introduced outlier detection methods with boxplot graphics in his study. Similarly, other researchers employed boxplot visuals to eliminate data noise (Islam Khan *et al.* 2022; Dawson, 2011). The boxplot approach was employed in several of these investigations to determine the extreme values of water quality parameters. The correction was accomplished by determining the extreme values (Tripathi & Singal, 2019; Horvat *et al.* 2021). The methodologies used are examined one by one below.

2.3.1. Boxplots

In the boxplot, a box shape is formed by drawing lines from the first quartile (Q1) to the third quartile (Q3); in this box shape, where the median of the data set is also marked, whiskers extending from the first quarter to the minimum and from the third quarter to the maximum are drawn, and a boxplot is formed. Outlier detection analysis was done on the data set in this study, and boxplots were utilized to extract data noise. **Figure 3** shows the creation of boxplots and the determination of outliers. Whiskers are typically much longer than the box; small whiskers may have a uniform distribution with sharp cuts (Dawson, 2011). Outliers are described as being far from whiskers.

Outlier detection analysis were done on the data set in this study, and boxplots were utilized to extract data noise. **Figure 3** shows the creation of boxplots and identification of outliers (with normalized values). This figure has been normalized so that the data can be aggregated.

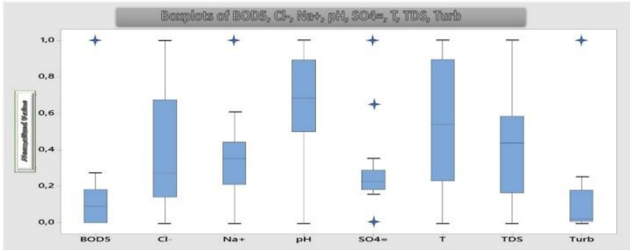


Figure 3. For outlier detection , BOD5, Cl$^-$, Na$^+$, pH, Turb, SO4, T, TDS, Turb parameters' boxplots graphics (with normalized value).

As seen in boxplot graphs, some graphs are also seen with '*' except for whiskers extending from the outside of the box to the minimum the first quarter (Q1) and from the third quarter (Q3) to the maximum (whiskers). As shown in the figure, the '*' value is far from the data values where the box is produced. BOD5, Na, SO4, and Turb values, in particular, appear to have contradictory values when compared to the data set structure. Table 2 shows the boxplot values for these (selected) parameters.

2.3.2. Missing Value

After outliers are detected by boxplot analysis, these parameters were examined by comparing them with other station data on the same river. It was observed that these extreme values were incompatible even with the seasonal changes of the river, and therefore these extreme values were removed from the data set. Missing value studies were carried out to replace the deleted values in the data set. However, since all data for a particular month was missing in the data set, no work was done to find a value to replace the data for that month, and the data for that month was completely removed from the analysis."

The data set was checked for missing values after it had been cleansed of extreme values by using outlier detection. The measurements were conducted with samples obtained from the same point of the river. The missing value is obtained sometimes owing to meteorological causes and sometimes from the sample (storage, transportation, etc.). Due to the issues that have arisen, it is not possible to examine and assemble data. However, the data in the data set must be complete in order to analyze and construct a new model. Researchers have developed new ways to achieve this goal by examining different methodologies for various data sets and methods for eradicating missing data in data sets (Betrie *et al.* 2016; Y. Zhang & Thorburn, 2022). As a method of replacing missing data in the data set, several studies calculate the median or mean values of the relevant data set. In reality, some researchers claim that it is acceptable to use arbitrary values in some data sets (Yang, 2022). Among these studies, the linear regression method, which is developed by the least squares method, produces very effective results for missing data elimination (Wang *et al.* 2012). The correlation between parameters is examined by this method, and a linear estimating model is formed by choosing the parameter with the highest correlation for missing data. Table 3 displays the correlation values of the parameters in the data set. The estimated value is calculated using the values in the same order as the missing data.

Table 3 Correlation values between parameters in the raw data set.

Parameters	BOD5	Cl	Na	pH	SO4	T	TDS	Turb
BOD5	1	-0.359	-0.286	-0.109	-0.118	-0.265	0.021	0.333
Cl	-0.359	1	0.669	-0.070	0.943	-0.173	0.813	-0.288
Na	-0.286	0.669	1	0.031	0.529	-0.274	0.450	-0.084
pH	-0.109	-0.070	0.031	1	-0.153	0.119	-0.026	0.509
SO4	-0.118	0.943	0.529	-0.153	1	-0.048	0.842	-0.141
T	-0.265	-0.173	-0.274	0.119	-0.048	1	0.107	-0.049
TDS	0.021	0.813	0.450	-0.026	0.842	0.107	1	-0.409
Turb	0.333	-0.288	-0.084	0.509	-0.141	-0.049	-0.409	1

Table 4 The statistical table of the final version of the data set, purified from outliers and imputed values for missing parameters.

Parameters	Unit	Mean	Max	Min	Variance	CV
BOD5	mg/L	2.242	4	1	0.581	33.988
Cl	mg/L	247.2532	377.52	166.76	3711.26	24.611
Na	mg/L	179.478	233.54	117.82	905.716	16.768
pH	-	7.786	8.5	6.6	0.293	6.952
SO4	mg/L	292.819	446.1	215.7	3693.55	20.755
T	°C	15.42	21	9	14.352	24.568
TDS	mg/L	1127.36	1524	875	23007.6	13.455
Turb	NTU	8.906	19.6	0.4	28.427	59.868

For the purpose of imputation, the missing parameter estimation was performed using linear regression between the two parameters with the highest correlations in this table (Chen *et al.* 2022). The data set deficiencies were removed by using this method. (Burchard-Levine *et al.* 2014). For example, for SO4 imputation, the Cl data with the highest correlation is chosen. The table shows that the correlation between SO4 and Cl is 0.94307. Similarly, for lack of TSD values, the SO4 data with the highest correlation was chosen (0.842242). The estimation model is built using the linear regression equations listed below (1-3) (Ortas *et al.* 2019). **Figure 4** displays the residual graph and linear regression graph for SO4 vs. Cl. Also, correlation is very high between these parameters.

$$\alpha = \frac{\Sigma\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\Sigma(x_i - \bar{x})^2} \quad (1)$$

$$\beta = \bar{y} - \alpha\bar{x} \quad (2)$$

$$y_i = \alpha x_i + \beta \quad (3)$$

α = slope

β = intercept

x_i = value of parameter

$\bar{x}$ = mean of parameter

y_i = value of parameter

$\bar{y}$ = mean of parameter

Likewise, a prediction model was developed with SO4 data within the TDS value. The graph prepared for the developed model is given below. **Figure 5** shows the residual plot and linear regression plot for TDS and SO4. In residual graph, only one value appears to be too far out.

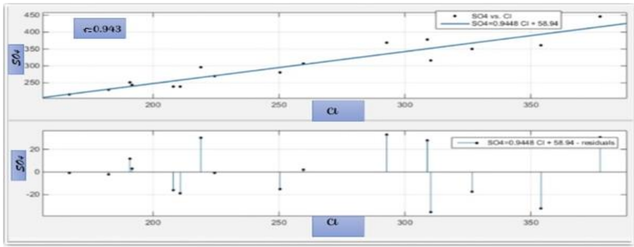


Figure 4. Linear regression and residual value plot between SO4 and Cl parameters.

Outliers and missing values are therefore deleted and rearranged in the data set. External values and defects in the entire data set are removed using this procedure. Table 4 shows the statistical distribution of the arranged data set. The variance of the TDS value is 23007.6, as shown in the table, whereas the CV value is estimated as 13.455, which is rather high. This value shows that the TDS data is distributed around the mean. Likewise, the variance of SO4 value is 3693.55, while the CV value is 20.755, indicating that the available data are distributed around the mean. Despite the fact that the variance of the turbidity value was not very high, the CV value was determined as the highest (59.868). Since a portion of the river basin where the study was done is in the Black Sea region, as is the station, the data is considered valid. There is abundant rainfall and dense forest cover in the Black Sea region. For this reason, many floods mix with the river along with rainfall. As a consequence, substantial CV in Turbidity values is expected. Furthermore, for compatibility, these data were compared to data from other stations on the river, and it was verified that the parameters were compatible with the data from other stations.

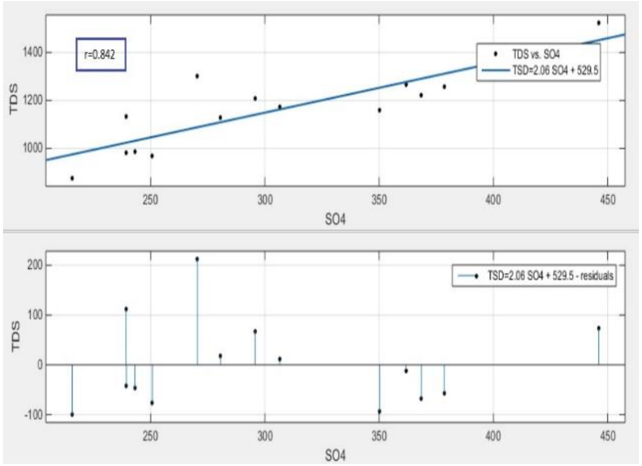


Figure 5. Linear regression and residual value plot between TDS and SO4 parameters.

Many methods have been developed by researchers for water quality monitoring. Some of them are modeling systems, while others are various systems ranging from Artificial Neural Network (ANN) to Genetic Algorithm (GA). In others, methodologies are used to investigate the correlations between water quality measures. In other methods, relationships between parameters of water quality are examined. Prediction models have been developed that allow water quality monitoring of one or more parameters (Burchard-Levine *et al.* 2014; Mishra, S. *et al.* 2016..Sakizadeh, 2016; Barcellos & Souza, 2022, Yeon L. S. *et al.* 2008).

3. Results

3.1. Multivariate statistics results

The correlation values of these parameters selected for multivariate statistical analysis after outlier analysis and missing value adjustment are given in Table 5.

Figure 6 (a) and (b) show the highly connected (Na and Cl) and (pH and Turb) plots. Despite the impressive correlation of this data, linear regression does not provide adequate efficiency in parameter estimation. Because the correlation coefficients for the Cl parameter are 0.947 for SO4 and 0.820 for TSD when checked in the correlation table. This demonstrates that the representation of some parameters with many parameters is high. **Figure 6** (c) depicts this situation with a graph of Cl - SO4 - TDS.

Table 5 Adjusted experimental data set correlation table.

Parameters	BOD5	Cl-	Na+	pH	SO4=	T	TDS	Turb
BOD5	1,0000	-0,4317	-0,3416	-0,1591	-0,2534	-0,2301	-0,2454	0,1840
Cl-	-0,4317	1,0000	0,7277	-0,0959	0,9471	-0,1852	0,8199	-0,2669
Na+	-0,3416	0,7277	1,0000	-0,1299	0,6166	-0,1432	0,6079	-0,1118
pH	-0,1591	-0,0959	-0,1299	1,0000	-0,1920	0,1022	-0,1965	0,6467
SO4=	-0,2534	0,9471	0,6166	-0,1920	1,0000	-0,1377	0,8613	-0,3077
T	-0,2301	-0,1852	-0,1432	0,1022	-0,1377	1,0000	0,0053	0,0126
TDS	-0,2454	0,8199	0,6079	-0,1965	0,8613	0,0053	1,0000	-0,3931
Turb	0,1840	-0,2669	-0,1118	0,6467	-0,3077	0,0126	-0,3931	1,0000

Table 6 Estimation model results for four selected parameters, as well as comparison statistics obtained from the data set.

Parameters	MSE	MAPE	r
SO4	151.061	3.731	0.979
Cl	118.716	3.627	0.984
Na	292.661	8.695	0.822
BOD5	0.224	20.737	0.784

As shown in the graphs and the correlation table, several parameters have more than one effective parameter. The correlation chart in Table 5 shows that the effect of T parameter (on other parameters) is not particularly strong. This parameter's maximum absolute correlation's value is about 0.23 (Table 4). Although the correlation value had a minor effect on the parameters, it was included in the model. The rationale for this is that it is the most easily ascertained parameter physically, and it has been chosen and included in the evaluation to help to the model's progress, albeit in a minor way.

Moreover, the coefficient ratios of all factors have different effects on the prediction models. A linear prediction model was created by analyzing the effect of other parameters on one parameter using multivariate statistical approaches (Liu *et al.* 2021; Maiolo & Pantusa, 2021; Betrie *et al.* 2016). In general, the linear model is given in **Figure 7** and in equation 4.

$$\hat{y} = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6 + \alpha_7 x_7 + \beta \qquad (4)$$

α_i = each parameters co efficients

x_i = each parameter

β = intercepts of errors.

3.2. Statistical tests results

According to the selected parameter, the coefficient in front of the parameters was determined one by one according to the least squares method. These are $\alpha_1 \ldots \alpha_7$ coefficients and β value, and a prediction model has been developed for 4 different parameters to examine the established modeling system. With the multi-statistical model created (for the selected parameter), $\hat{y}$(estimate) values were derived, and correlation coefficient (r) with equation 5, Mean Squared Error (MSE) with Equation 6, and Mean Absolute Percentage Error (MAPE) with equation 7 on the obtained values and existing (edited experimental) values were examined (Wang *et al.* 2012; Salman *et al.* 2017; Nicolson & Paliwal, 2019; Bagla *et al.* 2021; Jiang *et al.* 2021;Longqin &Shuangyin Liu ,2013)

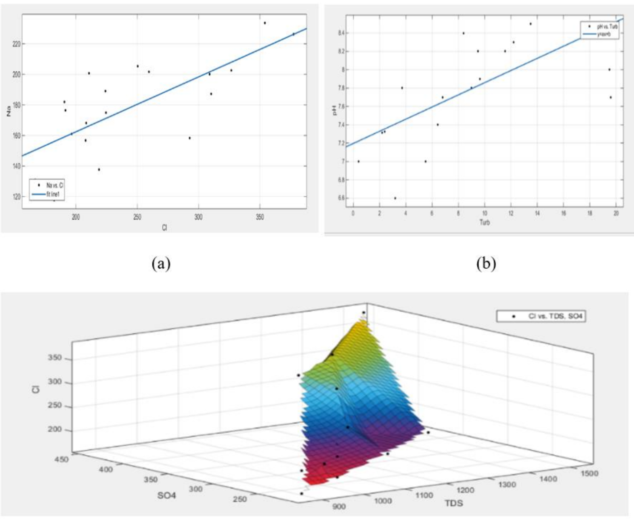
$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}} \tag{5}$$

$$MSE = \frac{\sum\left(y_i - y_i\right)^2}{n} \tag{6}$$

$$MAPE = \frac{\sum\left(\frac{Abs(residual)}{real}\right)}{n} \times 100 \tag{7}$$

residual = y_i-($\hat{y}_i$)

The statistical table for the r, MSE, and MAPE values derived for the SO4, Cl, Na, and BOD5 models is provided below.



(a) (b) (c)

Figure 6. a) Na vs Cl linear regression graph b) pH vs Turb linear regression graph c) Cl vs SO4, TDS linear regression graph.

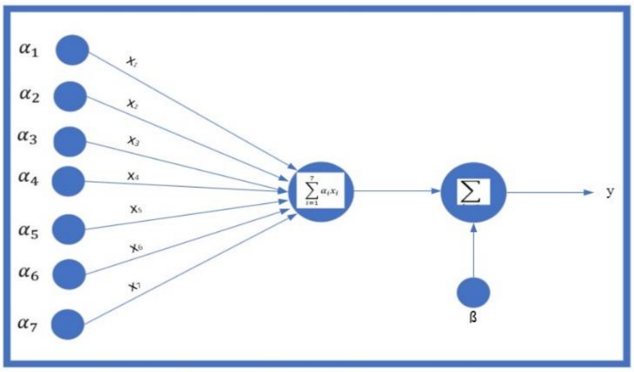


Figure 7. Linear prediction model general representation.

As seen in Table 6, the MAPE values of the models were SO4, Cl, and for Na with MAPE <10% it produces value for the parameter predicted with very high accuracy (MAPE < 10% in scientific studies for model studies: It gives excellent prediction performance). The MAPE value for BOD5 is roughly 20%, which is within safe ranges. Correlation values in these parameters between the experimental data and the prediction model demonstrated that the models were compatible for SO4, Cl, and Na. Experimental data and prediction model' results are also within acceptable limits for BOD5. In addition, the fact that the MSE value for BOD5 is quite small shows that the experimental data and the prediction model produce results that are quite compatible.

4. Conclusion

4.1. Conclusion for prediction model

Figure 8 and equation 8 shows the generated Cl prediction model. The coefficients in the models built individually for the other parameters were derived separately based on the parameters.

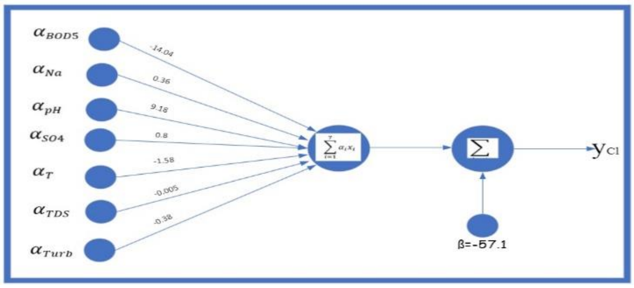


Figure 8. Cl prediction model with parameters and weight.

$$\hat{y}_{Cl} = -14.04\alpha_{BOD5} + 0.36\alpha_{Na+} + 9.18\alpha_{pH} + 0.8\alpha_{SO4} - 1.58\alpha_T - 0.005\alpha_{TDS} - 0.38\alpha_{Turb} - 57.1 \tag{8}$$

The agreement between the experimental data and the estimating moles created by the multivariate statistical analysis approach and the data acquired was shown using comparison charts. The Cl comparison graph is shown in **Figure 9**. When the graph is examined, the data are found to be quite compatible. Additionally, the percentage difference between the values can be seen in the graph.

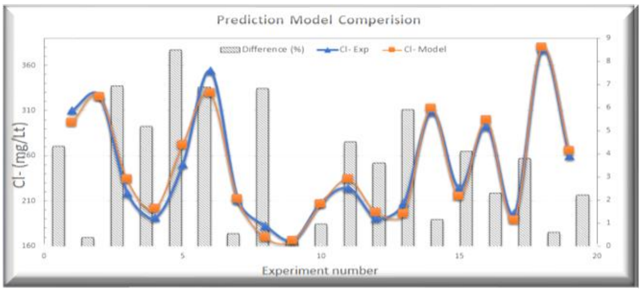


Figure 9. Experimental (Cl-Exp) and prediction model (Cl-Model) comparison plot for Cl with difference percentage.

Figure 10 shows the graph generated using the simulation model for Na in multivariate statistical modeling done using the least squares method. Although the MSE value on the tested data has the greatest value, the MAPE and r values demonstrate that the prediction model yields findings that are consistent.

Figure 11 shows the graph obtained using the BOD5 model defined for multivariate statistical modeling using the least squares method. It is the prediction model with the smallest MSE value and generates an increase and decrease trend prediction value that is generally accurate.

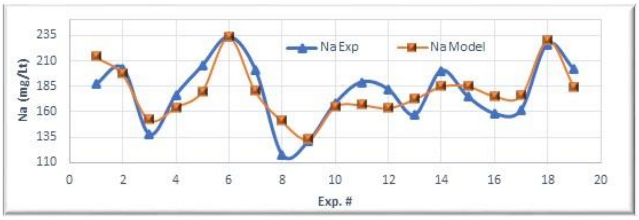


Figure 10. Experimental (Na Exp) and model (Na model) comparison plot for Na.

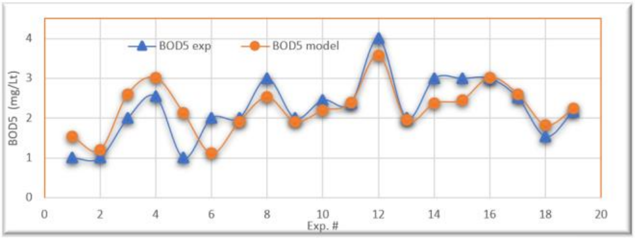


Figure 11. Experimental (BOD5 exp) and model (BOD5 model) comparison plot for BOD5.

The graphic generated by using model developed for SO4 is shown in **Figure 12**. in the multivariate statistical modeling carried out using the least squares method. The graph shows that there is a clear agreement between the forecast model findings and the actual data. Additionally, Table 6's correlation value for the data sets is very near to 1.

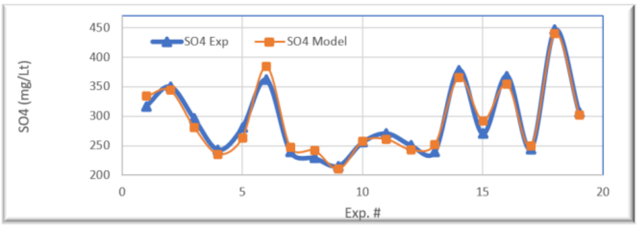


Figure 12. Experimental (SO4 Exp) and model (SO4 Model) comparison plot for SO4.

4.2. Graphical User Interface for prediction model

As stated above, the data set was edited and finalized after missing data and outlier analysis. After data editing studies, a water quality parameter estimation model was developed. The data obtained was tested statistically with real river data and the model studies carried out for four different parameters were found to be successful. However, the fact that the study has more than one phase makes it difficult for researchers to follow the process steps. For this reason, the process steps detailed above were combined in a single program and a user-friendly GUI was created using VBScripts to allow researchers to use the model in different data sets. Additionally, in order to simplify the GUI for researchers to understand more easily, a program that includes the working steps for only one water quality parameter (Cl) has been created in this article. The created model interface is seen in **Figure 13**. The process steps for the interface are as follows:

1. Reading the edited data set
2. Statistical analysis
a. Average
b. Maximum
c. Minimum
d. Variance
e.CV (Coefficient of Variation)
3. Multiple regression (coefficients)
4. Statistical tests (with actual values)
5. Water quality parameter estimation (Cl) with the model (default value) enter any value
6. Predicted Cl value

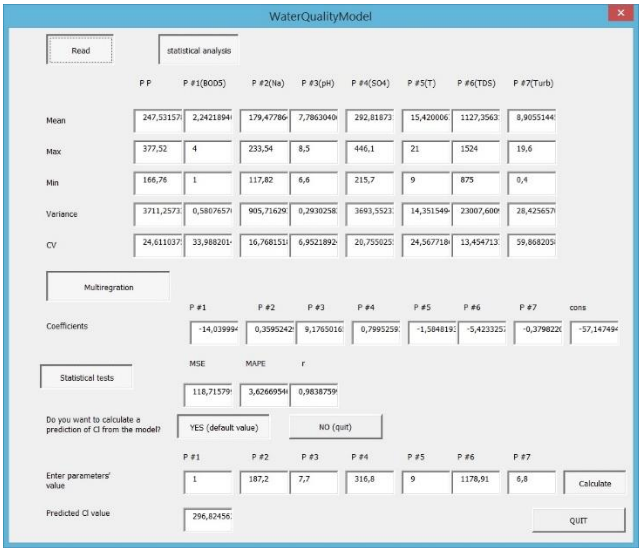


Figure. 13. Graphical User Interface for water quality parameter prediction model (for Cl).

4.3. Conclusion for normalized prediction model

The estimates for SO4, Cl, Na, and BOD5 that were generated using the multi-statistical model are thought to be quite compatible with the actual MSE, and MAPE, and correlation values. However, it is difficult for researchers to determine which parameter has the most impact on the chosen parameter because of the variations in the units and data ranges of the parameters. Additionally, the agreement appears to be low in the charts for the parameters with high agreement in the charts discussed earlier. This is because the ranges in the relevant parameters are different, for example, the parameter range values used for BOD5 are between 0 and 4, while the parameter ranges used for SO4 are between 210 and 410. This makes it difficult to interpret graphs and statistical analyzes together. In order for the researchers to make this interpretation more easily and to understand the effects of the parameters on each other much more clearly, the existing data were normalized (equation 9) and all the parameters were taken to the range of 0-1 (Lumb *et al.* 2011; Simões *et al.* 2008; Ni & Chen, 2013; Y. Zhang *et al.* 2022). For this;

$$X_{Ni} = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \tag{9}$$

all experimental data were normalized using the above formula, and the previous procedures were carried out again for the four parameters that were chosen. The data set's parameters were first correlated. Table 7 lists the correlation coefficients for normalized data. The estimation model's equation 10, which is reorganized using normalized values, is provided. Equation 11 derived for the Cl parameter is provided as an example for the coefficient evaluation.

Table 7. Correlations of parameters on normalized data.

Parameters	BOD5	Cl	Na	pH	SO4	T	TDS	Turb
BOD5	1	-0.432	-0.342	-0.159	-0.253	-0.230	-0.245	0.191
Cl	-0.432	1	0.728	-0.096	0.947	-0.185	0.820	-0.245
Na	-0.342	0.728	1	-0.130	0.617	-0.143	0.608	-0.090
pH	-0.159	-0.096	-0.130	1	-0.192	0.102	-0.197	0.652
SO4	-0.253	0.947	0.617	-0.192	1	-0.138	0.861	-0.277
T	-0.230	-0.185	-0.143	0.102	-0.138	1	0.005	0.002
TDS	-0.245	0.820	0.608	-0.197	0.861	0.005	1	-0.356
Turb	0.191	-0.245	-0.090	0.652	-0.277	0.002	-0.356	1

Table 8. Statistical table of prediction models (Model and normalized model (Norm model)).

Parameters	MSE		MAPE		r	
	Model	Norm_model	Model	Norm_model	Model	Norm_model
SO4	151.061	151.4883	3.731	3.748	0.979	0.979
Cl	118.716	118.716	3.627	3.627	0.984	0.984
Na	292.661	292.661	8.695	8.695	0.823	0.823
BOD5	0.224	0.224	20.737	20.737	0.784	0.784

$$\hat{y} = ((\alpha_{n1}x_{n1} + \alpha_{n2}x_{n2} + \alpha_{n3}x_{n3} + \alpha_{n4}x_{n4} + \alpha_{n5}x_{n5} + \alpha_{n6}x_{n6} + \alpha_{n7}x_{n7} + \beta n) * (y_{max} - y_{min})) + y_{min}. \tag{10}$$

$$\hat{y}_{Cl} = ((-0.2x_{BOD5} + 0.197x_{Na} + 0.083x_{pH} + 0.874x_{SO4} - 0.09x_{T} - 0.0167x_{TDS} - 0.0346x_{Turb} + 0.0867) * (377.52 - 166.76)) + 166.76 \tag{11}$$

Each of the coefficients derived from the developed Cl prediction model reveals its effects on the model more clearly. The SO4 concentration, for instance, has a coefficient of 0.874 and is the most useful parameter in the Cl prediction model. Table 5's correlation data show that the correlation between Cl and SO4 is 0.9471. The table below contains a comparison of experimental Cl data and Cl values (**Figure 14**). The normalized model (Norm_model) performance is also fairly strong, as seen in the graph.

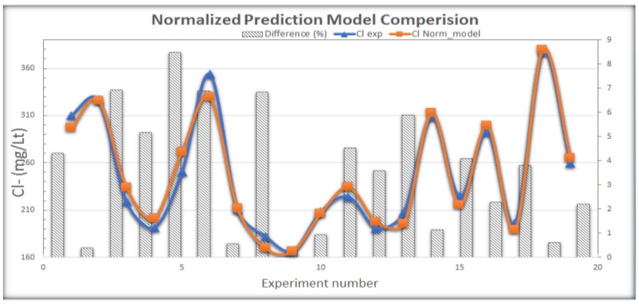


Figure 14. Experimental and Normalized prediction model comparison plot for Cl with difference percentage.

The data are highly comparable when compared to the Normalized prediction model (Norm model) graphs and the Cl prediction model (model) graphs shown in **Figures 9 and 14** above. In addition, other parameters (Na, SO4, and BOD5) were studied and the compatibility of graphics and models was observed. In order to understand the comparison, the aggregated table (Table 8) of the r, MSE and MAPE values of the models and normalized models is shown below. For researchers working on the topic, it will be simpler and clearer to evaluate the impact of each parameter on the parameter to be estimated in the estimation model created over normalized models.

The Kızılırmak River, where the investigation was done, was given the created equations and estimating models. The parameter values of each river and their connections to the parameters make it impossible to consider this equation to be a general one. This study's objective is to show the modeling steps that can be used for each station with the modeling steps completed.

5. Discussion

Water quality parameters are of great importance in monitoring river pollution. It is necessary to analyze these parameters regularly, follow the changes and obtain information about the river. However, compiling this data is a difficult process and requires high costs. Along with the cost of analysis materials, expert personnel expenses are also quite high. Some problems encountered during the collection of this data are:

1. In on-site analyses, the values of some parameters are recorded incorrectly due to seasonal factors.

2. Due to the same seasonal reasons, the sample used in the measurement of some parameters cannot be stored properly, resulting in the experiments not being carried out, which leads to incomplete data.

3. Difficulties in tracking or estimating may arise if on-site measurements of certain parameters are disregarded or not gathered because of expense.

In this research, various solution methods for the three problems mentioned above were examined and the most suitable methods were tested with real data. To address analysis errors, the widely accepted boxplot method developed by Tukey, J. W. (1977) in the literature has been used to clean possible outliers in the data set.

There are many methods in the literature to correct missing data. Among these methods, linear regression was considered to the most suitable for data with seasonal fluctuations. To complete the missing data, correlation analysis was performed to determine the relationship between the parameters, and an equation was created between the highly correlated parameters by determining the slope and intercept to create a linear regression equation with the existing data. Linear regression equation of more related parameter has been used as input for missing data. This method was repeated for each missing data and all deficiencies in the data set were completed.

At this stage, due to most of the data from February 2013 being missing, no imputation was done for this month, and all data for this month has been removed from the dataset The data set obtained after this stage was used for multivariate analyses.

Starting from the hypothesis of whether this parameter would be predictable if one of the parameters in the data set was missing, statistical analyzes were carried out again and parameters that could be estimated based on correlations were determined. At this stage, a multivariate regression model was created to determine the 8th parameter using 7 parameters. During the research, studies were conducted on the parameters SO4, Cl-, Na+, and BOD5. However, comparison tables are presented in the article only for Cl. In addition, the entire data set was normalized, the multivariate prediction model was run again, and comparison tables with real values were prepared so that the researchers could examine and interpret the parameters. It was examined whether the data obtained from the prediction model was within acceptable limits through statistical analysis.

The model obtained in this study was also examined with data obtained from different rivers, but the developed prediction model did not produce very good results on different rivers. This is because rivers have different characteristics from each other. However, each river can be modeled uniquely by performing prediction model studies with a good data set. It is possible to perform predictive modeling studies on all rivers and get positive results with new and comprehensive data sets, the process steps of which are presented in this study.

In this study, Minitab, Excel, and MATLAB program softwares were used for statistical analysis, calculations, and graphics. VBscript was user for GUI

6. Recommendation

In many nations, there are legislative regulations governing the routine monitoring of water resources, and these regulations are implemented by approved institutions and organizations. Experts have developed a variety of very different techniques for monitoring water resources. Water quality monitoring can be achieved using various models, from ANNs to genetic algorithms and artificial intelligence. In this study, data taken from one of the stations of the Kızılırmak River was used, and the analyzes were carried out by the state institution responsible for legal water monitoring. Because of data pollution, a problem that always exists with data sets, the series set was first checked for outliers using the boxplot approach, and the data new set was then purified of them. Values captured in the data set due to extreme values and seasonal factors were eliminated and imputation operations were carried out for missing data and the data set was arranged. Thus, a new and reliable data set was created.

8 parameters from the edited data set (BOD5, Cl, Na, pH, SO4, T, TDS, Turb) were selected for the study. A multivariate statistical model was developed using the least squares approach and a prediction model was created for SO4, Cl, Na and BOD5. MSE, r, and MAPE values were used to study the accuracy of the model and compare it with real values, and it was determined that the developed prediction model gave a usable prediction value. The data set's parameter values were normalized to make it easier for researchers to see the effects of the parameters and to interpret the graphs and statistical tables. All values were converted to the range of 0–1, and new models were established for the chosen parameters, even though the estimation model produced accurate results. These estimating models' output values passed the tests of statistical analysis (MSE, r, and MAPE) and were judged to be accurate.

The following is an important point that researchers should focus on because, the relationships between the parameters of each river vary and it is not possible to utilize the above - derived coefficients and provided models directly. The variables identified by this study are related to one another. Examining the properties of many rivers reveals that even the limit values and average values vary. In terms of techniques and coefficients, this paper offers significant support to academics working on ANN modeling and machine learning algorithms. The output values of these estimating models were found to be accurate and to have passed the statistical analysis tests (MSE, r, and MAPE).

The following is a crucial topic that researchers should focus on because, the relationships between the parameters of each river vary and it is not possible to utilize the above-derived coefficients and provided models directly. The variables identified by this study are related to one another. Examining the properties of many rivers reveals that even the limit values and average values vary. In terms of techniques and coefficients, this paper offers great support to academics working on ANN modeling and machine learning algorithms.

A post-study recommendation to the researchers might be to develop a multivariate statistical estimating model employing data from more stations along the same river, for this purpose, samples taken from more station locations in shorter time intervals should be taken into account in the calculations. Additional parameters affected by the values of the change rates of these parameters are expressed only in that river.

The computations for this should take samples from more station locations into account.

Additionally, by employing the above methods and adding more parameters to the models while using a larger data set, more successful models can be attained in parameter estimations of the river of interest.

Statements & declarations

There isn't a conflict of interest or duty sharing because this study has just one author. There is no private information in the article because the data used was tabulated.

Artificial intelligence was not used in this research.

Ethical approval

This study is a research carried out for “Water Quality Parameter Estimation Method". The purpose of this study is to show that a parameter that an unmeasured parameter can be estimated by a multivariate regression method with eliminating missing data. Research data were taken from State Water Works. The statistical values of the data were shared for study confidentiality reasons. There is no personal information in the study.

Consent to participate

No personal data was used in the study.

Consent to publish

“I have not submitted my manuscript to a preprint server before submitting it to Environmental Science and Pollution Research”

Authors contributions

There is only one writer.

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

The author did not receive any financial support related to this study.

References

Abyaneh, H. Z. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*, 12(40). https://doi.org/10.1186/2052-336X-12-40

Ahmad, S., Khan, I. H. and Parida, B. P. (2001). Performance of Stochastic Approaches for Forecasting River Water Quality. In *Water Research* (Vol. 35, Issue 18). https://doi.org/10.1016/S0043-1354(01)00167-1

Alemu Agegnehu, Bitew Woinitu, Anteneh Zelalem Liyew. (2024). Assessment of the Quality of Drinking Water Soursec in Bahir Dar City, Ethiopia. *Air Soil and Water Research* (Vol 17). https://doi.org/10.1177/11786221241301987

Anifowose, B. A. and Odubela, M. T. (2018). Oil facility operations: A multivariate analyses of water pollution parameters. *Journal of Cleaner Production*, 187, 180–189. https://doi.org/10.1016/j.jclepro.2018.03.044

Azhar, S. C., Aris, A. Z., Yusoff, M. K., Ramli, M. F. and Juahir, H. (2015). Classification of River Water Quality Using Multivariate Analysis. *Procedia Environmental Sciences*, 30, 79–84. https://doi.org/10.1016/j.proenv.2015.10.014

Bagla, P., Kumar, K., Sharma, N. and Sharma, R. (2021). Multivariate Analysis of Water Quality of Ganga River. *Journal of The Institution of Engineers (India): Series B*, 102(3), 539–549. https://doi.org/10.1007/s40031-021-00555-z

Barcellos, D. da S. and Souza, F. T. de. (2022). Optimization of water quality monitoring programs by data mining. *Water Research*, 221. https://doi.org/10.1016/j.watres.2022.118805

Betrie, G. D., Sadiq, R., Tesfamariam, S. and Morin, K. A. (2016). On the Issue of Incomplete and Missing Water-Quality Data in Mine Site Databases: Comparing Three Imputation Methods. *Mine Water and the Environment*, 35(1), 3–9. https://doi.org/10.1007/s10230-014-0322-4

Burchard-Levine, A., Liu, S., Vince, F., Li, M., & Ostfeld, A. (2014). A hybrid evolutionary data driven model for river water quality early warning. *Journal of Environmental Management*, 143, 8–16. https://doi.org/10.1016/j.jenvman.2014.04.017

Chen, X., Strokal, M., van Vliet, M. T. H., Fu, X., Wang, M., Ma, L., & Kroeze, C. (2022). In-stream surface water quality in China: A spatially-explicit modelling approach for nutrients. *Journal of Cleaner Production*, 334. https://doi.org/10.1016/j.jclepro.2021.130208

Dawson, R. (2011). How significant is a boxplot outlier? *Journal of Statistics Education*, 19(2). https://doi.org/10.1080/10691898.2011.11889610

Fathi, E., Zamani-Ahmadmahmoodi, R. and Zare-Bidaki, R. (2018). Water quality evaluation using water quality index and multivariate methods, Beheshtabad River, Iran. *Applied Water Science*, 8(7). https://doi.org/10.1007/s13201-018-0859-7

Gurjar, S. K. and Tare, V. (2019). Spatial-temporal assessment of water quality and assimilative capacity of river Ramganga, a tributary of Ganga using multivariate analysis and QUEL2K. *Journal of Cleaner Production*, 222, 550–564. https://doi.org/10.1016/j.jclepro.2019.03.064

Horvat, Z., Horvat, M., Pastor, K., Bursić, V. and Puvača, N. (2021). Multivariate analysis of water quality measurements on the danube river. *Water* (Switzerland), 13(24). https://doi.org/10.3390/w13243634

Isaac, R. and Siddiqui, S. (2022). Application of water quality index and multivariate statistical techniques for assessment of water quality around Yamuna River in Agra Region, Uttar Pradesh, India. *Water Supply*, 22(3), 3399–3418. https://doi.org/10.2166/WS.2021.395

Islam, Khan, M. S., Islam, N., Uddin, J., Islam, S. and Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University - Computer and Information Sciences*, 34(8), 4773–4781. https://doi.org/10.1016/j.jksuci.2021.06.003

Jiang, Y., Li, C., Sun, L., Guo, D., Zhang, Y. and Wang, W. (2021). A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks. *Journal of Cleaner Production*, 318. https://doi.org/10.1016/j.jclepro.2021.128533

Kazi, T. G., Arain, M. B., Jamali, M. K., Jalbani, N., Afridi, H. I., Sarfraz, R. A., Baig, J. A. and Shah, A. Q. (2009). Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. *Ecotoxicology and Environmental*

Safety, 72(2), 301–309. https://doi.org/10.1016/j.ecoenv.2008.02.024

Li, T., Li, S., Liang, C., Bush, R. T., Xiong, L. and Jiang, Y. (2018). A comparative assessment of Australia's Lower Lakes water quality under extreme drought and post-drought conditions using multivariate statistical techniques. *Journal of Cleaner Production*, 190, 1–11. https://doi.org/10.1016/j.jclepro.2018.04.121

Liu, Z., Zhu, H., Cui, X., Wang, W., Luan, X., Chen, L., Cui, Z. and Zhang, L. (2021). Groundwater quality evaluation of the Dawu water source area based on water quality index (WQI): Comparison between Delphi method and multivariate statistical analysis method. *Water* (Switzerland), 13(8), NA. https://doi.org/10.3390/w13081127

Longqin Xu and Shuangyin Liu. (2013) Study of Short-Term Water Quality Prediction Model Based on Wavelet Neural Network. *Mathematical and Computer Modelling*, (58), 807–813. doi:10.1016/j.mcm.2012.12.023

Lumb, A., Sharma, T. C. and Bibeault, J.-F. (2011). A Review of Genesis and Evolution of Water Quality Index (WQI) and Some Future Directions. *Water Quality, Exposure and Health*, 3(1), 11–24. https://doi.org/10.1007/s12403-011-0040-0

Ma, X., Wang, L., Yang, H., Li, N. and Gong, C. (2020). Spatiotemporal analysis of water quality using multivariate statistical techniques and the water quality identification index for the qinhuai river basin, east china. *Water* (Switzerland), 12(10). https://doi.org/10.3390/w12102764

Maiolo, M. and Pantusa, D. (2021). Multivariate analysis of water quality data for drinking water supply systems. *Water* (Switzerland), 13(13). https://doi.org/10.3390/w13131766

Mishra Saurabh, Kumar Amit and Prabhakar Shukla (2016). Study of water quality in Hindon River using pollution index and environmetrics, India, (57) 19121–19130. https://doi.org/10.1080/19443994.2015.1098570

Ni, J. and Chen, X. (2013). Steady-state mean-square error analysis of regularized normalized subband adaptive filters. *Signal Processing*, 93(9), 2648–2652. https://doi.org/10.1016/j.sigpro.2013.03.030

Nicolson, A. and Paliwal, K. K. (2019). Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Communication*, 111, 44–55. https://doi.org/10.1016/j.specom.2019.06.002

Noori, R., Abdoli, M. A., Ameri Ghasrodashti, A. and Jalili Ghazizade, M. (2009). Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: A case study of mashhad. *Environmental Progress and Sustainable Energy*, 28(2), 249–258. https://doi.org/10.1002/ep.10317

Noori, R., Sabahi, M. S., Karbassi, A. R., Baghvand, A. and Zadeh, H. T. (2010). Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. *Desalination*, 260(1–3), 129–136. https://doi.org/10.1016/j.desal.2010.04.053

Ortas, E., Burritt, R. L., and Christ, K. L. (2019). The influence of macro factors on corporate water management: A multi-country quantile regression approach. *Journal of Cleaner Production*, 226, 1013–1021. https://doi.org/10.1016/j.jclepro.2019.04.165

Pande, G., Agrawal, S. and Sinha, D.A. (2015) Impacts of Leachate Percolation on Ground Water Quality: A Case Study of Dhanbad City, *Global NEST Journal*, 17(1). https://doi.org/10.30955/gnj.001377.

Sadiq, Q., Ezeamaka, C. K., Daful, M. G. and Mustafa, I. A. (2022). Evaluation of the Water Quality of River Kaduna, Nigeria Using Water Quality Index. *Environmental Technology and Science Journal*, 13(1), 28–40. https://doi.org/10.4314/etsj.v13i1.3

Sakizadeh, M. (2016). Artificial intelligence for the prediction of water quality index in groundwater systems. *Modeling Earth Systems and Environment*, 2(1). https://doi.org/10.1007/s40808-015-0063-9

Salman, M. S., Kukrer, O. and Hocanin, A. (2017). Recursive inverse algorithm: Mean-square-error analysis. *Digital Signal Processing: A Review Journal*, 66, 10–17. https://doi.org/10.1016/j.dsp.2017.04.001

Setshedi, K. J., Mutingwende, N. and Ngqwala, N. P. (2021). The Use of Artificial Neural Networks to Predict the Physicochemical Characteristics of Water Quality in Three District Municipalities, *Eastern Cape Province*, South Africa. https://doi.org/10.3390/10.3390/ijerph18105248

Simões, F. dos S., Moreira, A. B., Bisinoti, M. C., Gimenez, S. M. N. and Yabe, M. J. S. (2008). Water quality index as a simple indicator of aquaculture effects on aquatic bodies. *Ecological Indicators*, 8(5), 476–484. https://doi.org/10.1016/j.ecolind.2007.05.002

Sun, X., Zhang, H., Zhong, M., Wang, Z., Liang, X., Huang, T. and Huang, H. (2019). Analyses on the temporal and spatial characteristics of water quality in a seagoing river using multivariate statistical techniques: A case study in the Duliujian river, China. *International Journal of Environmental Research and Public Health*, 16(6). https://doi.org/10.3390/ijerph16061020

Sundaray, S. K., Panda, U. C., Nayak, B. B. and Bhatta, D. (2006). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of the Mahanadi river-estuarine system (India) - A case study. *Environmental Geochemistry and Health*, 28(4), 317–330. https://doi.org/10.1007/s10653-005-9001-5

Tripathi, M. and Singal, S. K. (2019). Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India. *Ecological Indicators*, 96, 430–436. https://doi.org/10.1016/j.ecolind.2018.09.025

Tukey, J. W. (1977). Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley. ISBN:9780201076165, 0201076160

Uncumusaoğlu Aylin A. (2018) Statistical assessment of water quality parameters for pollution source identification in Bektaş Pond (Sinop, Turkey). Global NEST Journal, Vol. 20, No 1, pp 151–160. https://doi.org/10.30955/gnj.002369

Wang, J., Zhang, Y., Cao, H. and Zhu, W. (2012). Dimension reduction method of independent component analysis for process monitoring based on minimum mean square error. *Journal of Process Control*, 22(2), 477–487. https://doi.org/10.1016/j.jprocont.2011.11.005

Wenyan Li, Bingjun Li and Wenyan Ma. (2023) Prediction of agricultural grey water footprint in Henan Province based on GM(1, N)-BP neural network. *Environmental and Ecological Statistics*, 30, 335–354. https://doi.org/10.1007/s10651-023-00559-6

Yang, R. (2022). Analyses of Approaches to Deal with Missing Data in Water Quality Data Set. *Advances in Economics*, Business and Management Research, 622, https://doi.org/10.2991/aebmr.k.220405.184

Yeon, L. S., KIM, J. H. and JUN, K.W., (2008) Application of Artificial Intelligence Model in Water Quality Forecasting 29, 625 – 631. https://doi.org/10.1080/09593330801984456

Yidana, S. M. and Yidana, A. (2010). Assessing water quality using water quality index and multivariate analysis. *Environmental Earth Sciences*, 59(7), 1461–1473. https://doi.org/10.1007/s12665-009-0132-3

Zhang, Y., Li, C., Jiang, Y., Sun, L., Zhao, R., Yan, K. and Wang, W. (2022). Accurate prediction of water quality in urban drainage network with integrated EMD-LSTM model. *Journal of Cleaner Production*, 354. https://doi.org/10.1016/j.jclepro.2022.131724

Zhang, Y. and Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*, 128, 63–72. https://doi.org/10.1016/j.future.2021.09.033

Zhang, Z. M., Zhang, F., Du, J. L. and Chen, D. C. (2022). Surface Water Quality Assessment and Contamination Source Identification Using Multivariate Statistical Techniques: A Case Study of the Nanxi River in the Taihu Watershed, China. *Water* (Switzerland), 14(5). https://doi.org/10.3390/w14050778

Zhou, Y., Liu, Y., Wang, D., De, G., Li, Y., Liu, X. and Wang, Y. (2021). A novel combined multi-task learning and Gaussian process regression model for the prediction of multi-timescale and multi-component of solar radiation. *Journal of Cleaner Production*, 284. https://doi.org/10.1016/j.jclepro.2020.124710