Multiple Linear Regression Modeling for Analysis of Factors Affecting COD and BOD on River Water Quality in Yogyakarta, Indonesia

Muhammad Andang Novianta[1,2,*], Syafrudin [3,4], *Budi* Warsito [4,5]

[1]Students Study Program of Doctoral Environmental Science, School of Postgraduate Studies, Diponegoro University, Semarang 50275, Indonesia
[2]Department of Electrical Engineering, Faculty of Industrial Technology, Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia
[3]Department of Environmental Engineering, Faculty of Engineering, Diponegoro University, Semarang 50275, Indonesia
[4]Study Program of Doctoral Environmental Science, School of Postgraduate Studies, Diponegoro University, Semarang 50275, Indonesia
[5]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang 50275, Indonesia

Abstract. Many factors can affect the quality of river water in DIY, both the activities of the population and industry. Several river water quality parameters that can be used to determine the health condition of river water are Chemical Oxygen Demand (COD) and Biological Oxygen Demand (BOD). This study tested the effect of TSS and DO on BOD and COD in 10 rivers in DIY. The method used is multiple linear regression modeling. Based on hypothesis testing in multiple linear regression with a significance level of 5%, it is found that TSS and DO significantly affect BOD and COD conditions in 2019. Furthermore, in 2020 only DO significantly affects COD. The prediction result is that if TSS is high then BOD and COD will be predicted to have high value. If DO is high then COD and BOD will be predicted to be low.

1 Introduction

River water quality parameters consist of physical, chemical, and biological parameters. Chemical Oxygen Demand (COD) and Biological Oxygen Demand (BOD) are chemical parameters. COD indicates the total amount of oxygen required to chemically oxidize organic matter or indicates the level of inorganic waste as measured by the amount of oxygen required to break down inorganic waste. If the water contains a lot of inorganic waste, the amount of oxygen needed by microorganisms to break down the waste will be large, so the COD number will also be high. Meanwhile, BOD is a measure of the amount of oxygen used by microbial populations contained in waters in response to the entry of organic matter that can be decomposed. BOD indicates the amount of easily decomposed organic matter present in the waters. If the water contains a lot of organic waste, the amount of oxygen needed by microorganisms to break down the waste will be large, so the BOD number will also be high. Research on water quality was also carried out by previous researchers including using a clustering analysis algorithm. His research focuses on classifying rivers based on water quality classes [1] and clustering [2]. Furthermore, in China, research has been carried out on the water quality of the Yangtze River using Machine Learning [3]. The technique used is more modern because it utilizes machine learning data. Furthermore, in Banjarmasin, Indonesia, research was carried out on river water quality using K-Means Clustering [4]. Another research is by combining K-Means Clustering and Fuzzy Techniques combined with Genetic algorithms to identify groundwater quality [5].

Many factors affect the high COD and BOD. The utilization of land around the river that is used for hospitality activities will affect the quality of river water. Rivers can be polluted by wastes originating from hotels operating around the river. Waste generated from industrial activities can pollute rivers which are a source of water for daily needs and affect the development of biota in them. Meanwhile, it is stated that the roughness of the channel and the physical condition of the river have a big impact on pollutant concentrations based on the COD parameter. River water temperature also affects COD [6]. This is because the temperature will follow the movement of the flow and pollutant discharges that enter the water body by balancing the physical condition of the river which causes turbulence in the water body and has a direct impact with little effect on COD.

Factors that can affect COD and BOD are dissolved oxygen, organic matter, and other pollutant sources. Environmental parameters pH, Dissolved Oxygen (DO), and temperature have a very strong relationship and are inversely proportional to BOD5 and COD in Lake Maninjau, West Sumatra [7]. If DO is high, BOD and COD will be low. DO is the amount of dissolved oxygen in the water that comes from photosynthesis and absorption of the atmosphere/air. It is also said that BOD5, NA^{+}, T, DO, and $PO4^{-3}$ are important factors that can be relied upon to predict COD values as indicators of organic and non-organic pollution in rivers [8]. In research on the Riva River, Türkiye, COD values

* Corresponding author: m_andang@akprind.ac.id

increased due to the influence of NH4-N, TSS, and T are increased [9].

Analysis of factors affecting COD and BOD parameters is very important to evaluate river water quality. Many statistical methods can be used, one of which is multiple linear regression analysis [10-11]. The method is a type of modeling that produces a mathematical equation that shows the effect of the independent variable on the dependent [12]. This equation can also be used to predict the value of the dependent variable. The use of this method can also be used to analyze factors that influence COD [8]. The use of linear regression methods was also used to estimate the water quality index in the Yamuna River, India [13]. Many researchers use the linear regression analysis method, including using Modal Linear Regression to estimate the regression coefficient of modal linear regression [14]. Apart from that, research was also carried out on multiple linear regression [15]. Use of Multiple Linear Regression to find factors that can better predict an outcome [16]. Then, in other research, Multiple Linear Regression was used for regression experiments using random numbers which produced critical values (Fmax) which could be used to assess significance [17].

The multiple linear regression method will be applied to the analysis of river water quality in the Special Region of Yogyakarta (DIY). DIY has 10 rivers that flow in five regencies which have different qualities. The activities of the population and industry greatly influence it. Even the Department of Environment and Forestry for the Special Region of Yogyakarta (DIY) said that river water pollution is one of 17 DIY environmental issues or problems in 2021. In addition, it is also one of the three main issues that are a priority in improving environmental quality DIY with the issue of waste and land conversion that are not under spatial planning.

Many factors can affect the quality of river water in DIY. This study tested the effect of TSS and DO on BOD and COD. The method used is multiple linear regression modeling. With this analysis, it will be obtained whether TSS and DO significantly affect BOD and COD and find out what form their influence takes.

2 Method

The source of the data in this study was secondary data from the book Environmental Quality Index, by the Department of Environment and Forestry of the Special Province of Yogyakarta. From this secondary data, 149 sample points were taken for 2019 data and 210 sample points for 2020 data. In 2020, the sample points came from the Winongo River, Code River, Gadjah Wong River, Tambakbayan River, Kuning River, Konteng River, Bedog River, Belik River, Bulus River, and Oyo River.

The variables used in this study are several physical and chemical parameters of river water quality which are divided into dependent and independent variables. The independent variables are DO and TSS. While the dependent variable is BOD and COD. The data analysis steps are as follows.

1. Prepare river water quality data.
2. Determine the dependent and independent variables.
3. Exploration of data based on minimum, average, and maximum values, and their comparison with quality standards. This quality standard is based on Government Regulation of the Republic of Indonesia Number 22 of 2021 concerning the Implementation of Environmental Protection and Management.
4. Identification of the relationship between variables through the scatterplot and Pearson correlation test.
5. Regression analysis

Multiple linear regression analysis is an analysis to determine the effect of two or more independent variables on one dependent variable. The general form of the multiple linear regression model with the dependent variable (Y) and the independent variables $X_1, X_2, \ldots, Xp$ is presented as follows:

$$Y = ® + {}_1X_1 + {}_2X_2 + \square + ®_pX_p + \sum \quad (1)$$

With $\sum \sim (0, \sigma\ 2)$, β is the regression parameter, p is the number of independent variables. The matrix form in equation (2) is

$$Y = X® + \sum \quad (2)$$

This research will predict an equation model that shows the effect of TSS and DO on BOD and COD, respectively in 2019 and 2020. The form of the equation is as follows:

$$\text{BOD} = \beta_0 + {}_1\text{TSS} + {}_2\text{DO} + \varepsilon \quad (3)$$

$$\text{COD} = \beta_0 + {}_1\text{TSS} + {}_2\text{DO} + \varepsilon \quad (4)$$

The details of the modeling steps to obtain equations (3) and (4) are as follows:

1. Parameter estimation β uses the Ordinary Least Square (OLS) method, using equation (5) as follows:

$$\hat{\beta} = (X' X)^{-1} X'Y \quad (5)$$

2. Residual assumption test (identical, independent, normal distribution) using the Breusch Pagan test method, Durbin Watson Test, and Kolmogorov Smirnov
3. Test the significance of the parameters in each independent variable using the t test. the hypothesis used is

$H_0: \beta_i = 0 H_0 : \beta_p = 0$ H_0 : $®_p$ = 0 (There is no significant effect between the independent variables on the dependent variable)

H_1 : $®_p$ ¯0 (There is a significant effect between the independent variables on the dependent variable)

$$t_{test} = \frac{\hat{\beta}_p}{S(\hat{\beta}_p)} \quad (6)$$

With $S(\hat{\beta}_p) S(\hat{\beta}_p)$ is the standard error of the coefficient on the p-th observation

With a significance level (α) = 5%, conclusions are drawn by rejecting H_0 if $|t_{hitung}| > t_{\left(\frac{\alpha}{2}, n-p-1\right)}$ $|t_{test}| > t_{\left(\frac{\alpha}{2}, n-p-1\right)}$ or P value < α

4. Model interpretation and coefficient of determination.

3 Result and discussion

3.1 Data description

A description of the data used in this study is shown in Table 1. The average TSS of the 149 samples in 2019 was 35.052 mg/L. There are 45 sample points (30%) which are above the quality standard. Furthermore, in 2020, the average TSS of 210 samples was 17 mg/L, this data decreased compared to 2019. The number of sample points that were above the quality standard also decreased, namely 12 sample points (6%).

Table 1. Data Description

Characteristics	TSS (mg/L)	DO (mg/L)	BOD (mg/L)	COD (mg/L)
In 2019				
Minimum	0.800	2.390	0.100	1.390
Average	35.052	6.242	3.178	15.287
Maximum	101.800	12.440	11.560	61.034
In 2020				
Minimum	0.019	4.420	0.250	3.180
Average	17.000	7.908	4.722	23.841
Maximum	147.000	12.130	75.080	243.890
Quality Standard	50	4	3	25

DO conditions have also been good where the average in 2019 and 2020 has been above the quality standard. However, 2020 was better than 2019. Meanwhile, the BOD and COD conditions in 2019 and 2020 were still not good because there were still sample points that were above the quality standard. In 2019, there were 78 sample points (53%) that had a BOD above the quality standard and there were 15 sample points (10%) that had COD above the quality standard. In 2020, there are 102 sample points (49%) that have a BOD above the quality standard and there are 61 sample points (23%) that have a COD above the quality standard.

3.2 Relationship patterns

The initial stage before getting the modeling is to identify the pattern of relationships between the variables TSS, DO, BOD, and COD. This relationship pattern is identified through the scatterplot between variables in Fig. 1 and Fig. 2, as well as the Pearson correlation test in Table 2. Fig. 1 shows the data relationship pattern in 2019. It is known that TSS has a relationship that is comparable to BOD and COD, the higher the TSS number, the BOD and COD will be high. TSS which shows suspended solids in the water, has an impact on the decrease in natural dissolved oxygen in water so that the BOD and COD values are high. Meanwhile, the DO number has the opposite relationship with BOD and COD, namely the higher the DO number, the lower the BOD and COD. The greater the DO value in water, the higher the amount of dissolved oxygen and the water has good quality. This has an impact on low BOD and COD. Based on the correlation test in Table 2, the relationship between the variables TSS, DO, BOD, and COD is very strong. This is shown by testing the hypothesis with a significance level of 5%. The strong relationship shows that TSS and DO have a significant effect on BOD and COD.

Fig. 2 shows the data relationship pattern in 2020. The relationship pattern is slightly different compared to 2019. From Fig. 2 it can be seen that the higher the TSS number, the lower the BOD and COD. However, this relationship is not significant, which means that TSS does not really affect BOD and COD. This is different from the pattern in 2019. The higher the DO number, the lower the BOD and COD. The greater the DO value in water, the higher the amount of dissolved oxygen and the water has good quality. This has an impact on low BOD and COD.

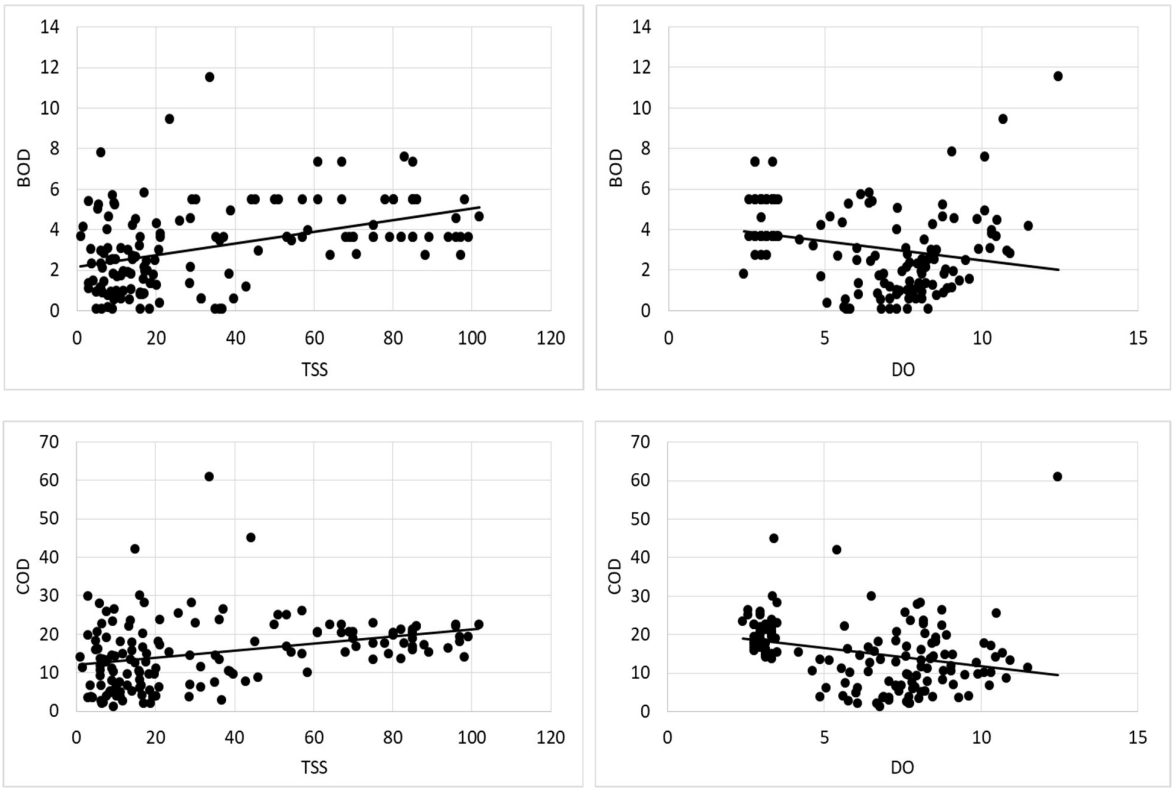

Fig. 1. Scatterplot Pattern of relationship between TSS, DO, BOD, and COD variables in 2019

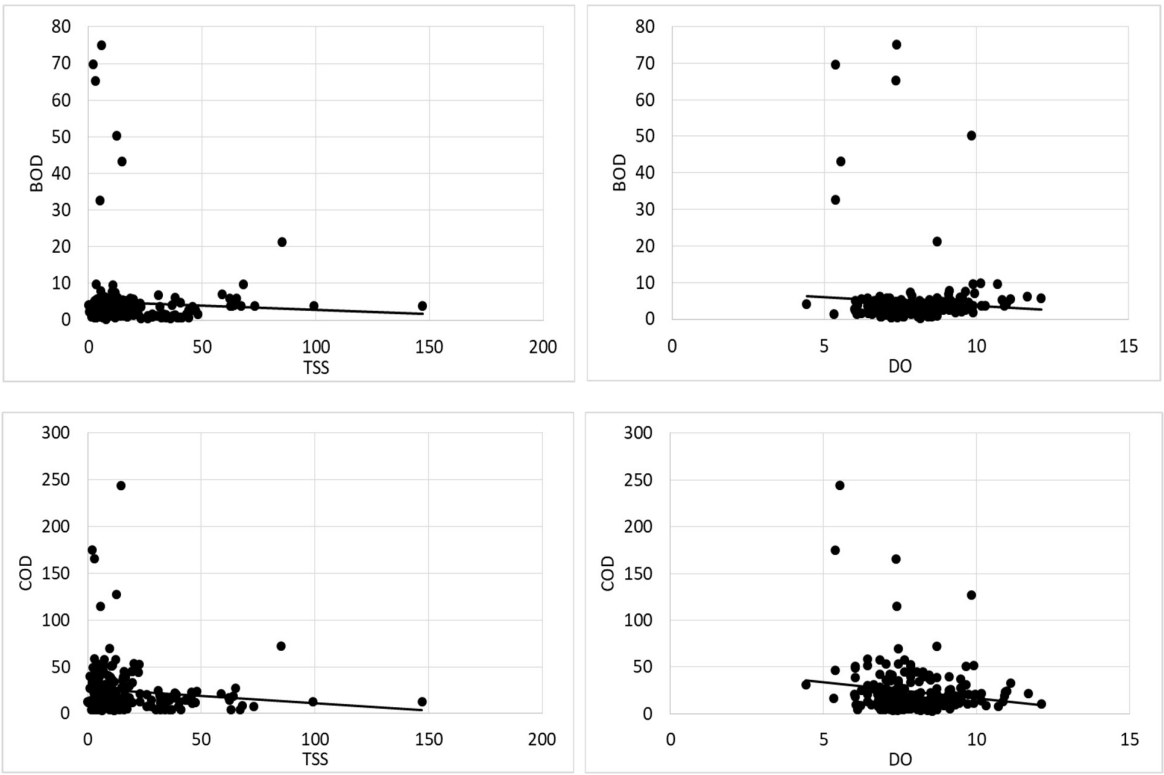

Fig. 2. Scatterplot Pattern of relationship between TSS, DO, BOD, and COD variables in 2020

Table 2. Pearson Correlation Coefficient Value

Variabel		BOD (mg/L)	COD (mg/L)
In 2019			
TSS	Pearson Correlation	0.432	0.316
	P-value	0.000*	0.000*
DO	Pearson Correlation	-0.235	-0.279
	P-value	0.004*	0.001*
In 2020			
TSS	Pearson Correlation	-0.048	-0.109
	P-value	0.218	0.115
DO	Pearson Correlation	-0.061	-0.160
	P-value	0.379	0.020*

Information: *) Significant at the 5% significance level with the null hypothesis is that there is no significant correlation between variables.

3.3 Regression analysis

Based on the identification of the relationship pattern, it can be seen that TSS and DO have a relationship with BOD and COD. To find out more detailed patterns of relationships and influences, multiple linear regression modeling was carried out. The results of model parameter estimation are presented in Table 3 and Table 4.

Table 3. Parameter Data Estimation in 2019

Variable	*Parameter Estimation*	*t-Test Statistics*	
		P-value	$\|t_{valus}\|$
Dependen BOD Variable			
Constant	1.710	0.011	2.56
TSS*	0.032	0.000	4.91
DO*	-0.054	0.009	-5.71
Dependen COD Variable			
Constant	15.700	0.000	5.25
TSS*	0.067	0.024	2.28
DO*	-0.445	0.028	-2.39

Information = *) significant at the 5% significance level with the null hypothesis is a variable that has no significant effect.

Table 4. Parameter Data Estimation in 2020

Variable	*Parameter Estimation*	*t-Test Statistics*	
		P-value	$\|t_{valus}\|$
Dependen BOD Variable			
Constant	8.546	0.046	2.01
TSS	-0.020	0.533	-0.62
DO	-0.439	0.411	-0.82
Dependen COD Variable			
Constant	52.61	0.000	4.36
TSS	-0.134	0.162	-1.41
DO*	-3.350	0.028	-2.21

Information = *) significant at the 5% significance level with the null hypothesis is a variable that has no significant effect

The resulting model equation is as follows:

In 2019:
BOD = 1.71 + 0.0321 TSS – 0.0545 DO
COD = 15.7 + 0.0669 TSS – 0.445 DO
In 2020:
BOD = 8.55 – 0.0209 TSS – 0.439 DO
COD = 52.61 – 0.134 TSS – 3.350 DO

In the 2019 data, the pattern of the relationship between TSS and BOD and COD is that if TSS increases by 1 mg/L, BOD will increase by 0.032 mg/L and COD will increase by 0.067 mg/L. Meanwhile, if DO increases by 1 mg/L, BOD will decrease by 0.054 mg/L and COD will decrease by 0.445 mg/L. Based on the significance test with a significance level of 5% it gives the result that TSS and DO really significantly affect BOD and COD. This model has a coefficient of determination of 18.9% for the BOD model and 11% for the COD model which shows that the variables TSS and DO affect BOD and COD at that rate. It can be said that there are many other factors that influence BOD and COD besides TSS and DO.
The results of regression modeling in 2020 are different from 2019. The pattern of the relationship between TSS and BOD and COD is that if TSS increases by 1 mg/L, then BOD will decrease by 0.02 mg/L and COD will decrease by 0.134 mg/L. However, this relationship is not significant. Meanwhile, if DO increases by 1 mg/L, BOD will decrease by 0.439 mg/L and COD will decrease by 3.350 mg/L. In this model only the DO variable has a significant effect on COD. This model has a coefficient of determination of 6 for the BOD model and 30% for the COD model which indicates that the TSS and DO variables affect BOD and COD at that rate. It can be said that there are many other factors that affect BOD and COD besides TSS and DO.
To find out whether the formed model meets the required assumptions, this study also tests the residual assumptions as shown in Table 5, namely normal, independent, and identical distributions.

* Corresponding author: m_andang@akprind.ac.id

Table 5. Residual Assumption Test Results

Model Name	*Test Name*	*P-Value*
	In 2019	
	Kolmogorov Smirnov	0.062
BOD	Brusch Pagan	0.121
	Durbin Watson	0.053
	Kolmogorov Smirnov	0.071
COD	Brusch Pagan	0.066
	Durbin Watson	0.031
	In 2020	
	Kolmogorov Smirnov	0.085
BOD	Brusch Pagan	0.271
	Durbin Watson	0.065
	Kolmogorov Smirnov	0.125
COD	Brusch Pagan	0.332
	Durbin Watson	0.021

Based on the hypothesis test, it can be seen that the P value in the Kolmogorov Smirnov and Breusch Pagan tests exceeds the 5% significance level. This shows that the residuals in all models have met the assumptions of normal distribution and are identical. However, the independent assumptions based on the Durbin-Watson test on the COD model do not meet the normal and identical distribution assumptions.

The prediction results for BOD and COD in 2019 and 2020 are presented in Fig. 3 and Fig. 4. Based on data in 2019, it can be predicted that if TSS increases by 1 mg/L then BOD will increase by 0.032 mg/L and if DO increases by 1 mg /L, the BOD will decrease by 0.054 mg/L. For example, if the TSS is 110 mg/L and DO is 13 mg/L, the BOD will be 5.95 mg/L. Meanwhile, it can be predicted that if TSS increases by 1 mg/L then COD will increase by 0.0699 mg/L and if DO increases by 1 mg/L then COD will decrease by 0.455 mg/L. For example, if the TSS is 110 mg/L and DO is 13 mg/L, the COD will be 17.274 mg/L.

Based on data for 2020, it can be predicted that if TSS increases by 1 mg/L then BOD will decrease by 0.0209 mg/L and if DO increases by 1 mg/L then BOD will decrease by 0.493 mg/L. For example, if the TSS is 150 mg/L and DO is 13 mg/L, the BOD will be -0.292 mg/L. Meanwhile, it can be predicted that if TSS increases by 1 mg/L then COD will decrease by 0.134 mg/L and if DO increases by 1 mg/L then COD will decrease by 3.350 mg/L. For example, if the TSS is 150 mg/L and DO is 13 mg/L, the COD will be -11.04 mg/L.

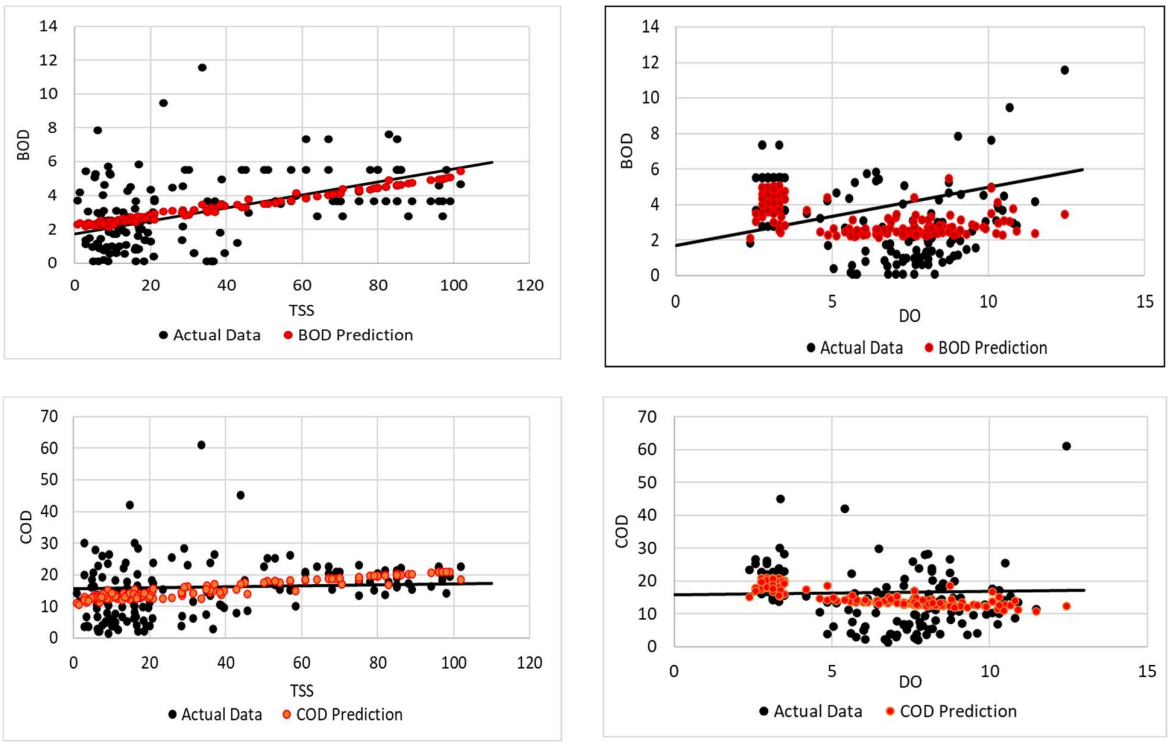

Fig. 3. BOD and COD predictions based on modeling results in 2019

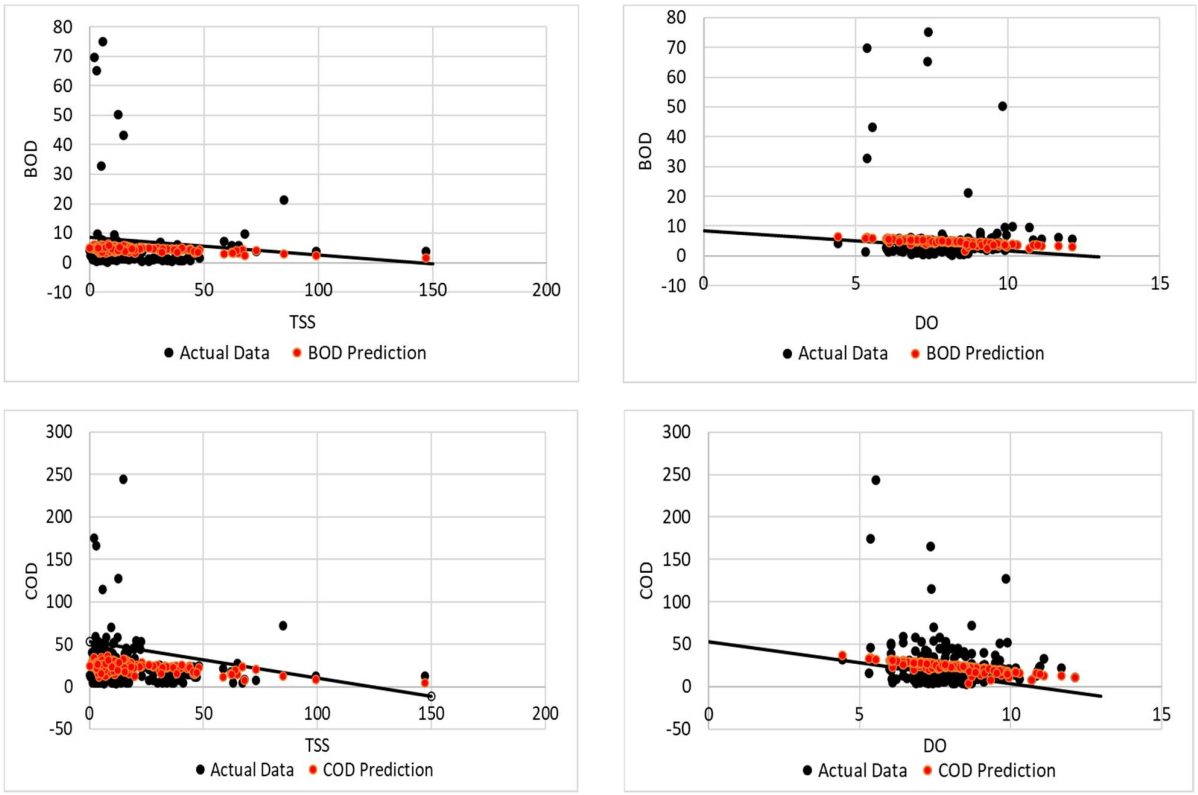

Fig. 4. BOD and COD predictions based on modeling results in 2020

4 Conclusion

The development of river water quality conditions in 2020 is higher than in 2019 based on the BOD and COD factors. This is indicated by the number of river water sample point locations which are above the quality standard. Based on multiple linear regression modeling, it can be seen that TSS and DO significantly affect BOD and COD conditions in 2019. Furthermore, in 2020 only DO significantly affected COD. The shape of the influence and the prediction is that if the TSS is high, the BOD and COD will be predicted to have high values. Conversely, if DO is high, COD and BOD will be predicted to be low.
However, the COD modeling results do not meet the assumption of independent residuals. The unfulfilled independent assumptions show that the residuals are still interconnected. This shows that the sample points are indeed related to each other. Future research can use other modeling as an alternative, for example, spatial regression or nonlinear regression.

Acknowledgement

This research was supported by the Directorate of Research and Community Service of the Ministry of Education, Culture, Research and Technology with contract No: 449A-06/UN7.D2/PP/VI/2023.

References

1. Warsito, B., et al. Evaluation of river water quality by using hierarchical clustering analysis. *IOP Conference Series: Earth and Environmental Science*. **896**(1), p. 012072, IOP Publishing, (2021)
2. Novianta, M. A., Syafrudin, and Warsito, Bo. K-Means Clustering for Grouping Rivers in DIY based on Water Quality Parameters. *JUITA: Jurnal Informatika* , **11**(1), p 155 (2023)
3. Di, Zhenzhen, Miao Chang, and Peikun Guo. Water quality evaluation of the Yangtze River in China using machine learning techniques and data monitoring on different time scales. *Water*, **11**(2), p. 339 (2019)
4. Zubaidah, Tien, Nieke Karnaningroem, and Agus Slamet. K-means method for clustering water quality status on the rivers of Banjarmasin, Indonesia. *ARPN Journal of Engineering and Applied Sciences*, **13**(6), p. 3692 (2018)
5. Mohammadrezapour, Omolbani, Ozgur Kisi, and Fariba Pourahmad. Fuzzy c-means and K-means clustering with genetic algorithm for identification of homogeneous regions of groundwater quality. *Neural Computing and Applications*, **32** (8), p.3763 (2020)
6. Marlina, Nelly, Hudori Hudori, and Ridwan Hafidh. Pengaruh Kekasaran Saluran dan Suhu Air Sungai pada Parameter Kualitas Air COD, TSS di Sungai Winongo Menggunakan Software QUAL2Kw. *Jurnal Sains & Teknologi Lingkungan* , **9** (2), p. 122 (2017)
7. Komala, P. S., A. Nur, and I. Nazhifa. Pengaruh Parameter Lingkungan Terhadap Kandungan Senyawa Organik Danau Maninjau Sumatera Barat. *Seminar Nasional Pembangunan Wilayah dan Kota Berkelanjutan*. **1**(1), p. 265 (2019)

8. Ali Abed, Salwan, Salam Hussein Ewaid, and Nadhir Al-Ansari. Evaluation of water quality in the Tigris River within Baghdad, Iraq using multivariate statistical techniques. *Journal of Physics: Conference Series*. **1294**(1), p.072025. (2019)
9. Oz, Nurtac, Bayram Topal, and Halil Ibrahim Uzun. Prediction of water quality in Riva River watershed. *Ecological Chemistry and Engineering S.* **26**(4), p.727 (2019)
10. Montgomery, D. C., Peck, E. A. and Vining, G. G. *Introduction to linear regression analysis*. John Wiley & Sons, (2021)
11. Poole, Michael A., and Patrick N. O'Farrell. The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*: **52** (1), p. 145 (1971)
12. Olive, David J., and David J. Olive. Multiple linear regression. *Springer International Publishing,* (2017)
13. Gaya, Muhammad Sani, et al. Estimation of water quality index using artificial intelligence approaches and multi-linear regression. *Int. J. Artif. Intell. ISSN* 2252 (2020)
14. Yao, Weixin, and Longhai Li. A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, **41**(3), p. 656 (2014)
15. Uyanık, Gülden Kaya, and Neşe Güler. A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences* , **106**, p. 234 (2013)
16. Pandis, Nikolaos. Multiple linear regression analysis. *American journal of orthodontics and dentofacial orthopedics* , **149** (4), p 581 (2016)
17. Livingstone, David J., and David W. Salt. Judging the significance of multiple linear regression models. *Journal of Medicinal Chemistry* , **48**(3), p 661(2005)