P. ISSN 2527-9629, E. ISSN 2548-6675
JOURNAL OF ENVIRONMENTAL ENGINEERING AND WASTE MANAGEMENT
In Collaboration With:
Managed by Environmental Engineering Undergraduate Program, School of Engineering, President University, Jl. Ki Hajar Dewantara, Kota Jababeka, Cikarang Baru, Bekasi 17550 – Indonesia.

Development of multiple linear regression model to predict cod concentration based on west tarum canal surface water quality data

Julio Putra David[1*] Rijal Hakiki[1]

[1]Department of Environmental Engineering, Faculty of Engineering, President University, Cikarang, 17550, Indonesia

Manuscript History

Received
05-02-2021
Revised
09-03-2021
Accepted
26-03-2021
Available online
26-03-2021

Keywords

multiple linear regression, predictive analysis, COD

Abstract. COD level indicates the organic matter pollution in water. COD level is normally measured using time-consuming and costly lab tests. A predictive analysis, such as Multiple Linear Regression, could be an option to make the COD measurement more effective. **Objectives:** This research aims to determine the parameter that can predict COD concentration using correlation analysis and develop a Multiple Linear Regression Model for predictive analysis on COD level in the West Tarum Canal surface water. **Method and results:** The surface water quality data used in this study are collected from the official website of PAM Jaya with a period from August 2017 to May 2020. The correlation analysis to determine the predictors is done in Microsoft Excel using the Pearson Product Moment Correlation Analysis. The predictors selected are TDS, SO_4, and Fluoride. The water quality dataset is inputted to the R Studio and made the MLR model. The model is validated using t-Test. The result showed that all models in all intake points are not showing good prediction results, and the predictors showed no effect on the COD level. **Conclusion:** The Multiple Linear Regression is not a fit tool for predicting the COD in the West Tarum Canal surface water.

* Corresponding author : julioputra0707@gmail.com

1 Introduction

The West Tarum Canal began operating in 1968 from the Curug Dam to the Ciliwung River of 69.8 km. It delivers raw water for Karawang Regency, Bekasi Regency, and Bekasi City's drinking, industrial and agricultural needs. It supplies 80% of the raw water to the citizens of Jakarta [1].

COD is defined as the number of oxygen equivalents absorbed by a strong oxidant in organic matter's chemical oxidation [2]. In the West Tarum Canal surface water, the COD level is reported exceeding the allowable level [3]. If the COD value is high, it indicates that the water is polluted by organic matter [4]. Four primary methods for measuring COD in water are included in APHA (1995): the titrimetric method, the closed reflux method, the open reflux method, and the closed reflux/colorimetric method [5].

This study would concentrate on designing statistical models to predict COD levels in order to make measuring COD levels more efficient and cost-effective. This study focuses on developing a Multiple Linear Regression model to do predictive analysis on COD level in the West Tarum Canal surface water using several predictors that will be determined later on.

2 Method

Maulani et al. (2016) had done a research about the effect of BOD, TSS, and Oil & Grease on COD level. They conducted a research using statistical tools to make an accurate prediction. They used correlation method and Multiple Linear Regression to do the prediction [6].

The first step of conducting this study was determining the idea of study and doing some literature review to strengthen the research base. After that, the author did a data collection and then pre-processed it, so the data is cleaned and ready to be analyzed. The next step is to run a correlation test of COD with all parameters to obtain the correlation coefficient used in the predictor selection process. Then the author ran a Multiple Linear Regression analysis to the predictor and response

variable to get the regression coefficient. Once the regression coefficient is obtained, it can predict the COD using the selected predictors.

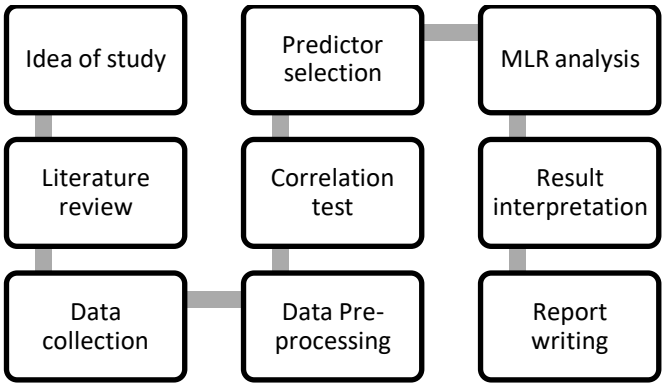

Fig. 1. Research Flow Diagram

2.1 Water Quality Data

The water quality data is a time-series data from August 2017 to May 2020, which was collected from the official website of PAM Jaya. The water quality data contains physical, chemical, and biological parameters. The intake points located in the West Tarum Canal, which are also used as the sampling point of this research, are Curug Dam, Bekasi Dam, Cawang, and Ciliwung River intake point.

2.2 Data Analysis

This research is using two software in the analysis, which is Microsoft Excel and RStudio. Microsoft Excel is used for correlation tests, while RStudio is used to carry out the regression analysis and t-Test.

2.2.1 Correlation Test

Correlation analysis is a statistical analysis used to assess the strength of the bond between two variables [7]. The correlation technique used in this research is Pearson product-moment correlation coefficient. Pearson product-moment correlation coefficient is the most widely used coefficient, the sign of which is *r* (usually called the Pearson *r*). This technique calculates the degree of linear correlation between

two variables. The coefficient of correlation varies from –1.00 to 1.00, where a value above 0 represents a positive correlation (direct correlation), and a value below 0 represents a negative correlation (inverse correlation), as can be seen in figure 2 [8].

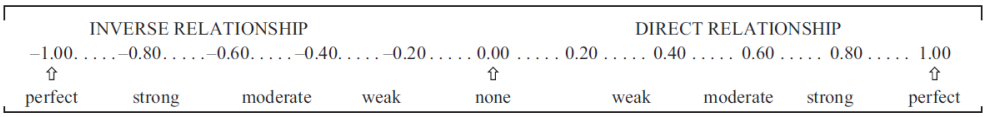

Fig. 2. Correlation Coefficient Strength Metric

Pearson product-moment correlation coefficient has also been used in the environmental field. Mustapha et al. (2012) used this technique to explain the relationship of biological and physicochemical parameters in the Jakara Basin, Nigeria.

Based on COD's characteristics, the level of COD in water is affected by organic matter. The higher the COD level, the higher the organic matter contained in water [4]. COD also has a relationship with Dissolved Oxygen, and higher COD levels mean a higher level of oxidized organic material, which decreases the levels of dissolved oxygen (DO) [8, 9].

However, in the water quality data collected from the PAM Jaya, DO data is not provided. Therefore, correlation coefficient analysis will be the tool to select COD predictors, and the chemical and physical characteristics of COD will be neglected.

2.2.2 Multiple Linear Regression Model and Predictive Analysis

Multiple Linear Regression is a statistical tool that uses several independent variables to predict a dependent variable's outcome. Multiple linear regression (MLR) is developed to model the relationship between the dependent and independent variables [11]. The independent and dependent variables are also called predictor and response variables, respectively [12]. The general equation of the Multiple Linear Regression model is shown below. Y is the response variable, X_k is the predictor variable, and β_k is the Xk regression coefficient.

$$Y = X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k + \varepsilon$$

Maulani et al. (2016) studied the influence of TSS, BOD, and Oil & Grease on COD using multiple linear regression. The study concluded that BOD and TSS significantly impacted COD with an R square of 83.7% [6].

2.3 Significance Test (t-Test)

The t-test is mostly used to test the null hypothesis of the observed difference between the two means [8]. One of the comparative tests (compare means) is the paired sample t-test. This test is useful for testing two interrelated/correlated samples or "paired samples" from populations with the same average [13]. The paired sample t-test is also called the repeated measures t-test or the related t-test. This t-test would be performed when the samples are related with commonly the same participants in each sample [14].

The t-test begins with determining the null hypothesis (H_0) and the alternative hypothesis (H_a). The null hypothesis says that there is no difference between the means, while the alternative hypothesis says that there is some difference between the means.

To test the hypothesis, a significant value is used. If the significant value is below 0.05, the null hypothesis is rejected, and the alternative hypothesis is accepted. If the significant value is above 0.05, the null hypothesis is accepted, and the alternative hypothesis is rejected.

3 Results and Discussion

3.1 The Predictors

The coefficients for Curug Dam, Bekasi Dam, Cawang, and Ciliwung River intake points are shown in Tables 1, 2, 3, and 4, respectively. The correlation coefficients are rounded to 3 decimal places.

According to the correlation test result, the $KMnO_4$, an Organic Matter, the correlation coefficient with the COD is very low. Thus the Organic Matter shall not

be selected as a predictor. Since the Dissolved Oxygen concentration is not provided in the dataset, COD characteristics will be neglected in the predictor selection.

Theoretically, $KMnO_4$, as an organic matter, is supposed to have a reasonable correlation with the COD because they are related compositionally. There might be some abnormalities in the raw data that causes this to occur.

Nonetheless, the coefficients are showing weak correlations between the parameters and the COD. Several parameters have a reasonably high coefficient value in one intake point but very low in another. Therefore, the parameters with the highest coefficient in all intake points will be selected as the predictors, TDS, SO_4, and Fluoride.

Table 1. Correlation Coefficient with COD in Curug Dam Intake Point

Parameter	Coeff	Parameter	Coeff
Color	0.003	Zn	0.000
Turbid	0.113	NH4	0.020
Temp	-0.123	KMnO4	0.039
Cond	0.152	SO4	-0.169
TDS	0.149	CN	0.000
pH	0.050	F	0.302
TH	0.162	Cr6+	0.000
Cl	0.043	Al	0.000
Mn	-0.128	TC	-0.064
Fe	#DIV/0!	FC	-0.078
N	0.144		

Table 2. Correlation Coefficient with COD in Bekasi Dam Intake Point

Parameter	Coeff	Parameter	Coeff
Color	-0.083	Zn	0.000
Turbid	-0.158	NH4	0.119
Temp	0.033	KMnO4	-0.264
Cond	-0.012	SO4	-0.093
TDS	0.120	CN	0.000
pH	-0.212	F	0.057
TH	-0.096	Cr6+	0.000
Cl	-0.049	Al	0.000
Mn	-0.030	TC	0.172
Fe	#DIV/0!	FC	0.090
N	-0.108		

Table 3. Correlation Coefficient with COD in Cawang Intake Point

Parameter	Coeff	Parameter	Coeff
Color	-0.023	Zn	0.000
Turbid	-0.062	NH4	0.124
Temp	0.081	KMnO4	-0.105
Cond	0.089	SO4	-0.137
TDS	0.120	CN	0.000
pH	0.022	F	0.280
TH	0.111	Cr6+	0.000
Cl	0.180	Al	0.000
Mn	0.117	TC	0.184
Fe	0.000	FC	0.180
N	-0.134		

Table 4. Correlation Coefficient with COD in Ciliwung River Intake Point

Parameter	Coeff	Parameter	Coeff
Color	0.201	Zn	0.000
Turbid	-0.083	NH4	-0.066
Temp	-0.011	KMnO4	0.052
Cond	0.136	SO4	0.319
TDS	0.239	CN	0.000
pH	-0.153	F	0.312
TH	0.133	Cr6+	0.002
Cl	0.142	Al	-0.072
Mn	0.151	TC	0.100
Fe	0.124	FC	-0.072
N	0.182		

3.2 MLR Model and Prediction Results

The predicted COD concentration will be obtained using the regression equation generated by the MLR models computed using R Studio. The package for MLR in R is `tidyverse`. Table 5 below shows the regression equations of Curug Dam, Bekasi Dam, Cawang, and Ciliwung River intake points. A

Table 5. Regression Equations

Intake Point	Regression Equation
Curug Dam	$\mathrm{COD} = (-3.38) + (0.076 \times \mathrm{TDS}) + (-0.046 \times \mathrm{SO_4}) + (54.266 \times \mathrm{F})$
Bekasi Dam	$\mathrm{COD} = 4.728 + (0.077 \times \mathrm{TDS}) + (-0.047 \times \mathrm{SO_4}) + (19.639 \times \mathrm{F})$
Cawang	$\mathrm{COD} = (1.842) + (0.061 \times \mathrm{TDS}) + (-0.057 \times \mathrm{SO_4}) + (70.629 \times \mathrm{F})$
Ciliwung River	$\mathrm{COD} = 9.153 + (0.032 \times \mathrm{TDS}) + (0.222 \times \mathrm{SO_4})$

In the Ciliwung River intake point, the equation excluded the Fluoride, and this happened because the significance value (1-tailed) of Fluoride is 0.

After the regression equations are generated, the next step is to substitute the predictor variables in the equation with the dataset's value. Figures 3, 4, 5, and 6 show the comparison chart between the actual COD concentration and the predicted COD concentration in Curug Dam, Bekasi Dam, Cawang, and Ciliwung River intake points, respectively.

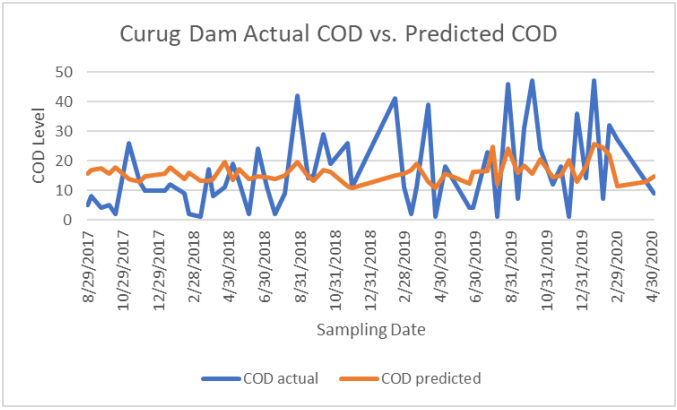

Fig. 3. Graph of COD Actual Value vs. Predicted Value in Curug Dam Intake Point

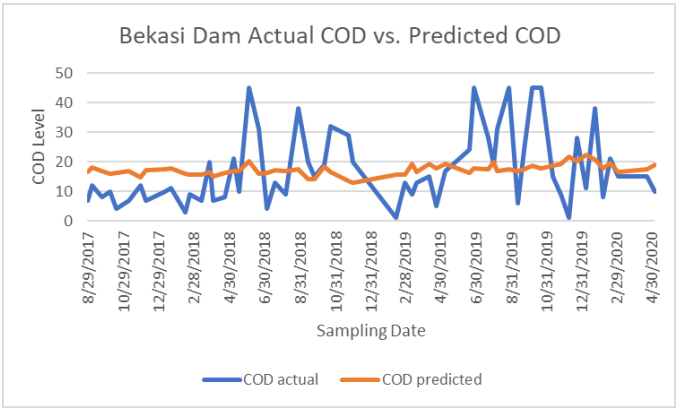

Fig. 4. Graph of COD Actual Value vs. Predicted Value in Bekasi Dam Intake Point

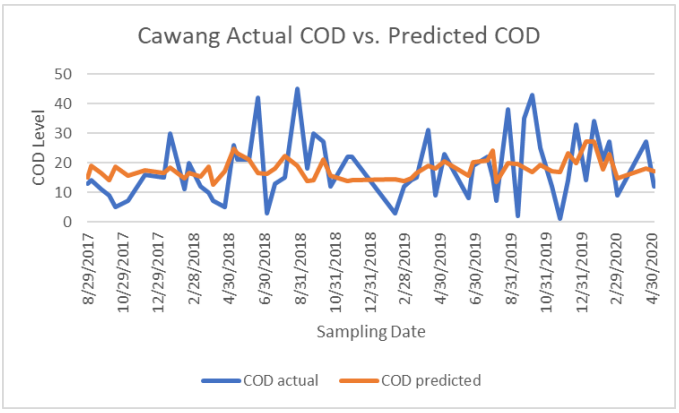

Fig. 5. Graph of COD Actual Value vs. Predicted Value in Cawang Intake Point

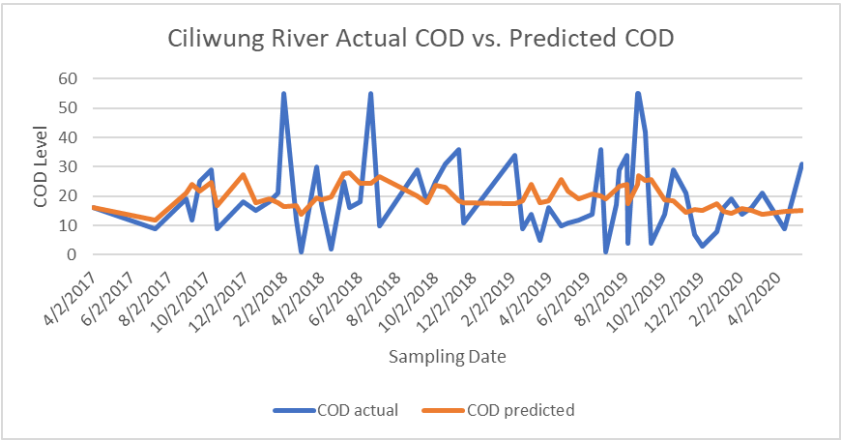

Fig. 6. Graph of COD Actual Value vs. Predicted Value in Ciliwung River Intake Point

The results of the Multiple Linear Regression model are not showing good predictions of COD concentration. The graphs above show different fluctuations between the actual and predicted value of COD concentrations. It is because the correlation between the predictors and COD is weak. Therefore the predictors do not have a significant effect on COD value. The results above show that the MLR modeling on West Tarum Canal surface water quality data is not reliable to be the predictive analytic approach in determining the COD concentration.

One of the reasons why the prediction did not show a good result is because the water quality parameters contained in the raw data were not measured consistently over standard timeframes. Several months show the measurement results twice in the raw data, but in some other months, the measurement is only reported once. There were even some months where the measurement data is not available. Therefore, it is likely that the measured pollutant matrix will also be different.

3.3 MLR Model Validations using t-Test

The means of actual COD and predicted COD concentrations at all intake points are shown in table 6. Those values were calculated using Microsoft Excel by using the average formula.

Table 6. Means of COD Concentration

CURUG DAM		BEKASI DAM		CAWANG		CILIWUNG RIVER	
Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
15.9138	15.9455	17.3103	17.3103	18.0175	18.0175	19.8667	19.8667

From the table, descriptively, there are no differences between the actual and predicted value at all sampling points. To statistically prove the hypothesis, the t value will be noticed. The R package for t-Test is `ggpubr`. Table 7 shows the t value obtained from the t-test.

Table 7. Comparison of t Count and t Distribution Table of COD Concentrations

Intake Point	df	t Count	t Distribution Table
Curug Dam	57	-0.019474	2.00247
Bekasi Dam	57	-2.1555e-10	2.00247
Cawang	56	3.8687e-10	2.00324
Ciliwung River	59	-3.0168e-10	2.00100

The hypothesis can be tested by comparing the t count in table 7 with the t value in the t distribution table under these conditions:

1. t is greater than or equal to t table, then H_a is accepted and H_0 is rejected
2. t count is less or equal to t table then H_0 is accepted, and H_a is rejected

Table 7 shows that the t Counts in all intake points are less than the t Distribution table value. Therefore, H_a is rejected, and H_0 is accepted, which means there are no differences between them. Since there are no differences between the means, it indicates no effects of the predictor variables on COD concentration as the response variable. Thus, the prediction is not reliable.

4 Conclusions

Based on the study's result, the parameters that are selected to be the predictor of COD level are Total Dissolved Solid (TDS), Sulfate (SO_4), and Fluoride (F). Those parameters are selected using the Pearson Correlation method, individual characteristics are neglected.

The MLR models are not showing a good prediction on COD concentrations of the West Tarum Canal surface water. In plain view, the differences between the actual and predicted value are quite large. The t-Test result is also showing that the predictors are not giving any effect to the COD level. Thus, the Multiple Linear Regression model is not a fit tool for predicting COD level in the West Tarum Canal surface water.

6 References

[1] M. Sumiarsih, D. Legono, and R. Kodoatie, "Rehabilitation of West Tarum Canal Related with Spatial Jabodetabek Region." 2015.

[2] Dhanjai, A. Sinha, H. Zhao, J. Chen, and S. M. Mugo, "Water analysis | determination of chemical oxygen demand," in *Encyclopedia of Analytical Science*, 2019.

[3] PAM Jaya, "Sumber Air Baku." http://pamjaya.co.id/id/service-info/raw-water/raw-water-source.

[4] D. Li and S. Liu, *Water quality monitoring and management: Basis, technology and case studies*. 2018.

[5] EPA, "Methods for Collection, Storage and Manipulation of Sediments for Chemical and Toxicological Analyses: Technical Manual Acknowledgments," *EPA 823-B-01-002. U.S. Environ. Prot. Agency, Off. Water, Washington, DC.*, 2001.

[6] D. Maulani and E. Widodo, "Analisis Pengaruh BOD, TSS dan Minyak Lemak Terhadap COD dengan Pendekatan Regresi Linear Berganda PT. X di Tangerang." 2016.

[7] M. Franzese and A. Iuliano, "Correlation analysis," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, pp. 706–721, 2018, doi: 10.1016/B978-0-12-809633-8.20358-0.

[8] M. L. Patten and M. Newhart, *Understanding research methods: An overview of the essentials, tenth edition*. 2017.

[9] K. Arina, "ANALISIS TINGKAT CHEMICAL OXYGEN DEMAND (COD), BIOCHEMICAL OXYGEN DEMAND (BOD), DAN TOTAL DISSOLVE SOLID (TDS) AIR LAUT DI PERAIRAN TELUK LAMPUNG," Universitas Negeri Lampung, 2015.

[10] Real Tech Inc., "CHEMICAL OXYGEN DEMAND (COD)," 2017. https://realtechwater.com/parameters/chemical-oxygen-demand/.

[11] C. D. Lewis, *Industrial and business forecasting methods : a practical guide to exponential smoothing and curve fitting*. Butterworths Scientific, 1982.

[12] N. J. Horton and K. Kleinman, *Using R and rstudio for data management, statistical analysis, and graphics, second edition*. 2015.

[13] I. Machali, *Statistik itu Mudah Menggunakan SPSS sebagai Alat Bantu Statistik*. Yogyakarta: Lembaga Ladang Kata, 2015.

[14] P. Hinton, *SPSS Explained*. 2014.