Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
4(2): 125–137.
DOI: 10.3934/aci.2024008
Received: 21 May 2024
Revised: 15 October 2024
Accepted: 16 October 2024
Published: 22 October 2024

*Research article*

# Using multiple linear regression for biochemical oxygen demand prediction in water

**Isaiah Kiprono Mutai[1,2], Kristof Van Laerhoven[3], Nancy Wangechi Karuri[1] and Robert Kimutai Tewo[1,*]**

[1] Department of Chemical Engineering, Dedan Kimathi University of Technology, Private bag 10143, Dedan Kimathi, Nyeri, Kenya

[2] Department of Mechanical Engineering, Dedan Kimathi University of Technology, Private bag 10143, Dedan Kimathi, Nyeri, Kenya

[3] Department of Ubiquitous Computing, University of Siegen, H-A 8110, Holderlin Str., Siegen, 57076, Germany

* **Correspondence:** Email: robert.tewo@dkut.ac.ke; Tel: +254723484716.

Academic Editor: Azlan Ismail

**Abstract:** Biochemical oxygen demand (BOD) is an important water quality measurement but takes five days or more to obtain. This may result in delays in taking corrective action in water treatment. Our goal was to develop a BOD predictive model that uses other water quality measurements that are quicker than BOD to obtain; namely pH, temperature, nitrogen, conductivity, dissolved oxygen, fecal coliform, and total coliform. Principal component analysis showed that the data spread was in the direction of the BOD eigenvector. The vectors for pH, temperature, and fecal coliform contributed the greatest to data variation, and dissolved oxygen negatively correlated to BOD. K-means clustering suggested three clusters, and t-distributed stochastic neighbor embedding showed that BOD had a strong influence on variation in the data. Pearson correlation coefficients indicated that the strongest positive correlations were between BOD, and fecal and total coliform, as well as nitrogen. The largest negative correlation was between dissolved oxygen, and BOD. Multiple linear regression (MLR) using fecal, and total coliform, dissolved oxygen, and nitrogen to predict BOD, and training/test data of 80%/20% and 90%/10% had performance indices of RMSE=2.21 mg/L, $r$=0.48 and accuracy of 50.1%, and RMSE=2.18 mg/L, $r$=0.54 and an accuracy of 55.5%, respectively. BOD prediction was better than previous MLR models. Increasing the percentage of the training set above 80% improved the model accuracy but did not significantly impact its prediction. Thus, MLR can be used successfully to estimate BOD in water using other water quality measurements that are quicker to obtain.

**Keywords:** machine learning; BOD; multiple linear regression; water treatment; contamination

## 1. Introduction

Water treatment plants (WTPs) provide safe and portable water [1]. These facilities treat the water to the appropriate standards before it can be released to the public. They also treat wastewater before it can be discharged to the surface water bodies including rivers, lakes, dams, and canals [2–4]. Studies have indicated effects of contamination from the wastewater discharged to the water bodies even after treatment [5]. Thus, monitoring surface water quality can help determine any danger of contamination.

Biochemical oxygen demand (BOD) is the amount of dissolved oxygen necessary for the microorganisms to break down organic matter in water [6]. BOD measurement is one way in which organic matter contamination is measured in WTPs. It is used as one of the water quality indices [7]. $BOD_5$ is the standard BOD measurement and takes five days to produce a result. This time lag may result in a delay in any corrective action necessary [8]. Machine learning (ML) is a fast-growing technology, expected to be one of the most transformative technologies of the 21st century [9]; therefore, its application in the field of water treatment is of great significance. Large volumes of datasets are generated in WTPs, but there is low utilization of the collected data due the lack of data science background for water treatment professionals, and the complexity of datasets generated [10]. The challenges associated with the intrinsic complexity of the processes in WTPs can be overcome through modeling with machine learning methods [11]. ML can predict $BOD_5$ based on the input of other water quality parameters that can be determined in a shorter time and provide predictive results in less than 5 days. Multiple linear regression (MLR) is a ML method that can be used for water quality parameter prediction [12]. There is an opportunity for applying MLR to model BOD prediction using water variables collected regularly from the surface water sources, which receive effluents from WTPs.

There are studies on applying MLR to predict water quality properties. Obasi et al. [13] used MLR and adaptive neuro-fuzzy inference system (ANFIS) to predict the performance of WTP in terms of conductivity, pH, iron content, BOD, chemical oxygen demand (COD), and total dissolved solids in wastewater. They found out that an MLR model is more accurate in determining WTP performance compared to ANFIS. Nourani et al. [14] carried out a study on WTP performance using MLR, ANFIS, feed forward neural network and support vector machine to predict the performance of a WTP in terms of effluent BOD, COD, and total nitrogen using different models of input and outputs. For the MLR model, they found out that a model with five inputs and one output was the best in predicting BOD and COD. Similarly, El Hammoudani and Dimane [15], successfully used MLR to characterize the elimination of micropollutants in a WTP. On the other hand, Rahmat et al. [16], used MLR to predict the water quality index using influent BOD, influent COD, and effluent COD. They found out that MLR model performed well and could be used to predict water quality index for WTPs. Abyaneh [17], used MLR to predict BOD using four water quality parameters and had promising results. Typically, the number of water quality parameters collected in a WTP may go up to 20 [18], and it is possible that a model with judicious choice of inputs from the array of parameters will have better prediction. There is a need for an MLR model that can predict BOD from the diverse water quality data and that considers the significance of the multiple water quality parameters.

The goal of this work was to develop a MLR model for BOD prediction that takes into account the significance of the multiple water quality parameters. This goal was addressed using principal component analysis (PCA) and Pearson correlation. Pearson correlation was used to identify the correlation within a dataset containing eight water quality parameters. T-distributed stochastic neighbor

embedding (t-SNE) and K-means clustering were used to identify the effect of BOD on data clustering. Dissolved oxygen, nitrogen, fecal coliform, and total coliform had strong correlations to BOD and were used to develop the MLR model. A comparison of the model results to the observed data showed that the model had better performance than some of the existing MLR models of BOD.

## 2. Materials and methods

### *2.1. Experimental data*

The data used in this work was obtained from the Kaggle database, which is freely accessible, and consists of water quality parameters for over 100 rivers and five lakes in the Indian states of Andhra Pradesh, Assam, Bihar, Goa, Gujarat, Himachal Pradesh, Jammu, Karnataka, Madhya Pradesh, Maharashtra, Punjab, Rajasthan, Tamil Nadu, Uttar Pradesh, West Bengal, Uttarakhand as well as the Union Territory of Delhi [19]. The database provides eight time averaged water quality parameters for each river and lake at multiple sample points. The water quality parameters provided in the dataset are temperature, dissolved oxygen, pH, conductivity, BOD, nitrogen, fecal coliform, and total coliform. Nitrogen was in the form of both nitrate and nitrite.

### *2.2. Principal component analysis (PCA) and data clustering analysis*

Data preprocessing was first done by dropping off all non-numeric data and accounting for the missing data values by imputing them with column feature mean values. After the preprocessing, PCA was carried out via Python code in Google Colaboratory. PCA was also used to detect the presence of data clusters in a plot of the first two principal components. A biplot was also generated from the PCA data and used to identify correlations to BOD. K-means clustering was used to determine the number of data clusters present in the data set. T-distributed stochastic neighbor embedding (t-SNE) is a powerful tool for identifying data clusters [20] and was used to identify clusters in the data based on BOD.

### *2.3. Parameter selection*

The selection of the parameters used in the ML was done through the analysis of the strength of association of each parameter to BOD, which was the target variable. This was achieved through the analysis of Pearson correlation of each independent variable to BOD in addition to the findings obtained from PCA analysis. The parameters that showed strong correlation were chosen for use in the ML model.

### *2.4. Machine learning*

Machine learning was done with multiple linear regression through a linear model imported from the scikit-learn library [21]. The training was done with both 80%/20% and 90%/10% training/test data split. The data splitting was done randomly, using the train_test_split function, from the sklearn.model_selection module [22], in the scikit-learn library [21]. The general form of the MLR model was

$$y = C_1x_1 + C_2x_2 + C_3x_3 + C_4x_4... + C_nx_n + \beta_0, \tag{2.1}$$

where y is the dependent variable, $\beta_0$ is the y-intercept, $C_1, C_2, C_3....C_n$ are the coefficients for the independent variables, and $x_1, x_2, x_3...x_n$ are the independent variables of the model. Most of the data (90.4% in total) had values of BOD between 0 and 11mg/L. The model was developed using this range of values to eliminate the effect of outliers in the model.

### *2.5. Evaluation criteria*

The evaluation criteria adopted in this study for evaluation of the model performance was the coefficient of correlation ($r$), the root mean squared error (RMSE), and accuracy. The coefficient of correlation ($r$) is a key common criterion for checking the goodness of the line of best fit [17]. It checks the fitness of the regression model to the data rather than the capability of the model in prediction. A well-fitting model results in predictions being close to the observed values. Nonetheless, it is worth noting that the coefficient of correlation does not work well for all data and hence cannot be relied on as the only measure of performance of the prediction model [23]. RMSE on the other hand indicates the absolute fitness of the model to the data and has the advantage of being expressed in the same units as the response variable. It is the best criterion for a fit when the main reason for the model is prediction. When RMSE is low and $r$ is high, the model is considered good [24].

Accuracy was used because it gives a glimpse of the performance of a model. It is however not a dependable option for the evaluation of the model performance [25]. Algorithms with lower accuracy could be preferred over those with higher accuracy upon consideration of the other factors of performance [26]. The average prediction accuracy for a model to be acceptable is 50% [27].

## 3. Results and discussion

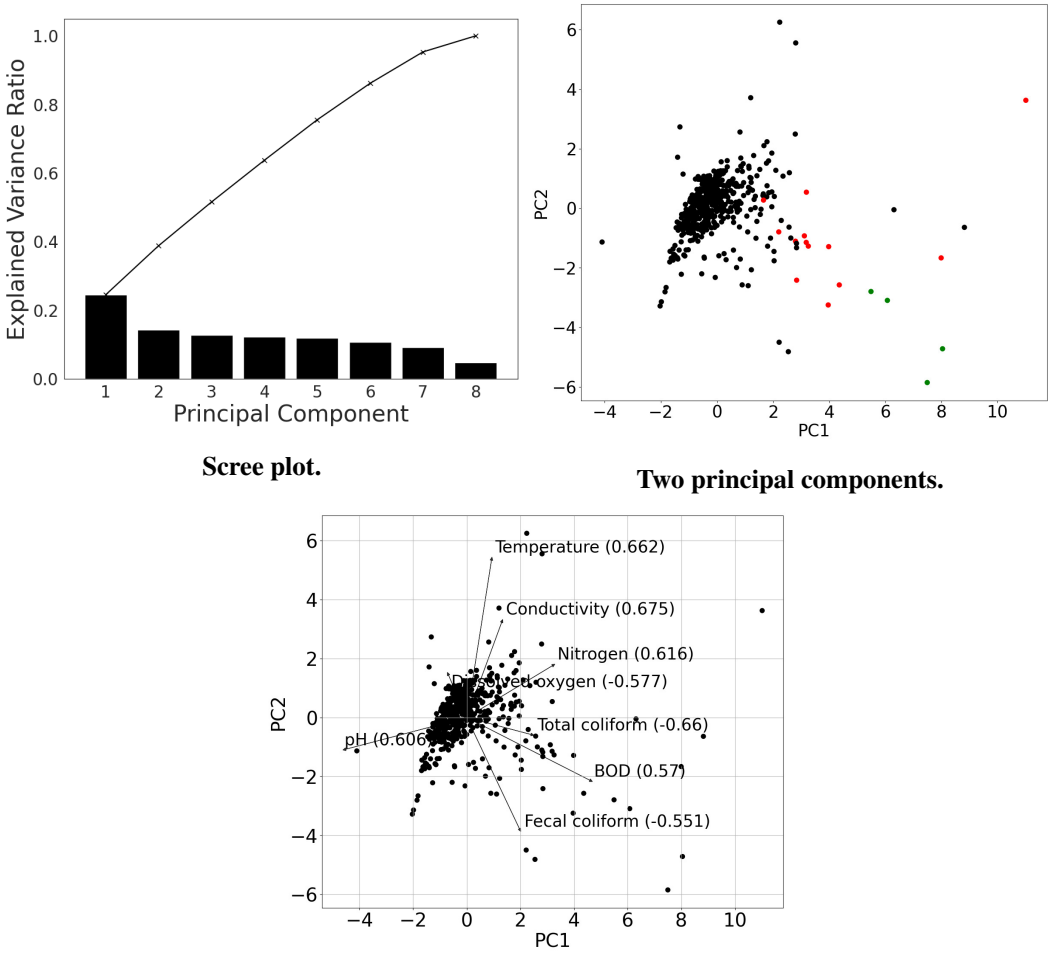
A ML algorithm was developed by first checking the consistency of the training data for water quality parameters available in the dataset used. The size of the dataset is 534 samples and each sample contain eight possible water quality parameters. A summary of the data obtained in its original form is shown in Table 1. As seen in Table 1, some measurements are missing. Of the eight water quality parameters available, there is only one case where more than 10% of the data is missing. The highest percentage of the missing measurements is found in fecal coliform and is 15.40% or 82 out of 534 measurements. The missing values were imputed with the mean value for each parameter before training the model.

**Table 1.** The structure of the dataset before imputation [19].

| Water quality parameters | Available measurements | Missing measurements |
|---|---|---|
| Temperature | 529 | 0.94 % |
| Dissolved oxygen | 532 | 0.37 % |
| pH | 534 | 0.00 % |
| Conductivity | 504 | 5.62 % |
| BOD | 528 | 1.12 % |
| Nitrogen | 532 | 0.37 % |
| Fecal coliform | 452 | 15.40 % |
| Total coliform | 495 | 7.30 % |

PCA was carried out for dimension reduction, to relate the different parameters to the level of variation and identify correlations to BOD. Figure 1 shows the the results of data visualization by PCA. The first two principal components, as shown in the scree plot, accounts for 38.8% of the variation in the data. The algorithm used to carry out PCA in Figure 1 identified three clusters in the data, in accordance with the plot for the first two principal components. However, there is no clear separation between the clusters. This may indicate that the parameters may be closely associated. The overall variation of the data is oriented towards a direction tilted to the right lower corner of the plot, away from the point of the major centroid.

The size of the vectors from the biplot of PC1 and PC2 indicate that BOD, pH, temperature and fecal coliform contribute the greatest in terms of variation in the data [28]. The orientation of the variation of data identified in the plot for the first two principal components is in the direction of BOD loading. The biplot further suggests the existence of a positive correlation between BOD, total coliform, nitrogen and fecal coliform. The correlation of pH, conductivity and temperature with BOD is weak. Dissolved oxygen shows a strong negative correlation to BOD. PCA suggests clustering in the data and correlations between BOD, dissolved oxygen, nitrogen, fecal coliform and total coliform.




**Scree plot.**

**Two principal components.**

**Biplot of PC1 and PC2.**

**Figure 1.** PCA analysis of the water data showing the scree plot, a plot of the first two principal components, and a biplot of the first and second principal components.

To explore existence of any unique information in the data that would inform on the modeling approach for BOD prediction, K-means clustering was applied to the data. K-means is a most widely used algorithm, applicable to clustering for numeric data [29]. Figure 2 shows the K-means results. According to elbow rule [30], we should select the number of clusters using the visual plot at the point in which the curve bends most. Following this, we select 3 clusters. This agreed with the PCA findings in Figure 1.

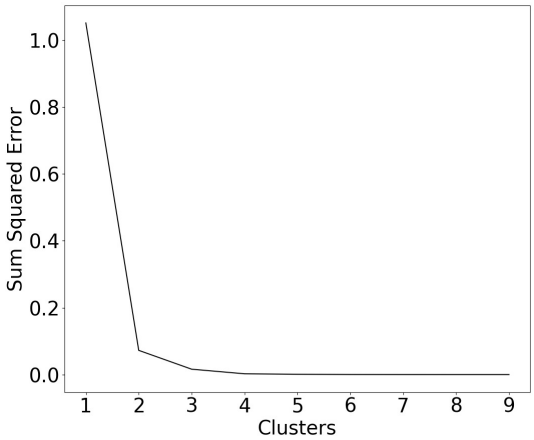
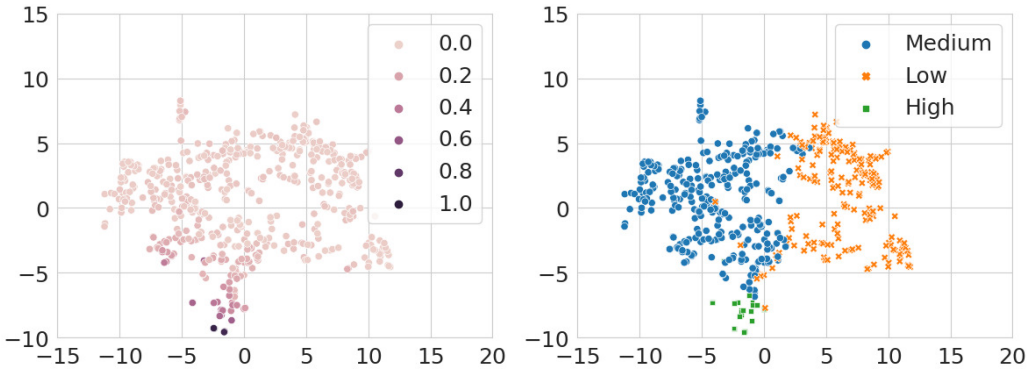



**Figure 2.** K-Means clustering of the data. Three clusters selected in the data structure in line with the elbow rule.

Further cluster analysis was carried out with t-SNE. The target variable for the t-SNE analysis was BOD because the orientation of the variation of the data was in the direction of BOD vector and BOD had significant contribution to the variation in the data (Figure 1). Figure 3 shows the results of t-SNE analysis on normalized BOD. From Figure 3, it can be seen that the clusters are easily identified on the basis of levels of BOD. Low values of BOD are in the right half of the plot. From the plot of normalized BOD data, we can identify three broad normalized BOD range of values where most of the data values lie. These are 0–0.3, 0.3–0.7 and 0.7–1.0. The data in the 3 levels of normalized BOD data plot is represented in terms of high, medium and low BOD values. There is clustering of data based on normalized BOD values and three clusters are easily identified. This supports the data obtained in Figures 1 and 2. Clearly, BOD has a significant impact in clustering of the water quality parameters and a significant effect on the variation in water quality data. This makes it an ideal target for a regression model.

Pearson correlation analysis was carried out to determine the strength of correlations between BOD with temperature, pH, dissolved oxygen, conductivity, nitrogen, fecal coliform, and total coliform. Table 2, summarizes the results of Pearson correlation analysis among all the variables in the dataset. The highest positive correlation coefficients obtained were between BOD and fecal coliform, nitrogen and total coliform indicating a positive correlation between these water quality parameters with BOD. The largest negative correlation coefficient was between BOD and dissolved oxygen, indicating a negative correlation between these two variables. These findings are in line with the findings from PCA.




**Normalized BOD data.** **3 levels of normalized BOD data.**

**Figure 3.** t-Distributed Stochastic Neighbor Embedding using normalized BOD values as the target. The transformation on the right is based on low (0–0.3), medium (0.3–0.7), and high (0.7–1.0) normalized BOD levels.

**Table 2.** Summary of the Pearson correlation coefficients for the variables in the dataset [19].

| | Temperature | pH | Dissolved oxygen | Conductivity | BOD | Nitrogen | Fecal coliform | Total coliform |
|---|---|---|---|---|---|---|---|---|
| Temperature | 1.000 | | | | | | | |
| pH | 0.018 | 1.000 | | | | | | |
| Dissolved oxygen | -0.185 | 0.066 | 1.000 | | | | | |
| Conductivity | 0.074 | 0.012 | -0.105 | 1.000 | | | | |
| BOD | -0.071 | -0.056 | **-0.522** | 0.099 | 1.000 | | | |
| Nitrogen | 0.089 | -0.019 | -0.269 | 0.084 | **0.285** | 1.000 | | |
| Fecal coliform | 0.004 | -0.013 | -0.080 | -0.001 | **0.299** | 0.018 | 1.000 | |
| Total coliform | -0.003 | -0.030 | -0.230 | 0.001 | **0.174** | 0.131 | 0.036 | 1.000 |

The data in Table 2 was further summarized with respect to the BOD and all the other water quality parameters. Table 3 shows the summary with respect to the strength of the correlations. From the results in Table 3, it is clear that three of the parameters had a medium to strong level of correlation with BOD. These are: dissolved oxygen (-0.522), nitrogen (0.285) and fecal coliform (0.299). Next was total coliform which had correlation coefficient of 0.174 with BOD. Temperature, conductivity and pH all showed very low strength of correlation with BOD ($\leq |0.1|$). The findings by Abyaneh, on the other hand, showed a high significance of pH in prediction of BOD [17]. This difference between our work and that of Abyaneh can be attributed to the difference in the parameters under review in this study and those selected by Abyaneh. Abyaneh used total suspended solids, temperature and pH for the study. Our study on the other hand has considered a broader set of parameters, that is dissolved oxygen, temperature, pH, conductivity, nitrogen, fecal coliform, and total coliform. We found out that dissolved oxygen had a negative correlation to BOD, which is in agreement with the work of others [6,31]. These studies show that the effects of high levels of BOD results is low dissolved oxygen concentration in water. The strength of these correlations agreed with the results of PCA analysis and informed on the principle water quality parameters to be used in the training of the linear model.

**Table 3.** The correlation coefficients of the individual parameters and the BOD. The correlation strength was classified with reference to [32] and [33].

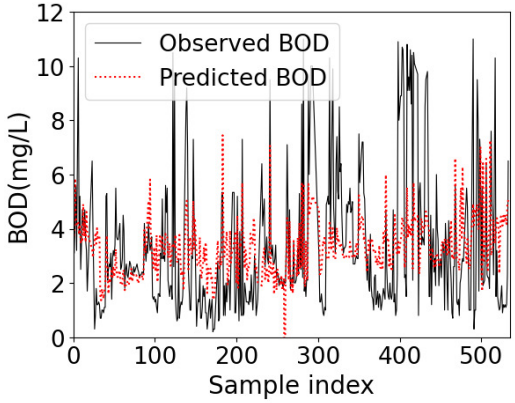
| Parameters | Temperature | pH | Dissolved oxygen | Conductivity | Nitrogen | Fecal coliform | Total coliform |
|---|---|---|---|---|---|---|---|
| Correlation | -0.071 | -0.056 | **-0.522** | 0.099 | **0.285** | **0.299** | **0.174** |
| Correlation Strength | Small Negative | Small Negative | High Negative | Small Positive | Medium Positive | Medium Positive | Small Positive |

Multiple linear regression was developed to determine the coefficients in Eq (2.1) using two sets of data split: 80%/20% and 90%/10% training to test data. Table 4 is a summary of the regression coefficients for the two sets of training/test data. The absolute values of the coefficients $C_1$–$C_4$ are indicative of the order of importance of the independent variables in their contribution in the model; with the highest absolute value implying relatively most important variable in the model [34]. From the absolute values of the coefficients $C_1$–$C_4$, it is evident that dissolved oxygen made the highest contribution to the model, followed by nitrogen, fecal coliform, and total coliform, respectively. This is in agreement with the findings of the Pearson correlation analysis as summarized in Table 3.

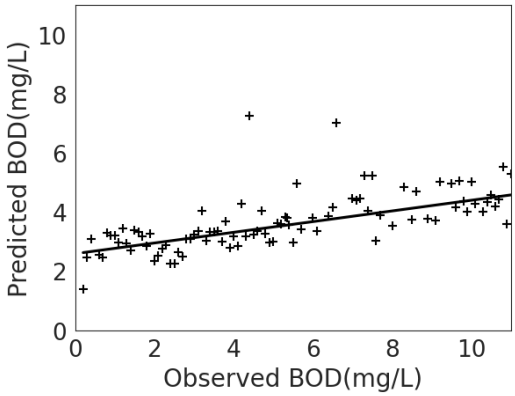
**Table 4.** Regression coefficients, $C_1$–$C_4$ and the y-intercept, $\beta_0$, for a linear model of BOD (Eq (2.1)) with dissolved oxygen ($x_1$), nitrogen( $x_2$), fecal coliform ($x_3$) and total coliform ($x_4$) respectively.

| Training/test data split ratio | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $\beta_0$ |
|---|---|---|---|---|---|
| 80%/20% | $-6.02 \times 10^{-1}$ | $2.50 \times 10^{-1}$ | $1.26 \times 10^{-5}$ | $-2.80 \times 10^{-6}$ | 6.92 |
| 90%/10% | $-6.34 \times 10^{-1}$ | $2.89 \times 10^{-1}$ | $6.76 \times 10^{-6}$ | $-1.11 \times 10^{-6}$ | 7.09 |

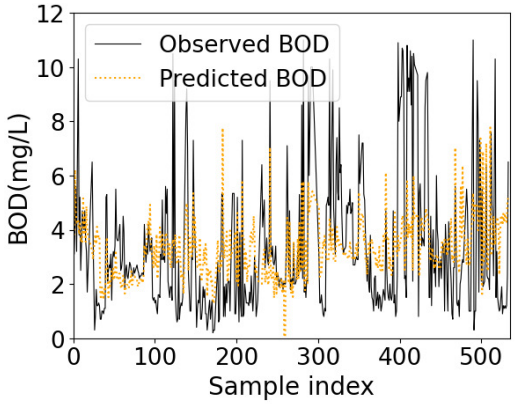
Hydrographs with scatter plots for both the 80%/20% and 90%/10% training/test data split were generated to determine how well the model agreed with the observed data, as shown in Figure 4. From the hydrograph for the 80%/20% training/test data split, it is clear that the predicted values of BOD closely mirror the observed values. The $r$ value for the 80%/20% scatter plot , which is an indicator of how well the model and the observed data agreed with each other, is 0.48. The RMSE value in this case was 2.21 mg/L. Similarly, the hydrograph for 90%/10% training/test data split shows a good agreement between the predicted and observed BOD values. The $r$ value of the predicted and observed BOD values was 0.54. The RMSE for the 90%/10% data split was 2.18 mg/L, which is a negligible change compared to that of the 80%/20% split. An increase in the training/test split ratio beyond 80%/20% does not significantly improve the prediction capability of the MLR model. Our results agree with the findings by Rácz et al. [35]. They used four split ratios of 50%, 60%, 70% and 80% and found that the 80%/20% split ratio achieved the best performance. In modeling water treatment plants, the 80%/20% data split ratio for model training/testing is a common ratio. Güçlü and Dursun [36] used approximately 80%/20% split ratio while modeling WTP using ANN. Similarly, Hamed et al. [37], obtained a good fit for BOD and suspended solids prediction with 80%/20% data split ratio.



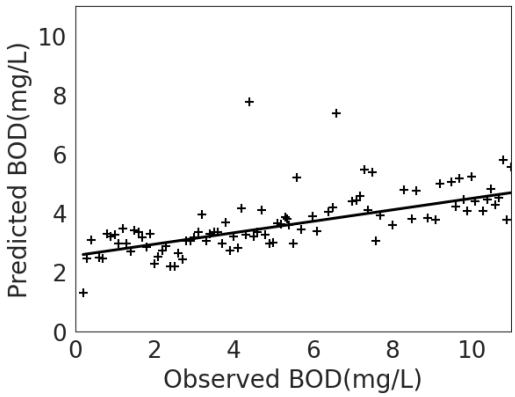

**80%/20% hydrograph.**




**80%/20% scatter plot.**




**90%/10% hydrograph.**




**90%/10% scatter plot.**

**Figure 4.** Comparison of observed and predicted BOD values for the 80%/20% and the 90%/10% training/testing data split. An accuracy of 50.1% and 55.5%, RMSE of 2.21 mg/L and 2.18 mg/L, and r values of 0.48 and 0.54 were achieved for the for the 80%/20% and 90%/10% training/test data splits respectively.

The accuracy of the two models obtained from the 80%/20% and 90%/10% training/testing data splits was assessed. The 80%/20% data split had an accuracy of 50.1% while that of 90%/10% data split had a higher accuracy of 55.5%. The model accuracy met the minimum acceptable average prediction accuracy of 50% [27]. It is evident from these results that the accuracy of a model increases by increasing the ratio of training/test set of the data.

Abyaneh [17], predicted BOD using MLR. The author found an $r$ value of 0.53 and a RMSE of 37.8 mg/L. The $r$ values achieved in this study were 0.48 and 0.54 for 80%/20% and 90%/10% split ratios, respectively. These values were in close agreement with the value obtained by Abyaneh (0.53). Lower RMSE values of 2.21 mg/L and 2.18 mg/L were obtained for the 80%/20% and 90%/10% split ratios, respectively. Notably, the lower RMSE values achieved in this investigation suggest that a judicious choice of input parameters contributes to enhanced prediction capabilities. MLR is an easy to develop and easy to implement technique for water quality prediction and can easily be deployed in surface water quality monitoring to check on any potential contamination even in low resource settings including the countries in the global south.

### *3.1. Limitations and future directions*

An understanding of the limitations of the MLR model developed in this study is crucial for refining predictive capabilities and advancing research in water treatment. Despite its simplicity, the model exhibits inherent constraints that warrant acknowledgment and consideration. The reliance on linear relationships may not have captured the complexity of BOD dynamics in real-world scenarios. Similarly, factors such as data quality and variable selection influence the model's predictive accuracy and could explain the low accuracy achieved in this study. Addressing these limitations necessitates a multifaceted approach, including data augmentation, feature engineering, and the exploration of alternative modeling techniques such as nonlinear regression [38, 39].

Additionally, we employed k-means clustering with the elbow method as an exploratory tool, however, we acknowledge the limitations of this approach in identifying natural clusters within the dataset. There is a need to conduct a more in-depth cluster analysis using alternative techniques, such as internal clustering validity indices [40], to accurately evaluate the presence and structure of potential clusters. Other algorithms such as Random Swap [41] or variants of k-means, like the M-algorithm [42], can also be employed for detailed clustering. This will ensure a more robust understanding of the underlying structure in the data and provide more insight in the modeling. The work presented here is essential for future studies on enhancing the robustness and reliability of predictive tools for BOD levels in water treatment.

## 4. Conclusions

We wanted to develop a predictive tool for BOD using measurements that were easier to obtain. We found that BOD influenced organization and clustering of the water quality data. Out of eight water quality parameters, four strongly correlated to BOD. A MLR regression model was used to develop a linear model for BOD prediction based on these four water quality parameters with two sets of training/test split ratios. In both cases, the model was able to account for over half the variation of the data. MLR is a simple model and the work herein establishes a foundation for the development of more complex models on BOD.

## Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work in this paper.

## References

1. T. Ahmad, K. Ahmad, M. Alam, Sustainable management of water treatment sludge through 3'R' concept, *J. Clean. Prod.*, **124** (2016), 1–13. http://dx.doi.org/10.1016/j.jclepro.2016.02.073
2. R. O. Carey, K. W. Migliaccio, Contribution of wastewater treatment plant effluents to nutrient dynamics in aquatic systems: a review, *Environ. Manage.*, **44** (2009), 205–217. http://dx.doi.org/10.1007/s00267-009-9309-5

3. G. Crini, E. Lichtfouse, Advantages and disadvantages of techniques used for wastewater treatment, *Environ. Chem. Lett.*, **17** (2019), 145–155. http://dx.doi.org/10.1007/s10311-018-0785-9

4. B. E. Igere, A. I. Okoh, U. U. Nwodo, Wastewater treatment plants and release: the vase of odin for emerging bacterial contaminants, resistance and determinant of environmental wellness, *Emerging Contaminants*, **6** (2020), 212–224. http://dx.doi.org/10.1016/j.emcon.2020.05.003

5. C. Holeton, P. A. Chambers, L. Grace, Wastewater release and its impacts on canadian waters, *Can. J. Fish. Aquat. Sci.*, **68** (2011), 1836–1859. http://dx.doi.org/10.1139/f2011-096

6. R. Jha, C. Ojha, K. Bhatia, Development of refined bod and do models for highly polluted kali river in india, *J. Environ. Eng.*, **133** (2007), 839–852. http://dx.doi.org/10.1061/(ASCE)0733-9372(2007)133:8(839)

7. P. Yu, J. Cao, V. Jegatheesan, X. Du, A real-time bod estimation method in wastewater treatment process based on an optimized extreme learning machine, *Appl. Sci.*, **9** (2019), 523. http://dx.doi.org/10.3390/app9030523

8. K. S. Ooi, Z. Y. Chen, P. E. Poh, J. Cui, Bod5 prediction using machine learning methods, *Water Supply*, **22** (2022), 1168–1183. http://dx.doi.org/10.2166/ws.2021.202

9. M. I. Jordan, T. M. Mitchell, Machine learning: trends, perspectives, and prospects, *Science*, **349** (2015), 255–260. http://dx.doi.org/10.1126/science.aaa8415

10. K. B. Newhart, R. W. Holloway, A. S. Hering, T. Y. Cath, Data-driven performance analyses of wastewater treatment plants: a review, *Water Research*, **157** (2019), 498–513. http://dx.doi.org/10.1016/j.watres.2019.03.030

11. D. Wang, S. Thunéll, U. Lindberg, L. Jiang, J. Trygg, M. Tysklind, et al., A machine learning framework to improve effluent quality control in wastewater treatment plants, *Sci. Total Environ.*, **784** (2021), 147138. http://dx.doi.org/10.1016/j.scitotenv.2021.147138

12. A. E. Bilali, A. Taleb, Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment, *Journal of the Saudi Society of Agricultural Sciences*, **19** (2020), 439–451. http://dx.doi.org/10.1016/j.jssas.2020.08.001

13. O. P. Okeke, I. I. Aminu, A. Rotimi, B. Najashi, M. Jibril, A. S. Ibrahim, et al., Performance analysis and control of wastewater treatment plant using adaptive neuro-fuzzy inference system (ANFIS) and multi-linear regression (MLR) techniques, *GSC Advanced Engineering and Technology*, **4** (2022), 001–016. http://dx.doi.org/10.30574/gscaet.2022.4.2.0033

14. V. Nourani, G. Elkiran, S. Abba, Wastewater treatment plant performance analysis using artificial intelligence—an ensemble approach, *Water Sci. Technol.*, **78** (2018), 2064–2076. http://dx.doi.org/10.2166/wst.2018.477

15. Y. El Hammoudani, F. Dimane, Assessing behavior and fate of micropollutants during wastewater treatment: statistical analysis, *Environ. Eng. Res.*, **26** (2021), 200359. http://dx.doi.org/10.4491/eer.2020.359

16. S. Rahmat, W. A. H. Altowayti, N. Othman, S. M. Asharuddin, F. Saeed, S. Basurra, et al., Prediction of wastewater treatment plant performance using multivariate statistical analysis: a case study of a regional sewage treatment plant in melaka, malaysia, *Water*, **14** (2022), 3297. http://dx.doi.org/10.3390/w14203297
17. H. Z. Abyaneh, Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, *J. Environ. Health Sci. Engineer.*, **12** (2014), 40. http://dx.doi.org/10.1186/2052-336X-12-40
18. K. S. Kumar, P. S. Kumar, M. J. R. Babu, Performance evaluation of waste water treatment plant, *International Journal of Engineering Science and Technology*, **2** (2010), 7785–7796.
19. *U. Agrawal, Water quality data*, Kaggle, 2020. Available from: `https://www.kaggle.com/datasets/utcarshagrawal/water-quality-data`.
20. D. Kobak, P. Berens, The art of using t-sne for single-cell transcriptomics, *Nat. Commun.*, **10** (2019), 5416. http://dx.doi.org/10.1038/s41467-019-13056-x
21. F. Pedregosa, Scikit-learn: machine learning in python fabian, *J. Mach. Learn. Res.*, **12** (2011), 2825.
22. A. Zollanvari, Supervised learning in practice: the first application using scikit-learn, In: *Machine learning with Python: theory and implementation*, Cham: Springer, 2023, 111–131. http://dx.doi.org/10.1007/978-3-031-33342-2_4
23. M. A. Razi, K. Athappilly, A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models, *Expert Syst. Appl.*, **29** (2005), 65–74. http://dx.doi.org/10.1016/j.eswa.2005.01.006
24. D. Chicco, M. J. Warrens, G. Jurman, The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation, *PeerJ Comput. Sci.*, **7** (2021), e623. http://dx.doi.org/10.7717/peerj-cs.623
25. A. Rechkemmer, M. Yin, When confidence meets accuracy: exploring the effects of multiple performance indicators on trust in machine learning models, *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, 535. http://dx.doi.org/10.1145/3491102.3501967
26. G. I. Webb, M. J. Pazzani, D. Billsus, Machine learning for user modeling, *User Model. User-Adap.*, **11** (2001), 19–29. http://dx.doi.org/10.1023/A:1011117102175
27. J. L. Lerios, M. V. Villarica, Pattern extraction of water quality prediction using machine learning algorithms of water reservoir, *Int. J. Mech. Eng. Rob. Res.*, **8** (2019), 992–997. http://dx.doi.org/10.18178/ijmerr.8.6.992-997
28. P. M. Kroonenberg, *Applied multiway data analysis*, Hoboken: John Wiley & Sons, 2008.
29. P. Fränti, R. Mariescu-Istodor, A. Akram, M. Satokangas, E. Reissell, Can we optimize locations of hospitals by minimizing the number of patients at risk? *BMC Health Serv. Res.*, **23** (2023), 415. http://dx.doi.org/10.1186/s12913-023-09375-x
30. T. M. Kodinariya, P. R. Makwana, Review on determining number of cluster in k-means clustering, *International Journal of Advance Research in Computer Science and Management Studies*, **1** (2013), 90–95.

31. E. Dogan, B. Sengorur, R. Koklu, Modeling biological oxygen demand of the melen river in turkey using an artificial neural network technique, *J. Environ. Manage.*, **90** (2009), 1229–1235. http://dx.doi.org/10.1016/j.jenvman.2008.06.004
32. P. Schober, C. Boer, L. A. Schwarte, Correlation coefficients: appropriate use and interpretation, *Anesth. Analg.*, **126** (2018), 1763–1768. http://dx.doi.org/10.1213/ANE.0000000000002864
33. W. Cui, Z. Sun, H. Ma, S. Wu, The correlation analysis of atmospheric model accuracy based on the pearson correlation criterion, *IOP Conf. Ser.: Mater. Sci. Eng.*, **780** (2020), 032045. http://dx.doi.org/10.1088/1757-899X/780/3/032045
34. G. K. Uyanık, N. Güler, A study on multiple linear regression analysis, *Procedia-Social and Behavioral Sciences*, **106** (2013), 234–240. http://dx.doi.org/10.1016/j.sbspro.2013.12.027
35. A. Rácz, D. Bajusz, K. Héberger, Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification, *Molecules*, **26** (2021), 1111. http://dx.doi.org/10.3390/molecules26041111
36. D. Güçlü, Ş. Dursun, Artificial neural network modelling of a large-scale wastewater treatment plant operation, *Bioprocess Biosyst. Eng.*, **33** (2010), 1051–1058. http://dx.doi.org/10.1007/s00449-010-0430-x
37. M. M. Hamed, M. G. Khalafallah, E. A. Hassanien, Prediction of wastewater treatment plant performance using artificial neural networks, *Environ. Modell. Softw.*, **19** (2004), 919–928. http://dx.doi.org/10.1016/j.envsoft.2003.10.005
38. S. A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, T. A. Mann, Data augmentation can improve robustness, *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2024, 29935–29948.
39. S. R. Shams, A. Jahani, S. Kalantary, M. Moeinaddini, N. Khorasani, The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting $SO_2$ concentration, *Urban Clim.*, **37** (2021), 100837. http://dx.doi.org/10.1016/j.uclim.2021.100837
40. Q. Zhao, P. Fränti, Wb-index: a sum-of-squares based index for cluster validity, *Data Knowl. Eng.*, **92** (2014), 77–89. http://dx.doi.org/10.1016/j.datak.2014.07.008
41. P. Fränti, Efficiency of random swap clustering, *J. Big Data*, **5** (2018), 13. http://dx.doi.org/10.1186/s40537-018-0122-y
42. P. Fränti, S. Sieranoja, K. Wikström, T. Laatikainen, Clustering diagnoses from 58 million patient visits in finland between 2015 and 2018, *JMIR Med. Inform.*, **10** (2022), e35422. http://dx.doi.org/10.2196/35422



AIMS Press