e-ISSN: 2689-4934

Available online at
http://uppp.uitm.edu.my/online-issues.html

ESTEEM Academic Journal Vol 21, March 2025, 21–30

ESTEEM
Academic
Journal (EAJ)

Development of regression models for predicting water quality index based on dissolved oxygen for river pollution assessment

Wan Mohamad Haziq Wan Roselan[1], Muhamad Irfan Ahmad Suhkri[1], Mohamad Faizal Abd Rahman[1], Moheddin Usodan Sumagayan[2], Mohd Suhaimi Sulaiman[1*]

[1]*Faculty of Electrical Engineering, Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia*
[2]*Mindanao State University - Iligan Institute of Technology Andres Bonifacio Ave, Iligan City, 9200 Lanao del Norte, Philippines*

ARTICLE INFO

Article history:
Received 5 February 2025
Revised 27 February 2025
Accepted 6 March 2025
Online first
Published 17 March 2025

Keywords:
Water Quality Index (WQI)
Dissolved Oxygen (DO)
Regression Modelling
Machine Learning
Water Quality Assessment
River Pollution

DOI:
10.24191/esteem.v21iMarch.5023.g3037

ABSTRACT

Water is an essential resource in Malaysia, playing a crucial role in sustaining human life, agriculture, and industry. However, rapid industrialization, urbanization, and development have significantly deteriorated river water quality, posing serious environmental and public health risks. Traditional water quality monitoring methods rely on manual sampling and laboratory analysis and are often time-consuming, labor-intensive, and inefficient. This study aims to overcome these challenges by developing regression-based predictive models to estimate the Water Quality Index (WQI) based on Dissolved Oxygen (DO) measurements. The research utilizes a dataset of 219 river water samples collected between June and November 2023 from the Kaggle database. Statistical validation techniques were applied to assess data distribution and accuracy, including normality tests and error bar plots. Multiple regression techniques were implemented using MATLAB and Python to determine the most effective model. MATLAB's Linear Regression model demonstrated superior performance among the tested approaches, achieving an R^2 value of 0.95397 and a Root Mean Square Error (RMSE) of 7.2728. These results highlight the potential of regression models in providing a fast, reliable, and cost-effective method for water quality assessment. By leveraging these predictive techniques, environmental authorities and policymakers can implement timely interventions, ensuring better management and protection of freshwater ecosystems in Malaysia.

[1*] Corresponding author. *E-mail address*: shemi@uitm.edu.my

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037

1. INTRODUCTION

Rivers in Malaysia are crucial for the nation’s freshwater supply, but the water quality has been compromised due to the increase in industrial, agricultural, and urban activities in the country. This has led to environmental degradation and pollution [1]. Subsequently, many complaints have been reported from firms, factories, and individuals to the Department of Environment (DoE) [2-3].

In light of these challenges, this project aims to construct regression models to forecast the Water Quality Index (WQI) through the lens of Dissolved Oxygen (DO) readings. WQI, as an aggregate representation of water quality, allows various stakeholders to monitor river health and assess if any pollution risks exist. This study focuses on the use of Artificial Neural Networks (ANNs) through MATLAB’s Regression Learner to construct models that predictably classify water as either clean or polluted [4].

The importance of DO as a key indicator of water quality has been well-documented. For instance, a recent study on the Hatirjheel urban lake in Dhaka highlighted that DO levels critically reflect aquatic ecosystem health. Low levels indicate severe water degradation and unsuitability for sustaining life [5]. A study on the Bhavani River demonstrated that DO is a fundamental parameter in calculating the WQI, with its levels inversely correlated with pollution indicators such as biological and chemical oxygen demand, emphasizing its role in assessing and predicting water quality effectively [6]. Moreover, machine learning approaches such as ANNs and Support Vector Machines (SVMs) have been shown to achieve high accuracy in predicting water quality parameters, allowing for early detection of pollution events [7]. Studies by Hoque et al. further emphasized that DO is a critical input feature in predictive models for WQI, where its incorporation enhances the reliability and precision of water quality assessments across diverse aquatic ecosystems [8].

Traditional water quality monitoring methods, such as manual sampling and laboratory analysis, are time-consuming and resource-intensive and often result in delayed pollution detection [9]. These limitations hinder authorities’ ability to monitor water quality and address threats before they escalate proactively.

Linear regression was chosen for this project due to its simplicity and computational efficiency. While other machine learning techniques—such as artificial neural networks and support vector machines—have shown promise, linear regression often outperforms them in simplicity and computational efficiency [10-11]. Multiple studies have demonstrated that linear regression models can achieve high prediction accuracy for water quality indices. For instance, one study reported an R^2 value of 1.0 and a low RMSE of 0.0025 when using multiple linear regression (MLR) for water quality index estimation [12]. In another analysis, linear regression provided the best estimates for pH levels among various regression models, highlighting its effectiveness in specific parameter predictions [10-13].

The Coefficient of Determination, denoted as R^2, indicates how effectively a model accounts for the variability present in a dataset, with values ranging from 0, indicating no explanatory power, to 1, signifying a perfect fit. A high R^2 value suggests that a considerable amount of the variability in the dependent variable can be attributed to the independent variable(s), thereby illustrating a robust relationship between them. The Root Mean Square Error (RMSE) provides a measure of prediction accuracy by calculating the average magnitude of errors in predictions. A lower RMSE signifies that the predicted values closely align with the actual values, indicating greater precision in the model's forecasts. While moderate RMSE values may reflect the complexities inherent in real-world data, they can still be deemed practically useful when considered alongside a high R^2. Collectively, these metrics affirm the model's dependability: R^2 highlights the strength of the relationship between the variables, whereas RMSE guarantees that the predictions are sufficiently precise for practical decision-making.

This project aims to develop reliable regression models for predicting WQI based on DO measurements. The study classifies water quality as clean or polluted by utilizing MATLAB’s Regression Learner for Artificial Neural Network (ANN) analysis, providing an accurate and data-driven approach for

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037

river pollution assessment. These models enhance the ability to monitor and evaluate water quality and support timely decision-making for environmental authorities and researchers.

This project bridges existing gaps in water quality prediction by offering a practical and effective tool for assessing river health. By applying regression models, the study contributes to environmental management by enabling accurate water quality predictions, thereby aiding in the proactive management of freshwater resources.

2. METHODOLOGY

This project involves a structured methodology, as shown in Fig. 1. The flowchart of the method begins with data collection and the categorization of Dissolved Oxygen (DO) and the Water Quality Index (WQI) obtained from Kaggle. Statistical analysis is conducted to distinguish between polluted and clean water conditions, ensuring data suitability for model training. Upon approval, linear regression models were developed using DO data from clean and contaminated water samples to predict the WQI, with MATLAB R2024a and Python employed for model implementation. Due to the straightforward nature of the linear regression model, which can easily elucidate the relationship between an independent variable (e.g., DO) and the dependent variable (WQI), this method was selected [14]. Multiple models were evaluated, and only those meeting accuracy criteria proceeded to the final stage, ensuring the optimal model was identified for accurate WQI prediction.

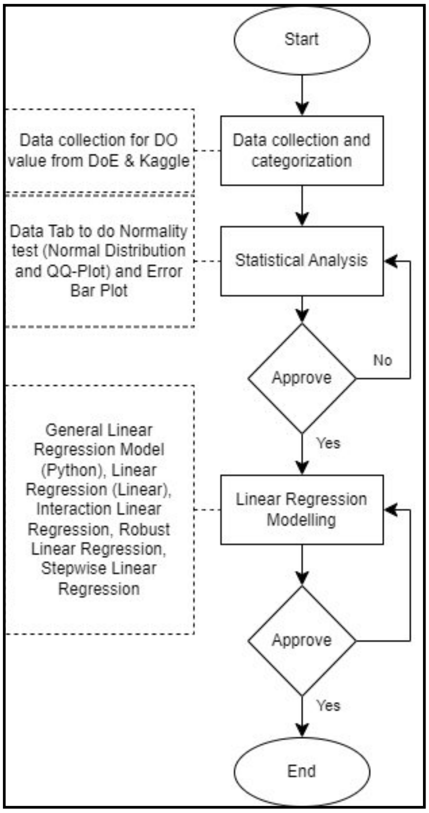
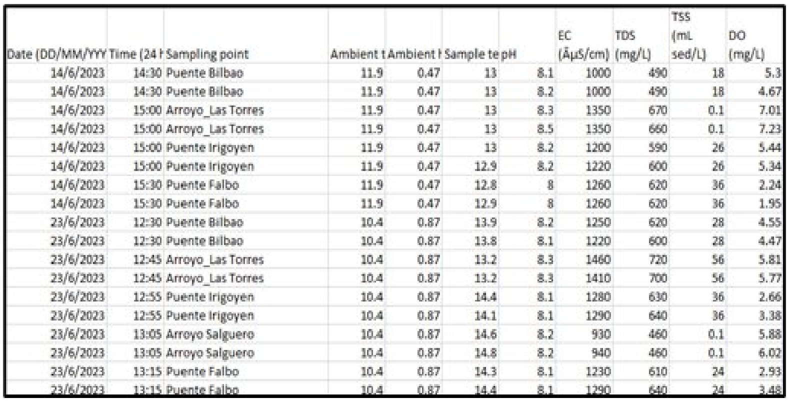

Start
Data collection for DO value from DoE & Kaggle
Data collection and categorization
Data Tab to do Normality test (Normal Distribution and QQ-Plot) and Error Bar Plot
Statistical Analysis
Approve
No
Yes
General Linear Regression Model (Python), Linear Regression (Linear), Interaction Linear Regression, Robust Linear Regression, Stepwise Linear Regression
Linear Regression Modelling
Approve
Yes
End

Fig. 1. Flowchart of the process flowchart

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037

3. DATA COLLECTION AND CATEGORIZATION

Fig. 2 shows a snippet of data collected from Kaggle. This dataset comprises 219 clean river water samples collected from June to November 2023. The data were arranged according to the standard for DO values. The data will be categorized into clean and polluted, with a threshold for clean water for DO above 5.0 mg/L [15]. In any research relating to water quality, the involvement of the Department of Environment (DoE) is essential because water quality regulations help minimize health risks and user conflicts caused by pollutants [16-17]. Clean Water Act mandates specific water quality standards, ensuring that water bodies are classified according to their intended use [18].



Date (DD/MM/YYY	Time (24 I	Sampling point	Ambient t	Ambient I	Sample te	pH	EC (ÂµS/cm)	TDS (mg/L)	TSS (mL sed/L)	DO (mg/L)
14/6/2023	14:30	Puente Bilbao	11.9	0.47	13	8.1	1000	490	18	5.3
14/6/2023	14:30	Puente Bilbao	11.9	0.47	13	8.2	1000	490	18	4.67
14/6/2023	15:00	Arroyo_Las Torres	11.9	0.47	13	8.3	1350	670	0.1	7.01
14/6/2023	15:00	Arroyo_Las Torres	11.9	0.47	13	8.5	1350	660	0.1	7.23
14/6/2023	15:00	Puente Irigoyen	11.9	0.47	13	8.2	1200	590	26	5.44
14/6/2023	15:00	Puente Irigoyen	11.9	0.47	12.9	8.2	1220	600	26	5.34
14/6/2023	15:30	Puente Falbo	11.9	0.47	12.8	8	1260	620	36	2.24
14/6/2023	15:30	Puente Falbo	11.9	0.47	12.9	8	1260	620	36	1.95
23/6/2023	12:30	Puente Bilbao	10.4	0.87	13.9	8.2	1250	620	28	4.55
23/6/2023	12:30	Puente Bilbao	10.4	0.87	13.8	8.1	1220	600	28	4.47
23/6/2023	12:45	Arroyo_Las Torres	10.4	0.87	13.2	8.3	1460	720	56	5.81
23/6/2023	12:45	Arroyo_Las Torres	10.4	0.87	13.2	8.3	1410	700	56	5.77
23/6/2023	12:55	Puente Irigoyen	10.4	0.87	14.4	8.1	1280	630	36	2.66
23/6/2023	12:55	Puente Irigoyen	10.4	0.87	14.1	8.1	1290	640	36	3.38
23/6/2023	13:05	Arroyo Salguero	10.4	0.87	14.6	8.2	930	460	0.1	5.88
23/6/2023	13:05	Arroyo Salguero	10.4	0.87	14.8	8.2	940	460	0.1	6.02
23/6/2023	13:15	Puente Falbo	10.4	0.87	14.3	8.1	1230	610	24	2.93
23/6/2023	13:15	Puente Falbo	10.4	0.87	14.4	8.1	1290	640	24	3.48

Fig. 2. Clean river data from Kaggle

*source: https://www.kaggle.com/datasets/natanaelferran/river-water-parameters/code

There are numerous water quality standards; however, since this project focuses on water quality in Malaysia, the official standards from the DoE were used to determine which WQI ranges are considered safe for drinking, irrigation, and aquatic life. Tables 1 and 2 represent the classes and corresponding WQI ranges set by the DoE of Malaysia.

Table 1. DoE WQI standards

Class	WQI	Irrigation	Aquatic Life	Drinking Water
I	>92.7	Suitable	Suitable	Suitable
II	76.5–92.7	Suitable	Suitable	Suitable
III	51.9–76.5	Suitable	Suitable	Suitable
IV	31.0–51.9	Not	Not	Not
V	<31.0	Suitable	Suitable	Suitable

Table 2. DoE water quality classification based on water quality index

Quality Index	Range	Description
Clean	91–100	Slightly Polluted
Polluted	80–90	Moderately Polluted
Heavily Polluted	0–79	Very Polluted

*source: https://www.doe.gov.my/en/national-river-water-quality-standards-and-river-water-quality-index/

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037

3.1 Statistical analysis

The categorized data were analyzed using statistical methods performed in DATAtab [19]. The analysis focused on several key parameters, including tests for normal distribution and the construction of error bar plots. First, a normality test was conducted to determine whether the data was appropriately distributed. Q-Q plots were also employed to assess the data's alignment with a normal distribution, providing insight into whether the normality assumption held or if data transformation was necessary [20]. Error bar plots were utilized to identify the variability and confidence intervals of the data, visually representing the mean values and the range within which the true values are likely to fall—thereby aiding in the assessment of data reliability and the comparison between groups.

Preparing the data for modelling is critical. Normal distribution charts help measure the data spread and highlight any outliers that might affect model performance [21]. Q-Q plots and Normal Probability Plots enable rapid detection of abnormalities that might indicate potential issues affecting the reliability of the regression model [22-23].

3.2 Linear regression modelling

After data screening and the discrimination between clean and polluted water, the data were imported into MATLAB R2024a and Python (scikit-learn) for linear regression modelling. In this study, the Dissolved Oxygen (DO) values are input features for predicting WQI. A supervised learning approach is employed, where the model is trained on labelled data comprising polluted and clean water samples.

In MATLAB, multiple regression models are implemented:

- Linear Regression
- Interaction Linear Regression
- Robust Linear Regression
- Stepwise Linear Regression

Each model is tailored for various analytical needs. Linear regression examines the relationship between predictors and one or more response variables, while Interaction Linear Regression assesses how combinations of predictors affect a given variable. Robust Linear Regression minimizes the influence of outliers, and Stepwise Linear Regression selectively includes variables to optimize the model. Similarly, Python's scikit-learn toolbox is used for linear regression analysis with methods such as Linear Regression, Polynomial Features, Huber Regressor, and Sequential Feature Selector, which explore different data analysis approaches without the need to categorize regression types explicitly.

The models are evaluated using performance metrics such as Root Mean Square Error (RMSE) and the coefficient of determination (R^2). R^2 is the proportion of variance in the dependent variable explained by the independent variables; a higher R^2 indicates a better fit. However, increasing model complexity (e.g., through varying coefficient models) may inflate R^2 by fitting noise rather than the underlying trend [24-26]. Then, the models undergo iterative refinement to optimize predictive performance, ensuring that the most accurate model is selected for predicting the Water Quality Index, as in Table 3.

Table 3. Platform and regression models

Platform	Regression Model	Description
MATLAB	Linear Regression	A simple linear model for prediction.
	Interaction Linear Regression	Includes interaction terms for features.
	Robust Linear Regression	Handles outliers by reducing their influence.
	Stepwise Linear Regression	Sequentially adds/removes predictors to optimize the model.
Python	Linear Regression (SKLearn)	A generalized linear regression model.

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037

4. RESULTS & DISCUSSION

4.1 Statistical analysis

The initial analysis involved performing preliminary statistical tests to determine whether the data could be used for linear regression modelling. This phase included normality tests and the construction of error bar plots. Descriptive analysis and comparisons of each dataset's distributional properties and sample sizes helped assess their suitability for further study. Data that passed the significance test ($p > 0.05$) were then visualized through error bar plots. When assessing dissolved oxygen (DO) levels in clean and polluted water, a p-value below 0.05 suggests that the observed difference is statistically significant and not due to random chance. This offers a strong foundation for making informed decisions about water quality. These plots verified that the datasets were distinguishable and reliable for subsequent modelling.

Test for normal distribution

Table 4 presents the normality test results for Dissolved Oxygen (DO) data in clean and polluted environments. For the clean DO data, the Kolmogorov-Smirnov test yields a p-value of 0.91, suggesting normality; however, the Shapiro-Wilk and Anderson-Darling tests show p-values of 0.003 and 0.001, respectively, indicating deviations from normality—possibly due to skewness or outliers. In contrast, the polluted DO data returned p-values greater than 0.05 across all tests, implying that these data are normally distributed and more appropriate for linear regression modelling.

Table 4. Normality test result for significant of dissolved oxygen (DO) clean and polluted

Parameter	Category	Type of Test	Statistic	P value
Dissolved Oxygen, DO	Clean	Kolmogorov-Smirnov (Lilliefors Corr)	0.16	0.91
		Shapiro-Wilk	0.91	0.003
		Anderson-Darling	1.51	0.001
	Polluted	Kolmogorov-Smirnov (Lilliefors Corr)	0.14	0.749
		Shapiro-Wilk	0.97	0.298
		Anderson-Darling	0.35	0.047

Error bar plot

The error bar plot in Fig. 4 demonstrates no overlap between the confidence intervals of the clean and polluted DO datasets [24]. The findings reveal a clear difference in dissolved oxygen (DO) levels between clean and contaminated water, as shown by the distinct 95% confidence intervals (CIs) that do not overlap. Clean water has an average DO of 4 mg/L with a CI ranging from 3 to 5 mg/L, while polluted water has a significantly lower average DO of 1 mg/L and a narrower CI of 0.8 to 1.2 mg/L. The absence of overlap in these intervals strongly suggests that the difference is statistically significant at the 95% confidence level and not just a result of random chance.

This result is not only statistically solid but also ecologically important, as the 75% drop in DO from clean to polluted water can significantly affect aquatic life, potentially creating hypoxic conditions where many organisms struggle to survive. These findings highlight the essential role of DO in evaluating water quality and point out the harmful effects of pollution, which likely adds organic matter that consumes oxygen during decomposition. This emphasizes the need for pollution control strategies, like reducing nutrient runoff and enhancing wastewater treatment, to keep DO levels healthy and safeguard aquatic ecosystems. Additionally, more research into the factors causing variability in clean water samples and formal hypothesis testing could enhance the conclusions and help set regulatory standards for DO management.

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037

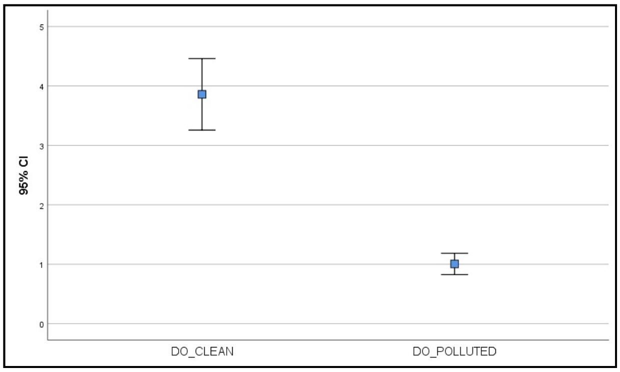


Fig.4. Error bar plot for clean and polluted data

4.2 Linear regression modelling

Table 5 presents the features of the linear regression model using 5 k-fold cross-validations as implemented in MATLAB and Python's scikit-learn. The models were evaluated by RMSE—which describes the average magnitude of prediction errors (lower is better)—and R², which indicates the proportion of variance explained by the model (higher is better). MATLAB's Linear Regression model was identified as the most effective, yielding the lowest RMSE (7.2728) and the highest R² (0.95397). The Robust Linear Regression model in MATLAB produced an RMSE of 7.3535 and an R² of 0.95294. Other MATLAB models, such as Interaction Linear and Stepwise Linear Regression, had higher RMSE values (7.5011 and 7.6872, respectively) and lower R², suggesting reduced accuracy. Conversely, the Python scikit-learn Linear Regression model performed poorly, with an RMSE of 25.8327 and an R² of 0.4040, indicating it failed to capture the underlying relationships. Since the study used the default settings of SKLearn, it can be concluded that the SKLearn model needs multiple adjustments in the settings to obtain the best results.

Table 5. Result of linear regression modelling of 5 folds

Software	Model	RMSE	R-Squared
MATLAB	Linear Regression	2.95307	0.95307
	Interaction Linear Regression	1.0511	0.95103
	Robust Linear Regression	7.6872	0.94587
	Stepwise Linear Regression	Unclear	0.94587
Python (SKLearn)	Linear Regression	25.8327	0.4040

Table 6 and Fig. 5 show the performance of the two selected MATLAB models—Linear Regression and Robust Linear Regression—evaluated using 5, 10, and 15 k-fold cross-validation. Both models maintained high R² values and low RMSE across different k-fold numbers. For the Linear Regression model, RMSE increased slightly from 7.2728 (5 folds) to 7.7507 (10 folds) before decreasing to 7.4822 (15 folds), while R² remained consistently high (0.95397, 0.94797, and 0.95277, respectively). The Robust Linear Regression model exhibited similar trends with marginally higher RMSE values. These results confirm the robustness and accuracy of the Linear Regression (Linear) model, which will be recommended for further predictive modelling.

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037

Table 6. Further modelling result with K-Fold from 5 to 15

Folds	Model	RMSE	R-Squared
(5 Folds)	Linear Regression	7.22728	0.95397
	Robust Linear	107.7507	0.94797
(10 Folds)	Linear Regression	7.3535	0.95294
	Robust Linear	7.7623	0.94781
(15 Folds)	Linear Regression	7.4822	0.95277
	Robust Linear	7.5293	0.95218

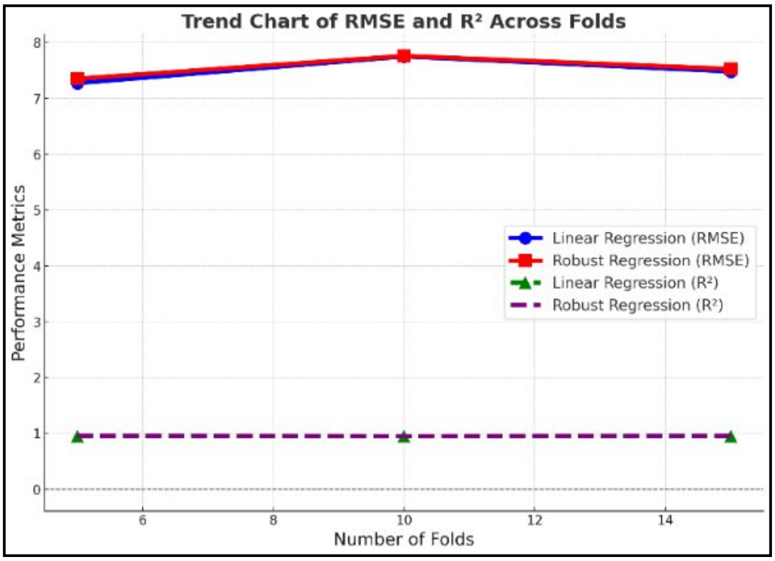


Fig. 5. Trend cart of RMSE and R^2

5. CONCLUSION

This study develops regression models to predict the Water Quality Index (WQI) using Dissolved Oxygen (DO) levels to assess river pollution, utilizing a dataset from Kaggle and the water quality standards based on those of the Department of Environment Malaysia. Traditional monitoring methods are slow and inefficient, requiring a data-driven approach. Various regression models were tested using MATLAB and Python, including Linear, Interaction, Robust, and Stepwise Regression. MATLAB's Linear Regression model performed best (R^2 = 0.95397, RMSE = 7.2728), while Python's model performed poorly (R^2 = 0.4040, RMSE = 25.8327). Statistical analysis confirmed that polluted DO data followed a normal distribution, supporting its use in regression models. Error bar plots validated significant differences between clean and contaminated water samples.

This study demonstrates that regression-based modelling is effective for real-time WQI prediction, supporting efficient water quality monitoring. Future work should explore neural networks and IoT-based monitoring systems. The findings align with Malaysia's DoE standards, aiding environmental policy and sustainable water management.

6. ACKNOWLEDGEMENTS/FUNDING

The authors would like to thank Electrical Engineering Studies, College of Engineering, Universiti Teknologi MARA, Cawangan Pulau Pinang for their support and assistance in completing this research work.

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037

7. CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this paper. All affiliations and funding sources supporting this study are transparently acknowledged, and no competing financial or non-financial interests exist that could have influenced the outcome of this research.

8. AUTHOR'S CONTRIBUTION

Muhamad Irfan Ahmad Suhkri: Contributed to the design of the study and data collection. **Wan Mohamad Haziq Wan Roselan:** Contributed to the literature review, assisted with the data analysis, and provided significant input in interpreting the results. **Mohamad Faizal Abd Rahman:** Supported the project by providing technical expertise and guidance on the analytical tools. **Moheddin Usodan Sumagayan:** Contributed to the technical expertise on water quality assessment. **Mohd Suhaimi Sulaiman:** Conceptualized the research, led the manuscript drafting, and coordinated the entire research team.

9. REFERENCES

[1] N. A. S. Abdullah *et al.*, "Water Quality Assessment of Tekala River Selangor," in *Proc. Int. Conf. Environ. Sci. Technol.*, 2017, pp. 101–106

[2] Z. S. Khozani, M. Iranmehr, and W. Mohtar, "Improving Water Quality Index prediction for water resources management plans in Malaysia: application of machine learning techniques," *Geocarto Int.*, vol. 37, no. 25, pp. 10058–10075, Dec. 2022. Available: https:// doi.org/10.1080/10106049.2022.

[3] R. Afroz, M. M. Masud, R. Akhtar, and J. B. Duasa, "Water pollution: Challenges and future direction for water resource management policies in Malaysia," *Environ. Urban. Asia*, vol. 5, no. 1, pp. 63–81, Mar. 2014. Available: https:// doi.org/10.1177/0975425314521544

[4] A. J. Smalley *et al.*, "Predictive Modeling Approach for Surface Water Quality: Development and Comparison of Machine Learning Models," *J. Environ. Manage.*, vol. 230, pp. 365–378,

[5] A. S. Selim, S. N. A. Islam, M. M. Moniruzzaman, S. Shah, and M. Ohiduzzaman, "Predictive Models for Dissolved Oxygen in an Urban Lake by Regression Analysis and Artificial Neural Network," *Total Environ. Res. Themes*, 2023. Available: https:// doi.org/10.1016/j.totert.2023.100066

[6] J. Smith and D. W. Jones, "Correlation Between Conductivity and Total Dissolved Solids in Various Types of Water," *Water Res.*, vol. 45, no. 4, pp. 1483–1495, 2018

[7] J. P. N. and M. S. Vijaya, "River Water Quality Prediction and Index Classification Using Machine Learning," *J. Phys.: Conf. Ser.*, vol. 2325, no. 1, 2022. Available: https:// doi.org/10.1088/1742-6596/2325/1/012011

[8] H. L. Chan *et al.*, "Correlation between Electrical Conductivity and Total Dissolved Solids in Natural Waters," *Int. J. Environ. Sci. Technol.*, vol. 16, no. 2, pp. 729–738, 2019

[9] J. M. Z. Hoque, N. A. Ab. Aziz, S. Alelyani, M. Mohana, and M. Hosain, "Improving Water Quality Index Prediction Using Regression Learning Models," *Int. J. Environ. Res. Public Health*, vol. 19, no. 20, p. 13702, 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/20/13702

[10] D. AM *et al.*, "Prediction of Water Quality Parameters of River Periyar Using Regression Models," in *Proc. 2nd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE)*, Greater Noida, India, 2022, pp. 53–57. Available: https:// doi.org/10.1109/ICACITE53722.2022.9823774

[11] S. Palabıyık and T. Akkan, "Evaluation of water quality based on artificial intelligence: performance of multilayer perceptron neural networks and multiple linear regression versus water quality indexes," *Environ. Dev. Sustain.*, 2024. Available: https:// doi.org/10.1007/s10668-024-05075-6

[12] B. Sharma and H. Kaur, "Parameters of Water to be Predicted using Regression Analysis," in *Proc. 11th Int. Conf. Syst. Model. Adv. Res. Trends (SMART)*, Moradabad, India, 2022, pp. 970–976. Available: https:// doi.org/10.1109/SMART55829.2022.10046732

[13] C. Lavanya, M. Nikitha, N. Swetha, K. Nikhitha, and L. Hussein, "Assessment and estimation of water quality using multi-linear regression," *E3S Web Conf.*, vol. 529, p. 03007, 2024. [Online]. Available: https://doi.org/10.1051/e3sconf/202452903007

[14] National Water Quality Standards for Malaysia, [Online]. Available: https://doe.gov.my/wp-content/uploads/2021/11/Standard-Kualiti-Air-Kebangsaan.pdf (accessed Feb. 28, 2025)

[15] C. E. Boyd, "Water Quality Regulations," in *Water Quality: An Introduction*, Cham: Springer Int. Publ., 2015, pp. 339–352

[16] G. F. Lee and A. Jones-Lee, "Clean Water Act, Water Quality Criteria/Standards, TMDLs, and Weight-of-Evidence Approach for Regulating Water Quality," in *Water Encyclopedia*, pp. 598–604

[17] J. W. Moore, "Water Quality Guidelines and Standards," in *Balancing the Needs of Water Use*, New York, NY: Springer New York, 1989, pp. 244–254

[18] DATAtab Team, *DATAtab: Online Statistics Calculator*, Graz, Austria, 2024. [Online]. Available: https://datatab.net

[19] F. B. Oppong and S. Y. Agbedra, "Assessing univariate and multivariate normality. A guide for non-statisticians," *Math. Theory Model.*, vol. 6, no. 2, pp. 26–33, 2016

[20] F. F. de Campos, O. A. B. Licht, and N. B. F. Campos, "PPlot, a webapp to partition geochemical data and isolate mixed subpopulations using probability plot modeling," *Geochim. Brasiliensis*, vol. 37, p. e-23002, Sep. 2023. Available: https:// doi.org/10.21715/GB2358-2812.202337002

[21] R. Jiang, P. Li, and K. Zhang, "Quantile-Quantile Plot of Folded-Normal Distribution and its Applications in Reliability and Quality Modeling," in *Proc. 10th Int. Symp. Syst. Secur., Safety, and Reliability (ISSSR)*, Xiamen, China, 2024, pp. 44–50. Available: https:// doi.org/10.1109/ISSSR61934.2024.00011

[22] M. S. Sameera and G. R. Kancharla, "The Selection of Best Fit Model Involving Correlation in Examination with QQ Plot," in *Proc. Int. Conf. Edge Comput. Appl. (ICECAA)*, Tamilnadu, India, 2022, pp. 814–817. Available: https:// doi.org/10.1109/ICECAA55415.2022.9936255

[23] Q.-Y. Peng, J.-J. Zhou, and N.-S. Tang, "Varying coefficient partially functional linear regression models," *Stat. Papers*, vol. 57, no. 3, pp. 827–841, Sep. 2016, doi: 10.1007/s00362-015-0681-3

[24] Y. C. Wu, J. Q. Fan, and H. G. Müller, "Varying-coefficient functional linear regression," *Bernoulli*, vol. 16, no. 3, pp. 730–758, Aug. 2010. Available: https:// doi.org/10.3150/09-bej231

[25] S. G. Schreiber *et al.*, "Statistical tools for water quality assessment and monitoring in river ecosystems – a scoping review and recommendations for data analysis," *Water Qual. Res. J.*, vol. 57, no. 1, pp. 40–57, Feb. 2022. Available: https:// doi.org/10.2166/wqrj.2022.028

https://doi.org/ 10.24191/esteem.v21iMarch.5023.g3037