INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS

ISSN(print): 2643-9840, ISSN(online): 2643-9875

Volume 06 Issue 10 October 2023

DOI: 10.47191/ijmra/v6-i10-36, Impact Factor: 7.022

Page No. 4815-4827

Predictive Modeling of WQI of Trans Amadi Industrial Layout, Port Harcourt, Rivers State Nigeria Using Multiple Linear Regression

Agori, J. E[1]. Atikpo, E[2], IWEMAH E.R[3]. Umukoro, O. L.[4]

[1,2,3,4]Department of Civil Engineering, Faculty of Engineering, Ole.h Campus, Delta State University, P. M. B. 22, Oleh, Nigeria

ABSTRACT: It is important to monitor water quality and understand factors influencing indices like the Water Quality Index (WQI). This study aimed to assess water quality and to develop a predictive model for WQI based on common ions and using multiple linear regression analysis. Water samples were collected from 15 locations in Trans Amadi Industrial Layout, Port Harcourt and analyzed for 19 groundwater quality parameters (Temp, EC, DO, Turbidity, NO_3, HCO_3, pH, TDS, Cl, SO_4, Fe, Na, Ca, Mg, Zn, Cu, Cr, Pb and Cd). Results of correlation analysis revealed that there were significant relationships between some of these parameters and the WQI. A multiple linear regression (MLR) model was developed to quantitatively examine relationships between the Water Quality Index (WQI) and key indicator variables and to predict the WQI based on the parameters. The model exhibited a strong coefficient of determination (R-squared = 0.9811), indicating a high level of accuracy in predicting the WQI. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values confirmed the model's predictive capability. Overall, this study contributed to understanding the factors influencing water quality and showed that linear regression is a reliable method for predicting WQI. The model accurately explains variation in WQI based on nitrate, chloride, sodium, calcium and magnesium levels.

KEYWORDS: Water Quality Index, linear regression, Prediction, groundwater quality parameters

1.0 INTRODUCTION

Water is a crucial component of the environment. Water quality impacts human and environmental health, but surface water and groundwater quality have long been deteriorating due to both natural and human-related activities.. Natural factors that influence water quality are hydrological, atmospheric, climatic, topographical and lithological factors (Magesh et al., 2013; Uddinet al., 2018). While anthropogenic activities that adversely affect water quality are mining, livestock farming, production and disposal of waste (industrial, municipal and agricultural), increased sediment run-off or soil erosion due to land-use change (Lobato et al., 2015) and heavy metal pollution (S´anchez et al., 2007). Water quality assessment is crucial for maintaining the health of aquatic ecosystems and ensuring safe water for human consumption. In regions with rapid industrialization and urbanization, such as the Trans Amadi Industrial Layout in Port Harcourt, Rivers State, Nigeria, water quality can be significantly affected due to the discharge of various pollutants into water bodies and on ground surface. To effectively manage and monitor water quality in such areas, the use of water quality indices (WQIs) has become a common practice.

Water quality assessment is crucial for maintaining the health of aquatic ecosystems and ensuring safe water for human consumption. In regions with rapid industrialization and urbanization, such as the Trans Amadi Industrial Layout in Port Harcourt, Rivers State, Nigeria, water quality can be significantly affected due to the discharge of various pollutants into water bodies. To effectively manage and monitor water quality in such areas, the use of water quality indices (WQIs) has become a common practice.

In this study, we focus on the application of multiple linear regression (MLR) to predict the WQI in the Trans Amadi Industrial Layout, Port Harcourt. MLR is a statistical technique used to model the relationship between a dependent variable (in this case, WQI) and multiple independent variables (water quality parameters). By identifying the key water quality parameters that influence the WQI, MLR enables us to develop a predictive model for assessing water quality (Singh et al., 2004).

1.1 The Water Quality Index (WQI)

The Water Quality Index (WQI) is a composite numerical expression that provides an overall assessment of the suitability of water for various uses based on multiple water quality parameters (Brown et al., 1970). It offers a convenient way to communicate complex water quality information to stakeholders and decision-makers. It is a tool developed by scientists to help evaluate the quality of water that helps in representing the overall quality of water based on various water quality parameters. (EPA, 2017). It's often used to communicate the health of water bodies and to make informed decisions regarding water usage. It summarizes large amounts of water quality data into a single "score" from 1 to 100, with lower scores indicating cleaner water (EPA, 2017). It offers a convenient way to communicate complex water quality information to stakeholders and decision-makers. The calculation of WQI involves the integration of several physicochemical and biological parameters, which makes it a multidimensional problem. The WQI combines data on several parameters into a single score representing quality. Understanding relationships between WQI and chemical measures aids management and remediation efforts. WQI models involve four consecutive stages; (1) selection of the water quality parameters, (2) generation of sub-indices for each parameter (3) calculation of the parameter weighting values, and (4) aggregation of sub-indices to compute the overall water quality index.

1.1.1 Assignment of weighing (w_i)

This is where a scale of importance is in play. Not all of the measured parameters are used to compute water quality index, hence there is need to select the needed parameters on the basis of their importance. Here, a factor is judged more important than the other based on the level of harm they can cause when present in groundwater. Those ones with high hazards are assigned high value while those with less harm are assigned lower value. Each parameter is assigned with a value on the basis of their relative importance which has a significant role in overall water quality, as per the guidelines of standardizing agencies. The parameters with low permissible limits can cause maximum extent of pollution even on slight fluctuation, while the high permissible limit allows relatively less chances of pollution. Two key factors considered in assigning unit weight are the parameter with the narrowest/lowest range of permissible limit and its influence on water quality as well as on the health risk index (HRI) e.g., nitrate is assigned the highest weight of 5 and down to 1 for the least important ones e.g., zinc (Singh et al., 2018). Further, the unit weight of each parameter is calculated and the mean assigned weights and WHO standard for each parameter is presented.

1.2 Correlation and Regression Analysis

Correlation and regression analysis are valuable statistical tools for exploring and understanding data relationships, making predictions, and informing decision-making. They are often used together to provide a more comprehensive analysis of data. However, it's important to interpret the results carefully and consider the underlying assumptions and limitations of these techniques.

Correlation analysis is used to measure the strength and direction of a relationship between two or more variables. It helps determine whether changes in one variable are associated with changes in another. The result of a correlation analysis is the correlation coefficient, which ranges from -1 to +1. Regression analysis on the other hand is used to model the relationship between a dependent variable and one or more independent variables (Agori et al.). It aims to find the best-fit line (or curve) that predicts the dependent variable based on the independent variables. The goal is to understand how changes in the independent variables affect the dependent variable. It could be Simple Linear Regression or Multiple Linear Regression. Regression analysis provides coefficients for each independent variable, indicating the strength and direction of their impact on the dependent variable. The model's accuracy can be assessed using metrics like the coefficient of determination (R-squared), mean squared error (MSE), and root mean squared error (RMSE).

2.0 WATER QUALITY INDEX PREDICTION MODEL

In this study, the multiple linear regression (MLR) was applied in predicting the WQI in the Trans Amadi Industrial Layout, Port Harcourt. MLR is a statistical technique used to model the relationship between a dependent variable (in this case, WQI) and multiple independent variables (water quality parameters). By identifying the key water quality parameters that influence the WQI, MLR was used to develop a predictive model for assessing water quality (Singh et al., 2004).

2.1 Model performance metrics

The performance of the MLR model is often times evaluated using The Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). Also comparing the RMSE to the range of WQI values provided context for the model's accuracy. If the RMSE is significantly smaller than the range of WQI values, it suggests that the model is performing well.

Mean squared error (MSE): The MSE measures the average squared difference between the predicted values and the actual values. A lower MSE indicates better predictive performance that the model's predictions are closer to the actual values. A higher MSE suggests that the model's predictions are farther from the actual values.

Root mean squared error (RMSE): The RMSE is the square root of the MSE and is often used to express the error in the same units as the target variable. Like the MSE, a lower RMSE value is desirable, indicating that the model's predictions are more accurate. The RMSE is the square root of the MSE and is expressed in the same units as the target variable.

The R-squared (coefficient of determination): R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (WQI in this case) that is explained by the independent variables (water quality parameters in this case) included in the regression model and it is used to measure the goodness of fit of the model. R-squared ranges: R-squared ranges from 0 to 1. A higher R-squared value indicates that a larger proportion of the variance in the dependent variable is explained by the independent variables, and therefore, the model provides a better fit to the data. A lower R-squared value suggests that the model doesn't explain much of the variability in the dependent variable.

3.0 MATERIALS AND METHOD

3.1 Description of study area

Trans Amadi industrial layout is in Obio-Akpo L.G.A. of Port Harcourt, Rivers State, Nigeria. It is located between latitude of 4º47'N and 4º48'N, and 7º1'E and 7º2'E longitude, with an estimated area of a thousand hectares (10.12km^2). It lies in the north and is bordered by D/line in the south west, Woji Township to the east and Rumola in the north west, Figure 1. The main abattoir of the city is also located here. It is the main industrial area of Port-Harcourt, densely populated with buildings of many national and trans-national firms. It also has diverse residential neighbourhood in the city of Port Harcourt. Unsuitable urbanization, population growth and changes in lifestyles have contributed to increasing per capita solid waste generation in Trans Amadi which has resulted in poor environmental conditions that have affected the quality of both surface and ground water. Effluents from the industries have also worsened the quality of both surface water and groundwater.

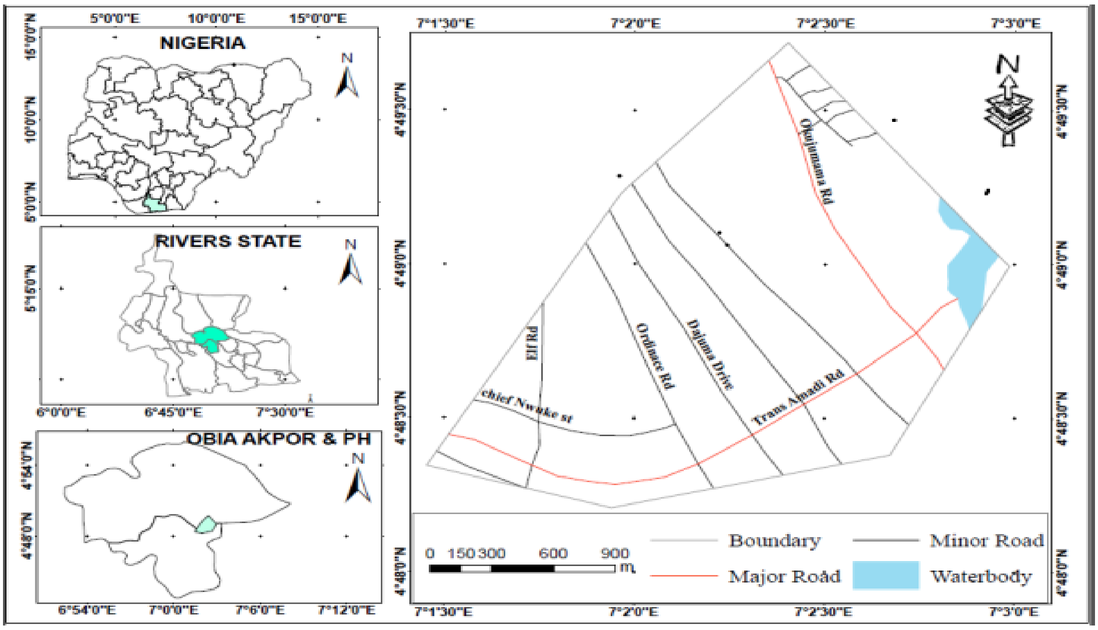

Figure 1: Trans Amadi Industrial Layout, Port Harcourt

It is located 5 km from the heart of Port Harcourt. The study area enjoys tropical hot monsoon climate due to its latitudinal position. The tropical monsoon climate is characterized by heavy rainfall from April to October ranging from 2000 mm to 2500 mm with high temperature throughout the year and a relatively constant high humidity. The study area is influenced by urbanization or urban sprawl whereby smaller communities have merged together and form a megacity. This is due to high influx of people, resulting in a rapid growth of the population in the study area. This in turn is largely due to the expansion of the oil and allied industries which have also attracted many varied manufacturing industries. The relief is generally lowland which has an average elevation of between 20 m and 30 m above sea level. The geology of the area is basically comprised of alluvial sedimentary basin and basement complex. The vegetation found in this area includes raffia palms, thick mangrove forest and light rain forest. The soil is usually sandy or sandy loam underlain by a layer of impervious pan and is always leached due to the heavy rainfall experienced in this area. The study area is well drenched with both fresh and salt water. The salt water is caused by the intrusion of sea water inland, thereby making the water slightly salty. Due to continuous heavy rainfall and river flow, the study area experiences severe flooding almost every year and the effects are extended to biological resources hence the city was chosen as the study area. According to available records depths to water levels in wells vary from 5m - 7m.

3.2 Dataset

3.2.1 Sampling locations, water samples collection and analysis

ArcMap 10.6. was used to obtain the satellite imagery of the study area and 15 collection locations B1 to B15 were identified based on spread, land use (high and low population density, industrial, agricultural, open solid waste dump sites and other anthropogenic activities. The depths of some boreholes were also determined using a plumb bulb and line.

Water samples were collected from the selected 15 existing boreholes during the wet season (August 2022 –October, 2022) using new high-density PET screw-capped containers of 1.5litres capacity. The PET containers and stoppers were thoroughly washed with distilled water and rinsed with water from the bore holes. The bottles were filled, allowed to overflow and immediately corked, properly labeled to avoid mix up, placed in an ice block chest and transported to the laboratory within a prescribed period of not more than three hours after collection. The APHA, 2017 standard method of testing water and waste water was used in analyzing the samples. All the samples were processed by using the analytical-grade chemicals of Merck and glassware of Borosil. Double-distilled water was used for the reagent preparation, instrument calibration and rinsing of glassware.

The non-conservative (sensitive) parameters; temperature, pH, electrical conductivity (EC), Total Dissolved Solids (TDS) and dissolved oxygen (DO) which change with storage time were measured in-situ with appropriate apparatus (Mercury thermometer, pH meter (PHS-25), Conductivity/TDS Meter (HI 2315, Hanna Instrument)and DO meter (DO analyzer JPG 607) respectively and recorded before other samples were transported to a water laboratory and preserved in refrigerators at 4^oC in the laboratory to keep the samples intact and later analyzed for other physical, chemical and bacteriological parameters.

The water samples were analyzed to obtain the concentration of eighteen (19) physico-chemical and bacteriological groundwater quality parameters (pH, temperature, turbidity, electrical conductivity (EC), dissolved oxygen (DO), chemical oxygen demand (COD), biochemical oxygen demand (BOD), salinity, nitrate (NO_3), iron (Fe), sodium (Na), calcium (Ca), magnesium (Mg), bicarbonate (HCO_3), sulphate (SO_4), total dissolved solid (TDS), zinc (Zn), copper (Cu), iron (Fe),lead (Pb), Cadmium (Cd) and Chromium (Cr). The results of the water quality concentrations are presented in Table 2. Maximum care was taken to minimize the errors during sampling, storage, handling and analysis of the samples. The built-in function of Microsoft Excel was used in obtaining the mean, range (minimum and maximum values) and standard deviation of each parameter and were presented in Table 3 with the WHO and NSFDWQ standard permissible limits.

3.2.2 Evaluation of water quality index (WQI)

Assignment of weighing (w_i)

Data from the 15 locations were used for the evaluation of WQI. Some parameters were chosen based on their relative importance (significant role in overall water quality, as per the guidelines of standardizing agencies). were assigned weights. Accordingly, the weights were assigned in a range of 1–5. The unit weight of each parameter was calculated and the mean assigned weights and WHO standard are presented in Table 1.

The assigned mean weight values of some selected parameters and WHO permissible standard values and their respective observed concentrations in table 2 were used to evaluate the water quality index (WQI) in the 15 locations using the Weighted Arithmetic method of computing WQI using Equations 1 to 4 and the Microsoft Excel spreadsheet result presented in Table 4.

Table 1: ***Mean assigned weights and WHO*** **permissible standard.**

Water quality parameter	Mean assigned weight	WHO permissible standard
pH	3.7	7.0-8.9
TDS (mg/l)	4.5	1500
Cl^- (mg/l)	3.5	250
SO_4^{2-} (mg/l)	4	500
NO_3^- (mg/l)	5	50.0
HCO_3^- (mg/l)	2.3	500
Na^+ (mg/l)	2.8	250
Ca^+ (mg/l)	2.3	75.0

Mg^+ (mg/l)	2.3	20.0
	$\sum W_i = 30.4$	

$$\text{Relative weight } (W_i) = \frac{w_i}{\sum_{i=1}^{n} w_i} \quad (1)$$

$$\text{Quality rating scale } (q_i) = \frac{c_i}{s_i} \times 100 \quad (2)$$

$$Sl_i = W_i \times q_i \quad (3)$$

$$WQI = \sum Sl_i \quad (4)$$

where;

n = Number of parameters

w_i = Weight of each parameter

q_i = Rating based on concentration of ith parameter

c_i = Observed concentration of each parameter in each water sample in mg/l

S_i = Standard value

SI_i = Sub − index of ith parameter

W_i = Sum of mean weights

3.2.3 Correlation and regression analysis

The dataset was imported to Microsoft Excel, ensuring that each column represented a different water quality parameter with the last column containing the computed WQI values at each location. The built-in function of Microsoft Excel was used to conduct the correlation analysis in order to examine bivariate relationships between WQI and each potential predictor variables. This generated a correlation matrix in the form of a chart (Table 5) and the correlation coefficient (R) of each parameter with the WQI was determined. The water quality parameters were taken as the independent variables and the actual WQI as the dependent variable. The parameters with R values equal to or greater than 0.5 (DO, NO_3, Cl, Na, Ca and Mg) were used as the predictor variables to prepare the Multi Linear Regression (MLR) Statistics and is presented in Table 6. These parameters showed moderate to strong positive correlations with WQI, with correlation coefficients ranging from 0.55 to 0.80. Based on these correlation results indicating meaningful linear associations between the response and predictor candidates, a stepwise multiple linear regression was then performed. This procedure sequentially selected the set of predictor variables (DO, NO_3, Cl, Na, Ca, Mg) that best predicted WQI while avoiding collinearity, based on their partial correlations with the response when controlling for other terms in the model.

4.0 PREDICTION OF WQI

From the obtained multiple linear regression table, the WQI for each location was predicted using the regression coefficients (intercept and slope) of each parameter.

$$\text{Predicted WQI} = \text{Intercept} + \text{Coefficient } 1 \times \text{Parameter } 1 + \cdots + \text{Coefficient } n \times \text{Parameter } n$$

Where: n is the total number of parameters considered in the regression analysis.

Predicting Equations for WQI = 13.0070DO - 0.5765NO_3 + 0.0894Cl - 0.0369Na + 0.0812Ca + 0.0855Mg + 25.5054 (5)

4.1 Evaluation of Model Performance

The model's performance was evaluated based on the magnitude of the MSE and RMSE values. From the calculated and predicted WQI, the residuals were calculated. The predicted WQI values were compared to the actual WQI values and the model's performance was evaluated based on the magnitude of the MSE, RMSE and R-squared values. Mean Squared Error (MSE) = $\frac{\sum(Actual\ WQI - Predicted\ WQI)^2}{n}$ (6)

a) Root Mean Squared Error (RMSE) = $\sqrt{MSE}$ (7)

b) $\text{R} - \text{squared } (R^2) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$ (7)

Where:

n is the number of data points.

y_i is the actual WQI value for the ith location.

$\hat{y}_i$ is the predicted WQI value for the ith location.

$\bar{y}_i$ is the mean of the actual WQI values.

These values represent the mean squared error and root mean squared error for the predicted WQI values compared to the actual WQI values across all locations. The R-squared value was obtained using the actual and predicted WQI values at the 15 locations and Equation 7.

5.0 RESULTS

The concentration of water quality parameters are shown in Table 2 and the descriptive statistics of water samples, WHO and NAFDAC permissible drinking water quality standards are presented in Table 3.

Table: 2: Water quality parameters at sampling locations

Loca-tions	Temp °C	EC µS/cm	DO mg/l	Turbid NTU	NO_3 mg/l	HCO_3 mg/l	pH	TDS mg/l	Cl mg/l	SO_4 mg/l	Fe mg/l	Na mg/l	Ca mg/l	Mg mg/l	Zn mg/l	Cu mg/L	Cr mg/l	Pb mg/l	Cd mg/l	Actual WQI
B1	25.9	209	5.9	0.47	10.86	120	7.0	496	671	6.28	0.08	56	280	135	0.02	0.000	0.01	0.010	0.001	188.39
B2	26.6	389	6.2	0.019	40.01	12.5	4.0	220	710	7.5	2.4	404	620	880	0.13	0.002	0.01	0.060	0.0	213.43
B3	27.1	573	4.2	0.027	28.04	18.5	5.3	684	243	6.83	1.8	138	60	90	0.06	0.010	0.02	0.025	0.001	52.70
B4	26.2	521	5.8	0.91	32.47	35	6.3	346	566	8.2	7.15	303.6	60	50	0.14	0.000	0.0	0.220	0.0	160.34
B5	29.3	895	6.40	0.062	18	20	6.2	431	24	6.2	0.01	47.8	92	25.9	0.09	0.002	0.0	0.002	0.0	62.87
B6	27.3	717	6.1	0.01	17.27	365	7.1	450	86.8	6.74	0.00	36.8	150	70	0.02	0.001	0.0	0.150	0.001	201.93
B7	26.5	236	5.2	0.97	15.66	38	7.15	210	75.1	8.55	9.6	46	120	40	0.17	0.000	0.0	0.200	0.00	93.35
B8	26.4	200	5.4	0.35	10.86	110	7.0	530	71	6.16	0.02	46	280	120	0.01	0.020	0.0	0.200	001	137.36
B9	27.6	339	6.2	0.024	14	29.5	6.2	410	399	9.4	2.25	372	187	820	0.18	0.030	0.02	0.050	0.0	219.10
B10	28.1	250	5.1	0.025	2.7	55.5	6.4	573	12.2	5	0.6	9.2	70	40	0.02	0.004	0.00	0.280	0.0	62.24
B11	28.1	343	6.2	0.028	40.01	13.6	5.3	320	419	7.5	2.2	278	564	820	0.09	0.002	0.0	0.075	0.001	256.65
B12	27.6	230	4.8	0.057	2.9	12	5.0	312	14.1	5.04	0.02	9.2	38	22	0.01	0.00	0.0	0.090	0.001	78.29
B13	26.3	1000	4.7	0.071	3.3	13	5.0	341	21.3	5	0.02	13.8	34.55	25.7	0.04	0.010	0.0	0.006	0.0	102.69
B14	28.3	300	3.9	0.06	1.7	14.5	6.8	370	7.1	16.16	0.02	4.6	48	42.4	0.02	0.002	0.01	0.i00	0.01	71.38
B15	29.5	296	1.1	0.08	2.7	14.5	6.6	369	9.1	17.16	0.03	7.6	58	42.4	0.03	0,001	0.02	0.100	0.0	71.34

Table 3: Groundwater descriptive Statistics, WHO and NAFDAC permissible limits

	Temp. °C	EC (µS/cm)	DO mg/l	Turbidity NTU	NO_3 mg/l	HCO_3mg/l	pH	TDS mg/l	Cl mg/l	SO_4 mg/l	Fe mg/l	Na mg/l	Ca mg/l	Mg mg/l	Zn mg/l	Cu mg/L	Cr mg/l	Pb mg/l	Cd mg/l
Mean	27.2	426.3	5.1	0.066	14.7	84.5	6.4	398.4	166.4	10.8	4.05	140.3	208.6	55.7	0.055	0.003	0.0043	0.119	0.000
Std. dev	1.3	265.7	1.6	0.026	11.2	152.3	0.4	177.6	208.5	14.2	5.19	157.1	246.1	37.2	0.05	0.005	0.0066	0.090	0.001
Min.	25.9	200	3.9	0.019	1.1	10.9	4.0	210	6.16	1.7	0.01	9.2	38	25.7	0.01	0.001	0.0	0.002	0.0
Max	29.5	1000	7.15	0.97	6.4	365	7.15	1000	10.86	40.01	32.47	880	820	135	0.22	0.02	0.02	0.280	0.001
WHO. Std	27.0	1200	6.0	< 5	50.0	500	7.0-8.9	1500	250	500	1.0	250	75.0	20.0	3.0	0.5	0.05	0.01	0.003
NAFDAC Std.	27.0	1000	5.0	6.0	10.0	100	6.5-8.5	500	100	100	3.0	200	75	20.0	5.0	1.0	0.05	0.01	0.003

Table 4: Water quality index and water status

Water Quality Parameter	pH	TDS	Cl	SO_4	HCO_3	NO_3	Ca	Mg	Na	QUALITY RATING (qi) = (ci/si)*100									PARAMETER INDEX (Si) = qi*Wi									WQI	WATER STATUS
WHO Standard	7.5	500	250	100	100	50	75	8.2	200																				
Relative Weight Wi	0.12	0.148	0.155	0.132	0.164	0.076	0.09	0.08	0.076																				
Sampling Location	MEASURED VALUES (Ci)									pH	TDS	Cl	SO4	HCO_3	NO_3	Ca	Mg	Na	pH	TDS	Cl	SO_4	HCO_3	NO_3	Ca	Mg	Na		
B1	7.00	496.00	671.00	6.28	120.00	10.86	280.00	135.00	56.00	93.33	500.00	268.40	6.28	120	21.7	373.33	1646.34	28.00	11.20	74.00	41.60	0.83	19.68	0.83	25.20	10.80	4.26	188.39	UNSUITABLE
B2	4.00	220.00	710.00	7.50	12.50	40.01	620.00	880.00	404.00	53.33	0.15	284.00	7.50	12.5	80.0	826.67	10731.71	202.00	6.40	0.02	44.02	0.99	2.05	3.04	55.80	70.40	30.70	213.43	UNSUITABLE
B3	5.30	684.00	243.00	6.83	18.50	28.04	60.00	90.00	138.00	70.67	0.00	97.20	6.83	18.5	56.1	80.00	1097.56	69.00	8.48	0.00	15.07	0.90	3.03	2.13	5.40	7.20	10.49	52.70	POOR
B4	6.30	346.00	566.00	8.20	35.00	32.47	60.00	50.00	303.60	84.00	496.00	226.40	8.20	35	64.9	80.00	609.76	151.80	10.08	73.41	35.09	1.08	5.74	2.47	5.40	4.00	23.07	160.34	UNSUITABLE
B5	6.20	431.00	24.00	6.20	20.00	18.00	92.00	25. 9	47.80	82.67	220.00	9.60	6.20	20	36.0	122.67	315.85	23.90	9.92	32.56	1.49	0.82	3.28	1.37	8.28	1.52	3.63	62.87	POOR
B6	7.10	450.00	86.80	6.74	365.00	17.27	150.00	70.00	36.80	94.67	684.00	34.72	6.74	365	34.5	200.00	853.66	18.40	11.36	101.23	5.38	0.89	59.86	1.31	13.50	5.60	2.80	201.93	UNSUITABLE
B7	7.15	210.00	75.10	8.55	38.00	15.66	120.00	40.00	46.00	95.33	346.00	30.04	8.55	38	31.3	160.00	487.80	23.00	11.44	51.21	4.66	1.13	6.23	1.19	10.80	3.20	3.50	93.35	VERY POOR
B8	7.00	530.00	71.00	6.16	110.00	10.86	280.00	120.00	46.00	93.33	431.00	28.40	6.16	110	21.7	373.33	1463.41	23.00	11.20	63.79	4.40	0.81	18.04	0.83	25.20	9.60	3.50	137.36	UNSUITABLE
B9	6.20	410.00	399.00	9.40	29.50	14.00	187.00	820.00	372.00	82.67	450.00	159.60	9.40	29.5	28.0	249.33	10000.00	186.00	9.92	66.60	24.74	1.24	4.84	1.06	16.83	65.60	28.27	219.10	UNSUITABLE
B10	6.40	573.00	12.20	5.00	55.50	2.70	70.00	40.00	9.20	85.33	210.00	4.88	5.00	55.5	5.4	93.33	487.80	4.60	10.24	31.08	0.76	0.66	9.10	0.21	6.30	3.20	0.70	62.24	POOR
B11	5.30	320.00	419.00	7.50	13.60	40.01	564.00	820.00	278.00	70.67	530.00	167.60	7.50	13.6	80.0	752.00	10000.00	139.00	8.48	78.44	25.98	0.99	2.23	3.04	50.76	65.60	21.13	256.65	UNSUITABLE
B12	5.00	312.00	14.10	5.04	12.00	2.90	38.00	22.00	9.20	66.67	410.00	5.64	5.04	12	5.8	50.67	268.29	4.60	8.00	60.68	0.87	0.67	1.97	0.22	3.42	1.76	0.70	78.29	VERY POOR
B13	5.00	341.00	21.30	5.00	13.00	3.30	34.55	25. 7	13.80	66.67	573.00	8.52	5.00	13	6.6	46.07	313.41	6.90	8.00	84.80	1.32	0.66	2.13	0.25	3.11	1.36	1.05	102.69	UNSUITABLE
B14	6.80	370.00	7.10	16.16	14.50	1.70	48.00	42.40	4.60	90.67	320.00	2.84	16.16	14.5	3.4	64.00	517.07	2.30	10.88	47.36	0.44	2.13	2.38	0.13	4.32	3.39	0.35	71.38	POOR
B15	6.60	369.00	9.10	17.16	14.50	2.70	58.00	42.40	7.60	88.00	312.00	3.64	17.16	14.5	5.4	77.33	517.07	3.80	10.56	46.18	0.56	2.27	2.38	0.21	5.22	3.39	0.58	71.34	POOR

Table 5.: Correlation matrix of water quality parameters

Parameters	Temp. °C	EC (µS/cm)	DO (mg/l)	Turbidity (NTU)	NO_3 (mg/l)	HCO_3 (mg/l)	pH	TDS (mg/l)	Cl (mg/l)	SO_4 (mg/l)	Fe (mg/l)	Na (mg/l)	Ca (mg/l)	Mg (mg/l)	Zn (mg/l)	Cu (mg/l)	Cr (mg/l)	Pb (mg/l)	Cd (mg/l)	Actual WQI
Temp. °C	1																			
EC (µS/cm)	0.0506	1																		
DO (mg/l)	-0.4130	0.1725	1																	
Turbidity NTU)	-0.5321	-0.2271	0.1252	1																
NO3 (mg/l)	-0.2300	0.1128	0.5029	0.1310	1															
HCO3 (mg/l)	-0.1981	0.1363	0.2676	-0.0203	-0.0472	1														
pH	0.0615	-0.2473	-0.1131	0.4285	-0.4274	0.4802	1													
TDS (mg/l)	0.0488	0.0529	-0.0711	-0.2752	-0.1766	0.2494	0.2566	1												
Cl (mg/l)	-0.4923	-0.1763	0.4686	0.2316	0.6932	-0.0541	-0.2947	-0.1475	1											
SO4 (mg/l)	0.4897	-0.2701	-0.6822	-0.0285	-0.2103	-0.2019	0.2924	-0.2167	-0.1447	1										
Fe (mg/l)	-0.3774	-0.1600	0.1922	0.8065	0.4267	-0.1885	0.1044	-0.4396	0.2827	0.0051	1									
Na (mg/l)	-0.2255	-0.0565	0.4654	0.0360	0.7829	-0.2220	-0.4666	-0.2705	0.7843	-0.0535	0.3757	1								
Ca (mg/l)	-0.2004	-0.2457	0.4622	-0.1250	0.6787	0.0057	-0.3817	-0.2912	0.6524	-0.1530	0.0350	0.6224	1							
Mg (mg/l)	-0.0384	-0.1854	0.4169	-0.2816	0.6127	-0.1833	-0.4828	-0.3078	0.6246	-0.0154	0.0816	0.8438	0.8182	1						
Zn (mg/l)	-0.1574	0.0321	0.3865	0.4091	0.5563	-0.3114	-0.1515	-0.4616	0.4515	0.0203	0.7664	0.7292	0.2545	0.5280	1					
Cu (mg/l)	-0.3995	-0.2076	0.2264	0.1956	-0.1243	-0.0999	0.1138	0.1705	0.1444	-0.1994	0.1524	0.2849	-0.0822	0.2121	0.3336	1				
Cr (mg/l)	0.3163	-0.0618	-0.4986	-0.2670	-0.0431	-0.2106	-0.0308	0.3366	-0.0097	0.4210	-0.0694	0.1948	-0.2101	0.1631	0.1826	0.3469	1			

Pb (mg/l)	-0.0995	-0.4152	-0.0283	0.4345	-0.1025	0.2284	0.4360	0.0193	-0.2142	-0.0050	0.3983	-0.1255	-0.1555	-0.2590	-0.0095	-0.0110	-0.2748	1		
Cd (mg/l)	-0.2423	-0.2504	0.0485	0.1219	-0.1098	0.1496	0.2652	0.2655	-0.1629	-0.1398	-0.1609	-0.1365	0.1487	-0.0809	-0.2616	0.4041	-0.1332	0.3075	1	
Actual WQI	-0.3462	-0.1069	0.5902	0.0139	0.5845	0.3224	-0.1269	-0.2801	0.7121	-0.1579	0.1166	0.6937	0.7615	0.7642	0.3563	0.1928	-0.1277	-0.1036	0.0191	1

Table 6: Multiple linear regression statistics water quality predictors

Regression Statistics of NO_3, Cl, Na, Ca and Mg	
Multiple R	0.8961
R Square	0.8041
Adjusted R Square	0.7532
Standard Error	6.1228
Observations	15.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5.0000	18403.6	3680.7	8.3159	0.002
Residual	9.0000	4489.1	498.7		
Total	14.0000	22892.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	25.5054	50.2932	0.5071	0.6257	-90.4709	141.4818	-90.4709	141.4818
DO (mg/l)	13.0070	10.7979	1.2046	0.2628	-11.8929	37.9069	-11.8929	37.9069
NO_3 (mg/l)	-0.5765	2.2485	-0.2564	0.8041	-5.7614	4.6085	-5.7614	4.6085
Cl (mg/l)	0.0894	0.1022	0.8748	0.4072	-0.1463	0.3251	-0.1463	0.3251
Na (mg/l)	-0.0369	0.3861	-0.0955	0.9263	-0.9273	0.8536	-0.9273	0.8536
Ca (mg/l)	0.0812	0.2146	0.3781	0.7152	-0.4138	0.5761	-0.4138	0.5761
Mg (mg/l)	0.0855	0.1598	0.5352	0.6071	-0.2830	0.4540	-0.2830	0.4540

Table 6: Predicted WQI and R^2 value

Locations	DO mg/l	NO_3 mg/l	Cl mg/l	Na mg/l	Ca mg/l	Mg mg/l	Actual WQI	Predicted WQI	Residuals	SSR	Actual WQI - Mean Value	SST	$R^2 = 1$-SSR/SST
B1	5.9	10.86	671	56	280	135	188.39	188.9	-0.51	0.26	56.92	3239.81	
B2	6.2	40.01	710	404	620	880	213.43	186.7	26.73	714.49	81.96	6717.33	
B3	4.2	28.04	243	138	60	90	52.7	63.97	-11.27	127.01	-78.77	6204.82	
B4	5.8	32.47	566	303.6	60	50	160.34	148.89	11.45	131.10	28.87	833.44	
B5	6.4	18	24	47.8	92	25.9	62.87	68.25	-5.38	28.94	-68.60	4706.05	
B6	6.1	17.27	86.8	36.8	150	70	201.93	194.6	7.33	53.73	70.46	4964.52	
B7	5.2	15.66	75.1	46	120	40	93.35	99.49	-6.14	37.70	-38.12	1453.19	
B8	5.4	10.86	71	46	280	120	137.36	136.64	0.72	0.52	5.89	34.68	0.973
B9	6.2	14	399	372	187	820	219.1	207.57	11.53	132.94	87.63	7678.90	
B10	5.1	2.7	12.2	9.2	70	40	62.24	67.74	-5.5	30.25	-69.23	4792.89	
B11	6.2	40.01	419	278	564	820	256.65	234.33	22.32	498.18	125.18	15669.87	
B12	4.8	2.9	14.1	9.2	38	22	78.29	73.6	4.69	22.00	-53.18	2828.18	
B13	4.7	3.3	21.3	13.8	34.55	25.7	102.69	97.08	5.61	31.47	-28.78	828.33	
B14	3.9	1.7	7.1	4.6	48	42.4	71.38	69.12	2.26	5.11	-60.09	3610.89	
B15	1.1	2.7	9.1	7.6	58	42.4	71.34	72.52	-1.18	1.39	-60.13	3615.70	
						SUM	1972.06	1909.40	62.66	1815.1	0.00	67178.58	0.9811
						MEAN	131.47	127.29	4.18	121.01			

N/B

Sum of squared difference between actual and predicted values (SSR) = $\sum_{i=1}^{15}(y_i - \hat{y}_i)^2$

Sum of squared difference between actual and mean values (SST) = $\sum_{i=1}^{15}(y_i - \bar{y}_i)^2$

Actual and Predicted WQI

Actual WQI: 188.39, 213.43, 52.7, 160.34, 62.87, 201.93, 93.35, 137.36, 219.1, 62.24, 256.65, 78.29, 102.69, 71.38, 71.34.

Predicted WQI: 188.90, 186.7,0 63.97, 148.89, 68.25, 194.60, 99.49, 136.64, 207.57, 67.74, 234.33, 73.60, 97.08, 69.12, 72. 52.

Model performance

i. MSE = 116.0771

ii. RMSE= 10.7733

iii. R^2 = 0.9811.

6.0 DISCUSSION AND CONCLUSION

A weighted arithmetic WQI approach factored for all designated uses seems appropriate to holistically assess status. Weights require adjusting over time as new data and hydrologic understanding develops. Selecting locally calibrated benchmarks aligned with all river uses would strengthen interpretation. Classifying results within standardized categories aids communicating outcomes and priorities. The actual WQI ranged from 52.70 – 256.65 indicating that the water quality status of the investigated boreholes were from bad to unsuitable. This is as a result of the adverse impact of anthropogenic activities and the wanton discharge of waste water and disposal of solid waste in the study area.

The correlation analysis identified DO, NO_3, Cl, Na, Ca and Mg as candidate predictor variables that exhibited moderate to strong linear associations with WQI. The MSE value of 116.0771 suggests that, on average, the squared difference between predicted and actual WQI values is relatively moderate. This means that, on average, the squared difference between the predicted and actual WQI values across all locations is 116.0771. The Sum of Squared Residuals (SSR) represents the sum of the squared differences between the predicted WQI and the actual WQI values, providing a measure of how well the model fits the data. The RMSE value of 10.7733 indicates that, on the average, the difference between predicted and actual WQI values is approximately 10.7733 units across all locations. This value is far lower than the WQI values, indicating that the model's predictions are more accurate. The RMSE value of 10.7733 is close to the standard deviation of the target variable and considering the range of WQI values, the model's performance seems to be reasonable and is therefore considered a reasonable model. The regression model exhibited a high coefficient of determination (R-squared) value between the actual and predicted WQI of approximately 0.9811 indicating that the model was able to explain around 98.11% of the variance in the WQI. These values further supported the accuracy of the regression model in predicting the WQI. The model's performance metrics validated the robustness and generalizability of the developed model. This indicates a high degree of correlation between the actual and predicted values, suggesting that the model performed well in explaining the variability in the data. The R-squared indicates how well the regression model fits the data, it implies that 98.11% of the variability in the Actual WQI can be explained by the predictor variables (DO mg/l, NO3 mg/l, Cl mg/l, Na mg/l, Ca mg/l, Mg mg/l) in the model. The other 1.89% of the variability is due to unknown factors not included in the model or random error.

This analysis demonstrated the utility of regression modeling for understanding relationships between water indicators. Further expansion of the dataset could refine parameter estimates and improve predictive accuracy. Nonetheless, the model presented provides a useful starting point to inform water monitoring and management decisions through quantitative evaluation of key water quality determinants impacting the WQI. In conclusion, the multiple regression approach successfully captured key relationships between WQI and predictor water quality variables. With further validation, the model could support management and monitoring needs. Future work could refine model using expanded or new datasets. Overall, while these results provide an assessment of the model's accuracy, it's important to also consider domain knowledge and the specific context of water quality prediction. Further refinement of the model, feature engineering, and potentially exploring other regression techniques could help improve its performance.

REFERENCES

1) Agori, J. E., Nwoke, H. U., Okoro, B. C. and Dike, B. U. (2021). Application of Correlation and Regression Models in Predicting the Physico-chemical Quality of Groundwater from Insitu Measured Parameters. European Journal of Engineering and Technology Research, 6(6), 19 – 38.
2) Brown, R. M., McClelland, N. I., Deininger, R. A., & Tozer, R. G. (1970). A water quality index—crashing the psychological barrier. Water, Air, & Soil Pollution, 1(3), 389-397.

3) Lobato, T.C., Hauser-Davis, R.A., Oliveira, T.F., Silveira, A.M., Silva, H.A.N., Tavares, M. R.M., Saraiva, A.C.F., 2015. Construction of a novel water quality index and quality indicator for reservoir water quality evaluation: A case study in the Amazon region. J. Hydrol. 522, 674–683.
4) Magesh, N.S., Krishnakumar, S., Chandrasekar, N., Soundranayagam, J.P., 2013. Groundwater quality assessment using WQI and GIS techniques, Dindigul district, Tamil Nadu, India. Arab. J. Geosci. 6, 4179–4189
5) Sanchez, ´ E., Colmenarejo, M.F., Vicente, J., Rubio, A., García, M.G., Travieso, L., Borja, R., 2007. Use of the water quality index and dissolved oxygen deficit as simple indicators of watersheds pollution. Ecol. Indic. 7, 315–328
6) Singh, K. P., Malik, A., & Mohan, D. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study. Water Research, (18), 3980-3992.
7) U.S. Environmental Protection Agency (2017). Freshwater Quality
8) Uddin, M.G., Moniruzzaman, M., Quader, M.A., Hasan, M.A., 2018. Spatial variability in the distribution of trace metals in groundwater around the Rooppur nuclear power plant in Ishwardi, Bangladesh. Groundw. Sustain. Dev. https://doi.org/10.1016/j. gsd.2018.06.002.