IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)
e-ISSN : 2278-1684, p-ISSN : 2320–334X
PP 65-70
www.iosrjournals.org

APPLICATION OF REGRESSION ANALYSIS FOR SURFACE WATER QUALITY MODELING

K. U. Ahamad[1a], B.K.Nazary[2a], P. Mudoi[3b], R. Rani[2a], V. Bharati[5a] and H. Singh[6a]

[1]Department of Civil Engineering, Tezpur University, Napaam 784028, Assam, India
[b]Department of Molecular Biology and Biotechnology, Tezpur University, Napaam 784028, Assam, India
**Corresponding author, E mail: kahamad@tezu.ernet.in; Phone: +91-3712-275959*

Abstract: *With the growth of human populations, commercial and industrial activities, surface water had received large amount of waste water which not only pollutes the system, but degrades the quality of these water bodies. Cconstant monitoring of water quality is needed to record any alteration in the quality and outbreak of health disorders as well as for water quality management. Present study aims to develop a water quality prediction model using linear regression model. Water from the lake situated inside the Tezpur University has been collected for a period of 4 months and various water quality parameters has been estimated. Regression analysis has been carried out using the estimated water quality parameters. Nine water quality parameters were selected to set up the regression model. The dependent variable was Dissolve oxygen and the independent variables were electrical conductivity, total solid, turbidity, pH, alkalinity, chloride, Biological oxygen demand, temperature and phosphorous. From these data, multiple regression analysis is done using data analysis tool of MS Excel at 75% confidence level and the regression equation relating DO with its dependent water quality parameters has been obtained. The obtained regression analysis matches the experimental values vey accurately with a R^2 value of 0.9235.*

Keywords*: Dissolved Oxygen, Regression analysis, Surface water quality, water quality modeling*

I. Introduction

Human being needs water for drinking, cooking and for personal hygiene and the quality of the water, therefore plays the most important role. With the growth of human populations, commercial and industrial activities, surface water had received large amount of waste water from various sources. The surface water bodies are used as a sink for the disposal of domestic, industrial and agricultural wastewaters, which not only pollutes but degrades these water bodies. Increasing surface water pollution causes not only the deterioration of water quality, but also threatens human health, balance of aquatic ecosystem, economic development and social prosperity. It require attention to prevent and control the surface water pollution and to have reliable information on its quality for effective management. The industrial units located at the outskirts of cities, intensive agricultural practices, and indiscriminate disposal of domestic wastes are the sources of contamination for these water bodies. Thus, constant monitoring of surface water quality is needed so as to record any alteration in the quality and outbreak of health disorders as well as for water quality management. One individual parameter cannot express the quality of water and it requires all the water quality parameters controlled by physical, chemical and biological compositions such as natural (precipitation, geology of the watershed, climate and topography) and anthropogenic (domestic, industrial activities and agricultural run-off) sources. Therefore, the water quality is assessed by measuring a broad range of parameters. Moreover in order to develop a predictive model large number of data has to be generated for simulation and for the generation of model equation. In the recent past many researchers have carried out studies in this direction to understand and model the process [1-9]. Adamu et al., 2012, has carried out a study to identify the pollution sources and their contribution toward water quality variation using principal component analysis and multiple linear regressions [10]. The major objective of the present study is to develop a water quality prediction model using linear regression model. Water from the lake situated inside the Tezpur University has been collected for a period of 4 months and various water quality parameters has been estimated. Regression analysis has been carried out using the estimated water quality parameters.

II. Methods

2.1 Sampling Location and Strategy

Water sample for the present study was collected from the natural lake situated inside the Tezpur University.Water level is high during rainy season and least water level is obtained during winter. Water of the lake is highly turbid throughout the year.For the experimental purpose water samples from the sampling site were collected in the month of January, February, March and April 2015. The samples were collected at an interval of five days. A 5L plastic container was used for sample collection. Prior to sampling the containers were washed carefully to remove any solids or impurities which may be present earlier. All samples were collected well away from the edges of the water body. Temperature of the sample was recorded at the site itself. All the experiments were carried out in environmental engineering laboratory of civil engineering department of Tezpur University. The estimated water quality parameters were dissolve oxygen, electrical conductivity, total solid, turbidity, pH, alkalinity, chloride, biological oxygen demand, temperature and phosphorous. All the experiment carried out as per the method specified in standard methods [11].

2.2 Regression Model

Regression analysis is used for explaining or modeling the relationship between a single variable *y*, called the response, an output or dependent variable, and one or more predictor, input, independent or explanatory variables, x_1, ………, x_p. When $p = 1$, it is called simple regression, but when $p > 1$ it is called multiple regression or sometimes multivariate regression. The simplest multiple linear regression equation becomes:

$$y_i = a + \varepsilon_i + \sum_{i=1}^{n} b_i x_i$$

where, quantities $b_1, b_2, \ldots, b_n$ are termed partial regression coefficient.

In multiple regression, as we are concerned with several variables, we can measures the effect of variations in the independent variable on those in the dependent variable by calculating an analogous measure *r*. This is called the coefficient of multiple determination. It can vary between -1 and 1 and it tells us what proportion of the observed variance of *y* is due to variations in $x_1, x_2, \ldots, x_n$. The relationship of water quality parameters on each other in data of water analyzed was determined by calculating Karl Pearson's correlation coefficient, R, by using the formula given:

$$R = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\,\Sigma(Y-\bar{Y})^2}}$$

Where, X (X= value of X-variable, $\bar{X}$ = average value of X) and Y (Y= value of Y-variable, $\bar{Y}$ = average value of Y)

A single water quality parameter is the dependent variable, and its variation is accounted for by the variation in two or more independent variables of physical, chemical, or biological water characteristics. The general equation is:

$$Y^{'} = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$$

Where Y' is the dependent variable, X's are the independent variables, k the number of independent variables in the equation, and 'a' is the regression constant. By choosing appropriate independent variables (water parameters), we seek to maximize the correlation between the predicted value of our water quality variable and the actual value of the variable.

III. RESULTS AND DISCUSSION

The quality of water is characterized by various physicochemical parameters. These parameters change widely due to many factors like source of water, type of pollution, seasonal fluctuations, etc. Ten parameters were analyzed for over a period of four months (January 2015 – April 2015). The analytical results are illustrated in Fig. 1, depicting the seasonal variation of selected parameters at the sampling sites.

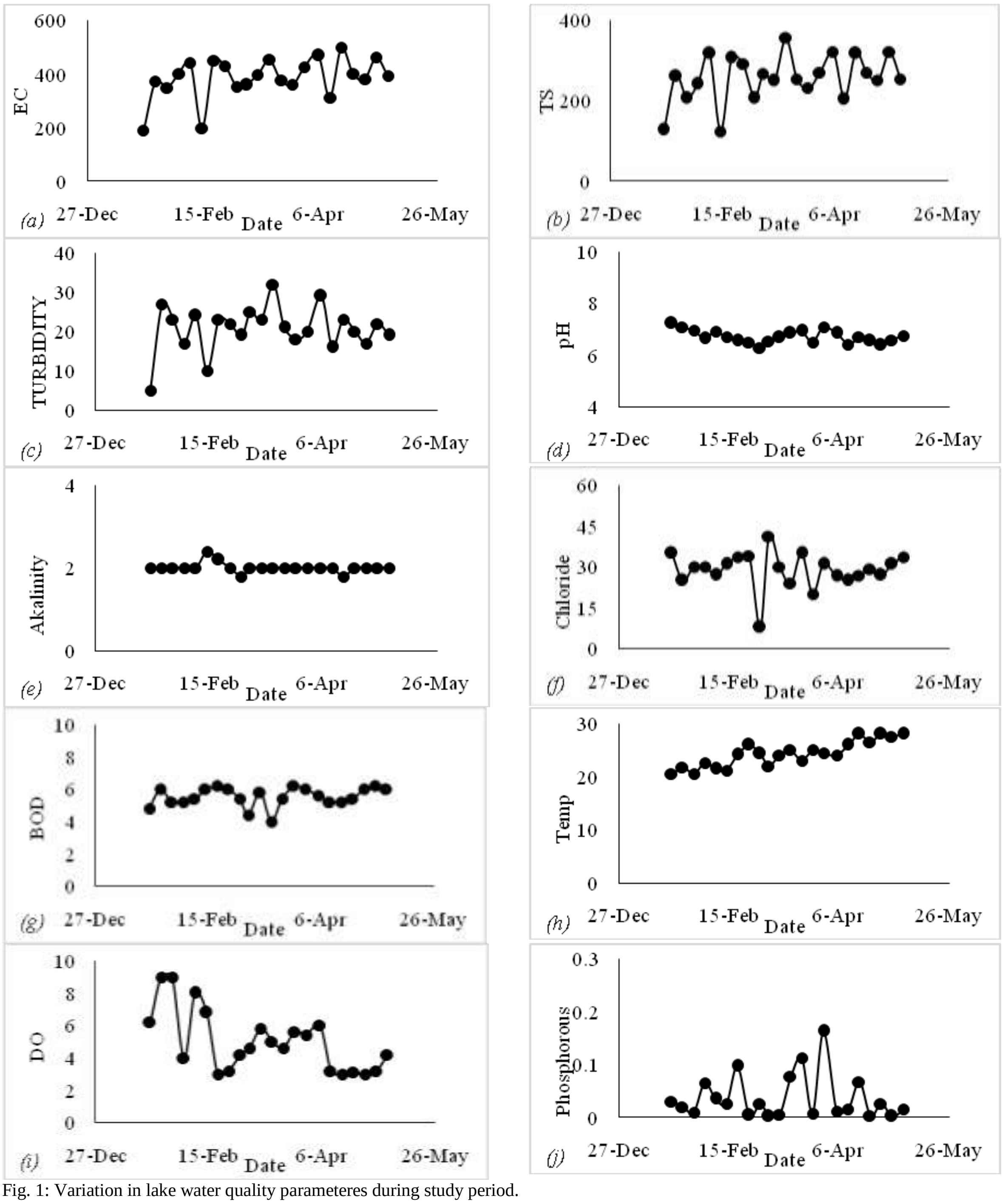

Fig. 1: Variation in lake water quality parameteres during study period.

The conductivity were found to be in the range between 190-500 µs/cm at Lake (Fig1*a*). The highest conductivity values observed in the month of April and lowest in the month of January. The conductivity is mostly affected by the bedrock and soil in the watershed. The Total solid values were ranged between 125 to 360 mg/L (Fig. 1*b*). The maximum values of total solid occurred during April while minimum during January month. One of the possible reason for possible increase in total solids is because of rainfall - which may bring the nearby soil and debris into the lakes. Turbidity mainly indicates the pollution level of the water bodies. The turbidity varied between 5 to 35 NTU. In January and February the values were almost identical, and fluctuates between 17 and 23 NTU (Fig. 1*c*). Increase in the turbidity is may be due to silt, clay and other suspended particles brought in the reservoir by surface run off. The pH value varied between 6.2 and 7.3. In study period pH were almost identical, and fluctuate between 6.4 and 7.3, the lowest values are recorded in February (Fig. 1*d*). Alkalinity of water samples varied from 1.8 to 2 mg/L (Fig. 1*e*). The mild alkalinity may be due to the seepage of effluent, domestic sewage etc. into the lake. Chloride levels in the lakes have ranged from 25 mg/L to 50 mg/L (Fig. 1*f*). Dissolve Oxygen in the lake varies from 3-8 mg/L (Fig. 1*i*). The average concentration of BOD in the lake ranged from 4-6 mg/L (Fig. 1*g*). A high concentration of BOD indicate low DO and low BOD concentration indicate high DO. The concentration of BOD in the lake is very low, which may be contributed by the organic decomposition of nearby vegetation which may fall in the lakes. Maximum and minimum temperature varies for 20°C (in January) to 28°C (in April) during the study period (Fig. 1*h*). Lakes surface temperature measurements are confined to the top portion of the water. During the daytime in high sunshine conditions may lead to the formation of a warm layer at the lake's surface. Average concentration of Phosphorous were observed as 0.002 to 0.163 mg/L (Fig. 1*j*). The increasing concentration of available phosphorus allows plants to assimilate more nitrogen before the depletion of phosphorus. The phosphorus in the water is contributed by the inflow of nearby residential and hostel wastewater into the lake.

3.1 Regression Model

A range of surface water quality data were generated and examined in the present work for a comprehensive water quality evaluation. An analysis of the data, based on multiple regression, was attempted. This was done to examine whether seasonal or spatial variation in different parameters can be explained and predicted based upon their interrelationship in terms of source and mobility. From analysis point of view, nine parameters were selected to set up the regression model. The dependent variable was Dissolve oxygen and the independent variables were electrical conductivity, total solid, turbidity, pH, alkalinity, chloride, Biological oxygen demand, temperature and phosphorous. From these data, multiple regression analysis is done using data analysis tool of MS Excel at 75% confidence level and the regression equation relating DO with its dependent water quality parameters has been obtained. Summary of regression analysis and the result is presented in Table 1. The estimated regression equation is

$$\text{Dissolved oxygen} = [-24.359 + 0.002(\textit{Electrical Conductivity}) - 0.014(\textit{Total Solid}) + 0.155(\textit{Turbidity}) + 4.894(\textit{pH}) + 1.796(\textit{Alkalinity}) - 0.081(\textit{Chloride}) + 0.548(\textit{BOD}) - 0.323(\textit{Temperature}) - 13.270(\textit{Phosphorous})]$$

For the analysis of the data sample using regression analysis, 75% confidence level has been fixed. Therefore p value should be within 25% confidence level i.e., p value should be less than 0.25 for the analysis to be accurate within the assumed confidence level. It has been observed from the Table 1 (*before standardization part*), the p value for Electrical Conductivity and Alkalinity is more than 0.25 (25%) i.e., p value for Electrical Conductivity is 0.784 and Alkalinity is 0.448. However p value for total solid is slightly more than 0.25, i.e 0.278 – may be considered for the analysis.

As the two water quality parameter falls outside the assumed confidence limit, therefore the regression equation observed using the parameters is not valid and necessary corrections has to be made in the equation. Hence further regression equation has been obtained by eliminating the mentioned two parameters, the modified linear empirical equation has been developed, where the basic approach taken is same as earlier. The final regression values are presented in Table 1 under the section *After Standardization*, and the final linear modified empirical equation is given below.

The final regression equation is

$$\text{Dissolved oxygen} = [-19.145 - 0.012(\textit{Total Solid}) + 0.160(\textit{Turbidity}) + 4.614(\textit{pH}) - 0.069(\textit{Chloride}) + 0.729(\text{BOD}) - 0.356(\textit{Temperature}) - 12.220(\textit{Phosphorous})]$$

Table 1: Regression values for the estimated water quality parameters

Before Standardization							
Regression Statistics		*ANOVA analysis*					
			df	*SS*	*MS*	*F*	*Significance F*
R Square	0.928	Regression	9.000	70.289	7.810	17.101	0.000
Adjusted R Square	0.873	Residual	12.000	5.480	0.457		
Observations	22	Total	21.000	75.770			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-24.359	9.408	-2.589	0.024
Electrical Conductivity	0.003	0.010	0.281	0.784
Total Solid	-0.014	0.013	-1.135	0.278
Turbidity	0.155	0.061	2.528	0.027
pH	4.894	0.899	5.442	0.000
Alkalinity	1.797	2.289	0.785	0.448
Chloride	-0.082	0.029	-2.777	0.017
(BOD)5	0.549	0.396	1.385	0.191
Temperature	-0.324	0.104	-3.102	0.009
Phosphorous	-13.271	4.497	-2.951	0.012

After Standardization							
Regression Statistics		*ANOVA analysis*					
			df	*SS*	*MS*	*F*	*Significance F*
R Square	0.923	Regression	7.000	69.973	9.996	24.142	0.000
Adjusted R Square	0.885	Residual	14.000	5.797	0.414		
Observations	22	Total	21.000	75.770			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-19.145	6.258	-3.059	0.008
Total Solid	-0.012	0.006	-1.844	0.086
Turbidity	0.160	0.057	2.801	0.014
pH	4.614	0.767	6.018	0.000
Chloride	-0.069	0.024	-2.886	0.012
(BOD)5	0.729	0.253	2.881	0.012
Temperature	-0.356	0.092	-3.845	0.002
Phosphorous	-12.220	3.817	-3.202	0.006

A scatter is plotted between actual DO and predicted DO is shown in Fig. 2. R^2 value of scatter plot shows the probability of existence of error. If R^2 is approximately 1 then probability of existence of error is negligible that means the predicted model is valid. R^2 for the present analysis is 0.9235, which shows that the analysis carried out by regression analysis matches the experimental values vey accurately.

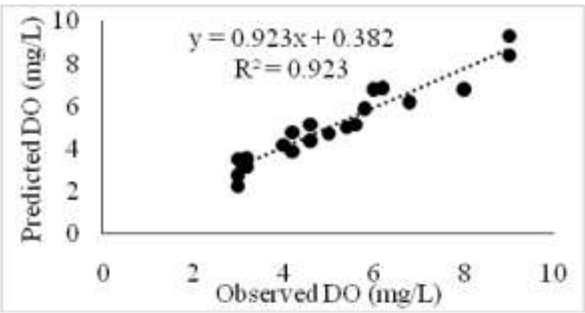

Fig. 2: Actual versus predicted DO values

IV. Conclusion

1. Water sample was collected from the lake situated inside the Tezpur University for a period of 4 months. The estimated water quality parameters were dissolve oxygen, electrical conductivity, total solid, turbidity, pH, alkalinity, chloride, biological oxygen demand, temperature and phosphorous.

2. Highturbidity in the lake was observed which may be due to silt, clay and other suspended particles brought in the reservoir by surface run off. The mild alkalinity may be due to the seepage of effluent, domestic sewage etc. into the lake. The average concentration of BOD in the lake ranged from 4-6 mg/L. The concentration of BOD in the lake is very low, which may be contributed by the organic decomposition of nearby vegetation which may fall in the lakes. Lakes surface temperature measurements are confined to the top portion of the water. Average concentration of Phosphorous were observed as 0.002 to 0.163 mg/L. The increasing concentration of available phosphorus allows plants to assimilate more nitrogen before the depletion of phosphorus. The phosphorus in the water is contributed by the inflow of nearby residential and hostel wastewater into the lake.

3. Nine water quality parameters were selected to set up the regression model. The dependent variable was Dissolve oxygen and the independent variables were electrical conductivity, total solid, turbidity, pH, alkalinity, chloride, Biological oxygen demand, temperature and phosphorous.

4. Multiple regression analysis is done using data analysis tool of MS Excel at 75% confidence level and the regression equation relating DO with its dependent water quality parameters has been obtained. The obtained regression analysis matches the experimental values vey accurately with a R^2 value of 0.9235.

References

[1] R. Aziz, and O. C. Muhammad, Regional interpretation of river Indus water quality data using regression model, African Journal of Environmental Science and Technology, 8(1), 2014, 86-90.

[2] H. Majid, O. Ehsan, M. Hamid and K. Ozgür, Development of a Neural Network Technique for Prediction of Water Quality Parameters in the Delaware River, Pennsylvania, Middle-East Journal of Scientific Research, 13(10), 2013,1367-1376.

[3] P. S. Kunwar, B. Ankita, M. Amrita, and J. Gunja, Artificial neural network modelling of the river water quality - A case study, Ecological Modeling, 220(6), 2009, 888–895.

[4] D. Emrah, S. Bulent, and K. Rabia, Modelling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique, Journal of Environmental Management, 90(2), 2009, 1229–1235.

[5] B. Ustun, S.Ekercin, and A. Oztopal, Investigation of water quality parameters by using multiple regression and fuzzy logic in the Istanbul Strait, Turkey, in Z.Bochenek (Ed.),New Developments and Challenges in Remote Sensing, (Millpress, Rotterdam, 1997).

[6] V. Simeonov, J. A. Stratis, C. Samara, G. Zachariadis, D. Voutsa, A. Anthemidis, M. Sofoniou, and T. Kouimtzis, Assessment of the surface water quality in Northern Greece, Water Research,37(17), 2003, 4119-4124.

[7] H. Wang, M. Hondzo, C. Xu, V. Poole, and A. Spacie, Dissolved oxygen dynamics of streams draining an urbanized and an agricultural catchment, Ecological Modeling, 160, 2003, 145-161.

[8] Z. Qing, and J. S. Stephen, Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modelling, Water Research, 31(9), 1997, 2340–2350.

[9] R. J. Wilcock, Agricultural runoff: A source of water pollution in New- Zealand,Journal of Agricultural Science, 20, 1986, 98-103.

[10] M. Adamu andA. Ado,Application of Principal Component Analysis and Multiple Regression Models in Surface Water Quality Assessment,Journal of Environment and Earth Science, 2(2), 2012, 16-23.

[11] APHA, Standard methods for the examination of water and wastewater, (20th Ed., American Public Health Association, American Water Works Association, Water Environment Federation, Washington, DC, USA, 1998).