© 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939

Modeling of Spatial and Temporal Variation in Nitrate-Nitrogen Concentration of Upper Lake Bhopal Using Best Subset Procedure

[1]Shubham Agarwal and [2]Harsangeet Kaur

[1]P.G. Research Scholar, [2]Associate Professor
[1,2]Department of civil engineering
[1,2] Maulana Azad National Institute of Technology, Bhopal

***Abstract*-** **Upper Bhopal Lake, the largest water lake in M.P., India is located in the heart of Bhopal town. A number of water channels enter the lake from the human settlements around the lake and discharge polluted water in the lake bringing several chemical as well as sediments load in the lake. Eutrophication is one of the major problems in some part of the lake. Water quality may become so degraded that the lake's original use are lost. Nitrogen is the limiting nutrients responsible for eutrophication of the lake. The physico-chemical status of Upper Lake (Bhopal, India) has been investigated during the year 2011-2015. Nitrate is the important nutrient in the lake loading through point and non-point pollution sources such as agricultural activities, domestic raw sewage, atmospheric precipitation and runoff.This nutrient support the fast growth of the aquatic plants as a result these plants lead to gradual shrinking of wetland area along with other complications like low light penetration, reduces oxygen concentration, clogging of water channels and lowers entertainment value of lake.**
In the present study, Best subset procedure based on R^2 and F values is used in model for predicting the Nitrate concentration.Regression model are developed in explaining variation in nitrate concentration using routinely measured parameters. The predicted value of nitrate is compared with observed values and reasonably good matching is obtained. Further, efficiency of the developed model was tested by error estimation using correlation statistics, standard error , t-testand ANOVA test. The developed model may be used for prediction of missing observed values of Nitrate concentration.

Keywords*- *Water quality parameters, Regression, Best subset, Nitrate modeling, Upper Bhopal Lake

I. INTRODUCTION

Nitrate and phosphate are twoimportant constituents that immensely help inthe growth of the plants where they present. Ifthey are present in lake and ponds they areexcessively promote the growth of aquaticweeds and polluting our aquatic resources.International studies on the nitrates in the surface waters of variousbodies of water have expressed their concernand drawn the attention of scientists around theglobe. This constituent is immensely helpin the growth of the macrophytes like waterhyacinth (Eichhornia crassipes) which is themost troublesome aquatic weed in the world.The major sources of nitrate in lakes and pondsare from the catchment area by rainfall, sewageeffluents, suspended organicmatter and agriculture runoff. When algae and other suspended microorganismsdie and settle down to the bottom; they carry nitrogen withthem, during decomposition. This nitrogen isreleased and becomes available for subsequentgrowth of aquatic biota (Singh, 1987).

Both modeling and monitoring are useful in attempting to ascertain a lake's response to Nitrogen loading from development. Monitoring provide present condition but is expensive, time consuming as lakes exhibit high variability in terms of algal response and nutrients concentration on both an annual and seasonal basis. Therefore there is a need to provide a cost effective alternative which can help to protect the water quality of lakes. Modeling provides such an alternative to interpretate data with prediction for the future.

Statistical models are developed for estimation of concentration of different water quality constituents using routinely-monitored water quality parameters. The best subset modeling procedure enables comparison between full models to subset models. Best subset procedure based on R^2 and F values can be used in model formation.

Sharma and Jain (2008) developed statistical models to estimate Phosphate concentration in upper lake of Bhopal city. Best subset procedure based on R^2 and F values was used in model formulation. It was found that potassium, fluoride, Secchi depth, pH, BOD and magnesium could be used as surrogate parameters for the prediction of phosphate and these parameters successfully explained 77.8 % variation in Phosphate concentration.

Phosphorous concentration = 0.088 + 0.040K - 0.512F - 0.033SD - 0.011Mg + 0.024pH + 0.007BOD

Muniz et al. (2000)developed a predictive model for nutrients below the surface layer of gulf of Lyon & Catalan sea; they carried statistical analysis for nitrate relationship with silicate, salinity & depth. The result of the statistical study for both nutrients (N & P) showed that the variation were different in time and space. (Nitrate = –187.86 + 0.0072 Depth+ 5.02 Salinity)

Pathak et al. (2012)studied about the impact of excess nutrients status in lake water ecosystem, they analyzed water sample for 26 water quality parameters from six sampling stations of sagar lake of sagar city (M.P.); regression analysis and multiple linear regression for total data point were carried out using SPSS 11.0 software, MLR method was used to evaluate relationship between dissolved oxygen & among other water quality parameters (Turbidity, BOD, pH, PO_4, NO_3, F, Fe, AmmN) which have significant and separate effect on the dependent variable. Result have shown that the sagar lake contain high concentration of nitrate and phosphate, which led to quick growth of algae and as well as death plants.
Saleem et al. (2012)developed a regression model for physic-chemical quality of ground water of Gulbarga-India; they tested water quality parameters for 150 bore well water sample & developed a regression for nitrate by using SPSS software as analytic tool; they revealed that 70% variation in nitrate by Na+ & Ca++.
In the present investigation, an attempt has been made to develop statistical model to predict Nitrate concentration using routinely monitored water quality parameters of upper lake, Bhopal (India) by the analytical tool SPSS (statistical package for the social science).

II. MATERIALS AND METHODS

The Upper Lake Bhopal has been selectedfor present investigation. This lake is an east westerly elongated shallow lake with irregular margins.Catchment area of the upper Lake Bhopal is about 372 km2. Upper Lake is 10.6 km while the width comes out to be 3.25 Km and the depth varies from 4 to 8 Meters. In order to develop the models detailed information is required on water quality parameters of upper Lake of Bhopal.The physico chemical analysis data of water samples collected in a month of years 2011-2015 from various locations of upper Lake Bhopal. State research laboratory-Bhopal analyzed various water quality parameters of upper lake like Temperature, turbidity, total solid (TS), dissolved solid (DS), suspended solid (SS), sulphate (SO_4), chloride (Cl), dissolved oxygen(DO), biochemical oxygen demand(BOD), chemical oxygen demand(COD), pH, Alkalinity (A), total hardness (TH), Ca-hardness, Mg hardness, potassium (K), magnesium (Mg), sodium (Na) and total coliform at various location of upper lake namely Yatch club, kamla park, Karbala, Bairagarh and Khanugaon. These secondary collected data was used for the statistical analysis. The best subset analysis was carried out for full data sets. SPSS 22.0 was used for carrying out the analysis.

Formulation of regression model

The general representation of statistical models is given by

$$Y_i=\sum^{k}_{j=0}\beta_j x_{ij}+e$$

With $x_{i0}= 1$;

Where x_{ij} is the independent variable for thei^{th} observation (various water quality constituents), Y_i is the dependent variable for the i^{th} observation, β_j are the unknown coefficients to be estimated, k+1 are number of coefficients to be estimated in the model and e is the error in the determination of Y_i which is generally assumed as having zero mean and constant standard deviation.
Before carrying out a statistical regression of the data set, preliminary analysis of data was done. This consists of initial filtration of data, partial visual inspection of data files and the creation of the scatters plots. Once the indentified input errors were removed, a general regression analysis between independent variables and dependent variable is formed. Relation between any two variable is explained by coefficient, commonly known as correlation coefficient (r), which describe the relation between the two variables. Square of correlation coefficient (R^2) indicate the contribution of independent parameter in explaining the variation in the dependent variables.Model is useful for the predictive purpose if it include many independent variables, so that reliable fitted value can be determined. We obviously want R^2 to be large, since R^2 gives the proportion of variation in the dependent variables; that is explained by the fitted regression model. On the other hand, because of the effort involved in the monitoring of a large number of independent variables, there is interest in include few independent variables. There is no unique statistical procedure for doing this (Draper and smith 1981). However many worker has suggested different statistical approach, such as all possible regression, backward elimination, forward elimination in stepwise regression, principal component regression and stage wise regression which may help in optimum model formulation (Draper and smith 1981; Weisberg, 1980). In the present case the best subset regression approach has been used to select the best set of independent variables.

Best subset regression

Different best subset of independent variables cab be selected using the proportion of variation explained in the dependent variable (that is R^2 information). Assessment of each subset was made on the priority of the value of R^2 achieved, F value and the number of observation used in developing in the model. The model obtained from the large data set and achieving higher values of R^2 and F values will always be preferred. There two criteria (R^2 and F values) which will used in model selection are briefly described by Weisberg (1980).

For formulation of model first we need to work on statistical analysis on water quality parameters. This statistical analysis has been performed using SPSS 22.0 version software package of IBM was used for executing both correlation and regression analysis in order to investigate and study relationship between two or more parameters, by calculating correlation coefficients between different pairs of parameters.

The correlation coefficient (r) measures the degree of association that exists between two variables, one taken as dependent variable and other is taken as independent variable. The greater the value of regression coefficient, the better is the fit and more useful the regression variables. Correlation is the mutual relationship between two variables. Direct correlation exists when increase or decrease in the value of one parameter is associated with a corresponding increase or decrease in the value of other parameter (K. Jothivenkatachalam, 2010).

III. RESULT AND DISCUSSION

Correlation analysis for Nitrate-Nitrogen

Regression analysis was conducted to investigate the individual relationship between Nitrate-Nitrogen and other physico-chemical properties. SPSS analytical tool used to carried out the correlation studies for Nitrate –Nitrogen, in which Nitrate-Nitrogen considered as dependent variable and other water quality parameters namely Turbidity, Total solid, pH, Dissolved solid (DS), suspended solid (SS), Dissolved oxygen (DO), Bio-chemical oxygen demand(BOD), potassium (K), sodium (Na), magnesium (Mg), alkalinity (A), calcium (Ca) and phosphate (PO_4) considered as independent variables. Total 124 observations (from April 2011 to March 2016) were used in this analysis.

Table 1: Individual contributions of various independent chemical variables to predict Nitrate-Nitrogen

Dependent variable	Independent variable	Coefficient of correlation (r)	Coefficient of variation (R^2)	Variance ratio(F)
NO_3-N	turbidity	0.109	0.012	1.218
NO_3-N	pH	0.064	0.004	0.501
NO_3-N	TS	.433	.188	28.225
NO_3-N	DS	.434	.188	28.269
NO_3-N	SS	.116	.013	1.66
NO_3-N	PO_4	.173	.03	2.574
NO_3-N	AmmN	.143	.020	1.587
NO_3-N	NO_2-N	.226	.051	6.305
NO_3-N	DO	.119	.014	1.643
NO_3-N	BOD	.072	.005	0.602
NO_3-N	K	0.031	.001	.088
NO_3-N	Na	.185	.034	.034
NO_3-N	Ca	.347	.121	16.202
NO_3-N	Mg	.261	.068	8.650

Selection of independent variables

It can be seen that Dissolved Solid (DS) is best single variable explaining 18.8 % variation in the concentration of Nitrate-N, where other water quality parameters namely TS, Ca, Mg, PO_4, AmmN, Na, Alkalinity if taken as independent variable explaining approximately 18.8%, 112.1%, 6.8%, 3%, 2%, 3.4%, 5.1%, respectively. In order to increase the R^2, various combination of water quality parameters with DS are attempted, However DS & TS both give same variation in nitrate which is 18.8%; but DS gives higher variance value (F) than TS; the model obtained from the large data set& achieving higher values of R^2& F values will always be preferred; describe by Weisberg (1980). To increase the R^2 value, other water quality parameters were included in the combination of DS and K. (DS+ K+ Ca) is selected as the best 3 parameter model. Similarly the 4-paramter model, 5-parameter model and 6-paramter model were selected which are (DS+ K+ Ca+ AmmN), (DS+ K+ Ca+ AmmN+ Alkalinity) and (DS+ K+ Ca+ AmmN+ Alkalinity+ PO_4), respectively. Various selected model are given in Table 2. Now to choose the best model among the various models, one has to keep in mind that the selected model should have minimum number of explaining variables and maximum R^2

value. There is a compromise between the two criteria. The decision can be made on the basis of the value of F statistics as explained by Weisberg (1980).

Table 2: various combination of model with their statistics

No. of variables	**Set of independent variables**	**R^2**	**F**
2	DS+PO_4	0.122	5.675
	DS+ Turbidity	0.252	16.994
	DS+ pH	0.195	14.683
	DS+ SS	0.196	14.773
	DS+ AmmN	0.179	8.173
	DS+ DO	0.187	13.135
	DS+ BOD	0.195	13.675
	DS+ K	**0.342**	**23.416**
	DS+ Na	0.342	22.396
	DS+ Ca	0.238	18.250
	DS+ Mg	0.235	18.017
	DS+ Alkalinity	0.236	17.942
3	DS+K+ PO_4	0.345	10.163
	DS+K+ Turbidity	0.443	20.417
	DS+K+ pH	0.351	16.010
	DS+ K+ SS	0.348	15.825
	DS+K+ AmmN	0.345	8.617
	DS+K+ DO	0.401	18.280
	DS+K+ BOD	0.367	15.663
	DS+K+ Na	0.351	15.317
	DS+K+ Ca	**0.469**	**25.007**
	DS+K+ Mg	0.443	22.575
	DS+K+ Alkalinity	0.433	21.657
4	DS+K+ Ca+ PO_4	0.367	8.273
	DS+K+ Ca+ Turbidity	0.471	16.904
	DS+K+ Ca+ pH	0.476	19.062
	DS+K+ Ca+ SS	0.476	19.078
	DS+K+ Ca+ AmmN	**0.535**	**12.632**
	DS+K+ Ca+ DO	0.430	15.254
	DS+K+ Ca+ BOD	0.485	17.898
	DS+K+ Ca+ Na	0.487	18.971
	DS+K+ Ca+ Mg	0.469	18.544
	DS+K+ Ca+ Alkalinity	0.469	18.547
5	DS+K+ Ca+ AmmN+ PO_4	0.530	6.082
	DS+K+ Ca+ AmmN+ Turbidity	0.533	9.361
	DS+K+ Ca+ AmmN+ Ph	0.535	9.896
	DS+K+ Ca+ AmmN+ SS	0.538	10.018
	DS+K+ Ca+ AmmN+ DO	0.439	6.420
	DS+K+ Ca+ AmmN+ BOD	0.544	10.272
	DS+K+ Ca+ AmmN+ Na	0.547	9.885
	DS+K+ Ca+ AmmN+ Mg	0.543	10.208
	DS+K+ Ca+ AmmN+ Alkalinity	**0.563**	**11.097**
6	**DS+K+ Ca+ AmmN+ Alkalinity+ PO_4**	**0.741**	**12.403**
	DS+K+ Ca+ AmmN+ Alkalinity+ Turbidity	0.575	9.005
	DS+K+ Ca+ AmmN+ Alkalinity+ SS	0.567	9.152
	DS+K+ Ca+ AmmN+ Alkalinity+ pH	0.566	9.145
	DS+K+ Ca+ AmmN+ Alkalinity+ DO	0.471	5.933

	DS+K+ Ca+ AmmN+ Alkalinity+ BOD	0.569	9.240
	DS+K+ Ca+ AmmN+ Alkalinity+ Na	0.565	8.603
	DS+K+ Ca+ AmmN+ Alkalinity+ Mg	0.582	9.765
7	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ Turbidity	0.755	10.559
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ pH	**0.812**	**15.431**
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ SS	0.743	10.349
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ DO	0.741	9.813
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ BOD	0.744	10.365
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ Na	0.754	10.064
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ Mg	0.767	11.788
8	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ pH+ Turbidity	0.861	17.859
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ pH+ SS	0.824	14.038
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ pH+ DO	0.812	12.414
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ pH+ BOD	0.813	13.03
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ **pH+ Na**	**0.873**	**18.947**
	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ pH+ Mg	0.824	14.085

Table 3: selected subset of independent variables for possible model development

No. of Variable	Set of Independent Variable	n	R^2	F
8	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ pH+ Na	124	0.873	18.947
7	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4+ pH	124	0.812	15.431
6	DS+K+ Ca+ AmmN+ Alkalinity+ PO_4	124	0.741	12.403
5	DS+K+ Ca+ AmmN+ Alkalinity	124	0.563	11.097
4	DS+K+ Ca+ AmmN	124	0.535	12.632
3	DS+K+ Ca	124	0.469	25.007
2	DS+K	124	0.342	23.416
1	DS	124	0.188	28.269

Model Result

As the regression analysis applied on the values entered in SPSS software and result are shown in table 5. We developed linear model for Nitrate-N using multiple linear regression with the value of R^2 is 87.3 % (87.3 percent variation in Nitrate-N concentration is explained by selected independent variables) as shown in Table 4. Table shows the result of MLR for Nitrate-N, in which Coefficient part of the output gives us the values that we need in order to write the regression equation. So model equation of predictive model for Nitrate-N is given below-

$$NO_3\text{-}N = -34.543 + 0.024\ DS - 0.054K + 0.124\ Ca - 2.976\ AmmN + 0.044\ Alkalinity + 2.386\ PO4 + 2.730\ pH + 0.011\ Na \quad \text{Eq. (1)}$$

Water qualities parameters are of great significance to explaining the variation in concentration of Nitrate-N.variation in Nitrate concentration is explained by the DS, K, Ca, AmmN, Alkalinity, PO4, pH and Na while other 12.3 % is influenced by other factorsThree parameters Dissolved solid (DS), Potassium (K) and Calcium (Ca) explain the nearly 47% variation in Nitrate concentration.

Table 4: Model summary for Nitrate-N linear model with ANOVA result

R-Squared	Adj. R-Squared	F	Significance
0.873	0.827	18.647	0.000

The significance of the regression coefficient of each independent variable in a regression equation is indicated by the t test statistic. The value of t is calculated by dividing the regression coefficient by its standard error of the regression coefficient. Result of t value of each coefficient or term can see in Table 5, in which p value of each term is less than 0.05. As the p value of each factor and constant is less than 0.05 so it evident that in this model the coefficient for the intercept and other terms are significantly different from zero, so they affect the dependent variable. If a coefficient is different from zero then it has a genuine effect on the dependent variables.

Table 5: Result of statistics for model parameters

Model Parameters	Coefficient	Standard error	t-stat	p-value
Constant	-34.543	5.949	-5.723	.000
DS	0.024	.004	6.629	.000
K	-0.054	0.038	-1.424	.048
Ca	.124	.031	4.022	.001
AmmN	-2.976	.484	-6.144	.000
Alkalinity	.044	.006	6.849	.000
PO_4	2.386	.510	4.676	.000
pH	2.730	.600	4.552	.000
Na	.011	.039	.275	.036

Model Performance

The performance of the model has been tested using standard error, correlation coefficient, ANOVA test and T-test. For this first we need to determine the predictive values of Nitrate-N by using the model equation given in equation (1). For validation of the above model, the physico-chemical data of the month ofOctober 2016 was used. It can be seen that the statistical models developed in this study perform well in computing the Nitrate-N concentration for the lake. The Nitrate-Nitrogen (NO_3-N) value obtained by the developed regression model indicates statistically significant regression.Scatter plot between actual Nitrate-N and Predicted Nitrate-N is shown in Fig.1 with the R^2 value of 0.859 which suggest good agreement in observed and predicted Nitrate-N concentration for the lake.

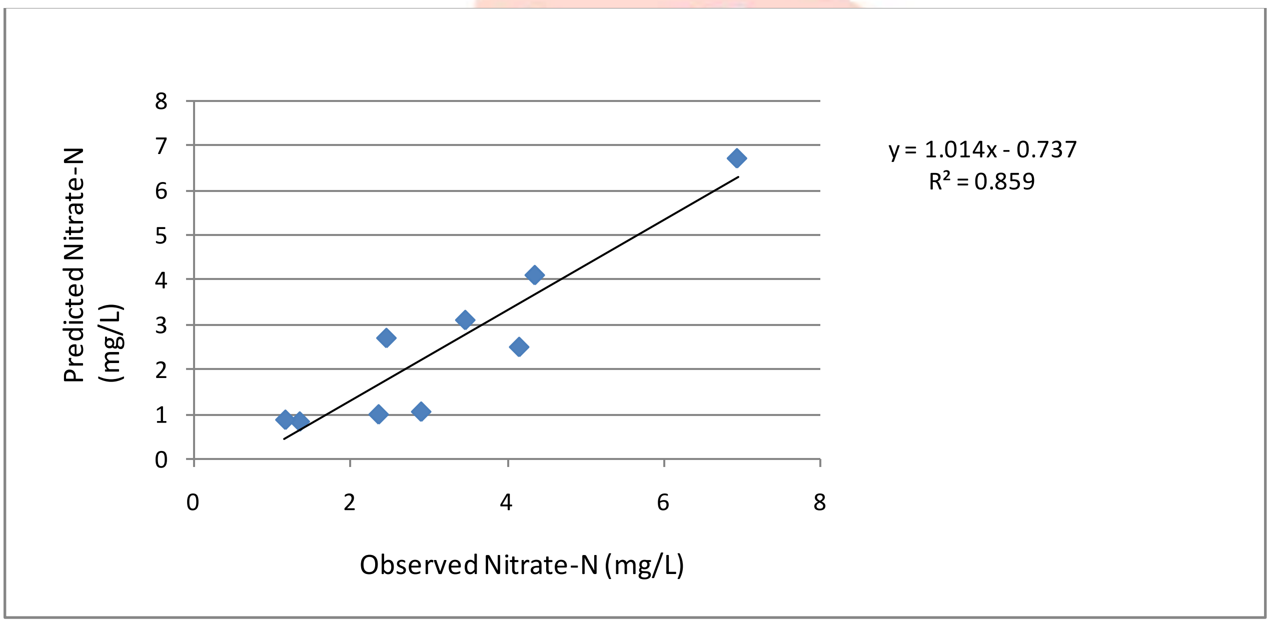

Fig. 1: scatter plot between actual (observed) NO_3-N with predicted NO_3-N concentration for lake

Standard error has been determined using the following equation

$$SE= \left[\sum_{i=1}^{N} \frac{(K_P-K_M)^2}{} \right]$$

Where N is the number of observation and K_P and K_M are the predicted and measured values. The SE value was found to be very low 0.83 for the year 2016. Low value of standard error indicates that the data points tend to be close to the mean or deviation of residual value is not too much from each other. Residual is the difference between the observed value and the estimated value.

IV. CONCLUSION

The statistical model developed in this study is successfully explaining the 87.8 % variation in Nitrate-Nitrogen concentration. R- square value obtained for the developed linear prediction model for Nitrate is 87.3 %; this shows that of the variation in Nitrate concentration is explained by the variation in DS, K, Ca, AmmN, Alkalinity, PO_4, pH and Na while other 12.3 % is influenced by other factors. Potassium is present in plant material and is lost from agriculture soil by crop harvesting in the surrounding area of the lake, both K and Phosphorous used in NPK fertilizer with ratio of 16:4:8 which removal and leaching resulting Eutrophication of lake. Ca used in fertilizer as macro nutrient so it also influences the amount of nitrogen by made organic compound Ca (NO_2). Ammonium based fertilizer used in agriculture field which also resulting into Eutrophication. Alkalinity comes from alkali which refers to ionic compounds (salts) containing alkali metal or alkaline earth metal elements (Na or K), these alkali metal made compound with Nitrate after combination that is sodium Nitrate ($NaNO_3$) and Potassium Nitrate (KNO_3). As the lake becomes enriched in anthropogenic CO_2, the resulting decrease in pH could lead to decreasing rates of nitrification (Ammonium compound converted into Nitrate). The performance of the model was tested by standard error test which is found very low and correlation coefficient analysis which is describes good correlation between actual and predicted value of nitrate concentration.

The developed model may be used for the prediction of missing observed value of Nitrate-Nitrogen concentration in upper Lake Bhopal.

ACKNOWLEDGEMENT

We highly indebted to my praiseworthy supervisor Prof. Harsangeet Kaur (Associate Professor) Department of Civil Engineering MANIIT Bhopal for her precious guidance, intellectual stimulation continuing and sustained encouragement, untiring help and his unflinching interest with great love and affection throughout the present paper.

REFERENCES

Draper, N. R. and Smith, H. (1981), Applied regression analysis, Second edition, John Wiley and Sons, New York

Jothivenkatachalam, K., Nithya, A., and Chandra Mohan, S. (2010). "Correlation analysis of drinking water quality in and around Perur block of Coimbatore District, Tamil Nadu, India" Rasayan Journal Chemistry, 3(4), pp 649654.

Muniz, K., Cruzado, A. and De Villa, C.R., (2001). Statistical analysis of nutrient data quality (nitrate and phosphate), applied to useful predictor models in the northwestern Mediterranean Sea. BOLETIN-INSTITUTO ESPANOL DEOCEANOGRAFIA, 17(3/4), pp.221-232

Official website of statistics solution available on http://www.statisticssolutions.com

Pathak, H**.,** (2012). Evaluation of ground water quality using multiple linear regression and mathematical equation modeling, Annals of the University of Oradea—Geography Series, 2, pp.304-307

Saleem, A., Dandigi, M.N. and Kumar, K.V. (2012). Correlation-regression model for Physico-chemical quality of groundwater in the South Indian city of Gulbarga, African Journal of Environmental Science and Technology, 6(9), pp.353-364

Sharma, M.K., Choubey, V.K. and Dwivedi, V.K., (2007). Phosphorous Modeling of Upper Bhopal Lake Using Best Subset Procedure in Proceedings of Taal2007: The 12th World Lake Conference (Vol. 851, p. 856)

Singh R. and Mahajan I. (1987) : Phytoplankton & water chemistry of Rawalsar & Renuka lake. H.P. India. J. Ecol. 14(2): 273-277.

Weisberg, S. (1980). Applied linear regression, John Wiley and Sons, New York, NY