REGRESSION AND NEURAL NETWORK MODELLING FOR RIVER YAMUNA

Mrs. Maninder Kaur[1], Ms. Tarushi Singh[2], Rishabh Bhargava[3], Akshay Yadav[4], Mohanil Kataria[5], Shubham Gautam[6]

[1]*Head of department, Civil Engineering Department, Northern India Engineering College*
[2]*Assistant Professor, Civil Engineering Department, Northern India Engineering College*
[3,4,5,6,]*Students, Civil Engineering Department, Northern India Engineering College.*

---***---

Abstract: *Water is the most used resource of planet earth. It is no secret that every human application depends on it in one way or the other. However, the quality of water is being degraded every year due to rapid industrialization and urbanization by man in his quest for development and also due to natural activities. Water Pollution in India is a major problem which calls for special attention by the state water departments and the local citizens. In this Research Paper, we aim to establish a regression co-relation between the chemical parameters affecting the water quality standards using various water modelling techniques like Artificial Neural Network(ANN) , Multi linear regression (MLR) and Factor Analysis (FA) and analyze and compare the results given by them.*

Key Words: ANN, MLR, FA, Regression, Chemical parameters.

1. INTRODUCTION

Water pollution of the river Yamuna is a serious setback for the city of Delhi, which with its ever growing population of more than 15 million depends on the quality of water for various purposes such as drinking, washing, cleaning, flushing etc. The chemical parameters that we have considered for our research are pH of the water, BOD (Biological Oxygen Demand), Total Coliform Bacteria of water, Dissolved Oxygen and COD (Chemical Oxygen Demand). We obtained the raw values of the chemical parameters for a duration of 31 months from January 2014 to July 2016 from the Central Pollution Control Board (CPCB) for 5 stations of the Yamuna, namely, Palla, Nizamuddin, Agra Canal, Okhla after meeting Shahdra Drain and Agra Canal at Madanpur Khadar. With the advent of technology and Computer Science, we now have softwares which have the ability to predict the regression co-relation of these chemical parameters with each other. We have used, Factor Analysis, ANN and MLR for this purpose. By performing the Factor Analysis we expressed values of the raw data as a function of a number of possible causes in order to find which chemical parameter is most important. We came to the observation that the most affected chemical parameters were DO and BOD. So, we kept the DO and BOD as single independent variables alternatively, and performed regression on them. Regression basically tells us how accurate our co-relation formed from these operations are.

2. OBJECTIVE

- To determine the variance in the data of yamuna river from different stations using multivariance regression and artificial neural network.
- To find out the similarities and dissimilarities within the different blocks of our study area based on water quality using artificial neural network.

3. Methodology

The necessity of building a model using the above mentioned three techniques is to differentiate between a precise format which would be able to predict the values of the pollutant concentration, with inputs being the meteorological parameters, in such a way so that the role of these parameters on water quality can be understood and the error between the predicted and the observed values is minimum using regression analysis. The collected data for all the five stations was preprocessed i.e. normalized in the range of 0.1-0.8. This normalization is done to tone down the input values in a specific range. Normalization is done to obtain precise results on the data set.

3.1 P-Value Test

The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true – the definition of 'extreme' depends upon how the hypothesis is being tested. . P is also described in terms of rejecting H0 when it is actually true, however, it is not a direct probability of this state. The null hypothesis is usually a hypothesis of "no difference". The only situation in which one should use a one sided P value is when a large change in an unexpected direction would have absolutely no relevance to the study, which is unusual. The term significance level (alpha) is used to refer to a pre-chosen probability and the term "P value" is used to indicate a probability that is calculated after a given study.

3.2. F.A. (Factor Analysis)

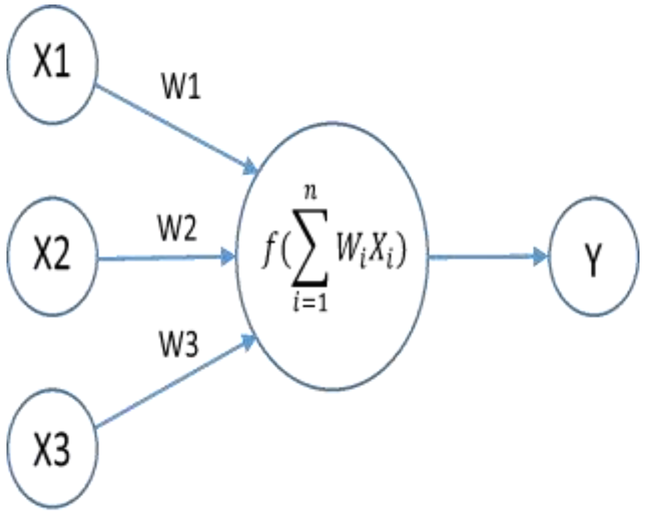
Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called **factors**. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved (underlying) variables. Factor analysis aims to find independent latent variables. In other words, using Factor Analysis we expressed values of the raw data as a function of a number of possible causes in order to find which chemical parameter or which 'factor' is most important. We came to the observation that the most critical chemical parameters were DO and BOD. So, we kept the DO and BOD as single independent variables alternatively, and performed regression on them.

4. Multiple Linear Regression (MLR)

MLR is a method used to model the linear relationship between a dependent variable (pollutant concentration in our case) and one or more independent variables (meteorological inputs in our case). The statistical technique is used to measure the relationship between the dependent variable and the independent variable up to a precision level where the predicted values are brought close to the observed ones at 0.05 significance level and 95% confidence interval.

5. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) are biologically inspired parallel computational models. They consists of simple highly inter connected processing elements which processes the input similar to human brain (Rumelhart and McClelland, 1986). ANN acquires, represents and computes mapping from one multivariate space to another (Wassermann, 1993) using self-learning. ANN has been used in almost all the fields of Civil Engineering. Here, in this fig, X1, X2, X3 are the inputs and W1,W2,W3 are the weights which have a threshold value to generate the signal and give us a single output Y.

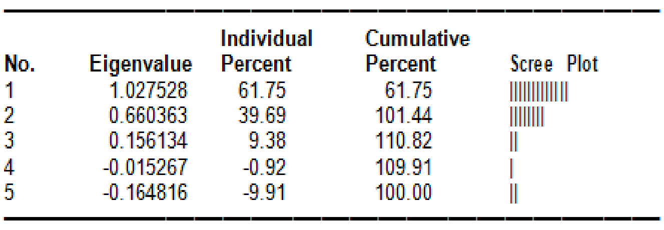


6. RESULTS AND DISCUSSIONS

FA RESULT

STATION 1:- Yamuna River at Palla

Eigenvalues after Varimax Rotation

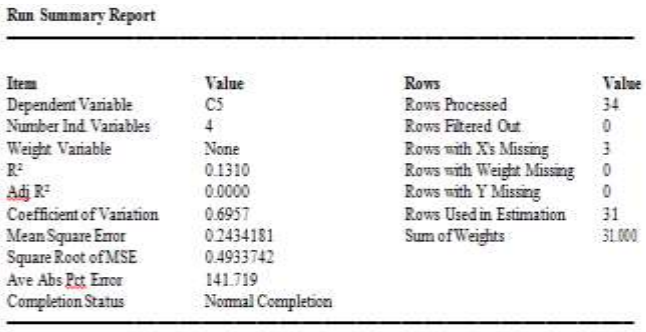


No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	1.027528	61.75	61.75	|||||||||||||
2	0.660363	39.69	101.44	||||||||
3	0.156134	9.38	110.82	||
4	-0.015267	-0.92	109.91	|
5	-0.164816	-9.91	100.00	||

MLR RESULT

BOD

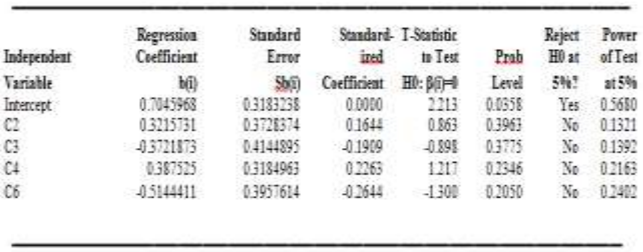
STATION 1:- Yamuna River at Palla



Run Summary Report

Item	Value	Rows	Value
Dependent Variable	C5	Rows Processed	34
Number Ind. Variables	4	Rows Filtered Out	0
Weight Variable	None	Rows with X's Missing	3
R^2	0.1310	Rows with Weight Missing	0
Adj R^2	0.0000	Rows with Y Missing	0
Coefficient of Variation	0.6957	Rows Used in Estimation	31
Mean Square Error	0.2434181	Sum of Weights	31.000
Square Root of MSE	0.4933742		
Ave Abs Pct Error	141.719		
Completion Status	Normal Completion		

Run Summary Report

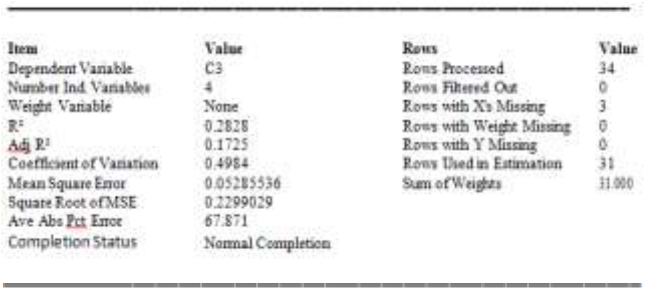


Independent	Regression	Standard	Standard-	T-Statistic		Reject	Power
	Coefficient	Error	ized	to Test	Prob	H0 at	of Test
Variable	b(i)	Sb(i)	Coefficient	H0: β(i)=0	Level	5%?	at 5%
Intercept	0.7045968	0.3183238	0.0000	2.213	0.0358	Yes	0.5680
C2	0.3215731	0.3728374	0.1644	0.863	0.3963	No	0.1321
C3	-0.3721873	0.4144895	-0.1909	-0.898	0.3775	No	0.1392
C4	0.387525	0.3184963	0.2263	1.217	0.2346	No	0.2163
C6	-0.5144411	0.3957614	-0.2644	-1.300	0.2050	No	0.2402

DO REPORT

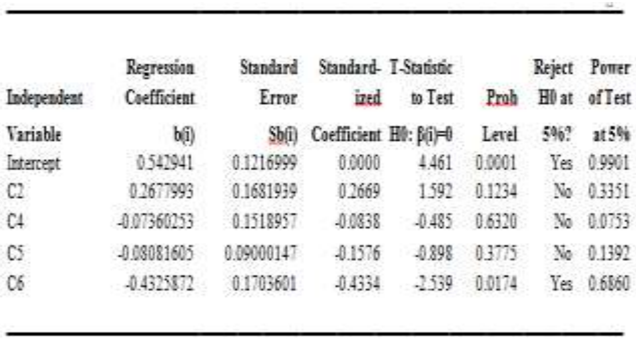
STATION 1:- Yamuna River at Palla

Run Summary Report



Item	Value	Rows	Value
Dependent Variable	C3	Rows Processed	34
Number Ind. Variables	4	Rows Filtered Out	0
Weight Variable	None	Rows with X's Missing	3
R^2	0.2828	Rows with Weight Missing	0
Adj R^2	0.1725	Rows with Y Missing	0
Coefficient of Variation	0.4984	Rows Used in Estimation	31
Mean Square Error	0.05285536	Sum of Weights	31.000
Square Root of MSE	0.2299029		
Ave Abs Pct Error	67.871		
Completion Status	Normal Completion		

Regression Coefficients T-Tests

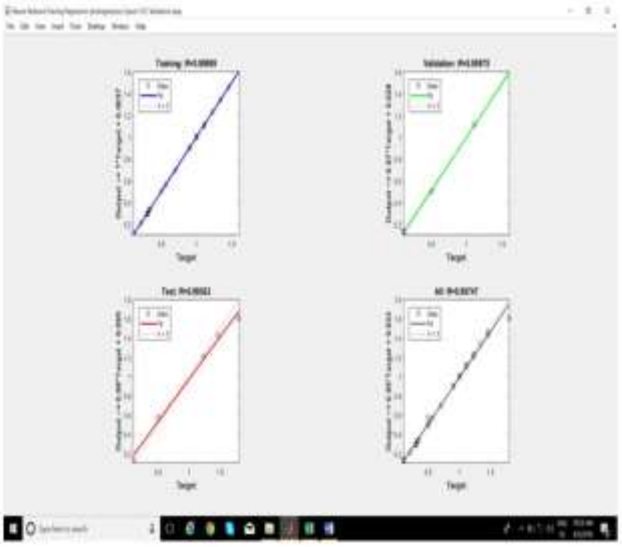


Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Standard- ized Coefficient	T-Statistic to Test H0: β(i)=0	Prob Level	Reject H0 at 5%?	Power of Test at 5%
Intercept	0.542941	0.1216999	0.0000	4.461	0.0001	Yes	0.9901
C2	0.2677993	0.1681939	0.2669	1.592	0.1234	No	0.3351
C4	-0.07360253	0.1518957	-0.0838	-0.485	0.6320	No	0.0753
C5	-0.08081605	0.09000147	-0.1576	-0.898	0.3775	No	0.1392
C6	-0.4325872	0.1703601	-0.4334	-2.539	0.0174	Yes	0.6860

ANN RESULT

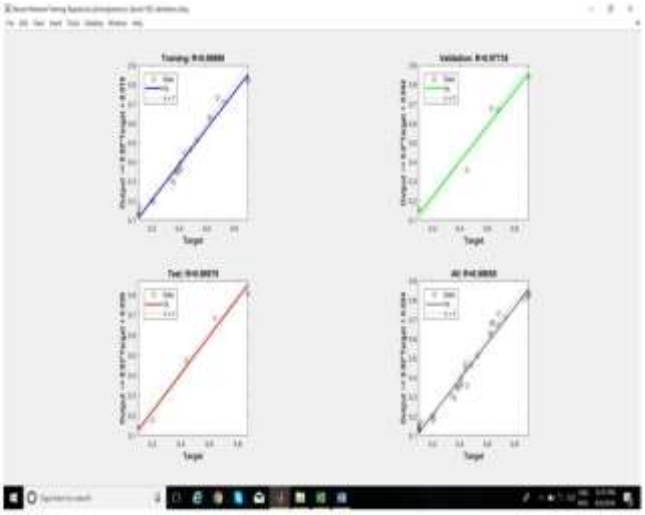
BOD

STATION 1:- Yamuna River at Palla



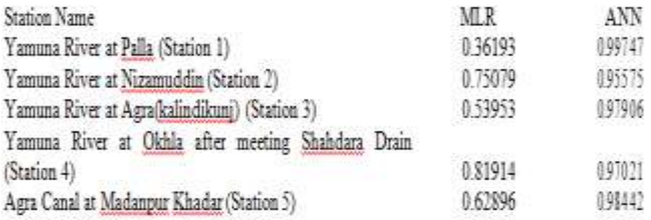
DO RESULTS

STATION 1:- Yamuna River at Palla



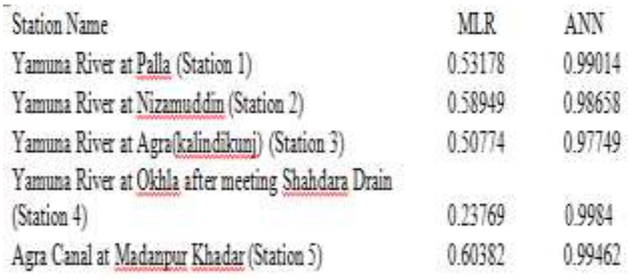
COMPARISION BETWEEN MLR ANN THROUGH R VALUES

BOD R VALUE

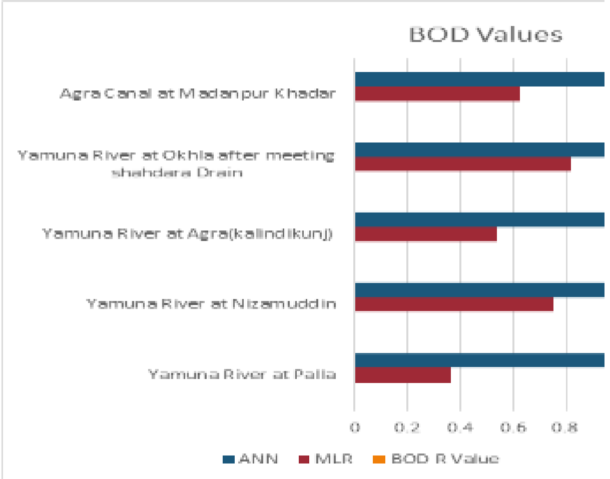


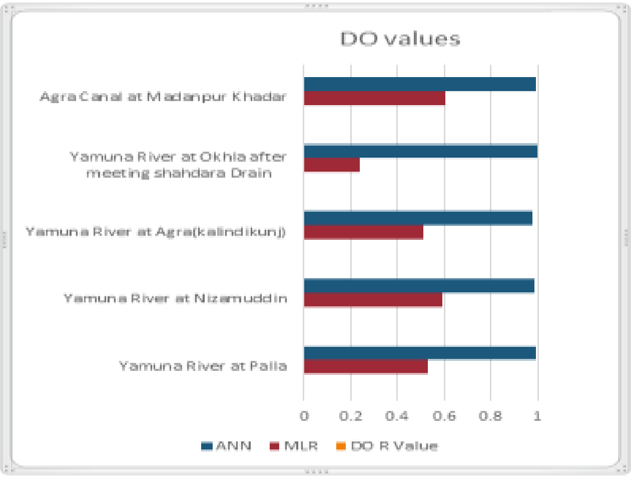
Station Name	MLR	ANN
Yamuna River at Palla (Station 1)	0.36193	0.99747
Yamuna River at Nizamuddin (Station 2)	0.75079	0.95575
Yamuna River at Agra(kalindikunj) (Station 3)	0.53953	0.97906
Yamuna River at Okhla after meeting Shahdara Drain (Station 4)	0.81914	0.97021
Agra Canal at Madanpur Khadar (Station 5)	0.62896	0.98442

DO R VALUE



Station Name	MLR	ANN
Yamuna River at Palla (Station 1)	0.53178	0.99014
Yamuna River at Nizamuddin (Station 2)	0.58949	0.98658
Yamuna River at Agra(kalindikunj) (Station 3)	0.50774	0.97749
Yamuna River at Okhla after meeting Shahdara Drain (Station 4)	0.23769	0.9984
Agra Canal at Madanpur Khadar (Station 5)	0.60382	0.99462





7. SUMMARY AND CONCLUSION

In this study two variables, were modelled using three different statistical techniques where, models obtained from neural network shows maximum correlation with the observed Values for all the five air monitoring stations. The data was divided as 70% for training and 30% for testing. The overall R value was found to be maximum in case of neural network, as high as 0.95, explaining the relationship between dependent and independent variables to utmost precision. Though there are numerous factors that affect performance of neural networks like lack of data or over training. The data collected from the central pollution control board (CPCB) is authentic but sometimes the readings are missed due to improper functioning of pollutant measuring instrument thus, proper care should be taken while computing the pollutant. Also, the meteorological measurements are taken at a different station but should be taken on the same station used for monitoring air quality, thereby reducing the chances of error. This study focusses primarily on the two variables where as if more variables like, ph, cod, temp, are also measured, predictions can be improved. Modelling approach serve a baseline for future research.

8. REFERENCES

Research Papers

1) Chang Shu, Donald H. Burn, et-al, "Artificial Neural Network Ensembles and Their Application in Pooled Flood Frequency Analysis" published on 4th September 2004.

2) Praveen Kumar, Pooja Sharma, et-al, "Artificial Neural Networks-A Study" International Journal of Emerging Engineering Research and Technology Volume 2, Issue 2, May 2014.

3) Behzad Saeedi Razavi, et-al, "Predicting the Trend of Land Use Changes Using Artificial Neural Network and Markov Chain Model (Case Study: Kermanshah City)" ISSN:2041-049 Published on April 20, 2014.

4) Xander Olsthoorn, et-al, "Cost Benefit Analysis of European Air Quality Targets for Sulphur Dioxide, Nitrogen Dioxide and Fine and Suspended Particulate Matter in Cities" Volume 14, Issue 3, published in October 1999.

5) Gabriel Ibarra-Berastegi, et-al, "Assessing spatial variability of SO2 field as detected by an air quality network using Self-Organizing Maps, cluster, and Principal Component Analysis" Published on 10th August 2009.