Journal of Science and Technology
ISSN: 2456-5660 Volume 5, Issue 4, July-August 2020

www.jst.org.in *DOI: https://doi.org/10.46243/jst.2020.v5.i4.pp268-275*

Analysis of Different Water Quality Parameters of Ganga River by Multivariate Tools

Smita Jain
(Associate Professor, Department of Mathematics, JECRC University)
Corresponding Author: smitajain.maths@gmail.com

To Cite this Article
Smita Jain, "Analysis of Different Water Quality Parameters of Ganga River by Multivariate Tools", Journal of Science and Technology, Vol. 05, Issue 04, July-August 2020, pp268-275

Article Info
Received: 07-04-2020 Revised: 10-07-2020 Accepted: 12-07-2020 Published: 14-07-2020

Abstract: *This Study Statistically analyzes the deteriorating water quality of the River Ganga. Statistical techniques such as Water Quality index (WQI), Cluster Analysis, Best Subsets Regression and Multiple Regression Analysis were applied to seven water quality parameters, collected from 21 sampling Stations in India. Water Quality Index identified the most polluted stations that are Kadaghat, Allahabad, Khurgi, Patna U/S, Bihar, Varanasi D/S (Malviya Bridge), U.P, Indrapuri, Dehri and Varanasi U/S (ASSIGHAT), U.P. Cluster Analysis for the different Stations showed a similarity of 99.99% between the stations Ganga D/S, Mirzapur , Varanasi D/S (Malviya Bridge) and Varanasi U/S (Assighat), U.P. Cluster Analysis for variables showed a 98.96% similarity of parameter BOD with WQI and 96.06% similarity between the parameters Total Coliform and Fecal Coliform. After applied the Best Subset Regression Analysis we get the highest Mallow c-p value with high R^2 for the parameters BOD, Nitrate, Total Coliform and Fecal Coliform. In the Regression analysis the p value for the estimated coefficients of BOD is 0.00, indicates that BOD is significantly related to WQI.In this paper we conclude that BOD is the most critical parameter and we study the comparison of water quality of river Ganga for different stations.*

Keywords*: Water quality parameters, WQI, Cluster Analysis, Regression analysis, Best subset Regression and Dendograms*

I. Introduction

River water quality is of great environmental concern since it is one of the major available fresh water resources for human consumption. Throughout the history of human civilization, rivers have always been heavily exposed to pollution, due to their easy accessibility to disposal of wastes. However, after the industrial revolution the carrying capacity of the rivers to process wastes reduced tremendously. Anthropogenic activities such as urban, industrial, and agricultural as well as natural processes, such as precipitation inputs, erosion, and weathering of crustal materials affect river water quality and determine its use for various purposes. The usage also depends upon the linkages (channels) in the river system, as inland waterways play a major role in the assimilation and transportation of contaminants from a number of sources. Besides linkages, the seasonal variation in precipitation, surface runoff, interflow, and groundwater flow, and pumped in and out flows also have a strong effect on the concentration of pollutants in rivers. In view of the limited stock of freshwater worldwide and the role that anthropogenic activities play in the deterioration of water quality, the protection of these water resources has been given topmost priority in the 21st century. Research-wise, one of the important stages in the protection and conservation of these resources is the spatiotemporal analysis of water and sediment quality of the aquatic systems. The nonlinear nature of environmental data makes spatio-temporal variations of water quality often difficult to interpret and for this reason statistical approaches are used for providing representative and reliable analysis of the water quality.

Water Quality Index may be defined as a rating, reflecting the composite influence on overall quality characteristics of water of individual quality parameters, which is being regarded as one of the most effective way to communicate water quality. This will give us only certain numerical value but for estimating exact quality of water an indexing system has been developed known as Water Quality Index, which gives us the idea of water of whole system.

Quality rating equal to zero means a complete absence of pollutants. While 0 to 100 implies that the pollutants are under consideration means within the prescribed standards. When quality rating is more than 100 it implies that the pollutants are above the standards. Similarly when WQI is 0 to 100 indicate that the water is under consideration and is safe for use. Multivariate statistical techniques such as cluster analysis (CA) and best subset regression analysis have been widely used as unbiased methods in analysis of water quality data for drawing out meaningful conclusions. In this paper we present a methodology for examining the impact of all the sources of pollution in Ganga River (Uttar Pradesh) and to identify the parameters responsible for spatiotemporal variability in water quality using CA.

II. Study Area

The **Ganges** or **Ganga** (Hindustani: [ˈɡəŋgaː]), is a trans-boundary river of the Indian subcontinent which flows through the nations of India and Bangladesh. The 2,525 km (1,569 mi) river rises in the western Himalayas in the Indian state of Uttarakhand, and flows south and east through the Gangetic Plain of North India. After entering West Bengal, it divides into two rivers: the Hooghly and the Padma River. The Hooghly, or Adi Ganga, flows through several districts of West Bengal and into the Bay of Bengal near Sagar Island. The other, the Padma, also flows into and through Bangladesh, and joins the Meghna river which ultimately empties into the Bay of Bengal.

The Ganges is one of the most sacred rivers to Hindus. It is also a lifeline to millions of Indians who live along its course and depend on it for their daily needs. It is worshipped in Hinduism and personified as the goddess Gaṅgā. It has also been important historically, with many former provincial or imperial capitals (such as Kannauj, Kampilya, Kara, Prayag or Allahabad, Kashi, Pataliputra or Patna, Hajipur, Munger, Bhagalpur, Baranagar, Murshidabad, Bah arampur, Nabadwip, Saptagram and Kolkata) located on its banks.

The Ganges is highly polluted. Pollution threatens not only humans, but also more than 140 fish species, 90 amphibian species and the endangered Ganges river dolphin. The Ganges is a major source of global ocean plastic pollution. The levels of fecal coliform bacteria from human waste in the waters of the river near Varanasi are more than 100 times the Indian government's official limit. The Ganga Action Plan, an environmental initiative to clean up the river, has been a major failure thus far, due to rampant corruption, lack of will on behalf of the government and its bureaucracy, lack of technical expertise, poor environmental planning, and lack of support from religious authorities.

Sampling and Analysis

Seven Parameters pH, BOD, Conductivity, Nitrate, Fecal Coliform, Total Coliform and DO were selected in order to study the Physicochemical Characteristics of River Ganga at 21 stations of Utter Pradesh in India. The study based on secondary data collected from various relevant government departments, published and unpublished reports. Software Minitab is used to evaluate the Descriptive Statistics, WQI, Cluster Analysis and Best subset Regression Method for the different parameters.

Analysis For Water Quality Index:

The basic equation for Water Quality Index is W.Q.I. = $\sum$qi wi

Where qi = quality ratings of parameters
wi = unit weight of different parameters
qi wi = parameters sub-index

Quality Ratings Of The Parameters

Quality ratings $qi = \frac{[V_O - V_I]}{V_S - V_I} * 100$

Where V_O = observed value of the parameter
V_S = standard value of the parameter
V_I = ideal value of the parameter

V_I for pH = 7 and for dissolved oxygen = 14.6 mg/l
While V_I for Nitrate, Conductivity, B.O.D, Total Coliform and for Fecal Coliform is zero.

Unit Weight For Various Parameters

$Wi = K/ Si$

Where $K = 1 \div 1/vs_1 + 1/vs_2 + 1/vs_3 + \text{-------} 1/vs_n$

Si = recommended standard value for the corresponding Parameter

vs_i = standard value for the corresponding parameter

Analysis For Cluster Analysis

Cluster Analysis (CA). Cluster analysis is a multivariate statistical technique, which allows the assembling of objects based on their similarity. CA classifies objects, so that each object is similar to the others in the cluster with respect to a predetermined selection criterion. Bray-Curtis cluster analysis is the most common approach of CA, which provides intuitive similarity relationships between any one sample and the entire dataset and is typically illustrated by a dendrogram (tree diagram). The dendrogram provides a visual summary of the clustering processes, presenting a picture of the groups and their proximity with a dramatic reduction in dimensionality of the original data.

Best Subset Regression Analysis

Best subsets regression is an exploratory model building regression analysis. It compares all possible models that can be created based upon an identified set of predictors. The results presented for best subsets, by default in Minitab, show the two best models for one predictor, two predictors, three predictors, and so on for the number of possible predictors that were entered into the best subsets regression. The output in Minitab presents R^2, adjusted R^2, Mallow's C_p, and S. To determine the best model, these model fit statistics will be used in conjunction with one another. $R^2$and adjusted R^2measure the coefficient of multiple determination and are used to determine the amount of predictability of the criterion variable based upon the set of predictor variables. Mallow's C_p is a measure of bias or prediction error. This the square roots of the mean square error (MSE).

III. Results & Discussion

1. Water Quality Index

Station	**WQI**
GANGA AT GARHMUKTESHWAR, U.P	65.05185
GANGA AT KANNAUJ U/S (RAJGHAT), U	71.59335
GANGA AT KANNAUJ D/S, U.P	70.38797
GANGA AT BITHOOR (KANPUR), U.P.	79.01873
GANGA AT KANPUR U/S (RANIGHAT), U	52.36029
GANGA AT KANPUR D/S (JAJMAU PUMPING STATION), U.P	81.42684
GANGA AT DALMAU (RAI BAREILLY), U	88.4441
GANGA AT KALA KANKAR, RAEBARELI	86.8147
GANGA AT ALLAHABAD (RASOOLABAD), U.P.	94.94404
GANGA AT KADAGHAT, ALLAHABAD	**129.414**
GANGA AT ALLAHABAD D/S (SANGAM), U.P.	93.85674
GANGA U/S, VINDHYACHAL, MIRZAPUR	91.51511
GANGA D/S, MIRZAPUR	96.65601
GANGA AT VARANASI U/S (ASSIGHAT), U.P	
GANGA AT VARANASI D/S (MALVIYA BRIDGE), U	**107.7238**
GANGA AT TRIGHAT (GHAZIPUR), U.P	70.56398

GANGA AT BUXAR,BIHAR	71.11775
GANGA AT BUXAR, RAMREKHAGHAT	83.88488
GANGA AT KHURJI, PATNA U/S, BIHAR	117.6773
GANGA AT INDRAPURI, DEHRI ON SONI	107.1365
GANGA AT THE CONFLUENCE OF SONE RIVER DORIGANJ, CHAPRA	61.44192

2. Cluster Analysis

Cluster Analysis of Variables: St.1, St.2, St.3, St.4, St.5, St.6, St.7, ...

Correlation Coefficient Distance, Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		Number of obs. in new cluster	Number of obs. in new cluster
1	20	99.9982	0.000035	13	14	13	2
2	19	99.9966	0.000068	8	9	8	2
3	18	99.9956	0.000087	13	15	13	3
4	17	99.9942	0.000117	11	20	11	2
5	16	99.9929	0.000142	11	19	11	3
6	15	99.9927	0.000146	11	12	11	4
7	14	99.9777	0.000445	6	16	6	2
8	13	99.9662	0.000677	7	8	7	3
9	12	99.9661	0.000678	7	13	7	6
10	11	99.9630	0.000740	7	11	7	10
11	10	99.9624	0.000752	6	7	6	12
12	9	99.9598	0.000803	3	4	3	2
13	8	99.9442	0.001116	10	17	10	2
14	7	99.9319	0.001362	6	18	6	13
15	6	99.9159	0.001683	3	5	3	3
16	5	99.7967	0.004067	1	6	1	14
17	4	99.4089	0.011822	1	10	1	16
18	3	98.4392	0.031216	1	3	1	19
19	2	80.5615	0.388771	1	21	1	20
20	1	79.8814	0.402371	1	2	1	21

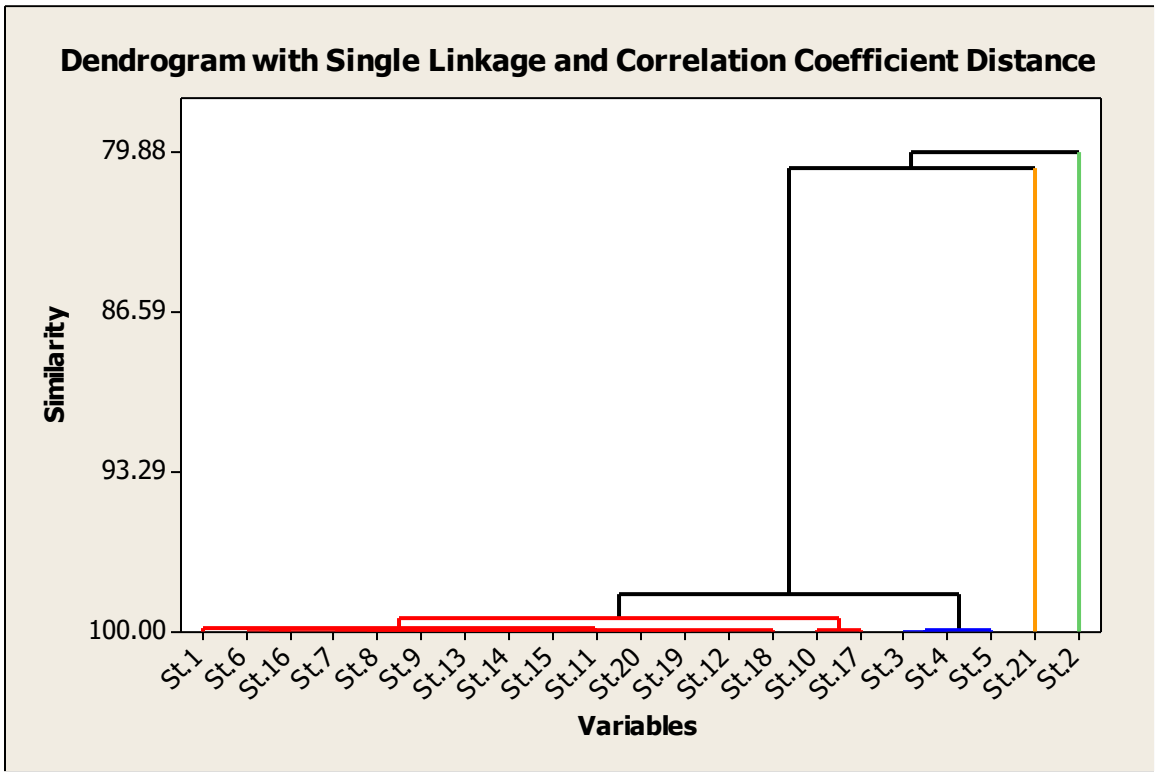
Final Partition

Cluster 1
St.1 St.6 St.7 St.8 St.9 St.10 St.11 St.12 St.13 St.14 St.15 St.16 St.17 St.18 St.19 St.20

Cluster 2
St.2

Cluster 3
St.3 St.4 St.5

Cluster 4
St.21



Cluster Analysis of Variables: DO, Ph, Cond., BOD, Nitrate, Fecal Coli, ...

Correlation Coefficient Distance, Single Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	7	98.9626	0.020748	4	8	4	2
2	6	96.0650	0.078701	6	7	6	2
3	5	93.7953	0.124095	4	6	4	4
4	4	89.7998	0.204005	3	4	3	5
5	3	83.7989	0.324022	2	3	2	6
6	2	71.8138	0.563724	2	5	2	7
7	1	69.5769	0.608462	1	2	1	8

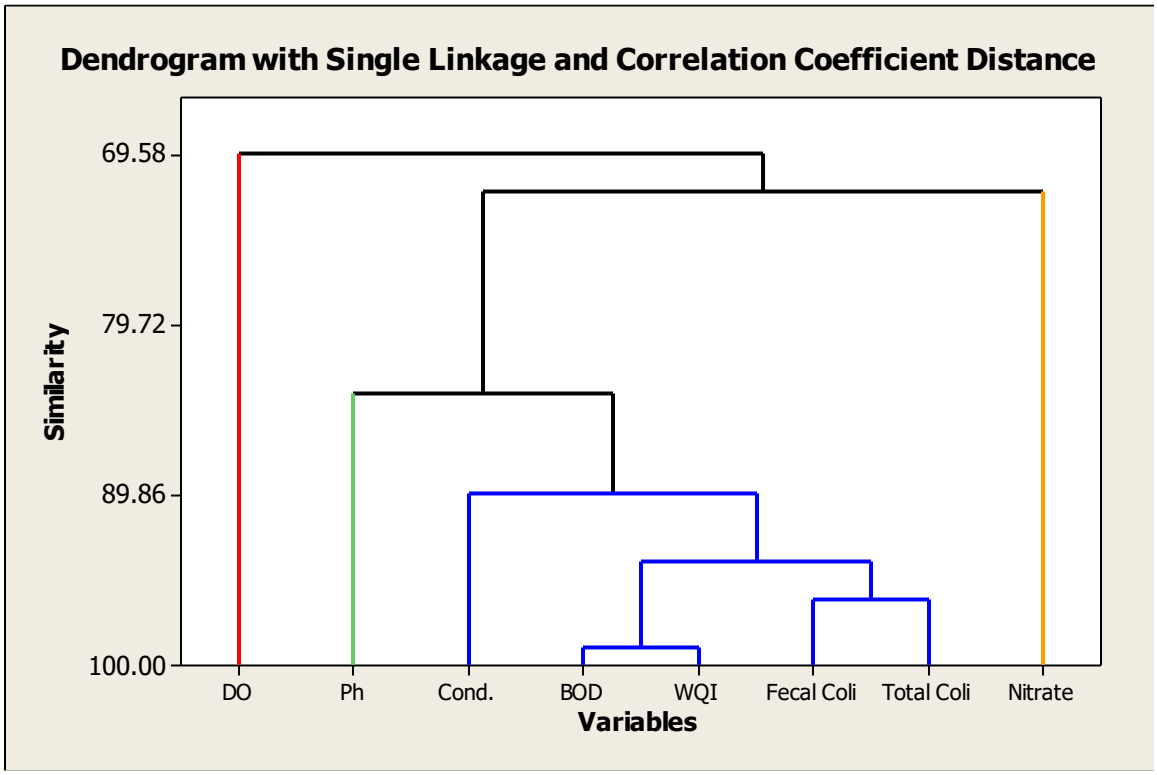
Final Partition

Cluster 1
DO

Cluster 2
Ph

Cluster 3
Cond. BOD Fecal Coli Total Coli WQI

Cluster 4
Nitrate



3. Best Subsets Regression: Stations versus DO, Ph, ...

Response is Stations

21 cases used, 1 cases contain missing values

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	DO	Ph	Cond.	BOD	WQI	Fecal Coli	Total Coli	Nitrate
1	46.5	43.7	0.5	4.6578		X						
1	22.5	18.4	8.3	5.6060						X		
2	48.6	42.9	1.8	4.6902	X	X						
2	47.2	41.3	2.3	4.7541		X		X				
3	51.2	42.5	3.0	4.7032		X		X				X
3	50.0	41.1	3.3	4.7603	X	X			X			
4	54.5	43.1	3.9	4.6815	X	X			X			X
4	52.8	41.0	4.4	4.7660	X	X		X				X
5	57.8	43.8	4.8	4.6521	X	X			X	X	X	X
5	55.9	41.2	5.4	4.7597	X	X		X	X			X
6	59.4	42.0	6.3	4.7249	X	X		X	X	X	X	X
6	58.1	40.1	6.7	4.8017	X	X	X		X	X	X	X

```
   7  60.2       38.8      8.0  4.8547  X X X X X X X
```

4. Regression Analysis: WQI versus BOD, Nitrate, Fecal Coli, Total Coli

The regression equation is

WQI = 24.6 + 18.9 BOD - 1.42 Nitrate + 0.000473 Fecal Coli - 0.000086 Total Coli

Predictor	Coef	SE Coef	T	P
Constant	24.576	4.803	5.12	0.000
BOD	18.860	1.799	10.49	0.000
Nitrate	-1.418	1.503	-0.94	0.360
Fecal Coli	0.0004732	0.0002511	1.88	0.078
Total Coli	-0.0000860	0.0001649	-0.52	0.609

S = 3.72793 R-Sq = 97.0% R-Sq(adj) = 96.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	7302.4	1825.6	131.36	0.000
Residual Error	16	222.4	13.9		
Total	20	7524.7			

Source	DF	Seq SS
BOD	1	7215.7
Nitrate	1	1.2
Fecal Coli	1	81.6
Total Coli	1	3.8

IV. Conclusion

Water Quality Index identified the most polluted stations that are Kadaghat, Allahabad, Khurgi, Patna U/S, Bihar, Varanasi D/S (Malviya Bridge), U.P, Indrapuri, Dehri and Varanasi U/S (ASSIGHAT), U.P.

Cluster Analysis for the different Stations showed a similarity of 99.99% between the stations Ganga D/S, Mirzapur , Varanasi D/S (Malviya Bridge) and Varanasi U/S (Assighat), U.P. Cluster Analysis for variables showed a 98.96% similarity of parameter BOD with WQI and 96.06% similarity between the parameters Total Coliform and Fecal Coliform. After applied the Best Subset Regression Analysis we get the highest Mallow c-p value with high R^2 for the parameters BOD, Nitrate, Total Coliform and Fecal Coliform.

In the Regression analysis the p value for the estimated coefficients of BOD is 0.00, indicates that BOD is significantly related to WQI.

In this paper we conclude that BOD is the most critical parameter and we study the comparison of water quality of river Ganga for different stations.

References

[1]Behrens, J. T., K. E. Dicerbo, N. Yel, and R. Levy. Exploratory data analysis. In Handbook of psychology. 2d ed. Vol. 2. Edited by J. A. Schinka, W. F. Velicer, and I. B. Weiner, 34–70. Hoboken, NJ: Wiley. (2013)

[2] D.N. Allen and G. Goldstein (eds.), Cluster Analysis in Neuropsychological Research: 13 Recent Applications, DOI 10.1007/978-1-4614-6744-1_2, © Springer Science Business Media New York (2013)

[3]J. Stewart, M. Miller, C. Audo, G. Stewart, Using cluster analysis to identify patterns in students' responses to contextually different conceptual problems, Phys. Rev. ST Phys. Educ. Res. 8, 020112 (2012)

[4] D. Hammer & L. K. Berland , Confusing Claims for Data: A critique of Common Practices for Presenting Qualitative Research on Learning, Journal of the Learning Sciences, 23, 37-46 (2014)
[5] M. T. H. Chi, Quantifying Qualitative Analyses of Verbal Data: A Practical Guide, The Journal of the Learning Sciences 6(3), 271-315 (1997)
[6]B. S. Everitt, S. Landau, M. Leese and D. Stahl, Cluster Analysis, (John Wiley & Sons, Ltd, Chichester, UK (2011)