TEEP @ASIA PLUS

DRONE AND AI-BASED REAL TIME POLLUTION MONITORING OF WATER

17/07/2023

IIT ROPAR

Author- SUJAL THAKUR

FINAL PROGRESS REPORT

SUPERVISOR- PROF. PAO-ANN HSIUNG

The major aim of this study is the collection of data-set from the Punjab Pollution Control Board for three years (2020, 2021 & 2022). Sutlej river data-set from various important locations in the Roop-Nagar to Harike area, analysis of the physicochemical parameters of the collected data-set, calculation of the WQI using these data and WAWQI method, and then prediction modeling of the water quality using machine learning techniques. The performance of the prediction models was assessed by using several algorithms.

Table of Contents

Abstract……03
1. Introduction……04
1.1Importance of Water……04
1.2Current Situation of Sutlej River and its Pollution……04
1.3Problems Caused by Sutlej River Pollution ……04
2. Water Quality……05
2.1Criteria & Overall Quality……05
2.2WQC Standards……05
2.3Classification Scheme……06
2.4Parameters……07
3. Methods……08
3.1WQI calculation methods……08
3.2OIP……10
4. Data Analysis……..11
4.1 Sampling stations……11
4.2 Formulation for the determination of WQI……..11
4.3 Calculations & Plots for WQI……14
4.4 Year-Over-Year Comparison……18
5. Statistical Analysis……..20
5.1 ANOVA……20
5.2 ANOVA table and observation……20
6. Sustainability……..23
6.1 U4SSC……23
6.2 KPI……24
6.3 Equations involved……25
6.4 Significance of SDG (6.3.1)……25
7. Results……26
7.1 Code explanation……26
7.2Model Comparison……31
8. Conclusions……32
9. Software used……33
10. References……34

Abstract

Various National and International Agencies involved in water quality assessment and pollution control have defined water quality criteria for different uses of water considering different indicator parameters. **Classification schemes** for water quality criteria/standards developed by these agencies differ in addition to terminologies used such as Action level, Guide level etc. in defining the concentration values in these classes. In the present report a general classification scheme viz. Excellent, Acceptable, Slightly Polluted, Polluted and Heavily Polluted water is proposed for surface water quality assessment. The concentration ranges in these classes are defined in Indian scenario considering Indian Standards and **CPCB** criteria. Standards by the European Community (EC), WHO etc. and the reported facts about the pollution effects of important water quality indicator parameters on the surrounding were also taken into account. The mathematical equations to transform the actual concentration values into pollution indices are formulated and corresponding value function curves are plotted. Based on the individual index values, 'Weighted Arithmetic Water Quality Index' **(WAWQI)** is estimated. The application of WAWQI is demonstrated at a few sampling stations on river **Sutlej** based on observed water quality data. The general classification scheme along with concentration ranges defined in these classes will be of immense use for determining the surface water quality status with reference to specific individual parameter, and the WAWQI for assessing the water quality status in Indian context.

We have used open datasheets of the year 2020, 2021 and 2022 of various sites to check the variation of WQI over time and predict the importance of such parameters in determining WAWQI by using some statistical analysis methods such as ANOVA (Analysis of Variance). ML model was trained using the information we gathered.

1. Introduction

1.1Importance of water

Water plays a vital role in numerous aspects of daily life, including drinking, cooking, personal hygiene, agriculture, and industrial processes. Unfit drinking water can lead to diseases like cholera, dysentery, and typhoid, compromising well-being and increasing mortality rates. Access to clean and safe water is crucial for maintaining good health, preventing the spread of diseases, and promoting overall well-being.

1.2Current Situation of Sutlej River and its Pollution

The Sutlej River, one of the major rivers in South Asia, faces a concerning state of pollution and degradation in its current situation. Spanning across India and Pakistan, the Sutlej River plays a crucial role in providing water for irrigation, supporting ecosystems, and serving as a source of livelihood for numerous communities. However, anthropogenic activities and inadequate environmental management have led to a deterioration of its water quality.

1.3Problems Caused by Sutlej River Pollution

The industrial sector along the Sutlej River contributes significantly to its pollution. Untreated industrial effluents containing heavy metals, chemicals, and other contaminants are discharged directly into the river, compromising its ecosystem health and the well-being of downstream communities.

Moreover, agricultural practices, including excessive use of fertilizers and pesticides, result in agricultural runoff, further exacerbating water pollution. The current situation of the Sutlej River demands immediate attention and concerted efforts from government bodies, environmental organizations, and local communities.

Implementing effective pollution control measures, such as establishing wastewater treatment plants, enforcing strict regulations on industrial discharges, and promoting sustainable agricultural practices, is crucial to mitigate the pollution and restore the river's health.

2. Water Quality

The quality of water is defined in terms of its physical, chemical, biological, and bacteriological parameters.

2.1Water Quality Criteria

Water quality criteria refer to the specific standards or guidelines set by regulatory agencies, organizations, or governments to assess and manage the quality of water resources. These criteria define acceptable levels or ranges of various physical, chemical, and biological parameters that indicate the suitability of water for specific uses or protection of aquatic ecosystems.

Acceptable levels of concentrations of water may vary with the type of use such as drinking, bathing and domestic use, irrigation, recreation and industrial use.

Overall Quality

Water quality parameters measured individually and is transformed into a single number (Index) representing the overall quality of water. Water quality indices, such as the Water Quality Index (WQI), are tools that combine multiple parameters and provide a single numerical value to represent overall water quality. These indices simplify the assessment process by integrating various factors and providing an easily understandable indicator of water quality.

2.2WQC Standards

a) **European Community (EC) Standards: -** The standards are expressed as Guide Level and Maximum Admissible Concentration.

b) **World Health Organization (WHO) Guidelines: -** are indicated in terms of Action Levels for important organic, inorganic and bacteriological parameters.

c) **Central Pollution Control Board Standards: -** CPCB has developed the standards for Inland Surface Water, considering the classification (A) to (E).

2.3Proposed Classification Scheme by CPCB

A classification scheme is proposed by CPCB in the present Indian context, which is: -

• **Excellent** means water quality is pristine –Class C1.

• **Acceptable** needs only disinfection – Class C2.

• **Slightly Polluted** requires filtration and disinfection – Class C3.

• **Polluted** requires special treatment and disinfection – Class C4.

• **Heavily Polluted** water cannot be used for any purpose – Class C5.

Table-1(Proposed classification of water quality)

Classification	Excellent C1	Acceptable C2	Slightly polluted C3	Polluted C4	Heavily Polluted C5
Class Index Score	1	2	4	8	16
Parameters			Concentration limit / ranges		
Turbidity (NTU)	5	10	100	250	>250
pH	6.5-7.5	6.0-6.5 and 7.5-8.0	5.0-6.0 and 8.0-9.0	4.5-5 and 9-9.5	<4.5 and >9.5
Colour	10	150	300	600	1200
DO	88-112	75-125	50-150	20-200	<20 and >200
BOD	1.5	3	6	12	24
TDS	500	1500	2100	3000	>3000
Hardness	75	150	300	500	>500
Cl	150	250	600	800	>800
NO(3)	20	45	50	100	200
SO(4)	150	250	400	1000	>1000
Coliform	50	500	5000	10000	15000
As	0.005	0.01	0.05	0.1	1.3
F	1.2	1.5	2.5	6.0	>6.0

Notes:

(a) Except for pH and DO, the maximum concentration value indicated is to be included in that class (and not in the next class).

(b) In case of DO, upper and lower limits are to be included in that class.

(c) In case of pH, lower and upper limit of C1 are included in C1. In all other classes (C2–C5) lower limit of lower range and upper limit of upper range are included in that class, but upper limit of lower range and lower limit of upper range are excluded from that class.

2.4Selection of Indicator Parameter

Physical parameters: Physical characteristics of water include temperature, color, turbidity (clarity), and odor. These factors provide insights into the visual and sensory aspects of water quality.

Chemical parameters: Chemical components of water include dissolved oxygen, pH, nutrients (such as nitrogen and phosphorus), heavy metals, pesticides, and various organic and inorganic substances. Monitoring these parameters helps identify potential contamination and the presence of pollutants that may have detrimental effects on water quality and the environment.

Biological parameters: Biological indicators in water quality assessments often involve the measurement of coliform bacteria, which serve as an indicator of fecal contamination and the potential presence of harmful pathogens. Other biological parameters can include the assessment of aquatic biodiversity, the presence of algae or other microorganisms, and the health of indicator species.

a) Field Observation- color, odor, temperature.

b) Core Parameters- pH, electrical conductivity, BOD, DO.

c) General Parameters- TDS, turbidity, hardness, salinity.

d) Metals- As, Cl, NO3, F, DDT etc.

3. Methods

At present, the most commonly used method to evaluate the quality of a water stream is the application of the Water Quality Index, which may be determined by using different methods. The main purpose of this study is to describe four methods for calculating the Water Quality Index with their advantages and disadvantages: NFS-WQI (National Sanitation Foundation-Water Quality Index), OWQI (Oregon Water Quality Index), WAWQI (Weighted Arithmetic Water Quality Index) and CCME-WQI (Canadian Council of Ministers of the Environment -Water Quality Index).

3.1WQI Calculation methods

(a) NFS-WQI=$\sum_{i=1}^{n}(Wi)(Qi)$

WQI-NFS is a numerical value between 0-100,**W(i)** is the weighting factor for each parameter, **Q(i)** is the sub-index of the quality parameter i, which is obtained from the conversion curve (curves that convert parameters determined by values from the interval 0-100).

- It uses only 9 parameters.
- It quickly, objectively, and reproducibly sums together data on the analyzed parameters into one number.
- The index value indicates the potential water use only.
- Some of the data get lost while being manipulated.

Water quality:

0-25	**Excellent**
26-50	**Good**
51-75	**Poor**
76-100	**Very Poor**
>100	**Unsuitable for drinking.**

(b) OWQI= $\sqrt{\frac{n}{\sum_{i=1}^{n} \frac{1}{(Si)^2}}}$

where:- **n** is the numbers of parameters (n=8), and **Si** is the sub-index of sub-index i(th) parameter. OWQI is a numerical value between 0-100.

- It uses only 8 parameters.
- The formula is sensitive to environmental changes and significant impacts on water quality,
- It does not provide conclusive data on how harmful substance concentrations have changed over time.
- It cannot evaluate all the toxic elements for health (bacteria, metal, toxics).

(c) WAWQI= $\frac{\sum Wi*qi}{\sum Wi}$

The next sections will provide a detailed explanation of this formula.

- It uses only 13 parameters.
- It encompasses the values of various parameters of water quality into a mathematical equation, which indicates the water ecological state.
- It can be used to describe the suitability of both surface and underground water sources for human consumption.
- This index may not provide enough information about the real situation of the water quality.

(d) CCME= $100 - \left(\frac{\sqrt{(F1)^2+(F2)^2+(F3)^2}}{1.732}\right)$

Where,

F(1) = (Number of failed variables / Total number of variables)*100,

F(2) = (Number of failed tests / Total number of tests)*100,

Excursion (i) = (Failed Test Value / Objective) –1,

nse = ∑ excursion / number of tests,

F3 = [nse / (0.01 nse) + 0.01].

- It has a high adaptability to different water uses.
- Easy to calculate but time consuming.
- The water quality gets described only partially.

Water quality: -

95-100	**Excellent**
80-94	Good
60-79	Fair
45-59	Marginal
0-44	Poor

3.2OIP (Overall index of pollution)

OIP is adopted in India for calculation of WQI. It will be of immense use for determining the surface water quality status with reference to specific individual parameter, and for assessing the overall water quality status in Indian context. However, comparative assessments of water quality at different places or at different times can be made only when the parameters included in the OIP are the same, and accordingly recommendations may be made regarding the specific use of water.

$$\mathbf{OIP} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Pi}, \qquad \mathbf{P(i)} = \frac{\boldsymbol{V(n)[Observed\ value\ of\ parameter]}}{\boldsymbol{V(s)[Standard\ value\ of\ parameter]}}$$

where, Pi = pollution index for ith parameter. (i = 1, 2, . . , n) and n = number of parameters. The numerical estimate of OIP corresponds to following classes:-

0–1 : Excellent (Class C1),
1–2 : Acceptable (Class C2),
2–4 : Slightly polluted (Class C3),
4–8 : Polluted (Class C4),
8–16 : Heavily polluted (Class C5).

4. Data Analysis

4.1 Sampling Stations

The locations that were primarily focused on to analyze, compare the WQI results produced using the WAWQI method, and perform ANOVA are from various areas of the Sutlej river. I obtained several raw open data sheets from the CPCB website. The data gathered spans the years 2020 to 2023. Three locations were selected from a pool of seven for this study. The sites are as follows: -

Sites	Name
Site-1	Ropar Head
Site-4	Ludhiana (D/S)
Site-7	Harike

4.2 Formulation for the determination of WQI

First, we took the original data sheets and extracted the site information. Parameters that aren't used to determine the WQI were omitted from the calculation.

Third, we produced site-specific tables by year. To do this, we used the Weighted Arithmetic Water Quality Index approach, which involves taking the values of the parameters from the data sheet and converting them into an index value. Graphed WAWQI against time to indicate the development of water quality.

Methodology to calculate WQI : -

WAWQI method is done in five steps:-

Step1- Collect data of various water quality parameters year wise.

Step2- Calculate Proportionality constant ‘K’ value using the given formula:-

$\mathbf{K= \left[\frac{1}{\frac{1}{\sum_{i=1}^{n} S(i)}}\right]}$, where ‘S(i)’ is standard permissible for nth parameter.

Step3- Calculate quality rating for nth parameter 'Q(n)' where there are 'n' parameters.

Q(n) = 100 {(V(n) – V(io)) / (S(n) – V(io))}

where, V(n) = estimated value of the nth parameter of the given sampling station, V(io) = ideal value of nth parameter in pure water, and S(n) =standard permissible value of the nth parameter.

Step4- Calculate unit weight for the nth parameters by:-

W(n) = [K / S(n)]

Step5- Calculated WQI using the given formula:-

WQI= [∑W(n)Q(n) / (∑W(n))]

Relative weights (Wn) of the parameters used for WQI determination:-

Parameter	S(n)	W(n)
pH	6.5-8.5	0.215
EC	300	0.0061
TDS	500	0.00366
TSS	500	0.00366
Hardness	300	0.0061
Chloride	250	0.00732
DO	5	0.366
BOD	5	0.366
Sulphate	150	0.0122
Alkalinity	120	0.01525

The numerical estimate of WAWQI corresponds to following classes:-

WQI	Water Quality Status
0-25	Excellent
26-50	Good
51-75	Poor
76-100	Very poor
>100	Unsuitable for drinking

Both VS Code and Excel were used for the bulk of the computations, while some were done manually.

Example:-

G3 | =((F3-E3)/(C3-E3))*100*D3

	A	B	C	D	E	F	G	H	I
1	**River Sutlej at Ropar Hea**	**9 Parameters**				**Jan-22**			
2	Parameters	WQI-No.	V(s)=w	W(n)	V(i)=V	V(n)=v	Q(n)*W(n	WAWQI	WQ-Status
3	**Ph**	0-25	8.5	0.215	7	8	14.3333		Excellent
4	**EC**	26-50	300	0.0061	0	331	0.67303	**39**	**Good**
5	**TDS**	51-75	500	0.00366	0	195	0.14274		Poor
6	**TSS**	76-100	500	0.00366	0	48	0.03514		Very Poor
7	**Cl**	>100	250	0.00732	0	15	0.04392		Unsuitable
8	**BOD**		5	0.366	0	0	0		
9	**DO**		5	0.366	14.6	8.8	22.1125		
10	**SO(4)**		150	0.0122	0	12	0.0976		
11	**Total Alk.**		120	0.01525	0	108	1.3725		
12	Sum=			**0.99519**			**38.81**		

WQI.py > ...

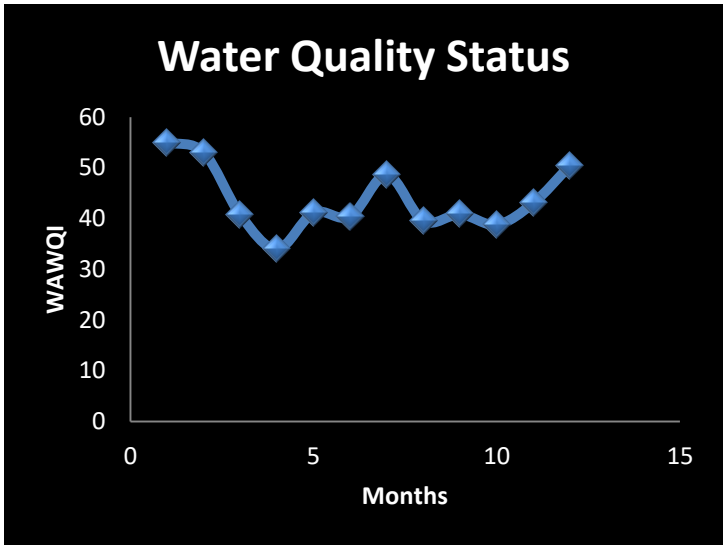
```
w= 8.5
W= 0.215
V= 7
v= 4.8
a = (v-V)
b = (w-V)
Q = (a/b)*100
I = Q*W
print(I)

t= 57.29
n=0.99519
A= t/n
print(A)
```

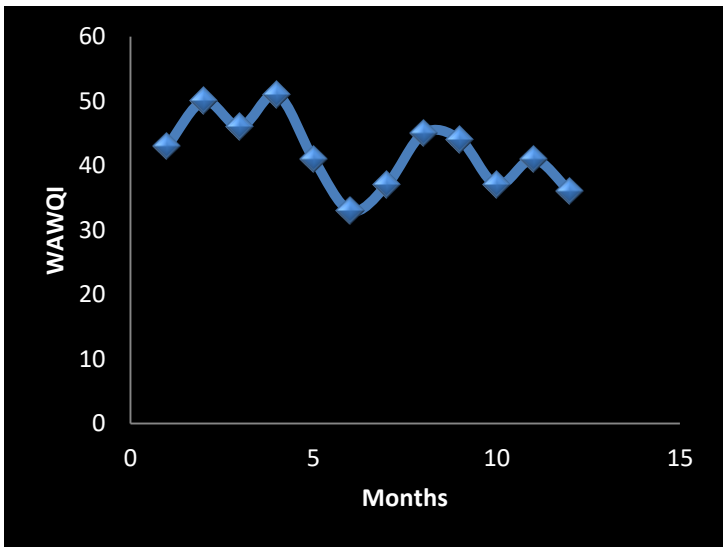
4.3 Calculations and WQI plot

Site-1 (Ropar head, Starting point, D/S)

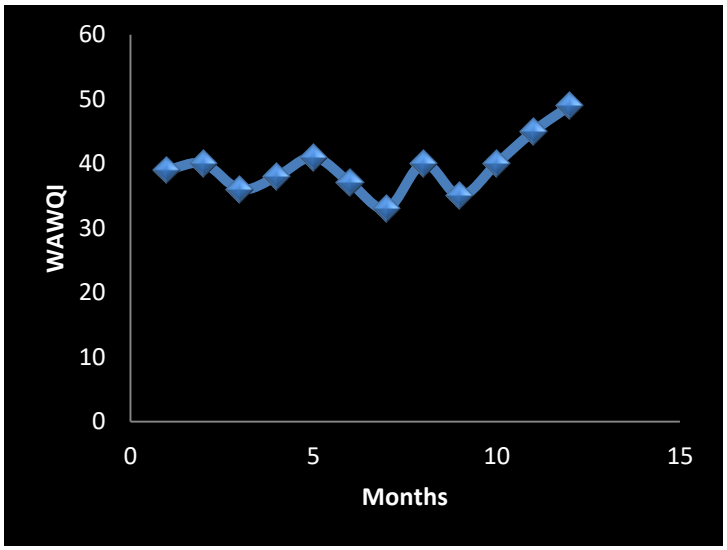
2020			
S.No.	**Month**	**WAWQI**	**Status**
1	Jan	55.01	Poor
2	Feb	53.01	Poor
3	Mar	40.73	Good
4	Apr	34	Good
5	May	41.17	Good
6	June	40.41	Good
7	July	48.6	Good
8	Aug	39.52	Good
9	Sep	40.85	Good
10	Oct	38.8	Good
11	Nov	43.08	Good
12	Dec	50.45	Good



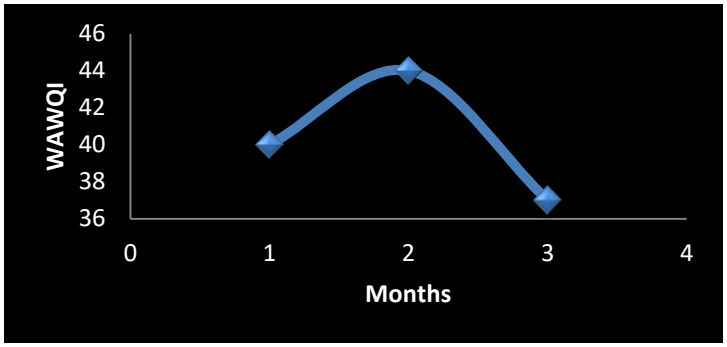
2021			
S.No.	**Month**	**WAWQI**	**Status**
1	Jan	43	Good
2	Feb	50	Poor
3	Mar	46	Good
4	Apr	51	Poor
5	May	41	Good
6	June	33	Good
7	July	37	Good
8	Aug	45	Good
9	Sep	44	Good
10	Oct	37	Good
11	Nov	41	Good
12	Dec	36	Good



2022			
S.No.	**Month**	**WAWQI**	**Status**
1	Jan	39	Good
2	Feb	40	Good
3	Mar	36	Good
4	Apr	38	Good
5	May	41	Good
6	June	37	Good
7	July	33	Good
8	Aug	40	Good
9	Sep	35	Good
10	Oct	40	Good
11	Nov	45	Good
12	Dec	49	Good

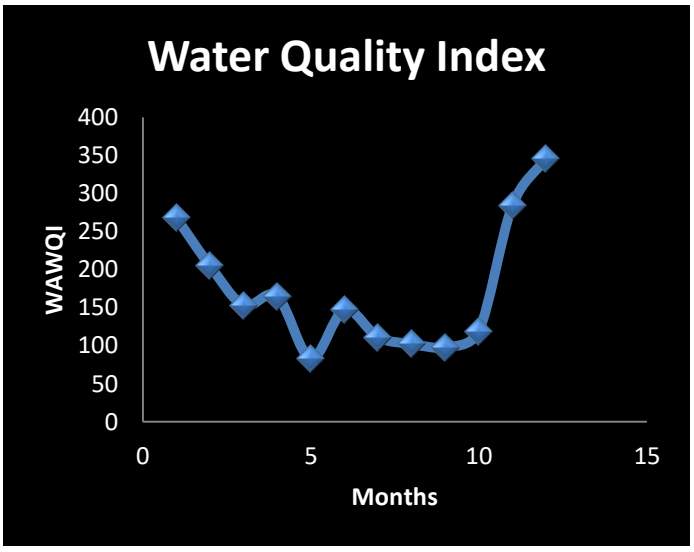


2023			
S.No.	**Month**	**WAWQI**	**Status**
1	Jan	40	Good
2	Feb	44	Good
3	Mar	37	Good

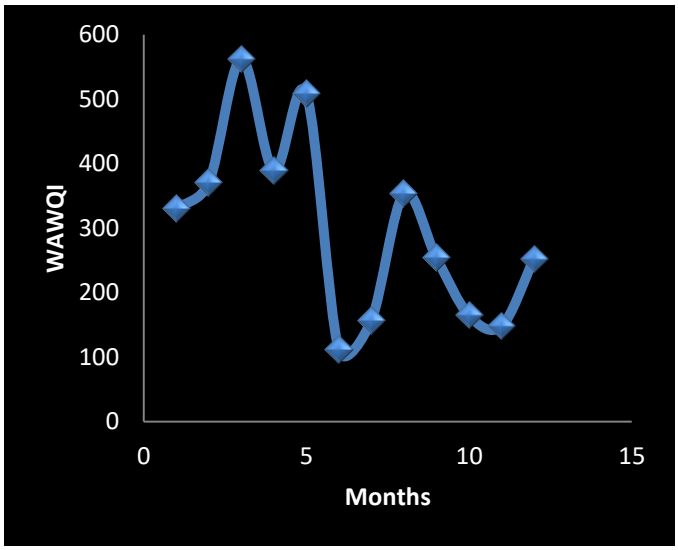


Site-4(Ludhiana, Mid-point, D/S)

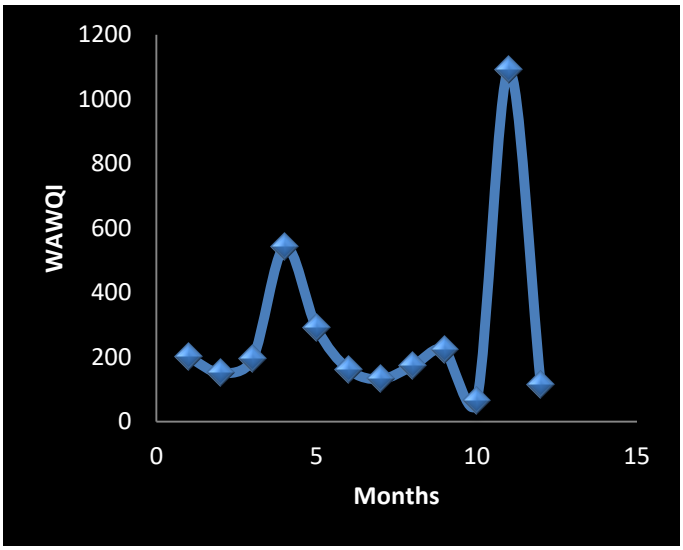
2020			
S.No.	**Month**	WAWQI	Status
1	Jan	268	Unsuitable
2	Feb	205	Unsuitable
3	Mar	153	Unsuitable
4	Apr	165	Unsuitable
5	May	83	Very Poor
6	June	147	Unsuitable
7	July	111	Unsuitable
8	Aug	102	Unsuitable
9	Sep	98	Very Poor
10	Oct	119	Unsuitable
11	Nov	284	Unsuitable
12	Dec	346	Very Poor



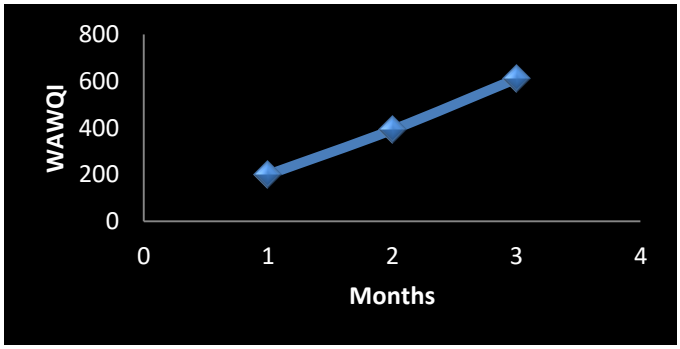
2021			
S.No.	**Month**	WAWQI	Status
1	Jan	330	Unsuitable
2	Feb	370	Unsuitable
3	Mar	562	Unsuitable
4	Apr	389	Unsuitable
5	May	508	Unsuitable
6	June	111	Unsuitable
7	July	157	Unsuitable
8	Aug	353	Unsuitable
9	Sep	254	Unsuitable
10	Oct	165	Unsuitable
11	Nov	148	Unsuitable
12	Dec	252	Unsuitable



2022			
S.No.	**Month**	WAWQI	Status
1	Jan	201	Unsuitable
2	Feb	151	Unsuitable
3	Mar	194	Unsuitable
4	Apr	542	Unsuitable
5	May	291	Unsuitable
6	June	162	Unsuitable
7	July	132	Unsuitable
8	Aug	173	Unsuitable
9	Sep	223	Unsuitable
10	Oct	65	Poor
11	Nov	1090	Unsuitable
12	Dec	113	Unsuitable

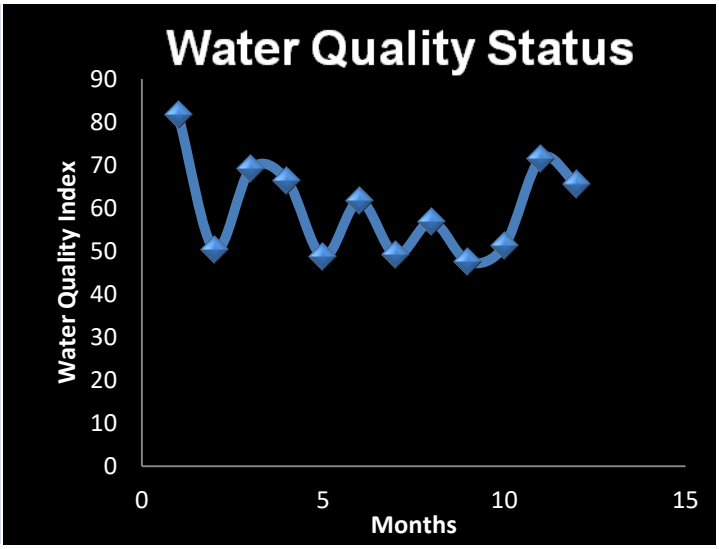


2023			
S.No.	**Month**	WAWQI	Status
1	Jan	199	Unsuitable
2	Feb	391	Unsuitable
3	Mar	610	Unsuitable

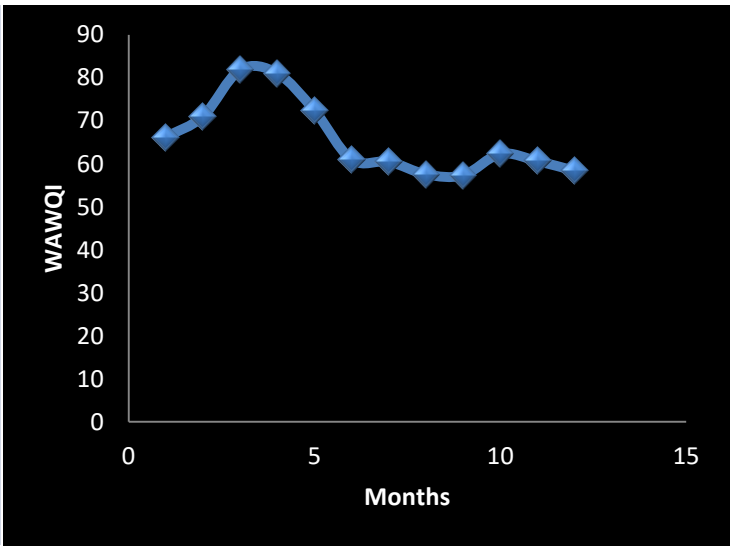


Site-7 (Harike, Last Station)

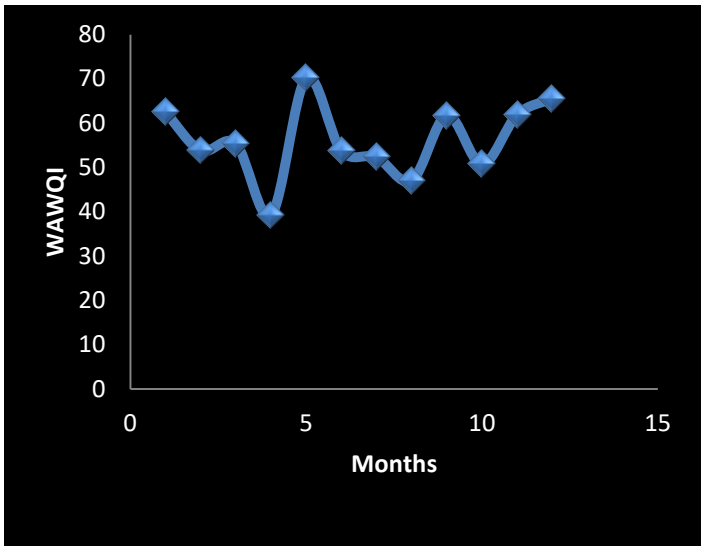
2020			
S.No.	**Month**	**WAWQI**	**Status**
1	Jan	81.74	Very Poor
2	Feb	50.39	Good
3	Mar	69.1	Poor
4	Apr	66.46	Poor
5	May	48.64	Good
6	June	61.66	Poor
7	July	49.14	Good
8	Aug	56.82	Poor
9	Sep	47.46	Good
10	Oct	51.18	Poor
11	Nov	71.5	Poor
12	Dec	65.53	Poor



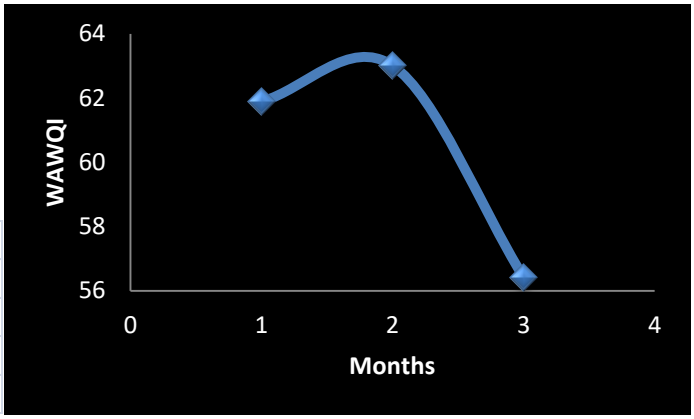
2021			
S.No.	**Month**	**WAWQI**	**Status**
1	Jan	69.1	Poor
2	Feb	71	Poor
3	Mar	81.4	Very Poor
4	Apr	81.03	Very Poor
5	May	72.39	Poor
6	June	60.96	Poor
7	July	60.49	Poor
8	Aug	57.56	Poor
9	Sep	57.33	Poor
10	Oct	62.31	Poor
11	Nov	60.62	Poor
12	Dec	58.38	Poor



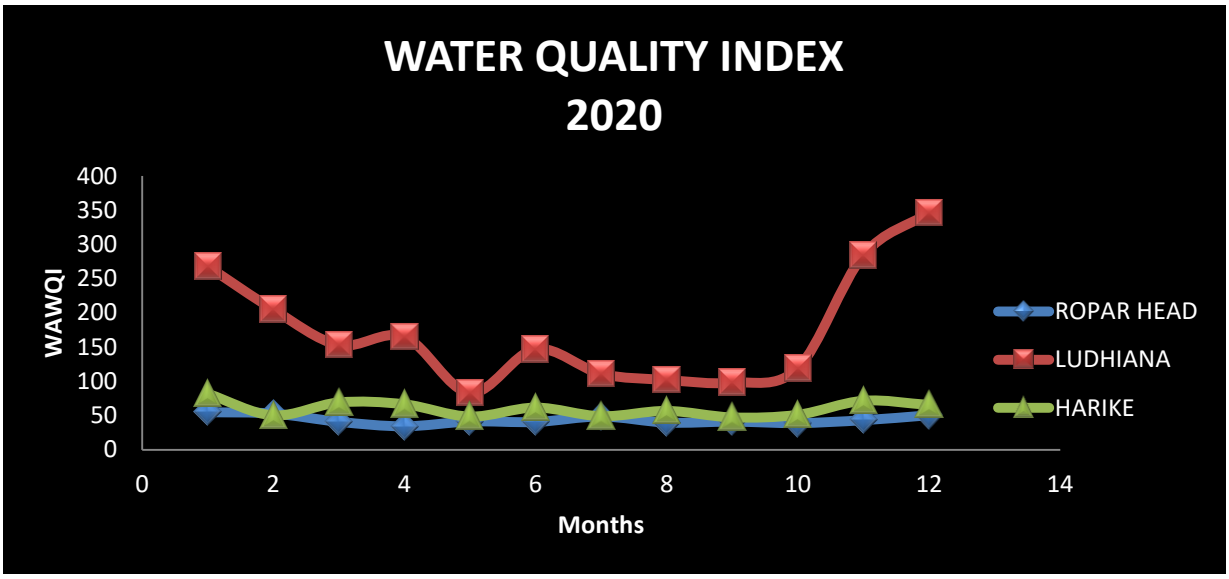
2022			
S.No.	**Month**	**WAWQI**	**Status**
1	Jan	62.53	Poor
2	Feb	53.86	Poor
3	Mar	55.4	Poor
4	Apr	39.24	Good
5	May	70.23	Poor
6	June	53.81	Poor
7	July	52.35	Poor
8	Aug	47.08	Good
9	Sep	61.77	Poor
10	Oct	50.76	Poor
11	Nov	61.93	Poor
12	Dec	65.62	Poor

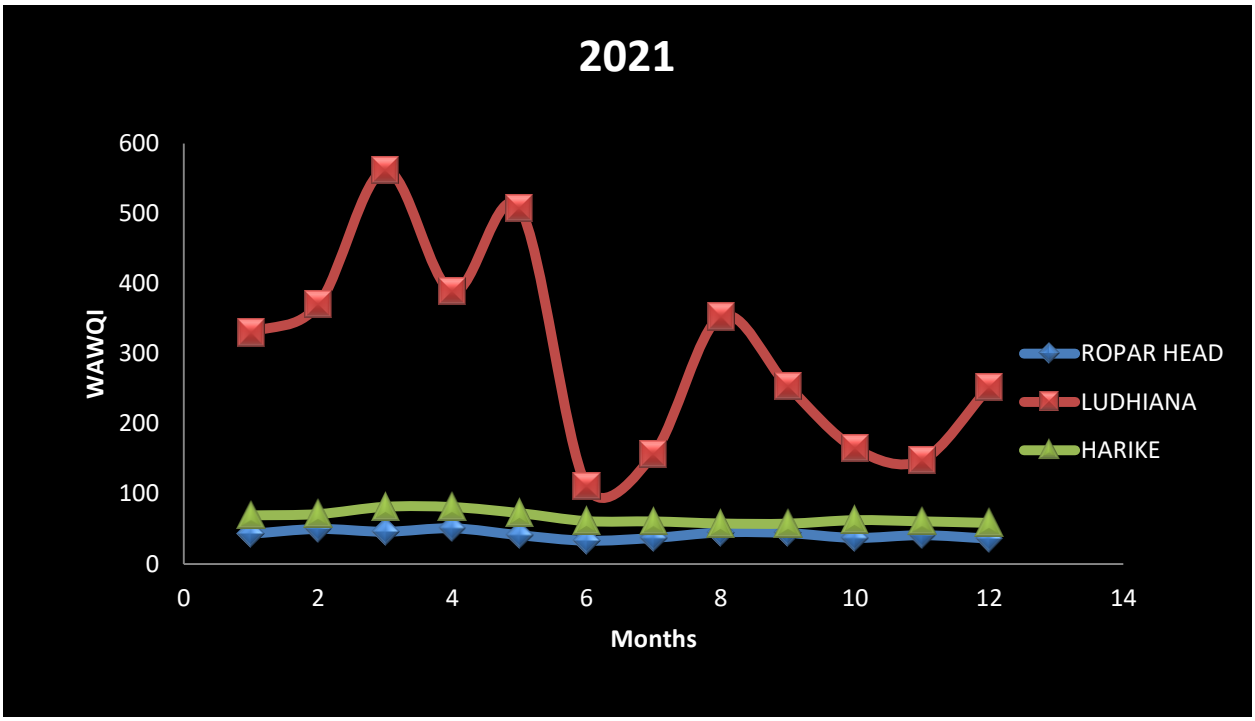


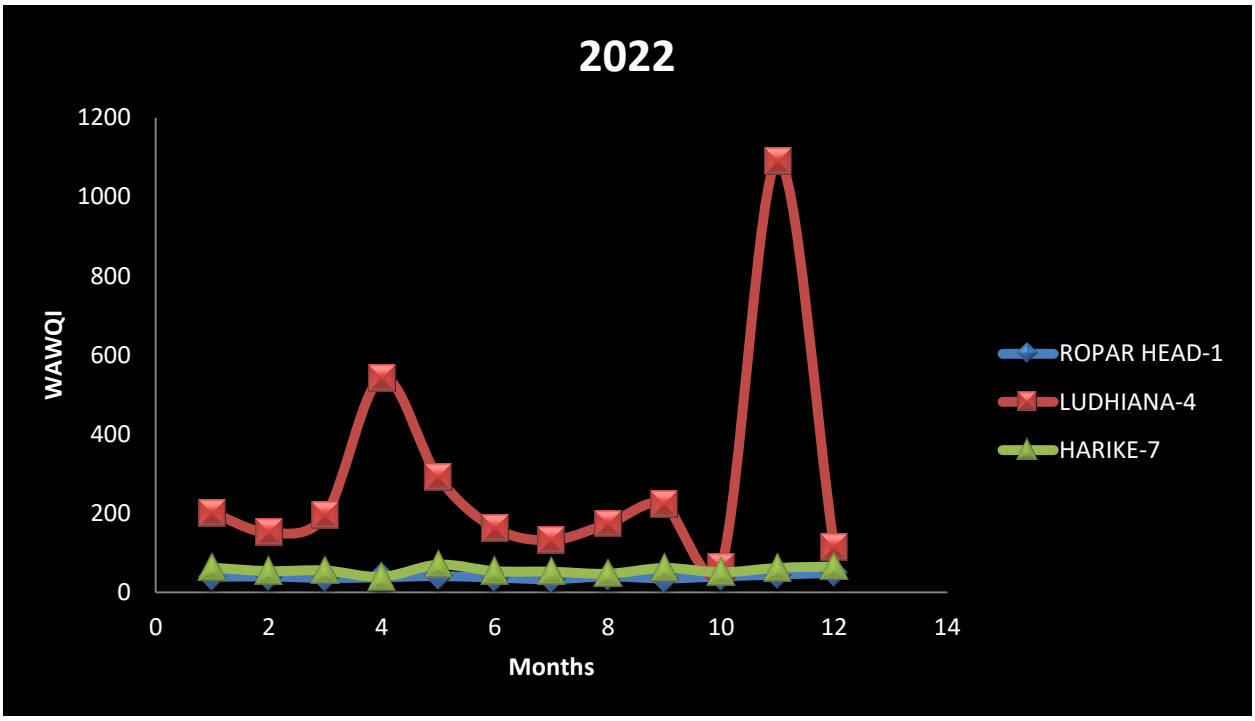
2023			
S.No.	**Month**	**WAWQI**	**Status**
1	Jan	61.89	Poor
2	Feb	63.01	Poor
3	Mar	56.42	Poor



4.4 A Year-Over-Year Comparison







Site-4 stands out as the most contaminated of the three options.

5. Statistical analysis

5.1 ANOVA

ANOVA (Analysis of Variance) is a statistical technique used to compare the means of two or more groups to determine if there are significant differences between them. It is often applied when the dependent variable is continuous, and the independent variable is categorical. ANOVA can be used in different scenarios:

- One-way ANOVA
- Two-way ANOVA
- Factorial ANOVA**(one we used)**

5.2 Table and Observation

Site-1								
2020	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	11.1496859	0.658381322	16.935	0.000448	9.054422737	13.2449491	9.05442274	13.2449491
EC	-0.0421757	0.061231766	-0.68879	0.540445	-0.23704248	0.15269114	-0.2370425	0.15269114
TDS	-0.004625	0.11172298	-0.0414	0.96958	-0.36017741	0.35092736	-0.3601774	0.35092736
TSS	-0.0298486	0.011159782	-2.67466	0.075391	-0.06536404	0.00566678	-0.065364	0.00566678
Cl	0.52276413	0.150390657	3.476041	0.04017	0.044153936	1.00137432	0.04415394	1.00137432
BOD	8.03925896	0.731805684	10.98551	0.001615	5.710326663	10.3681913	5.71032666	10.3681913
DO	-5.2269335	0.377605351	-13.8423	0.000816	-6.42864222	-4.02522471	-6.4286422	-4.02522471
SO(4)	-0.4160538	0.137888123	-3.01733	0.05688	-0.85487536	0.02276774	-0.8548754	0.02276774
Total Alk.	0.04528356	0.043636894	1.037736	0.375691	-0.09358851	0.18415564	-0.0935885	0.18415564
2021	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	9.6288552	1.208409145	7.968208	0.004124	5.783157981	13.4745524	5.78315798	13.4745524
EC	0.01091188	0.027004634	0.404074	0.713239	-0.07502892	0.09685268	-0.0750289	0.09685268
TDS	0.03373579	0.066445979	0.507718	0.646626	-0.17772497	0.24519655	-0.177725	0.24519655
TSS	-0.0049211	0.01237277	-0.39773	0.717437	-0.04429676	0.0344546	-0.0442968	0.0344546
Cl	-0.4237037	0.257151745	-1.64768	0.197976	-1.24207535	0.39466789	-1.2420754	0.39466789
BOD	10.7306954	1.237904937	8.668432	0.00323	6.791129369	14.6702614	6.79112937	14.6702614
DO	-5.3315078	0.934165664	-5.70724	0.01067	-8.30443982	-2.3585757	-8.3044398	-2.3585757
SO(4)	0.28547453	0.124641962	2.290357	0.105922	-0.11119182	0.68214089	-0.1111918	0.68214089
Total Alk.	-0.0024163	0.048909087	-0.0494	0.963703	-0.15806679	0.15323429	-0.1580668	0.15323429
2022	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	7.80106755	1.375684786	5.67068	0.010863	3.42302459	12.1791105	3.42302459	12.1791105
EC	0.06944363	0.068959588	1.007019	0.38811	-0.15001656	0.28890381	-0.1500166	0.28890381
TDS	-0.2661765	0.182866637	-1.45558	0.241522	-0.84813977	0.31578673	-0.8481398	0.31578673
TSS	0.02168261	0.031327244	0.692133	0.538614	-0.07801466	0.12137988	-0.0780147	0.12137988
Cl	-0.5872978	0.536307865	-1.09508	0.353526	-2.29406881	1.11947316	-2.2940688	1.11947316
BOD	0	0	65535	#NUM!	0	0	0	0
DO	-2.9208157	1.464950996	-1.9938	#NUM!	-7.58294358	1.74131219	-7.5829436	1.74131219
SO(4)	1.05626566	0.630965647	1.674046	0.192715	-0.95174863	3.06427995	-0.9517486	3.06427995
Total Alk.	0.24285085	0.116652677	2.081828	0.128777	-0.12839004	0.61409173	-0.12839	0.61409173

Observation-

From the ANOVA we can observe the importance of each of the parameters used to estimate WAWQI as:-

Year	Parameters
2020	Ph>BOD>Cl>Alkalinity
2021	BOD>pH>SO(4)>TDS
2022	pH>SO(4)>Alkalinity>EC

Site-4								
2020	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	7.4406721	1.175411727	6.330269	0.007971	3.69998739	11.18135681	3.699987392	11.18135681
EC	-0.0048206	0.009836122	-0.49009	0.657672	-0.0361235	0.026482312	-0.03612355	0.026482312
TDS	-0.0370788	0.044602433	-0.83132	0.466776	-0.1790237	0.104866013	-0.17902368	0.104866013
TSS	-0.0372185	0.030078147	-1.23739	0.303964	-0.1329405	0.058503633	-0.13294054	0.058503633
Cl	0.22299415	0.140829552	1.583433	0.211484	-0.2251883	0.671176641	-0.22518833	0.671176641
BOD	7.50197454	0.406883939	18.43763	0.000348	6.20708825	8.796860831	6.207088254	8.796860831
DO	-0.5322037	2.854984085	-0.18641	0.864015	-9.6180373	8.55362981	-9.6180373	8.55362981
SO(4)	0.10534764	0.131235304	0.802739	0.480833	-0.3123017	0.522996945	-0.31230167	0.522996945
Total Alk.	-0.0035312	0.0570769	-0.06187	0.95456	-0.1851753	0.178113013	-0.18517533	0.178113013
2021	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	8.9857214	0.750088424	11.97955	0.001251	6.59860526	11.37283753	6.598605263	11.37283753
EC	-0.0195452	0.028884388	-0.67667	0.547118	-0.1114682	0.07237782	-0.11146821	0.07237782
TDS	0.04173821	0.047317425	0.88209	0.442683	-0.108847	0.192323368	-0.10884696	0.192323368
TSS	-0.0087847	0.013346538	-0.6582	0.557412	-0.0512593	0.033689968	-0.05125931	0.033689968
Cl	0.01001681	0.036580513	0.273829	0.801987	-0.1063987	0.126432331	-0.10639871	0.126432331
BOD	7.29874288	0.11565191	63.10958	8.77E-06	6.93068688	7.666798868	6.930686884	7.666798868
DO	-4.9241005	2.879624093	-1.70998	0.185798	-14.08835	4.240148592	-14.0883495	4.240148592
SO(4)	-0.0891641	0.103095685	-0.86487	0.45073	-0.4172606	0.238932378	-0.41726059	0.238932378
Total Alk.	0.00364731	0.032010107	0.113942	0.916481	-0.0982231	0.105517757	-0.09822314	0.105517757
2022	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	9.61613971	0.765295926	12.56526	0.001087	7.18062651	12.0516529	7.180626514	12.0516529
EC	-0.0181945	0.03383785	-0.5377	0.628114	-0.1258816	0.089492664	-0.12588161	0.089492664
TDS	0.04017594	0.072729305	0.552404	0.619163	-0.1912812	0.271633045	-0.19128117	0.271633045
TSS	-0.0055147	0.016412125	-0.33601	0.75899	-0.0577454	0.046716023	-0.05774539	0.046716023
Cl	0.04039325	0.02791288	1.447119	0.243668	-0.048438	0.129224491	-0.04843799	0.129224491
BOD	7.35847894	0.031388709	234.4308	1.71E-07	7.25858606	7.458371819	7.258586059	7.458371819
DO	-3.9550912	0.395239786	-10.0068	0.002124	-5.2129206	-2.69726178	-5.21292057	-2.697261784
SO(4)	-0.2686829	0.195249764	-1.3761	0.262522	-0.8900548	0.352688959	-0.89005482	0.352688959
Total Alk.	-0.0186483	0.023161142	-0.80515	0.479631	-0.0923574	0.055060795	-0.09235738	0.055060795

Observation-

The ANOVA reveals the significance of each of the parameters used to estimate WAWQI as follows:-

Year	Parameters
2020	BOD>pH>Cl>SO(4)
2021	pH>BOD>TDS>Cl
2022	pH>BOD>Cl>TDS

Site-7								
2020	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	10.6607989	2.59541693	4.107548	0.026124	2.40102384	18.9205739	2.40102384	18.92057388
EC	0.0194284	0.143497549	0.135392	0.900876	-0.43724485	0.47610164	-0.4372448	0.476101641
TDS	0.06411589	0.236958925	0.270578	0.804263	-0.68999316	0.81822495	-0.6899932	0.818224946
TSS	0.00137261	0.018316914	0.074937	0.944982	-0.05691998	0.05966521	-0.05692	0.05966521
Cl	-0.2243262	0.339569871	-0.66062	0.556055	-1.30498913	0.85633664	-1.3049891	0.856336639
BOD	9.24272308	2.235179536	4.135114	0.025667	2.12938423	16.3560619	2.12938423	16.35606194
DO	-6.4385419	3.282605055	-1.96141	0.144653	-16.8852562	4.00817247	-16.885256	4.008172474
SO(4)	-0.0793231	0.280049359	-0.28325	0.795409	-0.97056511	0.81191899	-0.9705651	0.811918991
Total Alk.	-0.1443089	0.165208439	-0.8735	0.446682	-0.67007585	0.38145812	-0.6700758	0.381458125
2021	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	8.52570487	0.661320347	12.89194	0.001007	6.42108837	10.6303214	6.42108837	10.63032136
EC	-0.0342574	0.039747288	-0.86188	0.452139	-0.16075097	0.09223625	-0.160751	0.092236249
TDS	0.05820573	0.068111153	0.85457	0.455603	-0.15855436	0.27496581	-0.1585544	0.274965813
TSS	-0.0169525	0.008002476	-2.11841	0.12437	-0.04241996	0.00851494	-0.04242	0.008514944
Cl	0.0815812	0.077477214	1.05297	0.369674	-0.16498587	0.32814828	-0.1649859	0.328148277
BOD	5.44335573	0.586014621	9.288771	0.002641	3.57839567	7.3083158	3.57839567	7.3083158
DO	-3.1893807	0.499511517	-6.385	0.007779	-4.77904932	-1.5997122	-4.7790493	-1.599712161
SO(4)	-0.0313165	0.160422202	-0.19521	0.8577	-0.5418515	0.47921859	-0.5418515	0.479218588
Total Alk.	0.02554512	0.048271914	0.529192	0.633329	-0.12807766	0.17916789	-0.1280777	0.179167891
2022	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Ph	12.2564877	1.175388823	10.4276	0.001882	8.51587592	15.9970996	8.51587592	15.99709956
EC	0.04218972	0.041539584	1.015651	0.384581	-0.09000777	0.17438722	-0.0900078	0.174387219
TDS	-0.0597814	0.109468859	-0.5461	0.622986	-0.40816014	0.2885974	-0.4081601	0.288597395
TSS	-0.0119992	0.018268617	-0.65682	0.558185	-0.07013813	0.04613965	-0.0701381	0.04613965
Cl	0.03751119	0.099332539	0.377632	0.730838	-0.27860928	0.35363166	-0.2786093	0.353631659
BOD	4.69694455	1.456162418	3.225564	0.048379	0.06278584	9.33110326	0.06278584	9.331103258
DO	-5.1592294	1.370582964	-3.76426	0.032794	-9.52103609	-0.7974227	-9.5210361	-0.797422712
SO(4)	0.0999564	0.170934221	0.584765	0.599777	-0.44403257	0.64394538	-0.4440326	0.643945383
Total Alk.	-0.1533256	0.099488859	-1.54113	0.220939	-0.4699436	0.16329231	-0.4699436	0.163292311

Observation-

The significance of the various parameters used to determine WAWQI can be gleaned from the ANOVA as follows:-

Year	Parameters
2020	pH>BOD>TDS>EC
2021	pH>BOD>Cl>TDS
2022	pH>BOD>SO(4)>EC

6. Sustainability

The capacity to keep anything going for a long time is what we mean when we talk about sustainability. It's a sort of ethics that considers future generations, ensuring that our current activities do not compromise the quality of life for those who come after us. Sustainability is an interdisciplinary study that includes all aspects of human well-being: ecological, cultural, economic, and social.

6.1 U4SSC

U4SSC is a United Nations Initiative coordinated by ITU and UNECE that advocates for public policy to encourage the use of ICTs to facilitate and ease the transition to smart sustainable cities.

The U4SSC Initiative has developed a set of international key performance indicators (KPIs) for Smart sustainable cities (SSC) to establish the criteria to evaluate ICT´ s contributions in making cities smarter and more sustainable, and to provide cities with the means for self-assessments.

The U4SSC Key Performance Indicators are a vital policy tool for cities wishing to achieve the UN Sustainable Development Goals. U4SSC is currently working on the following deliverables:-

- Tools and mechanisms to finance SSC projects,
- Guidelines on strategies for circular cities,
- Principles for the application of AI in urban settings.

6.2 KPI

The KPIs support cities and communities worldwide in evaluating the role and contribution of ICTs and digital technologies in enabling smartness and sustainability.

There are 91 KPIs that cities can measure and report. Each KPI falls under one of three dimensions of smartness and sustainability in a city:

- Economy,
- Environment and
- Society and Culture.

Principles-

- Comprehensiveness- (should cover all the aspects of SSC).
- Availability- (historic and current data should either be available or easy to collect).
- Independence-(overlap of the KPIs should be avoided).
- Simplicity- (The concept of each indicator should be simple and easy to understand for the urban stakeholders).
- Timeliness-(ability to produce KPIs with respect to emerging issues in SSC construction).

Needs-

- Measure [like- new customers, % increase in sales],
- Target [1000, 22%],
- Source [Organization, CRM],
- Frequency [How we are reporting or running the monthly (quarterly) strategies].

Types-

- Raw [which tells the number(in form of a graph)],
- Progress [tells the % complete of the goal],
- Change [% increase in the sales compared to the last year].

6.3 Equations involved

S.No.	Dimension	KPI name	Methodology	Data Sources	SDG Reference
1	Economy	Wastewater Collection	[(No. of households served by wastewater collection) / (Total no. of households)]*100	Data should be collected from local utilities that operate wastewater facilities.	SDG Target- 6.3 by 2030.
2	Environment	Wastewater treatment	[(Total amount of wastewater that has undergone treatment) / (Total amount of wastewater collected)]*100	This information is usually known by municipal authorities and is available from the main water supply and treatment companies.	SDG indicator 6.3.1.

6.4 Significance of SDG Indicator 6.3.1

- **Human Health protection-** The indication aids in the fight against numerous water-borne infections.

- **Environmental conservation-** We can lessen the negative effects, such as water pollution, affects on biodiversity, and disruptions to aquatic ecosystem equilibrium.

- **Resource conservation-** Many useful materials, such as nutrients and energy, are found in wastewater; by treating wastewater, we may recover and utilize these materials, so increasing resource efficiency.

- **Sustainable development-** It addresses the environmental impacts of wastewater treatment and increases people's access to clean water and sanitation at a reasonable cost.

7. Results

The WAWQI approach was used to efficiently organize data from publicly available sources, and the WQI was then determined. Using the values of nine parameters from the same open datasheet, I have calculated the value of WAWQI and added a new column to the 'table' to display it. The time period that is covered by this data collection is from January 2020 through March 2023.ML model was trained in order to predict the water potability.

Features- pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity.

Target- Potability (0-Unsuitable, 1-Suitable).

7.1 Code explanation

River data in.csv format were extracted from the sorted data sheets.

- **Libraries and model used**

```
import pandas as pd
import numpy as np
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score

from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.ensemble import AdaBoostClassifier
from xgboost import XGBClassifier
```

- **Filling null values**
  The column's mean value was used to fill in any null values. In terms of pH, sulphate, and trihalomethanes, there were 491 781 and 162 null results, respectively.

```
df["ph"] = df["ph"].fillna(df["ph"].mean())
df["Sulfate"] = df["Sulfate"].fillna(df["Sulfate"].mean())
df["Trihalomethanes"] =
df["Trihalomethanes"].fillna(df["Trihalomethanes"].mean())
```

- **Visualization dataset and also checking for outliers**

```
fig, ax = plt.subplots(ncols=5, nrows=2, figsize= (20,10))

ax = ax.flatten()

index = 0

for col,values in df.items():
  sns.boxplot(y=col, data=df, ax = ax[index])

  index +=1
```

- **Correlation-** means that particular feature is how much co-related with the other features. And which feature has the highest and lowest correlation with the target (Potability).[Highest- Solids, Lowest- Sulfate]

```
plt.figure(figsize= (12,8))
sns.heatmap(df.corr(), annot=True)
```

- **Standard scaling-** keeping the data in one scale.

```
scaler = StandardScaler()
x = scaler.fit_transform(x)
x
```

- Data Preparations for Training, x-axis(features), y-axis(Target). Secondly, I have split the data into train and test data sets in which the test size is 20% of the total length of the data.

```
 x = df.drop("Potability", axis=1)

y= df["Potability"]
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2)
```

- I have used various Scikit learn machine learning classifier models to compare the accuracy scores obtained from the testing data-set. Firstly, I have trained a RandomForestClassifier model by using the training data set.

a) Random Forest Calssifier-

```
from sklearn.ensemble import RandomForestClassifier

model_rf = RandomForestClassifier()
model_rf.fit(x_train,y_train)
pred_rf = model_rf.predict(x_test)
accuracy_score_rf = accuracy_score(y_test, pred_rf)
accuracy_score_rf
cross_val_score(model_rf, x,y,cv=5)
model_rf_cross_val_score = np.mean(cross_val_score(model_rf,x,y,cv=5))
model_rf_cross_val_score
cm3 = confusion_matrix(y_test, pred_rf)
cm3
```

b) Logistic Regression

```
from sklearn.linear_model import LogisticRegression
model_lr= LogisticRegression()
model_lr.fit(x_train,y_train)
pred_lr = model_lr.predict(x_test)
accuracy_score_lr = accuracy_score(y_test, pred_lr)
accuracy_score_lr
from sklearn.model_selection import cross_val_score
cross_val_score(model_lr, x,y,cv=5)
model_lr_cross_val_score = np.mean(cross_val_score(model_lr,x,y,cv=5))
model_lr_cross_val_score
```

c) Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier

# Creating the model object
model_dt = DecisionTreeClassifier(max_depth = 4)
model_dt.fit(x_train, y_train)
pred_dt = model_dt.predict(x_test)
accuracy_score_dt = accuracy_score(y_test, pred_dt)
accuracy_score_dt
cm2 = confusion_matrix(y_test, pred_dt)
cm2
from sklearn.model_selection import cross_val_score
cross_val_score(model_dt, x,y,cv=5)
model_dt_cross_val_score = np.mean(cross_val_score(model_dt,x,y,cv=5))
model_dt_cross_val_score
```

d) kNN

```
from sklearn.neighbors import KNeighborsClassifier
for i in range(4,15):
  model_knn = KNeighborsClassifier(n_neighbors=i)
  model_knn.fit(x_train,y_train)
  pred_knn = model_knn.predict(x_test)
  accuracy_score_knn = accuracy_score(y_test, pred_knn)
  print(i, accuracy_score_knn)

 # Best accuracy in 4- 0.649390243902439
model_knn = KNeighborsClassifier(n_neighbors=4)
model_knn.fit(x_train,y_train)
pred_knn = model_knn.predict(x_test)
accuracy_score_knn = accuracy_score(y_test, pred_knn)
print(accuracy_score_knn)
```

e) SVM- (kernel- rbf)

```
from sklearn.svm import SVC

# Creating object of model
model_svm = SVC(kernel="rbf")
model_svm.fit(x_train, y_train)
pred_svm = model_svm.predict(x_test)
accuracy_score_svm = accuracy_score(y_test, pred_svm)
accuracy_score_svm

# For "Linear"
model_svm = SVC(kernel="linear")
model_svm.fit(x_train, y_train)
pred_svm = model_svm.predict(x_test)
accuracy_score_svm = accuracy_score(y_test, pred_svm)
accuracy_score_svm
```

f) AdaBoost Classifier

```
from sklearn.ensemble import AdaBoostClassifier

model_ada = AdaBoostClassifier(n_estimators=200, learning_rate=0.03)
model_ada.fit(x_train, y_train)
pred_ada = model_ada.predict(x_test)
accuracy_score_ada = accuracy_score(y_test, pred_ada)
accuracy_score_ada
```

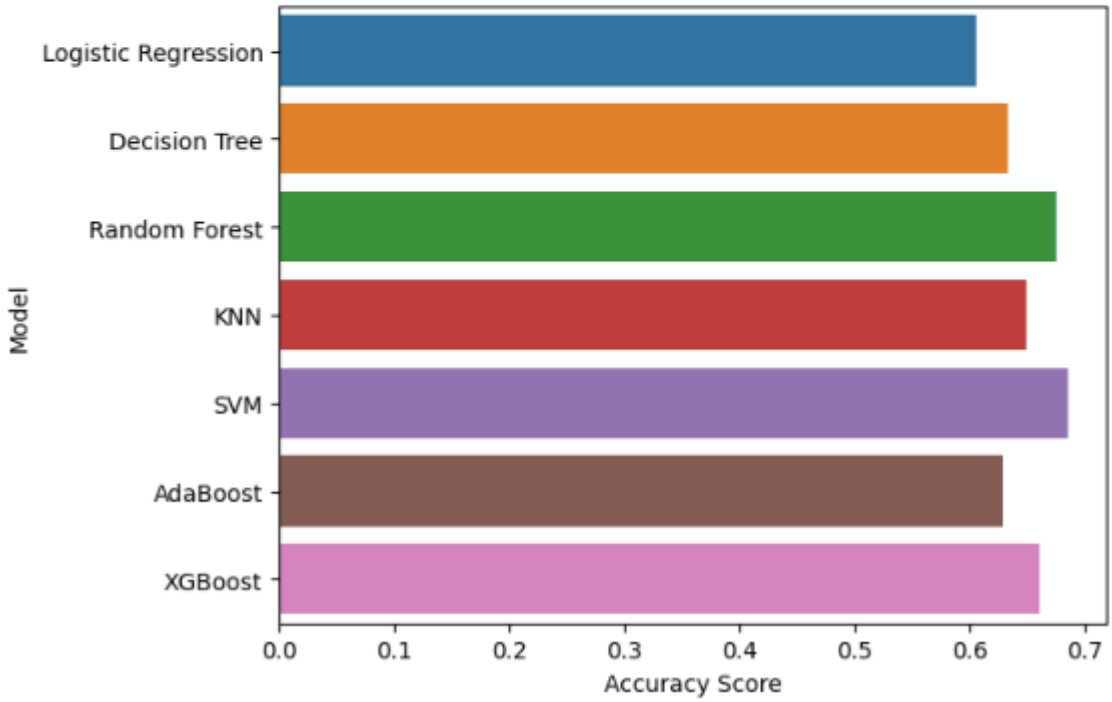
g) XGBoost Classifier

```
from xgboost import XGBClassifier
model_xgb = XGBClassifier(n_estimators=200, learning_rate=0.03)
model_xgb.fit(x_train,y_train)
pred_xgb = model_xgb.predict(x_test)
accuracy_score_xgb = accuracy_score(y_test, pred_xgb)
accuracy_score_xgb
```

7.2 Model Comparison

```
sns.barplot(x="Accuracy Score", y="Model", data= models)
models.sort_values(by="Accuracy Score", ascending= False)
```

	Model	Accuracy Score
4	SVM	0.684451
2	Random Forest	0.675305
6	XGBoost	0.660061
3	KNN	0.649390
1	Decision Tree	0.632622
5	AdaBoost	0.628049
0	Logistic Regression	0.605183



In the SVM model, I have achieved the highest accuracy score.

8. Conclusions

- After assigning the concentration ranges for each of the parameters in the proposed classes, an attempt is made to transform the information on water quality data in discrete terms.
- WQI is a single numerical value used for determining the quality of water for human consumption.
- **OIP** is used for calculation of WQI.
- WQI is also estimated by **WAWQI** method.

- India implements mainly two methods to determine the level of water pollution of reservoirs namely: OIP & WAWQI.
- Reasons for using WAWQI are: includes all the parameters in a single equation, units and dimensions are transformed to a common scale, reduces errors & approaches towards the better results.
- Reasons for using so many parameters are: point source pollution, disposal of untreated sewage, littering, disposing of ritual materials, human activities, oil seepage and agricultural pollutants, urbanization, domestic waste etc.
- Also one of the reasons could be the Mutagenic activities in the river due to the pesticide pollution in the rivers.

- **Site-1,**we can see that for all the three years the water quality status is falling under the Good category, which means that water can be used for drinking after disinfection.
- **Site-4,** we can see that for all the three years the water quality status is unsuitable for the use, which means that water can be used only for the industrial and agricultural purposes.
- **At Site-7,** we can see that the water quality status has been Poor for all three years, indicating that water can be used for consumption after conventional treatment and disinfection.

- It is also clear from the graph that there aren't many swings because the numbers are all quite close to one another. (for site-1 & site7).
- Site-4 may be contaminating other sites. Because site-4 values after computations are substantially higher than 100.The highest value recorded was 1090 in November 2022.
- Calculations are primarily affected by pH, electrical conductivity, and BOD, as shown by analysis of variance. Three main factors can cause high conductivity in the water: Salinity, Water Flow and Temperature.
- The final results matched the PPCB exactly. It is safe to say that water contamination in the Sutlej river is getting worse every year.

9. Software used

Jupyter Notebook,

VS code,

Excel,

Google colab,

Programming language used: - Python.

10. References

CPCB ENVIS | Control of Pollution,Water Quality | Punjab Pollution Control Board, Government of Punjab,

Water quality assessment in terms of water quality index (WQI): case study of the Kolong River, Assam, India | Applied Water Science (springer.com),

Development of an Overall Index of Pollution for Surface Water Based on a General Classification Scheme in Indian Context | Request PDF (researchgate.net),

(PDF) Water quality assessment using overall index of pollution in riverbed-mining area of Ganga-River Haridwar, India (researchgate.net),

Primary_Water_Quality_Criteria.pdf (cpcbenvis.nic.in),

Mutagenic activity of the Ganges water with special reference to the pesticide pollution in the river between Kachla to Kannauj (U.P.), India | Request PDF (researchgate.net),

(PDF) River water pollution, sources, and human health effect in India (researchgate.net),

(PDF) Analysis and Prediction of Ground Water Quality of Belpahar area, Odisha, India on performance of ANN Method (researchgate.net),

(PDF) Various methods for calculating the water quality index (researchgate.net),

Factor analysis and spatial distribution of water quality parameters of Aurangabad District, India. - ScienceDirect,

Water Quality | Punjab Pollution Control Board, Government of Punjab,

CPCB ENVIS | Control of Pollution,

U4SSC KPI – United for Smart Sustainable Cities (U4SSC) (itu.int),

itu.int/dms_pub/itu-t/opb/tut/T-TUT-SMARTCITY-2021-26-PDF-E.pdf,

A novel taxonomy of smart sustainable city indicators | Humanities and Social Sciences Communications (nature.com),

Collection Methodology for Key Performance Indicators for Smart Sustainable Cities | UNECE,

Choosing the right estimator — scikit-learn 1.2.2 documentation,

Search Publications | ResearchGate,ScienceDirect.com | Science, health and medical journals, full text articles and books.