EFFECT OF SAMPLING TECHNIQUES IN WATER QUALITY DATA

Vamsikrishna A* , Pragya Pandey, Sanika Mhamunkar, Rutuja Vartak, Shruti Deshmukh, Abhishek Singh, Pushpak Bhonde

Corresponding Student email-id* : vamsikrishna.a@ssi.edu.in

Group 5 : Ecominions

ABSTRACT

Water quality is one of the most critical factors in a healthy ecosystem. Many factors influence water quality. By observing and evaluating several water samples, we come across different results; however when identifying the correct sampling techniques and inferring the effects to the Population sample, we see the various changes happening across and we study these changes across other sampling techniques here. We used three different Sampling techniques Random Sampling, Systematic Sampling, and Convenient sampling. The results in the study show no absolute winner in the case of choosing the proper sampling technique. We conclude in the study that different sampling techniques depict different results, and no one sampling technique depicts the whole population.

INTRODUCTION

What is Sampling?

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the characteristics of the whole population. It is a time-convenient and cost-effective method of conducting research on the population.

For example, if a drug manufacturer would like to research the adverse side effects of a drug on the country's population, it is almost impossible to conduct a research study that involves everyone. In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug's behavior.

Why do we need Sampling?

Sampling is done to draw conclusions about populations from samples, and it enables us to determine a population's characteristics by directly observing only a portion (or sample) of the population.

- Selecting a sample requires less time than selecting every item in a population.
- Sample selection is a cost-efficient method.
- Analysis of the sample is less complicated and more practical than an analysis of the entire population.

Types of sampling: Sampling methods

There are two types of sampling used in research.

- Probability sampling :- This Sampling methodology uses randomization to ensure that every member of the population has an equal probability of being included in the selected sample.

All the members have an equal opportunity to be a part of the sample with this selection parameter.

- Non-probability sampling :- Unlike probability sampling, Non-Probability Sampling doesn't rely on randomization. This method is highly dependent on the researcher's ability to choose items for a sample. The outcome of sampling may be skewed, making it difficult for all aspects of the population to be included in the sample equitably.

Following are the types of Probability sampling:

Simple Random Sampling: This type of sampling is one of the best probability sampling techniques where every element has an equal probability of being chosen as a part sample. It is applied when we do not have any prior knowledge about the target demographic.

For example, out of 100 people, 20 are selected randomly. Here each person has an equal probability of getting selected. The probability of selection is 1/100.

One big advantage of this technique is that it is the most direct method of probability sampling. But it comes with a limitation – it may not select enough individuals with our characteristics of interest.

Systematic Sampling: In this type of sampling, the first individual is selected randomly and others are selected using a fixed 'sampling interval'. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.

For example, a researcher can give a survey to every fourth customer that comes in to the movie theatre.

Systematic sampling is more convenient than simple random sampling. However, it might also lead to bias if there is an underlying pattern in which we are selecting items from the population (though the chances of that happening are quite rare).

Stratified Sampling: In this type of sampling, we divide the population into subgroups (called strata) based on different traits like gender, category, etc. And then we select the sample(s) from these subgroups.

For example, a researcher looking to analyze people from different socioeconomic backgrounds can distinguish respondents into their annual salaries.

We use this type of sampling when we want representation from all the subgroups of the population. However, stratified sampling requires proper knowledge of the characteristics of the population.

Cluster Sampling: In a clustered sample, we use the subgroups of the population as the sampling unit rather than individuals. The population is divided into subgroups, known as clusters, and a whole cluster is randomly selected to be included in the study.

For example, if the United States government wishes to evaluate the number of immigrants living in the Mainland US, they can divide it into clusters based on states such as California, Texas, Florida, Massachusetts, Colorado, Hawaii, etc.

This way of conducting a survey will be more effective as the results will be organized into states and provide insightful immigration data. However, this type of sampling is used when we focus on a specific region or area.

Following are the types of Non-Probability sampling:

Convenience Sampling: This is perhaps the easiest method of sampling because individuals are selected based on their availability and willingness to take part.
For example, startups and NGOs usually conduct convenience sampling at a mall to distribute leaflets of upcoming events or promotion of a cause – they do that by standing at the mall entrance and giving out pamphlets randomly.

Convenience sampling is prone to significant bias because the sample may not be the representation of the specific characteristics such as religion or, say the gender, of the population.

Quota Sampling: In this type of sampling, we choose items based on predetermined characteristics of the population. Elements are chosen until correct quantities of particular sorts of data are achieved or until adequate data in various categories is gathered.

For example, if our population is composed of 65 percent females and 35 percent men, our sample should be composed of the same proportion of males and females.

In quota sampling, the chosen sample might not be the best representation of the characteristics of the population that weren't considered.

What is Water Quality?

Water quality describes the condition of the water, including chemical, physical and biological characteristics, usually with respect to its suitability for a particular purpose such as drinking or swimming. Poor water quality can also pose a health risk for ecosystems.
Drinking water quality varies from place to place, depending on the condition of the source water from which it is drawn and the treatment it receives. Water quality is measured by several factors, such as the concentration of dissolved oxygen, temperature, pH, bacteria levels, the amount of salt (or salinity), or the amount of material suspended in the water (turbidity).
Water, whether it is meant for business purposes, agriculture, domestic purposes, or is used by public municipalities and private homeowners must be tested regularly in order to keep the source of water safe and free from environmental risks and potential health disorders.

Importance of Water Quality and Testing

In the whole world, a majority of the people rely on the private water supply. This includes ponds, dugouts, and wells. A superior quality of water is crucial to the economic, health, and social well-being of the people. Monitoring the quality of your water and testing it regularly is very important to maintain reliable and safe water sources and eliminate the potential health risks related to water contamination.

Water testing is carried out to meet the regulatory requirements and adhere to the safety procedures that are needed for pollutant-free water. When the water is tested it offers the knowledge; we require to address the problem that is currently involved with the water quality. It will also ensure that the water quality is protected from every potential cause of contamination and an appropriate approach is involved with the treatment system.

It is vital to check the suitability of the water quality before its use. It can be for irrigation, livestock watering, drinking, or spraying. It will also help you in making an informed decision about how to use the water and what should be done about its purity.

OBJECTIVE

The main objective of this project were following :

1. Comparison of population data to that of sample using various sampling techniques namely :- Convenient, Random and Systematic Sampling.
2. Use of sampling techniques in Polynomial Regression.
3. Comparison of Mean Square Error versus Order of sample to that of population
4. Comparison of general characteristics of population with each of the sampling techniques.
5. Comparison of Correlation matrix of population and sample.

LITERATURE SURVEY

M.G. Kelly et al depict the principles and practice of sampling in relation to the main habitats in Europe. The authors have provided guidelines and recommendations that can be adapted for most of the river types in Europe. [1] Santosh Kumar Sarkar et al have examined the physicochemical characteristics and level of dissolved heavy metals at Babughat, Diamond Harbor, and Gangasagar. The authors suggest an integrated management approach to meet the needs. [2] T.R.Giriraj et al have studied the water quality assessment of a relatively small tributary of the Brahmaputra called the Bharalu River. According to the study, dissolved oxygen was absent in several locations irrespective of the season. [3] Md. J. B. Alam et al have analyzed various water quality parameters of the Surma river during dry and monsoon periods. The study concludes that the river was highly turbid during the monsoon season and acidic during the dry season. [4] Shweta Tyagi et al have attempted to review the WQI criteria for the appropriateness of drinking water sources. The study infers that the aim of WQI is to give a single value to the water quality for easy interpretation of a source. [5] The study presented by Arivoli Appavo et al reveals how the Cauvery river water is contaminated by effluents from small-scale industries and dumping of wastages from markets and domestic use wastages. [6] Gaganpreet Sharma has described different techniques and types of sampling along with their pros and cons. [7] Sanjay Datta has explained the concept of sampling methods and different types of sampling in an elaborate manner. [8]

LITERATURE GAP

In the literature, few areas were kept untouched. For the fact that we should use the proper sampling technique for the given population, we should have comparative studies on the sampling techniques on the practical data. Hence the three important that we did not find in the literature and that we explored in our work were:

1. Effect of a sample compared to the population
2. Effect of sampling techniques in regression analysis
3. Effect of sampling factors affecting the pH and water quality.

METHODOLOGY

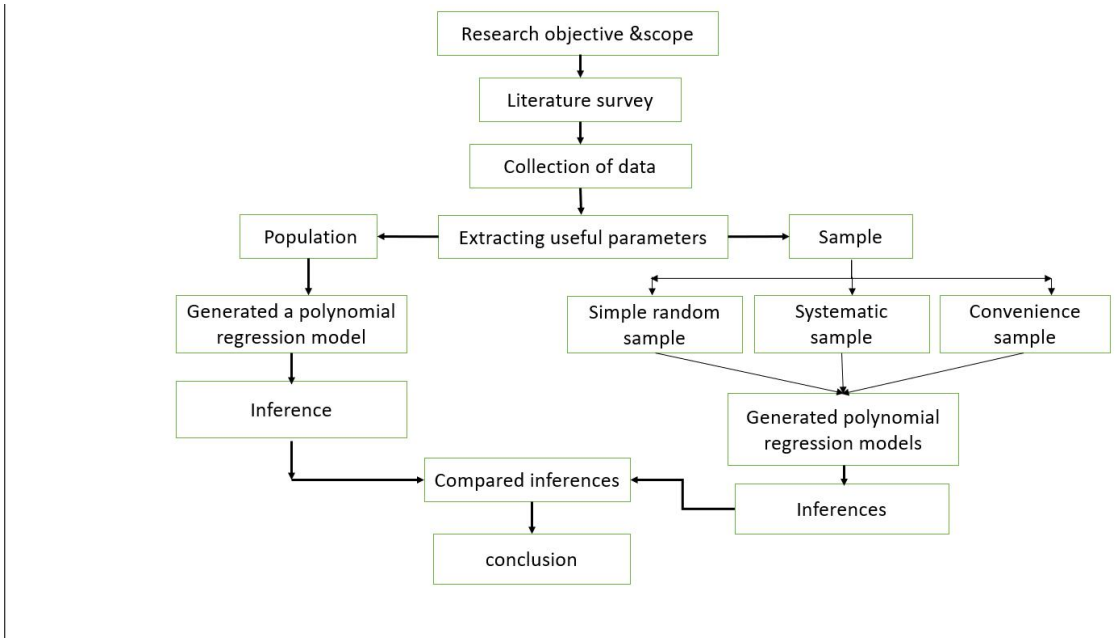


FIG 1: Methodology flowchart.

This project looks at the various research methodologies and research methods that are commonly used by researchers in the field of statistics. The research methodology and research method used in this project is acknowledged and discussed.

The project starts off by providing a comprehensive introduction to research. Then the research methodologies and research methods particularly used in sampling theory are discussed. The research methods in general are discussed, and the types of research methods suitable for sampling theory are explained.

In this project, our objective was to find out the various factors affecting the pH of river water via statistical sampling and to understand the effect of sampling in the population and not to get perfect results. We approached this project by firstly selecting a population data and then created a polynomial regression model for order two by using python programming language. After which we have derived general characteristics, compared distribution plots, calculated the mean and standard deviation for actual and fitted values of the population. We have also formed a correlation matrix. This correlation matrix gives us the correlation between the different variables from the data. And also plotted order vs MSE graph values of the population.

Furthermore, we started with a sampling process by clearly defining the target population. Population is commonly related to the number of people living in a particular country. After that we limited ourselves to select sampling techniques namely simple random sampling, systematic sampling and convenience sampling. And then we again created models for simple random sampling, systematic sampling and convenience sampling using the python programming language. After which, we have derived general characteristics, compared distribution plots, calculated the mean and standard deviation for actual and fitted values of the samples. We have also formed a correlation matrix, and also plotted order vs. MSE graph values of the samples. And then we compared general characteristics, distribution plots, correlation matrix and order vs. MSE graphs of simple random sampling, systematic sampling and convenience sampling with the population. And based on this comparison we concluded our experiment.

RESULTS AND DISCUSSIONS

Part 1 (Population Vs. Convenient): Convenient sampling method is used to sample from the population such that they are convenient for the statistical analyzer. In this project, we are limiting ourselves in taking a convenient sample from the population in such a way that we considered only the first 40 entries. It was considered in this manner because the objective of the study was to understand the effect of sampling in the population and not to get perfect results. (NOTE: Getting perfect results is for practical implementation purposes, but here we are just studying the effect of sampling techniques on the population).

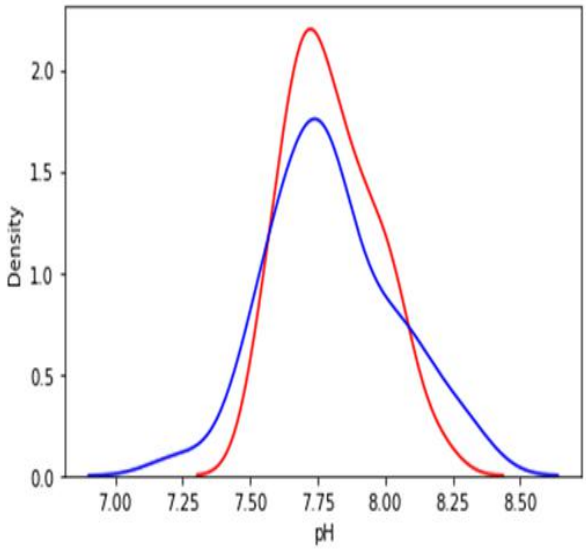


FIG 2.A : - Distribution plot for convenient sampling

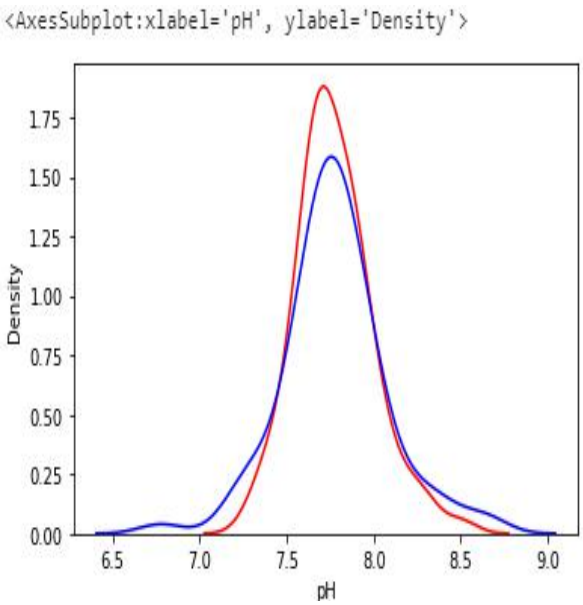


FIG 2.B- Distribution plot for Population; Red-Actual, Blue-Fitted

In Fig-2:A, the plot depicts that the population is spread closely symmetrical with σ(0.1645) standard deviations around the mean value (7.800375). The bell-shaped curve evidence that it is close

to normal distribution. Confidence intervals represent the range of values between which we are fairly sure that our population means lies. The lower limit being 7.3068 and upper limit 8.2938. Fig-2:B again is a normal distribution with mean 7.7742 and standard deviation 0.2202. Both the lower limit being 7.1133 and upper limit is 8.4351.

In both the graphs, we see a positive correlation till a set point is reached. The pH increases as the density increases because of the increase in dissolved oxygen, nitrate, BCOD etc. Both the graphs show that data near the mean are more frequent in occurrence than data far from the mean.

```
from sklearn.metrics import mean_squared_error

mean_squared_error(df['pH'],yhat)
```

```
0.03742109970092773
```

FIG 3.A: Mean Square Error for actual and predicted values of the Convenient Sampling

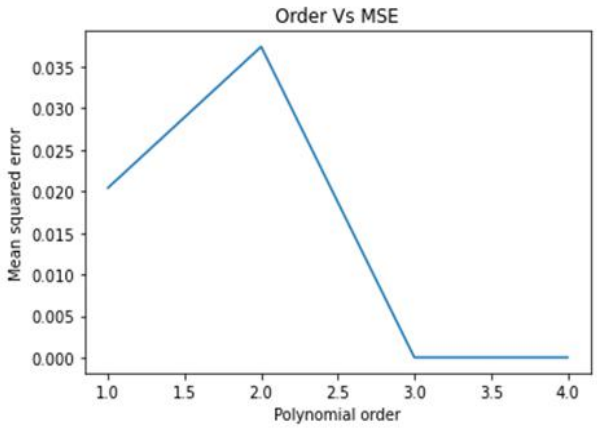
```
]: from sklearn.metrics import mean_squared_error

   mean_squared_error(df['pH'],yhat)
```

```
]: 0.06989782776773712
```

FIG 3.B: Mean Square Error for actual and predicted values of the population

The mean squared error (MSE) tells how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. The squaring is necessary to remove any negative signs. It can be observed that the MSE is 0.069 for the population while it is 0.037 for Convenient sampling. Since 0.037<0.069, the predicted values fit better for the convenient sample compared to the population.



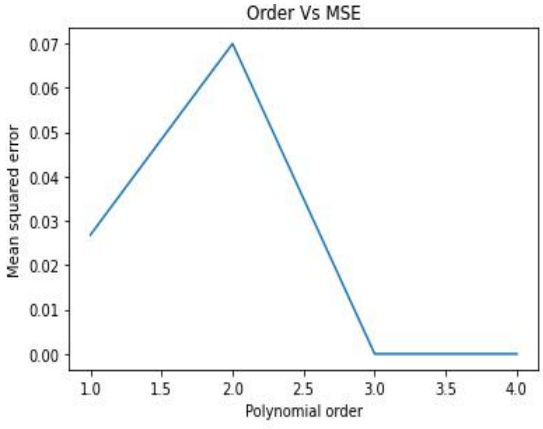


FIG 4.A: MSE vs Order - Convenient sample

FIG 4.B: MSE vs Order – Population

In the Order Vs. MSE graph it is clearly visible that both the convenient sample and the population show that MSE for polynomial order 2 and the trend the graph follows the same pattern. However, its MSE peak values are different at the order 2 as discussed before.

TABLE 1.A : General characteristics of the Convenient Sample

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.00000	40.000000
mean	15.002500	8.920000	7.800375	343.072500	0.603750	0.501500	127.93875	620.962500
std	5.062076	0.738571	0.164529	414.485551	0.480169	0.318317	320.70610	2021.576674
min	7.500000	7.600000	7.540000	77.000000	0.000000	0.100000	2.00000	2.000000
25%	10.400000	8.200000	7.700000	144.625000	0.250000	0.250000	13.87500	108.000000
50%	14.500000	8.900000	7.775000	215.750000	0.375000	0.485000	38.00000	232.500000
75%	20.000000	9.650000	7.900000	255.250000	1.087500	0.700000	89.25000	437.500000
max	26.000000	10.250000	8.200000	1453.500000	1.550000	1.300000	2000.00000	13000.000000

TABLE 1.B : General Characteristics of Population

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
count	81.000000	81.000000	81.000000	81.000000	81.000000	81.000000	81.000000	8.100000e+01
mean	16.071605	8.080247	7.774259	590.665432	4.308025	1.126049	15146.684568	3.441897e+04
std	5.225646	2.074935	0.220285	571.581527	13.643424	1.233695	97158.676076	2.060708e+05
min	7.500000	0.000000	7.300000	77.000000	0.000000	0.100000	2.000000	2.000000e+00
25%	11.000000	7.950000	7.650000	214.000000	0.250000	0.350000	22.000000	1.795000e+02
50%	16.500000	8.600000	7.750000	275.500000	0.600000	0.700000	55.900000	4.100000e+02
75%	20.500000	9.250000	7.900000	1012.500000	1.200000	1.300000	300.000000	8.470000e+02
max	26.000000	10.250000	8.500000	2630.000000	84.000000	6.350000	863500.000000	1.789500e+06

In the tables, we can see the general characteristics of the Population and Convenient Samples. The difference between them is clearly visible. We can see that the mean temperature, BCOD and nitrate have come down while for other columns it has risen. Even with respect to the std dev, we can see a lot of differences. This is due to the fact that a lot of data is lost while choosing the sample. This is one of the major drawbacks of any kind of sampling technique.

TABLE 2.A: Correlation matrix for Convenient Sample

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
temperature	1.000000	-0.722687	0.177139	-0.331792	0.301065	-0.082944	0.497782	0.340922
dissolved oxygen	-0.722687	1.000000	0.017978	-0.025608	-0.549710	-0.132889	-0.160198	0.002444
pH	0.177139	0.017978	1.000000	-0.050187	-0.308840	-0.092017	0.152604	0.192330
conductivity	-0.331792	-0.025608	-0.050187	1.000000	0.298225	0.693811	-0.102892	-0.030847
BCOD	0.301065	-0.549710	-0.308840	0.298225	1.000000	0.359633	0.025033	-0.150733
nitrate	-0.082944	-0.132889	-0.092017	0.693811	0.359633	1.000000	0.012156	0.019612
faecal coliform	0.497782	-0.160198	0.152604	-0.102892	0.025033	0.012156	1.000000	0.960693
total coliform	0.340922	0.002444	0.192330	-0.030847	-0.150733	0.019612	0.960693	1.000000

TABLE 2.B : Correlation Matrix for Population

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
temperature	1.000000	-0.581304	-0.228084	-0.356913	0.369034	0.351529	0.215403	0.227615
dissolved oxygen	-0.581304	1.000000	0.527503	-0.204633	-0.838554	-0.394366	-0.370665	-0.381674
pH	-0.228084	0.527503	1.000000	0.102397	-0.451255	-0.285431	-0.240923	-0.257665
conductivity	-0.356913	-0.204633	0.102397	1.000000	0.201083	0.061895	0.061737	0.058369
BCOD	0.369034	-0.838554	-0.451255	0.201083	1.000000	0.325552	0.559099	0.561209
nitrate	0.351529	-0.394366	-0.285431	0.061895	0.325552	1.000000	0.112182	0.122953
faecal coliform	0.215403	-0.370665	-0.240923	0.061737	0.559099	0.112182	1.000000	0.991509
total coliform	0.227615	-0.381674	-0.257665	0.058369	0.561209	0.122953	0.991509	1.000000

Though a polynomial regression model was constructed and evaluated with a target variable as pH, it is still necessary to know the major contributors that affect the pH values. From the tables we see that dissolved oxygen and BCOD has the highest influence on the pH value of the water for the population wherein the dissolved oxygen has a strong positive correlation while BCOD has a strong negative effect. Coming to the correlation matrix of the sample, we see that pH is most influenced by BCOD and Total coliform wherein the BCOD has a negative effect while the Total coliform has a positive effect.

Part 2 (Population Vs. Random)

Random sampling is a sampling method that allows for the randomization of sample selection and in which every member of the population has an equal chance of being selected. In this experiment, we selected a random sample of size 40 from a population of 81 using a simple random sampling technique. We generated random numbers with the help of EXCEL. The objective here is to properly understand the effect of sampling in the population and not to get the perfect results.

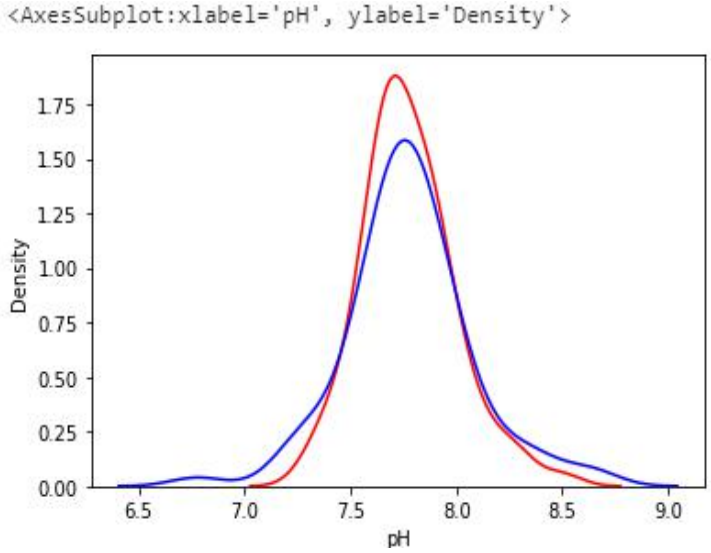


FIG 5. A: Distribution plot for Population; Red-Actual, Blue-Fitted

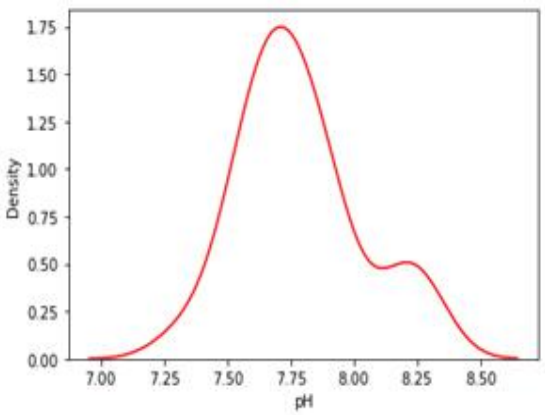


FIG 5.B.1: Actual distribution plot for simple random sampling

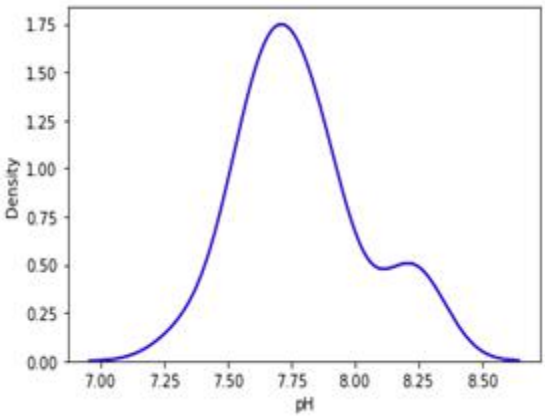


FIG 5.B.2-Fitted distribution plot for simple random sampling.

Fig-5:A is a normal distribution with a mean 7.7742 and a standard deviation 0.2202.).The lower limit being 7.3068 and upper limit 8.2938

In this graph, we see a positive correlation till a set point is reached. The pH increases as the density increases because of the increase in dissolved oxygen, nitrate, BCOD etc. In this graph, the data near the mean is more frequent in occurrence than data far from the mean.

Fig-5.B.1(ACTUAL), Fig-5.B.2(FITTED)-These plots depict that the population is spread asymmetrically with σ(0.240246) standard deviations around the mean value (7.785000). Confidence interval is a range of values so defined that there is a specified probability that the value of a parameter lies within it. Both the limits being minimum (7.300000) and maximum (8.300000). The pH increases as the density increases because of the increase in dissolved oxygen, nitrate. The plots would have appeared as a normal distribution excluding the range(8-8.5) where there is an abnormal distribution.

```
In [25]: from sklearn.metrics import mean_squared_error

         mean_squared_error(df['pH'],yhat)

Out[25]: 1.3785318047401062e-20
```

FIG 6.A: Mean Squared Error for actual and predicted values using simple random sampling

```
In [25]: from sklearn.metrics import mean_squared_error

         mean_squared_error(df['pH'],yhat)

Out[25]: 0.06989782776773712
```

FIG 6.B: Mean Squared Error for actual and predicted values of the population

The mean squared error (MSE) of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk

function corresponding to the expected value of the squared error loss. It can be observed that the MSE is 0.069 for the population while it is approximately 0 for simple random sampling. Since 0<0.069, the predicted values fit best to the simple random sampling compared to the population.

FIG 7.A: MSE vs Order - Random sample

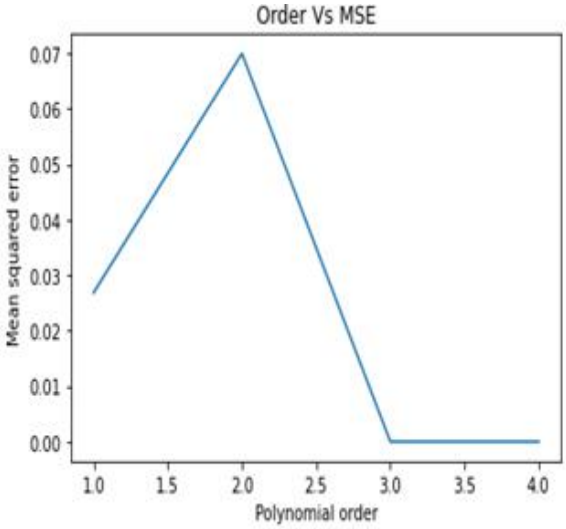


FIG 7.B: MSE vs Order - Population

From fig. 7.A and 7.B, it is visible that for population, MSE is at its peak for polynomial order 2 whereas for random sample it is lowest at the polynomial order 2. Thus, population and random samples show opposite patterns in the MSE vs Polynomial order graph.

TABLE 3-A: General characteristic table for population:

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
count	81.000000	81.000000	81.000000	81.000000	81.000000	81.000000	81.000000	8.100000e+01
mean	16.071605	8.080247	7.774259	590.665432	4.308025	1.126049	15146.684568	3.441897e+04
std	5.225646	2.074935	0.220285	571.581527	13.643424	1.233695	97158.676076	2.060708e+05
min	7.500000	0.000000	7.300000	77.000000	0.000000	0.100000	2.000000	2.000000e+00
25%	11.000000	7.950000	7.650000	214.000000	0.250000	0.350000	22.000000	1.795000e+02
50%	16.500000	8.600000	7.750000	275.500000	0.600000	0.700000	55.900000	4.100000e+02
75%	20.500000	9.250000	7.900000	1012.500000	1.200000	1.300000	300.000000	8.470000e+02
max	26.000000	10.250000	8.500000	2630.000000	84.000000	6.350000	863500.000000	1.789500e+06

TABLE 3-B: General characteristic table for random sample:

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000
mean	15.886250	8.021250	7.785000	681.100000	4.301250	0.877125	4797.396250	10512.532500
std	5.442909	2.289689	0.240246	538.416448	14.044378	0.849145	17127.340008	35717.700005
min	7.500000	0.000000	7.300000	77.000000	0.000000	0.100000	2.000000	40.500000
25%	11.000000	7.912500	7.637500	214.125000	0.250000	0.250000	25.250000	198.750000
50%	16.525000	8.675000	7.750000	291.500000	0.750000	0.700000	51.450000	432.500000
75%	20.500000	9.325000	7.900000	1321.000000	1.300000	0.877500	361.500000	905.000000
max	26.000000	10.250000	8.300000	1442.000000	84.000000	3.400000	93500.000000	179500.000000

From the above tables for population and random sample, for mean values, we can observe that only conductivity has increased while the other factors have decreased. In the case of standard deviation, only BCOD has increased while other factors have decreased.

TABLE 4-A: correlation matrix for population:

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
temperature	1.000000	-0.581304	-0.228084	-0.356913	0.369034	0.351529	0.215403	0.227615
dissolved oxygen	-0.581304	1.000000	0.527503	-0.204633	-0.838554	-0.394366	-0.370665	-0.381674
pH	-0.228084	0.527503	1.000000	0.102397	-0.451255	-0.285431	-0.240923	-0.257665
conductivity	-0.356913	-0.204633	0.102397	1.000000	0.201083	0.061895	0.061737	0.058369
BCOD	0.369034	-0.838554	-0.451255	0.201083	1.000000	0.325552	0.559099	0.561209
nitrate	0.351529	-0.394366	-0.285431	0.061895	0.325552	1.000000	0.112182	0.122953
faecal coliform	0.215403	-0.370665	-0.240923	0.061737	0.559099	0.112182	1.000000	0.991509
total coliform	0.227615	-0.381674	-0.257665	0.058369	0.561209	0.122953	0.991509	1.000000

TABLE 4-B: Correlation matrix for random sample:

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
temperature	1.000000	-0.534587	0.083077	-0.382594	0.326006	0.349088	0.323785	0.322567
dissolved oxygen	-0.534587	1.000000	0.529298	-0.200800	-0.799326	-0.639458	-0.765513	-0.728792
pH	0.083077	0.529298	1.000000	-0.202288	-0.444766	-0.514317	-0.466967	-0.462485
conductivity	-0.382594	-0.200800	-0.202288	1.000000	0.222960	0.195991	0.180187	0.154320
BCOD	0.326006	-0.799326	-0.444766	0.222960	1.000000	0.605829	0.924313	0.848311
nitrate	0.349088	-0.639458	-0.514317	0.195991	0.605829	1.000000	0.532998	0.484696
faecal coliform	0.323785	-0.765513	-0.466967	0.180187	0.924313	0.532998	1.000000	0.984770
total coliform	0.322567	-0.728792	-0.462485	0.154320	0.848311	0.484696	0.984770	1.000000

The factors dissolved oxygen and BCOD are most affecting the pH of the water in the population. From the correlation matrix for population, we can observe that dissolved oxygen and pH are positively correlated whereas BCOD and pH are negatively correlated. In random sample data, we can observe that factors dissolved oxygen and nitrate have a high influence on the pH of the water and the factors such as faecal coliform and total coliform are moderately affecting the pH of the water. In the sampled data only dissolved oxygen and pH are positively correlated and on the other hand nitrate and pH, faecal coliform and pH, total coliform and pH are negatively correlated.

Part 3 (Population Vs. Systematic)

Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval. In this experiment, our desired sample size was 40, so we got an interval of 2 by dividing population size by sample size. The objective here is to properly understand the effect of sampling in the population and not to get the perfect results.

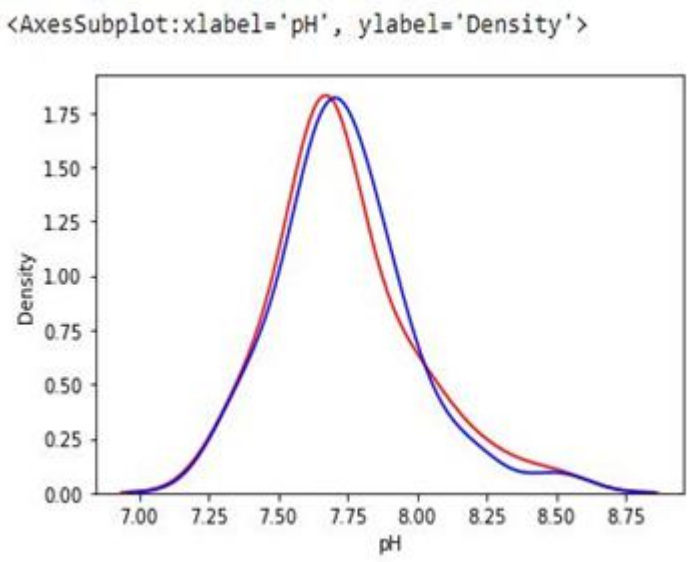


FIG 8.A: Distribution plot for systematic Sampling

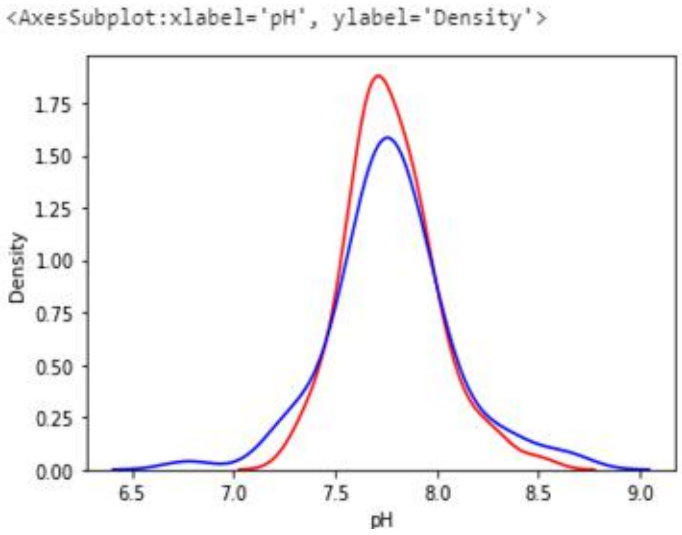


Fig 8.B - Actual distribution plot for Population

In Fig- 8.A, the plot depicts that the population is spread closely symmetrical with 0.2522 standard deviations around the mean value 7.7393 . The plots would have appeared as a normal distribution excluding the range(8.25-8.5) where there is an abnormal distribution. Confidence intervals represent the range of values between which we are fairly sure that our population means lies. The area between confidence intervals is called the acceptance region while the area outside is called the rejection region.

Fig-8.B again is a normal distribution with mean 7.7742 and standard deviation 0.2202. The lower limit is 7.1133 and upper limit 8.4351

In both the graphs, we see a positive correlation till a set point is reached. The pH increases as the density increases because of the increase in dissolved oxygen, nitrate, BCOD
etc.

```
from sklearn.metrics import mean_squared_error
```

```
mean_squared_error(data_df['pH'],yhat)
```

```
0.0076655021619517465
```

FIG 9.A: Mean Square Error for actual and predicted values of the Systematic Sampling

```
]: from sklearn.metrics import mean_squared_error

   mean_squared_error(df['pH'],yhat)
```

```
]: 0.06989782776773712
```

FIG 9.B: Mean Square Error for actual and predicted values of the Population

The mean squared error (MSE) tells how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line and squaring them. The squaring is

necessary to remove any negative signs. It can be observed that the MSE is 0.069 for the population while it is 0.007 for Systematic Sampling. Since 0.007<0.069, the predicted values fit better for the Systematic sample compared to the population.

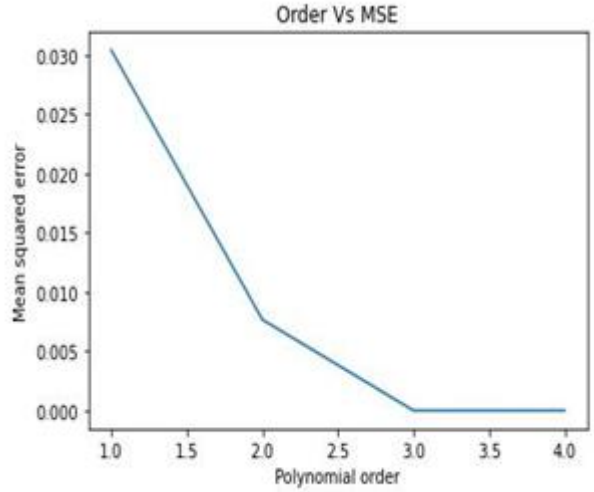


FIG 10.A: MSE vs Order - Systematic sample

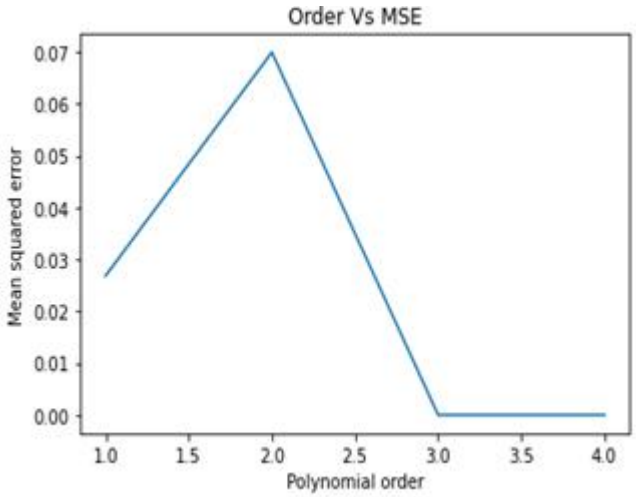


FIG 10.B: MSE vs Order - Population

From fig. 10.A and 10.B, it is visible that for population, MSE is at maximum for polynomial order 2 and starts decreasing from 2, whereas for Systematic sample as polynomial order increases, MSE decreases. Thus, population and Systematic samples show opposite patterns in the MSE vs. Polynomial order graph.

Table 5.A:- General Characteristics of Systematic Sample

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	4.000000e+01
mean	16.031250	7.956250	7.739375	629.650000	5.947500	1.039625	28945.381250	6.611194e+04
std	5.067059	2.225095	0.252226	586.439203	16.698031	1.170880	137564.842539	2.914164e+05
min	8.500000	0.000000	7.300000	77.000000	0.000000	0.115000	2.000000	2.000000e+00
25%	11.000000	8.100000	7.600000	219.125000	0.250000	0.375000	20.150000	1.172500e+02
50%	16.750000	8.625000	7.700000	281.500000	0.600000	0.700000	53.950000	3.350000e+02
75%	20.500000	9.025000	7.862500	1203.875000	1.200000	1.200000	257.625000	8.233750e+02
max	24.000000	10.250000	8.500000	2619.000000	84.000000	6.350000	863500.000000	1.789500e+06

Table 5.B:- General Characteristics for population

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
count	81.000000	81.000000	81.000000	81.000000	81.000000	81.000000	81.000000	8.100000e+01
mean	16.071605	8.080247	7.774259	590.665432	4.308025	1.126049	15146.684568	3.441897e+04
std	5.225646	2.074935	0.220285	571.581527	13.643424	1.233695	97158.676076	2.060708e+05
min	7.500000	0.000000	7.300000	77.000000	0.000000	0.100000	2.000000	2.000000e+00
25%	11.000000	7.950000	7.650000	214.000000	0.250000	0.350000	22.000000	1.795000e+02
50%	16.500000	8.600000	7.750000	275.500000	0.600000	0.700000	55.900000	4.100000e+02
75%	20.500000	9.250000	7.900000	1012.500000	1.200000	1.300000	300.000000	8.470000e+02
max	26.000000	10.250000	8.500000	2630.000000	84.000000	6.350000	863500.000000	1.789500e+06

In the tables, we can see the general characteristics of the Population and Systematic Sample. The difference between them is clearly visible. We can see that the mean temperature, dissolved oxygen, pH and nitrate have come down while for other columns, it has risen. Even with respect to the std dev, we can see a lot of differences. This is due to the fact that a lot of data is lost while choosing the sample.

Table 6-A:- Correlation matrix of Systematic Sample

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
temperature	1.000000	-0.605652	-0.290467	-0.346920	0.448231	0.304130	0.306161	0.326684
dissolved oxygen	-0.605652	1.000000	0.545686	-0.215948	-0.877890	-0.440434	-0.452613	-0.474665
pH	-0.290467	0.545686	1.000000	0.185518	-0.465888	-0.248185	-0.262527	-0.285224
conductivity	-0.346920	-0.215948	0.185518	1.000000	0.209144	0.122824	0.064980	0.061309
BCOD	0.448231	-0.877890	-0.465888	0.209144	1.000000	0.426503	0.610567	0.619628
nitrate	0.304130	-0.440434	-0.248185	0.122824	0.426503	1.000000	0.171603	0.192683
faecal coliform	0.306161	-0.452613	-0.262527	0.064980	0.610567	0.171603	1.000000	0.991475
total coliform	0.326684	-0.474665	-0.285224	0.061309	0.619628	0.192683	0.991475	1.000000

Table 6-B :- Correlation matrix of population

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
temperature	1.000000	-0.581304	-0.228084	-0.356913	0.369034	0.351529	0.215403	0.227615
dissolved oxygen	-0.581304	1.000000	0.527503	-0.204633	-0.838554	-0.394366	-0.370665	-0.381674
pH	-0.228084	0.527503	1.000000	0.102397	-0.451255	-0.285431	-0.240923	-0.257665
conductivity	-0.356913	-0.204633	0.102397	1.000000	0.201083	0.061895	0.061737	0.058369
BCOD	0.369034	-0.838554	-0.451255	0.201083	1.000000	0.325552	0.559099	0.561209
nitrate	0.351529	-0.394366	-0.285431	0.061895	0.325552	1.000000	0.112182	0.122953
faecal coliform	0.215403	-0.370665	-0.240923	0.061737	0.559099	0.112182	1.000000	0.991509
total coliform	0.227615	-0.381674	-0.257665	0.058369	0.561209	0.122953	0.991509	1.000000

Though a polynomial regression model was constructed and evaluated with a target variable as pH, it is still necessary to know the major contributors that affect the pH values. From the tables we see that dissolved oxygen and BCOD has the highest influence on the pH value of the water for the population as well as the Systematic Sample wherein the dissolved oxygen has a strong positive correlation while BCOD has a strong negative effect.

CONCLUSIONS

TABLE 7 : Conclusion

Parameters	Population	Convenient	Random	Systematic
Distribution Plot	Bell shaped curve where the actual values and the fitted values are slightly overlapping each other and the actual values are at the peak.	Bell shaped curve where the actual values and the fitted values are slightly overlapping each other	Bell shaped curve where the actual values and the fitted values are exactly overlapping each other	Bell shaped curves overlapping each other.
Mean Squared Error	0.06989782	0.37421099	1.37853180	0.00766550

General Characteristics		Mean, temperature, BCOD and nitrate have come down while for other columns it has risen up.	Only conductivity has increased while other factors have decreased. In the case of standard deviation, only BCOD has increased while other factors have decreased.	We can see that the mean temperature, dissolved oxygen, pH and nitrate have come down while for other columns it has risen.
Order vs MSE	MSE for polynomial of order 2 peaked at 0.07	MSE for polynomial of order 2 peaked at 0.035	MSE is the lowest at the polynomial of order 2	As polynomial order increases MSE decreases
Correlation Matrix	Dissolved oxygen and BCOD has the highest influence on the pH value of the water for the population wherein the dissolved oxygen has a strong positive correlation while BCOD has a strong negative effect.	The pH is most influenced by BCOD and Total coliform wherein the BCOD has a negative effect while the Total coliform has a positive effect.	The dissolved oxygen and nitrate have a high influence on the pH of the water and only dissolved oxygen and pH are positively correlated.	Dissolved oxygen and BCOD have the highest influence on the pH value of the water and the dissolved oxygen has a strong positive correlation while BCOD has a strong negative effect.

Therefore, we observe that for different parameters, different sampling techniques are representing the entire population. There is no one sampling technique that completely depicts the whole population. For instance, if we take the Correlation matrix in consideration, the systematic sampling has completely shown the properties of the population. The population and the systematic sampling are both affected by the dissolved oxygen and BCOD where the dissolved oxygen has a strong positive correlation, whereas the BCOD has a strong negative correlation. If we take the Order vs MSE plot into consideration, we observe that the convenient sampling has exactly the same graph to that of the population. Both are highest for the polynomial of order 2. So, we can not come to the conclusion that any particular type of sampling is the best for the sampling process.

LIMITATIONS

1. Though the target population was river data across India, the data has less representation of most rivers.
2. We were not able to find the coefficients of polynomial regression.
3. Other Sampling techniques were not explored.

REFERENCES

[1]Kelly, M.G., Cazaubon, A., Coring, E., Dell'Uomo, A., Ector, L., Goldsmith, B., Guasch, H., Hürlimann, J., Jarlman, A., Kawecka, B. and Kwandrans, J., 1998. Recommendations for the routine sampling of diatoms for water quality assessments in Europe. *Journal of applied Phycology*, *10*(2), pp.215-224.

[2] Sarkar, S.K., Saha, M., Takada, H., Bhattacharya, A., Mishra, P. and Bhattacharya, B., 2007. Water quality management in the lower stretch of the river Ganges, east coast of India: an approach through environmental education. *Journal of Cleaner Production*, *15*(16), pp.1559-1567.

[3]Girija, T.R., Mahanta, C. and Chandramouli, V., 2007. Water quality assessment of an untreated effluent impacted urban stream: the Bharalu tributary of the Brahmaputra River, India. *Environmental monitoring and assessment*, *130*(1), pp.221-236.

[4]Alam, M.J., Islam, M.R., Muyen, Z., Mamun, M. and Islam, S., 2007. Water quality parameters along rivers. *International Journal of Environmental Science & Technology*, *4*(1), pp.159-167.

[5]Tyagi, S., Sharma, B., Singh, P. and Dobhal, R., 2013. Water quality assessment in terms of water quality index. *American Journal of water resources*, *1*(3), pp.34-38.

[6]Appavu, A., Thangavelu, S., Muthukannan, S., Jesudoss, J.S. and Pandi, B., 2016. Study of water quality parameters of Cauvery river water in erode region. *Journal of Global Biosciences*, *5*(9), pp.4556-4567.

[7]Sharma, G., 2017. Pros and cons of different sampling techniques. *International journal of applied research*, *3*(7), pp.749-752.

[8]Datta, S., 2018. Sampling methods. *Biostatistics & Computer Application*.

[9]https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/

[10]https://www.h2olabcheck.com/blog/view/why-is-it-important-to-test-water

Appendix: (python codes)

```
In [1]: import pandas as pd
```

```
In [2]: df=pd.read_csv(r"C:\Users\Vamsikrishna\Desktop\WaterQuality.csv")
        df.head()
```

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
0	7.5	9.95	7.85	134.5	0.15	0.255	22.5	180.0
1	11.0	9.65	7.70	77.0	0.45	0.200	62.5	410.0
2	7.5	9.90	7.65	101.5	0.55	0.100	26.0	200.0
3	8.5	9.65	7.55	148.0	0.35	0.250	97.5	600.0
4	10.0	9.55	7.80	106.0	0.25	0.150	47.5	380.0

```
In [7]: from sklearn.linear_model import LinearRegression
```

```
In [8]: from sklearn.preprocessing import StandardScaler
```

```
In [9]: from sklearn.pipeline import Pipeline
        from sklearn.preprocessing import PolynomialFeatures
```

```
In [10]: Input=[('sclae',StandardScaler()),('Polynomial',PolynomialFeatures(degree=2)),('mode',LinearRegression())]
```

```
In [11]: pipe=Pipeline(Input)
```

```
In [12]: pipe.fit(df[['temperature','dissolved oxygen','conductivity','BCOD','nitrate','faecal coliform','total coliform']],df['pH'])
```

```
Out[12]: Pipeline(steps=[('sclae', StandardScaler()),
                         ('Polynomial', PolynomialFeatures()),
                         ('mode', LinearRegression())])
```

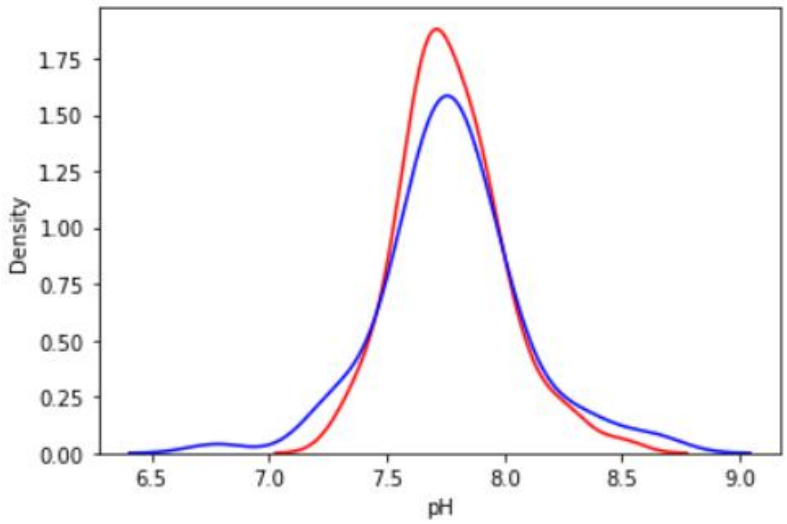
```
In [13]: yhat=pipe.predict(df[['temperature','dissolved oxygen','conductivity','BCOD','nitrate','faecal coliform','total coliform']])
```

```
In [14]: yhat
```

```
In [17]: import seaborn as sns

         ax1= sns.distplot(df['pH'],hist=False,color='r',label="actual value")
         sns.distplot(yhat, hist=False, color='b',label="Fitted Values",ax=ax1)
```

Out[17]: <AxesSubplot:xlabel='pH', ylabel='Density'>



In [18]:
```
df.corr() #pH is being predicted. clearly visible that 'dissolved oxygen' and 'BCOD' is influencing pH the most
```

Out[18]:

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
temperature	1.000000	-0.581304	-0.228084	-0.356913	0.369034	0.351529	0.215403	0.227615
dissolved oxygen	-0.581304	1.000000	0.527503	-0.204633	-0.838554	-0.394366	-0.370665	-0.381674
pH	-0.228084	0.527503	1.000000	0.102397	-0.451255	-0.285431	-0.240923	-0.257665
conductivity	-0.356913	-0.204633	0.102397	1.000000	0.201083	0.061895	0.061737	0.058369
BCOD	0.369034	-0.838554	-0.451255	0.201083	1.000000	0.325552	0.559099	0.561209
nitrate	0.351529	-0.394366	-0.285431	0.061895	0.325552	1.000000	0.112182	0.122953
faecal coliform	0.215403	-0.370665	-0.240923	0.061737	0.559099	0.112182	1.000000	0.991509
total coliform	0.227615	-0.381674	-0.257665	0.058369	0.561209	0.122953	0.991509	1.000000

In [33]:
```
df.describe()
```

Out[33]:

	temperature	dissolved oxygen	pH	conductivity	BCOD	nitrate	faecal coliform	total coliform
count	81.000000	81.000000	81.000000	81.000000	81.000000	81.000000	81.000000	8.100000e+01
mean	16.071605	8.080247	7.774259	590.665432	4.308025	1.126049	15146.684568	3.441897e+04
std	5.225646	2.074935	0.220285	571.581527	13.643424	1.233695	97158.676076	2.060708e+05
min	7.500000	0.000000	7.300000	77.000000	0.000000	0.100000	2.000000	2.000000e+00
25%	11.000000	7.950000	7.650000	214.000000	0.250000	0.350000	22.000000	1.795000e+02
50%	16.500000	8.600000	7.750000	275.500000	0.600000	0.700000	55.900000	4.100000e+02
75%	20.500000	9.250000	7.900000	1012.500000	1.200000	1.300000	300.000000	8.470000e+02
max	26.000000	10.250000	8.500000	2630.000000	84.000000	6.350000	863500.000000	1.789500e+06

In [25]:
```
from sklearn.metrics import mean_squared_error

mean_squared_error(df['pH'],yhat)
```

Out[25]: 0.06989782776773712

In [38]:
```
MSE=[]
order=[1,2,3,4]

for i in order:

    Input=[('sclae',StandardScaler()),('Polynomial',PolynomialFeatures(degree=i)),('mode',LinearRegression())]

    pipe=Pipeline(Input)
    pipe.fit(df[['temperature','dissolved oxygen','conductivity','BCOD','nitrate','faecal coliform','total coliform']],df['pH'])
    yhat=pipe.predict(df[['temperature','dissolved oxygen','conductivity','BCOD','nitrate','faecal coliform','total coliform']])
    MSE.append(mean_squared_error(df['pH'],yhat))

plt.plot(order,MSE)
plt.xlabel('Polynomial order')
plt.ylabel('Mean squared error')
plt.title('Order Vs MSE')
plt.show()
```