Analyzing Water Quality in India

Shashank Lokesh
PES University
Bangalore, India
shashanklokesh2000@gmail.com

Varun Kerenalli
PES University
Bangalore, India
varunk0404@gmail.com

Vishruth P Reddy
PES University
Bangalore, India
vishruthpreddy@gmail.com

Keywords - data analytics, water quality, regression, water quality index

I. Introduction

Water, a chemical compound, with the basic features of being colourless, odourless, transparent and tasteless is the most essential requirement for all living forms in the world. It is one of the sources of survival on planet Earth. It covers about 71 percent of our planet in various forms such as oceans, rivers, lakes, polar ice caps, groundwater etc. Although there is enormous amounts of water, only 3 percent of it is clean. 2.5 percent of the water is unavailable to us as it is hidden in polar ice caps, glaciers, atmosphere etc. Only 0.5 percent of the entire supply is available to us as fresh water.[1] Apart from quenching our thirst and using it for our daily needs, it is the transporter of oxygen and essential nutrients to various cells of living things, facilitates the regulation of body temperature, detoxifies the body and removes wastes. It is a home to more than a million species which include both plants and animals.

II. Background

Due to industrialization and urbanization, water has been polluted to unimaginable extents. Due to water pollution, the available 0.5 percent of freshwater is also being depleted of nutrients. This 0.5 percent is insufficient to quench the thirst of the huge population in the world, let alone using the water for other requirements. The areas with sufficient surplus don't know the importance of water and waste and pollute it as they will. An awareness has to be spread to show the desperate need for a clean supply of water.[2] If impure water is consumed, it leads to many short and long term complications and side effects which can affect our health and can even lead to death. Due to water pollution, many species in the marine world have been swept away without a trace and many are being poisoned to death everyday.

The condition of water in India can't be compared elsewhere. Not only are factories the main source of water pollution, but our festivals, surface runoffs with all sorts of insecticides and pesticides, improper sanitary cleansing and various other factors are the reasons for the enormous pollution in India.[3] Accounting several factors like pH, BOD, carbon dioxide content, dissolved chemical and metals and other parameters, we have calculated the water quality index and provided the condition of water quality in every state in India.

III. Overview

Our work includes Exploratory Data Analytics and Predictive Data Analytics of the water quality over the years. After doing the necessary preprocessing, we performed regression and constructed scatter plots and bar charts. Using the existing data, we even predicted the water quality for further years.

IV. Problem Statement

Understanding the various factors that are affecting the water quality in India and predicting the water quality in the coming years is our ultimate goal.

V. Existing Solution

The data contained all the pollutants and the variation of their quantity over a period of time. There was no data regarding the correlation between any of the factors causing the pollution.

Through our paper, we can see that the densely populated regions of our country have more water quality index. Water quality index is inversely related to the quality of water. We have provided a state wise water quality index for states showing a different trend over the years [4]. Using the existing data, a prediction of water quality index in the future years is shown and represented graphically. A state wise rank of water quality index is given in a sorted fashion for better visualization.

VI. Limitations

Although our paper provides valid correlation, there are many more factors to be considered in the real scenario to calculate water quality index. Factors such as human beings' contribution to pollution, surface runoffs, improper sewage treatment etc will be really hard to gather and tabulate but they are also the main reasons for water pollution.

Due to the missing data in the initial years and the generalization of the values in various regions and inaccessible parts of the country, it is not the perfect data to use for prediction of water quality index and hence, make conclusions about the overall water quality.

Data from many states and union territories isn't available so we can't generalize the correlations we get for the entire country.

VII. Proposed Solution

Preprocessing

Three forms of missing data and mismatch presented themselves and were solved using the following techniques:

1. Swapped Columns:-
   The 'locations' and 'state' columns were interchanged in certain locations and had to be swapped.This was done utilizing the fact that India only has a specific set of states and could be swapped if they matched the list of states.
   PH and conductivity were also swapped in certain indices and were reverted as PH ranges only between 0-14.

2. Absent States:-
   Records didn't have state values assigned in the 'State' column , but some of them had it present in the 'locations' column.This was stripped and then checked to see if it matched with the list of states and inserted.

3. Filling NAN Values:-
   For missing numeric values , replacing with a combination of the value in the upper and lower record seemed inappropriate, as did replacing it with the average of the column , which would corrupt the actual central tendency of the dataset. Therefore the only logical technique left remaining was replacing the numeric nan values with a list that retains the mean and standard deviation of the values in that column. We feel that this greatly helped our analysis as it retains the gaussian for the attributes that would have otherwise been present.
   Even this posed a problem as the random values then drifted into the negatives. This was handled by using a function which trimmed off the gaussian in the negative for the numeric data which conceptually could not be negative, i.e. all of them.

Model Building:-

After organising the data appropriately , in order to derive conclusions from it, we devised a singular metric that could be used to determine the quality of the water in the area.

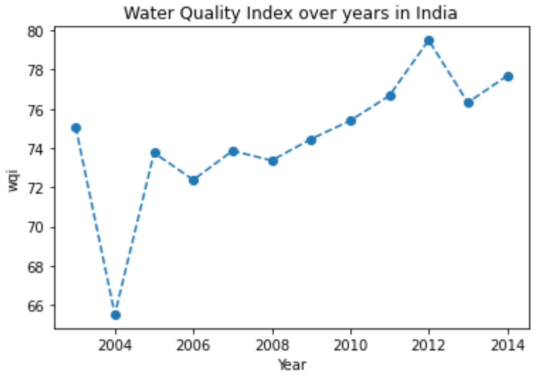
To successfully predict this metric WQI (Water Quality Index) for the future, we used a few techniques which included but were not limited to decision trees, random forest, Naive Bayes etc but all these algorithms were overfitting the data.

Regression was the only algorithm which gave us correct results. Regression analysis is a form of predictive modelling technique which investigates the relationship between dependent and independent variables. This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. This technique fits perfectly due to the fact that we are trying to predict the future of a particular state, which in this case is the Water quality index, using previous values. Other algorithms that we tried had underfitting issues and could not be trained efficiently.

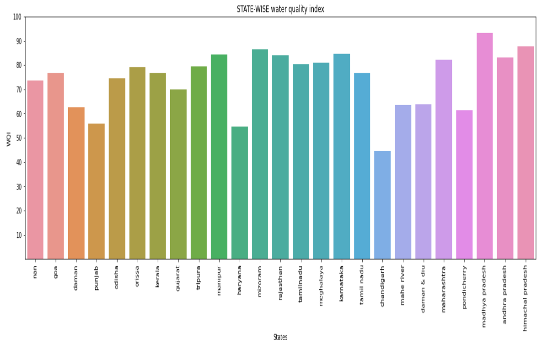
VIII. Experimental Results

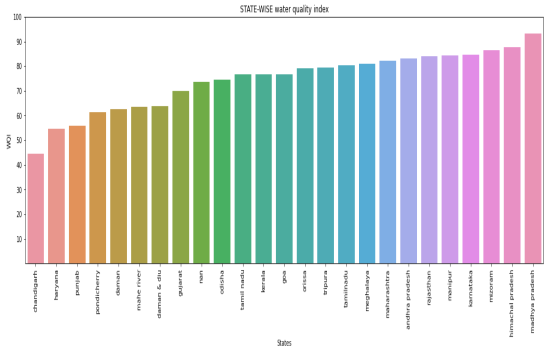
After testing our dataset, we came up with several notable insights related to water quality index:

1. From the scatter plot, we see that the overall water quality kept deteriorating over the years until around 2012. The graph gradually dips down till 2013 due to the reduction in the coliform content in the water bodies but again increases mostly due to the increase in coliform content.

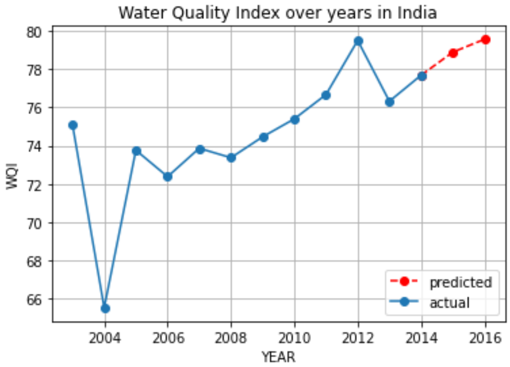


2. In the water quality index versus state bar graph, we can witness that Madhya Pradesh has the highest water quality index whereas, Punja has the least. When we sort the bar graphs based on height, we can clearly see that population is directly proportional to the water quality index. This shows that water quality has deteriorated and very less measures have been taken to better the quality in the densely populated states. Places like Chandigarh, Haryana, Punjab etc have sufficient good quality groundwater surplus whereas, places like Karnataka, Madhya Pradesh, Himachal Pradesh etc have very poor water quality supply.





3. Using the data, we also managed to predict the water quality for future years. The obtained results showed that the water quality index grew by a huge margin. This means that the water quality has decreased drastically on the whole in India. We evaluated this by checking the articles and journals which support our prediction.



IX. Conclusions

Although from our prediction we can see that the overall water quality has decreased, we cannot come to conclusions that the water quality is decreasing everywhere in the country. Due to the lack of inaccessibility to certain regions in the country, the data to be gathered is hard. Even the data gathered in particular states may not be the same for the entire state as the data could be limited to a particular region which could have been generalized. Nevertheless, there are several states that have worked towards improving the water quality while there are others that really don't seem to show any concern related to it. States like Goa and Himachal Pradesh have improved their water quality to a great extent. From the data of the water quality table, we can witness that these states have excellent surplus of clean groundwater. Certain states like Karnataka are gradually showing improvement and the government body is also taking measures to improve it. Gujarat had shown tremendous improvement in the quality of water but saw a sudden deterioration in water quality.

X. References

[1] Richa Gupta1, Prateek Srivastava1, Ambrina Sardar Khan and Ajay Kanaujia. "Ground Water Pollution in India- A Review"

[2] A. Maria,"The Costs of Water Pollution in India"

[3] Subodh Kumar, Hari Mohan Meena and Kavita Verma ,"Water Pollution in India: Its Impact on the Human Health: Causes and Remedies"

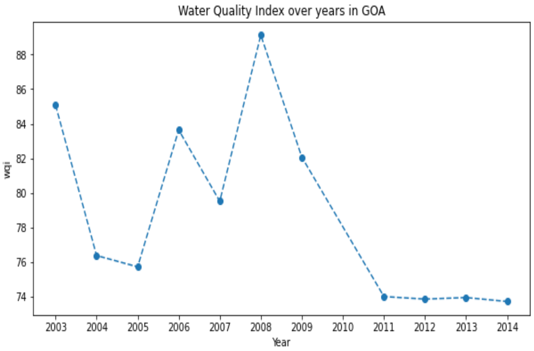
[4] Dwivedi, Anil. (2017). RESEARCHES IN WATER POLLUTION: A REVIEW. 10.13140/RG.2.2.12094.08002.

XI. Interesting Findings

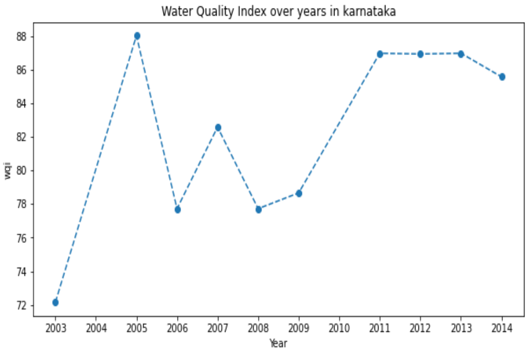
1. Goa has seen an amazing improvement in water quality when the new appointed chief minister banned all the nude beaches and restricted the access of water bodies by tourists. This not only improved the water quality but also brought order in the state.
2. Karnataka has seen a gradual improvement in water quality from 2013 as the government body went on a mass lake cleaning campaign which cleared out most of the lakes of its waste although there is a lot more to do.
3. Although Gujarat saw a good improvement of water quality, there was a sudden rise in the water quality index in the year 2014. The main reason was that the state of Gujarat overexploited the ground water which resulted in high salinity and high fluoride content that caused lots of damage to water quality and hence became impure.

XII. Appendix

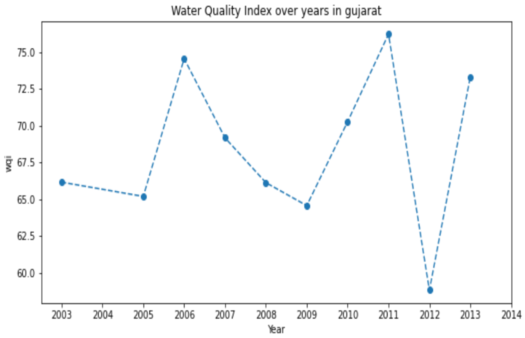
1. Water Quality Index in Goa versus Time.
Water quality has increased drastically and more importantly has been stable from 2011-14.



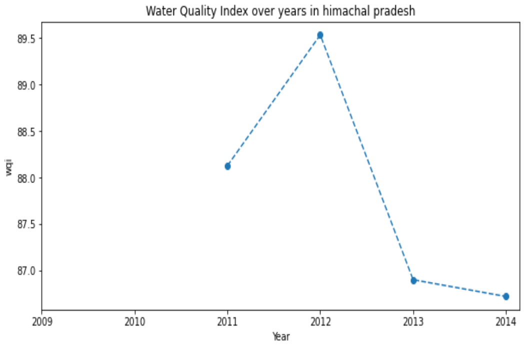
2. Water Quality Index in Karnataka versus Time
Water quality index , was erratic from 2003-2008, mainly because of missing data, but as the data became more abundant , we can see that the quality levelled off and has even gotten better in 2014



3. Water Quality Index in Gujarat versus Time
   Water quality has been getting worse from 2009 and the blip in the graph is mainly because of minimal data.



4. Water Quality Index in Himachal Pradesh versus Time
   Water quality was in a critical state in 2012 but has since gotten better and stayed that way up until 2014


Water Quality Index over years in himachal pradesh
wqi
Year