Regions at Risk of Water Contamination

Qiyue Zou[1], Dajun Luo[2], Zhe Wang[3], Brandon Wong[4]

[1]) qz298@cornell.edu [2]) dl938@cornell.edu
[3]) zw273@cornell.edu [4]) byw4@cornell.edu

1 Abstract

Water is an undeniable resource. It is essential for human wellbeing and is critical to various industry as well. However, as our economy grows and our industry develops, the by-products of expansion begin to accumulate. Access to safe water, even in a developed country such as the United States, is not guaranteed. Our model predicts how water quality will change in the future. This can be used to inform the at-risk population, reducing the hazardous effects of water contamination. Our techniques can be applied to informing the at-risk population and ultimately produce a predictive model with the aim to mitigate water contamination.

2 Team Toolset

Python: scikit-learn, pandas, Jupyter Notebook, Colaboratory Tableau, Excel, Box

3 Data understanding

The following core datasets were pre-cleaned and provided by Citadel and

- CorrelationOne:
- chemicals.csv
- droughts.csv
- earnings.csv
- educational_attainment.csv
- industry_occupation.csv
- water_usage.csv

Note that not all of them were relevent to our exploration. Since we want to study the relationship between changes in industry and water quality, we decided to use the chemicals.csv and industry_occupation.csv datasets.

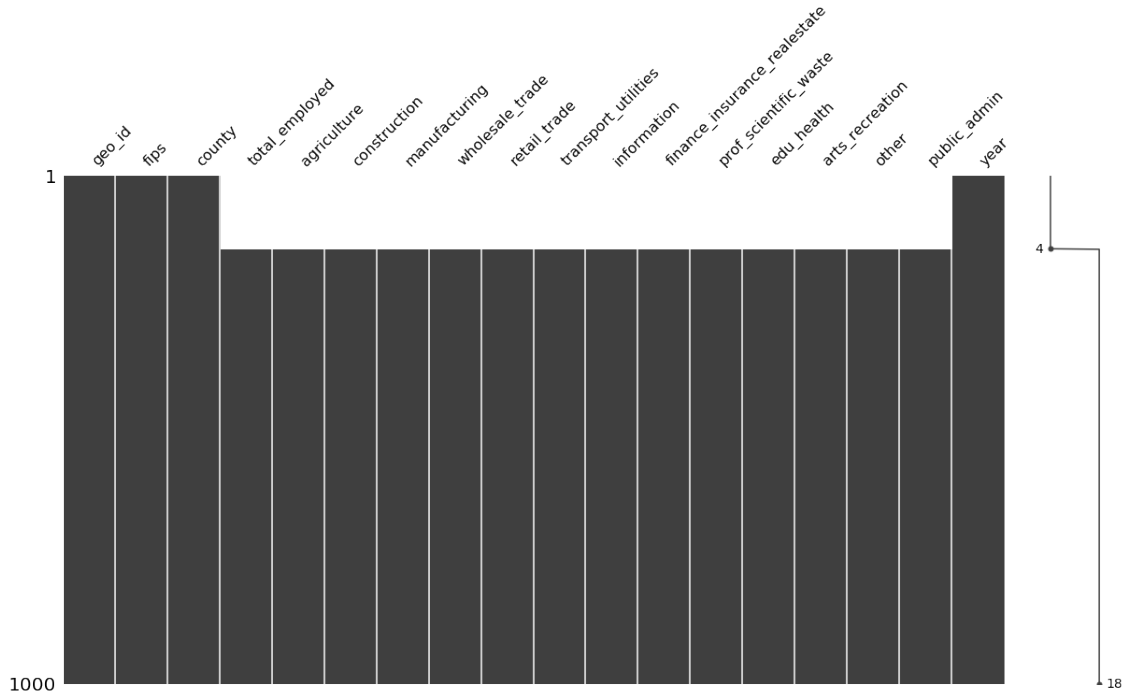

Figure 1: The spread of missing data in the industry_occupation.csv dataset. We see that the missing data almost happens at the same time.

4 Data Wrangling and Cleaning Process

We firstly check whether the two data sets (chemicals.csv and industry_occupation.csv) have missing data or not. We find that chemical.csv is complete, while industry_occupation.csv is missing a substantial amount of data. So, we took a closer look at how the missing data in industry_occupation.csv was spread. Fig. 1. shows that a county either has complete data, or it has no data about its industry and occupation. To avoid unnecessary complexity, we assume that the missing data points are MCAR (missing completely and random). As a result, we just drop the rows with missing values.

5 Exploratory Data Analysis

Next we want to see how employment is distributed across the major industry sectors within each county. Note that there are over a thousand counties and it is not feasible to analyze the distribution on a county by county basis, we aggregate the employment distribution across all counties. Fig. 2 shows that employment distribution is stable from 2010-2016. Henceforth, we will assume that future employment distribution will be similarly stable.

Now, we will take a closer look at the chemicals.csv file. We want to see if the unit of measurement used is consistent over the years. After confirming

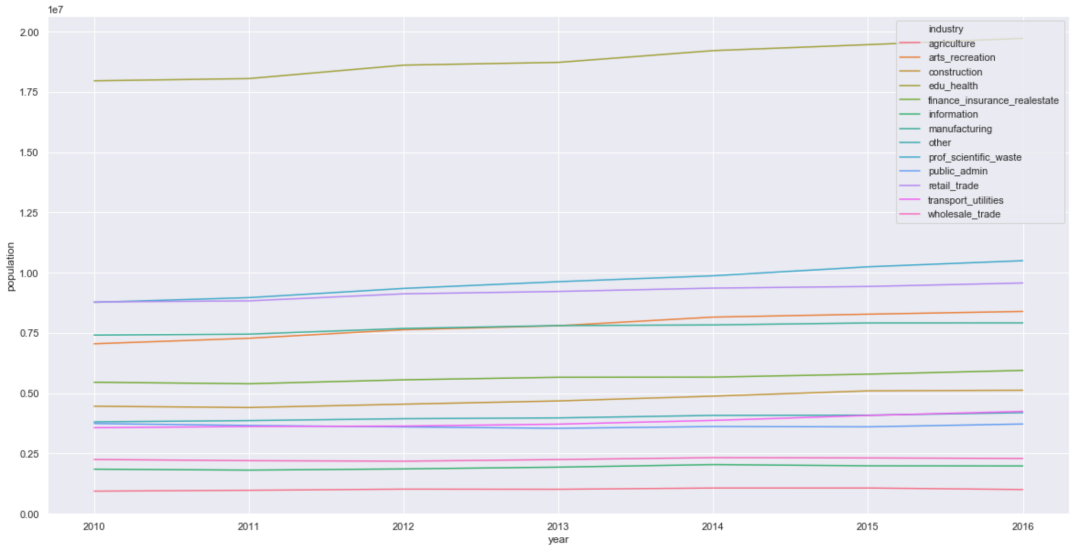

Figure 2: Number of population in each industry type as a function of year. We see that the employment distribution is stable from 2010-2016

that the unit of measurement is indeed consistent. We find that different counties have different standards for each of the six chemicals, as is shown in Fig. 3.

Because of this inconsistency, in future analysis, we will use the level instead of raw values. Later we show in Fig. 4 that there is correlation between different types of industires and the level of contamination for each of the six chemical species. For example, we see that "Nitrates" are correlated with most industries, but especially with agriculture. The left figure shows the non-timelagged correlation, while the right figure shows the correlation when timelag = 2. As the timelagged heatmap exhibits a greater correlation than the non-timelagged heatmap, we will use the timelag=2 heatmap.

6 Modeling

After creating the timelag=2 dataset to be used in our predictive model, we perform a train test split and reclassify the containment_level feature. Originally, the levels were:

- Less than
- Non-detect
- Greater than

Now we have

- Safe (less than and non-detect)
- Danger (greater than)

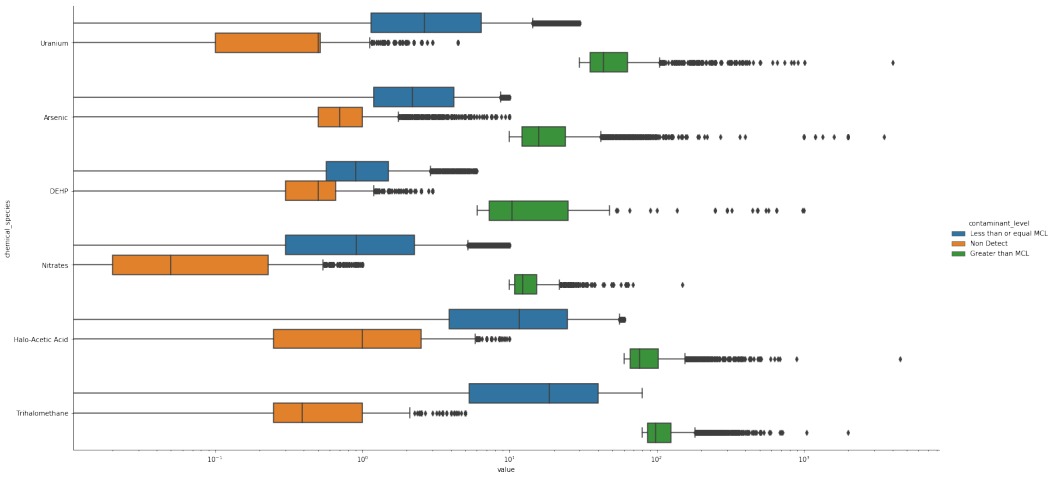

Figure 3: Contamination level for each of the six chemical species. The overlapping regions indicate that different counties have different standards.

Because the data is imbalanced, we will use AUC to measure the accuracy. We chose LightGBM as our prediction model because it is generally faster and has higher accuracy. LightGBM is a gradient boosting framework that uses tree-based learning algorithm. LightGBM grows tree vertically while other algorithm grows trees horizontally meaning that LightGBM grows tree leaf-wise while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm.

We fit the model to our training data and use the validation to test it. The accuracy score is around 92%. This shows that the model fits the data quite accurately. We find that population and location are very influential in the prediction of future polluted areas, as is shown in Fig. 5 . We also find that other interesting features, such as the number employed by the agriculture and transportation sectors, also play important roles. These findings are in agreement with common intuition.

7 Future Predictions

We know that our model is accurate, so we can now use that model to predict locations where water quality might fall in the future. We created a copy of the original dataset (2010-2016) and used that to train the predictive model (timelag=2). This model predicts the safe/danger classification of each chemical species in 2017 across all counties. (Reminder: we assumed that future industry structure will not change.) Fig. 6 shows the predicted water pollution around the counties for different types of chemical species.

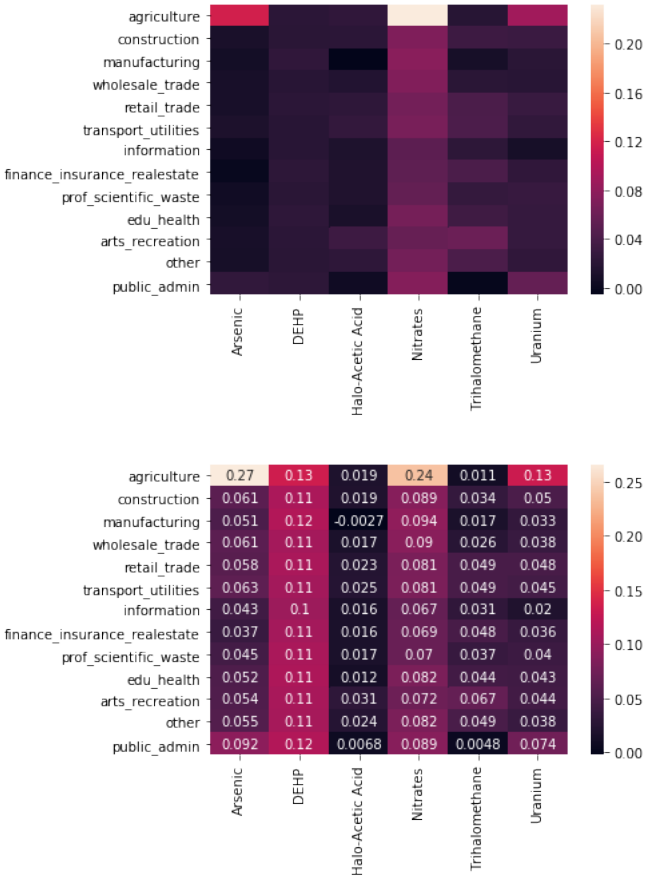

Figure 4: Correlation between different types of industires and the level of contamination. The upper figure shows the non-timelagged correlation, while the bottom figure shows the correlation when timelag = 2.

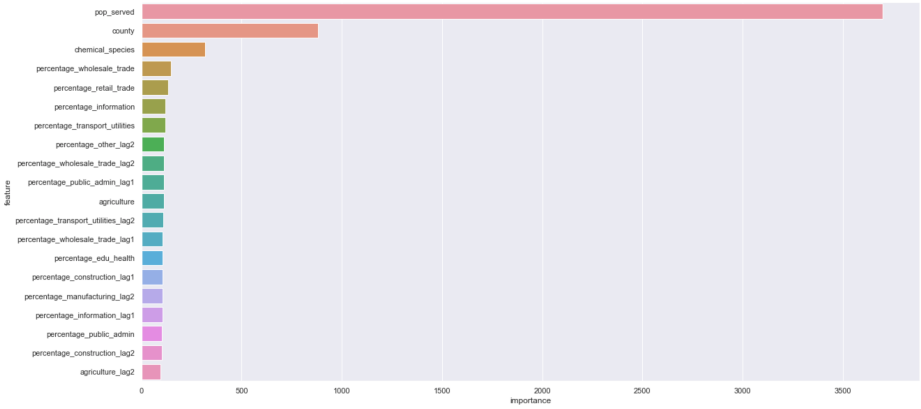

Figure 5: Feature importance in predicting water pollution.

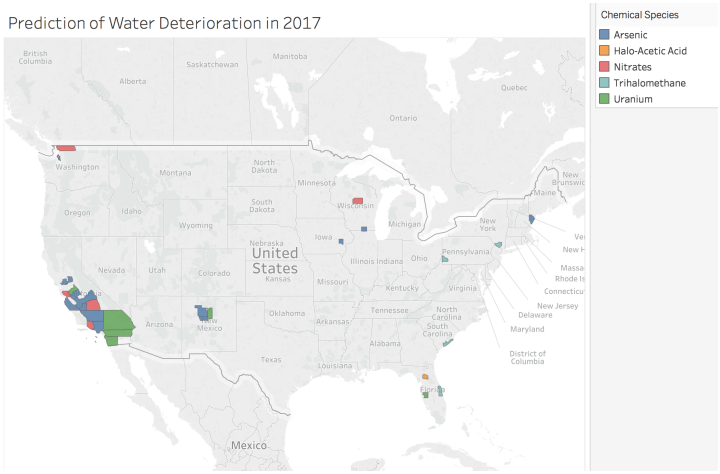

Figure 6: Predicted water pollution around the counties for different types of chemical species.

8 Summary

We explored the relationship between industry and water quality over time and built a predictive model with high accuracy (AUC=0.92). This model can be used to alert the at-risk population of potential deterioration in water quality due to industry in a specific county. We hope our disovery can have a positive impact on public policy as preventing water contamination will increase the overall wellbeing of humankind.

9 Reference

James Max Kanter, Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. IEEE DSAA 2015.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.