TECH BRIEF

Data Science Summer Fellows Program

Summer 2020

Modernizing Water and Wastewater Treatment through Data Science Education & Research

Goodyear Artificial Wetlands Project

Ivan Ko, Affiliation
Blake Loosley, Affiliation
Lauren Varnado, Affiliation

SUMMARY

In the city of Goodyear, AZ, artificial wetlands were created as an experimental water treatment facility with the goal of removing selenium from water. Treated water could then be dispersed in the nearby Gila River. The data contains measures of selenium and other chemical concentration. Best subsets were used to suggest predictors for lowering selenium levels using vegetation, media and other chemicals available in the data. The results have shown that certain vegetation and media types performed better than others in the removal of selenium.

INTRODUCTION

The city of Goodyear, AZ has an osmosis facility where they have created artificial wetlands in 7 bins to produce potable water. The primary objective is to determine whether these wetlands can remove regulated constituents, such as selenium, in order to discharge this water into the nearby Gila River. It is important to remove these regulated constituents before discharging it into the river due to their toxicity and the facility's duty to maintain wildlife safety.

There are seven bins with vegetation and media imitating artificial wetlands. Water can pass through one of four trains with a unique mix of bins, as seen in Figure 1. The goal is to determine which bin or bin train best removes selenium and possible predictors that help determine this.

FACILITY SYSTEM DESCRIPTION

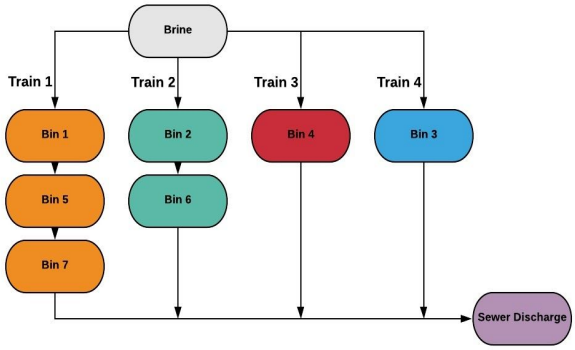

Figure 1: Diagram of Treatment Trains

The artificial wetland bins have treatment trains of other bins that water runs through to filter out selenium more effectively, where the goal is to have a discharge limit of selenium of 0.002 milligrams per liter(mg/L). As seen in Figure 1, bins 1,5 and 7 are in train 1. Bins 1 and 5 are vertical flow while bin 7 is

designed as a surface flow to increase the treatment residence prior to sewer discharge. Train 2 consists of bins 4 and 6, which are vertical flow. Trains 3 and 4 are single bin trains, which are also vertical flow.

DATA DESCRIPTION

The data provided consists of 29 variables and 3,504 observations. Data samples consist of two major categories: field data, collected at the facility, and lab data, collected monthly. Because of the numerous missing values (NAs), the dataset was reduced to 416 observations and sometimes down to only 17 entries in order to complete various calculations and/or modelling. Data is missing from mid 2014 to early 2015 due to a change in management. Due to the limited number of observations, the predictions cannot confidently be supported in some cases. Two additional sets of data were provided midway through the analysis: one with entries from Sept 2017 to Jul 2018 and another with data pertaining to soil analysis. Unfortunately, these could not be integrated with the original datasets due to time constraints. After initial exploratory data analysis, irrelevant variables were eliminated and the focus was placed on 13 variables. The Goodyear stakeholder is interested in the interaction of Temperature, pH, DO (dissolved oxygen), Nitrate and Arsenic. COD (chemical oxygen demand) is also included as a key variable.

EXPLORATORY DATA ANALYSIS

During exploratory analysis, boxplots were used to show the range of the log of Selenium concentrations amongst all bins and colored by Treatment Train as shown in Figure 1. One 3d plot (Figure 3) is included in this report.

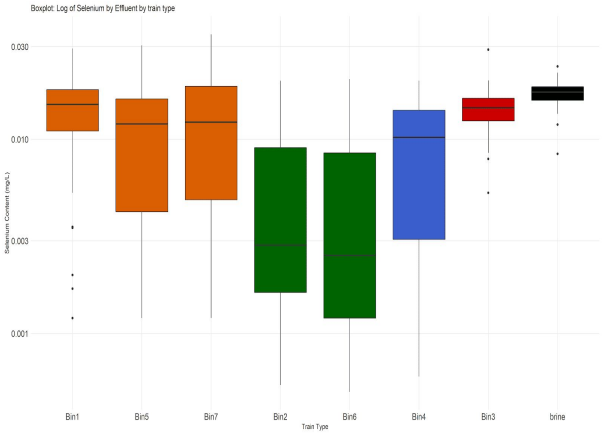

Figure 2: Selenium Treatment Train Boxplot

It is important to note that Bin 3 and Bin 1 have a small range with most of the data occurring well above the Selenium threshold. However, there are a couple outliers that produce more successful Selenium concentrations. Additionally, Bin 2 and Bin 6 have the most consistently low Selenium concentration values compared to the other bins. In other words, Bins 2 and 6 appear to be the only bins that are skewed towards higher values whereas the other bins are skewed toward the lower values. At face value, it appears that Bins 2 and 6 seem to be the best for removing Selenium since they have the lowest medians.

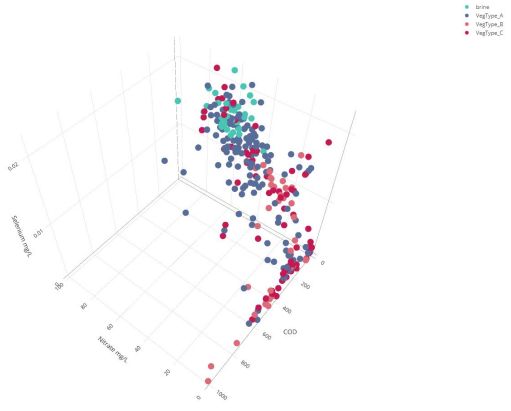

Figure 3: 3D Plot of Selenium level (z-axis) on Temp (x-axis) vs COD (y-axis)

Several 3D plots were used to explore the relationships between Temperature, Nitrate,

COD, DO and Selenium. The plots show that Temperature by itself is not a dominating factor related to Selenium level, however, when high Temperature is coupled with low Nitrate or DO, Selenium tends to be low. This trend is consistent among all Vegetation Types, and all Media Types except for Soil. In terms of Selenium removal, Vegetation Type A and B are more effective than Type C. It could be due to the system design as only Type C includes both Bin 6 and Bin 7, which have less Selenium to remove as they are not the first treatment bin. The same trend applies to Media Type Soil.

Estimated Marginal Means tests were run on available categorical variables to examine the effect of different covariates. One thought when identifying covariates was to find the least significant interaction terms and use those as covariates, but it was important to identify which variables the stakeholders can actually control. Most of the variables that would make good covariates, eg. Nitrate and DO, were mostly responsive to the system and couldn't be controlled so instead used Temperature as a covariate since it was an external factor collected in the dataset.

Primary findings from these tests showed that when controlling for temperature, there are no significant differences between the mean selenium content of the different Media Types and there was one significant difference between Vegetation Types, specifically the mean for Vegetation Type A is higher than the mean for Vegetation Type B. When controlling for Temperature, Bin 2, Bin 4 and Bin 6 all have significantly lower means than Brine, and Bins 2 and 6 have significantly lower means than Bin 3.

STATISTICAL ANALYSIS RESULTS

Best subset regression was used on the data to get the best variable model for predicting which variables suggest a relationship with selenium.

The procedure tests every combination of possible predictors and gives suggestions on subsets based on certain criteria. In particular, adjusted R^2 and Mallow's Criterion.

One of the primary problems in creating models with this data was a lack of complete observations, as stated above. The interesting dynamic with this project was finding a comfortable balance between testing all the predictors or having enough observations to get more accurate models. Field data was excluded from the model selection process since in most cases it reduced the degrees of freedom to 18, resulting in over-fit models.

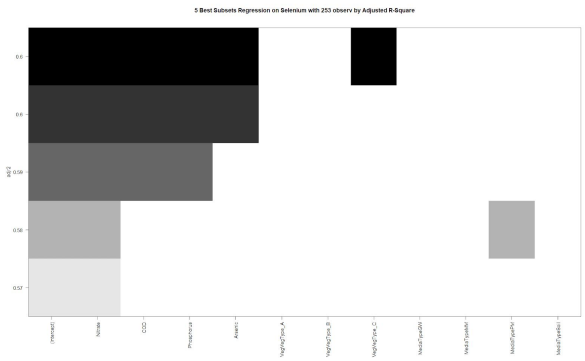

Figure 4: Best Subsets Regression Table of Models Ordered by Selection Criteria

The most successful model contained the following predictor variables: Nitrate Content, COD, Phosphorus Content, Arsenic Content, and a dummy variable showing whether or not the Vegetation Type from the Bin held Vegetation Type C. This subset of predictors was then mapped onto Selenium and had the highest possible Adjusted R^2, meaning this model accounted for the most variability of the data when compared to other models that best subsets tested.

One other interesting result from this test was the impact of the categorical variables for

Vegetation type and Media type. Two separate subsets procedures were run that only tested the impact of vegetation type and media type individually. Results showed that Vegetation Type C has the most significant impact on Selenium for Vegetation and Peat Moss has the most significant impact on Selenium for Media. When both Vegetation and Media are included in the procedure Vegetation Type C plays a bigger role in predicting Selenium than Peat Moss does as seen in Figure 4.

CONCLUSIONS

From the analysis, vegetation type appears to be influential in determining Selenium reduction. Even after controlling for Temperature, Vegetation has differing means across different Vegetation Types as shown by the Marginal Means Testing. One important question to raise is why Vegetation Type C appears to be a good predictor for Selenium in the model, when Vegetation Type B is the one Vegetation Type with the lowest Mean Selenium. Also, Vegetation Type C must be important because it's the only control variable that had significant results in best subsets regression so further analysis of the control variables is recommended.

Some recommendations we have to make the study more successful/conclusive are to get more consistent data collection, obtain better documentation of plant health and plant replacement, and also lining up field data with lab data. One of the biggest struggles in concluding the study was finalizing a model to predict Selenium. It was difficult to obtain conclusive results when the lab data and field data were taken on separate dates, creating a lot of incomplete observations. Also, when examining the effect of Vegetation Type on Selenium, the conclusion cannot be fully supported when Vegetation health isn't monitored in the data set. It skews the results making it difficult to tell if poor performance occurred as a result of the Vegetation itself or if the plants are dying out. Lastly, we had to exclude many lab data variables from model testing because of small amounts of complete observations. It forced models to be overfit and required the use of a smaller set of variables to test in the beginning.

REFERENCES

Waechter, Chris, Deborah Tosline, Katie Guerra, and Catherine Hoffman. Goodyear Pilot Wetlands: Developing Hydraulic Loading Rates, Hazardous Waste Disposal Requirements, and Optimum Operations for an Inland Concentrate Management Alternative, n.d. https://www.usbr.gov/research.

AUTHORS

Blake_Loosley1@baylor.edu

lvarnado@mymail.mines.edu

ivan_ko1@baylor.edu

ACKNOWLEDGEMENTS

Thank you to Dr. Amanda Hering, Dr. Doug Nychka, Maggie Bailey, and Kate Newhart, for their insight throughout the project. Special thanks to the stakeholders at Goodyear and everyone that helped in the collecting of the data.