# Searching for Predictors of Phosphorous Levels in the Chicago Water System

CS584 - Final Project

Grant Nikseresht, Travis Boltz

November 27, 2017

## 1 Introduction

Fresh water is the most important resource on this planet. As the population increases, so does the need for clean, sustainable sources of drinking water. The Great Lakes is the primary source of fresh drinking water for millions of Americans and Canadians, containing one fifth of the world's freshwater. The biggest threat to this precious resource are harmful algel blooms (HABs). These algae blooms can cover large areas of the lakes, destroy ecosystems, and release toxins into the water that can make it unsafe for consumption. The main ingredient required for these massive blooms is a large concentration of phosphorous in the lakes. Phosphorous spikes from agricultural fertilizer runoff has been identified as one of the major sources of phosphorous in lakes and rivers.

There are many groups that have formed to spread awareness and educate people about the challenges facing the Great Lakes region. One of these groups is the Cleveland Water Alliance, whose mission is to establish the infrastructure that's needed to understand and prevent occurances of events in the Great Lakes like HABs [1]. One of the issues they're facing is that phosphorous is hard to measure accurately and expensive to scrub from a waterway. One of the questions the Cleveland Water Alliance posed was is there is a sensor that could be used to measure the levels of phosphorous in rivers, lakes, and streams. In response to this question, we hypothesized that it could be possilbe to predict the level of phosphorous using other proxy predictors that are easier to measure and quantify such as turbidity, dissolved oxygen, acidity, or nitrogen.

## 2 Related Work

The U.S. Geological Survey in conjuction with Clean Water services used autosampler data to predict levels of phosphorous using a linear regression using factors such as turbidity, nitrogen, acidity, dissolved oxygen, E.coli, and Chlorine [2]. They were able to predict phosphorous using turbidity (cloudiness of the water sample), specific conductance, and flowrate with an adjusted $R^2$ of 0.808. Another paper from Mississippi State University, used a multiple regession model that used turbidity(NTU) and RFU(relative fluorescence) to predict phosphorous [3]. This paper was able to get an adjusted $R^2$ of 0.89. These datasets were of natural freshwater sources opposed to our dataset which comprises of municipal wasterwater (artificial source). Also we had 13 different variables while the other datasets only had 2 or 3 variables at most to predict. The dataset has been used previously in the context of understanding the conditions when sewage is dumped into the Chicago river. Our hope was to use it instead in the context of understanding factors associated with high phosphorous levels.

## 3 Approach

### 3.1 Dataset

We used datasets provided by the Metropolitan Water Reclamation District (MWRD) of Chicago that document the chemical makeup of the water that flows into their water treatment plants [4]. Influent data provides some of the most detailed information available on the water quality in Chicago's water systems. The datasets contained daily

measurements of water quality from 2011 to 2016 separated by plant. The datasets include point measurements of phosphorous in addition to information on water volume, acidity, nitrogen levels, and a variety of other chemical measurements.

### 3.2 Preprocessing

The datasets were spread across many locations, across many years, and contained a variety of documented measurement errors and factors that were only recorded in trace amounts in most samples. Chemicals that were rarely detected were removed. Factors such as nitrogen, oxygenation, and suspended solids that had more robust measurements were used for analysis. Daily measurements across all locations from 2011 to 2016 for the selected factors were consolidated into a single dataset with a location code added for each of the six plants to control for differences in plant volume. Our final dataset contained 13,152 observations of 10 variables including date.

Across the 13,152 observations, 61 observations contained missing data for factors other than `BOD5` and were removed. `BOD5` had 1,193 missing observations due to sensor reading errors noted in the original datasets. Due to its importance, we imputed the missing `BOD5` observations by training a linear model involving all predictors using the 11,943 recorded `BOD5` measurements and predicted the missing `BOD5` values using the model.

- Flow: the influent rate in millions of gallons per day
- pH: pH level, a measure of acidity
- BOD5: measure of biological oxygen demand (mg/l)
- SS: suspended solids (solids that float) (mg/l)
- TS: Total solids (mg/l)
- TKN: Total Kjeldahl Nitrogen (total amount of organic nitrogen and ammonia) (mg/l)
- P-TOT: Total phosphorous (mg/l)
- NH3N: Free ammonia or toxic ammonia (mg/l)
- Date: Month-Day-Year
- Location: A MWRD identifier; 1 = Calumet, 2 = Egan, 3 = Hanover Park, 4 = Kirie, 5 = Lemont, 6 = O'Brien

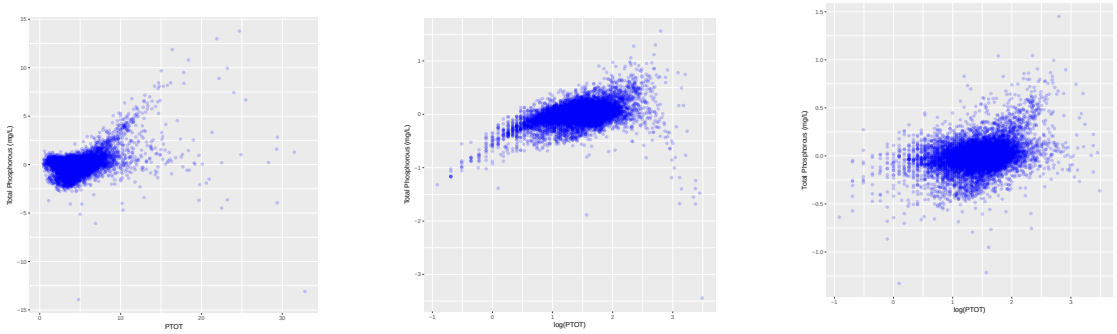
Observations from 2011-2015 were treated as the training set, while 2016 was kept as the holdout set. This split allowed for a balance in seasonality and served as a way to measure the dependability of measuring future phosphorous levels given past data.

### 3.3 Baseline and Metric

The mean value of phosphorous is our naive baseline for regression models. Nonetheless, our goal is to generate a model that explains a majority of the variance in phosphorous, so our initial metric for linear models will be $R^2$. $R^2$ will measure model performance and will improve as results get further from this naive baseline (where $R^2 = 0$). MSE on the test set will be utilized to compare across all models, but measures of variable importance like $R^2$ are vital. We'll also utilize variable importance and feature selection output from models like ridge regression, principal components regression, and gradient boosting to determine the factors that affect phosphorous.

### 3.4 Regression Methods

Several regression methods stood out as possibilities for predicting phosphorous. First, a simple linear regression model was used to gauge the importance of predictors. Log transformations of some predictors and the response variable were noted as necessary in prior work, so we compared the performance of the log transformed model to the original model. Further refinement of the linear model was done using ridge regression. Principal component regression (PCR) was a good comparision to the linear model because each of the principal components is a linear combination of all of the original features. An added benefit of PCR is that it is a concise and intuitive means of



(a) No log transformations

(b) Log transform of response

(c) Log transform of predictors and response

Figure 1: Heterscedasticity can be observed in the following residual plots where the errors are non-constant in the magnitude of phosphorous. Log transforming both phosphorous and several predictors removes the heteroscedasticity as can be seen in the residual plot in (c).

representing the impact of the various predictors. Finally, gradient boosting is performed using `xgboost`, which we use to search for more advanced ways to reduce MSE. `xgboost` has feature importance metrics, as well, that we can use to further infer which factors are the strongest predictors of phosphorous levels.

# 4 Results

## 4.1 Baseline

Our baseline prediction was the mean value of total phosophorous across all observations, which was 4.701. The baseline mean squared error on the 2016 data was 9.954. The baseline prediction for the log model is the mean of the natural logarithm of phosphorous, which is 1.447. Using this baseline the mean squared error on the test set with a log transform of phosphorous is 0.202.

## 4.2 Linear Regression and Ridge Regularization

Simple linear regression models helped significantly in trying to understand the relationship between phosphorous and our predictors. An initial model containing all predictors achieved a test MSE of 0.588 and an $R^2$ of 0.823. Model selection using ridge regression reduced the test MSE slightly to 0.584 with little impact on $R^2$.

Amongst single predictor models, three predictors had $R^2$ values above 0.4 - suspended solids, `BOD5`, and total Kjeldahl nitrogen - with $R^2$ values of 0.4, 0.53 and 0.59, respectively. This holds up to some of the conventional wisdom that nitrogen accompanies phosphorous in fertilizer polluted water.

Incidence of heteroscedasticity is apparent in the residual plots shown in Figure 1. Many of the properties of the water accelerate at exponential rates and there are some high leverage outliers included that need mitigation by a log transform. This confirms findings in the previous work that log transforms of phosphorous and various predictors such as suspended solids, BOD5, and TKN are necessary for a robust model.

Predicting the log of phosphorous using the log of TKN, log of BOD5, NH3N, log of SS, location and season resulted in the best performance among all linear models with an $R^2$ of 0.887 with all predictors being statistically significant. The test MSE for this model was 0.0257. Ridge regression using `cv.glmnet` performed slightly better achieving a test MSE of 0.0253. Ridge regression is helpful because it's likely that there is a degree of multi-collinearity in the predictors. Multiple measurements of solids and nitrogen are included, so ridge regression excludes some of these collinear predictors. Ridge regression for instance shrunk NH3N and TS, while maintaining TKN and SS. TKN and NH3N are somewhat collinear, as are TS and SS, so the ridge regression makes improvements by removing a collinear predictor.

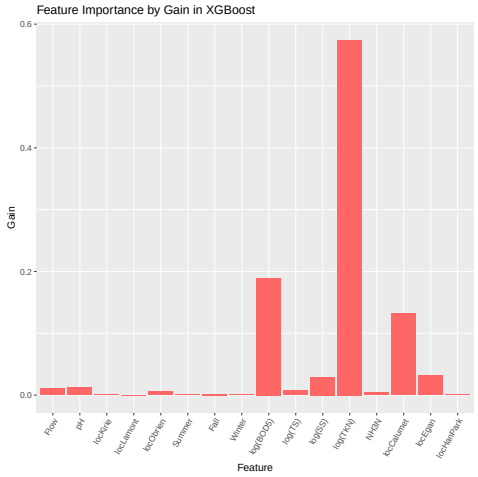



Figure 2: A plot of the gain metric as output by the gradient boosting model. A high gain implies that a large improvement in error was achieved by splitting on the corresponding feature. Total Kjeldahl nitrogen and BOD5 had the two highest gain metrics in the best boosting model.

### 4.3 Principal Components Regression

A log linear relationship between the response and predictors was suspected as the true relationship. Also, we suspected that most if not all the predictors had some relationship with total phosphorous. TKN, NH3N are correlated with each other and the same for TS and SS. Flow and pH were elminated due to a weak linear relationship with phosphorus. Dropping NH3N and TS from the PCR and using all of the other components resulted in a test MSE of 0.0251 that explained 88% of variance of total phosphorous. Principal component regression was useful in that it vectorized the attributes to make it easier to visualize patterns and relationships between the attributes.

### 4.4 Gradient Boosting

Although a log linear relationship is suspected between the predictors and the response variable, trying a method such as gradient boosting could further improve the test MSE. The `xgboost` implementation of gradient boosting was utilized, and cross-validation led to optimal parameters of a maximum tree depth of 5 and a learning rate of 0.01. The final test MSE found by gradient boosting was 0.0193, which outperforms all the prior linear models.

The primary metric to measure feature importance in `xgboost` is gain, which quantifies the improvement in the root mean squared error gained by splitting on a particular feature. Total Kjeldahl nitrogen and BOD5 were selected as having the greatest importance by gain in the boosting model. Interestingly, boosting identified whether or not an observation was made at the Calumet plant as the third most important factor. This could be because Calumet had a relatively large number of outliers that lets the boosting model capture a disproportionate amount of error in those outliers by splitting on Calumet.

## 5 Conclusion

### 5.1 Summary

In terms of performance using MSE, gradient boosting achieved the lowest mean squared error on the log transformed test set, significant beating the baseline MSE and the other linear models. However, our gaol was to find the elements in water that predict phosphorous levels independent of location. Total Kjeldahl Nitrogen had the highest gain in the gradient boosting model and achieved the highest $R^2$ value in its single feature regression model. TKN also had one of the highest coefficients in both the ridge regression model and the log linear regression model. Using TKN over the other nitrogen factor, NH3N, seems to improve performance. BOD5 was also considered the

| Model | Test MSE |
|---|---|
| Baseline (No Log) | 9.954 |
| Linear Regression (No Log) | 0.588 |
| Ridge Regression (No Log) | 0.584 |
| Baseline (Log) | 0.202 |
| Linear Regression (Log) | 0.0257 |
| Ridge Regression (Log) | 0.0254 |
| Principal Components Regression (Log) | 0.0251 |
| Gradient Boosting (Log) | 0.0193 |

Figure 3: Comparison of results using a common test MSE metric.

second most important factor in gradient boosting and the single predictor linear regression models. Suspended solids appeared to be more useful in predicting phosphorous than total solids, and had the third highest single predictor $R^2$ and was kept by ridge regression. Factors such as seasonality and location, while necessary for our dataset, did not appear to be the most important factors. Flow rate and pH also did not appear to be signficant throughout.

These findings support the conclusions in prior work that factors like nitrogen and some measure of turbidity (suspended solids) are associated with phosphorous levels. Furthermore, by including a large span of spatial and temporal data, we were able to control for factors such as seasonality and location and discover that predicting phosphorous seems to be somewhat independent of those factors. This has a variety of implications, for instance that the phosphorous is not coming from a localized source, but has instead spread throughout the Chicago water system, and that factors such as heavy rain or snow melt don't seem to have major impacts on the phosphorous levels.

### 5.2 Future Work

There were major unexplained outliers in our data that we were originally concerned about. While there effects on the analysis seem to be mitigated by the log transformations, it still remains to be seen what the cause of those spikes in factors such as BOD5 and phosphorous was. It's still possible that omitted variables such as pollution events could cause these seemingly random spikes but are unaccounted for. Furthermore, models that include categorical variables for the presence of the more trace elements could add more flexibility to the amount of ways to find phosphorous in a water system. The goal of predicting phosphorous is to find factors that may be easier to measure than phosphorous that could indicate the presence of the element in water. Our analysis suggests that turbidity (suspended solids) could be the most cost effective proxy for phosphorous, so more robust research into the relationship between turbidity and phosphorous in water systems could shine more light on the issue.

## References


[1] *Cleveland Water Alliance*, 2017. [Online]. Available: ⟨http://clevelandwateralliance.org/⟩

[2] C. Anderson and S. Rounds, "Use of continuous monitors and autosamplers to predict unmeasured water-quality constituents in tributaries of the tualatin river, oregon," *Scientific Investigation Report, 2010, 5008*, 2010. [Online]. Available: ⟨https://pubs.usgs.gov/sir/2010/5008/pdf/sir20105008.pdf⟩

[3] C. Andrews, R. Kroger, and L. Miranda, "Predicting nitrogen and phosphorus concentrations using chlorophyll-a," *Mississippi Water Resources Conference*, 2012. [Online]. Available: ⟨http://www.wrri.msstate.edu/pdf/andrews12.pdf⟩

[4] "Water reclamation plant data." [Online]. Available: ⟨http://www.mwrd.org/irj/portal/anonymous?NavigationTarget=navurl://9f766d4f820e9482d016681c86031b76⟩

# 6 Appendix

## 6.1 Code Repository

The code used for the project can be found at `https://github.com/grantnikseresht/predicting-phosphorous`. The `R` scripts used to generate our results is found in the `scripts` folder. Our final dataset is in `data/ALL_MWRD.csv`. Plotting output is saved in `data/plots`.

## 6.2 Map of Water Treatment Plants

A map of Chicago's water treatment plants is shown in full on the next page. The plants referenced in our dataset are spread across the Chicago area.

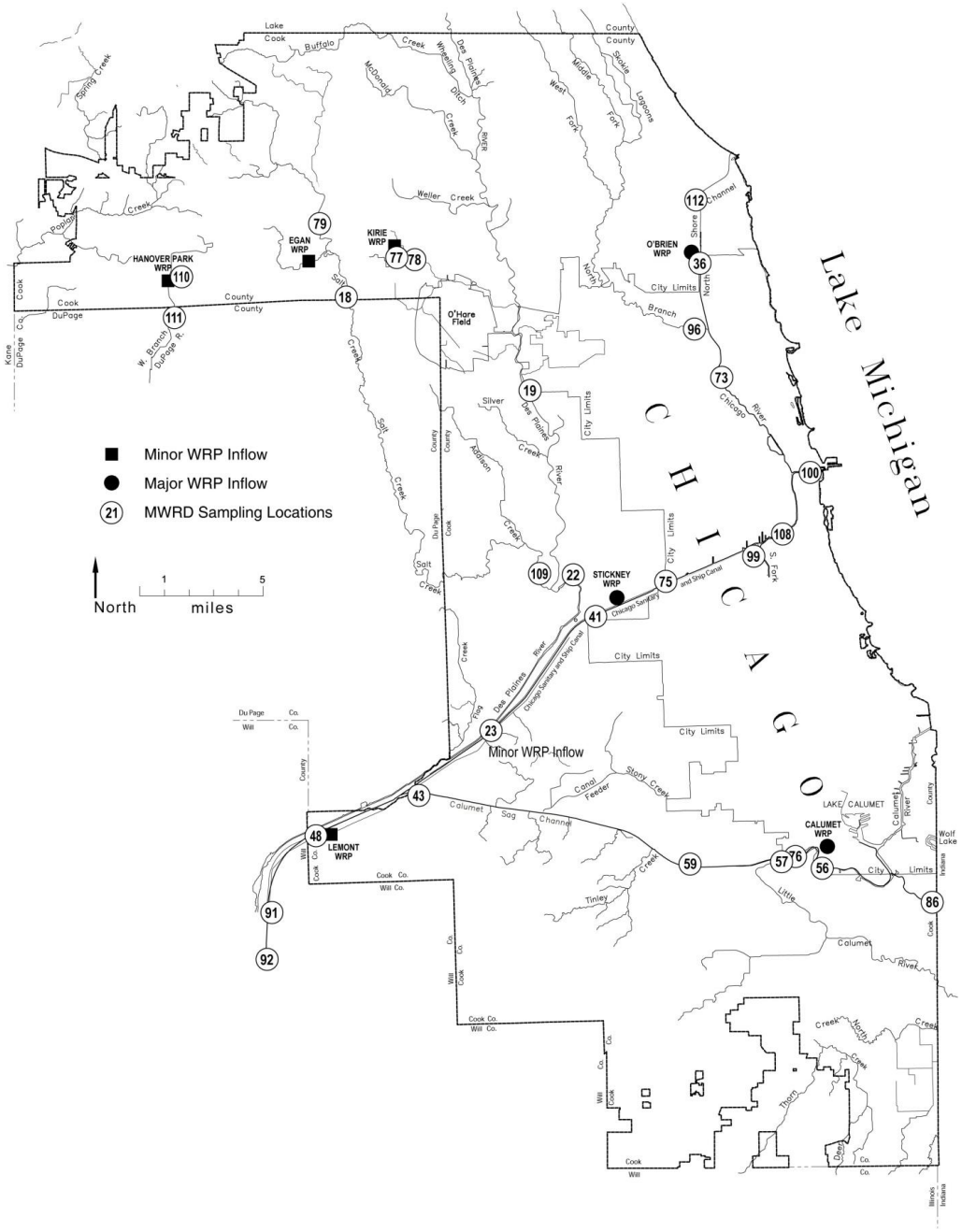



Figure 4: A map of the water treatment plants included in our datasets [4].