Predicting 5-Day Biological Oxygen Demand (BOD5) in Greater Chicago's Waste Water

(Math 571 Project Final Report)

Submitted to:
Prof. Adam McElhinney
Department of Data Science
Illinois Institute of Technology

Report Prepared By:

Travis Boltz A20389533

Yitao Ma A20380311

Yuqing Zhao A20392575

Yue Ning A20282614

May 3, 2018

Table of Content

1 Executive Summary

This report summarizes the analysis and modeling results associated with the Biological Oxygen Demand (BOD) prediction in wastewater study. The purpose of this report is to document both the implemented sampling design and all corresponding data modeling and inference techniques used during the subsequent statistical analyses.

The development of the sampling protocol, including both the data source and data cleaning strategy are discussed in Section 2. The basic statistics that summarize the contamination data associated with the analyzed compounds are given in Section 3. A total of 10 variables were analyzed for eight locations for this study. Five compounds concentration measurements include TKN, NH3.N, P.TOT, SS., and FLOW. Three variables are added in this study include Rainfall, population density, season and location. BOD5 is the target. The statistics summary and several plots were performed in this study.

Section 4 presents the analysis of the demographic, based on the 8 waste water treatment plants where data were acquired. ANOVA-test p-value approach were used to compare data for BOD5 (target variable) to see if the means are equal. Regression models and Algorithm background are discussed in Section 5. Section 6 introduces model evaluations. Finally, in section 7, discussion and conclusion are given.

2 Introduction

Untreated Wastewater is a significant problem for the environment when it is discharged into the body of water. The impact of dumping highly concentrated organic matter into the waterways creates an unfriendly environment in which the rate of oxygen consumption is greater than the supply of oxygen present in the water. It is highly related to the microorganism in water ecosystem. Microorganism growth is limited by the presence of nitrogen and phosphorus, which is in high concentration in wastewater. The organisms go into overdrive feeding and in the process consume oxygen at a high rates, much faster than it is supplied. This causes stress on the aquatic environment and left unchecked creates an anoxic environment which essentially turns a lake into a swamp. One of the common and most important metric used to measure the impact of wastewater on the environment is the 5-Day Biological Oxygen Demand(BOD5)[1]

2.1 Research Problem

Can we use Metropolitan Water Reclamation District (MWRD) public wastewater data to predict (within a reasonable error) the 5-Day Biological Oxygen Demand (BOD5) on Day 0?

2.2 What is the 5-Day Biological Oxygen (BOD5)?

5-Day Biological Oxygen Demand amount of dissolved oxygen needed/required by organisms to consume organic matter present in a sample of water that is isolated at a stable temperature for 5 days.

BOD5=BOD mg/L=[DO0-DO5)- seed concentration] x dilution factor*
DO0=Dissolved Oxygen Day 0
DO5=Dissolved Oxygen Day 5

Essentially, a sample of water and measure the dissolved oxygen (DO) is taken on Day 0 and put in an incubator for 5 days, the DO is measured again on day 5. The DO difference between the Day 0 and Day 5 is what we call the Biological Oxygen Demand. BOD5 is not the most accurate measure, the longer the more accurate you can be about the biological demand. Some countries require 7 days or BOD7. The 5-day waiting period is a balance between accuracy and time to action[2].

2.3 Issues with BOD5

A major issue with BOD5 is that it takes 5 days in order to know what the BOD was on Day 0. There is no way to speed up the procedure. This is inconvenient when you need to know the concentration on Day 0 to make a decision. Also, it is possible for the sample to get contaminated during the 5 days period (listwise deletion of observations when BOD5=N/A was 10% of our raw data)[3].

2.4 Why predict BOD5?

To explain why we would want to predict BOD5 we need a little background on Chicago and Combined Sewage Overflows (CSOs). Chicago has Combined Sewage Overflows

[1] See Reference 1
[2] See Reference 1
[3] See Reference 1

(CSO's) that can empty untreated sewage into the natural environment*. This is an old system that ensures that the sewage lines don't back up. Once a sewage reaches a certain level in the pipes, it overflows into the CSO and empties in the environment. This is a holdover from early wastewater treatment infrastructure. Modern wastewater infrastructure no longer utilize this system (no one likes sewage dumping into rivers and streams). During heavy rains Chicago needs to close the canal gates to keep the sewage from CSO's from contaminating the lake. Sometimes in the worse case, they need to re-open them to relieve the pressure from the sudden influx of rain water.

Predicting BOD5 on Day 0 could allow city officials to know what the BOD5 level is on Day 0 instead of on Day 5. They could make a more informed decision to close or open the canal gates into the lake. Also, we can could impute BOD5 on days where the sample was contaminated. A prediction could also replace the test, but this is not really feasible in the current regulatory environment.

3 Data Source and Cleaning

3.1 Predictor Variable Selection

We decided to use the following parameters to predict BOD5 based on Travis Boltz's (i.e. one of our team members) domain knowledge from his undergraduate education as an environmental engineer. The SS, NH3.N, TKN tests all can be done on Day 0, BOD5 is the only test that takes 5 days[4][5].[6]

Suspended Solids (SS)	Solids that don't settle readily in water.
Ammoniacal Nitrogen (NH3.N)	Measure for the amount of ammonia, a toxic pollutant found in waste water and leachate from landfills.
Total Phosphorus(P.TOT)	Sum of all phosphorus compounds in Water.
Total Kjeldalahl Nitrogen(TKN)	concentration of organic nitrogen and ammonia in water.
Flow Rate(FLOW)	measure of flow over a given period of time.
Population Density (Pop.density)	number people per square mile.
Rainfall (Rainfall)	inches of rain/melt.
Season(Season)	Winter, Spring, Summer, Fall.

Table 2.1 Selected variables

[4] See Reference 2
[5] See Reference 3
[6] See Reference 4

4. Waste Water Data

All of the data is publicly available. The wastewater physical and chemical properties were obtained from Greater Chicago's Metropolitan Water Reclamation district (MWRD). We used years 2011-2017. MWRD manages 8 modern water reclamation plants throughout Cook County. The purpose behind the reclamation plants is to provide treatment for residential and industrial wastewater. Before the water is discharged in the natural environment, the chemical and biological properties of the water must meet strict standards of the Environmental Protection Agency's (EPA) National Pollutant Discharge Elimination System permit program. All water (including storm/rain water) is processed in one of the 8 plants.

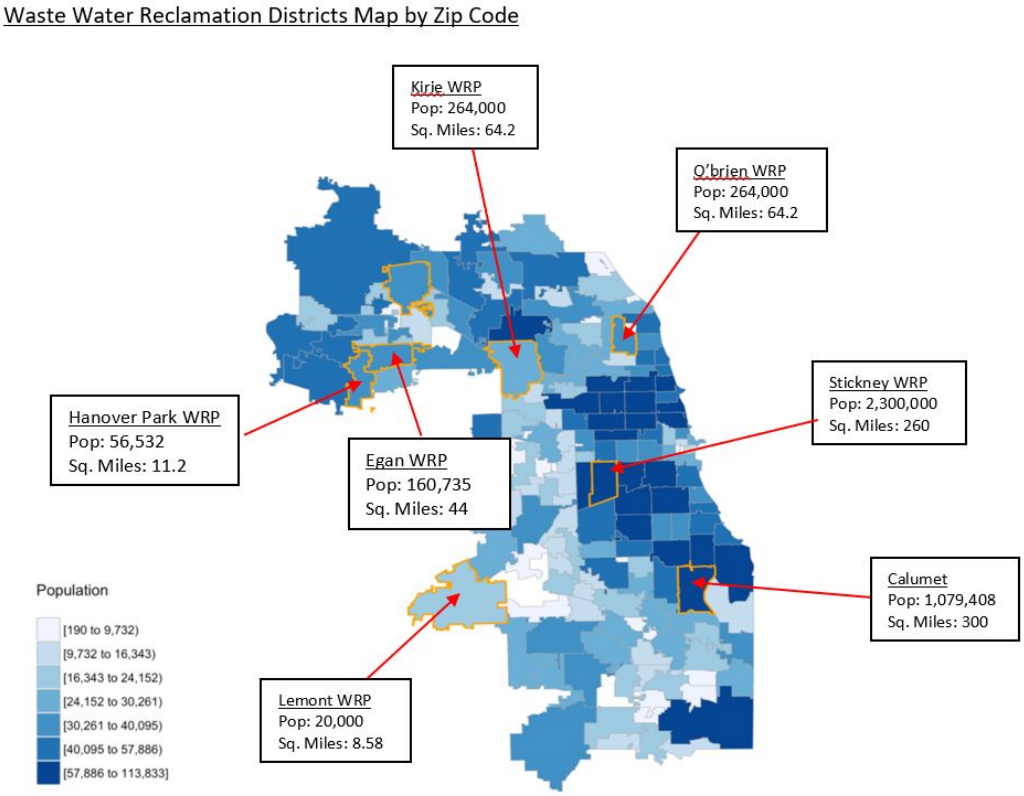


Figure 2.2 An overview statistic summary for numeric variable

We collected chemical properties of the waste water from the effluent data located on the MWRD Portal :

Calumet Outfall 2011-2020

Egan Outfall 2011-2020

Hanover Park Outfall 2011-2020

Kirie Outfall 2011-2020

Lemont Outfall 2011-2020

O'Brien Outfall 2011-2020

Stickney Outfall 2011-2020

We collected the flowrate of the waste water from the influent data located on the MWRD Portal (to obtain flow rate):
Calumet Raw 2011-2020
Egan Raw 2011-2020
Hanover Park Raw 2011-2020
Kirie Raw 2011-2020
Lemont Raw 2011-2020
O'Brien Raw 2011-2020

Rainwater Data

We obtained our rainwater data from the National Oceanic and Atmospheric Administration's (NOAA) Climate Data Online. We used O'hare International Airport's rain gauge for year's 2011-2017. We assumed that the rain was uniform across all Waste Reclamation Plant (WRP's).

Population Density

We derived the population density from the attached Factsheets for each of WRP's. They provided population served and the service coverage by square mile in each fact sheet. We divided the population by square mile coverage to obtain population density.

4.1 METADATA

From the data sources above we created a dataframe with the following METADATA:

Location(factor)	1=Calumet, 2=Egan, 3=Hanover,4=Kirie,5=Lemont,6=Obrein,7 =Southwest(Stickney), 8=Westside(Stickney)
Season (factor)	1=Winter, 2=Spring,3=Summer,4=Fall
BOD5 (mg/L)	continuous numeric variable
TKN(mg/L)	continuous numeric variable
NH3.N (mg/L)	continuous numeric variable
P.TOT (mg/L)	continuous numeric variable
SS(mg/L)	continuous numeric variable
Rainfall (in)	continuous numeric variable

FLOWRATE (MGD)	continuous numeric variable
Pop.density (Population/sq.miles)	continuous numeric variable

Table 2.3 Metadata for variables

5. Exploratory Data Analysis

In this study, 18379 observations of 10 variables were acquired from 8 wastewater treatment plants locations. Each sample was analyzed for the following compound: BOD_5(5-day Biochemical Oxygen Demand), TKN(Total Kjeldahl Nitrogen), NH3.N(ammonia), P.TOT(Total of Phosphorus), SS(Suspended Solids). Flow (MGD: million of gallons of water used per day) and Rainfall (in) data were collected for each sample.

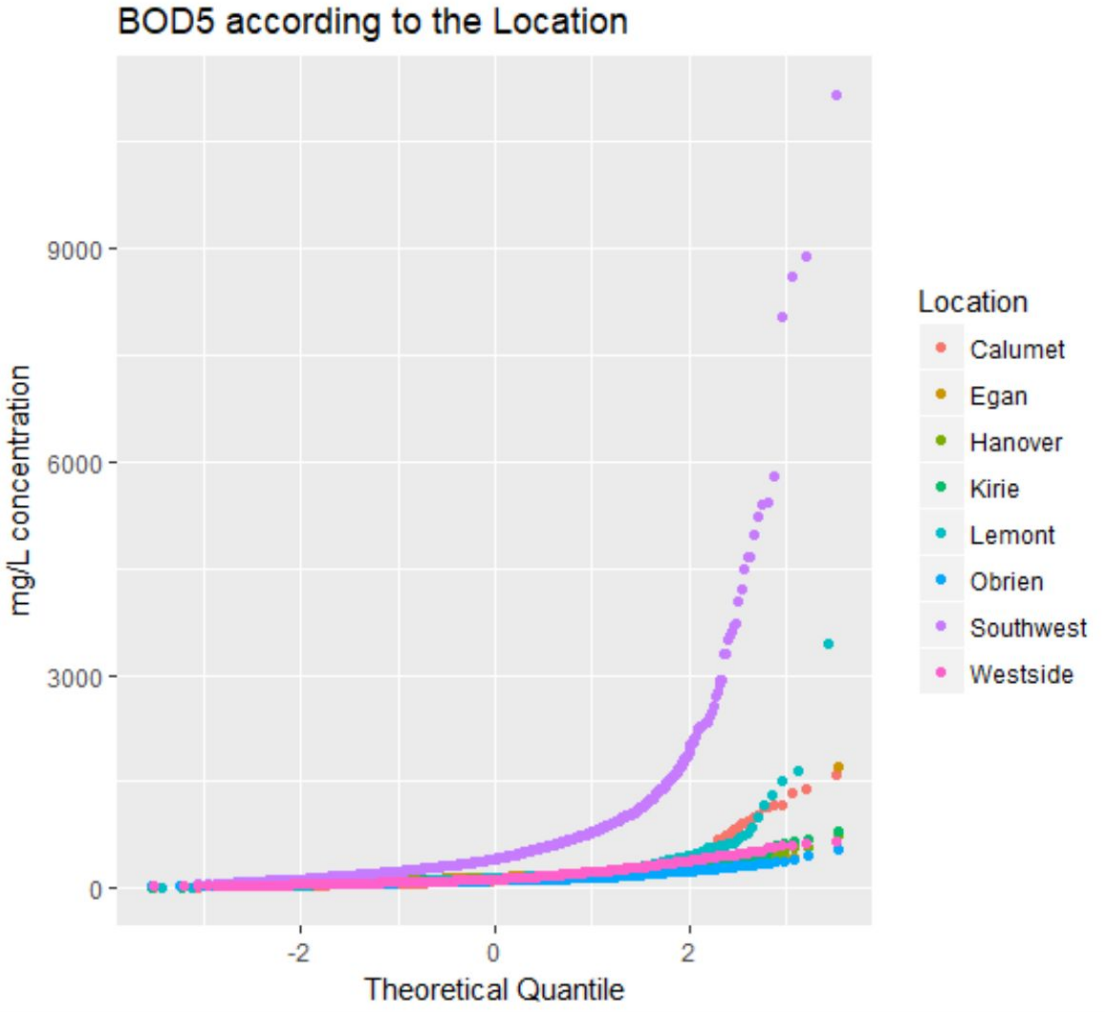


Table 5.1 QQplot of BOD5 in each location

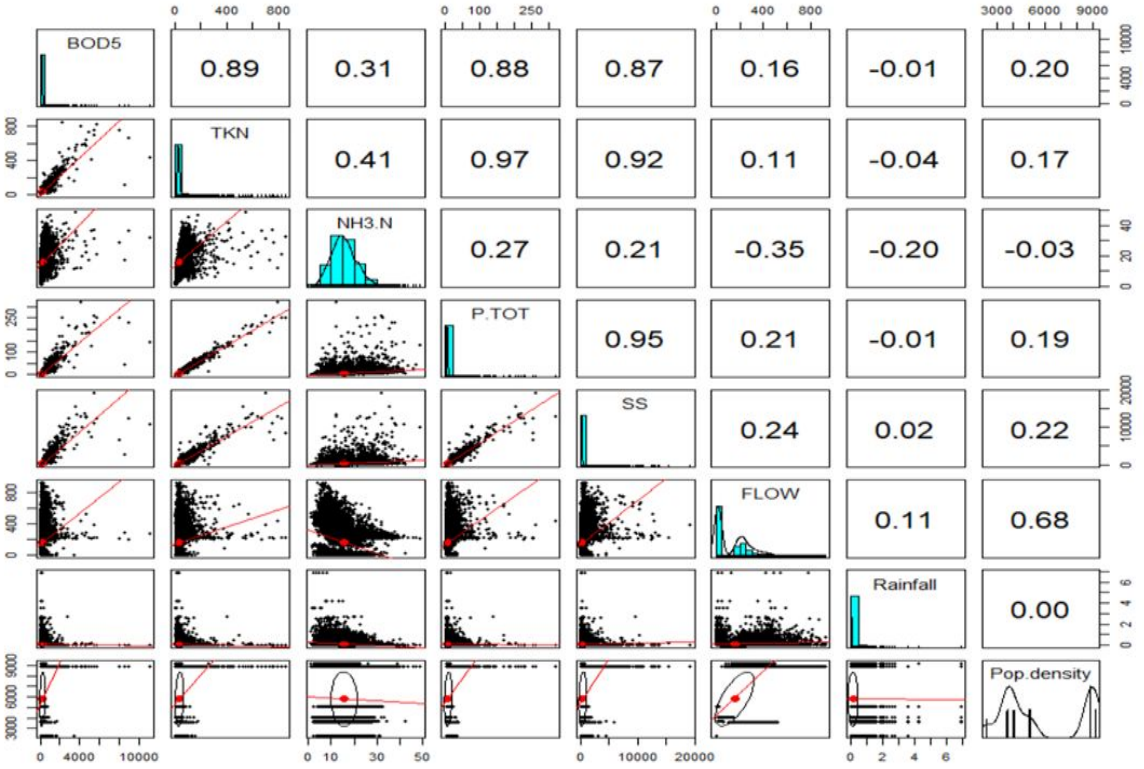


Table 5.2 Pairs Panels (pairs.panels::pysch): Lower Triangle are xy plots for each variable combination. Upper Triangle is the correlation of each of the numeric variables

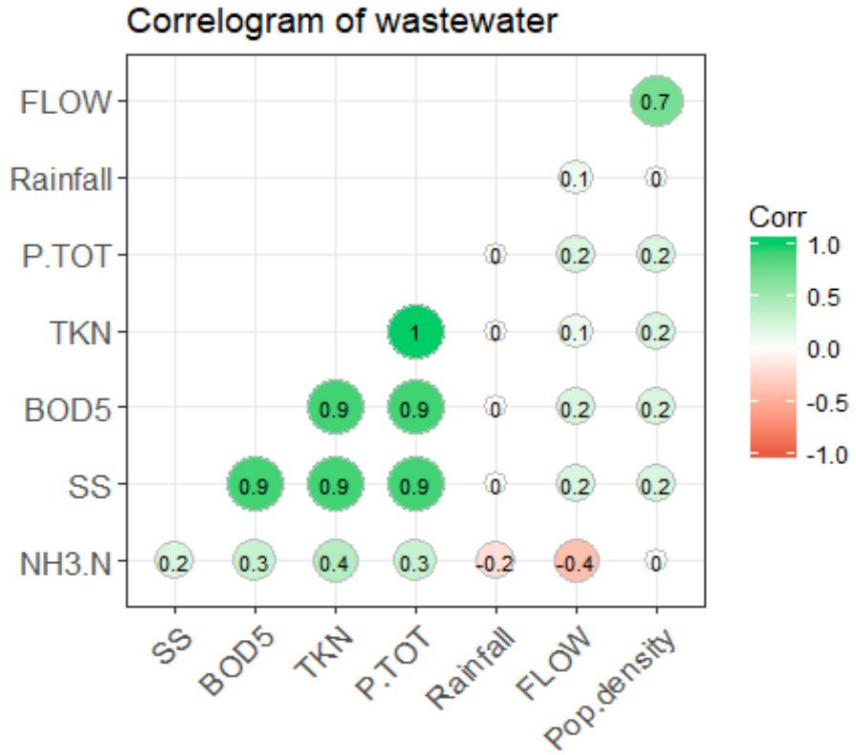


Table 5.3 Correlogram of wastewater

Strong correlations between BOD5 and TKN, P.TOT, SS. There are also strong correlations between TKN and P.TOT, SS and P.TOT, NH3.N and SS. All of these strong correlations are explainable. BOD5 would be correlated strong with TKN, P.TOT, and SS because these are the food stock for aerobic organisms in the water. So the more TKN and P.TOT, and SS the larger the biological oxygen demand. A large part of SS is organic matter, so it makes sense that it would be correlated TKN(Nitrogen) and P.TOT (Total Phosphorus), because when organic matter breaks down it mostly breaks down largely into these two components.

ANOVAThe final aspect of our exploratory data analysis is to determine if the Locations have equal means. We decided that a ANOVA test using the P-value as our metric. Let H_0 to be the mean of BOD5 of different locations are equal and H_a to be the mean of BOD5 of different locations are not all equal. Do ANOVA test on BOD5 based on different locations.

	Df	Sum sq	Mean sq	F value	P value
Location	7	347375724	49624961	938	<2e-16
α value	0.10	0.075	0.05	0.025	0.01

Table 4.1 anova result based on different locations

The result of ANOVA is that the P value=2e-16, which is less than 0.05. Therefore reject H_0 in favor of H_a which is that the means of BOD5 for each location are not equal. This lead us to use Location as factor in our models.

6. Data Analysis

Our target is to predict BOD5 based on other variables. Linear models are used to describe a relationship between one or more independent variables and a single target. Regression models take training dataset with build in learning algorithm to identify the model which fits the relationship between features and class labels. Here, multiple regression models including linear models, random forest and gradient boosting are applied to predict the BOD5 in wastewater.

All data is scaled first based on each variable to avoid variable with large magnitude to be dominated over other variables. The total dataset is randomly split data into (80%) training set and (20%) testing set. To avoid over-fitting problem, 10-fold cross-validation technique is applied on training dataset to train the model. Then the model is subsequently applied to the testing dataset. Finally, the results from model prediction and the original value from test data are compared using the Mean Squared Error. Adjusted R square is calculated for the test data to evaluate the models. Mean squared error is one way to measure the model

performance. It shows how much different are there between the real target value and the predicted one. Lower the value, lower the error. R-Squared is another method to measure the model performance since higher value explains more proportions of variance in dependent variables by independent variables. Equations are as followed:

$$R^2_{adj} = 1 - (\frac{(1 - R^2)(n - 2)}{n - q})$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

Square root transformation are applied on numeric variables to improve performances. All models used R built in functions. No code was modified for constructing models and cross-validation technique.

6.1 Regression models

Linear regression is applied first and used as baseline to be compared with other models.

To represent a linear regression model, a general form of linear function is used as combinations of intercept, coefficients variables and error term.

$$y = b_o + b_p x_{pi} + e_i$$

Where y is response variable with p independent variable x. b0 is intercept (i.e. constant term) and ei is error term. Model is developed by finding values for b to minimize the total error in the model. The basic way to build a model is to find all coefficients based on all data points using least square method. Five assumptions are made for all linear regression models.

Linear regression with variable selections are applied. To avoid unnecessary variables in the model, only significant variables are selected during the modeling process. Forward selection method starting with a null model with no variables builds a model by sequentially adding variables into the model. The selection of variables to be added depends on the value of the performance metric. The variable with best performance metric is added to the model and stayed in the model permanently. The process is repeated until the performance metric cannot be improved anymore. Backward selection method is similar. It starts with a full model with all variables builds a model by sequentially removing variables from the model.Stepwise selection method combines previous two and starts with forward selection. After each step of forward selection, backward selection is applied to see if removing any variables can improve the performance metric.

Linear regression with regularization is also used. Instead of selecting variables, a penalty can be added to the model to limit the number of variables for a model. A shrinkage parameter called lambda is used to the regularization amount. Lasso regression method builds with a full model and gives zero weight to unimportant variables. The linear regression function is subject to a regulation function where L1 norm of coefficients is

constrained. Ridge regression method builds with a full model and gives small weight to unimportant variables. The linear regression function is subject to a regulation function where L2 norm of coefficients is constrained. Elastic regression method builds with a full model and combines previous two methods by giving both small and zero weight to unimportant variables. The linear regression function is subject to a regulation function where both L1 norm and L2 norm of coefficients are constrained. Portions of L1 and L2 norm can be varied.

Two ensemble methods are used. Random forest regression is applied as a non-linear regression trail. It applied by constructing many decision tree models at training time on sub-samples and use the average to improve model performance. Gradient Boosting builds an additive model in a forward stagewise fashion. In each stage a regression tree is fit on the negative gradient of the given loss function.

Principal component analysis can construct the new variables from original variables in new dimension. It allows to reduce a data complexity to lower demission and in turn reduce the model complexity.

7. Results and Discussion

After applied models to the test data set, the model evaluation values are obtained as following:

	MSE	AdjR.2
Linear	0.2588944	0.5859705
Lasso	0.2591812	0.5855727
Elastic	0.2597431	0.5847936
Backward	0.2609373	0.5831395
Forward	0.2609373	0.5831395
StepWise	0.2609373	0.5831395
Principal Component	0.2701893	0.5704043
Ridge	0.2707909	0.5695811
Random Forest	0.6066091	0.2024749
Xtreme Gradient Boosting	0.6222013	0.1899130

We cannot see a obviously with feature selection and regularization. This is likely due to that all features we have now are important since they are carefully selected. Each feature has an important coefficient. The best model we have here is linear regression model. The nonlinear model does not show a good result since the relationship between target and variables tends to be linear in nature.

In order to further improve the model performances, the square root transformations are

applied to the numeric variables. Lower MSE and higher adjusted R square are obtained as shown below:

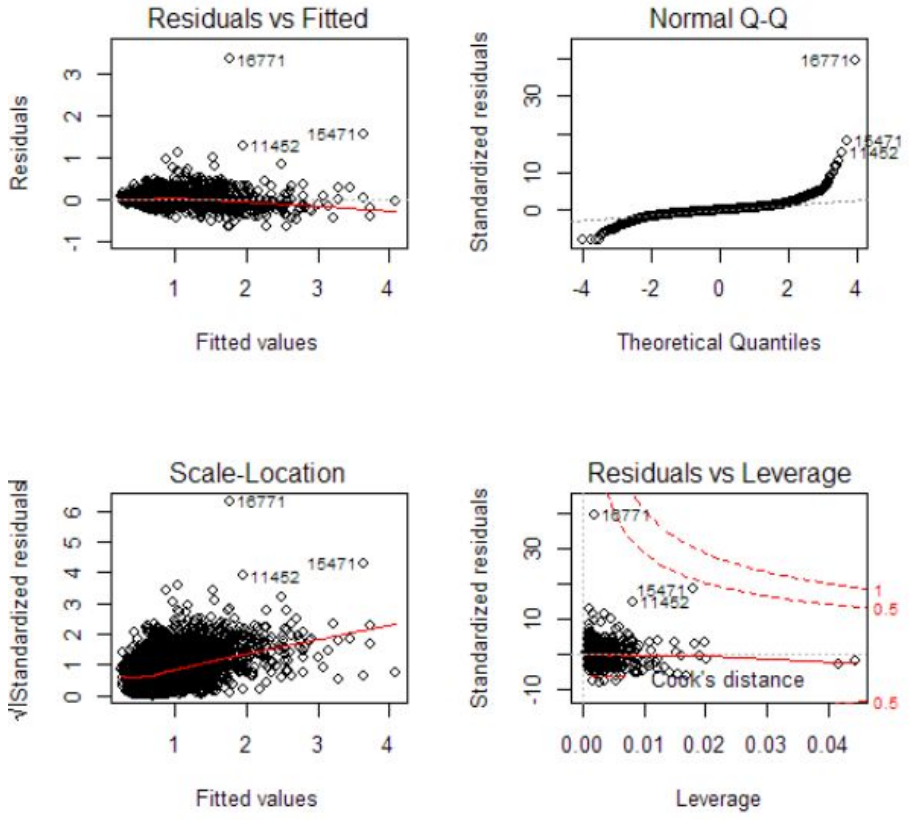
	MSE.sqrt	AdjR2.sqrt
Xtreme Gradient Boosting	0.008256975	0.8281607
Random Forest	0.008466140	0.8240174
Linear	0.008972446	0.8140310
Lasso	0.008981778	0.8138475
Elastic	0.008983543	0.8138128
Forward	0.009334066	0.8069356
Backward	0.009334066	0.8069356
Ridge	0.009499616	0.8036976
StepWise	0.009719772	0.7994017
Principal Component	0.019288777	0.6237751

This time best model changes from linear regression to gradient boosting. We are not sure about the exact reason while usually the gradient boosting gives a better result since it comes with an ensemble of several models. The ensemble of models, even though each model is not the best, it would lead to a better result.

When applying to practical case, the linear regression model will still be used since it is simple to interpret and less time consuming to run.

7.1 Analysis of linear regression model

We performed an analysis on linear regression model with transformed variables. The diagnostic plot of model is shown below to check if linear regression assumptions are met.



From the plot, we can see that residuals have a linear pattern and normal Q-Q plot is closely to normally distributed outliers. However, outliers may have a negative influence on the other two plots. The scale-location seems not spread out equally which indicates that a non-equal variance. The outlier influence can be confirmed by the residuals vs leverage plot. Three outliers are outstanding there at locations 16771, 15471, and 11452. We examined the outliers, and determine to keep them in the model. This is because if there was an error in the measurements it would have been notated in the comments portio of the excel file.

To summarize the model, the coefficients importance is shown below. We can see that all features are important and SS is the most important feature. It does make sense, since SS measure the solids that don't settle readily in water which would affect the oxygen content in water. The NA shown here for the population density is likely from the linear relationship between population density and the location here. In our case, the population density is changing with location in the exactly same pattern. That is for each location, only one population density is used. Thus, R treated these two variables as identical. Thus, one of population density and location can chosen for the modeling and the other one can be eliminate without affecting the modeling result here. For other applications, when location and population density is changing differently, both variables should be used.

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.189649   0.010506  18.052  < 2e-16 ***
TKN          0.249271   0.015159  16.444  < 2e-16 ***
NH3.N        0.114625   0.007939  14.438  < 2e-16 ***
P.TOT        0.096774   0.012153   7.963 1.80e-15 ***
SS           0.479451   0.007207  66.528  < 2e-16 ***
FLOW        -0.109003   0.006384 -17.073  < 2e-16 ***
Rainfall     0.007225   0.001471   4.912 9.13e-07 ***
Location2   -0.047765   0.004846  -9.857  < 2e-16 ***
Location3   -0.069033   0.005603 -12.322  < 2e-16 ***
Location4   -0.074277   0.004837 -15.356  < 2e-16 ***
Location5   -0.082605   0.006387 -12.934  < 2e-16 ***
Location6   -0.009677   0.003257  -2.972  0.00297 **
Location7    0.082678   0.004113  20.101  < 2e-16 ***
Location8    0.046990   0.003414  13.764  < 2e-16 ***
Season2     -0.009859   0.002008  -4.910 9.21e-07 ***
Season3     -0.043483   0.002030 -21.424  < 2e-16 ***
Season4     -0.019492   0.001995  -9.770  < 2e-16 ***
Pop.density        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8 Possible future works

Next step for data collection would be adding more years. Currently, only 8 years wastewater data is used. We can definitely collect more data to see if the model still works. We also want to determine how generalizable the models are by evaluating them in a different aquatic environments. For modeling, we can try other variable transformations and imputing missing values. Next, we can pair down the models to just 3-5 variables to determine if the MSE and R^2 are comparable.

We would like to have our results have some real meaning. Checking with the domain expertise to see if the prediction is useful is also important. We would also improve the user interface such that people can use their waste water data to get predicted BOD5 value.

9 References

[1] G.C. Delzer and S.W. McKenzie, "FIVE-DAY BIOCHEMICAL OXYGEN DEMAND,"*FIELD MANUAL,*2003.[Online]. Available: <https://water.usgs.gov/owq/FieldManual/Chapter7/NFMChap7_2_BOD.pdf>

[2] *Environmental Express,* "Total Kjeldahl Nitrogen (TKN) FAQs". [Online] Available:<http://www.envexp.com/pdf/FAQs/TKN%20FAQ.pdf>

[3] Fisher Scientific,"Standard Operating Procedure for: Total Suspended Solids",Revision 2 ,2007. [Online]. Available:<https://beta-static.fishersci.com/content/dam/fishersci/en_US/documents/programs/scientific/technical-documents/white-papers/apha-total-suspended-solids-procedure-white-paper.pdf>

[4] Missouri State University, and Ozarks Environmental and Water Resources Institute (OEWRI),"Total Phosphorus Analyses Using Genesys 10S UV-Vis(Total P Absorbance genesys R01.doc)", 2010. [Online] Available:<https://oewri.missouristate.edu/assets/OEWRI/Total_P_Absorbance_Genesys_R01.pdf>

10 Appendix:

A. **R Code:** https://github.com/Qingzz7/18S571project
B. **Summary Statistics**

	BOD5	TKN	NH3.N	P.TOT	SS	FLOW	Rainfall	Pop.density
median	140.0	27.00	15.200	4.600	154.0	132.0	0.0000	5047.50
mean	195.8	32.35	15.687	6.791	306.4	156.2	0.1070	5792.73
SE.mean	2.0	0.24	0.042	0.081	5.1	1.2	0.0023	18.99
CI.mean.0.95	3.9	0.46	0.083	0.158	9.9	2.4	0.0045	37.23
var	71758.6	1024.05	32.743	119.616	472101.0	26559.0	0.0967	6630152.20
std.dev	267.9	32.00	5.722	10.937	687.1	163.0	0.3110	2574.91
coef.var	1.4	0.99	0.365	1.610	2.2	1.0	2.9065	0.44

Table 3.1 An overview statistics summary for numeric variables

	Location	BOD5	TKN	NH3.N	P.TOT	SS	FLOW	Rainfall	Pop.density
1	Calumet	130	22	12	5.7	166	239.6	0.11	3598
2	Egan	164	29	16	5.8	181	24.6	0.11	3653
3	Hanover	151	29	18	4.6	153	8.8	0.11	5048
4	Kirie	141	27	16	4.2	183	35.4	0.11	4049
5	Lemont	164	31	17	4.6	197	2.4	0.11	2331
6	Obrien	117	23	14	3.3	142	228.2	0.11	9189
7	Southwest	560	76	20	22.3	1238	361.4	0.11	8846
8	Westside	153	24	12	4.2	216	337.2	0.11	8846

Table 3.2 Mean of numeric the variables in each location

	Location	BOD5	TKN	NH3.N	P.TOT	SS	FLOW	Rainfall	Pop.density
1	Calumet	112	11.2	4.2	3.4	274	89.76	0.32	0
2	Egan	58	7.6	4.1	1.7	119	6.34	0.32	0
3	Hanover	52	8.5	5.8	1.4	80	2.82	0.32	0
4	Kirie	58	8.5	5.1	1.6	136	17.98	0.31	0
5	Lemont	139	13.4	5.4	2.8	268	0.79	0.28	0
6	Obrien	46	7.6	3.7	1.1	71	63.99	0.31	0
7	Southwest	619	74.1	8.1	25.7	1621	167.31	0.31	0
8	Westside	85	8.5	3.2	2.3	189	105.67	0.31	0

Table 3.3 Standard Deviation of the numeric variable in each location. Standard Deviation of Population Density should be 0, as each location has the same population density for each of its observations.

	Location	BOD5	TKN	NH3.N	P.TOT	SS	FLOW	Rainfall	Pop.density
1	Calumet	62	10.0	6.0	3.4	78	136.00	0.05	0
2	Egan	61	8.0	5.4	2.0	76	5.70	0.05	0
3	Hanover	58	12.0	9.3	1.9	62	3.80	0.05	0
4	Kirie	55	10.0	7.5	1.8	96	11.90	0.05	0
5	Lemont	79	12.0	8.0	1.9	110	0.96	0.05	0
6	Obrien	41	7.0	4.6	1.1	62	64.00	0.05	0
7	Southwest	356	41.0	13.1	15.7	970	237.00	0.05	0
8	Westside	94	8.9	3.9	2.2	168	172.00	0.05	0

Table 3.4 IQR of the numeric variables in each location

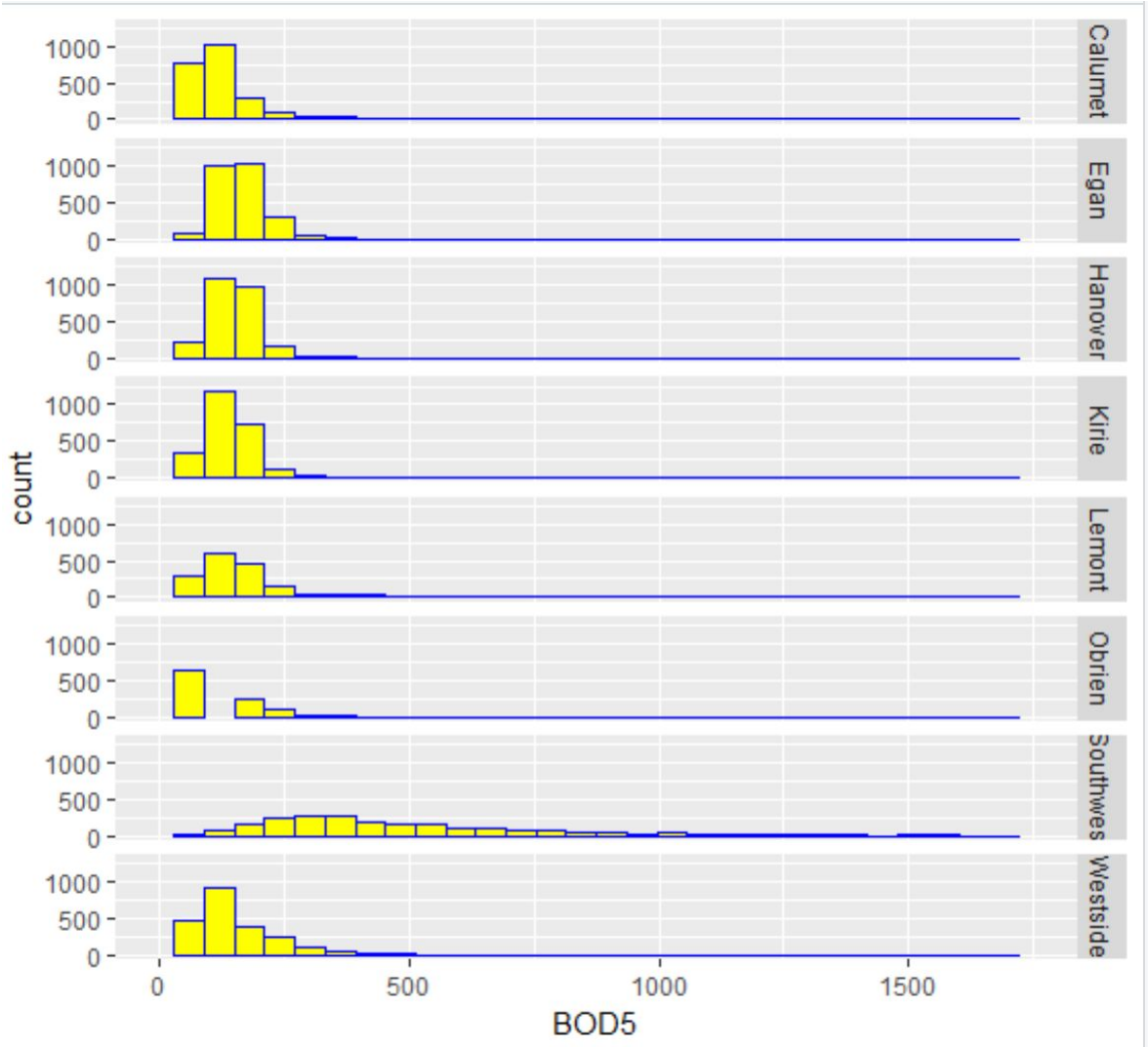
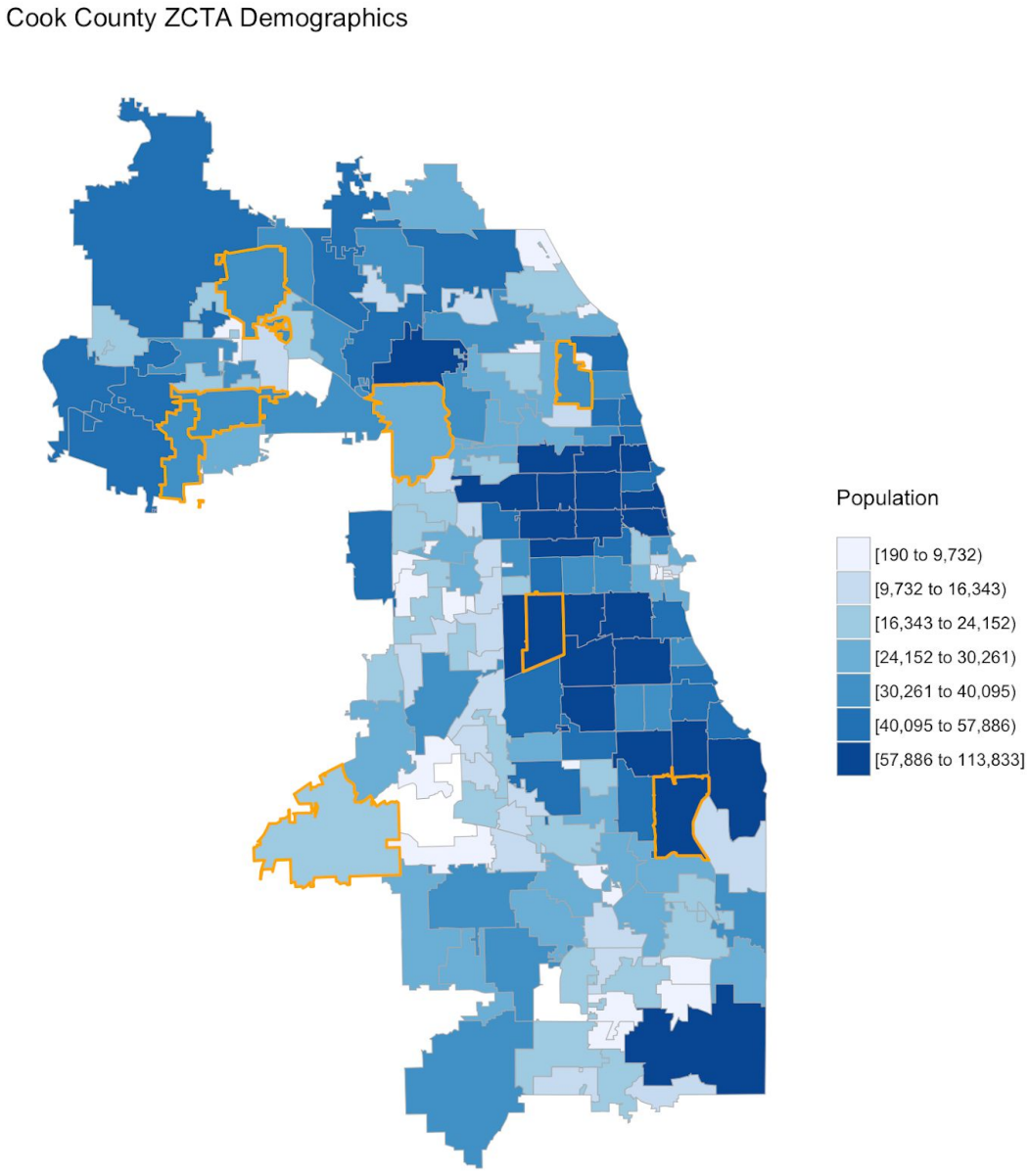


Table 3.5 Histogram of BOD5 in each location

C. **Analysis of Population**

We've already had the zip code and address of these water plants. We want draw Cook county demographics based on zip code. The main target is to see the relationship between and population distribution.



Draft county map by FIPS county code as a original graph . Then draw the border based on ZIP Code. By the density of population, make a heat map on the original graph based on population data with ZIP Code. Highlight the block which contains water plant. Finally, we can get cook county ZCTA demographic map.Figure 4.1 Cook County Demographic.