TECH BRIEF

Data Science Summer Fellows Program
Summer 2021

Modernizing Water and Wastewater Treatment through
Data Science Education & Research

Metro Wastewater Reclamation District – An Investigation of Sludge Settleability and its Causes

Austin Blanchard, Baylor University
Julia Eiken, Colorado School of Mines
Claire Teng, Baylor University

SUMMARY

The addition of activated sludge is a process in water treatment involving a high concentration of microorganisms for contaminant removal. Sludge bulking is a problem when sludge does not settle quickly and can reduce the efficiency of a water treatment plant. Sludge bulking is a problem when sludge does not settle quickly and can reduce the efficiency of a water treatment plant. A linear model created with four parameters (primary effluent total phosphorus, BioReactor #2 - zone 2A dissolved oxygen, stepfeed daily average, RAS daily average flow) and effluent temperature predicts SVI levels better than a linear model with all water quality parameters, a regression tree with lasso selected parameters, a random forest with lasso selected parameters, and a binary logistic regression with lasso selected parameters. We recommend Metro Wastewater consider the aforementioned water quality measurements as a useful set of parameters for predicting sludge settleability.

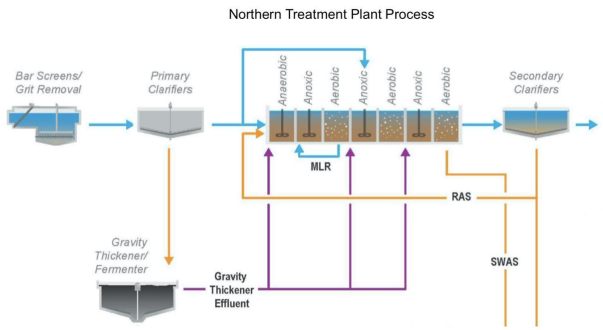
INTRODUCTION

Metro Wastewater Reclamation District's Northern Treatment Plant in Denver, Colorado occasionally experiences issues with sludge bulking, which is poor liquid-solid separation in water clarifiers. Bulking is associated with slow settling of sludge mass and is quantified by the sludge volume index (SVI) as the volume of sludge (mL) occupied per gram of settled sludge after 30 minutes. Sludge bulking is important to avoid to protect apparatuses downstream in the plant, maintain plant efficiency, and cut down on operating costs. Metro observes the most issues with sludge settling in cold weather months but has spikes year round with no explanation. Thus, Metro seeks to understand which parameters in the plant and water quality increase SVI in primary, secondary, and tertiary stages of the plant. Metro seeks to diagnose the issue with SVI to enable real-time control of SVI and maintain levels under 150.

FACILITY SYSTEM DESCRIPTION

The Northern Treatment Plant process begins with the raw influent entering the system, and solids are removed via bar screens and grit removal (see Figure 1). Next, primary clarifiers separate solids from the water. The solids enter the gravity thickener and fermenter. Carbon is introduced to the bioreactors where nutrient removal takes place. This process further separates the solids from the water. The secondary clarifiers separate waste from sludge and this is where the SVI levels are measured. After nutrient removal, the water is nearly clean and separated from the microbes in the sludge. Lower SVI levels result in fewer cycles used to purify the water, and the downstream process of post-aerobic digestion can function more

smoothly. Concentrate from the post-aerobic system is returned to the head of the plant with the nutrient load that impacts the secondary clarifier.
Figure 1.



This diagram depicts the movement of wastewater from the primary to secondary clarifiers. A variety of filtering methods are used to separate solids from the water.

DATA DESCRIPTION

Metro provided an original data set of 116 parameters and 1206 observations. The data ranged from September 2019 to March 2021. The variables included water quality data from influent, primary effluent, secondary effluent, tertiary, and other downstream processes as well as pump flow data. Later in the project Metro provided us with eight additional water quality and flow variables, including effluent temperature. All data were sampled once per day but not every day and the measurements were recorded at 1:00 AM. Some of the data were collected as twenty-four-hour averages and others were grab samples. Only a few of the total 124 parameters had complete columns of data and on the whole, measurements for each parameter were sparsely populated. However, the data set was organized and easy to understand. The variable in question, SVI, was divided into two parts due to it being sampled from one BioReactor in the plant for several years and then another for the remainder of the given time interval. Despite being two different variables in the dataset, the measurements for SVI represented one continuous record. For ease of analysis, we concatenated the two columns. Thus, we had 165 measurements of SVI to consider for our future modelling process. Each column of the data set was also renamed to its description. Metro did identify 38 variables that would have the greatest chance of having a relationship with SVI levels, which helped to narrow down this large data set.

EXPLORATORY DATA ANALYSIS

Early analysis of the data set included basic exploration of linear relationships between parameters and SVI. Each of the 38 variables deemed "important" were linearly modeled against SVI and the residuals were analyzed. This had the goal of finding variables with some correlation to SVI. Unfortunately, none of the variables showed any strong linear relationship. We also considered the time series of SVI and each of the 38 variables to identify patterns over the course of the data collection period. Again, no obvious patterns were discovered. These models and plots are featured in the RShiny App called *Finding Correlation with SVI*. Exploratory analysis also included a logistic transformation of SVI and other parameters but did not reveal much about the data. A pairwise plot was used for the 38 important variables and SVI (divided into primary, secondary, and tertiary sections for ease of viewing) to identify relationships. Relationships observed in this activity were not particularly helpful in our future modeling.

STATISTICAL ANALYSIS and RESULTS

INTERPOLATION

To fit predictive models for SVI, we narrowed down the data set to only include observations for which we had an actual measurement of SVI and chose to focus on the 38 variables deemed most important by Metro. Thus, our modelling was all based on a data set with 39 variables and 165 observations. In order to fill out each column such that our data set did not have any missing values, we used a cubic

smoothing spline to approximate the values of each parameter over time. With our data set complete, we began modeling SVI.

LINEAR MODELS AND LASSO

The first step in our process was to create a comprehensive linear model that predicted values for SVI based on the 38 parameters (see figure 3 for model statistics). This model became the reference model to compare other models against. The next step was to narrow down the number of variables in the model. In order to accomplish this, we used the lasso variable selection technique from the *glmnet* package. The lasso function computes coefficient values for an "optimal" value of lambda (a tuning parameter controlling the number of variables used in the linear model). Nonzero coefficients indicate impact in the linear model. The first iteration of lasso selection gave no nonzero coefficients based on its automatic selection of a lambda value. Thus, we fixed a lambda value of 2 to select four variables with the highest nonzero coefficients, suggesting they had the most impact in the linear model. The parameters and their lasso coefficients are listed:

- Primary Effluent TP : -2.41
- BioReactor #2 - Zone 2A DO : -3.88
- BioReactor #2 Stepfeed Daily Average : -1.89
- BioReactor #2 RAS Daily Average Flow : 7.69

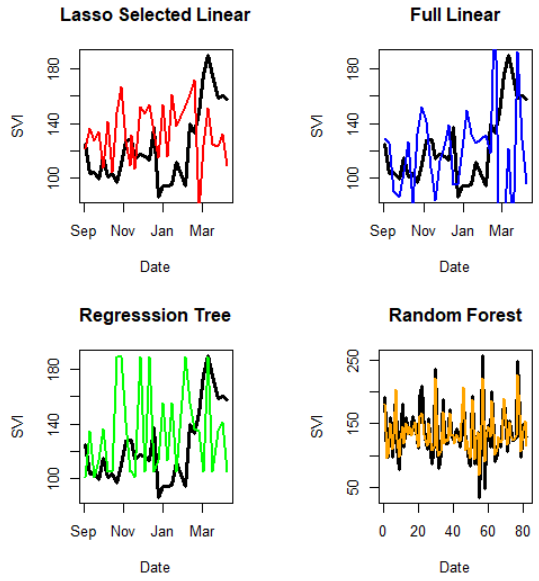
Having isolated a few variables with impact in the model, we then created another linear model with these four parameters plus temperature. This model fits the data much better (see figure 3) but still did not explain all the variability in SVI. Thus, we decided to move to a nonlinear approach using the variables selected by the lasso.

REGRESSION TREES AND RANDOM FORESTS

We first used the *rpart* and *rpart.plot* package to build a regression tree model. This method puts data frames into smaller groups and fits a model for each of those subgroups, creating a tree with a single node that branches into various outcomes. We created an RShiny app to build regression trees with all variables as a way to explore their impact.

To further investigate the relationships between the response variable SVI and the predictor variables, we used the random forest method included in the *randomForest* package to construct multiple decision trees. Using the random forest model, we created a variable importance plot that ranks the four predictor variables from the lasso model according to prediction power. This plot represents how much prediction accuracy of the model is lost by removing each of the variables. The plot indicated Bioreactor #2 RAS Daily Average Flow has the most prediction power. Although this technique is most appropriate for data sets with more observations, we found it useful to compare with our other models (see figure 2).

Figure 2.



This figure compares the predictions over time for each of the four aforementioned models with the actual SVI data. Since the random forest function does not take in date objects, the fourth plot uses a larger testing set and integer surrogate for date.

CATEGORICAL EXPLORATION

We also considered SVI as a categorical variable by classifying SVI as true if value is below 150 (the plant's desired SVI level), and false if otherwise. This approach seemed

reasonable because the actual value of SVI is not as important as whether it exceeds the plant's SVI standards for health, safety, and cost. Accordingly, we fit a binary logistic regression model to predict SVI as a binary variable (greater or less than 150). We used this model to predict whether SVI would be greater or less than the desired level with leave one out cross validation. We trained the model with the entire data set minus one row of observations and then tested the model on the one isolated row of observations. Results are listed in the table below.

Figure 3.

Model	RMSE	Accuracy
Full Linear	49.06	79.3%
Linear	42.13	81.1%
Regression Tree	47.94	78.7%
Random Forest	22.38	78.0%
Binary Logistic	------	81.1%

All root mean square error values (RMSE) in the table were averaged over 100 trials of random 80/20 training/testing splits. The accuracy rate describes the number of correct predictions (using SVI as a categorical value and leave one out cross validation) divided by the total number of predictions. AIC is similar to adjusted R^2 with a penalty.

CONCLUSIONS

Our five separate models each give unique indications of the impact of the four variables selected by the lasso selection criteria. Analyzing the effectiveness of each variable with respect to the others suggests that the linear model including the lasso selected variables and effluent temperature fits the data the best and also has the most accuracy in its predictions. Thus, we would recommend that Metro consider the four variables selected by the lasso as the most impactful water quality and flow parameters that affect sludge bulking the most.

Our models could have been improved had our data set been complete. Consistency in measurements would have increased the accuracy of all our models and increased our faith in our recommendations for the plant. With more time, we would have liked to explore the regression trees and their potential for forming models with all 38 important variables. We would also have liked to better understand the physical role of each of the four parameters selected by the lasso model, and understand which parameters from the original data set correlate to the selected variables.

REFERENCES

James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*. 2nd ed., Springer, 2021.

AUTHORS

Austin Blanchard is a Computer Science major at Baylor University who loves to play guitar in his free time.

Julia Eiken is studying Computational and Applied Mathematics with a minor in Computer Science and is also a varsity volleyball player for Mines. She loves the outdoors and trying new foods.

Claire Teng is a Data & Environmental Science major at Baylor University who has a small custom sticker business. She enjoys baking and making coffee.

ACKNOWLEDGEMENTS

First and foremost, we thank Thomas Worley-Morse of Metro Wastewater Reclamation District for providing and cleaning the data. We also thank Dr. Amanda Hering, Dr. Douglas Nychka, Aurora Waclawski, Sweta Rai, and Hattie Means for their guidance.