TECH BRIEF

Data Science Summer Fellows Program

Summer 2021

Modernizing Water and Wastewater Treatment through Data Science Education & Research

J.D. Phillips Water Resource Recovery Facility –Colorado Springs Utilities

Brighton Garrett, Baylor University
Alex Hilgefort, Colorado School of Mines
Jamie Smith, Baylor University
Yuyang Wu, Baylor University

SUMMARY

Wastewater treatment facilities use different methods to treat water to meet state and federal discharge regulations. Operated by Colorado Springs Utilities, the J.D. Phillips Water Resource Recovery Facility (JDP) treats wastewater from various parts of the City of Colorado Springs. To comply with Regulation 85, JDP must keep their annual effluent median total phosphorus (TP) below a limit of 1.0 milligram per liter (mg/L). The removal of biological-phosphorus (bio-P) is facilitated by both biological and chemical processes in wastewater treatment plants. This report studies the factors that impact efficient phosphorus removal at the JDP facility through data wrangling, exploratory analysis, and the implementation of statistical methods for modeling.

INTRODUCTION

Effluent water must comply with Regulation 85 from the JDP facility. Over the past 8 months, JDP measured TP levels in the treatment plant effluent. Our goal is to analyze the events contributing to changing concentrations of TP and to identify variables of importance to determine the processes that affect JDP's ability to treat water to desired bio-P levels.

FACILITY SYSTEM DESCRIPTION

The JDP facility treats the water that enters their facility through preliminary (grit removal), primary (sedimentation removal), secondary (biological nutrient removal (BNR)), and disinfection (ultraviolet) treatment.

The plant uses biological filters to remove nutrients such as phosphorus and nitrogen. In this process, microorganisms are manipulated to uptake additional phosphorus by controlling the environment in the bioreactor. The microorganisms accumulate phosphorus within their biomass and are removed from the system as solids waste. Removal of nitrogen occurs using a

nitrification-denitrification process and converts influent nitrogen (ammonia) into nitrogen gas.

The JDP facility was created with anaerobic, anoxic, and aerobic zones in the secondary treatment. However, they are not capable of removing all nutrients, especially bio-P. The facility is carbon-limited which is essential in bio-P removal. Supplemental carbon is added to the treatment processes as whey, a by-product of cottage cheese production that is high in biological oxygen demand (BOD). Since Feb 2020, deliveries of whey have been inconsistent and limited. To combat this shortage, operators at JDP have made several changes to the facility including dosing with acetic acid, changing the primary clarifier sludge blanket levels, and other modifications with the goal to efficiently lower bio-P levels. This project investigates all of the changes JDP has made, and analyzes them to see which are lowering bio-P.

DATA DESCRIPTION

The data consists of three types: daily laboratory data, fifteen-minute probe data, and facility dosing data.

For reference, table A.1 in the Appendix lists each variable by type. The provided data ranges from August 2020 to the end of February 2021. The samples from influent, primary effluent, and final effluent laboratory observations were measured one day after they were sampled. The frequency of these measurements varied. Additionally, YSI, Inc./ Xylem Inc. sensors were used in the basin in Zones 2, 5, and 7.

Flow of acetic acid dosing was provided between Jun 2020 until dosing stopped on Feb 2nd, 2020. Beginning in August 2020, JDP bulk dosed whey into the primary influent channel by dumping 5000 gallons in at once. In late October 2020, JDP began reducing whey input by adding it at varying rates each hour.

EXPLORATORY DATA ANALYSIS

Figure 1 indicates plotting phosphorus over time and shows changing TP concentrations in final effluent water. Identifying facility changes at times when TP concentrations changed from over-compliant levels to under-compliant levels and vise-versa was used as a strategy to determine which factors lead to better bio-P removal. An evident portion on the graph where the bio-P levels are ideal is shown in late July/ early August. Unideal conditions are shown in in October.

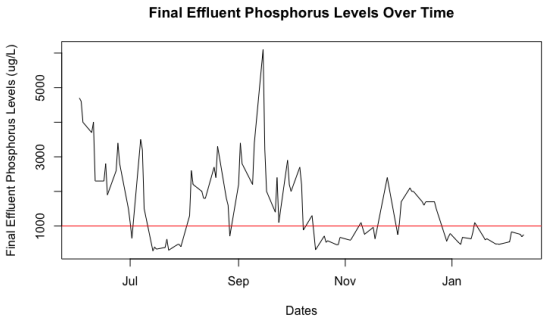


Figure 1. Phosphorus levels over time with the regulation level shown in red

SHINY APP

A shiny app (an interactive web app) was created for data analysis and exploration of the relationships among variables comparing relationships among the models to specific data points, assessing the accuracy of different models, and tracking the relationship of different variables over time.

The app has two scatter plots. The left-hand plot can be used to investigate how multiple variable trends over time and compare these to the five different models. Each model has information to validate it such as an r-squared value, residuals plot, and more. The plot on the right-hand side allows the capability to determine the correlations between variables. There are options to change the plot type, add a line of best fit, show the correlation, log transform, and show the outliers.

DATA WRANGLING

To begin wrangling our data, we organized it into two datasets: 15-minute data and daily data.

In the daily dataset, the facility dosing data (whey and acetic acid) was converted into gallons per day by adding up the rates over time.

Most of the fifteen-minute data was recorded with sensors that the stakeholder confirmed may report inaccurately. To address this problem, outlier analysis was performed on the 15-minute data through a shiny app. For each variable, with the clear exception of anaerobic oxidation-reduction-potential (Anaerobic ORP), it was discovered that observations followed a diurnal cycle. Most outliers that were flagged appeared to be a regular part of the local, day-to-day observations. Because of this, outliers in the 15-minute dataset were identified and removed on a day-to-day basis. For Anaerobic ORP, measurements below -400 and large values associated with probes being cleaned were changed to -400 and removed respectively.

Because the target variable was recorded on a daily scale, to allow data-compatibility in the model, the 15-minute data was aggregated into daily averages. This resulted in one daily data set (53 variables, 313 daily observations). Due to missing values in the data, some variables were not compatible with the models and had to be removed. Variables were eliminated with consideration of what might affect wastewater treatment processes and affect or indicate influent and effluent TP concentrations.. There were also limited date ranges of final effluent phosphorus levels, so the data set was condensed to cover just these dates (36 variables, 107 observations).

Finally, missing data values (NA values) had to be removed. All variables that had more than 35 NA values, except volatile fatty acids and acetic acid primary lab, were removed. These two variables were retained because they have a known significant effect on bio-P removal. NA values for these variables were linearly interpolated. The final modeling data set was 29 variables and 67 observations.

STATISTICAL ANALYSIS

The first step was to select a useful subset of variables that can predict TP. Four methods were used: lasso, all variable selection, backwards stepwise, forward stepwise, and leap exhaustive. The best fit for the data from the linear models was with backwards stepwise selection. Next, leave-one-out cross-validation was used to quantify the predictive skill of these models. The backwards stepwise model had the highest r-squared value (0.67) and lowest mean absolute error (MAE) value (0.37) from cross validation. The predictor variables chosen to model the response variable (phosphorus) are shown in Figure 2.

Predictor:	t Value :	Predictor:	t Value :
Total Whey Added	2.84	PTS Sludge Level 1	3.52
TSS of the Influent Lab	2.39	Anaerobic ORP	-2.85
Daily WAS Flow	2.02	Zone 2 DO	-3.13
Acetic Acid Primary Effluent Lab	-3.78	Zone 7 DO	2.00
Primary Effluent Phosphorus	2.92	Anoxic ORP	4.88
Total Volatile Fatty Acids	3.33	RAS Flow	2.30
Nitrate as Nitrogen Final Effluent Lab	2.34	Average Temperature	4.50
TSS Final Effluent Lab	3.58		

Table 1. Backwards stepwise variables with t-value

To verify the results, a non-linear model was examined in addition to the linear model. The non-linear generalized additive model (GAM) shows the possible variables that indirectly affect the efficiency of phosphorus removal. The GAM model selected Basin 1/2 mixed liquor suspended solids return activated sludge (MLSS RAS), average temperature of effluent channel, phosphorus in effluent lab, and acetic acid in effluent lab to have a high effect on the response variable (phosphorus).

RESULTS

The results of the backwards stepwise (linear model) and the GAM (nonlinear model) are very different. Because of the particularity of the two models and the limited data, the Akaike Information Criterion (AIC) value is a good basis for judging the quality of the models against each other. The smaller the absolute value of the AIC, the better the model is. The AIC for the linear model is 1106.941 and the AIC for the nonlinear model is 1086.2.

With this, the GAM model statistically is a better model. However, when looking at the variables selected for the backwards stepwise model and the GAM model, it was found that the backwards stepwise model had variable selections that had more logical reasoning in affecting the levels of final effluent phosphorus.

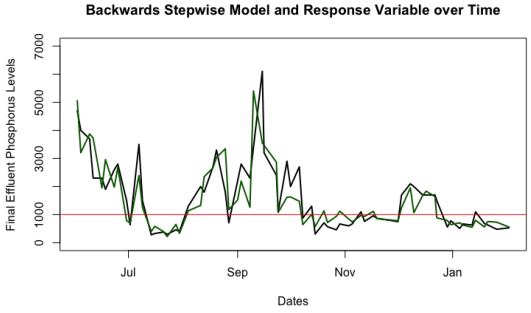


Figure 2.. Backwards stepwise model plotted over time in green with response variable plotted in black and regulation levels in red

CONCLUSIONS

The goal of the project was to identify the variables that directly correlated with efficient bio-p removal. Through the backwards selection model, we identified variables that influenced the target variable and their coefficients. With the exploratory shiny app, JDP will be able to compare variables and visually look at effects they have on final effluent phosphorus levels.

To the J.D. Phillips Water Treatment Recovery Facility, we recommend taking all measurements on the same timeline, so one can use more of the data collected. This could potentially save resources and allow for more precise data analysis.

For future work, it is useful to further examine the correlation between different carbon sources as well as investigate nitrogen measurements within the final effluent lab. Also, one should further investigate the strengths of our linear and nonlinear models to have a more confident conclusion on which model is better. To strengthen the backwards model, one should apply the GAM fitting

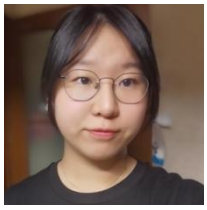
to the subset found by the backwards selection.

AUTHORS

Brighton Garrett is currently a sophomore at Baylor University studying Bioinformatics on the Pre-med track. She hopes to pursue a career in Pediatric Medicine.

Alex Hilgefort is a recent graduate in Applied Mathematics with a minor of Biology from the Colorado School of Mines. He aspires to learn more about the water treatment field.

Jamie Smith is currently a Junior at Baylor University studying Data Science. She hopes to pursue a masters in statistics following in the footsteps of her mother.



Yuyang Wu is currently a sophomore at Baylor University and studies data science. She hopes she can find more interesting things in the data science field.

AWKNOWLEDGEMENTS

We would like to give a special thanks to:
Amanda Hering PhD., Baylor University
Doug Nychka PhD., Colorado School of Mines
Michael Poor PhD., Baylor University
Tzahi Cath PhD., Colorado School Mines
Sweta Rai, Colorado School of Mines
Aurora Waclawski, Colorado School of Mines

J.D. Phillips Water Resource Recovery Facility, *Colorado Springs Utilities*
Rachel Knobbs,
Tara Kelly,
Kirk Olds &,
Shaun Thompson

We greatly appreciate the time you have dedicated to this program and helping us grow as data scientists.

APPENDIX

Table A.1

Lab samples (Final Effluent)	Lab samples (Influent)	Lab samples (Primary Effluent)	15-minute measurements	15-minute measurements (YSI probes)	Dosing records
Alkalinity	Alkalinity	Acetic Acid	influent_flow	Zone 2 NH3	Acetic Acid
Ammonia (Total) as Nitrogen	CBOD	Ammonia (Total) as N	PTS_sludge_level1_ft	Zone 2 DO	Volume of Whey
CBOD	COD	CBOD	Primary Effluent BOD	Zone 7 DO	
Nitrate as Nitrogen	Ortho_Phosphate	COD	Primary Effluent COD	Zone 5 NO3	
Nitrite as Nitrogen	Phosphorus	Ortho-Phosphate	RAS flow gpm	Zone 1 pH	
Nitrite + Nitrate as N	TSS	Phosphorus (total)	WAS flow gpm	Zone 7 pH	

Ortho-Phosphate	MLSS SVI Settle	Total Volatile Fatty Acids	WAS Solids Conc. Mg/L	Anaerobic ORP	
Phosphorus (Total)	Basin 1 and 2 MLSS RAS	TSS	Daily WAS Flow (Averaged)	Aerobic ORP	
Total Inorganic Nitrogen	Basin 1 and 2 MLSS SVI	RAS MLSS			
TSS					
Temp. effluent channel 1 and 2					