Modernizing Water and Wastewater Treatment through Data Science Education & Research

TECH BRIEF

Data Science Summer Fellows Program

Summer 2022

Sentry Water Tech – Anaerobic Digestion

Isaac Recio, Baylor University
Tara Williams, Baylor University
Kayla Balkcum, Baylor University

SUMMARY

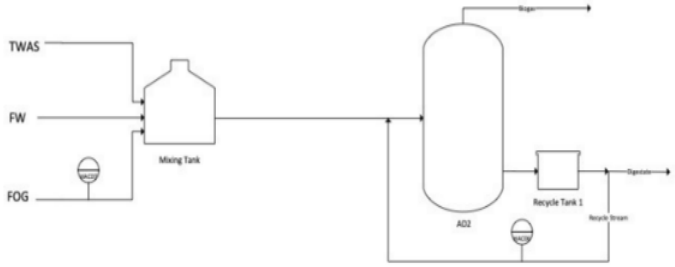
With the effects of climate change already taking place, it is important to discuss the methods being used to fight it. Biogas, a source of renewable energy consisting of methane and carbon dioxide, is one component that is gaining interest because of its ability to reuse waste and combat the effects of greenhouse gasses. Sentry Water Tech, headquartered in Prince Edward Island, Canada, monitors these wastes and biogas production in a process known as anaerobic digestion. Anaerobic digestion takes organic matter such as food and animal waste and converts it into heat and power. This brief will discuss facility processes and the use of statistical modeling to establish relationships between biogas production and other facility components.

INTRODUCTION

While anaerobic digestion and its production of biogas can bring great benefit to our environment, it is very costly and any malfunction in the system can cause financial repercussions. Among the wastes discussed earlier are fats, oils, and grease (FOG). The more FOG in the system, the more biogas produced. However, too much FOG can unbalance the system and create a greater risk of plant malfunction. By establishing relationships between biogas and wastes such as FOG and food waste (FW), we can find out the optimal levels of waste that provide maximum biogas production and prevent system imbalances.

FACILITY SYSTEM DESCRIPTION

Anaerobic digestion facilities take in thickened waste activated sludge (TWAS), primary sludge (PS), FW, and FOG into a tank where they are mixed and put into a feed line. Incoming waste is sent to a digester tank where microorganisms break down the waste. The leftover waste goes into a recycle tank and is recirculated with new incoming waste. Sentry sensors are placed throughout the system. Microbes that grow in the digester tanks grow on the biofilter in the sensors. The microbes release electrons during the oxidation of organics that send signals to a computer telling us the state of the microorganisms. Figure 1 below shows a map of the digester system with the circles attached to the feed lines representing sensor locations and AD2 representing the anaerobic digester tank.

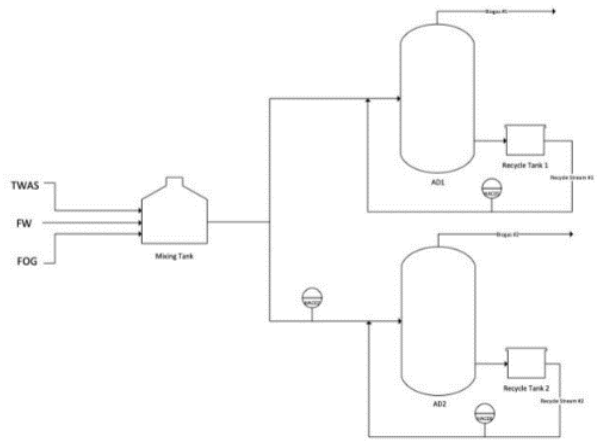


[1]Figure 1: SENTRY Full Scale Study digestion system

DATA DESCRIPTION

The data provided was collected in two separate studies: a pilot study spanning 2019-2020 and a full-scale study spanning 2021-2022. For the pilot study, the system was fed daily. Conversely, in the full-scale study, the system was fed these continuously. In addition, the pilot study was fed more FOG in

an attempt to test the limits of the system and it had three sensors and two tanks as seen in Figure 2 below as opposed to the full scale study represented in Figure 1. The data was a mix of feed and tank information provided by a wastewater treatment company partnered with SENTRY and carbon consumption rate (CCR) data collected using SENTRY's sensors. Our team was provided with six data sets, three of which contained data concerning the pilot study, and the remaining three contained data concerning the full-scale study. Sampling methods, frequency, and variables measured varied significantly between each provided dataset. In the provided data, there were also a considerable number of observations of each variable listed as "NA," which caused some relationships to be more difficult to visualize.



[2]Figure 2: SENTRY Pilot Study digestion system

EXPLORATORY DATA ANALYSIS

Initially, exploratory data analysis (EDA) focused on visualization of present relationships to determine their individual significance. It was noted that carbon consumption rate was expected to have a significant impact on the biogas production for each digester, so this was one of the relationships we first investigated. Despite this expectation, there did not appear to be a significant linear relationship between CCR and biogas. We also tried lagging CCR to account for the delay in the change of biogas after a CCR measurement. Ultimately this still did not result in a linear relationship between CCR and biogas.

To visualize these relationships, we created two separate Shiny apps [3], one for the pilot study, and one for the full-scale study. These apps created interactive web applications. Though the Shiny app made it easier to visualize these relationships, one significant issue we encountered was related to the number of missing values across the data set. These missing values led to interpretability issues in our plots viewed in our Shiny app, making it difficult to confirm any definite correlation between our variables. To solve this, we decided to visualize where missing values take place to determine which variables contain the least missing observations. Using this visualization, we saw that we had no complete observations in our 2019 data, leaving us with only 2020 data to utilize in our statistical models.

Another important issue encountered in our analysis was multicollinearity. Multicollinearity occurs when two or more independent variables are strongly correlated. Variables are perfectly multicollinear if the correlation value is +/- 1. Having multicollinear variables can affect the performance accuracy of regression models. To detect multicollinear variables, we used a correlation matrix and used a simple linear regression model to produce a plot displaying the variance inflation factor (VIF) values for the predictor variables. A VIF value of 10 suggests a strong correlation with another predictor. We noticed that if the VIF values of two variables are within a close range, it is most likely that the pair of variables are strongly correlated. Using these components, we found that several pairs of variables were highly correlated with one another. An example of multicollinearity in our data frame would be the correlation between PS feed volume (gal) and TWAS feed volume (gal), which happens to be perfectly multicollinear (+1). From each pair, we then removed one of the variables from the data set until all variables had a VIF value lower than 10.

STATISTICAL ANALYSIS and RESULTS

The goal of our statistical analysis was to establish variables of importance in relation to biogas production. In the pilot study, both digesters run parallel to each other and produce similar observations. Therefore, we only used data from the first digester of the system (AD1/PD1), to construct our models. All of the models were created using random sampling with 75% of our data being used to train the models and 25% being used to test the models.

Linear Regression

With CCR expected to be a significant predictor for biogas, we began by making a baseline model with CCR as the only predictor and biogas as the response variable. This model did not perform very well. The model's r-squared value, which is the squared correlation between actual and predicted values, was 0.01. We want this value to be as close to 1 as possible. Therefore, CCR alone was not an adequate predictor of biogas.

From there, we implemented forward and backward variable selection models. The forward variable selection model begins with an empty model and adds predictor variables one by one, with each step adding the variable that best improves your current model. The backward selection model, on the other hand, starts with all predictor variables and, in each step, removes the variable that contributes least to the response variable. The forward and backward stepwise models performed similarly, with both models concluding that volatile acids (VA), alkalinity, chemical oxygen demand (COD), and CCR as the most significant variables. Both models concluding CCR as a significant predictor was interesting since we found no indication of a strong correlation between biogas with CCR alone. The models also had some differing significant variables such as volatile solids (VS) in the forward selection model and pH in the backward selection model. This led us to explore variable importance using another type of regression model.

Random Forest Regression

We also applied a random forest regression model to determine variables of importance. This model was to be used for comparison to the linear regression model. Using random forest regression, we produced a variable importance plot. In our plot, we noticed that five predictors were significantly higher in importance in comparison to other predictors. Among these were VS, COD, organic loading rate, food loading rate, and pH.

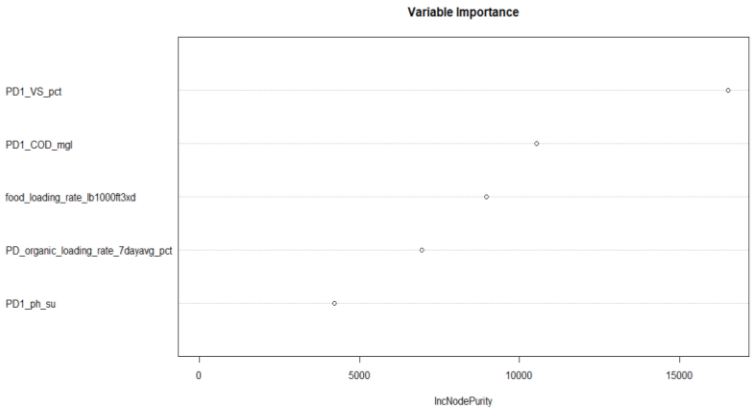


Figure 3: Variable importance plot with only the top 5 variables

In addition, the random forest models performed significantly better than the linear models with the same data when it came to making predictions. Figure 4 shows the root mean squared error (RMSE) of our random forest model with all 24 non-collinear variables versus the RMSE of our random forest with the top 5 variables and the RMSE of linear models with the same variables imputed into them. Root mean squared error is a way of signifying the accuracy of the model's predictions where a lower RMSE signifies a closer prediction to the actual value than a higher RMSE.

Type of Model	RF RMSE	Linear RMSE	Difference
All variables	16.8017933283396	17.1732018712992	-0.371408542959557
Top 5 variables	14.8508607124835	17.2371814212993	-2.38632070881586

Figure 4: Performance metrics for Random Forest and Linear Regression models

A weakness of the model though could be that, due to the random nature of the training testing split on such small amounts of data,

there could be some variance in how accurate the random forest is.

CONCLUSIONS

The models indicate that VS, COD, and food loading rate among other variables are important to predicting biogas; however, due to the limited scope of the data observations and the nature of machine learning becoming more accurate as more observations are included these variables could very well change as more data is added to the set.

We recommend longer term studies with grab samples at the same intervals. One problem with the dataset was that not all variables were taken on the same day. Because the missing data would appear at some variables more frequently than others, entire days were lost if we used those variables, which in a dataset with so few observations is important.

Alternatively, having a shorter study with more frequent measurements could also be helpful. The CCR data from SENTRY provided minutely measurements while the samples taken from the facilities at both studies were taken daily. If provided more frequent measurements to accompany the minutely CCR data, there would be more data to work with even if the study was over a smaller period of time.

REFERENCES

[1][2] These charts were given to us by the SENTRY team to help with understanding of the system

[3] https://kaylafbalkcum.shinyapps.io/Sentry_AD_shiny/
https://kaylafbalkcum.shinyapps.io/Sentry_AD_shiny_full_scale/

AUTHORS

Isaac Recio is a Statistics major, minoring in Business Administration at Baylor University.

Tara Williams is a Data Science major, minoring in Economics at Baylor University.

Kayla Balkcum is studying Data Science with minors in Spanish and Statistics at Baylor University.

ACKNOWLEDGEMENTS

We would like to thank Dr. Hering, Dr. Nychka, Luke Durell, and Natalie DeBonoPaula. We would also like to thank SENTRY Monitoring for providing us with the data used in our analysis.