Mo(Wa)²TER
Modernizing Water and Wastewater
Treatment through Data Science
Education & Research

TECH BRIEF
Data Science Summer Fellows Program
Summer 2022

Hazen and Sawyer – Optimizing Efficiency for Clean-In-Place Occurrences In Reverse Osmosis System

Cyril Pillai, Baylor University
Ella Higginbotham, Baylor University
Henry Burch, Baylor University
Lauren Walker, Baylor University
PJ Williams, Baylor University

SUMMARY

In treating water to healthy drinking standards, not only does the wastewater need to be cleaned, but the filters associated with the treatment facility need to be cleaned as well. Irregular or incorrect cleaning practices pose a serious threat to human health and safety. Water treatment membranes can become clogged or contaminated after excessive use or irregular cleanings, so it is crucial to perform cleanings at necessary times. Hazen and Sawyer, an environmental consulting firm, has a client who is looking to optimize their facility cleaning efficiency. This brief analyzes these cleaning measures and their importance through investigating various trends in wastewater data, identifying past cleanings, and predicting future cleanings through extensive exploratory analysis and logistic regression modeling.

INTRODUCTION

Hazen and Sawyer focuses on helping provide safe drinking water to its clients and controlling water pollution and its effects on the environment. One of their clients is a Reverse Osmosis (RO) facility interested in optimizing the frequency of their system cleanings.

A Clean-In-Place (CIP) is a crucial aspect of the RO water treatment process because it ensures that the system stays unclogged from debris and filters the water to the best quality. Hazen and Sawyer is interested in investigating when the RO facility performed CIPs in the past, along with predicting when CIPs should happen in the future, in order to optimize time and money.

When a CIP happens to a particular part of the system (a train), the train is off and water is not being treated. Therefore, it is important that the facility is performing CIPs when absolutely necessary.

FACILITY SYSTEM DESCRIPTION

This particular RO facility treats brackish water to make it drinkable. The source of the water and location of the facility are confidential to us.

Feedwater enters the system through one of five trains, where sediment and carbon filters remove larger debris and organic compounds. Water passes through the RO membranes and removes salt. Unlike Trains 1 and 2, Trains 4 and 5 each contain a booster pump between stages 1 and 2, which provides additional pressure to push the water through the system. Desalinated water, regarded as permeate, flows out of these five trains to another facility that was not specified to us.

As water flows through each train, data is collected. This information provides the operator with an idea of how well the train is performing. Each of the five trains in the system eventually requires necessary CIPs as compounds build up in the membranes.

DATA DESCRIPTION

Data was collected using all five trains from July 21st, 2020 to July 6th, 2021. Measurements were taken every hour, with a few incidents where the data was recorded at a particular minute and second, instead of the top of the hour.

Across the five trains, there were 206 variables: 113 raw variables and 93 calculated variables. There were varying amounts of NA values present in this dataset, but most of the variables had at least several NA values. The variables include each of the steps in the cleaning process, some of which include pressure, conductivity, flow rate, and salt passage. To look at all trains separately, we split the initial dataset into five dataframes: one dataframe per train. After creating new dataframes, there were 32 variables per each train's dataframe.

EXPLORATORY DATA ANALYSIS

Initial Analysis and Shiny App

We began our exploratory analysis by trying to discover when the RO facility had performed CIPs in the past, as we were not provided with this information. We first looked at different pressure variables over time to look for consistent, dramatic drops in pressure. We did not find any repeating trends of significant drops in pressure, so we looked for other trends in specific flux, permeate conductivity, salt passage, feedwater pressure, net driving pressure, and normalized differential pressure as well. We created a Shiny app[1], an interactive tool that allowed us to quickly produce graphs of different variables over time within each train.

Exploring NA Values

When exploring the NA values in the data, we discovered that they correspond to when the train was off. With our initial knowledge that when a CIP occurs, the train is off, we began to investigate the trends that correspond to the NA gaps in the data. We put vertical lines to represent these gaps on the plots of important variables, shown in Figure 1. In seeing consistent trends among variables before NA gaps, we assumed that the NA gaps corresponded to past CIPs.

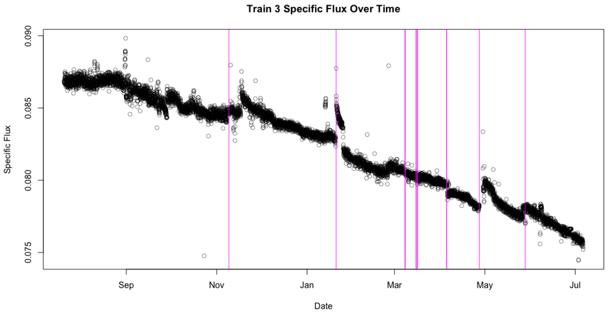


Figure 1: Plot of specific flux for Train 3 over time, and potential CIPs (when train is off) represented with magenta vertical lines.

Narrowing Down Potential Past CIPs

After our first stakeholder meeting, we were informed that Train 1 is old and Train 2 has faulty sensors, resulting in both of these trains having unreliable data. With this information, we focused our analysis on Trains 3, 4, and 5. We initially moved forward in our analysis by assuming that a CIP occurred when the train was off, using the binary variable in our dataset classifying when the train was on or off. We applied the same technique of adding vertical lines representing when the train was off to plots of important variables. When looking at these plots, we discovered that the time frames where the train is off varies drastically. With these varying time periods,

[1] https://pj-williams.shinyapps.io/Hazen__and_Sawyer_Shiny_App/

we realized that not all instances of a train being turned off were due to a CIP. However, since we know that the duration of a CIP typically ranges from 2-8 hours, we decided to only include times where the train was off between 2-8 hours. We moved forward using these parameters to mark all potential CIP events.

Exploring Trends in the Data

During our analysis, we received a guide explaining when CIPs should occur based on 3 parameters [1]:

> “Elements should be cleaned when one or more of the below mentioned parameters are applicable:
>
> - The normalized permeate flow drops 10%
> - The normalized salt passage increases 5-10%
> - The normalized pressure drop (feed pressure minus concentrate pressure) increases 10-15%”

To investigate this further, we created a dataframe that contained all the times the train was on, as this would allow us to analyze trends in the data between potential CIPs. With these separated time frames of when the train was on, we looked at percent change, total change, and rate of change for the three variables listed above. In studying the change of these three variables over time, along with our knowledge that CIPs occurred only when the train was off, we discovered that the trends we saw in our data were mostly inconsistent with what the source outlined. With this, we began to explore trends in other variables to see if they followed a more consistent trend between potential CIPs. We found that specific flux, net driving pressure, and feedwater pressure displayed consistent patterns before potential CIP occurrences among all trains. Figure 2 is one example of a consistent pattern, showing specific flux decreasing between potential CIPs.

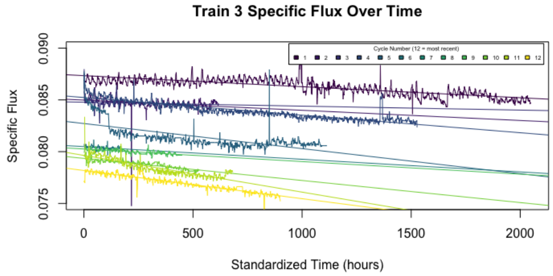


Figure 2: Plot of specific flux for Train 3 for each time period when the train is on (between CIPs).

STATISTICAL ANALYSIS/RESULTS

Preparing Data for Modeling

Using the variables that displayed the most consistent trends, we isolated the time frames between potential CIP events and calculated the percent change of each of the variables within them. Then, we averaged the percent changes for each of the variables. With these averages, we identified ranges when all the variables exhibited changes exceeding their respective mean. Out of the potential CIPs, we established these as the CIP occurrences we would model with. We consolidated our results into a column in our dataframe to indicate whether a CIP happened after that time period.

Logistic Regression Model

With the calculated percent change values from a random sample of 70% of the data from Trains 3, 4, and 5, we fit a logistic model that uses the following independent variables to predict the dependent variable. The independent variables are specific flux, net driving pressure, feedwater pressure, percent recovery, and normalized permeate conductivity. The dependent variable is the new CIP prediction variable we created. This model outputs a probability value between 0 and 1. A probability value of 0.55 or higher indicates that a CIP should be performed shortly after the time period. In practice, the sensor data is continuously fed into the model. Ergo, we suggest a CIP be conducted as the model output approaches 0.55. We were unable to predict specifically how long until a CIP should be performed due to our small sample size.

Logistic Model Equation:

$$P = \frac{a}{(1 + a)}$$

$$a = e^{-7.77 - 72.61 * SpFl + 65.05 * NtDP + 58.46 * Pf - 58.94 * PerRa + 1490.63 * NCp}$$

Note: All variables are percent changes during each non-cleaning time period.
SpFl = Specific Flux
NtDP = Net Driving Pressure
Pf = Feedwater Pressure
PerRa = Percent Recovery
NCp = Normalized Permeate Conductivity

Results

With our logistic model, we tested the model on the rest of the data (30%) not used in building our model. When testing this specific model with 4 of 13 total CIPs, we were able to correctly predict the CIP variable 100% of the time.

Note: Due to the constrained time period under which the given data was recorded, we were only able to isolate 13 CIP events. This makes our results and accuracy subject to error.

CONCLUSIONS

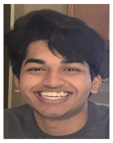
Our logistic regression model predicted CIP events with high accuracy. With a small sample size and instances of inconsistent data, the model reflects the data that was given to us. Since we were not provided information on past CIPs, we used time frames when the train was off between 2-8 hours as our past CIPs. We then built a model based off of these past CIPs and variables that had consistent trends over time to predict when CIPs should happen in the future, and how soon into the future it should occur.

The trends in our data were not consistent with the parameters provided from a reliable source. Due to this, as well as lack of information of past CIPs, we have concluded that either our assumption of CIPs occurring when the train is off is incorrect or the RO facility is performing CIPs at the incorrect times. Instead of predicting the exact timing of future events, we shifted our focus to predicting how long until a CIP may need to be performed based on trends of several variables. Given information on when past CIPs had occurred, we could apply this data to our model to build a more robust and reliable one.

REFERENCES

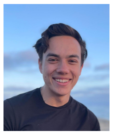
[1] DOW FILMTEC™ Membranes Cleaning Procedures for DOW FILMTEC FT30 Elements. (n.d.).

AUTHORS



Cyril Pillai is a junior at Baylor University, studying Bioinformatics.

Ella Higginbotham is a senior at Baylor University, studying Environmental Science and Economics.



Henry Burch is a junior at Baylor University, studying Data Science and Statistics.

PJ Williams recently graduated from Baylor University with his Bachelor's in Chemistry.

Lauren Walker is a sophomore at Baylor University, studying Data Science with a minor in Economics.

ACKNOWLEDGEMENTS

We would like to thank Dr. Douglas Nychka, Matthew Hofkes, Dr. Tzahi Cath, Dr. Amanda Hering, Dr. Luke Durell, and Natalie deBonoPaula. We would also like to thank Hazen and Sawyer professionals Javad Roostaei and Ben Stanford.