Mo(Wa)²TER

Modernizing Water and Wastewater
Treatment through Data Science
Education & Research

TECH BRIEF

Data Science Summer Fellows Program

Summer 2021

Hazen and Sawyer – Drinking Water

Daniela Cortes Bermudez, Baylor University
Dante Hart, Baylor University
Cassie Nataro, Baylor University
Sydney Toler, Colorado School of Mines
Ty Wicks, Baylor University

SUMMARY

Drinking water must be rigorously treated to meet potability standards in the US. Hazen and Sawyer, an environmental consulting company committed to creating solutions to water engineering challenges, provides safe drinking water to communities across the country. This brief will explore the water treatment processes and data collection methods of an unknown facility provided by Hazen and Sawyer. It will then demonstrate the data cleaning, online data visualization tools and results of statistical methods used to model relationships between water quality characteristics and treatment processes.

INTRODUCTION

Many drinking water treatment plants utilize filters to treat water by removing contaminants through physical processes. These filters are maintained with backwashing which removes particles in the filter by reversing the flow with clean water. Backwashes expend time and clean water, so minimizing the number of backwash occurrences would optimize the facilities operations and resources. The two factors that trigger a backwash event are head loss — a loss of energy per unit weight of a fluid — and turbidity — the clarity of the water. By identifying water quality variables that have the greatest impact on head loss and turbidity, the number of required backwashes may be monitored and reduced, improving the facility's efficiency.

FACILITY SYSTEM DESCRIPTION

This facility is a conventional drinking water plant with a flow of approximately 40 million gallons per day. The water predominantly comes from one reservoir, but also receives flow from a secondary reservoir.

The water begins treatment by entering a rapid mixer where chemical dosing occurs. Then it travels to four flocculation and four sedimentation basins, which split into sub-basins A and B. Next, the water flows through one of sixteen different filters, with basin 1 favoring filters 1-4, basin 2 and 3 favoring filters 5-12 and basin 4 favoring filters 13-16. At this facility, a backwash is triggered every 108 hours, but it can also be triggered when head loss accumulation is greater than eight feet or when turbidity is greater than 0.3 NTU. Lastly, the effluent water moves to a secondary treatment and disinfection before exiting the facility.

DATA DESCRIPTION

The data consisted of four parts: Filter, Chemical, Master Basin, and Alum Dosing. All of the data were gathered from 2017 to 2020.

Water sampling was collected and analyzed in a lab for the Chemical data. These were daily measurements; observations from the rest of the datasets relied on online sensors, which collected hourly data.

The Filter dataset contained head loss (feet), effluent turbidity (NTU), flow in million gallons per day (mgd), and the corresponding filter index. Several duplicate dates and times were contained in this dataset. Additionally, for filters 10-15 there was a loss of data from December 2019 to March 2020 due to repeated observations.

The Chemical data set contained over 40 variables regarding water quality. Additional variables refer to operational data, such as which filters were in use and temperature.

The Master Basin data set contained flow, turbidity and pH variables for each of the four basins. There were a total of 26,929 observations, with scattered periods of basin flow data missing. Basin 4 contained the most missing flow data.

The Alum Dosing data set contained the flow (mgd) and dosage of alum (mg/L) for each of the four basins. All basins were missing numerous alum dosing observations.

DATA WRANGLING

Due to missing observations across the various data sets and the need to use variables from different data sets, the data needed to be cleaned and unified. We created a column to identify when the filter was in use or turned off. We combined water quality data from all four data sets.

Given the missing data from December 2019 to March 2020, the date time observations were restricted from December 2018 to December 2019. Missing observations for the Alum data were omitted. Additionally, runtime calculations were derived and appended from the dataset along with head loss accumulation rate, which is the difference between the minimum and maximum head loss over the runtime.

To simplify results, the final data set was summarized to: (i) the mean of the Chemical and Filter observations between backwashes; (ii) the runtime per filter; and (iii) the filter index.

SHINY APP & VISUALIZATION

During the exploratory phase, the main goal was to visualize the relationships between the variables. We created Shiny apps, a tool tool allowing the user to interact with the graph, for this purpose. We were able to plot numerous variables over time to compare the relationship between head loss, turbidity, and other water quality measurements.

One plot we created displayed the relationship between settled water pH and settled water turbidity. The facility maintained a pH around 6, but we discovered that increasing the pH above 6.5 was consistent with a sizable decrease in turbidity. Along with this relationship, we also found that the pH levels within each sub-basin would occasionally differ from each other, suggesting there could be an inconsistency in the rapid mix process.

In the app[1] looking at head loss, we chose to look at how water temperature affects the filter head loss. This app did indicate that as water temperature rose, head loss would also rise. However, this relationship was only prevalent in the final year of observations. This prompted us to look at just the final year in some of the subsequent models.

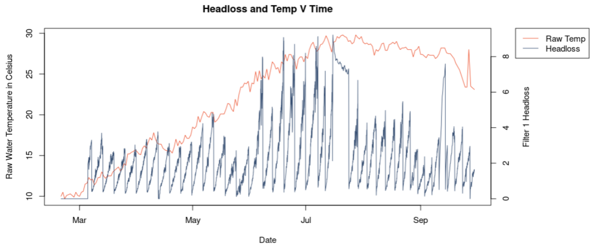


Figure 1: Head loss (blue) and Temperature (red) relationship

Within the turbidity app, we compared each of the four basins with the filters that they are most predominantly associated with. This helped to view how basin turbidity affected filtered turbidity. Along with this, it helped to show the filter use rotation along with individual filter performance.

https://mowater-hs.Shinyapps.io/app_-_combined

STATISTICAL ANALYSIS & RESULTS

Settled Water Turbidity Model:

A linear model was fit to estimate settled water turbidity using three predictor variables: raw water turbidity, alum dose and settled water pH (a surrogate for lime dosing). The residuals versus fit plot suggested that a linear model was reasonable. This model only included the past year of data provided in order to take into account changes made at the facility. The results showed a statistically significant relationship between settled water turbidity and all three predictor values as indicated by their p-values less than 0.05.

When looking at the predictors' coefficients, the model revealed for a one unit increase in raw water turbidity there was a 0.203 increase in settled water turbidity. Because turbidity indicated the presence of inorganic and organic materials in the water making it cloudy or murky, we expected there to be a positive relationship between the raw and settled water turbidity. For the chemical dosing variables, the model indicated for a one unit increase in alum dose there was a 0.029 decrease in settled water turbidity, and for a one unit increase in settled water pH (lime dose) there was a 0.066 decrease in settled water turbidity. The model had an R^2 of 0.5959, showing the model does a satisfactory job at fitting the observed settled water turbidity.

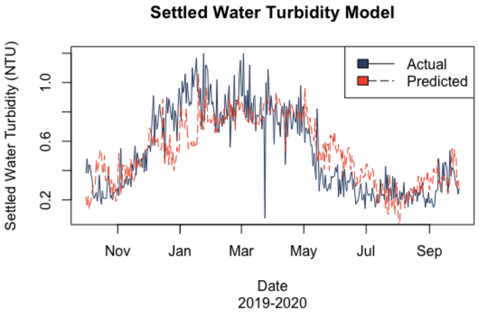


Figure 2: The actual versus predicted values from the linear regression settled water turbidity model.

A time series plot (Figure 2) was created to visualize how well our model predicts settled water turbidity when holding the other variables constant. When comparing the two lines, the model underpredicts in February and overpredicts in June.

Head Loss Accumulation Model:

Linear regression was also used to investigate different head loss values between filters. This was done using the final head loss accumulation dataset referenced in the Data Wrangling section. The response variable was the log head loss accumulation rate. The predictor variables were: (i) mean effluent filter turbidity; (ii) mean alum dosage per basin; and (iii) raw water temperature in Celsius.

There was a 7.158 (ft/hr) increase in head loss accumulation rate for a one unit increase in mean effluent filter turbidity. One unit increase in mean temperature resulted in a 0.022 increase in head loss accumulation rate. The p-value for both variables is less than 2e^{-16}, making them the most significant predicting variables. For each unit increase in mean alum there is a -0.009 decrease in head loss accumulation rate. This predictor variable is also significant with a p-value of 0.049.

We built a time series plot to explore how well our predictions based on the linear model fit the real head loss accumulation rate values. An interactive version of this plot is available in the Shiny App. This plot revealed that there was an underprediction of the real values (Figure 3). Therefore, we can conclude that the predictor variables used in the model are not the only contributing factors in the head loss accumulation rate.

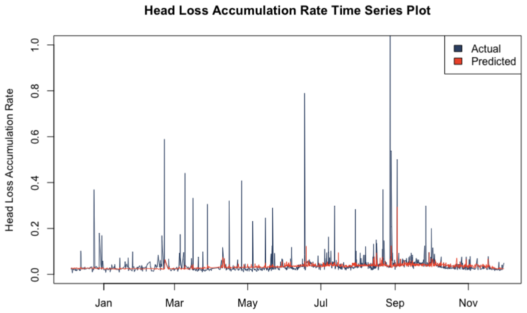


Figure 3: Head loss accumulation rate time series from December 2018 to December 2019.

Random Forest Results:

We applied a random forest regression model to predict filter head loss. The random forest model resulted in a RMSE of 1.386852. Variable importance was measured and we found that influent flow and turbidity resulted in the highest mean decrease in accuracy (60% and 45%, respectively) when they were left out of the model.

A second random forest model was fit to predict the head loss accumulation. The predictor variable was log (head loss accumulation rate). The RMSE was 0.22577. The most important variable was the mean temperature resulting in a mean decrease in accuracy of 40%.

CONCLUSIONS

The first goal of this project was to visualize the large quantity of data provided. Through the online Shiny app, Hazen and Sawyer will be able to compare influential factors across the filters and basins.

Second, we investigated which variables impact settled water turbidity. Our modeling indicated significant relationships between settled water turbidity and raw water turbidity, alum dose and settled water pH.

We created a model to see which process changes influence head loss accumulation. We found that effluent filter turbidity had a significant relationship with head loss accumulation.

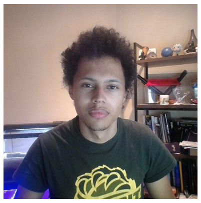
We fit two random forest regression models which suggested that flow and mean temperature were influential in estimating filter head loss.

Additionally, investigating the data collection methods is suggested to look at the causes of multiple duplicate entries occurring throughout the data sets.

AUTHORS

Daniela Cortes Bermudez

Studying a Data Science major and a Biology minor at Baylor University. Daniela enjoys live music and dancing.



Dante Hart

Studying Computer Science at Baylor University. Enjoys painting and playing games whether they're online or tabletop.

Cassie Nataro

Studying Statistics and Public Relations at Baylor University, Cassie enjoys listening to podcasts and swimming.

Sydney Toler

Recent Colorado School of Mines graduate with a major in Applied Mathematics and a Computer Science minor. Sydney enjoys hiking and skiing.

Ty Wicks

Studying Data Science and Geology at Baylor University. Ty enjoys watching and playing sports.

ACKNOWLEDGEMENTS

Amanda Hering PhD., Baylor University
Doug Nychka PhD., Colorado School of Mines
Tzahi Cath PhD., Colorado School of Mines
Luke Durell, Baylor University
Tayler Elwell, Colorado School of Mines

Hazen and Sawyer Professionals,
Ben Stanford
Troy Walker
Evan Ged
Tyler Hudson