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

TECH BRIEF

Data Science Summer Fellows Program

Summer 2022

Aeration Basin Optimization Using Sentry Sensors

Derik Boonstra, Baylor University
Kate Pogue, Baylor University
Pedro Prudencio, Baylor University

SUMMARY

One of the many stages of wastewater treatment involves aerating the wastewater in order to remove dissolved contaminants. This is an expensive process due to energy consumption. Sentry Water Technologies have developed the Sentry sensor to monitor *Carbon Consumption Rate* (CCR) in order to optimize the aeration process and prevent over and under-aeration. This tech brief explores the processes and methods used to analyze the given data and create a model to forecast influent airflow in an aeration basin in order to maintain a stable level of oxygen in the water.

INTRODUCTION

Wastewater treatment is a complex process consisting of distinct stages that prepare water to be returned to the environment. The stage this project focuses on is the aeration tank, where oxygen is pumped into the water so microbiota can metabolize and remove certain contaminants. Ideally, aeration is managed such that *Dissolved Oxygen* (DO) concentration stays within a certain range and the water is not over or under-aerated. Over-aeration is problematic because it reflects money being wasted on energy usage. Under-aeration, though less common in this case, damages the biological process by failing to meet the *Biological Oxygen Demand* (BOD) of the microbiota at a given time. This project aims to solve these problems by utilizing data provided by Sentry sensors which measure metabolic activity in water flowing through the treatment plant. To do this, we developed a model that could forecast the estimated BOD of the water and used it to predict the appropriate airflow that would keep DO within an acceptable range.

FACILITY SYSTEM DESCRIPTION

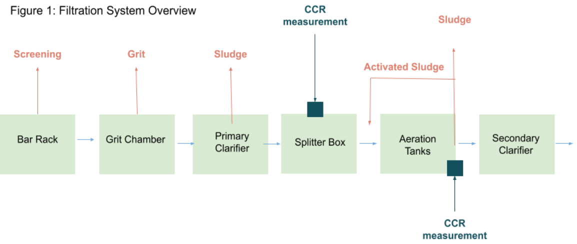


Figure 1: Filtration System Overview

Figure 1 was created based on Sentry's process diagram. The above utility consists of 6 consecutive water purification units: the bar rack, grit chamber, primary clarifier, splitter box, aeration tank, and secondary clarifier. Measurements of CCR are taken in the splitter box just before water enters the aeration tank and inside the aeration tank just before the water exits, as shown in Figure 1. The tank has 3 identical lanes, each consisting of at least 3 zones or "pockets". Only data from pockets 2 and 3 were given.

MoWaTER Baylor University MINES NSF

DATA DESCRIPTION

The original data contained seven variables measured between November 1, 2019 and November 18, 2020:

- Datetime: Date + time the measurements were taken
    - Initially contained two columns due to both CCR being collected more frequently than the rest of the data
- CCR (Sensors 1 and 2): CCR before and after aeration– measure of metabolic activity in the water before and after aeration, respectively, using the Sentry probes
- Ammonia: Concentration of effluent ammonia in the water, measured in mg/L
- Airflow: Measure of the amount of air being pumped into the water in m^3/hr
- DO (Pockets 2 and 3): Concentration of dissolved oxygen in two of the zones in mg/L
- Waterflow: Measure of influent water into the aeration basin in L/s
- MLSS: Mixed liquor suspended solids– concentration of suspended solids in the aeration basin in mg/L

The two CCR measurements were taken minutely, while the rest of the variables were measured every 15 minutes.

Some points to note regarding unusual values:

- There was a gap in the month of February in both CCR variables following a period of inconsistent measurements
- There appeared to be a gap in the DO Pocket 3 measurements, where from late May to late July flatline at a value of 0.28
- A few of the Ammonia measurements gave negative values
- MLSS had no data before May 2020 and was the only variable measured after October 2020

EXPLORATORY DATA ANALYSIS

Data Wrangling

Before beginning any analysis, any unusual values were first confirmed to be due to sensor error and were replaced with NAs. All data after September 30, 2020 was then eliminated since MLSS was the only variable being measured after that point. Next, both CCR variables were averaged into 15 minute intervals using their respective means to allow for a single date/time column across all variables. Lastly, the *estimated BOD* (BOD est.) was created using,

BOD est. = Waterflow • CCR (Sensor 1)

Though not an exact value of BOD, this can state roughly how much oxygen would be required based on the metabolic rate in a given volume of water.

Exploratory Data Analysis (EDA) and Conclusions

To help with the EDA, we made an interactive shiny app to compare the relationships between the different variables across different time scales.

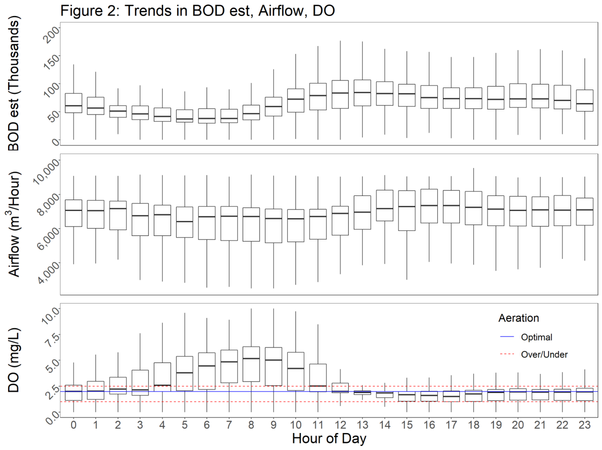


Figure 2– The main takeaway from the EDA was the trend in hours 5-10 among the three above variables. BOD est. is lower during this time, and airflow decreases to accommodate for this. However, airflow is not being lowered enough, which results in over-aeration occurring as high as 79% of the time. This led to the idea of creating an airflow model to further reduce aeration energy usage.

STATISTICAL ANALYSIS and RESULTS

More Data Wrangling

Based on the EDA, two columns were added for month and hour into the dataset as categorical variables to account for the temporal patterns of the data. The data was also further reduced to measurements after March 5, 2020 because the inconsistencies in the CCR data prior to that date were disrupting the BOD est. data and compromising the accuracy of the models.

Linear Regression for BOD est.

The first step was to develop a linear model to forecast BOD est. All of the quantitative variables were lagged by one step, allowing BOD est. to be forecasted 15 minutes in advance. To analyze the forecast performance, the data was split into training and testing sets by randomly selecting 70% of the data as a training set. After generating models with every possible predictor combination, a model was picked based on a combination of the *Root Mean Squared Error* (RMSE) and the number of variables used. In other words, the model should require as few variables as possible while still maintaining a low enough RMSE that it can effectively explain the data. The final regression model is as follows:

$$BOD\_est_t = 38{,}480 + 0.608 \bullet (BOD\_est_{t-1}) + \textit{monthly coefficients} + \textit{hourly coefficients}$$

Month	Coeff
Apr	-4,964
May	-10,999
June	-13,547
July	-14,331
Aug	-18,850
Sep	-21,093

Hour	Coeff	Hour	Coeff
1	-879	13	8,255
2	-4,395	14	7,643
3	-5,873	15	5,099
4	-8,231	16	3,647
5	-7,825	17	4,053
6	-7,963	18	3,324
7	-8,894	19	4,046
8	-4,482	20	5,033
9	479	21	5,660
10	3,237	22	3,798
11	6,029	23	980
12	7,255		

The final regression model had a RMSE of 19,430 relative to a mean of 69,490.

Random Forest Modeling for BOD est.

Wanting to improve on the regression model, we decided to incorporate random forest (RF) regression. The same process as earlier was carried out for the RF models, but now three more steps of lag for BOD were added for a total of four incrementally lagged BOD est. variables. As before, the most parsimonious model was selected, which now included the four lagged BOD ests along with month and hour. This model had an RMSE of 15,710, but it could not be easily interpreted because random forest is a black box method. For this reason, we are showing both models to our stakeholders despite the RF model better explaining the data.

Aeration Modeling

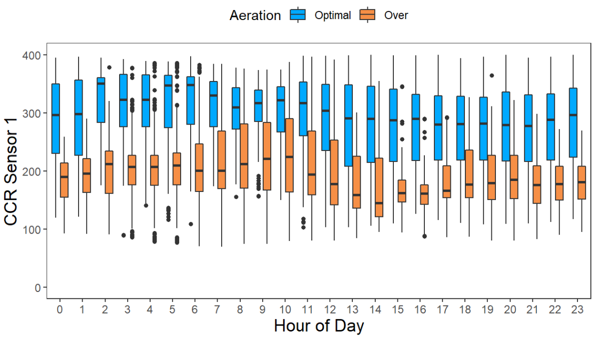
After creating a model for BOD est, a linear regression model for aeration was created based on instances of optimal airflow, defined as when DO is between 1 and 2.5 mg/L. Using the forecasted BOD est. as a predictor, the RMSE of the model was 685 relative to a mean predicted airflow value of 6,917m^3/hr, and had an R^2 value of 0.66. A RF model was also created using the same criteria and had an RMSE of 560 with the same mean as the linear model.

Analysis of Optimal vs. Over-Aeration

Both models were then applied to instances of over-aeration in the hopes that they would predict a lower airflow than the actual value. However, both models performed poorly by predicting higher airflow on average. The median and mean residuals of the airflow for the regression and random forest models was -178 and -200, and -24 and -17, respectively. These results pointed towards modeling airflow by using more specific instances of optimal airflow.

'Super Optimal' Airflow Modeling

Figure 3: CCR and Dissolved Oxygen Aeration



Based on Figure 3, over-aeration occurs when CCR is low, so subsetting by only optimal airflow results in higher CCR which is not representative of over-aeration. So, in a final effort to improve the model, we narrowed down the optimal airflow model to instances where the optimal aeration CCR was lower than the 75th percentile of the over-aerated CCR in a given hour. This created a subset of 'super optimal' airflow that could better account for periods of over-aeration. A regression and a RF model were run for this new data set and had a RMSE of 722 and 355, respectively. Both models had a mean of around 6,650 m^3/hr and the regression model also had an R^2 of 0.74.

The super-optimal models were then applied to the over-aerated data. In units of m^3/hr, the median and mean residuals of the airflow for the regression and random forest models were approximately -17 and -63, and 15 and 69, respectively. Based on the table, the models most often predict a decrease in airflow usage, with the difference being as high as 12% or higher in certain cases.

Super-Optimal to Over-Aeration Median Predicted

Hour	Actual	Regression	Random Foret
0	7,205	6,879 (-4.5)	6,776 (-6.0)
1	7,061	6,836 (-3.2)	6,809 (-3.6)
2	7,174	6,759 (-5.8)	6,602 (-8.0)
3	6,617	6,549 (-1.0)	6,411 (-3.1)
4	6,604	6,476 (-1.9)	6,413 (-2.9)
5	6,168	6,631 (7.5)	6,299 (2.1)
6	6,720	6,745 (0.4)	6,584 (-2.0)
7	6,758	6,542 (-3.2)	6,285 (-7.0)
8	6,756	6,255 (-7.4)	6,008 (-11.1)
9	6,806	6,157 (-9.5)	6,134 (-9.9)
10	6,924	5,587 (-19.3)	6,080 (-12.2)
11	6,868	5,985 (-12.8)	6,141 (-10.6)
12	6,771	6,484 (-4.2)	6,445 (-4.8)
13	7,048	6,788 (-3.7)	6,769 (-4.0)
14	7,529	7,553 (0.3)	7,752 (3.0)
15	7,619	7,915 (3.9)	8,289 (8.8)
16	8,050	7,797 (-3.1)	8,273 (2.8)
17	7,928	7,964 (0.5)	7,131 (-10.1)
18	7,490	7,261 (-3.1)	8,005 (6.9)
19	7,509	7,068 (-5.9)	7,757 (3.3)
20	7,350	7,033 (-4.3)	7,115 (-3.2)
21	7,351	7,018 (-4.5)	7,372 (0.3)
22	7,482	7,000 (-6.4)	7,013 (-6.3)
23	7,256	6,942 (-4.3)	7,134 (-1.7)

( ) Denotes Percentage Change

CONCLUSIONS

The overarching goal of this project was to be able to effectively forecast airflow 15-60 minutes in advance using the predicted BOD est. in order to account for periods of over and under-aeration. While the airflow prediction model applied to over-aeration data showed relative success, we were not able to apply the model to periods of under-aeration due to time constraints. This would be the next logical step with these findings, though under-aeration is much less common than over-aeration. Other modeling approaches such as LASSO regression, leave-one-out cross-validation, and rolling window verification may aid in future analysis as well.

A major takeaway from this project is that CCR plays an important role in determining DO and a utility's ability to meet BOD at a given time. Almost all instances of over-aeration occur during low CCR intervals. This implies that the Sentry sensors' measuring processes can prove highly valuable for regulating airflow in the aeration tank, which is good news for our stakeholders.

AUTHORS

Derik Boonstra just graduated with majors in: Business Honors, Accounting, Finance, and Statistics. He will be attending Baylor to pursue a PhD in Statistical Sciences with an interest in Financial Statistics

Kate Pogue is a Public Health and Statistics major at Baylor who is interested in research for population health applications.

Pedro Prudencio is a biochemistry major at Baylor University with an interest in data science and statistical analysis.

ACKNOWLEDGEMENTS

We would like to thank Dr. Mandy Hering, Dr. Douglas Nychka, Natalie De Bono Paula, and Luke Durell for overseeing and guiding this project. We would also like to thank our stakeholders Colin Ragush, Jack Ambler, Mahnaz Zare, and Kash Bahri for providing the data and helping direct our efforts.