Modernizing Water and Wastewater
Treatment through Data Science
Education & Research

TECH BRIEF

Data Science Summer Fellows Program

Summer 2022

Denver Water – Investigating CSMR

Nathan Hontz, Baylor University
Emily Warwick, Baylor University
Maya Maes-Johnson, Colorado School of Mines

SUMMARY

Denver Water provides water to over 1.5 million people in the city of Denver and surrounding suburbs. We analyzed water entering and exiting two of their treatment plants, Marston Lake and Foothills. Our main area of focus was investigating the chloride to sulfate mass ratio (CSMR) in the water, which has been correlated with the levels of lead in the water. By analyzing the relationship CSMR has with other factors in the water, we created linear models and used other statistical methods to predict CSMR over time.

INTRODUCTION

The chloride to sulfate mass ratio (CSMR) in water has a high correlation with lead precipitation. While high CSMR itself isn't an issue, the correlation it has with high lead levels could pose a problem. At the Denver Water facilities, lead has been removed from the delivery systems to homes, but it could still cause an issue in older homes that have outdated piping. By finding variables that are related to CSMR, we can predict when CSMR may spike, and when it is most variable. We were interested in finding relationships between CSMR and other measurable variables such as Total Organic Compounds (TOC), pH, and chlorine dosages. We also wanted to investigate the differences between the CSMR in water entering the plants (influent) vs exiting the plants (effluent), as well as the variations between the two different water treatment plants, Marston and Foothills.

FACILITY SYSTEM DESCRIPTION

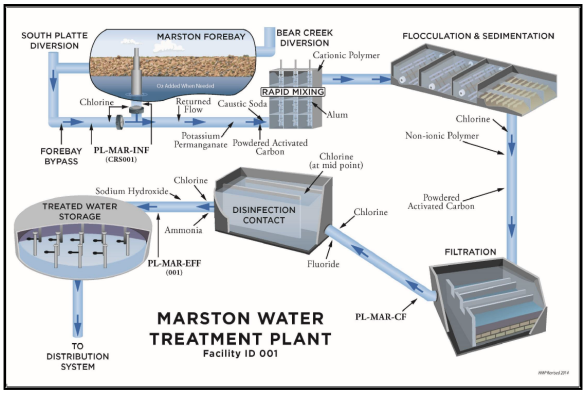

Fig. 1: Marston Water Treatment Plant, Denver Water

Denver Water treats from two different sources: Marston Lake and Foothills water run-off. Influent water is treated with chlorine and a variety of other cleaning chemicals. It then undergoes rapid mixing to thoroughly combine the coagulant (alum) with the dirty water. Flocculation and sedimentation then occur, removing suspended particles from the water. This is followed by filtration and more chlorine treatment until the water is ready for distribution. While the water supply may be different at each plant, the treatment processes for the Foothills and Marston Water Treatment Plants are nearly identical.

DATA DESCRIPTION

Denver Water provided our team with data from their Marston and Foothills Water Treatment Plants from 2015 to 2021. The data contained grab sample information about the anions (Br, F, Cl, SO_4, NO_2, NO_3) and existing TOC in the water. Daily samples of the existing pH, alum doses, and chlorine treatments in the water were also collected.

The provided dataset did not include CSMR data. We calculated CSMR by dividing chloride(Cl) by sulfate(SO4) for each entry. We did this separately for both plants, and for both the influent and effluent data at each plant. The data was also missing information about influent anions at the Marston Water Treatment plant from June 2018 onward, but after speaking with Denver Water about this, our team did receive data from 2019 to 2021. There is still a gap in the data from June 2018 to 2019, but having data through 2021 helped our analysis and comparisons to the effluent water at Marston.

EXPLORATORY DATA ANALYSIS

Our exploratory data analysis (EDA) included investigating patterns of CSMR over time, exploring the relationship between CSMR and other important variables, and comparing the CSMR at Marston and Foothills. When looking at CSMR over time, we compared the patterns of influent CSMR to effluent CSMR. This comparison allowed us to discover that influent and effluent CSMR were closely correlated. However, we were originally missing data on influent anions at Marston to complete our EDA. To solve this, we created predictive models to compensate for the missing data. By creating these predictive models, we gained more insight on how CSMR is different at Foothills compared to Marston. For example, CSMR at Foothills seemed to generally be higher and more variable than CSMR levels at Marston. Influent CSMR at Foothills also had an average difference of about 29% compared to influent CSMR at Marston, with a 27% average difference for effluent. Also, at both Foothills and Marston, CSMR levels were steadily dropping over time, indicating that the cleaning processes used were reducing CSMR in the water.

Additionally, when performing our EDA, we encountered a problem regarding the times when the water was sampled. The dates of the sampling did not match when comparing the influent and effluent water at the same treatment plant. To solve this problem, we created new data frames and adjusted all the variables to be on the same time scale. The time scale we chose was a sample once every seven days, which we believed was a reasonable amount of time and a sufficient number of samples to capture the patterns of the influent and effluent variables over the span of our data. Once all the variables were on the same time scale, we were able to understand the relationship between influent CSMR and effluent CSMR as well as their relationship with other variables.

STATISTICAL ANALYSIS and RESULTS

The models we created had a 50/50 split for the training and testing sets with the first half of the data being the training set and the second half the testing set. The goal of these models was to predict effluent CSMR based on influent CSMR and other significant variables.

Linear Regression

In order to predict Effluent CSMR at each water treatment plant, we created linear regression models using the Akaike Information Criterion (AIC) backward stepwise selection method. When applied to the variables for Foothills, we discovered that the significant variables for predicting effluent CSMR are influent CSMR, TOC, alum dosage, and influent fluoride. [See Table 1]

This produced our strongest model, with a high R-squared of 0.96 for our training data, suggesting that the model accurately predicts the actual CSMR values. As shown in Fig 2, we were able to predict much of the variation in effluent CSMR at Foothills. Our model has a mean absolute percent error of 7.83% for the training set and 6.87% for the testing set. The blue line represents where we split our data for training and testing.

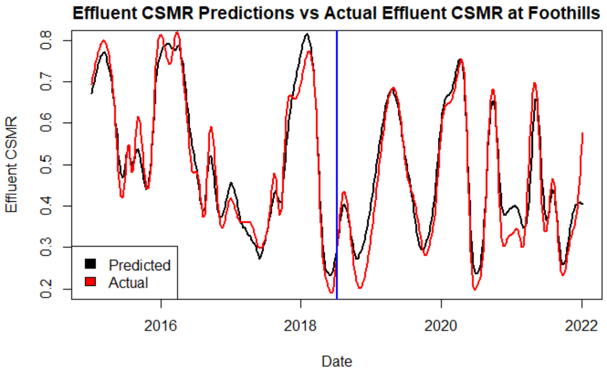

Fig. 2: Linear Regression for Foothills

Foothills:

Predictors	R-Squared Value (Training)	R-Squared Value (Testing)
Influent CSMR	0.87	0.94
Influent CSMR + TOC	0.96	0.94
Influent CSMR + TOC + Alum	0.96	0.94
Influent CSMR + TOC + Influent F	0.96	0.94

Table 1: Foothills Models for Predicting Effluent CSMR

A model was similarly constructed to predict effluent CSMR at Marston. Notably, the AIC backward stepwise selection concluded the variables significant to predicting effluent CSMR at Marston are not the same as the variables for Foothills. The variables deemed significant at Marston are influent CSMR, TOC, alum dosage, influent bromine, influent chlorine, and influent fluoride. [See Table 2]

Our model predicting effluent CSMR at Marston, as shown by Fig 3, has a mean absolute percent error of 8.14% for the training set and 10.44% for the testing set. For the training set, we were able to predict most of the variation in effluent CSMR whereas in the testing set our model struggled to predict the extremities of effluent CSMR. Additionally, Table 2 illustrates the difference in the R-squared values between the training and testing sets with training R-squared values being consistently higher. Table 2 also highlights the beneficial impact of additional variables to the predictive power of the model in the testing set compared to the training set.

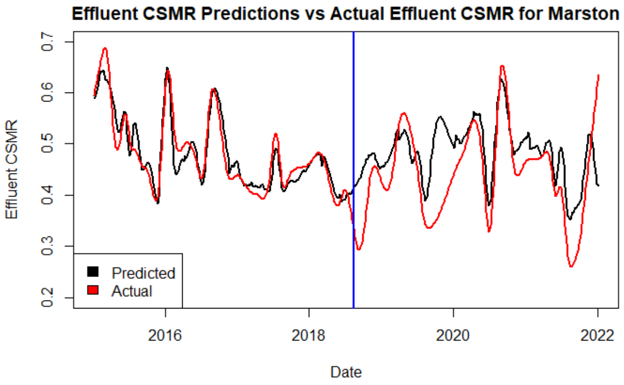

Fig. 3: Linear Regression for Marston

Marston:

Predictors	R-Squared Value (Training)	R-Squared Value (Testing)
Influent CSMR	0.84	0.54
Influent CSMR + TOC	0.86	0.55
Influent CSMR + TOC + Chlorine	0.86	0.68
Influent CSMR + TOC + Chlorine + Influent F + Influent Br + Alum	0.88	0.71

Table 2: Marston Models for Predicting Effluent CSMR

CONCLUSIONS

Denver Water tasked our team with exploring CSMR at the Marston and Foothills plants, its variability, and its relation to other variables in the water treatment process. Our initial exploratory data analysis suggested that the cleaning process does appear to reduce levels of CSMR in the water since effluent CSMR tended to be lower than influent CSMR. It also showed that Foothills tends to have higher CSMR in comparison to Marston.

Our models found that while significant predictor variables differ for the plants, influent CSMR is the greatest predictor of effluent CSMR for both the Marston and Foothills. TOC also seemed to be a variable of interest when making predictions of effluent CSMR. This suggests that TOC and influent CSMR are important to look at in water to gauge effluent CSMR leaving the plant

A logical next step would be to improve models to account for the changes in coagulant used at both plants. We suggest any future work also include an interactive visualization component for Denver Water to use. Another step would be to explore additional lead data provided by Denver Water as it relates to CSMR.

AUTHORS

Nathan Hontz

Currently pursuing a B.S. in Data Science with a minor in Business Administration at Baylor University

Maya Maes-Johnson

Currently pursuing a B.S. in Computational and Applied Math with a minor in Public Affairs

Emily Warwick

Currently pursuing a B.A in Economics and Political Science at Baylor University

ACKNOWLEDGEMENTS

We would like to thank Dr. Doug Nychka, Dr. Amanda Hering, Dr. Tzahi Cath, Dr. Luke Durell, Matthew Hofkes, and Natalie deBonoPaula. We would also like to thank Denver Water for their contribution of the data.