Modernizing Water and Wastewater
Treatment through Data Science
Education & Research

TECH BRIEF

Data Science Summer Fellows Program

Summer 2021

Denver Water – Evaluating Taste and Odor

Logan Blackwell, McLennan Community College
Amber Rexwinkle, Colorado School of Mines
Yang Zhang, Baylor University

SUMMARY

In water treatment for potable use, it is crucial to consider the final product's taste and odor; an odd taste or smell will limit customer confidence in the quality and safety of the water they are drinking. One primary compound responsible for taste and odor events in reservoirs is geosmin. Geosmin is produced by algae and bacterium and causes water to take on an earthy or musty smell or flavor. The Marston Forebay Reservoir serves Denver, CO with potable water and has seen increasing rates of algal blooms since 2012. This Tech Brief focuses on exploring the relationship between geosmin concentrations and basic water quality parameters and suggests methods to predict geosmin through statistical modeling.

INTRODUCTION

The Marston Forebay has had problems with taste and odor (T&O) compounds detectable to humans. The chief compound responsible for T&O episodes in this reservoir, geosmin, is detectable at concentrations as low as 5 ng/L. Geosmin is not directly harmful to humans, but the presence of the unpleasant taste or odor that it causes diminishes the consumer's confidence regarding the quality and safety of this drinking water.

Testing for geosmin can take up to 48 hours, during which the treatment plant, Denver Water, may serve millions of gallons of drinking water to its customers. However, removing all geosmin from water is a very expensive process, and treating for geosmin is inefficient and uneconomical when elevated concentrations are not present. Because of this, Denver Water is interested in investigating the relationships between easily measured physical and chemical water quality parameters (temperature, pH, etc.) and geosmin, and leveraging those relationships as possible indicators of increased T&O compound concentrations.

FACILITY SYSTEM DESCRIPTION

Denver Water uses Marston Reservoir to capture mountain runoff before being sent to the water treatment facility. Several changes have been made regarding the water collection and testing processes at Marston Reservoir over the past several years. These changes include changes in sampling locations and the installation of a Multilevel Outlet Tower, which allows water to be drawn and blended from various depths [1], and. In 2008, a speece cone (a mechanism to add oxygen to the water) was installed in the center of the reservoir. Three different types of sampling methods occur at Marston Reservoir: automatic sonde measurements, manual sonde measurements, and grab samples. The type, location, and time frame of measurements vary.

DATA DESCRIPTION

Data has been collected at various sites throughout the reservoir from 2012 to 2020. Our team was provided with seven data sets, five containing only water quality data, one containing T&O-related data and nutrient levels, and one containing both water quality and T&O data. Sampling frequencies, methodology, and location varied considerably across data sets. For example, some data sets contained full vertical profiles of the reservoir whereas others contained measurements taken at a fixed depth. Some data sets contained hourly measurements and others contained monthly measurements, and one dataset measured depth from the bottom to the surface rather than vice versa. Additionally, one data set completely lacked depth measurements.

The original data contained measurement errors, likely due to faulty sensors. For example, two dates in late 2020 measured depths greater than 300 meters and recorded pH values well above the standard limit of 14, neither of which are possible in the Marston Forebay.

EXPLORATORY DATA ANALYSIS

Our exploratory data analysis (EDA) centered on investigating visual relationships between geosmin and common water quality parameters. The first, and most significant, issue we encountered was a lack of synchrony between geosmin measurements and water quality measurements. To solve this, we matched geosmin measurements to the water quality measurement taken at the nearest point in time. Because we did not know how stable the water quality parameters were over time, we chose to create five datasets containing geosmin and water quality measurements taken within 30 minutes, 1 hour, 2 hours, 12 hours, and 24 hours of each other, respectively. Furthermore, due to the lack of depths in one data set, we averaged all individual profiles of the reservoir across their depths.

Another issue we encountered in our EDA was that high geosmin levels strongly distorted our graphs, making it impossible to see any clear relationships. To solve this, we used the logarithm of the geosmin values in our plots. However, this still failed to present any clear relationship between the variables. We proceeded to use boxplots to explore possible seasonality of water quality components and/or geosmin. The results of this were generally inconclusive, though we were able to observe that specific conductivity and dissolved oxygen measurements seemed to reach their local extreme values during months when geosmin levels were at their highest.

STATISTICAL ANALYSIS and RESULTS

All of the models were created using stratified random sampling with a 75%/25% test/train split. To begin exploring the relationship between geosmin and water quality measurements, we defined a geosmin event as any geosmin reading above 5 ng/L. Water quality parameters used for modeling are specific conductivity (uS/cm), dissolved oxygen (DO) (mg/L), pH, temperature (°C), and turbidity (NTU). Each modeling method that we employed was repeated for each of the 5 time frames we split the data into.

In order to determine model performance, we collected three different metrics; (1)accuracy, the proportion of correct predictions, (2)sensitivity, the proportion of samples correctly predicted to be below 5 ng/L, and (3)specificity, the proportion of samples correctly predicted to be above 5 ng/L.

Linear Regression

Our first attempt to model geosmin employed multiple linear regression. We used the test data set to predict geosmin levels, and then classified the results as events or non-events to determine model accuracy. As shown by the low values of R-squared and accuracy in Table 1, this model's performance appeared inadequate, which led us to explore non-parametric models.

Time Frame	∓30 min	∓ 1 hr	∓2 hr	∓ 12 hr	∓ 24 hr
Multiple R-squared	0.016	0.015	0.021	0.035	0.043
Accuracy (%)	48	42	38	48	43

Table 1: Evaluation of Multiple Linear Regression Performance

Random Forest Regression

Our second attempt was to build a Random forest regression model. Random forest regression uses a bootstrapping sampling technique, where subsets of observations are created from the original data set, and then each subset uses a subset of predictor variables to create a decision tree. Our model used 5,000 decision trees and then averaged the outputs to find the most likely result. Figure 1 shows the logarithm of the measured geosmin and the predicted log geosmin for the 30 minute data, as well as the boundaries for classification.

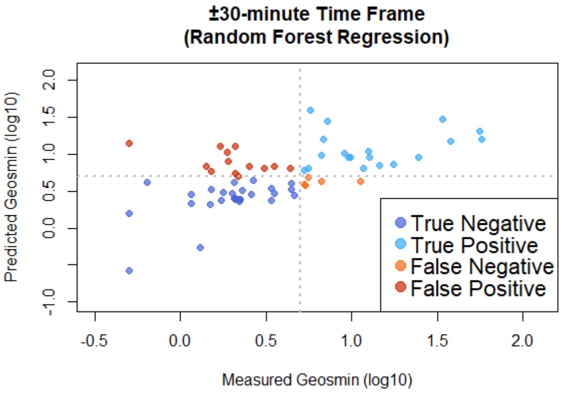


Figure 1: Random Forest Regression Performance

Table 2 shows the accuracy of the results. This model sees a significant increase in accuracy compared to the multiple linear regression model, but it still fails to classify nearly 1 out of every 4 events correctly.

Time Frame	∓30 min	∓ 1 hr	∓2 hr	∓ 12 hr	∓ 24 hr
Accuracy (%)	77	76	76	74	75

Table 2: Random Forest Regression Performance

Random Forest Classification

The final model we chose to use was a random forest classifier. Random forest classification is set up the same as regression, but rather than averaging the output values, the classification that is predicted the most across all trees is chosen as the result. Our model predicted whether a value of geosmin was an event, over 5 ng/L. Table 3 shows the model fits well overall, as proven by the high accuracy, sensitivity, and specificity values.

Time Frame	∓30 min	∓ 1 hr	∓2 hr	∓ 12 hr	∓ 24 hr
Accuracy (%)	97	90	93	92	93
Specificity (%)	98	85	95	93	95
Sensitivity (%)	96	100	90	88	88

Table 3: Evaluation of Random Forest Classification Performance

Based on the random forest classification model, Figure 2 shows that conductivity and DO are the most important variables to be considered when predicting geosmin. Temperature and turbidity are relatively less important variables to consider, however all of the water quality parameters are significant to the accuracy of the model.

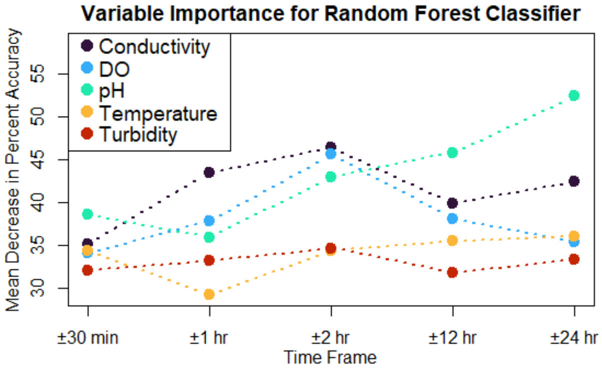


Figure 2: Random Forest Classifier Variable Importance

Across all models, the 30 minute time frame resulted in the highest prediction performance. There was not a noticeable breaking point at which the time elapsed became too large to create a beneficial model.

We created an interactive Shiny app dashboard to visualize the data and results of our models. The dashboard consists of 5 parts. We included a plot of the water quality data and allowed the variables to be visualized over time and subset by year. We then look at the relationship between the variables at each of the 5 time frames by creating plots, summary statistics, and allowing a comparison between two plots. We next plot the relationship between variables from each sampling location, allowing a smoothing feature and a third variable to allow an additional plotting dimension. Our last visualization compares the variables at different locations in the reservoir, allowing for a comparison between grab data and profile data. Lastly, we used our two random forest models to provide a section to forecast geosmin based on user-input water quality measures. The prediction section provides a predicted geosmin measurement (ng/L) and whether or not it is predicted to be an event.

CONCLUSIONS

The models we produced indicate that there is a relationship between geosmin concentrations and basic water quality parameters. Although the precise nature of this relationship remains unclear due to the non-parametric methods we used, the Marston treatment team should be able to leverage water quality data to anticipate and prepare for taste and odor events.

We recommend synchronous measurements of water quality at all the times that geosmin samples are recorded. This allows for less variance caused by the lag between measurements, and allows trends to be purely based on relationships with geosmin and water quality. In addition to the synchronous time measurements, we advise taking all the measurements at various locations across the reservoir to account for possible heterogeneity of the water. Lastly, we would recommend measuring additional easily attainable water quality parameters, such as ORP, depth, chlorophyll at all of the locations and times, that can be included in the modeling of the data. In order to avoid making false assumptions with N/A measurements, we were not able to include several parameters that could have been important in modeling geosmin levels.

Overall, the models produced did well in predicting the values of geosmin. Though the models are still inclined to predict false positives, they are highly accurate and provide a guideline as to when Denver Water should be cautious of high geosmin levels in the water and can be used as a notification to test geosmin more frequently.

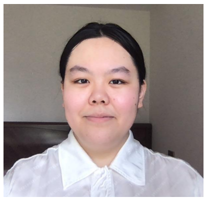
REFERENCES

[1] Adams, Jay. New 46-Foot Tower Makes a Big Splash at Marston. Denver Water, 3 Sept. 2015, www.denverwater.org/tap/new-46-foot -tower-makes-a-big-splash-at-marston

AUTHORS

Logan Blackwell is currently a student at McLennan Community College, and he intends to get his B.S. in Environmental Geoscience at Texas A&M University.

Amber Rexwinkle is getting her B.S. in Statistics and is starting her M.S. Statistics at Colorado School of Mines. In her free time, Amber loves to play with her dog, Olive, and cook.



Yang Zhang is studying Mathematics and Economics at Baylor University. In her free time, Yang loves to play the piano and cook.

ACKNOWLEDGEMENTS

We would like to thank Dr. Amanda Hering, Luke Durell, Tayler Elwell, Hattie Means, Dr. Tzahi Cath, Sweta Rai, Aurora Waclawski, and Dr. Douglas Nychka. We would like to also thank Denver Water for providing the data to further improve our data science skills.