Modernizing Water and Wastewater Treatment through Data Science Education & Research

TECH BRIEF

Data Science Summer Fellows Program

Summer 2021

Southern Nevada – Classifying Experimental, Contaminated Water Events in Lake Mead

Diego Curbelo, Baylor University
Ashley Gray, Colorado School of Mines
Sierra Shetler, Baylor University

SUMMARY

Southern Nevada Water Authority (SNWA) monitors the water quality of Lake Mead and ensures contaminants are efficiently removed before distribution. Thus, our main objective was to create a classification model that utilizes water quality (WQ) parameters as indicators of a water contamination event: tertiary wastewater effluent, stormwater runoff, or dry-weather runoff. By conducting exploratory data analysis and using random forest variable importance features, we developed classification models using random forest and k-nearest neighbors (KNN) methods that produced high accuracy, sensitivity, and specificity. Additionally, a real-time monitoring system was implemented. These results give SNWA forewarning to treat and monitor periods of contaminated water.

INTRODUCTION

Supplying drinking and irrigation water to over 40 million homes, Lake Mead serves as the main source of water consumption for 90% of the Las Vegas Valley; therefore, maintaining the water quality of Lake Mead is imperative. Creating classification models is important in identifying water contamination events quickly so SNWA may efficiently treat the water. To create these models, data were collected during a controlled experiment that simulated events by adding contaminants to the untreated water. By analyzing variable importance over nine different water quality characteristics, we created classification models that determine when a contamination event is occurring versus when the water is at normal quality levels. A monitoring system was also developed to add a real-time component to the classification model's performance. This system marked a given observation when the combination of WQ variables was unusual with respect to the normal conditions expected.

FACILITY SYSTEM DESCRIPTION

The water for this experiment was sampled from the drinking water intake deep in the center of Lake Mead. A pipe travels from the intake location of Lake Mead to the treatment plant. A slipstream from the water intake line provided water from Lake Mead for the experiment.

DATA DESCRIPTION

A detailed description of the experimental design used to collect the data can be found in the STAR Lab Phase I Manuscript.

The data set included 1,561 total observations. Arranged in a data frame with thirteen columns, each observation included its pH level, oxidation reduction potential (ORP) (mV), temperature (F), turbidity (NTU), conductivity (µS/cm), algae (µg/L), total organic carbon (TOC) (mg/L), UVA254 (1/cm), and unitless tryptophan-like fluorescence. The first eight days of the data were designated as the training set, with the remaining eight days as the testing set. As suggested by SNWA, measurements taken while the plant was under maintenance were removed from the dataset.

EXPLORATORY DATA ANALYSIS

Our exploratory data analysis focused on determining each WQ variable's importance in classifying water contamination events, with the goal of selecting the optimal variable combination to yield the strongest model. Maintaining high accuracy, specificity and sensitivity were crucial in developing models that correctly determined when a contamination event was occurring and rarely misclassified normal water as contaminated or vice versa (i.e. a false positive or false negative classification). However, it is most important to maintain a high specificity because too many false positives could prove to be disruptive to the SNWA plant operators. All exploratory data analysis was performed exclusively with the training data.

In order to establish a hierarchy of importance among our nine WQ variables, the correlation between each pair of variables was analyzed with scatter plots. TOC was highly positively correlated to UVA254 (0.9551), and temperature was negatively correlated to both pH (-0.4411) and oxidation reduction potential (-0.8319).

Each variable was plotted over time to compare their values during a water contamination event with those of normal water. Observations with normal water quality were connected with a black line, and each event type was plotted with a color-coordinated dot point to juxtapose the event versus normal water quality variable characteristics. Through this exploratory analysis, we determined that UVA254 (see Figure 1), turbidity, and pH are likely to be very important in classifying events. Their levels remained relatively stable during normal observations and spiked dramatically during water contamination events. This exploratory analysis served as a useful starting point in creating meaningful statistical models.

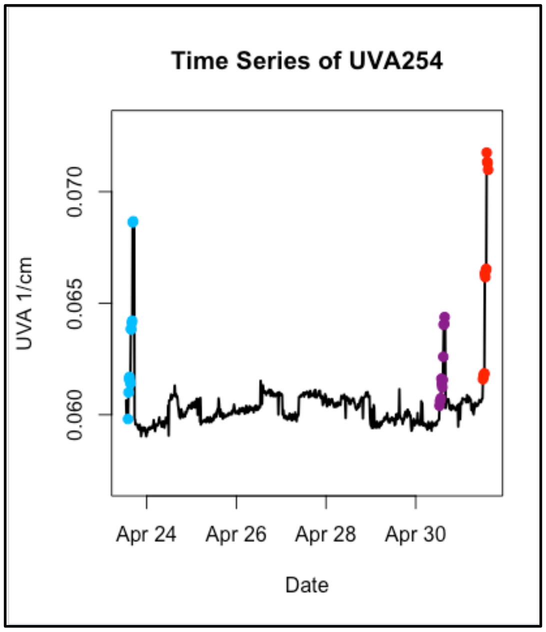

Figure 1. *UVA254 concentration plotted over time shows a small range of values for normal water situations versus spikes during water contamination events; blue points are tertiary wastewater effluent, purple points are stormwater runoff events, and red points are dry-weather runoff events.*

RANDOM FOREST

We sought to build binary classifiers for two levels: contaminated water event vs. non-event. A random forest (RF) model was chosen because of its excellent performance across a wide range of classification and regression predictive modeling problems in machine learning. Random forest is a classifier that contains a number of decision trees on several bootstrapped subsets of the given dataset. An average is taken of these decisions in order to improve the predictive accuracy of that dataset.

Utilizing all nine WQ variables, we compared four combinations of tree counts (501, 1001, 2001, and 3001) and three variable sets per split size (three, four, and five) for a total of twelve unique models. The best performing model used 2001 trees and three variables per split, with an overall accuracy of 99.1%, sensitivity of 91.9%, and a specificity of 99.6%.

Another important contribution from the RF model was the ability to extract measures of variable importance. The most important variables for classification can be assessed by calculating how much accuracy the model would lose when excluding each variable. The most important variables for this classification model were UVA254, pH, and turbidity, with 0.0178, 0.0120, and 0.006 mean reduction in accuracy, respectively. The RF variable importance supports the initial findings of our exploratory analysis.

K-NEAREST NEIGHBORS

Another method that proved useful in event classification was k-nearest neighbors. KNN classifies observations based upon the characteristics of the surrounding observations. A variety of variable combinations and number of neighbors were tested, with varying degrees of success. These include a (i) full model utilizing all nine WQ variables; (ii) three models based upon the important variables identified in the RF analysis; and (iii) a model that was reduced using a combination of Principal Component Analysis (PCA) and selecting variables that minimize collinearity. For each of these models, k-values of one through 100 were tested to find the k that yielded the best result.

The best performing model was the one whose variables were selected through PCA. The variables algae and conductivity were eliminated as unimportant variables. Next, UVA254 was eliminated due to its high correlation with TOC. This resulted in a KNN model composed of the following WQ variables: pH, ORP, temperature, turbidity, TOC, and trypto-interp. Using this variable combination and a k-value of five, this model identified whether or not the water was contaminated with an accuracy of 97.94%, a sensitivity of 97.72%, and a specificity of 97.97%. The sensitivity of this model was higher than the RF model, but the accuracy and specificity were lower.

HOTELLING'S T^2

In addition to classification, a Hotelling's T^2 control chart was built as a real-time monitoring device. The Hotelling's T^2 statistic can be used to signal unusual combinations of variables. Large values of T^2 indicate that a given observation is unusual with respect to the normal conditions expected. To apply this method, a training set that contains no events is used to establish normal conditions. The mean and covariance of the training data is then used to create a threshold for the T^2 value, and any T^2 value above this threshold in the testing set indicates an observation with non-normal water quality.

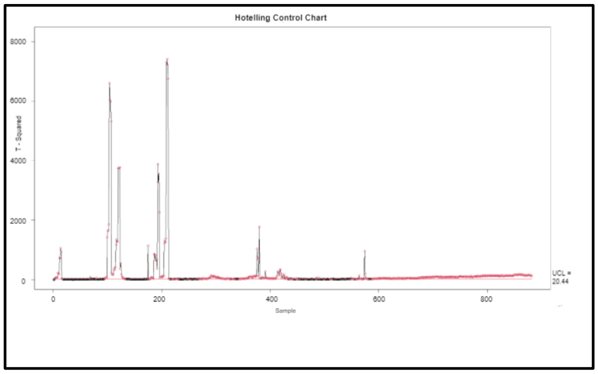

Figure 2. *Hotelling's Control Chart utilizing all nine WQ variables with an alpha value of 0.01.*

This type of method could serve to alert operators at Southern Nevada when the water quality has strayed from its normal state. Three different variable combinations were attempted: (i) all nine WQ variables; (ii) pH, ORP, temperature, turbidity, TOC, and trypto-interp; and (iii) pH, turbidity, and UVA254. Additionally, the alpha parameter, which can increase or decrease the threshold, was set to 0.01, 0.05, and 0.001 for each variable combination.

Of these nine models, those with an alpha of 0.001 performed better than their counterparts with larger alpha values. The model utilizing pH, turbidity, and UVA254 with an alpha of 0.001 performed the best, with an accuracy of 57.71%, a sensitivity of 96.67%, and a specificity of 54.87%. Overall, Hotelling's T^2 does perform well; however, due to the comparatively poor performance compared to our classification methods, and because these data do not meet the method's assumptions of normality and independence, further exploration is needed to find a good real-time monitoring method.

CONCLUSION

With the goal of creating an effective classification model, our RF and KNN models were both successful in classifying normal water versus contaminated water. While the RF model yielded a higher accuracy rate, the KNN model yielded a higher sensitivity. These differences will allow SNWA to weigh their classification needs and choose a model that best fits their system's parameters.

If this experiment were to be repeated, the time necessary for contamination to leave the system should be considered. Practically speaking, some observations were labeled as normal, when in actuality, the WQ seemed to indicate contamination, which led to errors in our models' performance due to a lag time in the data. In developing a model that alerts SNWA of water contamination events in real time, Hotelling's T^2 proved unsuccessful; however applying other multivariate statistical quality control methods to the data would be an intriguing focus for future research.

REFERENCES

Hannoun , Deena, et al. "The Potential Effects of Climate Change and Drawdown on a Newly Constructed Drinking Water Intake: Study Case in Las Vegas, NV, USA ." *Water Utility Journal* , xx, no. xx, 2021. *Xx*.

Hurt, Jon, and Claudio Cimiotti. "Lake Mead Intake No. 3." *Engineering*, vol. 3, no. 6, 2017, pp. 880–887.

Thompson, Kyle and Dickenson, Eric. "Supervised machine learning classification and an array of online instrumentation to detect de facto reuse and urban runoff in surface water." Water Quality Research and Development, Southern Nevada Water Authority, 1299 Burkholder 5 Blvd., Henderson, United States.

AUTHORS

Diego Andres Curbelo is a Data Science student at Baylor University.

Ashley Gray is a student at Colorado School of Mines, majoring in Applied Mathematics and Statistics and minoring in Data Science.

Sierra Shetler is a double major student at Baylor University studying both Mathematics and History.

ACKNOWLEDGEMENTS

Data is provided by Eric Dickenson (eric.dickenson@lvvwd.com) and Kyle Thompson (kyle.thompson@lvvwd.com) of Southern Nevada Water Authority.

Faculty Leadership:

Dr.Hering: Mandy_Hering@baylor.edu
Luke Durell: Luke_Durell1@baylor.edu
Tayler Elwell: tjelwell@mymail.mines.edu
Hattie Means: hattie_means1@baylor.edu