Predicting and Comparing Drinking Water Quality Across California Public Water Systems Using Random Forest Modeling

Bryson Patterson, Bethany O'Shea

Nearly one million Californians get their water from a failing water system. Failures in water treatment are complex but include numerous challenges based on the landscape characteristics of their catchment areas, climatic variables such as wildfires and precipitation, and resources available for treatment.

The goals of this study were twofold: develop a predictive model for the impact of climatological variables on drinking water analyte concentrations and assess the relative performance of public water systems. Data for the predictive variables included daily precipitation data from the PRISM climate group at Oregon State University, catchment-level landscape characteristics from the StreamCat dataset, and fire data from the Monitoring Trends in Burn Severity interagency program. Upstream watershed delineation was made using catchment connectivity and boundary data from the National Hydrography Dataset. All zonal statistics for determining the percentage of each catchment burned and cumulative rainfall between tests for each analyte were performed in R.

To develop the predictive model based on the input data, a Random Forest machine learning algorithm was used in R. This approach enables hundreds of variables to be used in developing a model while reducing the concern of overfitting the model to the training data. To create a baseline by which to compare the performance of public water systems, the most important input variables for each analyte, as determined by the model, were used to match public water systems based on similarity.

By developing this analysis tool with entirely open-source software, future work will be able to continue to enhance the model accuracy with new data as they become available. The ability to identify catchments of particular trouble and predicting the effects of climate change on water quality will enable regulators to focus resources and attention where it is needed now and where it will be needed going forward.

Presenting Author: Bryson Patterson