Problem Statement Formation

Can cyanobacteria abundance be reliably estimated in Lake Fictitious using water quality parameters collected by an i*n situ* probe and can less informative parameters be excluded from sampling, resulting in less time, effort, and cost in the field?

Context

Harmful algae blooms (HABs) occur in lakes across the USA every summer. These HABs, which consist primarily of blue-green algae, are visibly unappealing and can release toxins which can kill fish, waterfowl, livestock, and pets and make humans very ill. Two methods exist for quantifying blue-green algae. The first, sending samples to a specialized lab for identification, is time consuming and the results are usually not received until after the HAB has passed. The other method is to measure phycocyanin, which is a pigment found in cyanobacteria. This sensor is often installed in situ and programmed to continuously collect data. Lake operators, such as a state park, will monitor these readings and shut down recreation on a lake if blue-green bloom is detected. However, these closures are immediate, no warning is available to the public or staff, and all management actions are reactionary. Therefore, lake operators would be very interested in a method to predict blue-green algae in a lake before implementing a shutdown, allowing for planning and allocation of resources.

Criteria for Success

Accurately predict the quantity of phycocyanin and identify the most important predictors of water quality metrics for making these estimates.

Scope of Solution Space

Water quality data, including phycocyanin, was collected at Lake Fictitious every 30 minutes for three years using an AlgaeTracker manufactured by AquaRealTime. This data, in combination with domain knowledge, will be used to develop machine learning models to predict phycocyanin concentrations multiple days in advance. Data will be cleaned, explored, engineered, and modeled. The best model will be chosen and features that contribute the most to the model will be identified. With this model and data, the lake operator can predict the magnitude of a HAB before it occurs which would avoid unnecessary shutdowns. In addition, a reliable model would allow for much less frequent data collection and ultimately, a phycocyanin sensor may not be necessary, thus saving money for the operator.

Deliverables to the client will include a presentation including a slidedeck, a project report with management recommendations, and GitHub repository containing developed models.

Constraints

Some erroneous data (temporarily faulty sensors) exist in the dataset. These values will have to be replaced. Sufficient data exists for machine learning, however, more data would be ideal to

develop a more robust model. Also, nutrients, which have been shown to stimulate algal growth, were not analyzed. Their inclusion would have been beneficial to the model, however, those sensors are very expensive and may contradict a goal of this model.

Stakeholders

AquaRealTime
Lake Fictitious Basin Water Quality Authority (CCBWQA)
Lake Fictitious Homeowners Association

Data Sources

AquaRealTime has provided a login to their data portal for access to Lake Fictitious data.