Competitive Learning to Develop a Biomarker Forecasting Tool for Classifying Recreational Water Quality

Dominic Boccelli

University of Cincinnatti

Problem and Research Objectives

Recreational users of surface waters can be at risk when there are elevated pathogens in the water. In urban areas, such as Cincinnati, non-point source contamination can occur through increased runoff (due to impervious surface area) and direct discharge of storm water, or combined storm water and sewage, into local surface waters. Indicators such as E. coli and fecal coliforms are used as water quality surrogates due to their relative ease of measurement. Unfortunately, complete analysis and data reporting requires, at a minimum, 24 hours, thus limiting the utility of observed data to provide information to the population on water quality aspects in a timely fashion. However, the ability to predict microbial outbreaks in recreational waters would provide engineers, managers, regulators, and public health officials an important tool in disseminating pertinent public safety information in a timely fashion. Data-driven modeling approaches, such as linear regression or artificial neural networks, seek to capture the important forcing factors associated with microbial concentrations within a simple modeling framework. These data-driven models are then used to predict the microbial concentrations, which are then classified with respect to water quality standards. However, these approaches may still suffer from high rates of false-positives and false-negatives regarding classification.

The objective of the proposed study is to develop a Recreation Management Program tool capable of providing water quality classifications to the public regarding the safety of recreational waters. Previous research efforts have focused on quantifying microbial concentrations, prior to classification, using multivariate linear regression or artificial neural networks (used as a “black box” model). The proposed tool utilizes a type of neural network based on self-organizing maps entitled Learning Vector Quantization (LVQ). Rather than estimate the microbial concentration, the tool to be developed will predict the water quality classification directly, thus potentially eliminating the impact of errors in estimating the microbial concentrations. The LVQ approach will be compared to the more “typical” data-driven approaches such as linear regression and neural networks for microbial concentrations with emphasis on comparing the true and false classification rates.

Methodology

The approach for this study has utilized hydrologic and water quality data collected by the Charles River Watershed Association (CWRA) to develop a tool capable of providing a water quality indexing system for recreational water at the Larz Anderson bridge sampling location. CWRA has collected E. coli samples as well as flow and precipitation

data for approximately two recreational seasons (May through October) at multiple locations in the watershed that will be used in model development.

Previous research studies developed models that estimate the microbial indicator concentration first, which is then transformed into a classification. However, these approaches result in measurable false-positive and false-negative rates. Since classification of the water quality is of most importance, the neural network based approach of LVQ is proposed to use the available data to develop a tool that, given the appropriate hydrologic and meteorologic data, will directly produce a classification. This approach removes the reliance on adequate prediction of microbial concentrations.

To adequately compare the performance of the LVQ algorithm to other approaches, equivalent versions of linear regression and artificial neural network (ANN) models based on previous studies will be developed to represent the same data set. For simplicity, the explanatory variables used in the LVQ algorithm will be the same used to develop the linear regression model for the CWRA data (Eleria and Vogel, 2005). However, the dependent variables in the comparative models will be the actual E. coli concentrations with classifications performed after estimation. Comparisons between the different modeling approaches will be made based on the classification characteristics (i.e., true/false positive/negative rates) of each algorithmic approach.

Principal Findings and Significance

The linear regression, ANN, and LVQ modeling approaches have been developed to represent the Larz Anderson bridge monitoring data using E. coli concentrations as the dependent variable, and the antecedent rainfall during the previous 24- and 168-hours and lag-1 E. coli concentration data used as the independent variables. These independent variables were selected based on previous work performed by Eleria and Vogel (2005). The resulting model classifications were evaluated with respect to the ability of three modeling approaches to satisfy a primary and secondary contact recreation standard (200 and 1000 colony forming units/100 mL of sample).

While there are differences in the classifications from each algorithm, each individual algorithm showed little difference when comparing performance associated with the boating and swimming standards. In fact, all three algorithms performed well with respect to the true negative rates (>92% in all cases; equivalent false positive rates <8%) regardless of the standard.

With respect to the linear regression and ANN approaches, the ANN algorithm performed slightly better than the linear regression. The ANN model produced a true positive rate about 10 percentage points higher than the linear regression model (true positive rates for the linear regression were 45%/50% and for the ANN were 52%/62% for the swimming/boating standards, respectively; false negative rates are equivalent to 100-true positive rate). The LVQ algorithm, however, showed significant improvements for representing the true positive rates (82%/87% for the swimming/boating standards, respectively).

These results suggest that the LVQ approach for direct classification is capable of eliminating the uncertainty associated with classifications based on "concentration prediction first, classification second" approaches. This preliminary data analysis is very encouraging and additional studies associated with varying the independent variables may further improve the results.